A Study on the Use of Keywords in a Graph-based Image Caption Generation System

Brandon Birmingham

Supervisor: Prof Adrian Muscat

November 2022

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

L-Università ta' Malta
Faculty of Information & Communication Technology

L-Università ta' Malta

University of Malta Library – Electronic Thesis & Dissertations (ETD) Repository

The copyright of this thesis/dissertation belongs to the author. The author's rights in respect of this work are as defined by the Copyright Act (Chapter 415) of the Laws of Malta or as modified by any successive legislation.

Users may access this full-text thesis/dissertation and can make use of the information contained in accordance with the Copyright Act provided that the author must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the prior permission of the copyright holder.

Abstract

A long-standing goal of Artificial Intelligence is to have agents capable of understanding and interpreting the visual world using natural language. Research at the intersection of Computer Vision and Natural Language Processing is currently booming and the automatic generation of image captions has gained significant popularity. Several ideas and architectures have been proposed to machine generate human-like sentences that describe images, but all fall short of reaching human-level quality. In general, the task of image caption generation involves the selection of salient objects, attributes and relations depicted in the image which are then combined into a natural language sentence. While the state-of-the-art architectures attempt to do this task in one step, this PhD studies the generation of sentences out of a set of discrete keywords. The attentional behaviour of the human brain while processing the visual world inspired this PhD study and has led to the hypothesis that captions can be generated through a relevant set of keywords which are then connected through a path traversal in a knowledge graph derived from a language dataset.
This novel combination acts as a gangboard between the vision and language modalities, where keywords are represented as graph nodes, while the sequence between keywords is reflected by directed edges. As opposed to the current popular end-to-end learning approach, the proposed model reduces the dependency of large scale paired image-caption datasets which are very laborious and expensive to collect. To test this hypothesis, this study develops and studies KENGIC, a Keyword driven and N-gram Graph-based Image Captioning framework which exploits n-gram sequences as found in a given text corpus to construct sub-knowledge graphs for query images. By having a set of predicted image keywords considered as nodes, the proposed system is designed to probabilistically connect these nodes to form a directed graph through overlapping n-grams. The system infers the most likely captions by maximising the most probable n-gram sequences constructed from the predicted keywords. The study, investigates the generation of image captions under different configuration setups based on (a) keywords extracted from gold standard captions and (b) from automatically detected keywords. Both quantitative and qualitative analyses demonstrated the effectiveness of KENGIC. As spatial relations (SRs) are inherently more difficult to be predicted from

whole images due to their highly polysemous, locative, explicit, and ambiguous nature, this research also contributes to the problem of SR prediction by investigating the problem from a multi-label perspective. However, the explicit use of SRs was not found to improve the quality of the generated captions as evaluated on automatic metrics. The performance achieved by KENGIC is very close to that of current state-of-the-art image caption generators that are trained in the unpaired setting. The analysis of this approach could also shed light on the generation process behind current top performing caption generators trained in the paired setting and in addition, provide insights on the limitations of the current most widely used evaluation metrics in automatic image captioning.

To my parents, Lilian and Stephen;
and my girlfriend, Amy

For their love and endless support.

Acknowledgements

This research would not have been possible without the support and contribution of several persons.

First and foremost, I would like to express my sincere gratitude towards my supervisor Prof Adrian Muscat, who guided me throughout this PhD programme. I am so grateful for his expertise which helped me reaching my research ambitions. I thank him for all the support he showed in my proposed work, and for the direction he gave me during the countless meetings we had. I am deeply indebted to the feedback he provided me on my research and writing, and above all, for teaching me how to think.

Many thanks goes to all evaluators who participated in this research. Without their voluntary help, this work would not have been evaluated holistically. I would also like to acknowledge NVIDIA Corporation for donating a TITAN Xp GPU for this research.

Furthemore, I would like to extend my sincere thanks to Dr Charles Caruana and Ms Joan Burlò for proofreading this thesis. Their time and rigorous effort was greatly appreciated.

A special thanks goes to my parents, Stephen and Lilian, who supported me in each and every step of my life and who were indispensable throughout this whole journey. I thank them from the bottom of my heart for their love and for all they have done for me.

I also thank my brother Dylan, my sister Marilyn, and my friends Matthew Sacco and Andrew Sammut for supporting me throughout this journey.

Last but not least, a heartfelt thanks goes to my girlfriend, Amy, for her unwavering support, motivation and unconditional love. She was instrumental not only in helping me concluding this PhD thesis, but also, in fulfilling my life with abundant love, joy and happiness.

“The eye sees only what the mind is prepared to comprehand.”

- Henri Bergson

Publications

Birmingham, B. and Muscat, A. KENGIC: Keyword-driven and N-gram Graph-based Image Captioning. *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Sydney, Australia, November 30 - December 2, 2022. IEEE.

Birmingham, B. and Muscat, A. Multi spatial relation detection in images. *Spatial Cognition & Computation*, 22(3-4):293–327, 2022. doi: 10.1080/13875868.2021.1957897.

Birmingham, B. and Muscat, A. Clustering-based model for predicting multi-spatial relations in images. In *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics, ICINCO 2019 - Volume 2, Prague, Czech Republic*, July 29-31, 2019, pages 147–156, 2019. doi: 10.5220/0008123601470156.

Birmingham, B., Muscat, A., and Belz, A. Adding the third dimension to spatial relation detection in 2d images. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 146–151, 2018.

Birmingham, B. and Muscat, A. The use of object labels and spatial prepositions as keywords in a web-retrieval-based image caption generation system. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 11–20, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2002.

Muscat, A., Belz, A., and **Birmingham, B.** Exploring different preposition sets, models and feature sets in automatic generation of spatial image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, Berlin, Germany, August. Association for Computational Linguistics*, pages 65–69, 2016.

Belz, A., Muscat, A., **Birmingham, B.**, Levacher, J., Pain, J., and Quinquenel, A. Effect of data annotation, feature selection and model choice on spatial description generation in French. In *Proceedings of the 9th International Natural Language Generation Conference*, pages 237–241, Edinburgh, UK, September 5-8 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-6639.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence . . . 1
AMT Amazon Mechanical Turk . . . 62
ANN Artificial Neural Network . . . 4
ANOVA Analysis of Variance . . . 85
ASG Abstract Scene Graph . . . 44
AVS Attentional Vector Sum . . . 22
BLEU Bilingual Evaluation Understudy . . . 63
BRNN Bidirectional Recurrent Neural Network . . . 37
C-GAT Conditional Graph Attention Network . . . 58
CIDEr Consensus-based Image Description Evaluation . . . 65
CNN Convolutional Neural Network . . . 14
COCO Common Objects in Context . . . xiii
CRF Conditional Random Field . . . 25
CV Computer Vision . . . 1
D-CNN Deep Convolutional Neural Network . . . 39
DCCA Deep Canonical Correlation Analysis . . . 19
DETR DEtection TRansformer . . . 18
DGDN Deep Generative Deconvolutional Network . . . 41
DL Deep Learning . . . 4
DNN Deep Neural Network . . . 4
DPM Deformable Part-based Model
DT-RNN Dependency Tree Recursive Neural Network . . . 38
EM Expectation-Maximisation . . . 19
FCL Fully Connected Layer . . . 17
FPN Feature Pyramid Network . . . 17
G-CNN Grid-based Convolutional Neural Network

GAN Generative Adversarial Network 45
GAT Graph Attention Network 44
GCN Graph Convolutional Network 8
GNN Graph Neural Network 44
GPU Graphics Processing Unit
GRU Gated Recurrent Unit 7
HCP Hypotheses-CNN-Pooling 20
HMM Hidden Markov Model 33
HOG Histogram of Oriented Gradients
HOI Human Object Interaction 58
HSD Honest Significant Difference 120
ILP Integer Linear Programming 30
IOU Intersection Over Union 56
KCC Kernel Canonical Correlation 31
KD Knowledge Distillation 54
KENGIC Keyword-driven and N-Gram Graph-based Image Captioning 8
KG Knowledge Graph 67
kM-C k-Means Clustering 143
LAD Language Acquisition Device 2
LSTM Long-Short-Term Memory 7
m-CNN multimodal Convolutional Neural Network 39
m-RNN multimodal Recurrent Neural Network 40
MAN Mutual Attention Network 61
METEOR Metric for Evaluation of Translation with Explicit Ordering 64
MIL Multiple Instance Learning 19
ML-NN Multi-label Neural Network 150
MLE Maximum Likelihood Estimation 42
MT Multimodal Transformer 67
N-GG N-Gram Graph 77
NLG Natural Language Generation 32
NLP Natural Language Processing 1
NTM Neural Turing Machine 37
POS Part-of-Speech xiii
PVT Pyramid Vision Transformer 18

R-CNN Region-based Convolutional Neural Network 16
R-FCN Region-based Fully Convolutional Network
ReLu Rectified Linear Unit 14
ResNet Residual Network 61
RF Random Forest 123
RL Reinforcement Learning xiii
RNN Recurrent Neural Network 7
ROI Region of Interest 17
ROUGE Recall-Oriented Understudy for Gisting Evaluation 64
RPN Region Proposal Network 21
SCS Semantic-Constrained Self-Learning 61
SDG Scene Description Graph 46
SG Scene Graph 67
SGAE Scene Graph Auto-Encoder 53
SGAE-KD Scene Graph Auto-Encoder with Knowledge Distillation 69
SGC Scene Graph Captioner 52
SGD Stochastic Gradient Descent 15
sGPN sub-Graph Proposal Network 55
SIFT Scale Invariant Feature Transform
SSD Single Shot MultiBox Detector 17
SPICE Semantic Propositional Image Caption Evaluation 66
SPPNet Spatial Pyramid Pooling Network 17
SpRL Spatial Role Labelling 123
SR Spatial Relation 11
SRN Spatial Regularisation Network 20
SVM Support Vector Machine 16
TF-IDF Term Frequency-Inverse Document Frequency 65
UIC Unpaired Image Captioning 59
UP-DETR Unsupervisedly Pre-train object DEtection with TRansformers
VDG Visual Dependency Grammar 24
VDR Visual Dependency Representation 24
VGG Visual Geometry Group
VOC Visual Object Classes 13
VQA Visual Question Answering 171

VRD Visual Relationship Detection 170
XML Extensible Markup Language
YOLO You Only Look Once 17

List of Symbols

n	n-gram size
$\mathcal{N}_n$	Set of n-grams of size n
I	Image
$\mathcal{K}$	Keywords
G	Graph
$G_{I,K}$	Graph G for image I based on keywords K
G_n	n-gram graph
T	Text Corpus
k	Number of parents for each node found in graph G
$\mathcal{P}$	The set of parents of a node in graph G
p	Parent $p \in \mathcal{P}$
h	Number of hops
$\mathcal{Q}$	List of paths traversed in graph $G_\mathcal{K}$
q	Path $q \in \mathcal{Q}$
q_n	Total number of paths considered during graph traversal
c	Child node in graph G_K
n_2	n_2-gram size used during optimisation
e_f	Minimum edge frequency in graph G_K

o_p	Optimal paths considered during graph traversal
m	Number of matched keywords in the generated caption
N	Number of extra nouns mentioned in a generated caption
l	Caption length
v	Vocabulary set used in Multi-label training
$\mathcal{V}$	Set of all vocabularies used in Multi-label training
w	Top frequent words used in vocabulary v

1 Introduction

Following the pioneering vision of the English computer scientist, Alan Turing (1912-1954) in which "machines will eventually compete with men in all purely intellectual fields" (Turing, 1950), the Artificial Intelligence (AI) research community is still exploring ways of enabling machines to understand the intersection between vision and language. Researchers from both Computer Vision (CV) and Natural Language Processing (NLP) communities are currently narrowing the gap between the two domains in order to find solutions for machines to understand the visual world and to describe it using natural language. The long term goal is to enable human-to-machine interaction and to provide assistance for the visually impaired through spoken feedback (Lu et al., 2018).

1.1 Language

The human language is one of the most mysterious and dynamic natural systems that the human beings inherit naturally and which keeps evolving through each generation to serve as a communication channel between them (Hauser et al., 2014). From a very young age, infants unconsciously start formulating the building blocks and the linguistic structure needed for a verbal communication system by means of a continuous and natural flow of linguistic information coming from their senses of hearing (sound) and sight (vision). By parsing sequential utterances and associating parts of speech to the corresponding visual domain, infants instinctively start making sense of their surrounding world through their visual and auditory perception, while gradually keep broadening and enriching their vocabulary. During this phase, infants can be compared to young statisticians, fine-tuning a distributional and statistical model which maps co-occurrences as found between vision and language modalities. Infants continue developing their language skills and eventually begin to interact with others and explain their thoughts and emotions by fine granular sentences, such that others can contextually visualise and understand the content of their spoken messages (Peters, 1983). This ability makes the human being unique and distinct from all other species found on this planet (Dunbar, 2009). Although

L-Università ta' Malta

University of Malta Library – Electronic Thesis & Dissertations (ETD) Repository

The copyright of this thesis/dissertation belongs to the author. The author's rights in respect of this work are as defined by the Copyright Act (Chapter 415) of the Laws of Malta or as modified by any successive legislation.

Users may access this full-text thesis/dissertation and can make use of the information contained in accordance with the Copyright Act provided that the author must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the prior permission of the copyright holder.

the neuroscience community provides both insights of how language is processed in the human brain and clues about how humans acquire language (Kemmerer, 2014), language acquisition still remains a mystery and so far only theories from psycho- and cognitive linguistics shed some light on how this phenomenon occurs (Bowerman et al., 2001; Vetter and Howell, 1971).

1.2 Language Acquisition

Although the nature versus nurture debate continues to dominate in the psycholinguistics field, it still remains unclear how the first language originated and how languages continue passing from one generation to another with limited supervision. Up till now, only theories and some supporting evidence shed light on how this complex biological process occurs. The mentalist theorists hypothesise that language is acquired genetically, similar to any other human cognitive ability (Chomsky, 1995; Lenneberg, 1967), while the environmentalists argue that language is learned through an interactive process (Skinner, 1985). Although, to date, no exact and complete evidence has been found of how this perplexing process takes place, current neuroscience-based research is providing evidence that supports earlier proposed theories (Kuhl, 2010). One of the earliest theories was proposed by Skinner (1985), who was one of the pioneers of Behaviorism. The behaviorists theorise that infants acquire language based solely on behavioural reinforcement principles, such that when infants correctly associate words with corresponding meanings, they are positively reinforced by their caregivers. Through the eyes of the behaviorists, as infants continue being rewarded for their correct associations, they will be naturally constructing and adapting to the linguistic structure for their native language. This theory, however, fails to adequately explain how the human language is perceived and comprehended by the human brain. The nativist theorists seek to answer this by arguing that language is a biological instinct which was created by a single chance mutation in one individual about 100,000 years ago. This triggered the language faculty in the human brain or the so-called Language Acquisition Device (LAD), as proposed by one of the pioneers in the field of psycholinguistics and the father of modern linguistics, Chomsky (1995). This theory also hypothesises that humans are born with a universal grammar, consisting of structural rules and instinctive knowledge that serve as a backbone for language acquisition, even in cases where infants have limited language exposure. For the nativists, this is the primary reason why infants are in a position to acquire language at a very fast pace. The mentalists see language as an innate phenomenon for all human beings, while the cognitive theorists argue that language can be seen as another aspect of a child's overall intel-

lectual development and behavior that involves thinking and reasoning. Conversely, the interactionists (Tomasello, 2003; Vygotsky, 1978) combine ideas from sociology and biology, and emphasise the point that language emerges from the involved social interaction between infants and the outside world in the form of communication. This environmentalist approach prioritise the social interaction over the reward-based approach of the Behaviorists. In fact, Tomasello (2003), one of the critics of Chomsky's universal grammar, rejected the idea of an innate universal grammar and instead proposed a functional theory of language development, which revolves around the importance of semantic and communicative functions of language rather than the linguistic structure. Furthermore, the social-pragmatic theory of language acquisition states that children in reality do not need any specific linguistic constraints to learn words, but most importantly, they need flexible and powerful social-cognitive skills which allow them to understand the communicative intentions of others in different interactive scenarios (Tomasello, 2000). These opposing theories are merged together by the Emergentist theory which states that language is a product of several sources combined by internal associations working together simultaneously (MacWhinney, 1998). This theory bridges the disconnection between the nature versus nurture debate which essentially divides the nativists and social interactionist perspectives. The emergentist approach states that language acquisition cannot be reduced to being the result of isolated and independent causes, but emerges from complex and dynamic interactions between an infinite number of variables, which include but are not limited to, our genetic, neurological wiring and social-psychological evolutionary variables. Thelen and Smith (1996) argue that language is an evolution of a system which starts from an initial unstable phase and moves to higher levels of organisation, each with successive and better stability. To complement the Nativist theory, the psycholinguist Elman (1995), suggested that innateness can also takes place on several levels and in an emergent approach. Rather than being born with a pre-wired structure, ready to absorb language, Elman (1995) suggested that innateness itself emerges from several constituting units, including but not limited to, the architecture of brain cells and the arrangements of neurons in the cortical and subcortical areas of the brain. Also, Emergentism sees the importance of the distributional and statistical extraction by infants as a key characteristic for the comprehension of words and grammar. Research in this area confirms that language learning is heavily dependent on the amount and quality of language input, while highlighting the importance of engagement, processing and understanding of the same language input. As suggested by this theory, the input in language development cannot be simply viewed as passive speech bombardment addressed to infants, but on the contrary, children must actively process the input speech. From the emergentist point of view, children learn languages by continuously adjusting their learnt structure each

time their speech output differs from the input. This perspective is the inspiration behind Connectionism (Ellis, 1998).

1.3 Connectionism

The idea of having a system which automatically adjusts its structure and tunes its response based on the input it receives and the output it generates, was the primary foundation behind the connectionist approach. Connectionism is a subfield in cognitive science that aims to understand and study the complexity of human cognitive abilities with computational models that are inspired by the architecture of the human brain (McCloskey, 1991). These models are commonly referred to as Artificial Neural Networks (ANNs) which are composed of connected nodes to represent nodes and synapyses of the human brain. This is intended to emulate the signal propagation which occurs in brain synapses during cognitive tasks. This architecture was proposed to automatically infer statistical mappings between input and output training samples via a computational approach that continuously minimises the difference between the generated and the actual output for a given input sample. This approach has been used since the invention of the "Perceptron" by Rosenblatt (1958), but it was not until recently that it made major breakthroughs. With the sheer amount of data and high computational power available, learning complex non-linear functions for non-trivial problems has become possible in the last decade through the use of Deep Neural Networks (DNNs). This has marked the beginning of the Deep Learning (DL) era which is currently dominating the field of AI with state-of-the-art results across many research areas and applications, including one of the current popular and challenging problems of image caption generation (Wang and Chan, 2019).

1.4 Image Caption Generation

"Imagine, for example, a computer that could look at an arbitrary scene, anything from a sunset at a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does to finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers".

– David Stork, *HAL's Legacy (2001)* on Azriel Rosenfeld's vision (Stork, 1997).

This quote, which is attributed to Azriel Rosenfeld (1931-2004), who is one of the pioneers in Computer Vision (CV), highlights how difficult it is to computationally model

the human ability of seeing, understanding and describing the visual world using natural language descriptions. The involved complex visual recognition and the way visual perception is translated and communicated by natural language are intrinsic to most human beings. Through an incremental learning experience, most individuals can effectively and naturally describe any visual content using natural language. This process is the result of accumulated knowledge which combines visual perception and human language originating from word learning ability during language acquisition. Children eventually learn how to communicate with others and to provide meaningful representations of our world we live in through a natural and endless learning experience that continuously matures, improves and adapts. Although this is an effortless task for most human beings, automatically generating descriptions for visual content, has been one of the long-standing ambitions in Computer Vision (CV) since the late 1960s. At that time, it was believed that this could be simply achieved in a summer project by engineering a system that can perform automatic background and foreground segmentation, and extraction of non-overlapping objects from real-world images (Papert, 1966). After more than 50 years, we are still far from realising this dream.

Automatic visual understanding is performed on digital images which are encoded as large matrices of numerical data representing colour intensities at each single point. From these thousands or even millions of colour-coded pixels, computer vision algorithms are designed to transfer patterns of pixel values into semantic meanings that effectively describe the content of images. The recognition of objects in images which plays a vital role for visually analysing an image has seen major and rapid advancements in recent years. In fact, current state-of-the-art object recognition models built on top of deep neural architectures, are now capable of classifying thousands of object classes with human-comparable accuracy (Wang et al., 2014). While object detectors can be portrayed as the fundamental building blocks for automatic image captioning, such detectors can only produce descriptions as a laundry list of object categories which pale in comparison with the linguistic structure and naturalness of human descriptions. In contrast, humans through their active and exploratory eye gazing mechanism can select the main important aspects of scenes after recognising the objects (e.g., “person”, “boat”), their attributes (e.g., “wooden”) and how these objects relate to each other through verbs (e.g., “painting”) or prepositions (e.g., “next to”). After these composing entities are selected, humans instinctively use their high-level knowledge to transform these entities into well formed sentences that follow a particular language grammar.

1.5 Research Problem

The problem of automatic generation of concise natural language descriptions for images has gained huge popularity in both academia and industrial key players (Bernardi et al., 2016). The conventional process of automatically describing an image fundamentally involves the visual analysis of the image content such that a succinct natural language statement, verbalising the most salient image aspects, can be generated. The generation of image captions also depends extensively on natural language generation methods for constructing linguistically and grammatically correct sentences. Describing image content is very useful in applications for image retrieval based on detailed and specific image descriptions, including caption generation for enhancing the accessibility of existing image collections, for enabling human to robot interaction and more importantly, as an assistive technology for the visually impaired (Kulkarni et al., 2013b).

Arguably, the most difficult aspect in image description generation is bridging the gap between image visual analysis and its corresponding linguistic description. In fact, this task is an emerging research initiative which attempts to provide more relevant and human-like image descriptions (Yang et al., 2011). Numerous approaches have been applied in an attempt to tackle the complexity behind this problem but no single solution has yet been identified that matches human quality. First generation systems caption images by either reusing sentences from collections of image and sentence pairs (Mason and Charniak, 2014; Ordonez et al., 2011), or else by constructing sentences through direct generation pipeline, where sentences are constructed by using templates, grammars or hard-coded rules (Kulkarni et al., 2013a; Yang et al., 2011). With the help of recent advancements made in computational power and with the sheer amount of multimedia content available, Deep Learning (DL) paved the way for the second generation systems (Vinyals et al., 2015; Xu et al., 2015). These models are developed to automatically learn the mapping between images and corresponding sentences via deep neural architectures. These are trained end-to-end on large scale paired image-caption datasets which can be very expensive and laborious to collect. These neural-based models are designed to learn sequential linguistic patterns as conditioned by corresponding visual information. However, evidence shows that these models can be biased in the linguistic domain (Hendricks et al., 2018) and hence limit the compositionality and naturalness for less frequent scenarios (Nikolaus et al., 2019). This often results in syntactically correct descriptions but with semantically irrelevant captions and which lack diversity. Such models are also limited in terms of their explainability and their applicability across different domains.

The encoder-decoder framework (Vinyals et al., 2015) is the core pillar underpinning

(a) Boat on a lake *with* a Christmas tree (img).

(b) Boat on a lake *near* a Christmas tree (Meyering, 2020).

Figure 1.1: Two captions generated by KENGIC for the given keywords: "*boat*", "*lake*", "*Christmas*", "*tree*".

almost every deep learning based image caption generator. The encoder part is responsible for encoding the image into a fixed sized feature vector to represent the most salient aspects of the image. This is normally carried out by the use of Convolutional Neural Networks (CNNs). On the other other hand, recurrent-based neural networks, such as Recurrent Neural Network (RNN) (Mikolov et al., 2010), Gated Recurrent Unit (GRU) (Cho et al., 2014a) or Long-Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) are used to turn the image embeddings into sequences of words by learning sequence of patters as conditioned by the visual domain. However, this conventional encoder-decoder pipeline showed itself to be brittle in cases where images are composed of infrequent combinations, such as a "*boat on a lake with a Christmas tree*" as shown in Fig. 1.1(a) which happens to be a Greek tradition.

To mitigate this problem, attention mechanisms have been introduced in image captioning models to simulate the human attentive nature when describing visual content (Anderson et al., 2018; Lu et al., 2017; Xu et al., 2015). This was initially carried out at the decoding phase, where attention was given to the most informative image regions upon generating each corresponding word. A two-staged attention mechanism was later proposed to first detect the informative regions through object detectors (bottom-up) and the second phase (top-down) attends the most relevant detected image regions upon the generation of each word (Anderson et al., 2018). The more recent works handles attention through the use of Transformers (Vaswani et al., 2017) which were first introduced to generate machine translated text. Transformer-based models are designed to

perform end-to-end attention via dot-product to implicitly learn informative regions (Yu et al., 2020; Zhang et al., 2021b). Scene graphs (Xu et al., 2019; Yang et al., 2020) have recently been introduced to explicitly represent the relationship between the detected image regions via visual and semantic relation embeddings or through Graph Convolutional Networks (GCNs) trained to automatically encode the relationship between the detected image regions. However, although improvements have been made with this approach when using high-quality scene graphs, the quality of captions based on publicly available scene graph generator models is unsatisfactory due to the lack of quality in the generated scene graphs (Milewski et al., 2020).

Recently, researchers started focusing on the long-term viability of image caption generators by reducing the dependency of large paired image-caption datasets. To achieve this, models trained in an unpaired setting, where no association between images and captions is made during the training phase, were proposed (Ben et al., 2022; Cao et al., 2020). In this PhD, a Keyword-driven and N-Gram Graph-based Image Captioning (KENGIC) approach was proposed. This was purposely developed to (a) reduce the dependency of paired image and caption datasets and (b) to investigate the role of visual keywords in image captioning. While this approach can be used to project insights on how captions can be generated from a set of relevant image keywords, in the future it can also be exploited for more explainability and traceability in image caption generation.

1.6 Keyword-driven and N-gram Graph-based Image Captioning (KENGIC)

The KENGIC framework is designed to build graphs for images by linking keywords through frequent and overlapping n-grams as found in a text corpus. These graphs are then traversed to search for relevant captions. Salient keywords are generated explicitly from images, whilst other words are inserted during the graph traversal.

To further improve the grounding of the generated captions, KENGIC makes use of explicit spatial relations detected in images to validate any implicitly generated spatial relations in the captions. For instance, when taking into consideration the given set of keywords of Fig 1.1 (i.e., "*boat*", "*lake*", "*Christmas*", "*tree*"), the framework generated two possible captions which make use of two different spatial relations (i.e., "boat *with* tree" and "boat *near* tree"). To better ground captions in images, KENGIC validates or corrects the spatial relations that are generated implicitly after taking into consideration the trajector and landmark objects mentioned in the captions.

1.7 Hypotheses

Inspired by the emergent property of the human brain which structures the complexity of cognition into finer interactions between neurons, and how the latter are excited when humans interact with the visual world via connectionism, this research hypothesises that (a) high quality image captions can be generated out of relevant image keywords. Out of these keywords, this research hypothesises that (b) captions can be formulated through an n-gram graph by connecting the neighbouring words commonly used with the extracted keywords. Finally, this research hypothesises that (c) captions can be generated by traversing the graph which combines the extracted image keywords.

1.8 Research Questions

Based on the hypotheses of this PhD, this research aims to provide answers for the following questions:

1. *Can image caption generation be cast as a graph search problem through a keyword-based n-gram graph?*
   This question aims to investigate whether image descriptions can be grounded in image keywords through the use of an n-gram graph-based data structure. This also provides a novel framework to study the implicit and explicit generation of keywords in image captions and how these effect their quality.

   a) *What is the role of image keywords in KENGIC?*
   This question aims to find out which keywords are most important for the generation of n-gram graphs as evaluated with automatic metrics. For example, are graphs better generated with nouns only, or do graphs composed of nouns and attributes lead to better results?

   b) *What visual detectors are required?*
   Since conventional DL based models generate visual keywords implicitly, KENGIC framework offers the possibility to study the interplay between the generation of explicit and implicit keywords in a non end-to-end manner. As KENGIC constructs sentences based on keywords, this framework sheds light on what type of keywords are most important for the generation of graphs and what keywords can be generated implicitly by KENGIC.

c) *What is the quality of the generated captions?*
This question is set to find out the quality of the generated captions as evaluated using automatic metrics and human evaluation.

2. *How does the selection of keywords affect the evaluation performance in image caption generation as measured by current automatic metrics?*
Since image captions are generally evaluated by automatic metrics, this research questions how keywords affect the evaluation performance based on such metrics. For instance, it seeks to answer whether the widely used metrics influence the choice of keywords mentioned in the generated captions or whether captions which make use of rich keywords are penalised over generic and less specific vocabulary.

3. *How does spatial relation detection contribute in automatic image captioning?*
Given that the majority of current image caption generators produce captions without the explicit use of spatial relation detectors, this study aims to investigate the role of spatial relations in the context of image caption generation.

1.9 Objectives

To provide answers for the aforementioned questions, this research has the following objectives:

1. Review the literature which covers models used in image captioning and how graphs are used in image captioning to position KENGIC in the literature.
2. Review the literature on visual detectors used for the detection of objects, attributes, verbs and spatial relations.
3. Develop KENGIC framework.
4. Perform a preliminary study on KENGIC based on human keywords.
5. Study the performance of KENGIC based on detected objects and predicted image multi labels.
6. Perform both quantitative and qualitative analysis on KENGIC.
7. Handle the grounding of spatial relations in generated captions.

1.10 Contributions

This research presents the following as contributions to knowledge in image caption generation and spatial relation detection for better grounding in captions:

1. KENGIC: A novel keyword and *n*-gram graph based image caption generation framework. This was proposed to generate image captions from a set of detected image keywords. This approach was also intended to reduce the dependency of large scale paired image-caption datasets while paving the way for more explainable and traceable image caption generation. Both quantitative and qualitative analysis confirmed the efficacy of this approach.
2. A study on what keywords benefit image evaluation metrics. This study confirmed that nouns play the most important role in image captioning. Furthermore, it was found that current popular metrics pay more attention to the mentioned keywords rather than the structure of the sentences.
3. Spatial relation detection in images was studied to better handle the use of spatial prepositions in image captioning. The quality of the captions based on single-label based spatial relation detection confirmed that the quality of captions slightly reduces and therefore image captioning benefits from multi spatial relation detection.

1.11 Outline

The rest of this thesis is organised as follows. Chapter 2 provides the background related to this topic and reviews the literature relevant to image captioning. Chapter 3 presents the methodology based on KENGIC framework. Chapter 4 reports and discusses the conducted experiments and results. Chapter 5 presents the integration of Spatial Relation (SR) detection in KENGIC, while Chapter 6 presents a study on multi SR detection. Finally, the thesis ends by summing the conclusions and future work in Chapter 7.

2 Background and Literature Review

It has been a long standing problem for Computer Vision (CV) researchers to automatically understand and describe the visual content of images. Research on associating text with images goes back at least to the 1960s with early works focusing on object and region labelling (Rosenfeld, 1978). Image description proper, however, starts where a summarising description of the whole image is aimed for. Image description aims to produce a summarising description of the most important aspects of an image which typically involves the prioritisation of the most salient objects and their relationships. This problem is difficult to model because image captions written by humans are conditioned by language-specific uses and constraints (Herskovits, 1997) and on the cues humans pick up from a perceived 3D world, whereas an image is a 2D projection. This chapter reviews the main literature that contributed towards understanding image visual content, starting from image understanding to image caption generation.

2.1 Image Recognition

Early research in CV started from the simple but yet challenging problem of image classification, where images were classified based on their most prominent single object found in images (Vailaya et al., 2001). Since images can be classified under more than one category (for example an image can have both a “person” and a “boat”), researchers also contributed in multi-label image classification to assign multiple classes for a single image (Li et al., 2004). Despite being an easy task for most human beings, this problem is difficult for machines because of the various forms and sizes of object classes which can be either occluded or visible under different viewpoints and even due to illuminations and intraclass variations among others. Building up on that knowledge, researchers were then able to localise the main single object in bounding boxes (Harzallah et al., 2009; Lampert et al., 2008; Lowe, 2004). Later, the field moved to multi object detection and image segmention (Hariharan et al., 2014, 2015; He et al., 2017a) where multiple objects were not only detected in bounding boxes but also delineated from the rest of the image. Examples

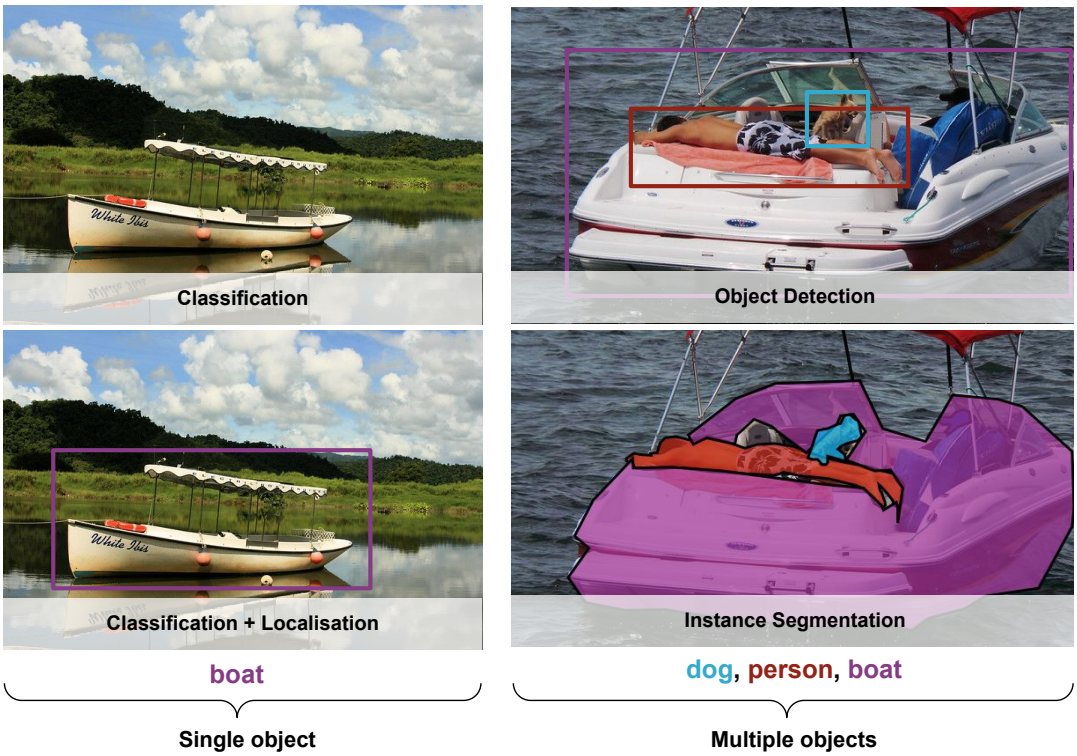


Figure 2.1: Computer Vision: From Single object classification to multiple instance segmentation.

of these sub-tasks in image recognition are illustrated in Fig.2.1.

2.1.1 Traditional Detectors

Early researchers opted for a direct region proposal generation pipeline, whereby sliding window mechanisms were used to (a) select informative image regions, and then (b) recognise the objects through handcrafted features like SIFT (Lowe, 2004), HOG (Dalal and Triggs, 2005) and Haar-like (Lienhart and Maydt, 2002) features. These hand-engineered features were used to learn discriminitive features from raw pixels through machine learning classifiers. Generally, this was handled by the use of SVM (Cortes and Vapnik, 1995), AdaBoost (Freund and Schapire, 1997) and DPM (Felzenszwalb et al., 2010) which was found to be the most effective model in the PASCAL[1] Visual Object Classes (VOC) detection competitions (Everingham et al., 2010) at that time.

[1]PASCAL corresponds to pattern analysis, statistical modelling and computational learning.

2.1.2 Deep Learning Approach

With the recent uptake of Deep Learning (DL) thanks to the availability of large-scale annotated data and high-performance parallel computing infrastructures such as GPU clusters, the use of Convolutional Neural Network (CNN) (LeCun et al., 1989) revolutionised the field of Computer Vision. Nowadays, traditional handcrafted features have been replaced by embeddings learned by networks while being trained to classify images or detect objects. These neural based architectures achieved impressive performance when compared to earlier works and have even surpassed human-level performance on single-label image datasets such as the MNIST[2] (Lecun et al., 1998) and ImageNet (Deng et al., 2009).

2.1.2.1 Convolutional Neural Network

Inspired by the neocognitron (Fukushima and Miyake, 1982), LeCun et al. (1989) proposed the first CNN framework to classify handwritten digits using the back-propagation algorithm (Hecht-Nielsen, 1992). Numerous deep architectures followed, with the most popular ones being AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2015), GoogleNet (Szegedy et al., 2015), and ResNet (He et al., 2016). Generally, a CNN is composed of three types of layers. These include the *convolutional*, *pooling*, and *fully-connected* layers. The *convolutional* layer consists of multiple convolutional kernels that are used to extract feature maps. This is handled by having each neuron within each feature map connected with a neighborhood of neurons in the previous layer. A feature map is obtained by convolving the input with an already learned kernel and by passing each convolved result through a non-linear activation function. The kernel is shared by all spatial locations of the input, while several kernels are used to obtain the complete set of feature maps. More formally, the feature value $z^l_{i,j,k}$ located in position (i, j) of the kth feature map found in the lth layer is computed by:

$$z^l_{i,j,k} = w^{l^\top}_k x^l_{i,j} + b^l_k, \tag{2.1}$$

where w^l_k and b^l_k are the weight vector and bias term for the kth filter in the lth layer, respectively. To detect non-linear features, CNNs make use of non-linear activations which can include the sigmoid (σ), hyperbolic tangent ($\tanh$) and Rectified Linear Unit (ReLu) which are defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

[2]MNIST stands for Modified National Institute of Standards and Technology.

$$\text{tanh(x)} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.3}$$

$$\text{ReLu}(x) = \max(0, x), \tag{2.4}$$

where x is the convolutional feature $z^{l}_{i,j,k}$.

In order to reduce the resolution of feature maps while keeping them approximately translation invariant (i.e., features should not be dependent on their location), *pooling* layers are used for the extraction of higher abstracted features (Goodfellow et al., 2016). This layer is typically situated between two *covolutional* layers. The pooling operation works by sliding a filter (generally of size 2×2 with a stride of 2) over an output feature map and this filter computes an output from the receptive field (i.e. feature map being processed). The most commonly used pooling techniques include *max pooling*, where the filters simply select the maximum pixel value found in that receptive field; and *average pooling* which instead calculates the average values. Having a number of *convolutional* and *pooling* layers following each other makes the network *deep* (i.e., D-CNN). Following these layers, *fully-connected* layers are used to connect neurons from previous layers to single neurons of the following layers. The final layer is considered as the *output* layer as it handles the final classification and is trained to output the class probabilities. Normally, this layer uses either a sigmoid layer for multi-label classification or softmax activation function for single-label classification as follows:

$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}, \tag{2.5}$$

where x is the output scores of the K classes and $\text{softmax}(x)_i$ is the probability of the ith class.

The optimal parameters (θ) of a CNN (i.e., weight matrices and biases) are obtained by minimising a loss function $\mathcal{L}$ with an optimiser such as Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951), or the more popular Adam optimiser (Kingma and Ba, 2015). Generally, for a given set of input and output pairs $\{(x_i, y_i) \mid i \in \{1, \ldots, N\}\}$, corresponding outputs $\hat{y}$, and loss ℓ for one pair, the total loss $\mathcal{L}$, is defined by:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, \hat{y}_i; \theta) \tag{2.6}$$

Cross-entropy (CE) loss, which is used to measure the dissimilarity of two probability distributions, is the widely used loss function when optimising models for classification

since it leads to better generalisability and faster training (Bishop and Nasrabadi, 2006). This is defined by:

$$\ell_{CE}(y_i, \hat{y}_i) = -\sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k}, \tag{2.7}$$

where y_i and $\hat{y}_i$ are the probabilities of the ground-truth (gt) and predicted class labels for the class label k of the ith training instance, respectively. The total cross-entropy loss can therefore be computed by:

$$\begin{aligned} \ell_{\text{total}CE} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k} \\ &= -\frac{1}{N}\sum_{i=1}^{N} \log \hat{y}_{i,k_{gt}} \qquad \text{(since only } k_{gt} \text{ is non zero in } y_k\text{)} \end{aligned} \tag{2.8}$$

2.1.2.2 Deep Learning based Detectors

The second generation Deep Learning (DL) based models recognise objects by either (a) following the traditional region proposal pipeline, or else (b) by casting the problem as as regression or classification problem through one holistic framework designed to learn both the categories and locations directly. Region proposal models mainly include the R-CNN (Girshick et al., 2014), SPPNet (He et al., 2015), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016), FPN (Lin et al., 2017a), and Mask R-CNN (He et al., 2017a). On the other hand, the classification-based models include the MultiBox (Erhan et al., 2014), AttentionNet (Yoo et al., 2015), G-CNN (Najibi et al., 2016), YOLO (Redmon et al., 2016), SSD (Liu et al., 2016) and RetinaNet (Lin et al., 2017b). More recently, the Transformer (Vaswani et al., 2017) architecture which is currently dominating the NLP field, has been introduced for object detection due its long-range representation and high performance. Transformer-based architectures for detection include the DETR (Carion et al., 2020), Deformable DETR (Zhu et al., 2021), the UP-DETR (Dai et al., 2021) and the PVT (Wang et al., 2021).

2.1.2.3 Region Proposal based Architectures

The Region-based Convolutional Neural Network (R-CNN) works by extracting a set of object proposals (over 2000 boxed from one image) through a selective search (Van de Sande et al., 2011). Each object candidate box is then scaled to a fixed size image and passed to a CNN model trained on ImageNet (Krizhevsky et al., 2012) for feature extraction. Detections are carried out by linear Support Vector Machine (SVM) classifiers. Due to overhead computation of features for a large number of overlapping bounding boxes,

SPPNet was introduced to generate fixed-length CNN features regardless of the image and region sizes. Fast R-CNN was then proposed to simultaneously train the detector and bounding box regressor under the same network setup unlike in R-CNN and Spatial Pyramid Pooling Network (SPPNet) which required independent configurations. Faster R-CNN was shortly after proposed as an end-to-end object detector which can handle both the generation of object proposals as well as their classifications. This was achieved by integrating most of the used object detection modules (e.g., proposal detection, feature extraction, bounding box regression, etc.) into one end-to-end framework. Feature Pyramid Networks (FPNs) were later proposed to improve the problem of scale invariance. In contrast to previous architectures which run detection only on the top layer of the network, FPN is based on a top-down architecture with lateral connections for building high-level semantics at all scales. The Mask R-CNN is an extension to the Faster R-CNN which also performs instance segmentation by adding a branch for predicting segmentation masks on each Region of Interest (ROI) using a small Fully Connected Layer (FCL) in a pixel-to-pixel manner.

2.1.2.4 Classification based Architectures

You Only Look Once (YOLO) was the first detector which performs object detection in one phase (i.e., does not follow the paradigm of generating region proposals followed by classification). This networks works by dividing the image into regions and predicts bounding boxes and likelihood of object classes simultaneously. Later modifications were proposed in YOLOv2 (Redmon and Farhadi, 2017) and v3 (Redmon and Farhadi, 2018) to improve detection performance while keeping its high efficiency. To enhance the recognition of small objects, Liu et al. (2016) proposed the Single Shot MultiBox Detector (SSD) architecture which makes use of multi-reference and multi-resolution techniques. Despite the efficiency of classification-based detectors, detection performance was not reaching the performance of the two staged architectures. Lin et al. (2017b) found that the main reason behind this problem was due to the extreme foreground-background class imbalance found in the training data. For this reason, the authors introduced a focal loss function in RetinaNet (Lin et al., 2017b) to focus the learning on hard negative examples.

2.1.2.5 Transformer based Architectures

The Transformer (Vaswani et al., 2017) architecture is an encoder-decoder model designed for sequence-to-sequence learning based on attention. The visual transformer typically splits the input image into small patches and makes use of input tokens which include patch tokens, class tokens and position embeddings. The architecture is normally

composed of a series of stacked transformer modules which includes normalization layer, multi-head self-attention, skip attention layer, multilayer perception or feed-forward network, and a post-processing module. Carion et al. (2020), were the first to exploit transformers for object detection. The authors proposed the DEtection TRansformer (DETR) where first they decomposed the output feature map of a CNN into patches and mapped them into one-dimensional vectors. These were passed into several transformer encoders together with positional embeddings. The decoder receives two inputs, namely the object query and the output of the encoder. Through the multi-head self-attention, this process aims to find patterns in image features as queries for objects. The output of the decoder is then passed on to the class and bounding box modules to predict the category and position of the objects respectively. To mitigate the problem of slow convergence and limited feature spatial resolution of DETR, Zhu et al. (2021) proposed the Deformable DETR which does not consider all values to calculate the attention of one query and attention scores are learned by a network instead of multiplying the queries and keys. Dai et al. (2021) proposed the random query patch detection for DETR as an unsupervised pretraining method while Wang et al. (2021) presented the Pyramid Vision Transformer (PVT) for both detection and segmentation which performs spatial reduction on the key K and value V while maintaining the spatial size of the query Q (Xu et al., 2022).

2.2 Image Classification

The problem of image classification from images involves the prediction of labels that are relevant to images (Lanchantin et al., 2021). Images can either be simply assigned to one category (single label), or else to multiple classes (multi-label). Single label classification makes use of machine learning classifiers to learn to discriminate engineered features or embeddings for each class. This can be cast either as a binary-classification problem, where images can be classified into two classes (e.g., whether an image belongs to a "person" category or not), or else as a multi-classification problem where images can be categorised in more than two classes.

The most naïve approach to assign multi-labels for images is to train a binary classifier for each label and all predictions which exceed a given threshold or which fall under a specific ranking criteria are considered relevant to an image (Zhang and Zhou, 2014). However, this approach does not exploit the interdependence relationship between labels, something which could be leveraged in order to predict labels with better accuracy and relevancy. For example, classes like "*lake*" and "*nature*" are more probable than "*office*" and "*computer*" in context of Fig. 2.1. This is commonly known as the binary relevance

problem (Zhang et al., 2018b) and the majority of the work conducted in this area aims to provide solutions for this problem.

2.2.1 General Multi-label Classification

There is a series of works which cast the problem of multi-label classification as a conditional prediction problem by estimating the true joint probability of output labels given the input using the chain rule and by predicting one label at a time. For example, Dembczyński et al. (2010) used probabilistic classifier chains to capture pairwise label correlations, while Read et al. (2011) linked binary classifiers along a chain. Each classifier was responsible for learning and predicting the binary association for each label based on the feature space and all prior predictions. Ghamrawi and McCallum (2005) explored the use of Conditional Random Field (CRF) models to directly parameterise label co-occurrences, while Guo and Gu (2011) developed a conditional dependency network model to provide intuitive representation for the dependencies among labels. A different approach in multi-label classification is to project the input features and the corresponding output labels in one shared and latent embedding space. For example, Yeh et al. (2017) derived a deep latent space through a deep neural network architecture which makes use of Deep Canonical Correlation Analysis (DCCA) and autoencoder structures.

2.2.2 Multi-label Image Classification

Early works which addressed the problem of multi-label image classification viewed it as multi-label image annotation problem and were inspired by machine translation techniques. Barnard and Forsyth (2001) used a generative hierarchical model which is a combination of the asymmetric clustering models that maps documents into clusters and the symmetric clustering model which models the joint distribution of documents and features. Duygulu et al. (2002) later proposed a machine translation method which annotates image regions with words by organising region types using a variety of features through Expectation-Maximisation (EM) algorithm. Later, image annotation was formulated as a classification-based problem based on generative models[3] that learn parametric functions to model class distributions. Due to the complex distribution found in image tagging, most works shifted to the discriminitive Multiple Instance Learning (MIL) paradigm. For instance, Makadia et al. (2008) proposed a nonparametric nearest-neighbor-based

[3] In classification, generative models learn the distribution of classes whereas discriminitive models learn the boundaries between classes

tag transfer, while Guillaumin et al. (2009) proposed TagProp which is a discriminatively trained nearest neighbor model.

Works which tackled multi-label classification directly in images include that of Xue et al. (2011) who proposed a Correlative Multi-Label Multi-Instance image annotation where the input image was segmented and viewed as a bag-of-regions. Inter-label correlations were then captured via a co-occurrence matrix of concept pairs. Probably, Gong et al. (2014) were the first who leveraged CNN features in multi-label image classification, while making use of top-k ranking objectives.

Wang et al. (2016) proposed to jointly model full images and labels using a CNN-RNN framework which predicts the multi-labels as an ordered prediction path. The CNN was used to extract the semantic representations from images, while the RNN was employed to model the image/label relationship and label dependencies. This design pattern became popular and was further improved by other researchers. For instance, Liu et al. (2017a) proposed to use a semantically regularised embedding layer as the interface between the CNN and RNN, while Zhu et al. (2017a) proposed a Spatial Regularisation Network (SRN) to generate class-related attention maps to capture both spatial and semantic label dependencies via simple learnable convolutions.

Since these works only consider the global representation of full images, they may not be optimal for multi-label images containing objects of different categories, scales and locations. Furthermore, this straightforward approach does not take into account the relationship between the semantic labels and the local image regions and can also be affected by noisy backgrounds. To this end, works were proposed to extract object proposals as the the informative regions for multi-label classification.

Fang et al. (2015) made use of a Multiple Instance Learning (MIL) (Maron and Lozano-Pérez, 1998) approach to predict relevant keywords for images by learning discriminative visual features for each word. MIL takes *positive* and *negative* bags of bounding boxes as input sets for each image. A bag is considered as positive if a keyword found in the dictionary is found in the image description, and negative otherwise. The MIL paradigm works by iteratively selecting instances found in the positive bags and re-training the detector using the updated positive labels. The authors trained visual detectors to predict words which commonly occur in image captions including nouns, attributes and verbs by taking the CNN features extracted from image sub-regions. Wei et al. (2016) proposed the Hypotheses-CNN-Pooling (HCP) approach which also considers images sub-regions as hypothesis. These are taken as inputs and a shared CNN is connected with each hypotheses to aggregate their outputs through category-wise max-pooling. Yang et al. (2013) considered this problem as a multi-class multi-instance problem, where each image was treated as a bag and object proposals were treated as instances. Zhang et al. (2018a)

also made use of CNN-based image proposals by proposing the Regional Latent Semantic Dependencies model (RLSD) which, unlike traditional Region Proposal Networks (RPNs) that try to predict the proposals with a single object, is designed to localise regions that may contain multiple highly-dependent labels. These are then propagated to an RNN to extract the latent semantic dependencies at the region level.

To eliminate the extraction of object proposals, Chen et al. (2018) used the advantage of attention-based models which were first proposed to automatically extract relevant regions in visual tasks (Anderson et al., 2018; Ghosh et al., 2019; Xu et al., 2015; You et al., 2016; Zhou et al., 2016). Attention based models generally employ RNNs to iteratively search for important regions in conjunction with RL to optimise the model with a delayed reward. The authors proposed a recurrent attention reinforcement learning framework to discover a sequence of informative regions related to different semantic objects and to predict label scores conditioned on the same regions. From the perspective of Yu et al. (2019), both global and local image information are important for classification and therefore the authors proposed the deep Dual-stream nEtwork for the mulLTi-lAbel image classification (DELTA) which is composed of a multi-instance and global priors network. While the former network was intended to extract the multi-scale and class-related local instance features, the latter was proposed to capture the global image priors.

The graph data structure provided a semi-superivised learning approach for multi-label image classification. For instance, Wang et al. (2009) first proposed the multi-label correlated Green's function approach to label images over a graph which uses the prior information of the training images. Later Wang et al. (2011), the same authors made use of the labelled graph to calculate the similarities between images. To consider the complimentary nature between multimodal features and correlated labels, Xu et al. (2014) proposed a semi-supervised label diffusion process on a bi-relational graph. However, since graphs do not always capture the information of individual features, Hamid Amiri and Jamzad (2015) suggested the formation of a graph based on sub-graphs generated for various types of visual features. Gao et al. (2015) and Song et al. (2016) proposed an optimal graph structure according to label information on image parts and various visual feature information. Ding et al. (2016) proposed a context-aware multi-instance multi-label learning (MIML) model to integrate the instance context and label context into a general framework, while Lei et al. (2016) presented a social diffusion analysis approach which exploits the abundant social diffusion records about how images are shared in online social networks. Recently, several studies have used graph neural networks to model label dependencies. These models rely on knowledge-based graphs generated from label co-occurrence statistics. For instance, Chen et al. (2019) proposed a model which builds a directed graph over the object labels represented by word embeddings and a GCN then

learns to project the label graph into a set of inter-dependent object classifiers. These were then applied to the image descriptors extracted by another sub-network.

The built-in attention mechanism of the Transformer model (Vaswani et al., 2017), which was first proposed for machine translation, has recently opened up another opportunity to adaptively extract the desired visual features for multi image labels. By taking each label as query in the Transformer decoder, recent models (Lanchantin et al., 2021; Liu et al., 2021; Ridnik et al., 2021c) pool related image features via cross-attention and then classify these features via binary classification. Building on this approach, Lanchantin et al. (2021) proposed the Classification Transformer (C-Tran) which consists of a Transformer encoder that is trained to predict a set of target labels given an input set of masked label embeddings and visual features extracted from a CNN. On the other hand, Ridnik et al. (2021c) proposed the ML-Decoder, an attention based classification head for the transformer-decoder. This achieved top results while improving the efficiency and scalability of the previous works (Lanchantin et al., 2021; Liu et al., 2021) with the elimination of self-attention and the introduction of group decoding.

2.3 Spatial Relation Detection

As reviewed in Birmingham and Muscat (2022), studies on spatial relations in two disjointed fields, namely cognitive linguistics and computer science, are briefly presented. In the psycho- and cognitive-linguistics literature, the spatial recognition problem was mainly addressed by manually developed spatial templates. Such works include that of Herskovits (1980) which categorises spatial language into use cases based on object and contextual features and typicality, while Logan and Sadler (1996) classify geometric scenes using spatial templates. These models require an exhaustive list of use cases/templates, an approach for selecting the correct template and an account of what modifications can be made to fit an imperfect template to a scene. The Attentional Vector Sum (AVS) computational model (Regier and Carlson, 2001) is able to produce human-like results for several different prepositions. This computational model was further extended by Terry et al. (2005) to account for functional information by focusing the attention on the functional parts of the objects. This approach predicts the acceptability rate of spatial terms in contrast to categorizing stimuli. Coventry et al. (2005) and Cangelosi et al. (2005) developed a model that addresses the constraints of the functional geometric framework developed by Coventry and Garrod (2004) for the set of prepositions that includes: over, under, above, and below. This was consistent with the level of acceptability rated by human evaluators when it comes to describing scenes involving both geometric and func-

tional information. Martinez et al. (2001) developed another model based on a neural network which takes descriptions of visual scenes as input. In Kelleher and Kruijff (2005), the choice for the “most" appropriate preposition was addressed by considering the minimum cognitive load (least effort). Such works were based on data gathered from controlled experiments, using 2D and 3D synthetic diagrams, where humans were asked to rate the acceptability of a given preposition depicted in a given configuration. Early models concentrated on the geometric features that predict prepositions. However, further work emphasized the language and geometrical bias of prepositions (Carlson-Radvansky and Radvansky, 1996; Coventry et al., 2001; Dobnik and Kelleher, 2014), and other work studied how perceptual features, such as occlusion, modify the spatial templates (Kelleher et al., 2011).

The spatial relation detection problem is even more difficult when considering images of the physical world. Sadeghi and Farhadi (2011) were probably the first to deal directly with relation detection in real-world images and treated the problem as object detection. Due to the combinatorial nature of the problem, this method does not scale because of the large number of unique relations that exist when adding new objects to the dataset. The most obvious and natural way is to compute spatial features in addition to language and visual properties of the objects under attention. Two approaches are considered when dealing with spatial features obtained from images; (a) methods based on image features learnt via deep neural networks, mainly CNNs (Dai et al., 2017b; Lu et al., 2016), and (b) methods based on manually defined geometrical or topological features (Belz et al., 2015; Ramisa et al., 2015), or a mix of both (Ramisa et al., 2015; Yu et al., 2017). These geometric and visual features have also been used in other tasks, such as in term disambiguation to improve spatial role labelling (Rahgooy et al., 2018). In most models, language features are used to represent the object labels by using either indicator vectors (one-hot encodings) or pre-computed word embeddings such as *word2vec* (Mikolov and Dean, 2013). Conversely, when spatial relations are extracted from textual image captions, linguistic features are also computed from natural language processing tools, which include among others, dependency parsing and semantic role labelling (Kordjamshidi and Moens, 2015).

Machine learning models are trained to learn all the steps at one go by selecting one or more plausible prepositions based on spatial or geometrical features, as modified by perceptual properties and filtered by linguistic knowledge. In addition, these models are expected to select an appropriate frame of reference (Carlson-Radvansky and Logan, 1997; Logan and Sadler, 1996) which, unless a set of rules are strictly followed, results in dataset noise which in turn manifests itself in errors at the output.

In contrast to template models, machine learning based classifiers are trained from crowd-sourced data, which is normally incomplete in terms of both the images depicting

all possible spatial configurations, as well as their corresponding human annotations. Due to this limitation, these models are normally trained in the single-label classification mode, i.e., the output is a softmax type that only ranks the output classes, without taking into account that multiple relations may be equally suitable in a given configuration.

One of the early works which tackled the problem of spatial relation detection as an independent sub-task is that of Elliott and Keller (2013). The authors proposed a framework that combines the structure of an image with the corresponding description structure in the form of Visual Dependency Representations (VDRs). As a complement to this structure, the authors introduced the hand-engineered Visual Dependency Grammar (VDG) to recognise a limited set of spatial prepositions between image objects in the form of a dependency graph. The grammar is defined in terms of three geometric properties consisting of pixel overlap, angle and distance between image regions. The relation between this dependency graph and the image syntactic tree provided a way of generating prepositions for image descriptions.

Muscat and Belz (2015) explored how spatial prepositions between two objects can be predicted from language and visual information by using probabilistic approaches. Three different models were used and combined. The *prior model* was used to capture the probabilities of prepositions for ordered pairs of object labels L_s and L_o. This language model was constructed as a probabilistic classifier which gives the preposition most likely to occur v_{OL} by:

$$V_{OL} = \underset{v \in \mathbf{V}}{\arg\max}\, P(v|L_s, L_o), \tag{2.9}$$

where v is a preposition found in set $\mathbf{V}$ containing all available prepositions, while L_s and L_o are the two object labels from the language domain.

The *likelihood model* was constructed to predict prepositions based on a set of six geometric features, extracted from the image size and the bounding boxes of the respective objects. These features include the area of the two bounding boxes, the ratio of the two areas, the distance between them, the overlapping area between the two objects and the position of the first object in relation to the second object. Based on these geometric features, this model predicts preposition v_{ML} by:

$$V_{ML} = \underset{v \in \mathbf{V}}{\arg\max} \prod_{i=1}^{6} P(F_i|v), \tag{2.10}$$

where F_i is the ith geometric feature.

Finally, the Naive Bayes classifier which is derived from the maximum-a-posteriori Bayesian model was employed to combine both language and visual information by the following model:

$$V_{NB} = \underset{v \in \mathbf{V}}{\arg\max}\, P(v|L_s, L_o) \prod_{i=1}^{6} P(F_i|v) \tag{2.11}$$

Belz et al. (2015) compared this Naive Bayes classifier with the hard-wired VDG proposed by Elliott and Keller (2013) to predict English and French spatial prepositions. It was verified that the combination of both language and geometrical features improves the generation of spatial prepositions. In fact, the Naive Bayes classifier which takes both language and vision information, achieved the greatest prediction accuracy.

Muscat et al. (2016) and Belz et al. (2016) investigated the generation of spatial relationships and how it is affected by varying the different aspects of the automatic generation method, including different preposition sets, models and feature sets. It was found that optimising the preposition and feature sets improves the previous best accuracy of the former work. This combination of language and visual features was also adopted by Ramisa et al. (2015). In fact, they combined geometric relations from image objects as well as encoded textual features from image annotations and visual information together to predict spatial prepositions. Ramisa et al. (2015) defined an 11-dimensional vector of geometric features, whilst textual image labels were encoded as an indicator vector and with a word2vec (Mikolov and Dean, 2013) feature vector, respectively. In addition to the aforementioned work, Ramisa et al. (2015) included high-level visual representations for image objects extracted from the final layer of a CNN. Given geometric and textual features, prepositions were predicted by using a multi-class logistic regressor, while a three-node chain Conditional Random Field (CRF) model was used to predict prepositions when having both visual and geometric features. Hürlimann and Bos (2016) suggested the use of an extension of first-order models to represent images of realistic situations. In their work, the authors concentrated particularly on predicting and integrating three spatial relations namely, *part-of*, *touching* and *supports* into first-order models borrowed from logic. These three relations were distinctly selected since they are well-defined and less fuzzy as opposed to the prepositions considered in previous studies (Muscat and Belz, 2015; Ramisa et al., 2015). In this study, the authors considered the spatial relation prediction task as a classification problem where each instance belongs to a set of six disjointed classes. The classification was based on spatial, lexical and corpus features. As spatial features, they exploited the overlap area between two bounding boxes, and examined whether the first object is contained within the second object or vice versa. Moreover, the object sizes and the occlusion present between images were part of the spatial feature set. To combine

linguistic knowledge, the authors incorporated the meronymy (part-whole relation) and hypernymy as two lexical features. Furthermore, prepositions and verbs occurring between the lemmas of the two objects were collected from large text collections and used as corpus features. Word embeddings were exploited by calculating word2vec (Mikolov and Dean, 2013) feature vectors for each synset as an average across all lemmas' vectors. This study proved that the combination of spatial and lexical feature groups significantly outperforms predictions based on independent feature groups.

2.4 Visual Relationship Detection

Understanding the visual relationship adds further insight on how the subject and the object of images are related to each other. This was formalised as a task by Lu et al. (2016), where they proposed subjects to be linked with predicates which include not only spatial relationships, but also verbs such as "running" or a combination of both such as "falling off". To tackle this problem, Lu et al. (2016) published the first Visual Relationship Detection (VRD) dataset and performed object and predicate prediction separately. The union of the two objects was used as the visual input to predict the predicates, while language priors and the likelihood of relationships were used to augment the visual module. Plummer et al. (2017) fused multi visual features, including appearance, size, bounding boxes, and linguistic cues such as adjectives. Liang et al. (2017) detected relationships and attributes with a reinforcement learning framework, while Li et al. (2017) proposed the Visual Phrase-guided Convolutional Neural Network (ViP-CNN) which leverages the visual feature level connection among the subject, predicate and object. Dai et al. (2017b) predicted visual relationships based on a Deep Relational Network trained to exploit the statistical dependencies between objects and their relationships. Their approach was to first run an object detector and then apply a network to select promising pairs of interacting object detections. Yu et al. (2017) proposed a linguistic knowledge distillation framework which extracts linguistic knowledge from training captions and public textual data to distill knowledge into an end-to-end deep neural network. Recently, researchers have integrated attention mechanisms and graphs in visual relationship detection. For instance, Kolesnikov et al. (2019) proposed a Box Attention mechanism which allows the modelling of pairwise interaction using standard object detection pipelines and Tripathi et al. (2021a) used spatially aware word embedding through scene graphs and joint feature representations that contain visual, spatial and semantic embeddings to better represent the semantic relationship between image objects. Li et al. (2021) proposed a relationship graph learning network (RGLN) to specifically learn correlations between objects'

relationships. This was handled by considering every pair of detected objects as relationship proposal and nodes in a graph. Graph attention subnetworks detected relationships based on visual and semantic information.

2.5 Image Captioning

The task of automatic image caption generation not only involves the detection of the most salient objects, relations and attributes in images but also the ability to formulate well structured sentences. Early researchers viewed this task either as retrieval-based problem where descriptions were reused from image and caption pairs (Mason and Charniak, 2014; Ordonez et al., 2011), or as a template-based problem where captions were generated through grammars or rule-based approaches (Kulkarni et al., 2013a; Yang et al., 2011). Nowadays, the majority of the research conducted in this area is inspired by the work carried out in neural machine translation (Bahdanau et al., 2014), with most recent contributions being based on the encoder-decoder model. Generally, these models are trained to maximise the probability of the ground-truth caption words given the paired image embeddings by using recurrent neural networks (RNNs). This approach was proposed to model the sequential dependency between words. Later, this has been improved by the introduction of attention mechanisms (Xu et al., 2015; Yang et al., 2016) to dynamically focus on different image regions during the decoding phase. This was proposed to simulate the natural attentive behavior that humans perform instinctively when describing images. More recently, novel objects (Huang et al., 2019) and semantic concepts have been introduced in image captioning. For instance, models were developed to exploit objects (Lu et al., 2018), attributes (Yao et al., 2017), and relationships (Yao et al., 2018). To further exploit the structural representation of images, researchers proposed the graph data structure to combine the semantic embeddings (Li and Jiang, 2019; Yang et al., 2019). Reinforcement Learning has been applied in image captioning due to the exposure bias (Ranzato et al., 2016) and loss-evaluation mismatch in sequence prediction (Ren et al., 2017) and lately, researchers started taking advantage of the Transformer (Vaswani et al., 2017) architecture by applying self-attention on words and cross-attention on the output of the last encoder layer (Herdade et al., 2019; Huang et al., 2019; Li et al., 2019a). Recently, to reduce the dependency of costly large scale datasets consisting of image and human authored caption pairs to train current image caption generators, researchers started looking into the problem of unpaired image captioning (Feng et al., 2019; Gu et al., 2018b, 2019).

2.5.1 Overview

Generating automatic descriptions from images requires an in-depth understanding of how humans describe images. An image can be analysed and described from different perspectives (Jaimes and Chang, 1999; Shatford, 1986). However, most of the current research follows Hodosh et al. (2013) viewpoint in such a way that descriptions are generated with respect to the visual information and ignore background details or components that are not present in an image, with such details being the location of where the image was taken or who took the picture. This section presents a detailed survey on existing image caption generation models.

Automatic image captioning originally attempted to produce only simple descriptions for images taken under extremely constrained conditions. As an example, Kojima et al. (2002) used concept hierarchies of actions, case structures and verb patterns to describe human activities in office environments, while Hède et al. (2004) used a dictionary of objects and language templates to describe unnatural images composed of just objects in backgrounds. It was not until recently, that work intended to generate captions for natural and generic real life images started being proposed (Farhadi et al., 2010; Gupta et al., 2012; Ordonez et al., 2011; Yang et al., 2011). Work in this area started off by following two lines of research, i.e., template and retrieval based image captioning. The first group of models generates textual descriptions by primarily analysing the composition of an image in terms of image objects, attributes, scene types and event actions, extracted from image visual features. These models subsequently exploit the extracted visual information to derive an image description by driving a natural language generation model such as *n*-grams, templates and grammar rules. On the other hand, the second group formulate descriptions by finding visually similar images to the query image from a collection of already described images. The novel image is then described by reusing the description of the most similar retrieved image, or by amalgamating the descriptions of the set of retrieved descriptions of visually similar images. Retrieval-based models can be further organised based on the technique used for representing and computing image similarity. There are methods which either use a *visual space* for finding related images or else a *multimodal space* that combines both textual and visual information in one single space. Since template and retrieval based methods perform captioning either by re-using existing descriptions from the training set or by relying on preset language structures, these first generation models are quite rigid and constrained, thus leading to limited expressiveness in the generated descriptions. Despite the complexity involved in image captioning, recent advances in deep neural architectures opened a new and effective way of how image captions can be generated. In fact, deep learning in image captioning has demonstrated

state of the art results. In the following sub-sections, a comprehensive overview of image captioning methods organised in the aforementioned categories is presented.

2.5.2 Retrieval based Image Captioning

Image description models casting the generation as a retrieval- or ranking-based based problem are designed to reuse textual descriptions of visually similar images taken from a collection of already captioned images (Bernardi et al., 2016). This can either be a direct transfer from one single caption describing an image or by linking a set of descriptions retrieved from a pre-specified sentence pool. As opposed to direct generation models, retrieval-based approaches require large and diverse datasets of pre-captioned images.

The approach taken by these models is to first extract visual features from the query images. Based on a visual similarity measure dependent on the extracted features, a candidate set of related images is retrieved from a collection of previously captioned images. Retrieved descriptions are then re-ranked based on visual and textual information extracted from the retrieved set.

One of the early works considering this approach is Im2Text proposed by Ordonez et al. (2011). In this work, visually similar images were retrieved by computing a global image similarity based on the GIST feature vector (Friedman, 1979). As a re-ranking step, the authors used a set of detectors to detect objects, background stuff, people and actions present in the retrieved images. This was mainly intended to better capture the visual content of the retrieved images as well as to represent the images by these detectors. The final re-ranking stage intended to select the most relevant image was carried out using a classifier trained over the aforementioned semantic features together with the evaluated description score.

Mason and Charniak (2014) considered the image captioning task as an extractive summarisation problem. Final descriptions were selected by exclusively taking into consideration the textual information in the final re-ranking step after representing images with scene attributes proposed by Patterson et al. (2014). Image descriptions were generated by first estimating the conditional probabilities of observing a word in the query description based on the set of retrieved images. Subsequently, the output was generated by exploiting two distinct extractive summarisation techniques, namely the SumBasic model (Nenkova and Vanderwende, 2005) and the one based on Kullback-Leibler divergence (Perez-Cruz, 2008) between the word distributions of the query image and the retrieved descriptions.

From this unrealistic assumption of finding complete descriptions relevant to query images, Kuznetsova et al. (2012) proposed a holistic data-driven approach that extends

the previous work conducted by Ordonez et al. (2011). Kuznetsova et al. (2012) proposed a framework that starts by running similar detectors and classifiers which were used in the first re-ranking step of Im2Text. Similarly, this step was carried out for extracting and representing the semantic meaning of a query image. Instead of performing a single retrieval process, as was performed by Ordonez et al. (2011), the designed approach was intended to perform a separate retrieval process for each detected image entity to collect additional descriptions. Specifically, this step aimed to retrieve four different types of phrases. Noun and verb phrases were retrieved for each query object detection using a visual similarity measure computed as a combination of histogram distances based on colour, texton, HOG (Dalal and Triggs, 2005) and SIFT (Lowe, 2004) visual features. Similarly, the authors retrieved region and stuff prepositional phrases (e.g., "*in the street*") by measuring appearance similarity and geometric arrangements. Additionally, scene prepositional phrases (e.g., "*on a rainy day*") were collected based on global scene similarity computed by the distance between scene classification score vectors. The final and composite image description composed of the retrieved phrases was generated via Integer Linear Programming (ILP) (Schrijver, 1998). With a similar concept, the authors proposed a tree-based method for composing image descriptions by making use of already annotated web images. After retrieving visually similar images and extracting their relevant captions, the authors considered the extracted phrases as tree fragments and modeled the description composition as a constraint optimisation problem, encoded by ILP and solved by using CPLEX solver[4]. Ordonez et al. (2015) later constructed a large collection of human-annotated images retrieved from the web-based Flickr[5] photo collection. This was the first attempt in using the Web as an intermediate source for obtaining human-annotated image descriptions. As in previous studies, descriptions were formulated by sythesising descriptions of visually similar images to the query image.

By prioritising linguistics over CV, Gupta et al. (2012) used a phrase-based approach for retrieving and composing image descriptions. For retrieving visually similar images, the authors used simple RGB and HSV colour histograms for extracting colour features, while Gabor (Kamarainen et al., 2006) and Haar (Lienhart and Maydt, 2002) descriptors were used to extract texture properties. Similar to previous works, GIST (Friedman, 1979) features were used for the extraction of scene characteristics and SIFT (Lowe, 2004) were used to capture shape properties. Rather than applying object detectors and scene classifiers for extracting the semantic aspect of the input image, the authors relied completely on retrieved descriptions of visually related images. These retrieved descriptions were then parsed into specific phrase structure. Typical examples include: *(subject, verb), (sub-*

[4] `https://www.ibm.com/analytics/cplex-optimizer`
[5] `https://www.flickr.com`

ject, *prep*, *object)*, *(verb*, *prep*, *object)*. From these phrase-based descriptions, the best description was selected with a joint probability model, built on image similarity and Google search counts. After representing images as triplets in the form of *((attribute1*, *object1)*, *verb)*, *((verb*, *prep)*,*(attribute2*, *object2))*, *(object1*, *prep*, *object2)*, image descriptions were generated from the top-scoring triplet based on a fixed template. Additional syntactic and predicate grouping rules were applied for better quality.

Retrieval-based approaches can also retrieve captions from a learned common multimodal space, intended to combine visual and textual data from a collection of image-description pairs. From this joint representation, models are capable of performing cross-modal retrieval (Bernardi et al., 2016). One of the first studies which incorporates the visual and textual domains was proposed by Farhadi et al. (2010). The authors suggested a multimodal space of image meanings, consisting of triplets in the form *(object*, *action and scene)*. This intermediate representation was limited to a set of pre-defined discrete values for each slot in the triplet. To eliminate this limitation, Hodosh et al. (2013) utilised Kernel Canonical Correlation (KCC) (Bach and Jordan, 2002; Hardoon et al., 2004) to induce a common space by finding linear projections from the two domains. This technique proved particularly successful in combining images (Hardoon et al., 2004) or image regions (Socher and Fei-Fei, 2010) with specific words or a list of tags.

Although retrieval based models generate grammatically sound and fluent descriptions, these methods are constrained by already existing captions ready to be reused. Images having concepts which are not present in the available collection cannot be described effectively with the most appropriate vocabulary. Under extreme conditions, such models can even go out of scope and thus render irrelevant image captions. To address this problem, Birmingham and Muscat (2017) developed a web-retrieval framework to caption images by text found from the web. The proposed approach was to exploit Google search by image to find visually similar images to the query ones. Descriptions were then extracted from the web pages from where the visually similar images were retrieved. In the next section, methods exploiting deep neural architectures are presented and organised according to how deep learning is applied.

2.5.3 Template based Image Captioning

The models that cast image captioning as a generation-based problem are constrained by a syntactic and semantic process. Typically, these methods first predict the visual content of the images by making use of different visual detectors. At this stage, different CV techniques are applied to extract the scene type, detect the occurring event actions and recognise the objects visible in an image, together with the attributes and relationships

between them. Natural Language Generation (NLG) models are followed to turn visual detectors' outputs into a natural language descriptions.

Generation-based description models differ in two main aspects, namely in the way that images are represented and in the linguistic approach adopted to generate textual descriptions. Traditional language models turn triplets composed of objects, attributes and relations (e.g., (wooden, boat) in (large, field)) to sentences like: "*A wooden boat parked in a large field*" by adding functional words to the triplets. Another approach is to compose captions by using templates with linguistic constraints. In this case, visual detectors detect triplets following a specified template, for example (attribute_1, object_1, preposition, attribute_2, object_2) which can therefore generate captions such as "*A [attribute_1] [object_1] [preposition] [attribute_2][object_2]*". Since this method is very rigid, a "randomised local search" (Chisholm, 2002) was proposed as an iterative method which repetitively selects a position to edit (insert, delete or replace). An edit is kept the score is improved, otherwise the next step is executed until convergence or a maximum number of iterations is reached. Another approach is to train a language model on labelled images to obtain word's statistical information based on n-grams (Yang et al., 2011). This method works by evaluating the probability of generating word ω_i given the preceding words and the remaining objects as defined in Equation 2.12

$$P_r(\omega_i|\omega_{i-1},\ldots,\omega_1,V_{i-1}) = \frac{\exp\left[\sum_{k=1}^{K}\lambda_k f_k(\omega_i,\omega_{i-1},\ldots,\omega_1,<s>,V_{i-1})\right]}{\sum_{v\in V\cup</s>}\exp[\sum_{k=1}^{K}\lambda_k f_k(v,\omega_{i-1},\ldots,\omega_1,<s>,V_{i-1})]}, \tag{2.12}$$

where $<s>$ and $</s>$ are the start and end tokens of the caption respectively, $f_k(\omega_i,\omega_{i-1},\ldots,\omega_1,<s>,V_{i-1})$ calculates the maximum entropy of the k-th feature, while and λ_k is the corresponding weight. The language model is trained by maximum log-likelihood estimation as follows:

$$\mathcal{L}(\theta) = \sum_{s=1}^{S}\sum_{i=1}^{|(s)|}\log P_r(\omega_i^{(s)})|\omega_{i-1^{(s)}},\ldots,\omega_1^{(s)},<s>,V_{i-1}^{(s)}), \tag{2.13}$$

where (s) is a sentence found in the training data, $|(s)|$ is the length of the current sentence s and θ represents the model parameters.

One of the works which follows a template-based approach is the work presented by Yang et al. (2011), where the quadruplet composed of nouns, verbs, scenes and prepositions was used as a sentence template for generating descriptions. Images were first described by using detection algorithms (Felzenszwalb et al., 2010; Oliva and Torralba, 2001) to extract the main image objects and to understand the scene type of an image.

The authors opted to train a language model (Dunning, 1993) on the Gigaword corpus[6] to predict verbs, scenes and prepositions. The probability for verb v was obtained by $P_r(v|n_1, n_2)$ given the two nouns based on the corpus. Similarly, scenes were predicted according to the objects and verbs. Based on the probabilities computed for all the elements, the best quadruplet is chosen by Hidden Markov Model (HMM) inference which is then transformed to the final image description by filling the full sentence structure.

Kulkarni et al. (2013b) generated descriptions by utilising a CRF to determine the main image concepts to be mentioned in the generated caption. Graph nodes corresponded to objects, object attributes and spatial relationships between objects, in which unary potential functions of nodes were obtained by visual models. On the other hand, pairwise potential functions were obtained statistically from a collection of existing descriptions. Image content to be mentioned in the description is selected from the output generated after performing CRF inference. This output is then used to generate the final description based on a sentence template. In a similar fashion, Li et al. (2011) and Mitchell et al. (2012) represented images in the form of tuples to capture different image concepts, including the scene type and the detected image objects together with their attributes and spatial relationships. Specifically, Li et al. (2011) employed visual models for detecting objects, attributes and spatial relationships from images. The output of the detectors was then encoded in a triplet of the form *((adj1*, *obj1)*, *prep*, *(adj2*, *obj2))*. To capture such triplets, the authors opted for web-scale *n*-gram data to provide frequency counts of possible valid sequences that match the triplet format. Phrase fusion subsequently ends this process by finding the optimal compatible set of phrases via dynamic programming. Similarly, Mitchell et al. (2012) employed detectors to represent images in triplets of the following structure: *(objects*, *actions*, *spatial relationships)*. Based on the visual recognition results, image descriptions are formulated as a tree-generating process. Starting from clustering and ordering object nouns, the authors proposed to generate sub-trees for each object noun. A trigram language model (Koehn, 2005) was then used to select the final sequence from the generated trees as the description for the query image.

To explicitly combine the visual representation of an image with the linguistic sentence structure, Elliott and Keller (2013) proposed the Visual Dependency Representation (VDR). The main idea is to conceptualise image descriptions by modeling the spatial relationship between image objects in the form of a dependency graph relating to the syntactic dependency tree of image descriptions. Descriptions were then generated by traversing the VDRs to fill in sentence templates. Inspired by this cohesive image-textual representation, Lin et al. (2015) described indoor images by representing visual content as

[6] https://catalog.ldc.upenn.edu/LDC2003T05

scene graphs. Similar to VDRs, scene graphs were used to represent images based on the depicted objects, their attributes and relations between them. Multi-sentence descriptions were finally generated by parsing scene graphs with a semantic grammar. Inspired by this work, Gilberto Mateos Ortiz et al. (2015), described abstract scenes by translating VDRs to corresponding textual descriptions using a machine translation-based model.

Yatskar et al. (2014) managed to generate descriptions from human densely-annotated images describing the salient image regions. A maximum entropy language model was then used to handle the output generated from word detectors. In contrast to this region labelling, Fang et al. (2015) used multi-instance learning to train visual detectors for detecting words that occur frequently in image captions, including many linguistic components such as nouns, verbs and adjectives. The words predicted by the visual detectors are then passed to a language model for generating image descriptions. These revised methods are all based on detected image concepts described by single words that are connected to formulate template based sentences. Instead of structuring image descriptions based on single words, Ushiku et al. (2015) proposed to generate captions composed of phrases. The authors presented the Common Subspace for Model and Similarity to learn phrase classifiers directly for image captioning. This was achieved by extracting sequence of words from training captions as phrases. Phrase features and corresponding image features are mapped in one space, where similarity based and model based classification are integrated to learn a classifier for each phrase. The description is then formulated by connecting estimated phrases through multi-stack beam search. Another beam search based approach that was applied for generating image captions is the Grid Beam Search (GBS) (Hokamp and Liu, 2017). This algorithm was proposed to allow the inclusion of pre-determined lexical constraints during sequence generation. Constraints can be single or multi-word tokens and can be applied simultaneously. By using this approach, particular phrases determined by the visual detectors can be specified to be present in the generated captions. Similarly, Anderson et al. (2017) proposed to enforce the inclusion of selected tag words at test time by using a constrained beam search.

Template based image captioning is able to generate both well structured sentences and relevant captions for images. However, such models suffer from constraints and limitations imposed by design. Being highly dependent on visual models for capturing the main image concept, template based descriptors tend to generate captions with limited coverage. Furthermore, the compositional and structural rigidness make the sentences generated far less natural when compared to human authored image captions.

2.5.4 Deep Learning in Image Captioning

Inspired by the advances in the field of DL, hand-engineered feature extraction and shallow models used in early works, started being replaced by deep neural architectures. Building on top of retrieval-based methods, several researchers explored deep models to cast both retrieval and generation based techniques as a multi-modal embedding and re-ranking problem. During this phase, new language models were proposed to mitigate the problems of earlier models which were difficult to train and which occupied considerable storage space in case of n-gram based models. Neural-based language models include the RNN (Mikolov et al., 2010), LSTM (Hochreiter and Schmidhuber, 1997), GRU (Cho et al., 2014a), and the Transformer model (Vaswani et al., 2017). In contrast to previous language models, these neural-based models are more flexible and can be trained to generate diverse image captions in an end-to-end manner using image-caption pairs.

2.5.4.1 Recurrent Neural Network (RNN)

RNN-based models are types of neural networks which use predicted output as input during training (Mikolov et al., 2010). These recurrent-based neural networks keep an internal memory state, known as the hidden state vector, to keep track of previous inputs. The RNN layer takes an input to remember (e.g., a word in a sentence) and the previous hidden state and outputs a new state. This approach leads to a final hidden state which represents the full sentence. Formally, an RNN can be represented with the following recurrence formula with function f_W with parameters W applied at each timestep t as follows:

$$h_t = f_W(h_{t-1}, x_t), \tag{2.14}$$

which receives a hidden state h_{t-1} at iteration timestep $t-1$, current input vector x_t to generate current state h_t and bias b.

The simplest RNN (commonly referred to as Vanilla RNN) is composed of a single hidden state which uses a function that updates the state h as a function of the previous state h_{t-1} and the current input x_t. Weight matrices W_{hh} and W_{xh} are used to project the previous state and the current input. These are then summed and squashed with $\tanh$ function so that state h_t is updated at timestep t. This is formally defined as:

$$h_t = \tanh(W_{hh} \cdot h_{t-1} + W_{xh} \cdot x_t + b). \tag{2.15}$$

Due to the inherent vanishing and exploding gradient problems the RNN cannot capture long dependencies. This happens as gradients during back-propagation keeps diminishing during each timestep when using squashing functions or explodes when eliminating such function. Hence, modeling long sequences of inputs gets impractical as it can get extremely slow due to insignificant gradient updates or early inputs being ignored. To mitigate these problems, ReLu activations were used and gradients were clipped up to a certain value (Pascanu et al., 2013).

2.5.4.2 Long Short-Term Memory (LSTM)

To specifically address the vanishing and exploding gradient problems, the LSTM (Hochreiter and Schmidhuber, 1997) network, which is more sophisticated than the RNN was proposed. In addition to the hidden state h_t, an LSTM has an additional cell state c_t to store long-term information. Intuitively, this recurrent network learns to read, erase and write information to and from the cell state. This is carried out via three gates, namely the *input* (i), *forget* (f), and *output* (o) gates which their values vary between 0 (closed) to 1 (open) by a sigmoid function. These control how much information is passed to the cell state. Similar to the cell state c_t, these gates are vectors of size n. For a given input vector x_t, previous hidden state h_{t-1} and previous cell state c_{t-1} the LSTM during each timestep t, the next hidden and cell states h_t and c_t, respectively as follows:

$$
\begin{aligned}
f_t &= \sigma(W_{hf} \cdot h_{t-1} + W_{xf} \cdot x_t + b_f) \\
i_t &= \sigma(W_{hi} \cdot h_{t-1} + W_{xi} \cdot x_t + b_i) \\
o_t &= \sigma(W_{ho} \cdot h_{t-1} + W_{xo} \cdot x_t + b_o) \\
g_t &= \tanh(W_{hg} \cdot h_{t-1} + W_{xg} \cdot x_t + b_g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
\tag{2.16}
$$

where $\odot$ is the pointwise operator and g_t is an intermediary cache used to calculate c_t.

2.5.4.3 Gated Recurrent Unit (GRU)

The GRU is a variant of the LSTM network which combines the *forget* and *input* gates in one gate, commonly referred to as the *update* gate. This architecture also combines the

cell and hidden states as follows:

$$
\begin{aligned}
z_t &= \sigma(W_{zh} \cdot h_{t-1} + W_{zx} \cdot x_t + b_z) \\
r_t &= \sigma(W_{rh} \cdot h_{t-1} + W_{rx} \cdot x_t + b_r) \\
\tilde{h}_t &= \tanh(W_{\tilde{h}h} \cdot (r_t \odot h_{t-1}) + W_{\tilde{h}x} \cdot x_t + b_z) \\
h_t &= (1 - z_t) \odot \tilde{h}_t + r_t \odot h_{t-1}
\end{aligned}
\tag{2.17}
$$

Apart from these two popular RNN-based variants, there is a plethora of LSTM-based architectures in the literature. In fact, Jozefowicz et al. (2015) evaluated more than $10,000$ networks and found that none of these variants outperform the LSTM and GRU consistently in all their experiments. Other LSTM-based recurrent neural networks are either purely based on the LSTM architecture or else integrated-based LSTM networks. The latter consist of LSTM networks and other components such external memory unit (e.g., in Neural Turing Machine (NTM) (Graves et al., 2014)) and CNNs (e.g., in C-LSTM (Zhou et al., 2015)), while the former are other variants of the LSTM. These include the Stacked LSTM Network which adds capacity and depth to the network and the Bidirectional LSTM Network (Graves and Schmidhuber, 2005) which, like the Bidirectional Recurrent Neural Network (BRNN) (Schuster and Paliwal, 1997), is trained in both directions simultaneously with separate hidden layers.

2.5.4.4 Transformer

Recently, the Transformer (Vaswani et al., 2017) network has become the most popular sequence-to-sequence model. As opposed to RNN, this network is not recurrent and hence can be parallalised and trained much faster. This network is based on an encoder-decoder architecture and relies completely on attention. The encoder maps the full input sequence into higher abstracted representation by N blocks composed of multi-head self-attention and a simple position-wise fully connected network. These are connected by a residual connection (He et al., 2016) and followed by layer normalization (Ba et al., 2016). Conversely, the decoder has a third sub-layer which performs multi-head attention over the output of the encoder in addition to the two sub-layers of the encoder. The Transformer works by calculating attention weights between each and every token simultaneously. The network learns three matrices, namely the query (W_Q), key (W_K), and value weights (W_V). For each token i that is encoded with word embedding x_i and positional information through $\sin$ and $\cos$ functions, the query ($q_i = x_i W_Q$), key vector ($k_i = x_i W_K$) and value vector ($v_i = x_i W_V$) are computed and attention weights are calculated by taking the dot product between the query (q_i) and key (k_j) vectors for the respective two tokens i and j, respectively. To stabilise the gradients during training the attention weights are

normalised by the dimension of the key vectors ($\sqrt{d_k}$) and passed through as softmax function. The attention computation can be formally defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V \tag{2.18}$$

This self-attention based mechanism attends to tokens that are relevant to each corresponding token. The transformer model leverages multiple attention heads to perform attention based on different definitions of relevancy. A multi-head attention module is therefore defined as follows:

$$\begin{aligned} \text{MultiHead(Q,K,V)} &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \end{aligned} \tag{2.19}$$

where QW_i^Q, KW_i^K, and VW_i^V are learnable projections matrices.

2.5.4.5 Deep Learning in Retrieval and Template based Models

Yagcioglu et al. (2015) proposed to retrieve descriptions from visually similar images through an average query expansion approach dependent on distributional semantics. In this work, the authors represented images by the activations of the seventh hidden layer of the Caffe deep learning architecture (Jia et al., 2014) trained on ImageNet[7]. To select the most ideal description, the authors used a query expansion as an average of the distributed semantics of the retrieved descriptions weighted by image similarity.

Socher et al. (2014) used neural networks as a representational model for images and descriptions. More specifically, the authors used a Dependency Tree Recursive Neural Network (DT-RNN) for generating language vectors and the end result of a nine-layer neural network trained on ImageNet for representing images. These multimodal features were then mapped into one common space by using a max-margin objective function. Once the model is trained, correct image and sentence pairs in the common space end up with larger inner products. Image descriptions were then retrieved based on similarities between image representations and sentences as found in the projected common space.

Rather than exclusively mapping entire images and sentences in one space, Karpathy et al. (2014) extended this multi-modal space by mapping image and sentence fragments into one common space. They used dependency tree relations as sentence fragments and the output result of a R-CNN as image fragments. By representing both image and sentence fragments as feature vectors, the authors used a structured max-margin objec-

[7] `https://image-net.org`

tive to map the two domains in one space. Similarity between images and sentences was based on fragment similarities which makes sentence ranking more fine-grained.

To find similarities between images and sentences having different levels of interactions between them, Ma et al. (2015) proposed a multimodal Convolutional Neural Network (m-CNN). This architecture is based on three distinct components. It includes image CNNs to encode visual data, matching CNNs to combine both visual and textual data, and multi-layer perceptrons to evaluate the compatibility between vision and text. The final matching score between images and captions was computed by an ensemble of m-CNNs. To map images and sentences, Yan and Mikolajczyk (2015) proposed to use DCCA (Andrew et al., 2013). Visual features were extracted via Deep Convolutional Neural Network (D-CNN) while a stacked network was employed for textual feature extraction from TF-IDF. Correlation analysis was used to combine both visual and textual features in a joint latent space by maximizing the correlation between paired features.

Deep learning was also used for template based methods. In fact, Lebret et al. (2015) leveraged soft templates to generate captions using deep neural models. The authors used SENNA software[8] to extract sentences from training data and to obtain statistics from the extracted phrases. Phrases were encoded by high dimensional vectors and images by features extracted form D-CNNs. Query images were described by phrase inference using a bilinear model trained as a metric between image and phrase features.

Although the introduction of deep learning improved the performance in both retrieval and template based methods, the main limitations of both models still make the methods susceptible to generate constrained and less natural image captions.

2.5.4.6 Multimodal Learning

The limitations imposed by retrieval and template based methods were mitigated by approaches based on multimodal learning. Such methods were designed to generate captions by not relying on existing captions or assumptions about sentence structures. These approaches tend to generate more expressive captions with better linguistic structure. Generating image captions using multimodal neural networks is one of these approaches.

The general pipeline of multimodal learning is to first extract image features using visual feature extractors, such as CNNs. These features are then passed to a neural language model intended to map image features into a common space and perform sentence generation. This is handled by predicting sequence of words conditioned on the extracted image features and previous predicted words. Kiros et al. (2014) proposed to use a log-bilinear neural language model conditioned on image visual features. In this

[8] `https://ronan.collobert.com/senna`

method, image visual features were extracted via a D-CNN and joint image-text feature learning was performed by back-propagating gradients from the loss function through the multi-modal neural network model. Based on this principle, Mao et al. (2015), used RNN language model to directly model the probability of generating words conditioned on a given image and previously generated words. To better align image regions and sentence segments, Karpathy and Fei-Fei (2015) presented an approach to learn a multimodal Recurrent Neural Network (m-RNN) model to specifically generate descriptions for image regions. After performing the required encoding for both modalities, a structured objective function was used to map visual and textual data into a common shared space. Since RNNs are not designed to learn long term dependencies, they can render themselves limited in image captioning (Bengio et al., 1994; Mikolov et al., 2010). To address this problem, Chen and Lawrence Zitnick (2015) proposed to dynamically build a visual representation of the image for which a caption is being generated so that long term visual concepts can be referred to during the entire captioning process.

2.5.4.7 Encoder-Decoder Framework

Inspired by neural machine translation (Cho et al., 2014b; Kalchbrenner and Blunsom, 2013), researchers view the generation of image captions as sequence-to-sequence problem and proposed the encoder-decoder framework (Vinyals et al., 2015; Xu et al., 2015). It is argued that the process of image caption generation (G) can be cast as a translation problem (Wu et al., 2016), where the input is an image (I) and the output is its corresponding sentence (s) as follows:

$$G : I \rightarrow s \tag{2.20}$$

These models use CNNs to encode images by visual embeddings extracted from the output of the last fully connected layer or from the last hidden layer of the network. These visual embeddings are used to guide RNNs while being trained on word embeddings computed for corresponding ground-truth captions. A model M is designed to generate captions by maximising the probability $P(s|I)$ by:

$$G(I) = \underset{s}{\text{argmax}}\, P(s|I) \tag{2.21}$$

In order to optimise the parameters for $P(s|I)$ the model is trained to minimise the cross-entropy loss by:

$$\mathcal{L} = -\sum_{i}^{N}\sum_{w}^{W} \log P(s_{i,w}|s_{i,0:w-1}, I_i), \tag{2.22}$$

where i is the ith instance in training data, while W corresponds to the total number of words found in sentence s_i.

As discussed in Tanti et al. (2018), an encoder-decoder model can integrate the image and word embeddings either in (a) init-inject (b) pre-inject (c) par-inject, or (d) merge approach. In init-inject, the initial hidden state of the RNN is initialised with the visual features, while the pre-inject mechanism uses the image vector as the first input to the language model. The par-inject models take both image and linguistic embeddings as input, while the merge-based architectures leaves the visual embeddings out of the RNN and combines both modalities before predicting each word. For instance, Kiros et al. (2015) introduced an encoder-decoder pipeline intended for both image description generation and phrase re-ranking. The model was designed to learn a multi-modal joint embedding representational space from images and corresponding descriptions. The authors encoded descriptions using a LSTM network and images by features extracted from a CNN. The second phase of this model was intended to generate novel descriptions by decoding the multi-modal space via their suggested neural language model. On a similar approach, Vinyals et al. (2015) also used a CNN-based encoder and LSTM to decode the extracted visual features. This framework generates descriptions by predicting the probability of a sentence conditioned on input images. Donahue et al. (2015) similarly proposed a LSTM neural-based architecture. Rather than encoding the visual space into the embedding space of hidden LSTM states, the proposed model passes a copy of the static image and the previous words to a stack of four LSTMs. Another closely related LSTM-based model was proposed by Jia et al. (2015), who included additional semantic image features as input to the LSTM.

Given the fact that captioned images are far less accessible when compared to uncaptioned images, Pu et al. (2016) proposed a semi-supervised learning method that follows the encoder-decoder framework. The method makes use of a CNN to encode images and Deep Generative Deconvolutional Network (DGDN) to decode latent image features. This encoder is proposed to provide an approximation to the distribution of the latent features of the deep network and link the latent features to generative models for image captioning. Once the training is performed, a query image can be described by an average across the distribution of the latent features.

2.5.4.8 Attention-based Image Captioning

In image captioning it is important that the main salient image contents are described clearly whilst possibly leaving out unnecessary details of less important aspects. Inspired

by the human visual attention mechanism, image caption generators based on visual attention mechanisms were recently proposed (Anderson et al., 2018; Lu et al., 2017; Xu et al., 2015). In these approaches, attention mechanism was added to the encoder-decoder framework so that captions are conditioned based on an attention criteria. In such methods, attention can be directed to different image regions at each time step during the generation process. As an example, Xu et al. (2015) proposed an attention-based encoder-decoder framework to dynamically prioritise salient image regions. By arguing that an attentive encoder-decoder models lacks global modeling abilities because of their sequential information processing behavior, Yang et al. (2016) proposed a review network to enhance the encoder-decoder framework. To address this limitation, the authors introduced a reviewer module to perform review steps on the hidden states of the encoder and to generate a thought vector at each step. During this process, attention mechanism is applied to determine weights assigned to hidden states. Based on this approach, the information encoded by the encoder can be reviewed and learned by the thought vectors to capture global properties of the input and to be used for word prediction by the decoder. In this work, the authors used the VGGNet (Simonyan and Zisserman, 2014) CNN to encode the image, while a LSTM neural network was used as a reviewer to generate the thought vectors. Other attention-based methods are based on the Transformer architecture. For instance, the Object-Semantics Aligned Pre-Training (Oscar) (Li et al., 2020) method first detects object labels to align corresponding visual features with semantic information. The motivation is that textual labels generally relate to the most prominent objects in the images and therefore, by explicitly adding image labels to the input, the visual features can be better attended during the generation process. The Oscar method was then extended by VinVL (Zhang et al., 2021a) to learn object instance-centered relationships between the visual and language domains using an adaptive pretraining scheme.

2.5.4.9 Image Captioning based on Reinforcement Learning

Recently, Reinforcement Learning (RL) (Sutton and Barto, 1998) has been introduced in image captioning to mitigate the exposure-bias and loss-mismatch problems of RNN-based models. RL aims to learn a policy that decides sequential actions by maximising the cumulative future rewards. Several RL methods have been proposed to solve computer vision related problems, such as visual tracking (Yun et al., 2017).

Encoder-decoder based image captioning models were mostly trained by Maximum Likelihood Estimation (MLE) to maximise the probability of sequential words conditioned by visual features. This limits such models to render captions by imitating word-by-word patterns as found in the available training data. As a result, captions generated from these

models tend to be templated and generic. RL with evaluation metrics (e.g., CIDEr (Vedantam et al., 2015)) as a reward allows the captioning model to explore more possibilities in the sample space and gives a better supervision signal compared to MLE.

Existing RL-based image captioning methods mostly rely on a single policy network and a reward function. Ren et al. (2017) proposed a policy network and a value network to collaboratively generate captions with a reward defined by visual-semantic embedding. Rennie et al. (2017) and Liu et al. (2017b) directly optimised image captioning systems by test rewards. The key problem of RL lies in correlating the policy and the reward parts for joint learning. To tackle this problem, Liu et al. (2018) co-train both parts in the traditional RL framework and introduced a guidance term. The latter being intended to minimise the distance between the vision-language reward and the sentence-level policy by calculating the mean squared loss. Since the vision-language reward is pre-trained with ground truth, it can be considered as the expert to measure the correlation between images and sentences.

2.5.4.10 Graph-based Image Captioning

To enhance the quality of conventional encoder-decoder models, researchers recently integrated graphs in image captioning. Rather than just encoding images using high level CNN features, researchers started encoding the structural information of images using scene graphs. This type of graph is an abstraction of objects and their complex relationships which provides rich semantic information about images. To extract this structural representation, researchers attempted to build scene graphs either (a) directly from scratch by capturing the objects, their semantic and spatial pairwise relationships using learnt detectors (Aditya et al., 2017; Xu et al., 2019; Yao et al., 2018), or (b) by extracting syntactic dependency trees and transforming them into scene graphs using rule-based methods (Gao et al., 2018; Gu et al., 2019; Yang et al., 2019), or (c) by scene graph parsers which are normally trained end-to-end through context propagation (Lee et al., 2019; Li and Jiang, 2019; Milewski et al., 2020; Wang et al., 2019; Zhong et al., 2020). Alongside the CNN extracted image features, these generated scene graphs are generally encoded to condition the caption prediction in two different ways. The simpler and less computationally demanding approach is applied by encoding the semantic relation triplets using a combination of word embeddings. For example, (i) by applying mean pooling over the word2vec (Mikolov and Dean, 2013) feature vectors of the relationship triplets (Li and Jiang, 2019); (ii) by concatenating the objects, attributes and relationships embeddings into one feature vector (Xu et al., 2019); (iii) or simply by projecting the detected visual relations of the scene graph parser into a lower dimensional feature vector (Lee et al.,

2019). The second approach is to apply Graph Neural Networks (GNNs) like GCN or Graph Attention Network (GAT) as seen in Gu et al. (2019); Milewski et al. (2020); Wang et al. (2019); Yang et al. (2019); Yang et al. (2020); Yao et al. (2018); Zhong et al. (2020). These neural networks are designed to encode the graph-internal context by encoding nodes with the information of their neighboring nodes. Apart from conditioning the sequence of image captions based on detected information, researchers are now exploiting terms that are semantically related to the objects depicted in images from external knowledge graphs, such as ConceptNet (Liu and Singh, 2004), to also condition the sequence of words on non-visible entities as performed in Huang et al. (2020); Zhou et al. (2019). Semantically related words are also being used to broaden the vocabulary set of the trained image captioning models by being injected directly during sequence prediction (Huang et al., 2020), while in Zhang et al. (2021b), a knowledge graph for every word in the vocabulary set was built to connect its semantically related words. This was used to enhance a Transformer based image captioning framework that rather than encoding words with their embeddings only, it represented them also by their neighboring embeddings. Both scene and knowledge graphs were used in Aditya et al. (2017) to generate captions based on scene description graphs generated using trained visual detectors while applying common sense reasoning from a constructed knowledge base.

Apart from directly introducing the structural information of images and the semantically relevant terms for better sequence modeling, graph-based image captioning systems increase the interpretability and explainability of the generation process, while enabling the possibility for error back-tracking and the support of VQA (Aditya et al., 2017). The combination of both scene and knowledge graphs opens the opportunity for image captioning systems to internally support logical and common sense reasoning. For example, questions involving object counting and visual relations can be handled through knowledge inferred from scene graphs, while the more complex and logically-oriented questions involving commonsense reasoning can be resolved using the interplay between scene and knowledge graphs as proposed by Aditya et al. (2017). The introduction of graphs in image captioning has not yet shown any major improvements over the performance of current state-of-the-art image captioning systems, except in controllable image captioning with the recently proposed Abstract Scene Graph (ASG)-based framework (Chen et al., 2020a). In fact, Milewski et al. (2020) found no significant enhancements between models that use scene graphs over models which use object detection features only. This was particularly noticed since the quality produced by the scene graph generation process was very low and therefore introduced considerable noise in the overall captioning process. When compared to the Bottom-Up Top-Down (Up-Down) (Anderson et al., 2018) attention-based model, which is one of the current benchmark and state-of-the-art attention-based

models, the authors report that the quality of the image captions when generated based on high quality scene graphs improved by up to 3.3 CIDEr points. This confirms the findings of Wang et al. (2019) which show that state-of-the-art scene graph parsers when trained on good quality graphs can boost performance almost as much as the ground-truth graphs. This was further confirmed by Tripathi et al. (2021b) who showed that an encoder-decoder framework conditioned solely on repurposed scene graphs can reach the state-of-the-art models which use image visual features as well, indicating that scene graphs are very promising to represent images for image caption generation. These studies point out that the main bottleneck of the image captioning process depends more on the overall approach rather than the encoding process.

The captioning models are reaching a performance plateau possibly because they are generally being built on the contemporary encoder-decoder framework and consequently are inheriting its major limitations. Although the conventional approach has been improved in several ways, ranging from the inclusion of attention mechanisms and graphs to the application of RL and Generative Adversarial Networks (GANs), these systems are still heavily dependent on the current neural-based sequence models that are trained to map the extracted features using a maximimisation-based decoding approach such as beam search, which most often produce generic and bland sequences of words (Holtzman et al., 2020). Since these models are generally trained to learn the distributional characteristics of the captions found in the training data by maximising the likelihood of each consecutive word, generally they end up producing syntactically plausible captions but at the same time tend to assign low probabilities to less frequent and unseen combinations of image visual concepts, for example in scenes including, a “man wearing a horse mask”, or “person standing on the wing of a plane”. In such cases, the trained models need to go far beyond just “remembering” and reusing sequence of words as were encountered during training. In complex circumstances or environments that are out of the training context, models that are only trained using this approach find it difficult to apply common sense reasoning, and hence most of the time end up producing syntactically correct but semantically irrelevant captions with hallucinated content (Rohrbach et al., 2018) and with poor compositional generalisation (Nikolaus et al., 2019).

These issues so far were addressed by leveraging unpaired out-of-domain data (Chen et al., 2020b), by the detection of novel objects using external datasets (Agrawal et al., 2019), and through the use of both RL and GANs. Although, knowledge graphs have also been used to extend current captioning systems with commonsense knowledge (Aditya et al., 2017; Huang et al., 2020; Zhou et al., 2019), no major improvements were noted. To better encode the structural representation of images while broadening the vocabulary set used in end-to-end models, researchers recently started exploiting the graph data

structure in image captioning.

Scene-graphs (Johnson et al., 2015) were proposed as a graph structure to model the relationship between objects, attributes and relations. A scene graph is a data structure of interconnected nodes representing objects grounded in images with corresponding visual attributes connected as other nodes. Related objects are linked via pairwise relationships represented by directed edges connected between graph vertices. More formally, for a given set of object classes $\mathcal{C}$, set of attribute types $\mathcal{A}$, and a set of relationships $\mathcal{R}$, a scene graph G is a directed graph defined by tuple $G = (O, E)$, where $O = \{O_1, \ldots, O_n\}$ is the set of all object instances which can include people ("*man*"), places ("*lake*"), things ("*boat*"), or parts of other objects ("*arm*"), and $E \subseteq O \times R \times O$ is the set of directed edges that reflect the relationship between objects. Examples of relationships include geometry ("*person on boat*"), actions ("*man driving boat*"), and object parts ("*boat has engine*"). Each constituent object is defined as $o_i = (c_i, A_i)$, where $c_i \in \mathcal{C}$ denotes the object class, while $A_i \subseteq \mathcal{A}$ are the attributes of that specific object which can describe its color ("*boat is white*"), shape ("*hull is round*"), and pose ("*man is bent*"). This representation for encoding semantic features has been widely used in visual tasks, which include, but not limited to, image retrieval (Johnson et al., 2015; Wang et al., 2020), image generation (Herzig et al., 2019; Johnson et al., 2018), Visual Question Answering (VQA) (Ben-Younes et al., 2019; Ghosh et al., 2019), and image captioning (Li and Jiang, 2019).

Another type of graphical structure proposed in image captioning is the general knowledge graph. This was proposed as an intermediary representation to infer both direct and indirect image concepts which cannot be simply detected by object recognition modules. By exploiting relational information retrieved from connected semantic concepts, researchers suggested the use of general knowledge graphs to better automate visual reasoning as well as to infer commonsense facts about visual scenes by leveraging readily available background knowledge. This data representation consists of graph nodes that represent general concepts for semantic labels that are linked by directed or undirected edges to encode relational facts between pairs of concepts from unstructured data.

2.5.4.11 Scene Description Graph

The first attempt where the graph data structure was used in automatic image captioning was in the work proposed by Aditya et al. (2017). To enhance the flexibility of image captioning, the authors suggested the extraction of Scene Description Graph (SDG) through visual image understanding and commonsense reasoning applied over an automatically constructed knowledge base. Common sense reasoning related to natural activities was particularly applied on (a) detections retrieved from existing visual perception modules, (b)

a “commonsense” knowledge base which was constructed from image annotations using natural language processing, and on (c) lexical ontological knowledge bases such as WordNet (Miller, 1998), that links words into semantic relations including synonyms (e.g., *car* is a synonym of *automobile*), hyponyms (e.g., *sparrow* and *eagle* are hyponyms of the *hypernym* bird), and meronyms (e.g., *finger* is a meronym of *hand* because a *finger* forms part of a *hand*). Inspired by the fact that the human visual processes continuously interact with high-level knowledge during understanding, the authors proposed to decompose their captioning framework in three interacting modules, namely the (1) visual detection, (2) knowledge base, and (3) logical reasoning modules. To model the early stages of human understanding, the authors integrated deep learning-based vision and state-of-the-art concept modeling from the constructed commonsense knowledge base. Specifically, the deep learning based visual module was responsible for (a) the detection of objects and regions (e.g., *man*, *wooden floor*), scenes (e.g., *beach*, *stadium*), visual relations including verbs and spatial prepositions between two objects or an object and a scene (e.g., *a man holding a ball*; *a man standing on the floor*), attributes and (b) visual attention.

Visual information was extracted by deep object and scene recognition modules and scene constituent recognition. The latter was trained on constituent annotations collected using Amazon Mechanical Turks (AMT) in free-text format, including not only the detected objects within images, but also what each object is doing and its properties. A multi-label SVM (Boser et al., 1992) was trained on deep visual features extracted from images using a pre-trained CNN model (Krizhevsky et al., 2012). Based on these visual detections, the authors proposed to generate a SDG, which is a directed labelled graph that represents the semantics of a scene through the interaction between Entities (objects, regions), Events (actions, linking verbs), Traits (attributes of objects and regions), and inferred scene constituents. This graph-based structure was also intended to represent the semantic relations (from KM-Ontology (Clark et al., 2004)) between Entity-Event and -Trait pairs, and the spatial orientation between entities as was suggested by Elliott and Keller (2013). This intermediate representation was introduced to move beyond visual analysis and tackle Event-Entity based analysis, enhance the generation of image captions, perform visual question answering, and reason beyond what can be seen in images. Image captions were generated using a template-based approach by applying the SimpleNLG (Gatt and Reiter, 2009) package to produce captions from the SDG.

2.5.4.12 Knowledge Graphs in Image Captioning

To extend image captioning with general and commonsense knowledge, Zhou et al. (2019) proposed to integrate knowledge graphs alongside the visual features extracted from im-

ages. The authors in this work hypothesised that if background knowledge is integrated in image captioning, such models can produce captions with information that is not explicit in the image and therefore makes the generated captions more effective. For example, when considering a photograph of a boat near a slipway, it could be effectively described by "*A boat ready to be loaded on a trailer*", even if the trailer is not present in the image. Generating these type of captions requires not just visual recognition but also the need for external knowledge to embed additional information. To take advantage of readily available background knowledge, the authors leveraged information encoded in knowledge graphs and integrated it with an already existing neural image captioning method introduced in Vinyals et al. (2015); You et al. (2016). The standard approach of first having a CNN trained to encode images into a fixed length vector space representation and then used as an initial state vector to train an RNN to produce corresponding sequences of words was extended with knowledge graphs in two aspects. Rather than just encoding images with CNN features, the authors also detected a set of objects from images using the YOLO9000 (Redmon and Farhadi, 2017) object detection framework. ConceptNet (Liu and Singh, 2004; Speer et al., 2017), a labelled knowledge graph that represents the commonsense relationship between words and phrases of natural language, was used to infer both direct and indirect terms related to the detected objects based on a cosine distance similarity score between their semantic vector representations as generated by retrofitting (Faruqui et al., 2015) on ConceptNet.

Direct terms were retrieved based on the individual objects, while indirect terms were collected based on the whole set of detected objects. The related terms, the detected objects and the image features were used to pre-train an RNN network based on corresponding ground-truth captions. The output from this intermediate caption generation along the two terms, the detected objects and the visual embeddings, were then used as an initial state of an LSTM-based RNN language module intended to decode these embeddings into corresponding image captions. The evaluation of this proposed framework revealed that the introduction of external concepts gathered from a semantic knowledge graph to a neural image captioner, in some metrics achieves better performance when compared to state-of-the-art captioning models (Donahue et al., 2017; Vinyals et al., 2015; Xu et al., 2015; Yao et al., 2017; You et al., 2016; Zhou et al., 2016) which do not make explicit use of external background knowledge.

In contrast, Huang et al. (2020) used the ConceptNet (Liu and Singh, 2004) knowledge graph to inject semantically related information related to the detected objects into the output stage of the caption generator by augmenting the probability of some latent meaningful words at each decoding step. This allows the system to generate more novel and meaningful captions. In this work, a new text dependent word attention was added

to the uniform visual attention model. Its calculation depends only on the internal annotation knowledge as found in the training data which provides rich semantic information to guide the generation of visual attention to handle the common discordant matching problem between regions in image and words in captions. To guide the model on the visual domain, the authors used the region proposal network proposed by Ren et al. (2015) to generate the region proposals which were then fed to the ROI pooling layer and 3 fully connected layers to obtain a vector representation of each image region. From the qualitative results, it was found that the use of knowledge graph can bring more benefits than word attention. This showed that by the incorporation of external knowledge, the model can discover more important cues to describe a given image. Furthermore, from the qualitative analysis it was shown that this proposed model can generate more fine-grained captions that reveals more implicit aspects of images which are normally difficult to be discovered by machines. However, it was confirmed that like to most existing models which are limited by the captioning length, this model does not perform well on complex images having multiple objects.

Instead of using the conventional RNN-based models for sequence modeling, Zhang et al. (2021b) proposed to use the Transformer (Vaswani et al., 2017), a model architecture which eschews recurrence and instead relies on an attention mechanism to draw global dependencies between input and output which makes it highly parallelisable and faster to train. To be able to use the Transformer model for image caption generation, the authors used the pre-trained bottom-up attention features (Anderson et al., 2018) as input after it was reshaped through a linear layer. To further improve the performance of the Transformer-based model, the authors proposed to leverage a knowledge graph. Rather than using only the embeddings of each single word found in the vocabulary as the input, the authors also exploited the neighbouring words as embeddings so that the model can leverage the information of the related words during training. For each word, a knowledge graph was constructed with nodes corresponding to the word's top most related words as measured by cosine similarity. The embedding of each word was then replaced by the combined embedding composed from the embedding of the word itself and the embeddings of its neighbours which were then projected into a single joint fixed-length vector using a fully connected layer.

2.5.4.13 Scene Graphs in Image Captioning

Numereous works have recently proposed the scene graph data structure as an intermediate representation between images and corresponding natural language descriptions. The visually grounded scene graphs are generally generated by following a RNN-based

approach which predicts image objects and their corresponding relationships. Normally, the first step towards generating the scene graph is to generate a set of initial bounding boxes for the image. For each object proposal, the object category as well as its bounding box offsets are predicted and the pairwise relationships are considered to construct the scene graph. As pointed out in the literature (Li and Jiang, 2019), relationship triplets (i.e., $< object >< relationship >< subject >$) are used since (a) a triplet corresponds to two entities and an edge is generally treated as the basic element; (b) a triplet can be considered as a small subgraph of an entire scene graph which provides discriminitive and informative cues for the generation of captions; and (c) individual predicates of relationships are ambiguous while a triplet conveys more representative visual content.

For instance, Gao et al. (2018) proposed a framework that is split in two separate phases, namely the concept cognition and sentence construction, where the former was intended to build a vocabulary of 267 semantic concepts extracted from the vocabulary of Fang et al. (2015) which includes nouns, verbs, adjectives and several prepositions. A subset of the words which were collected from image annotations were mapped to the extracted vocabulary set (e.g., "*baby*" was mapped to "*children*"), while a collection of words were left unmapped to limit the number of semantic concepts and to reduce noise during the generation of scene graphs, since on average, each image was estimated to approximately have nine semantic concepts. The first phase was also intended to generate the high-level semantic representation in the form of a scene-graph-based sequence through the use of a novel CNN-RNN-SVM framework.

Sequences were generated in three steps. In the first step, captions were parsed in scene-graph tuples using SPICE (Anderson et al., 2016). Afterwards, these tuples were combined with corresponding semantic concepts to build reference scene graph tuples via a rule-based approach such that the following conditions are met: (a) all words in tuples which have a corresponding semantic concept in the vocabulary must be mapped (e.g., "man" is replaced by "people"), otherwise they are not considered to be part in the final scene graph; (b) scene graph must conform to a tree-based structure to eliminate the possibility of having multiple relationships between two objects; and (c) limits the subjects of all objects to not have any relationship with other objects. Since it is difficult to train images with scenes directly (Anderson et al., 2016), the authors proposed to transform the generated scene graphs into sequences. The sequence that gave best results was composed of: *"the subject*, *attributes of the subject*, *relationship between the subject and object1*, *object1*, *attributes of object2*, *relationship between the subject and object2*, *object2*, *attributes of object2*, *…"*. The graph-based sequences together with corresponding images were then passed to their proposed novel framework to extract visual features from the CNN module and model concept relationship and dependency by an LSTM module.

By taking the embedding of the predicted concept at each time step and maintaining a hidden state, this module was responsible for recognising concept co-occurrence information. The a-priori probability of a semantic concept given the previously predicted labels were computed according to their dot product in relation to the sum of image and recurrent embeddings. Path prediction was performed by obtaining the product of the prior probability of each label given the previous concepts in the predicted path.

Both vision and language embeddings were projected to the same low-dimensional space as concept embedding through a data fusion layer. The authors introduced the SVM classifier to detect and filter out semantic concepts which were not considered suitable for the scene-graph based sequence. Such concepts include those adjectives which are weak-dependent on others. These concepts were automatically classified based on their CNN features and semantic concepts were predicted by the CNN-RNN module. The output of the SVM was integrated into the graph based sequence to generate a bit vector with a length equal to the vocabulary size. The second phase was responsible for the sentence reconstruction based on low-level CNN visual features and high-level semantic representation through the use of an LSTM. The effectiveness of having a rich and diverse vocabulary was confirmed by the authors when evaluated the system using current popular evaluation metrics on two vocabularies, one containing nouns only, and another with the full set of concepts, including verbs, adjectives and prepositions. The authors confirmed that the RNN module was exploiting statistical frequency between phrases and therefore common words like “people” and “play” had a greater chance of being included in the prediction irrespective of the extracted image features. To mitigate this problem and to enforce attention on visual features, the authors trained the framework on random semantic sequences which led to an improved performance.

Xu et al. (2019) have also proposed the scene graph as intermediate representation in their image captioning framework. The training was fully supervised in such a way that all images had corresponding scene graphs and bounding boxes reflecting the objects, attribute and relationships found in images. Scene prediction was handled by training two models based on CNN (VGG-16 (Simonyan and Zisserman, 2014)) to predict attributes and relationships for objects. At test time, the Faster-RCNN (Ren et al., 2015) was employed to detect objects, while two fine-tuned 16-layer VGG network models (Simonyan and Zisserman, 2014) were used to predict attributes and object pair-wise relationships to construct scene-graphs. The top-1/5 accuracy reported on the intersection between Visual Genome (Krishna et al., 2017) and COCO (Lin et al., 2014) for the attribute and relationship modules were $13.5\%/33.25\%$ and $11.75\%/29\%$ respectively. This shows the difficulty in predicting the two concepts as both can be humanly subjective and can therefore lead to a high degree of inaccuracy in scene graph generation. To embed constructed

scene-graphs in an image caption generator, the authors proposed the Scene Graph Captioner (SGC), a module responsible for the integration of semantic concepts, the toplogical structure and the attention region of scene graphs. The SGC was split in three phases.

The first phase was responsible for capturing both the semantic and structural information of images and model it in two individual graph embeddings, namely the concept and topology vectors. The second phase of the pipeline was dedicated to attending the most important regions within images, whilst the third module was the language model designed to decode the extracted semantic concepts, topological structure and the attention region into textual descriptions. During the first phase, a semantic vocabulary that is not tense or plural sensitive (e.g., *ride* and *riding* are considered as one concept) was constructed from the available scene graphs. The concepts were then divided into three components, namely objects, attributes and relationships. A total of 4096 were selected as the most common words which included 2000 objects, 1000 attributes and 1096 relationships. The prediction probabilities from all concepts were then aggregated to construct a multi-label concept representation. To embed the scene graph, the authors also extracted a topological fixed-vector to represent the structure of the scene graph.

This was handled by first generating an adjacent matrix where the objects and relationships of the graph are used as vertices and edges respectively. For consistency across all images, the authors proposed to construct a matrix size of $m \times m$, where $m = 2000$ to reflect the number of all possible objects. The prediction probabilities of relationships between two objects was reflected as a weight value in the adjacent matrix. The matrix was then transformed into a pseudo-colour map having the corresponding colour intensities for the adjacent matrix. Its toplology representation was then extracted from the output of the last fully-connected layer (conv8) after feeding the extended adjacent matrix to a shallow convolutional neural network CNN-M-128 (Chatfield et al., 2014).

The second phase of the pipeline was dedicated to attend the most important regions within images. In this phase, an attention mechanism module was employed to extract high internal homogeneity and external inhomogeneity among the nodes of the constructed scene graph which corresponds to the region where the nodes tend to dense cluster together in a graph. The proposed attention extraction model was meant to extract attention graphs by taking into consideration the edge structure of the scene graph. Based on the dominant clustering approach, a cluster is generalised for the entire vertex set of an input graph. All the related visual regions are absorbed to compose the attention bounding box from the image to guide the caption generator with the most salient image features as extracted from a CNN network (VGG-16 (Simonyan and Zisserman, 2014)).

The third module was finally trained to maximise the probability of correct captions conditioned by the images. In contrast to previous models of this type, this model, apart

from the CNN features extracted from the whole image and the attention region; semantic concepts and topological structure representations were additionally used. These were employed to condition the sequences of words by an LSTM-based language module.

The results based on standard captioning metrics showed that the larger the concept vector is, the higher the quality of the generated captions. A comparison between the generation of captions based solely on the nonextended and extended adjacent topological matrices revealed that the latter slightly improves the quality of the former but still does not reach the accuracy obtained when using the concept feature vector by a high margin, confirming that the topological vector is not very effective. When combined together, both concept and topological vectors did not outperform the model trained on a 4096 concept feature vector. The overall SGC showed incremental improvements over the baseline model which generated captions conditioned only on extracted image visual features. The visual features combined with the attention mechanism slightly improved the baseline results, while when the image features were combined with both concept and topological feature vectors the results continued improving. This confirms that information extracted from concept and topological features of scene graphs is more important than the visual information extracted from the visual attention region. Furthermore, when using the attention region features together with the topological and concept features, the performance went down which surprisingly showed the importance of the whole image features over the attention region. When including all vectors, which consist of the CNN features for the whole image and the attention region, the concept vector and the topological feature vector, the SGC improved its generated captions by an overall average increase of 2.32% over all computed metrics.

Yang et al. (2019) proposed incorporating the inductive bias of language generation into the conventional encoder-decoder framework to leverage the strengths of both symbolic reasoning and end-to-end multi-modal feature mapping. This framework was proposed to generate more human-like captions by reducing any dataset bias through the exploitation of language inductive bias. Specifically, the authors proposed the Scene Graph Auto-Encoder (SGAE) which is based on a sentence self-reconstruction network to learn the feature representation of the language inductive bias. This was carried out by transforming sentences ($\mathcal{S}$) into corresponding sentence graphs ($\mathcal{G}$) using the SPICE (Anderson et al., 2016) metric. After transforming the graph node embeddings into a new set of context-aware features through the use of spatial graph convolution, the authors proposed to embed the language inductive bias in language composition using a trainable dictionary ($\mathcal{D}$). Motivated by the use of working memory to preserve a dynamic knowledge base for run-time inference which is widely used in textual QA (Sukhbaatar et al., 2015), VQA (Xiong et al., 2016), and one-shot classification (Vinyals et al., 2016), a dic-

tionary was proposed in the sentence self-reconstruction network.

Sentences were reconstructed by decoding this dictionary using a trainable RNN-based language decoder (Anderson et al., 2018) with RL-based training strategy (Rennie et al., 2017) with the following overall pipeline: $\mathcal{S} \rightarrow \mathcal{G} \rightarrow \mathcal{D} \rightarrow \mathcal{S}$. By having this trainable dictionary also shared in the encoder-decoder pipeline ($\mathcal{I} \rightarrow \mathcal{G} \rightarrow \mathcal{D} \rightarrow \mathcal{S}$), the language prior was also integrated to guide the end-to-end image captioning during language decoding. The $\mathcal{I} \rightarrow \mathcal{G}$ step was handled by a visual scene graph detector composed of a Faster-RCNN (Ren et al., 2015) to detect the objects, MOTIFS relationship detector (Zellers et al., 2018) to classify the relationships and their own attribute classifier based on a single hidden layer network with ReLu activations (i.e., fc-ReLu-fc-Softmax). The $\mathcal{G} \rightarrow \mathcal{D}$ denotes a multi-modal GCN which is added to integrate necessary visual cues not detected by the visual detection module. This step was included to modulate scene graphs into visual representations by fusing detected label embeddings with visual features extracted from ROI extracted from Faster R-CNN.

The same authors later extended this work and in Yang et al. (2020) presented their proposed framework with a more fine grained dictionary. Specifically, it was partitioned into three distinct dictionaries, where each dictionary was responsible for handling the object, attribute and relation inductive bias. Furthermore, to transfer the inductive bias from the pure language domain to the vision-language domain, the authors proposed to use the concept of Knowledge Distillation (KD) (Hinton et al., 2015) through a Kullback-Leibler divergence (Kullback and Leibler, 1951) objective. This was computed between the word probability distributions generated by the RNN decoder of the sentence auto-encoder and the overall image captioning decoder. Also, to continue improving the scene graph encoding, both GCNs were replaced with attention-based graph convolution networks to generate better scene graph representations.

Li and Jiang (2019) fused visual and semantic features extracted from scene graphs by mean pooling to obtain the integral representation. Once these features were obtained, a hierarchical-attention-based module was used to discriminate features for word generation at each time step. The first level attention was used to selectively attend to different visual and semantic features to form the weighted visual and semantic context vectors. Instead of simply pooling these two vectors into a single one and without taking into account their inherent structures and differences between them, the attention module was proposed to learn relevance scores for these two modalities in order to obtain the final integrated context vector with the second-level attention.

In contrast to all previous works, Zhong et al. (2020) rather than using the full scene graphs as an intermediate layer between images and corresponding sentences, they proposed to perform image captioning by decomposing the generated scene graphs of im-

ages extracted by the MotifNet (Zellers et al., 2018) neural network into a set of sub-graphs to capture the different semantic components of images. The backbone of this method selects important sub-graphs and decodes a single caption from the chosen sub-graph. The presented deep learning based model was trained to choose the important sub-graphs using a sub-Graph Proposal Network (sGPN) and decode them in their target sentences in order to attend to different image regions. After sampling the sub-graphs, meaningful graphs were identified by first combining the visual and textual features on the scene graph nodes and edges respectively, followed by an integration of contextual information using a GCN. The GCN was used to combine information from the neighborhood within the graph and update node and edge features. With this updated scene graph and a set of sampled sub-graphs, the model was designed to learn a score function to select and rank sub-graphs for image captioning based on ground-truth captions to guide the learning process. The decoding process incorporates an attention mechanism on the sub-graph nodes during the generation of each caption word. This approach lead to the generation of accurate, diverse, grounded and controllable captions simultaneously for the first time. This means that controlling the number of sub-graphs results in a more diversified image captioning.

For better user intention controllability, Chen et al. (2020a) proposed the novel ASG as a fine-grained control signal to condition controllable caption generation. The ASG is a directed graph which consists of three abstract nodes grounded in images, including the object, attribute and relationship without having any actual semantic labels. The chosen ASG is transformed to an image caption by their proposed ASG2Caption model which is based on an encoder-decoder framework. This model was designed to (a) capture both intentions and semantics from the graph through a proposed role-aware graph encoder used to differentiate fine-grained intention roles of nodes and includes graph context in nodes to improve the semantic representation. Apart from controlling the content to describe via graph nodes, the ASG was also responsible to (b) implicitly decide the order of description through the way nodes are connected without omitting and repeating any content during the process. In this work, instead of using a fully detected scene graph, the ASG was adopted as a control signal for the generation of intention-aware and diverse image captions. ASGs are convenient for user intractability and controllability, while they are easier to generate automatically when compared to full scene graphs. Since the ASG is a graph layout without any semantic labels it can be easily generated manually or automatically by having an off-the-shelf object proposal network and a binary relationship classifier trained to detect whether two objects are related or not. Therefore, to generate diverse captions users can simply select sub-graphs from full ASGs or else an automatic sampling strategy can be used to automate the generation of different cap-

tions involving varied image aspects. In order to generate image captions based on the encoded ASG, a language decoder was specifically designed. This includes a graph-based attention mechanism which attends both the graph semantics through graph content attention, and structures via their proposed graph flow attention. Moreover, in contrast to previous attention-based models (Lu et al., 2017; Xu et al., 2015), this graph-based decoder was also designed to keep track of what has been attended to during the decoding process via an 'erase' followed by 'add' operations inspired by the Neural Turing Machine (NTM) (Graves et al., 2014) to prevent content omission and duplication during the decoding process.

Yao et al. (2018) explored the explicit use of visual relationships via scene graphs for the generation of image captions. By connecting the extracted visual embeddings of the main image objects with both semantic and spatial relations, GCNs were used to produce relation-aware region-level representations. In this work, the authors used the Faster-RCNN (Ren et al., 2015) to detect objects within images to encode the images into a set of salient regions containing objects. Both semantic and spatial relation graphs were then constructed over all the detected image regions based on the corresponding semantic and spatial relationships in the form of *subject-predicate-object*, where the latter describes the geometric orientation between the subject and object, and the former describes the action or interaction between pairs of objects. The spatial graph was constructed by classifying the geometrical orientation between each pair of detected objects by exploiting the Intersection Over Union (IOU) of the two objects, the relative Euclidean distance as measured from the two bounding boxes' centroids and the relative angle between the two centroids. These extracted features were used to categorise the spatial linkage within the spatial graph into 11 spatial categories. Furthermore, the semantic relations were predicted by using a simple deep classification model based on the union of two bounding boxes which covers the two respective objects. The regions as well as their relationships were encoded using a GCN to produce relation-aware region representations. The standard GCN was modified to preserve graph directions and labels, while an additional edge-wise gate unit was added to automatically focus on important edges. The relation-aware region representations were then passed to a two-layer LSTM-based captioning framework which enables region-level attention in image captioning. The semantic and spatial graphs were linked using a late fusion scheme by linearly fusing the predicted word distributions from the two separate decoders.

Similarly, Lee et al. (2019) extended the top-down image captioner (Anderson et al., 2018) by adding relation features extracted from a neural scene graph generator. In this work, the Stacked Motif Network (Zellers et al., 2018) was used to predict graph elements by specifically staging bounding box predictions, object classifications, and relationships

in such a way that the global contextual encoding of all previous stages provides rich context for predicting the following stages. The top-down captioner was modified so that the input vector of the attention-based LSTM at each time step is replaced by the concatenation of the mean-pooled relation feature, the mean-pooled region feature, the previous output of the language LSTM, and an encoding of the previously generated word.

To specifically study to what degree scene graphs have on the performance of image caption generation, Wang et al. (2019) incorporated scene graphs generated by Factorizable Net (Li et al., 2018) into the bottom-up top-down attention based image captioning architecture (Anderson et al., 2018). This effectiveness was analysed in the context of both the predicted and ground-truth scene graphs. From this work, similar to other reviewed works, it was re-confirmed that scene graphs can improve the generation of image captions. Furthermore, it was found out that although scene graphs generated by current state-of-the-art models are still limited in the number of objects and relations categories, the results produced are not way off from those produced when using the ground-truth scene graphs. This shows that the main bottleneck in current scene-graph-based image captioners is not owing to inaccurate scene graph generation but it is more related to the overall process used in captioning models. To come to this conclusion, the authors integrated the off-the-shelf scene graph parser with the attention-based image captioning framework proposed by Anderson et al. (2018) while including information from the original image through a set of region features obtained by an object detection module to improve performance. The scene graphs were encoded into contextual hidden vectors using the GCN proposed by Marcheggiani and Titov (2017) which incorporates directions and edge labels in the encoding, while also allowing edge-wise gating to let the network learn to prune invalid connections found in the generated scene graph. The used architecture was composed of an attention-based LSTM that is responsible for tracking contextual information from the inputs and which incorporates information from the decoder. Specifically, the used attention-based LSTM takes contextual information after concatenating the previous hidden state of the decoder, the mean-pooled region-level image features which include bounding box coordinates obtained from the Faster-R-CNN (Ren et al., 2015), the mean pooling of the scene-graph node features from the GCN, and the previous generated word. Words were decoded from the inputs of the previous hidden state of the attention-based LSTM, attention weighted scene graph node embeddings and attention weighted image features. When evaluating this work on standard evaluation metrics, the introduction of scene graph features with visual information it was shown to improve the captioning results when compared to the results generated based on image features only or graph features only. The authors also evaluated this model with ground-truth scene graphs and showed that the most notable improvement of 2.1% was noted on the

SPICE (Anderson et al., 2016) evaluation metric since the ground-truth scene graphs have a larger vocabulary.

Furthermore, as an attempt to study whether scene graphs are currently good enough to be used in image captioning and whether they can improve the quality of the generated captions, Milewski et al. (2020) explored the use of different graph-based architectures to fuse both object and relation information from images. In this work, the authors presented an extension for GAT (Veličković et al., 2018) by presenting a novel Conditional Graph Attention Network (C-GAT), which in contrast to a standard graph attention layer, is designed to condition scene graph updates on the current state of the image captioning decoder. The overall pipeline for the image captioning process was to first predict the scene graphs for images by making use of the pretrained scene graph generator through iterative message passing as proposed in Xu et al. (2017) with a relation proposal network (Yang et al., 2018) to obtain and inject relational information into the image captioning framework. Scene graphs were encoded by using both GAT and C-GAT, while flat versus hierarchical attention mechanisms were used. The authors found that most scene graphs had extremely low quality and because of their noise resulted in a reduction in the performance of the captioning process. For this main reason, models that leverage scene graphs had no significant difference in the reported performance than those which are not based on scene graphs. However, when the quality of the generated captions was analysed according to the quality of the generated scene graphs, it was confirmed that the quality of captions was improved with high quality scene graphs.

To address the ambiguity of whether scene graphs are generally useful (Wang et al., 2019; Yang et al., 2020) or not (Li and Jiang, 2019; Milewski et al., 2020) in image captioning, (Tripathi et al., 2021b) suggested to repurpose the visual scene graphs for caption generation and proposed the SG2Caps architecture which is based on an encoder-decoder framework that is conditioned on scene graphs only and which does not make use of any extracted image features. To enhance the effectiveness of the visual scene graphs that were generated using MotifNet (Zellers et al., 2018) after being trained on Visual Genome (Krishna et al., 2017), the authors proposed to extend these graphs with Human Object Interaction (HOI) information. This extension was primarily introduced to highlight the main regions concerned with human-to-object interactions, since humans normally tend to describe images involving persons by focusing only on the involved human-to-object interaction and excluding any other background information (Tripathi et al., 2021b). Furthermore, in contrast to all previous works, the authors extended the visual scene graphs with the spatial information of the nodes extracted from the corresponding object bounding boxes and post-processed the predicted scene graphs to reduce their noise to make them more suitable for image caption generation. This was

achieved by eliminating less confident graph nodes predictions, applying non-maximum suppression in object detection, and by manually mapping the detected object categories to the closest word string in the captioning dataset. Furthermore, in contrast to all previous models, no image or object-level visual features were used to train the proposed SG2Caps model. This framework encodes the visual scene graph to generate a context-aware embedding using five spatial graph convolutions for the object, bounding box, relationship, and attribute embeddings as was performed in Yang et al. (2019). This was then passed to an LSTM-based language decoder followed by reinforcement learning optimisation. From their evaluation, it was confirmed that a repurposed visual scene graph can provide enough information for generating high-quality image captions while substantially reducing the trainable parameters needed for image caption generation.

2.5.4.14 Unpaired Image Captioning

Most of the reviewed contributions in image captioning assume datasets composed of image and caption pairs. For instance, retrieval-based mechanisms reuse captions of visually similar images, while deep learning based models are trained on image and caption pairs. The same applies to the traditional encoder-decoder frameworks which make use of both image and caption pairs as inputs during training. Given the difficulty of collecting such data to train scalable image caption generators, recently researchers started addressing the problem of Unpaired Image Captioning (UIC) which assumes to have both image and caption pairs coming from the same domain.

The first contribution which addressed UIC was the pivot-based approach presented in Gu et al. (2018b). Despite not having the image-sentence pairs in their target language, this framework used a paired image-caption dataset in the pivot language (Chinese) and a machine translation dataset to translate the pivot-language and the target-language (Chinese-English). The pivot-language sentences were connected in different domains by shared word embeddings. Recently, scene graphs have also been proposed in UIC (Cao et al., 2020; Gu et al., 2019; Liu et al., 2019a). For instance, Gu et al. (2019) proposed a framework that is composed of an off-the-shelf image and sentence scene graph generators, a scene graph encoder, an attention-based sentence decoder composed of two LSTM layers, and a feature alignment module which maps graphs encoded in the image domain to the textual domain. The first step was to extract scene graphs from the sentence corpus and train the scene graph encoder and the sentence decoder based on textual information. The encoding was performed by minimising the cross-entropy loss and fine-tuning by RL based on the CIDEr metric (Vedantam et al., 2015) as the reward. Furthermore, the encoding was optimised by minimising the negative expected rewards

as proposed in (Anderson et al., 2018; Rennie et al., 2017).

To align both image and sentence scene graphs, the authors included CycleGAN (Zhu et al., 2017b), a cycle-consistent feature alignment module used to generate the data correspondence between the two domains. This was achieved by having two mapping functions that map image to sentence embeddings and vice-versa for objects, relations and attributes. These were implemented as fully connected layers with leaky ReLu activations. Furthermore, two discriminators were trained to distinguish the *real* features of the original modality from the *fake* features that were mapped by the respective functions which were trained to fool the corresponding discriminators though adversarial training. Given the unpaired image and scene graphs, this was accomplished by first encoding the two graphs using the scene graph encoder trained on the textual domain. The two encoded graphs were then mapped through feature alignment by an unsupervised cross-modal feature mapping. By this mapping mechanism, the encoded image scene graphs based on text corpora end up being close to the corresponding sentence modality which is consequently used as input to the sentence decoder for the generation of image captions. Both image and sentence scene graphs were generated in the exact way as performed previously in Yang et al. (2019), where the image scene graph was generated by the Faster R-CNN (Ren et al., 2015) and MOTIFS (Zellers et al., 2018) object and relationship detectors respectively, while object attributes were identified by the same module proposed in Yang et al. (2019).

The sentence scene graphs were generated by first transforming sentences into syntactic trees using (Anderson et al., 2016) which internally uses a syntactic dependency tree built by Klein and Manning (2003). The tree was then transformed into a scene graph by using a rule-based method (Schuster et al., 2015). Again, both graphs were encoded in a similar fashion to how the scene graphs were encoded in Yang et al. (2019) by exploiting three different spatial graph convolutional encoders to encode the three types of graph nodes by taking into account their neighbouring information. Since these three embeddings differ in their size and represent different information, their importance for decoding varies. For this reason, three independent attention modules were used to extract the most relevant context from each embedding. The attention vectors were then combined in a triplet embedding by a neural network which was then fed to an RNN-based decoder. Liu et al. made use of semantic concepts and their relationships to combine the vision and language modalities. Their proposed approach first extracts visual concepts, including nouns, attributes and relations from both images and captions. Concepts from images were extracted via a weakly-supervised method of MIL, while words used in the ground-truth captions and a pre-defined set of concepts were used as semantic concepts for captions. A semantic relationship explorer module was then proposed to

explore the relationship between the two sets of concepts. An attention-based LSTM decoder was used to decode the embeddings of the semantic relationships extracted via multi-head attention (Vaswani et al., 2017). Ben et al. (2022) recently proposed the Semantic-Constrained Self-Learning (SCS) framework which iteratively generates pseudo captions via a pre-trained captioner and re-trains the captioner which makes use of adversarial training (Goodfellow et al., 2014). Objects detected in images guide both stages of the framework and hence strengthen the alignment between the images and the output captions. The authors proposed to use an object inclusion and adverserial rewards to favour captions with predicted objects and the generation of human-like captions by using generative networks.

2.5.4.15 Unsupervised Image Captioning

To further reduce the problem of image-caption pairs, recently researchers proposed models for unsupervised image captioning. In contrast to the unpaired setting (discussed in Section 2.5.4.14), unsupervised image captioning assumes to have the image and captions coming from different domains. Feng et al. (2019) were the first to propose a fully-unsupervised framework which generates pseudo image-sentence pairs by training a model that maps visual concepts of images to sentences collected from the Web through image and sentence feature alignment using adversarial text generation (Fedus et al., 2018). Laina et al. (2019) built a joint embedded space composed of image and sentence features based on visual concepts. A language model was first trained to encode captions into semantically structured embeddings. Image features that are mapped to this embedding space were decoded into captions using the same language model. Cao et al. (2020) after encoding the images using their proposed Residual Network (ResNet)-based architecture, they extracted relations between objects using their proposed Mutual Attention Network (MAN). These features were used to align images and captions crawled from Shutterstock[9] in an adversarial way.

2.5.5 Datasets

Current image caption generators mainly use the **COCO** (Common Objects in Context) (Lin et al., 2014) dataset which was created by Microsoft in 2014. This large-scale dataset is not only used for image captioning but can also be used for tasks including image recognition, object detection and semantic segmentation. The dataset features 80 main object categories found in everyday life images. COCO comprises of $82,783$ training im-

[9] Shutterstock is a stock photography website hosting millions of images with corresponding human authored descriptions

ages, $40,504$ images for validation and $40,775$ test images which are not publicly available. Each image has five or in some cases, six human authored captions collected via Amazon Mechanical Turk (AMT). Other commonly used datasets found in the literature are the **Flickr8K** (Hodosh et al., 2013) and **Flickr30K** (Plummer et al., 2015) datasets which are based on images collected from the photo sharing Flickr[10] website. Flickr8K consists of $8,000$ images which mainly feature humans and animals. Five corresponding descriptions were also collected via crowdsourcing using AMT. On the other hand, Flick30K extends Flickr8K and consists of $31,783$ images with five corresponding captions per image. Other datasets which are used in image captioning include the **Visual Genome Dataset** (Krishna et al., 2017) which is composed of $108,077$ images, where each image contains an average of 35 objects, 26 attributes, and more than 5.4 million descriptions for image regions. Less commonly used datasets include, the **Pascal1K** dataset (Rashtchian et al., 2010) which is a small-scale dataset that consists of images that were selected from the Pascal 2008 object recognition dataset (Everingham et al., 2010). Each image is described with five human authored descriptions collected using AMT. The **SBU1M** web-scale dataset (Ordonez et al., 2011) has one million images collected from Flickr together with its user-provided descriptions. Furthemore, **IAPR-TC12** (Grubinger et al., 2006) is a collection of $20,000$ images which have to five descriptions in multiple languages, including English, German and Spanish.

2.5.6 Evaluation Metrics

Evaluating the quality of the generated captions is a critical and difficult problem in image caption generation. Assessing the quality of captions can be carried out by human evaluation; however, this is expensive and can take considerable time and effort. Furthermore, human judgments may vary and can lead to inconsistent evaluation. Therefore, in recent years, automatic metrics have been proposed to evaluate the generation of text. These include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) metrics which were adopted from machine translation and document summararisation, whilst CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) were later proposed specifically for image captioning. The BLEU metric measures the n-gram precision, ROUGE considers the n-gram recall, and METEOR takes into account the precision, recall and synonyms. On the other hand, CIDEr makes use of TF-IDF to weight n-grams and calculates cosine similarity between captions. In order to measure the semantic relatedness which n-gram based metrics do not consider, SPICE constructs scene graphs

[10] `flickr.com`

of reference and candidate captions and compares them based on an F-score computed over triplets composed of objects, attributes and relationships.

2.5.6.1 BLEU

The Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) metric computes the similarity between the generated and ground-truth captions by computing the precision p of the overlapping n-grams, where n usually varies between 1 and 4. The precision between a candidate (c) and a reference (r) caption is computed by:

$$p = \frac{\text{count}(n\text{-gram})}{\text{len}(r)}, \tag{2.23}$$

where count$(n$-gram$)$ is the number of matched n-grams and len(r) is the length of the reference caption.

However, as this can lead to overvalued precision when over-generating commonly used words (e.g., "the" in caption: "the the the the the the the"); the precision is modified by first counting the number of overlapping n-grams by count$(n$-gram$)$. Secondly, the maximum number of times a given n-gram is found in any given reference caption is found so that each candidate n-gram count is clipped to that maximum count. These counts are added and divided by the total number of unclipped candidate n-grams. This is formally defined as:

$$p_n = \frac{\sum_{c \in C} \sum_{n\text{-gram} \in c} \text{count}_{\text{clip}}(n\text{-gram})}{\sum_{c \in C} \sum_{n\text{-gram} \in c} \text{count}(n\text{-gram})} \tag{2.24}$$

Since precision does not cater for short sentences which do not capture the most relevant aspects of captions, BLEU uses a brevity penalty (BP) to penalise short captions that lack completeness by:

$$\text{BP} = \begin{cases} 1 & \text{if len}(c) > \text{len}(r) \\ e^{(1-len(r)/len(c))} & \text{if len}(c) \leq \text{len}(r) \end{cases}, \tag{2.25}$$

The BLEU score is then computed by:

$$\text{BLEU} = \text{BP} \cdot \exp\Big(\sum_{n=1}^{N} w_n \log p_n\Big) \tag{2.26}$$

where N is normally set to 4 and w_n to $1/N$.

Given that image captioning datasets normally consist of images with corresponding reference captions (R), where the latter are generally sets of five captions per image, the precision p_n for the generated captions C is calculated as follows:

$$p_n(C,R) = \frac{\sum_i \sum_k \min[h_k(c_i), \max_{j \leq m} h_k(r_{ij})]}{\sum_i \sum_k h_k(c_i)}, \tag{2.27}$$

where i is the ith image and j is the jth out of m reference captions of image i. $h_k(c_i)$ is how many times n-gram ω_k is found in the generated caption c_i, while $h_k(r_{ij})$ is the number of n-grams found in the reference caption r_{ij}. The BP and final BLEU score can then be computed by Equations 2.25 and 2.26 respectively.

2.5.6.2 METEOR

In contrast to the BLEU score, the Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Denkowski and Lavie, 2014) takes ordering into consideration when computing the similarity between matching words. Apart from the precision, METEOR computes also the recall and harmonic average. For instance, if the reference captions has w_r words and the candidate caption has w_c and m is the number of common words between the two captions, then precision P is calculated by $P = \frac{m}{w_c}$ and recall $R = \frac{m}{w_r}$. Therefore the harmonic mean is calculated by $F_{mean} = \frac{PR}{\alpha P + (1-\alpha R)}$ between the best candidate and reference caption. Since this considers only the matching of single words, this score introduced a penalty factor (pen) to give weight for longer matched chunks:

$$METEOR = (1 - pen) \times F_{mean}, \tag{2.28}$$

where the penalty pen equates to $\gamma(\frac{ch}{m})^\theta$ given that m and ch is the number of matched unigrams and chunks, respectively; while α, γ and θ are default parameters which are normally set to 3, 0.5, and 3, respectively.

2.5.6.3 ROUGE-L

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) is a collection of metrics proposed to assess the quality of text summarisation. The most popular metric of ROUGE is ROUGE-L. This considers the longest common subsequence (LCS) without a pre-defined length of n-gram between the candidate (c) and reference (r) captions as follows:

$$\text{ROUGE} - L = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}}, \tag{2.29}$$

where $R_{LCS} = \frac{LCS(c,r)}{|r|}$,$P_{LCS} = \frac{LCS(c,r)}{|c|}$ and $\beta = \frac{P_{LCS}}{R_{LCS}}$.

2.5.6.4 CIDEr

In contrast to the previous metrics which were first proposed for machine translation, the Consensus-based Image Description Evaluation (CIDEr) (Vedantam et al., 2015) metric was the first metric which was specifically intended to evaluate the quality of image caption generation. This metric first converts words into their stem or root form and considers each sentence as a set of n-grams, where n ranges from one to four. The intuition behind CIDEr is to quantify the number of important overlapping n-grams found between the candidate c_i and reference r_{ij} captions for image i, while giving less priority to common n-grams found in the dataset. This is achieved by weighting each n-gram ω_k found in the jth reference caption of image i (r_{ij}) by a Term Frequency-Inverse Document Frequency (TF-IDF) score $g_k(r_{ij})$ as follows:

$$g_k(r_{ij}) = \frac{h_k(r_{ij})}{\sum_{\omega_l \in \Omega} h_l(r_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(r_{pq}))}\right), \tag{2.30}$$

where $h_k(r_{ij})$ is how many times an n-gram ω_k is found in the jth reference caption of image i, Ω is the set of all n-grams, while I is the total number of images.

The first term of Equation 2.30 computes the TF of ω_k to give weight on n-grams which occur frequently in the reference captions, while the second logarithmic term evaluates the IDF to give less weight to n-grams commonly found in the dataset. The CIDEr$_n$ for n-grams of size n is calculated by the average cosine similarity between the candidate and reference captions as follows:

$$\text{CIDER}_n(c_i, R_i) = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(r_{ij})}{||\mathbf{g}^n(c_i)||\ ||\mathbf{g}^n(r_{ij})||}, \tag{2.31}$$

where $\mathbf{g}^n(c_i)$ is the vector composed of $g_k(c_i)$ for all n-grams of size n and $||\mathbf{g}^n(c_i)||$ is the magnitude of $\mathbf{g}^n(c_i)$. The same applies for $\mathbf{g}^n(r_{ij})$.

Variable n-grams (i.e., $1 < n \leq 4$) are used to capture the grammatical correctness of the generated captions. These are combined as one CIDEr metric by:

$$\text{CIDER}(c_i, R_i) = \sum_{n=1}^{N} w_n \text{CIDER}_n(c_i, R_i), \tag{2.32}$$

where empirically, it was found that $N = 4$ and $w_n = 1/N$ give best results.

2.5.6.5 SPICE

The Semantic Propositional Image Caption Evaluation (SPICE) (Anderson et al., 2016) metric is another metric which was proposed to specifically assess the quality of candidate caption c against a set of reference captions R. This metric prioritises the semantic propositional context of images by first transforming captions into a scene graph by using a syntactic dependency parser. The scene graph of a candidate caption is expressed by $G(c)$, while the scene graph of the set of reference captions is denoted by $G(R)$ after taking the union of scene graphs $G(s_i)\ \forall\ s_i \in S$. Captions are parsed into scene graphs as:

$$G(c) = \langle O(c), E(c), K(c) \rangle, \tag{2.33}$$

where $O(c)$ is the set of mentioned objects in caption c, $E(c)$ is the set of relations between objects, and K is the set of attributes related to the objects. The similarity between the candidate and reference scene graphs is handled by considering the scene graphs as conjunction of logical propositions or tuples by using the definition of: $T(G(c)) \triangleq O(c)\ \cup E(c)\ \cup K(c)$. Based on a binary matching operator ($\otimes$) which gives the number of tuples matched between two scene graphs,SPICE is computed based on F_1 score as follows:

$$P(c, R) = \frac{|T(G(c)) \otimes T(G(R))|}{|T(G(c))|} \tag{2.34}$$

$$R(c, R) = \frac{|T(G(c)) \otimes T(G(R))|}{|T(G(R))|} \tag{2.35}$$

$$\text{SPICE} = F_1(c, R) = \frac{2 \cdot P(c, R) \cdot R(c, R)}{P(c, R) + R(c, R)} \tag{2.36}$$

2.5.7 Discussion and Results

This section presents and discusses the performance of the reviewed image caption generators which are most relevant to the work carried out in this thesis, i.e., Keyword-driven and N-Gram Graph-based Image Captioning (KENGIC) framework. Since KENGIC decouples the generation of image captions into two phases which first proposes visual concepts/keywords related to images and the other phase translates the keywords into captions using an n-gram graph based approach, it positions itself with both unpaired and graph-based image caption generators. Given the fact that the keywords generation module can be trained independently from any image-caption pairs (as they can be collected either from human-labelled keywords or from readily available scene graphs) and

connected via a graph-based approach, KENGIC can be juxtaposed with models trained in unpaired setting and with models that make use of graphs. The reviewed models were evaluated on the current most widely used metrics including BLEU-4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), CIDEr (Vedantam et al., 2015), ROUGE-L (Lin, 2004) and SPICE (Anderson et al., 2016) as illustrated in Table 2.1. The models are categorised in two (i.e., paired (✓) and unpaired (×) as found in the "Paired" column). The presented table illustrates whether a Knowledge Graph (KG), Scene Graph (SG), or a combination of both (KG/SG) was used in each respective graph-based image captioning framework and each model is accompanied with a comment which briefly describes the method. As depicted in the table, image caption generators trained in the paired setting generally outperform those trained in the unpaired setting. However, recently proposed UIC generators are reaching competitive performance with early models. As tabulated in Table 2.1, the graph-based image captioning models are compared to three state-of-the-art and benchmark models. These include one of the pioneering works based on the conventional encoder-decoder model built on top of a CNN and a bidirectional BRNN (Karpathy and Fei-Fei, 2017) and the Bottom-Up Top-Down (Up-Down) (Anderson et al., 2018) attention-based model which is optimised on the non-differentiable CIDEr evaluation metric through RL. This model uses region-based bottom-up attention features extracted from a pre-trained object detector instead of the conventional CNN features. Furthermore, the graph-based models are compared to the Multimodal Transformer (MT)-based image captioning framework (Yu et al., 2020). This is one of the current state-of-the-art image caption generators which replaces the LSTM-based language model by using an extension of the Transformer (Vaswani et al., 2017).

Table 2.1: Results metrics of graph based models when trained on cross-entropy loss, or a combination of both (+RL) as evaluated on COCO dataset

Model	Paired	BLEU-4	METEOR	ROUGLE-L	CIDEr ↓	SPICE	Graph Type	Comment
SDG Aditya et al. (2017)	✓	5.0	10.0	-	-	-	SG/KG	Scene description graphs constructed from deep visual detection and knowledge base modules
Language-Pivoting Gu et al. (2018b)	×	5.4	13.2	-	-	-	-	Unpaired image captioning using language pivoting
Adversarial+Reconstruction Feng et al. (2019)	×	18.6	17.9	43.1	54.9	11.1	-	Unpaired image captioning using adversarial training
USGAE Yang et al. (2020)	×	17.1	19.1	43.8	55.1	12.8	SG	Unpaired Scene Graph Auto-Encoder
Multimodal Embeddings Laina et al. (2019)	×	19.3	20.2	45.0	61.8	12.9	-	Unpaired image captioning using multimodal embeddings
IGGAN Cao et al. (2020)	×	21.9	21.1	46.5	64.0	14.5	SG	Unpaired image captioning using Interactions Guide acGAN
BRNN Karpathy and Fei-Fei (2017)	✓	23	19.5	-	66	-	-	Image captioning based on Bidirectional Recurrent Neural Network
Graph-Align Gu et al. (2019)	×	21.5	20.9	47.2	69.5	15.0	SG	Unpaired image captioning via scene graph alignments through adversarial training
SGC Xu et al. (2019)	✓	23.9	21.8	48.8	73.3	-	SG	Scene graphs generated by detecting object, attributes and relationships
SCS Ben et al. (2022)	×	22.8	21.4	47.7	74.7	15.1	-	Unpaired image captioning with Semantic-Constrained Self-learning
CNN-RNN-SVM Gao et al. (2018)	✓	26.1	22.3	-	76.0	-	SG	A framework which generates a scene-graph-based sequence in the form of a bit sequence
CNet-NIC Zhou et al. (2019)	✓	29.9	25.6	53.9	107.2	-	KG	ConceptNet used to infer direct and indirect related terms to condition sequence prediction
RSG Wang et al. (2019)	✓	34.5	26.8	55.9	108.6	20.3	SG	Graph Convolution Network-based image captioning with edge-wise gating on graphs
SG2Caps Tripathi et al. (2021b)	✓	32.8	26.0	55.5	109.7	19.2	SG	Encoder decoder conditioned on repurposed scene graphs without using any image visual features
HA-SG+C-GAT Milewski et al. (2020)	✓	35.5	-	56.0	109.9	19.8	SG	Hierarchical attention with scene graph generator trained with conditional graph attention
KMSL (+RL) Li and Jiang (2019)	✓	33.8 (36.3)	26.2 (27.6)	54.9 (56.8)	110.3 (120.2)	19.8 (21.4)	SG	Know More Say Less: Image captioning using a pre-trained graph generator
Trans[D2GPO+MLE]+KG Zhang et al. (2021b)	✓	34.39	27.1	-	112.6	-	KG	Transformer enchanced with Data-Dependent Gaussian prior objective and KG
Up-Down (+RL) Anderson et al. (2018)	✓	36.2 (36.3)	27.0 (27.7)	56.4 (56.9)	113.5 (120.1)	20.3 (21.4)	-	The benchmark Bottom-Up Top-Down attention model trained by Reinforcement Learning
WA-KG (RL) Huang et al. (2020)	✓	(37.3)	(27.3)	(57.4)	(121.2)	-	KG	Knowledge graph used to inject semantically related words in captions
Sub-GC Zhong et al. (2020)	✓	36.2	27.7	56.3	115.3	20.7	SG	Sub Graph Captioning
SGAE (+RL) Yang et al. (2019)	✓	36.9 (38.4)	27.7 (28.4)	57.2 (58.6)	116.7 (127.8)	20.9 (22.1)	SG	Scene Graph Auto-Encoder that incorporates the language inductive bias framework
GCN-LSTM (+RL) Yao et al. (2018)	✓	37.1 (38.3)	28.1 (28.6)	57.2 (58.5)	117.1 (128.7)	21.1 (22.1)	SG	Uses Graph Convolutional Networks for semantic and spatial relationship graphs
SGAE-KD (+RL) Yang et al. (2020)	✓	37.3 (38.8)	28.1 (28.8)	57.4 (58.8)	117.1 (129.6)	21.3 (22.4)	SG	Scene Graph Auto-Encoder that incorporates the language inductive bias through knowledge distillation
MT (+RL) Yu et al. (2020)	✓	37.4 (40.7)	28.7 (29.5)	57.4 (59.7)	119.6 (134.1)	-	-	Multimodal Transformer based image captioning
HA-SG *Milewski et al. (2020)*	✓	*38.1*	-	*57.6*	*129.8*	*20.9*	*HQ SG*	*Hierarchical Attention with scene graph generator trained on high quality graphs*
ASG Chen et al. (2020a)	✓	23.0	24.5	50.1	204.2	42.1	SG	Abstact Scene Graph based controlled image captioning

The table of results clearly shows that the exposure bias can be mitigated by training models through the minimisation of the cross-entropy loss followed by the optimisation on the CIDEr metric. Although Aditya et al. (2017) took advantage of both scene and knowledge graphs when building SDGs for images, the applied SimpleNLG (Gatt and Reiter, 2009) language generator was not good enough to transform the generated graph into high quality captions. In fact, overall it was the least effective method among all the presented models and performed worse than the BRNN model. Interestingly, although the Graph-Align framework (Gu et al., 2019) was the only model which was trained in an unsupervised way using adversarial training on scene graphs extracted from unpaired image caption pairs, outperformed the benchmark BRNN model. In line with this work, (Yang et al., 2019) proposed the SGAE model to incorporate the language inductive bias into the conventional encoder-decoder image captioning framework by learning a trainable shared dictionary between image and sentence scene graphs. This framework was later extended in Yang et al. (2020) through a three-partition based dictionary, where each partition was responsible for separately handling the object, attribute and relations inductive bias. Also, the SGAE model was extended with the introduction of Knowledge Distillation (KD) to better transfer the extracted inductive bias from the sentence self reconstruction network decoder to the image captioning decoder (Scene Graph Auto-Encoder with Knowledge Distillation (SGAE-KD)). By being trained in a supervised way, the SGAE-based approach improved over the Graph-Align framework and surpassed the Up-Down method by a wide margin. This model slightly outperforms the Sub-GC model which is designed to caption images by automatically selecting the optimal sub-graph that best describes the main salient regions but slightly underperforms the GCN-LSTM model which employs Graph Convolutional Networks to integrate both the semantic and spatial relationship graphs in an image encoder. However, the extended version (i.e., SGAE-KD), despite being trained with a batch size of 100 (i.e., smaller than the 1024 batch size used by GCN-LSTM), it outperformed the GCN-LSTM model. Also, it is worth noticing that the SGAE-KD achieved comparable performance to the Multimodal Transformer-based model (MT), with the exception of when being evaluated using CIDEr metric. It is interesting to note that despite having a Transformer-based language model, the Trans[D2GPO+MLE]+KG method performs worse than SGAE-KD and the benchmark Up-Down method. Furthermore, the table lists (*in italics*) the performance of the HA-SG framework when being evaluated on the intersection between high quality Visual Genome scene graphs and their corresponding MSCOCO image captions. This model overall ends up improving slightly over the MT benchmark framework but at the same time exceeding the same benchmark with a wide margin of 10.2 CIDEr points. Further improvements in CIDEr score was achieved when the ASG was introduced for controlling image captioning. Although the model scores very

low in BLEU-4 (23.0), METEOR (24.5) and ROUGE-L (50.1), the controlled image captioner performed best in CIDEr (204.2) and SPICE (42.1) with 57.3% and 101.4% improvement respectively over the HA-SG model. The high CIDEr score achieved when training the latter model on high quality scene graphs confirms that with the help of better scene graph parsers that can accurately extract image structures, image captioning can improve a lot but still not enough to address the compositionality generalisation issue faced by current systems.

2.5.8 Summary

The previous sections reviewed the research in image captioning while outlining the main strengths and weaknesses of each respective technique. Models that cast image captioning as a generation based problem have the main advantage of generating novel descriptions. However, these heavily depend on accurate visual detectors and restricted to complex grammars. Since visual analysis is highly dependent on computer vision detectors, such models can be prone to generate irrelevant descriptions due to inconsistent visual content analysis. Another main problem of such models is the requirement of sophisticated natural language generation models for generating fluent and grammatically correct sentences. In contrast, retrieval-based models from visual space generate grammatically sound and humanlike image descriptions, as images are described by reusing captions of other visually related images. The main drawback of these methods is that they are dependent on a large and diverse collection of human-annotated image descriptions which might not be very accessible or too time-consuming for collection. Similarly, models that cast image description as a retrieval problem from multi-modal space are also capable of generating humanlike image captions. This is guaranteed as they are also designed to retrieve descriptions from a large collection of already annotated images.

State-of-the-art image captioning generators follow an encoder-decoder framework trained on image and caption pairs. Conventional encoder-decoder based models are trained to translate image CNN features into a sequence of words using RNN-based language models without leveraging the structural information about the images (Xu et al., 2017). Such background knowledge has been shown to be quite useful in multiple applications ranging from information retrieval to question answering (Zhou et al., 2019). Besides applying attention mechanism (Anderson et al., 2018; Gu et al., 2018a; Xu et al., 2015) to allow sentence decoders to dynamically focus on specific regions during the caption generation process, other works have applied different architectures to model the language generation process. For example, Gu et al. (2017) used a CNN-based language model to predict the sequence of words. Another theme of improvements was

to apply RL and GANs to address the exposure bias (Ranzato et al., 2016) and the loss-evaluation mismatch (Li et al., 2019b) problems in sequence prediction. The exposure bias occurs when sequence-based models are trained to maximise the likelihood of the next ground-truth word given the previous words using back-propagation, an approach which has been referred to as the "Teacher-Forcing" (Bengio et al., 2015). Consequently, at inference time, such image captioning models tend to accumulate prediction errors as they end up generating words based on their previous predicted words which they have never been exposed to during training. In RL, a recurrent model can be viewed as an "agent" that interacts with an external "environment" (words and image features). By its network parameters, it defines a policy that results in an "action" which leads to the the prediction of word sequences. After each action, the agent updates its internal "state" (cells and hidden states of the LSTM, attention weights, etc.). Once the model generates the end-of-sequence (EOS) token, the agent observes a "reward" which normally is the CIDEr score. The goal of the training is therefore to minimise the negative expected reward (Rennie et al., 2017). On the other hand, GANs generate captions by considering the production of each word as an "action", for which a reward is given from an evaluator (Dai et al., 2017a). Researchers have recently proposed the use of Transformers (Vaswani et al., 2017) in image captioning for its powerful and scalable attention-based architecture which currently is the state-of-the-art approach.

2.5.9 Outlook

Current state-of-the-art image caption generators attempt to learn the relationship between image and captions in an end-to-end manner at the expense of large-scale datasets consisting of image and caption pairs. In this PhD, KENGIC, a Keyword and N-Gram Graph-based Image Captioning approach, is proposed and developed to study the interplay between the explicit and implicit use of keywords in automatic image captioning. In contrast to current popular Deep Learning (DL) based architectures, KENGIC, casts image captioning as a search problem in an n-gram graph-based data structure by using keywords relevant to query images. Given a set of keywords and a text corpus, this approach constructs a knowledge-graph that features the visual keywords and traverses the graph to search for the best candidate caption which mentions the given set of keywords. Since this approach decouples the vision and language domains, it is comparable to direct generation-based approaches which use visual keywords as input to language modules. It also compares well to retrieval-based mechanisms as captions are being retrieved from a corpus of text but distinguishes itself from traditional-retrieval models as it does not retrieve captions from image-caption pairs. KENGIC, therefore, provides a framework for

both unsupervised image caption generation when the corpus and images are from different domains and for unpaired image captioning when both images and the corpus are of the same domain.

3 Methodology

3.1 Introduction

This chapter presents KENGIC, a Keyword-driven and N-Gram Graph based Image Captioning framework. This was purposely proposed to investigate the role of visual keywords in image captioning, while projecting insights on their importance in automatic evaluation and how these can be used to generate captions.

Inspired by how neurons are fired to activate neural pathways when humans interact with the visual world and by how mental images are constructed in the human brain (Kreiman et al., 2000), an image caption generator is proposed under the hypothesis that given a set of keywords, a caption can be constructed through an *n*-gram graph search based approach. When humans try to construct a mental image, with for example the keyword "*boat*", the mental image could vary with boats of various sizes and contexts such as "speed *boat* in a race", a "luxury *boat* at a dock", or maybe even a "fishing *boat* parked in a field near the mountains". With the introduction of further keywords, the mental image is further refined and adjusted to the original context of the subject. In this sense, the richer and relevant the keywords set is, the more accurate the mental images are constructed in the human mind. Based on this idea, a model designed to automatically connect keywords relevant to images in an n-gram graph based approach would offer an alternative research avenue in automatic image captioning. This system therefore was developed to provide answers for the following research questions:

1. Can we cast the generation of image captions based on keywords through an n-gram graph search problem?
2. What is the role of image keywords in KENGIC?
3. What quality can be obtained from such generator?

3.2 Architecture

The main objective behind this hypothesis is to develop an automated system that generates captions through the use of relevant keywords, such that the same keywords can be probabilistically linked together to form a directed graph through overlapping n-grams. The architecture is designed to first connect the main keywords by generating a knowledge graph through other intermediary n-grams. This graph is then traversed to search for paths which visit the given keywords. Nodes visited during graph walks are considered as phrases for candidate captions. Relevant captions are then selected based on a cost function which takes the following into consideration: (a) the fluency of captions by measuring how probable the sequences of words are, (b) the length of the captions, (c) the number of keywords found in the generated captions, and (d) the number of nouns which have been mentioned but not found in the given keywords set. The high-level architecture of KENGIC is split into two modules:

Vision: This module is responsible for the extraction of a set of keywords ($\mathcal{K}$) that are relevant to the query image (I). This set serves the basis for the generation of a knowledge graph ($G_{\mathcal{I},\mathcal{K}}$) which corresponds to image I based on keywords $\mathcal{K}$. Keywords that are grounded in images can be detected by either individually trained visual detectors, by scene graph generators trained to predict grounded scene graphs in images, or by multi-label models designed to predict image labels including nouns, attributes and verbs.

Language: This module, which is the core contribution of this work, handles the generation of knowledge graphs by probabilistically linking keywords $\mathcal{K}$ of image I through n-grams as found in a text corpus T. This module is designed to traverse the graph to find the most relevant caption that best describes the image based on the given keywords. A high-level architecture of KENGIC is illustrated in Fig. 3.1. The following section details the caption generation process.

3.2.1 N-Gram Graph

The proposed KENGIC approach is based on an n-gram graph data structure. In NLP, an n-gram refers to sequences of words (or characters) containing n elements as found in a given sentence. N-grams have been used extensively in various NLP applications, including sentiment analysis (Dey et al., 2018; Kouloumpis et al., 2011), text summarisation evaluation (Denkowski and Lavie, 2014; Lin, 2004; Papineni et al., 2002; Vedantam et al., 2015) and in probabilistic language models (Bickel et al., 2005; Pauls and Klein, 2011). For example, the word unigrams (i.e., 1-grams) of the phrase "*a person on a boat*" are {*"a"*, *"person"*, *"on"*, *"a"*, *"boat"*}, while the word bigrams (i.e., 2-grams) are {"*a person*", "*person on*",

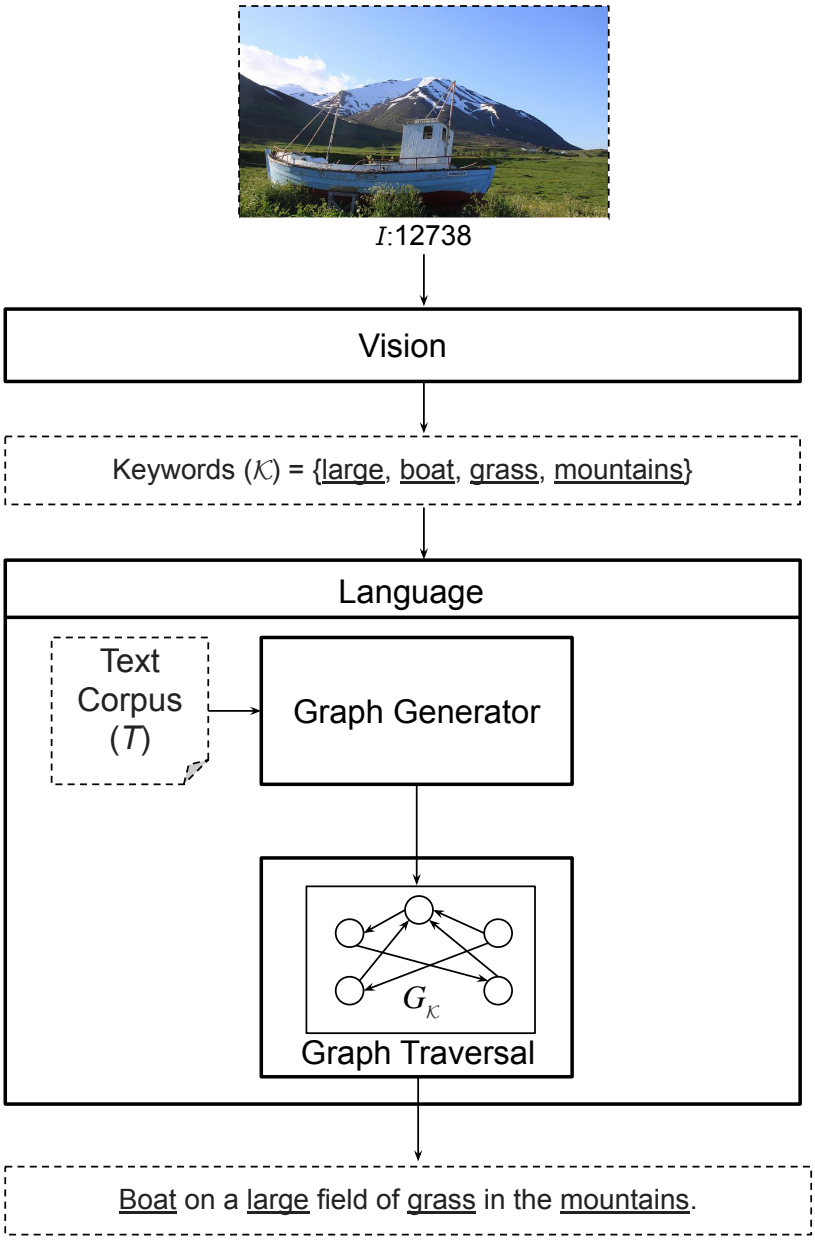


Figure 3.1: High-level architecture of KENGIC.

"*on a*", and "*a boat*"}. The extraction of n-grams $\mathcal{N}_n$ from text T is carried out as indicated in Algorithm 1:

Algorithm 1 N-Grams Extraction

Require: T ▷ Text
Require: $n > 0$ ▷ n-gram size
1: $\mathcal{N}_n \leftarrow \emptyset$ ▷ Initialise n-grams set
2: **for all** $i \in \{0,\ |T|\ - n + 1\}$ **do**
3: $\quad \mathcal{N}_n \leftarrow \mathcal{N}_n \cup T^{i:i+n}$
4: **end for**
5: **return** $\mathcal{N}_n$

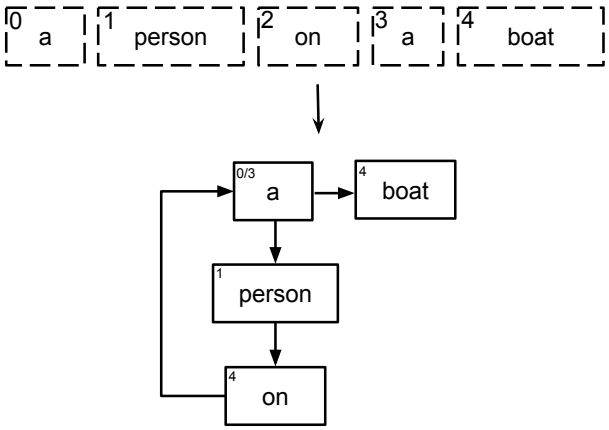
The N-gram graph was initially proposed in Giannakopoulos et al. (2008) as an automatic summarisation evaluation method. This was intended to associate pairs of n-grams with edges to denote how closely each pair is related. This data-structure was then applied in sentiment analysis (Aisopos et al., 2011), language identification (Tromp and Pechenizkiy, 2011) and even in molecular representation (Liu et al., 2019b), while to date no attempts were made in image caption generation. Formally, the n-gram graph is a graph $G_n = \{V, E, L\}$, where V is the set of vertices consisting of phrases extracted from n-grams, E is the set of directed edges which connect phrases represented by vertices (v_1, v_2), and L is a function that assigns a label to each vertex v_i after combining and filtering out overlapping n-grams. The vertices are connected based on whether the last token of each vertex (v^{-1}) overlaps with the first token (v^0) of the remaining vertices in V. For instance, $G_{n=1}$ for the phrase "*a person on a boat*" is illustrated in Fig 3.2(a) and defined as follows:

$$V = \{\text{“}a\text{”}, \text{“}person\text{”}, \text{“}on\text{”}, \text{“}boat\text{”}\},$$
$$E = \{\{\text{“}a\text{”}, \text{“}person\text{”}\}, \{\text{“}person\text{”}, \text{“on”}\}, \{\text{“}on\text{”}, \text{“}a\text{”}\}, \{\text{“}a\text{”}, \text{“}boat\text{”}\}\}$$

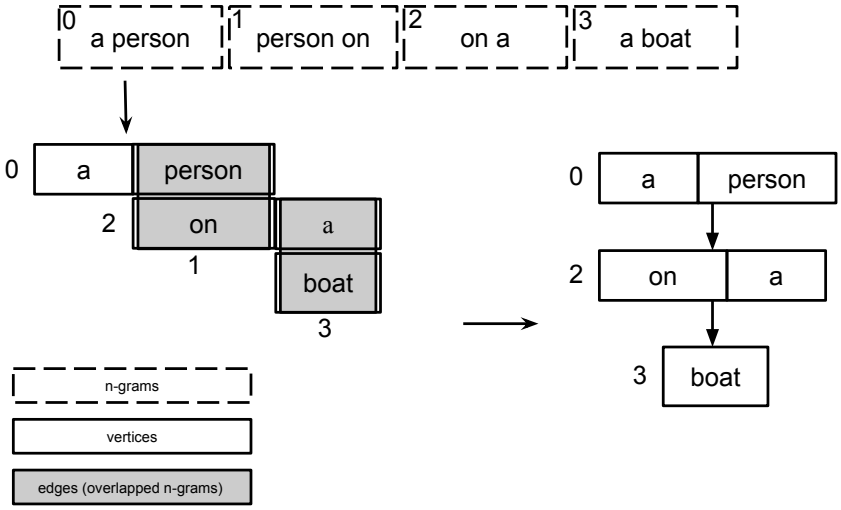
Similarly, G_2 is illustrated in Fig 3.2(b) and is defined as follows:

$$V = \{\text{“}a\ person\text{”}, \text{“}on\ a\text{”}, \text{“}boat\text{”}\},$$
$$E = \{\{\text{“}a\ person\text{”}, \text{“}on\ a\text{”}\}, \{\text{“}on\ a\text{”}, \text{“}boat\text{”}\}$$

More formally, given that $\mathcal{N}_n$ is the set of n-grams extracted from text T, vertices V is equal to $L(\mathcal{N}_n)$, where L combines and filters the n-grams as outlined in Algorithm 2.



(a) 1-gram graph.



(b) 2-gram graph.

Figure 3.2: N-gram graph construction for the phrase "*a person on a boat*" for $n = \{1, 2\}$.

3.2.2 Graph Generator

The language module based on an N-Gram Graph (N-GG) is designed to construct graphs in a bottom-up and top-down approach given a set of query image keywords. For each keyword that is not considered as a stop word[1], the top k frequent n-grams that end with the word w are considered as parents for keyword w, in such a way that n-gram$^{0:n-1}$ is

[1] A stop word is a commonly used word in a language which is not considered important for text-retrieval based applications. Examples of stop words include determiners which are used to mark nouns (e.g., "a", "the"), coordinating conjunctions which connect words, phrases and clauses (e.g., "for", "but"), and prepositions which are used for temporal (e.g., "before") or spatial relations (e.g., "in").

Algorithm 2 $L : \mathcal{N}_n \rightarrow \mathbb{L}$

Require: $\mathcal{N}_n$ ▷ Set of n-grams
$l \leftarrow |\mathcal{N}_n|$
$V \leftarrow \emptyset$
for all $i \in \{0 \dots l\}$ **do**
  if $i \,\%\, 2 = 0$ **then**
    $V \leftarrow \mathcal{N}_n^i$ ▷ i^{th} n-gram
  else if $i = l$ **then**
    $V \leftarrow \mathcal{N}_n^{(1:n)}$ ▷ n-gram without first token
  end if
end for
return V

connected to keyword w. This is repeated for h hops, where each parent which does not start with a start token $\langle t \rangle$ (i.e., $p \in \mathcal{P} \mid p^0 \neq \langle t \rangle$) is connected to its p^{h+1} ancestors in a bottom-up approach.

For instance, the corresponding five topmost 4-gram parents at $h = 0$ for the keyword "*boat*" (i.e., $\mathcal{P}^{h=0}_{w="boat"}$) can be found in Table 3.1. These five 4-gram sequences are then connected with the most probable n-gram parents which have their n^{th} word identical to the first word in the n-grams found in set $\mathcal{P}^h_w$. This is repeated for each keyword in set $\mathcal{K}$ up to a specified number of hops (h). As an example, the next hop ($h = 1$) for the first 4-gram sequence found in $\mathcal{P}^{h=0}_{w="boat"}$ that does not start with a start token $\langle t \rangle$ (i.e., "*sitting on a boat*") is generated through the extraction of the most probable 4-grams that end with the word "*sitting*" as shown in Table 3.2 and in the illustration of Fig. 3.3.

Table 3.1: The top five 4-gram parents ($\mathcal{P}^{h=0}_{k="boat"}$) for the keyword "*boat*" at hop=0.

n-grams			
token index			
0	1	2	3
<t>	<t>	a	**<boat>**
<t>	a	small	**<boat>**
<t>	a	large	**<boat>**
sitting	on	a	**<boat>**
next	to	a	**<boat>**

Once the top-level parents are reached at hop=h, the graph nodes are connected in a top-down approach. All unconnected nodes that form relevant phrases are connected by a directed and unweighted edge. This is handled by taking into consideration the fre-

Table 3.2: The top five grandparents ($h = 1$) for the first non-root parent of $\mathcal{P}^{h=0}_{k="boat"}$.

$h=1$				$h=0$		
n-grams						
token index						
0	1	2	{3,0}	1	2	3
<t>	a	man	**<sitting>**	on	a	**<boat>**
<t>	a	woman	**<sitting>**	on	a	**<boat>**
<t>	a	cat	**<sitting>**	on	a	**<boat>**
group	of	people	**<sitting>**	on	a	**<boat>**
<t>	a	person	**<sitting>**	on	a	**<boat>**

quency count of such connections and if they are found more than e_f times in the text corpus, the corresponding vertices are connected. For instance, both vertices with labels: "*boat*" and "*group of people*" can be linked to the vertex having the phrase "*next to a*", given that both combined phrases "*boat next to a*" and "*group of people next to a*" occur at least e_f times in the text corpus. This constraint is added to reduce rarely occurring connections in the graph generation, as well as to limit the graph complexity. Formally, $G_{\mathcal{K}}$ is generated as outlined in Algorithm 3.

3.2.3 Graph Traversal

Image captions are retrieved from the generated graph ($G_{\mathcal{K}}$) by traversing the graph to search for the most relevant caption that best mentions the set of given keywords. The process is formally outlined in Algorithm 4. The search is carried out in breadth-first approach by keeping a list of paths $\mathcal{Q}$ which can be considered as relevant captions. The search starts by initialising the list of paths with the set of keywords $\mathcal{K}$ (i.e., $\mathcal{Q} = \mathcal{K}$). Each child c of the last vertex of $q \in \mathcal{Q}$ is appended with q to form path $q + c$. Path q is removed from set $\mathcal{Q}$ while all appended paths are added to the set for future concatenation. To reduce the time complexity of the graph traversal, the search process considers a total of q_n paths, while $\mathcal{Q}$ is always kept with the top o_p optimal paths based on one of the following cost functions ($\text{fn}_i \mid 1 \leq i \leq 4$):

1. **F**=$\sum_{i=1} \log P(n_2\text{-gram}_i)$: This is used to compute the fluency (F) by calculating the total $\log$ probability based on each i^{th} n_2-gram according to the text corpus T in order to favour frequently used phrases.

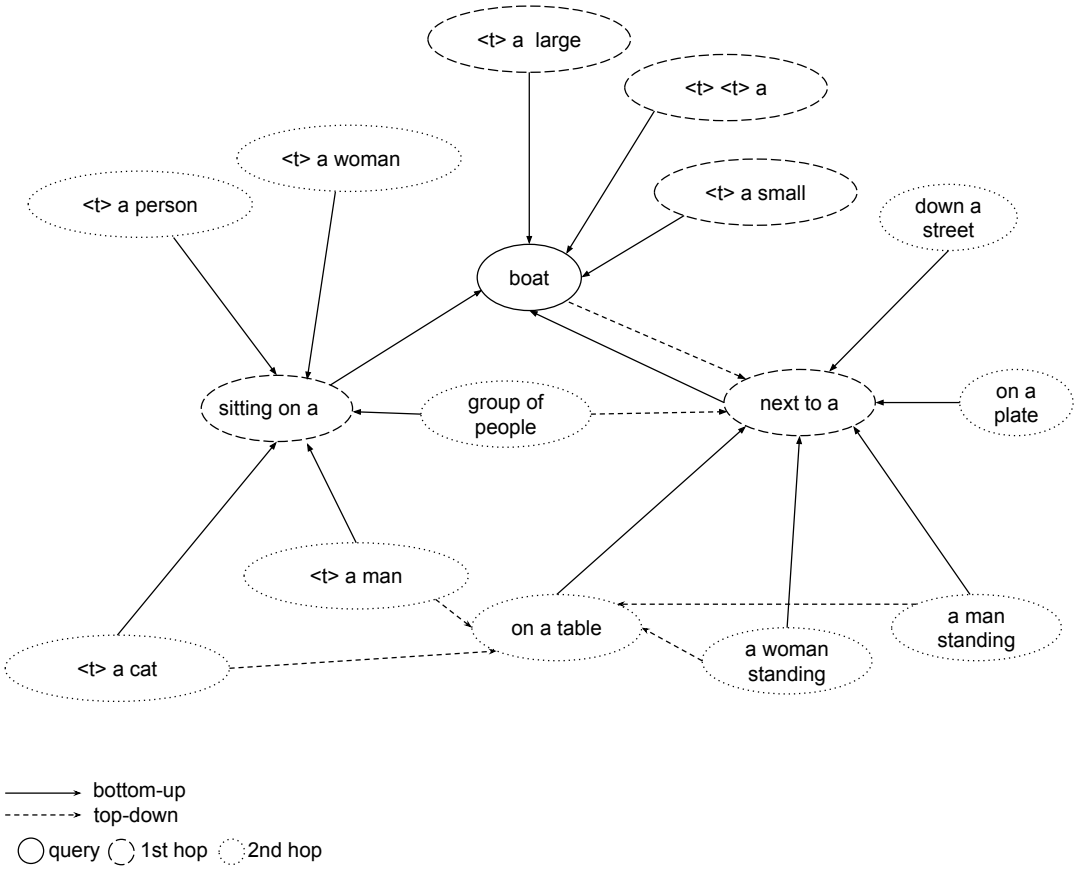


Figure 3.3: Constructed 4-gram graph for the query keyword "boat" with $p = 5$ and $h = 2$. The graph illustrates two examples which were connected in a top-down approach marked with dotted edges.

2. **F + M** = $\sum_{i=1} \frac{\log P(n_2\text{-gram}_i)}{m}$: This takes into consideration both the fluency and keywords matching (M) by considering the number of mentioned keywords m in a given caption. This is to penalise captions that do not mention the query keywords.
3. **F + M + L** = $\sum_{i=1} \frac{\log P(n_2\text{-gram}_i)}{m \times l}$: This is used to favour fluent and long captions with the maximum number of mentioned keywords. Long captions were favoured since fluent captions tend to be short captions.
4. **F + M + L +** N = $\sum_{i=1} \frac{\log P(n_2\text{-gram}_i)) \times N}{m \times l}$: This favours captions with the highest fluency, number of matched keywords m, caption length l and captions which have the lowest number of extra nouns N mentioned which are not found in the given keywords set.

Algorithm 3 Graph Generation

Require: $\mathcal{T}$ ▷ Text Corpus
Require: $\mathcal{K}$ ▷ Set of image keywords
Require: $\mathcal{S}$ ▷ Set of stop words
Require: $n > 0$ ▷ n-gram size
Require: $p > 0$ ▷ Number of parents
Require: $h > 0$ ▷ Number of hops
$G_{\mathcal{K}} \leftarrow \emptyset$ ▷ Initialise graph G based on keywords $\mathcal{K}$
$\mathcal{Q} \leftarrow \mathcal{K}$ ▷ Initialise queue $\mathcal{Q}$ with $\mathcal{K}$
$h' \leftarrow 0$ ▷ Initialise current hop h'
while $h' < h$ **and** $|\mathcal{Q}| > 0$ **do**
  $gn \leftarrow \mathcal{Q}^0$ ▷ First graph node in $\mathcal{Q}$
  $\mathcal{Q} \leftarrow \mathcal{Q}^{1:|\mathcal{Q}|}$ ▷ Removing first element from $\mathcal{Q}$
  if $gn \notin \mathcal{S}$ **then** ▷ If gn is not a stop word
    $\mathcal{P} \leftarrow getP(\mathcal{T}, gn^0, n, p)$ ▷ Gets p n-grams for the 1st token of gn from $\mathcal{T}$
    $G^{h'}_{k=gn} \leftarrow \mathcal{P}$ ▷ Sets $\mathcal{P}$ as parents of gn in G at hop=h'
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathcal{P}$ ▷ Add parents $\mathcal{P}$ to $\mathcal{Q}$
  end if
  $h' \leftarrow h' + 1$
end while
return $G_{\mathcal{K}}$

If we consider for example Fig. 3.4(a) as query image and the keywords *dog*, *skateboard* and *leash* relevant to the image, the caption is generated in the trace listed in Algorithm 5 with respect to the constructed 3-gram graph that is illustrated in Fig. 3.4(b).

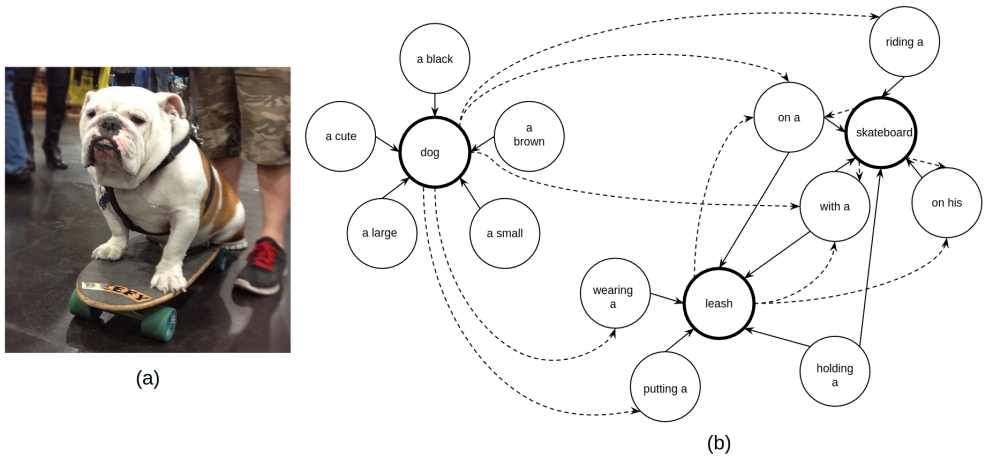


Figure 3.4: A query image with its corresponding 3-gram graph based on the keywords *dog*, *skateboard* and *leash*.

Algorithm 4 Graph Traversal

Require: $G_{\mathcal{K}}$ ▷ Graph G based on keywords $\mathcal{K}$
Require: $q_n > 0$ ▷ Maximum paths to traverse
Require: $o_p > 0$ ▷ Top o_p ranked paths to consider
Require: fn_i ▷ Cost function
$\mathcal{S} \leftarrow \emptyset$ ▷ Set of captions
$\mathcal{Q} \leftarrow \mathcal{K}$ ▷ Set of paths considered as captions
$q_i \leftarrow 0$ ▷ Current total paths considered
while $|\mathcal{Q}| > 0$ **and** $q_i < q_n$ **do**
  $\mathcal{Q} \leftarrow rank(\mathcal{Q}, \text{fn}_i, o_p)$ ▷ Top n ranked paths found in $\mathcal{Q}$ based on fn
  $q \leftarrow \mathcal{Q}^{(0)}$ ▷ First path found in set $\mathcal{Q}$
  $t \leftarrow q^{(|q|)}$ ▷ Last token found in path q
  $\mathcal{C} \leftarrow G^t_{\mathcal{K}}$ ▷ Children nodes of token t
  if $|\mathcal{C} \cap q| \neq |\mathcal{C}|$ **then** ▷ If not all children $\mathcal{C}$ are in q
    for $c \in \mathcal{C}$ **do**
      if $c \notin q$ **then**
        if $|\mathcal{K} \cap q| = |\mathcal{K}|$ **then** ▷ If all keywords $\mathcal{K}$ are in path q
          $\mathcal{S} \leftarrow \mathcal{S} \cup \{q\}$ ▷ Add path q to set of captions S
        else
          $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q + c\}$ ▷ Concatenating child c with current path q
        end if
      end if
    end for
  end if
  $\mathcal{Q} \leftarrow \mathcal{Q} \setminus \{q\}$ ▷ Remove current path q from set $\mathcal{Q}$
  $q_i \leftarrow q_i + 1$
end while
$\mathcal{S} \leftarrow rank(\mathcal{S}, \text{fn}_i, o_p)$ ▷ Rank o_p captions based on cost function fn_i
return $\mathcal{S}$

3.3 Dataset

In this study, the large-scale and most widely used COCO dataset (Lin et al., 2014), released in 2014, was used. This contains a total of $164,062$ images each captioned with five or six human authored captions. This dataset was officially split into training ($82,783$), validation ($40,504$) and testing ($40,775$) sets. However, since the captions of the test images are not publicly available and for consistency with previous studies, the common third-party split of Karpathy and Fei-Fei (2017) was adopted, where the original validation set was further split into validation ($5,000$) and testing ($5,000$) set. The remaining images were added to the original training set which led to a total of $113,287$ images. The human captions found in the training set were used as text corpus for the extraction of

Algorithm 5 Example of a graph traversal based on the keywords *dog*, *skateboard*, and *leash* with respect to the query image and its corresponding 3-gram graph as shown in Fig. 3.4.

$\mathcal{S} \leftarrow \emptyset$
$\mathcal{Q} \leftarrow$ {dog, skateboard, leash}
$q_0 / t_0 \leftarrow$ dog ▷ Connecting *dog*
$\mathcal{C} \leftarrow$ {riding a, on a, with a, wearing a, putting a} ▷ Children nodes of *dog*
$\mathcal{Q} \leftarrow$ {[dog riding a, dog on a, dog with a, dog wearing a, putting a], skateboard, leash}
$q_1 / t_1 \leftarrow$ skateboard
$\mathcal{Q} \leftarrow$ {[dog riding a, dog on a, dog with a, dog wearing a, dog putting a],
[skateboard on a, skateboard with a, skateboard on his], leash}
$q_2 / t_2 \leftarrow$ leash
$\mathcal{Q} \leftarrow$ {[dog riding a, dog on a, dog with a, dog wearing a, dog putting a],
[skateboard on a, skateboard with a, skateboard on his],
[leash on a, leash with a, leash on his]}
$\mathcal{Q} \leftarrow$ {dog riding a, dog on a, dog wearing a, skateboard with a} ▷ Top 4 phrases
$q_3 \leftarrow$ dog riding a ▷ First phrase in queue $\mathcal{Q}$
$t_3 \leftarrow$ riding a ▷ Last node of first phrase
$\mathcal{Q} \leftarrow$ {[dog riding a skatebord], dog on a, dog wearing a, skateboard with a}

▷ The process continues by concatenating each phrase with its children.

$q_4 \leftarrow$ dog on a
$t_4 \leftarrow$ on a
$\mathcal{Q} \leftarrow$ {[dog riding a skatebord], [dog on a leash, dog on a skateboard],
dog wearing a, skateboard with a}
$q_5 \leftarrow$ dog wearing a
$t_5 \leftarrow$ wearing a
$\mathcal{Q} \leftarrow$ {[dog riding a skatebord], [dog on a leash, dog on a skateboard],
[dog wearing a leash], skateboard with a}
$q_6 \leftarrow$ skateboard with a
$t_6 \leftarrow$ with a
$\mathcal{Q} \leftarrow$ {[dog riding a skatebord], [dog on a leash, dog on a skateboard],
[dog wearing a leash], [skateboard with a leash]}
$q_n \leftarrow$ dog wearing a leash
$t_n \leftarrow$ leash
$\mathcal{Q} \leftarrow$ {[...], [...], [dog wearing a leash on a, ...], [...]}
$q_{n+1} \leftarrow$ dog wearing a leash on a
$t_{n+1} \leftarrow$ on a
$\mathcal{Q} \leftarrow$ {[...], [...], [dog wearing a leash on a skateboard, ...], [...]}
$\mathcal{S} \leftarrow$ {dog wearing a leash on a skateboard, ...} ▷ Top ranked captions

n-grams needed to generate each n-gram graph per query image. On the other hand, the testing images and their corresponding ground-truth captions were used for the analyses of this study.

3.4 Metrics

To quantitatively measure the quality of the generated captions, the standard and most popular evaluation metrics were used. The BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) metrics had been adopted from machine translation and document summararisation, while CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) metrics were later proposed specifically for image captioning. All metrics, except SPICE, measure the n-gram overlap between the generated and ground-truth captions. While BLEU measures the n-gram precision and ROUGE considers the n-gram recall, METEOR takes into account the precision, recall and synonyms. CIDEr makes use of TF-IDF to weight n-grams and calculates cosine similarity between captions. On the other hand, to measure the semantic relatedness which n-gram based metrics do not consider, SPICE constructs scene graphs of reference and candidate captions and compares them based on an F-score computed over triplets composed of objects, attributes and relationships.

3.5 Experiments

This section describes the experiments which were carried out during this study. Two types of keyword sets were used for the generation of captions for images found in the testing set. The first type of keywords is based on human extracted words from gold captions which were used for preliminary studies. This set includes (a) human keywords (HK) extracted from the first caption (HC-0) as found in the set of ground-truth captions and (b) the most frequently used human keywords (HK-f) as found in all human captions per image. On the other hand, the second type of keywords are based on machine detected words, which include (c) automatically detected image objects (Objs) by an off-the-shelf object detector, and (d) image keywords detected by a multi-label model (ML-keys).

3.5.1 Human Keywords (HK)

The motivation behind these experiments was to analyse the quality of the captions that are generated based on keywords used by humans. To project a sufficient upper

bound, while simultaneously assessing the quality of the five human-authored captions, this study evaluated the quality of the five human authored captions (HC-i) on the most widely used evaluation metrics. Each human authored caption was compared against the remaining four captions and a test for any statistical significance between the evaluated captions was conducted. Given that no significant difference was noted in the evaluated ground-truth captions, the first set of experiments was based on human-authored keywords extracted from the first set of ground-truth captions (i.e., HC-0). When considering both CIDEr and SPICE scores no significant difference was noted when using one-way Analysis of Variance (ANOVA) (Girden, 1992) (CIDEr: $F(4, 4995) = 1.09, p > 0.05$); SPICE: $F(4, 4995) = 0.54, p > 0.05$). Keyword sets were extracted by using the POS tagger based on the Penn Treebank tagset (Marcus et al., 1993) of the Natural Language Toolkit (NLTK)[2] library after tokenizing the captions. The generated captions were evaluated based on the remaining ground-truth captions (i.e., HC-$\{1-4\}$). The human keyword sets which were composed of nouns, attributes, prepositions and verbs were used in a composite and non-composite way. The composite sets included phrases composed of grouped keywords such as an attribute ("*large*") followed by a noun ("*boat*") or a noun followed by a verb such as ("*boat navigating*"). Composite keywords were used to restrict the model by constraining it to mention such keywords in that specified order without leaving any room for discontinuity between keywords during the graph generation and path traversal. Furthermore, this experiment also sheds light on whether the generation of composite keywords would improve the quality of the generated captions. These experiments were split into the following six categories to reflect the combinations of nouns, attributes, prepositions, and verbs:

1. **HK-n**: Human keywords consisting of nouns only (e.g., "*boat*", "*person*").
2. **HK-na**: Human keywords consisting of nouns and attributes (e.g., "*large*", "*boat*").
3. **HK-nap**: Human keywords consisting of nouns, attributes and prepositions (e.g., "*large*", "*boat*", " *near*").
4. **HK-napv**: Human keywords consisting of nouns, attributes, prepositions and verbs (e.g., "*large*", "*boat*", " *near*", "*navigating*").
5. **HK-(na)**: Composite human keywords composed of an attribute followed by a noun (e.g., "*large boat*").
6. **HK-(nv)**: Composite human keywords composed of a noun followed by a verb (e.g., "*boat navigating*").

[2] `https://www.nltk.org/`

3.5.2 Frequency-based Human Keywords (HK-fi)

Rather than using human extracted keywords from one single caption, this experiment was set to examine the quality of the generated captions based on the most salient keywords that humans choose when describing images. This was carried out by selecting the most commonly used keywords found in the set of corresponding ground-truth captions per image. This was performed by taking into consideration keyword sets with cumulative frequency count per word. Each set was denoted by HK-fi, where i corresponds to the minimum frequency count of each word. For example, HK-f2 consists of keywords which occur at least twice in the set of ground truth captions. Similarly, HK-$f4$ includes keywords with a frequency count of four and five, while HK-f5 consists of keywords which are common in all five ground-truth captions. For consistency with HK, the generated captions were compared against HC-$\{1-4\}$. In this analysis, no POS tagging filtering was performed, which means that keywords were not restricted to nouns, attributes, prepositions and verbs. Captions simply composed from frequent keywords (HK-fi(kw)) were also evaluated to assess their effect on the evaluation metrics.

3.5.3 Detected Objects (Objs)

The third experiment was conducted to investigate the role of detected objects as keywords within the proposed image caption generator. For this purpose, objects were detected using a pre-trained Mask R-CNN (He et al., 2017b) which was trained to detect up to 80 COCO object classes. The average number of detected objects per image was 3.4. Since the human-authored captions were compared against four captions, in this experiment, the generated captions were evaluated on both four and five reference captions. Given that 15 out of the 80 detectable objects consisted of two words (e.g., "*dining table*"), an additional experiment was performed to analyse whether splitting (+sp) these keywords improves the quality of the generated captions. Multiples (+multi) were also taken into consideration when multiple objects of the same class were detected.

3.5.4 Multi-label Keywords (ML-keys)

The fourth experiment was set to generate captions based on the output of a multi-label model trained to predict multi labels that are relevant to query images. Having a model trained to predict labels that can include nouns, attributes and verbs can provide richer keywords for images. For this reason, the ML-Decoder (Ridnik et al., 2021c) model which follows a Transformer-based encoder-decoder pipeline was used for this experiment. This model predicts multi labels by first encoding images using a TResNet (Ridnik et al., 2021b)

backbone which offers higher performance when compared to the plain ResNet (He et al., 2016) architecture. Image embeddings are then transformed into prediction logits for each label through a Transformer architecture with an attention-based classification head. A pre-trained ML-Decoder trained to detect the 80 COCO object labels was used for the generation of multi label objects to be compared with the experiment of Section 3.5.3 (i.e., Objs). The model was trained on images with a resolution of 224×224 and with RGB channels scaled between 0 and 1 and encoded with the ResNet-M (i.e., the medium sized architecture in terms of depth and number of channels) backbone with pre-trained weights based on Open Images dataset (Kuznetsova et al., 2020) as was performed in Ridnik et al. (2021a).

Furthemore, the ML-Decoder model was fine-tuned using the same hyper-parameters used in Ridnik et al. (2021c) on vocabularies extracted from COCO image captions. The fine-tuned model was trained for 40 epochs using Adam optimizer and 1 cycle policy with maximal learning rate of $2\text{e-}4$. A cutout factor of 0.5 and a true-weight decay of $1\text{e-}4$ and auto-augmentation were used. The model was trained with full decoding and the number of token embeddings was set to 768 by adjusting the backbone embedding output via a 1×1 depth-wise convolution. The asymmetric loss function (ASL) for multi-label classification proposed in Ridnik et al. (2021a) was used to dynamically downweight and hard-threshold easy negative samples by using $\gamma_- = 4$, $\gamma_+ = 0$ and $m = 0.05$. The fine-tuning was performed for the following sets of vocabularies $\mathcal{V}$:

1. **Cleaned-w (C-w)**: The top w frequent words which are not considered as stop words and which do not contain any numbers in the human captions.
2. **Cleaned + Lemmatised-w (CL-w)**: The top w frequent cleaned and lemmatised (L) words found in the human captions. This is used to reduce the complexity of the used vocabulary set by using base words which are commonly referred to as lemmas.
3. **Cleaned + Lemmatised + POS Filtering (CLP-w)**: The top w cleaned, lemmatised and Part-of-Speech (POS) filtered words found in the human captions. This is used to further reduce the vocabulary set by words consisting of nouns, attributes, and verbs.

These three sets of vocabularies $v \in \mathcal{V}$ for $w \in \{1000, 2000, 3000\}$ were used as sets for the extraction of keywords $\mathcal{K}$ for each query image. To predict the salient image keywords, the ML-Decoder was explicitly trained to predict keywords according to their frequency as found in the corresponding ground-truth image captions. In the same way HK-fi based keywords were extracted (refer to Section 3.5.2), the ML-Decoder was trained to predict labels based on their cumulative frequency ($f_{ML} \mid 1 \leq f_{ML} \geq 5$). For

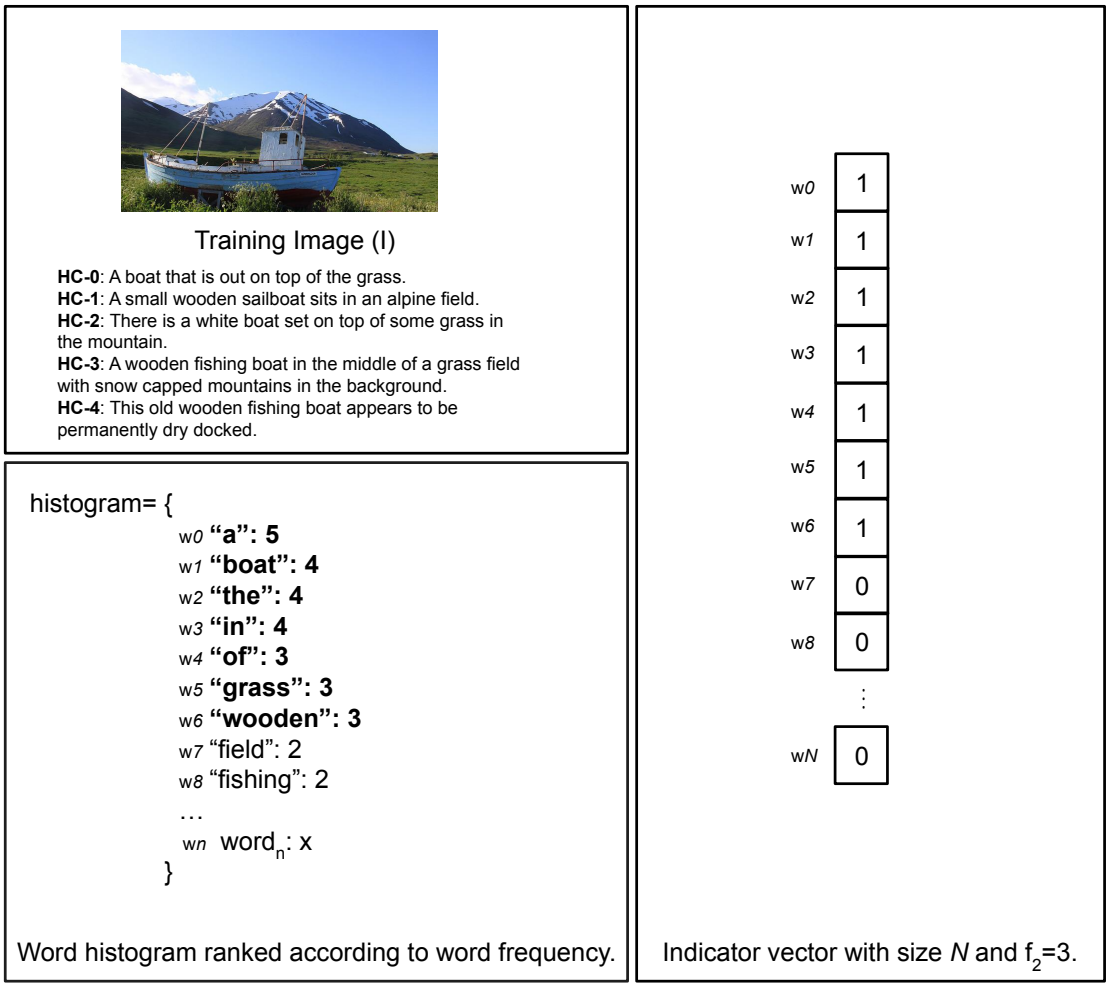


Figure 3.5: Multi-label training with vocabulary size $n = N$ and $f_2 = 3$

instance if $f_{ML} = n$, the ML-Decoder was trained to predict words $\mathcal{W}$ which occur at least f_{ML} times across all corresponding captions. The training setup is illustrated in Fig. 3.5.

3.5.4.1 Metrics

The quality of the multi-label output was evaluated by both example- and label-based metrics. The training of the multi-label output was validated on the conservative label-based mean average precision@i (mAP@i) metric which takes into consideration the ranking order of the multi-label output. This means that true positives are penalised according to their predicted ranking order. This metric is computed over each class as follows:

Let Y be the vector consisting of n ground-truth values of a given label $j \in J$,

$\hat{Y}$ the corresponding predicted labels with values ranging between 0 and 1 and, $\hat{Y}_s, Y_s = \text{sort}(\hat{Y}, Y)$ be the sorted pair of vectors based on $\hat{Y}$ in descending order;

then,

$$\begin{aligned} \text{AP@}i_j &= \frac{\left(\sum_{i=0}^{n} Y_s^i \times \text{rank}\left([\hat{Y}_s \times Y_s] > 0, \hat{Y}_i\right)\right) / (i+1)}{[|\hat{Y}_s \times Y_s] > 0| + \epsilon} \\ \text{mAP@}i &= \frac{1}{|J|} \sum_{j=0}^{|J|} \text{AP@}i_j, \end{aligned} \tag{3.1}$$

where j is the jth label in set J, rank(l, v) is a function which gives the rank order of a value v found in list l starting from 1, otherwise 0 if not found, and ϵ is a constant set to $1\text{e-}8$ to eliminate dividing by 0.

For example, if $Y = [0, 1, 0, 1]$ and $\hat{Y} = [0.75, 0.3, 0.2, 0.8]$, then $\hat{Y}_s, Y_s = [0.8, 0.75, 0.3, 0.2], [1, 0, 1, 0]$.

Therefore,

$$\begin{aligned} [\hat{Y}_s \times Y_s] &= [0.8(1), 0.75(0), 0.3(1), 0.2(0)] \\ &= [0.8, 0, 0.3, 0] \end{aligned}$$

$$[\hat{Y}_s \times Y_s] > 0 = [0.8, 0.3];$$

$$\begin{aligned} \text{AP@}i &= \frac{\frac{1(1)}{1} + \frac{0(0)}{2} + \frac{1(2)}{3} + \frac{0(0)}{4}}{2 + 1\text{e-}8} \\ &= \frac{1 + \frac{2}{3}}{2 + 1\text{e-}8} \\ &= 0.83 \end{aligned}$$

On the other hand, to test the multi-label output at testing time, the example-based metrics which include the accuracy (Acc), precision (P), recall (R) and F-score (F) (Zhang and Zhou, 2014) were used. These metrics were computed between the ground-truth Y

and predicted labels $\hat{Y} = \{h(x_i), \forall x_i \in X\}$, over each test instance as follows:

$$Acc = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap h(x_i)|}{|Y_i \cup h(x_i)|}; \tag{3.2}$$

$$P = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap h(x_i)|}{|h(x_i)|}; \tag{3.3}$$

$$R = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap h(x_i)|}{|Y_i|}; \tag{3.4}$$

$$F = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{2 \times P_i \times R_i}{P_i + R_i}\Big), \tag{3.5}$$

where n is the number of test instances, Y_i is the ground-truth keyword set for the i^{th} instance, x_i is the i^{th} image feature, and $h(x_i)$ is the predicted keyword set for image i.

Furthermore, precision and recall per label were computed to analyse the testing performance per label. When considering the following quantifications for each label y_j:

$$\begin{aligned}
TP_j &= |\{x_i \mid y_j \in Y_i \wedge y_j \in h(x_i), 1 \leq i \leq n\}|;\\
FP_j &= |\{x_i \mid y_j \notin Y_i \wedge y_j \in h(x_i), 1 \leq i \leq n\}|;\\
TN_j &= |\{x_i \mid y_j \notin Y_i \wedge y_j \notin h(x_i), 1 \leq i \leq n\}|;\\
FN_j &= |\{x_i \mid y_j \in Y_i \wedge y_j \notin h(x_i), 1 \leq i \leq n\}|,
\end{aligned}$$

then, the macro average precision and recall per label were computed by:

$$\text{Precision per label} = \sum_{j \in J}\frac{TP_j}{TP_j + FP_j} \tag{3.6}$$

$$\text{Recall per label} = \sum_{j \in J}\frac{TP_j}{TP_j + FN_j} \tag{3.7}$$

3.5.4.2 Validation Results

The ML-Decoder was validated on the validation set using the mAP@i score (Eqn. 3.1). The validation results for the different vocabulary sets ($\mathcal{V}$), their number of words (w) and the word frequency f_{ML} used while training the ML-Decoder are reported in Table 3.3. It was evident that the mAP@i decreases when the number of labels in the vocabulary is increased. This provides an indication of the complexity and ambiguity behind such multi-label learning. This was confirmed across all vocabularies and f_{ML} thresholds used. It was

Table 3.3: Maximum mAP@i recorded on the validation set while training the ML-Decoder on different vocabularies, f_{ML} and number of words (w).

		Max Val mAP@i		
		Num. of words (w)		
Vocabulary	f_{ML}	1000	2000	3000
Cleaned	1	30.49	23.11	19.13
Cleaned	2	27.40	20.05	15.53
Cleaned	3	22.88	15.11	11.16
Cleaned	4	17.21	10.44	7.58
Cleaned	5	10.72	6.05	4.16
Cleaned + Lemmatised	1	38.21	28.25	23.44
Cleaned + Lemmatised	2	34.47	23.88	17.57
Cleaned + Lemmatised	3	27.18	17.19	12.04
Cleaned + Lemmatised	4	19.48	11.54	7.80
Cleaned + Lemmatised	5	11.82	6.46	4.36
Cleaned + Lemmatised + POS	1	37.96	28.52	23.35
Cleaned + Lemmatised + POS	2	34.25	23.46	17.35
Cleaned + Lemmatised + POS	3	26.97	16.97	11.73
Cleaned + Lemmatised + POS	4	19.57	11.37	7.96
Cleaned + Lemmatised + POS	5	11.68	6.18	4.31

also clear that when increasing the threshold f_{ML} (i.e., restricting the number of output labels), the mAP@i decreases. This confirmed the difficulty of constraining the multi-label output according to the actual set of keywords used in the ground-truth captions.

3.5.4.3 Test Results

The test results are presented in tables 3.4 to 3.6. The results are split according to the size of the vocabulary v used (i.e., $w \in \{1000, 2000, 3000\}$). For validation purposes, the ML-Decoder was trained and tested to predict object labels found in the COCO dataset as was performed in Ridnik et al. (2021c) (Refer to COCO Objs (trained) in the tables). This was carried out in order to provide confidence in the generated results. As tabulated in the results tables, the scores obtained for the ML-Decoder when trained on the COCO objects were highly comparable to those obtained when using the pre-trained model. When considering both the precision and recall via the F-score measure, it was found that the performance of the ML-Decoder peaks at $f_{ML} = 3$ followed closely by $f_{ML} = 2$ as shown in bold and italics respectively. The general F-score trend across all tested scenarios follows a hill-climb trend between $1 \leq F_{ML} \leq 3$ and a downward trend between $3 \leq F_{ML} \leq 5$. Although no significant decrease in performance was noted when increasing the number of words in the vocabulary sets, the F-score recorded a slight decrease between sets consisting of larger vocabularies. For instance, the top F-score when

using the cleaned vocabulary set of size 1000 (i.e., 57.59 at $f_{ML} = 3$) dropped to 57.01 after increasing the same set with 2000 words. Performance reduction was less noted between sets consisting of 1000 and 3000 words. The results also confirm that the ML-Decoder performed slightly better on the Cleaned + Lemmatised + POS vocabulary set most probably because it is less ambiguous. However, this was not pronounced since the best F-score of the cleaned set at $w = 1000$ (i.e., 57.59) just reached a maximum of 57.92 when using the former set. Similar minor improvements were achieved on the vocabulary consisting of 2000 and 3000 words, respectively. All vocabulary sets with $f_{ML} = \{2, 3\}$ were considered for KENGIC experiments.

Table 3.4: ML-Decoder test results metrics based on the top 1000 frequent words (w) found in the different vocabulary sets ($\mathcal{V}$) with varying frequency count (f_{ML}) and COCO Objs labels.

Vocabulary	w	f_{ML}	Acc	P	R	F	P/label	R/label
COCO Objs (pretrained) (Ridnik et al., 2021c)	80	N/A	75.98	86.82	83.68	83.41	51.77	46.51
COCO Objs (trained)	80	N/A	75.61	85.18	84.93	83.21	42.25	47.31
Cleaned	1000	1	35.49	59.26	46.74	51.10	51.91	35.92
Cleaned	1000	2	42.33	62.31	56.09	*56.63*	50.35	31.88
Cleaned	1000	3	45.69	62.76	59.55	**57.59**	47.35	31.00
Cleaned	1000	4	48.21	59.08	57.83	53.83	46.33	27.96
Cleaned	1000	5	51.43	38.56	54.39	36.47	44.45	23.41
Cleaned + Lemmatised	1000	1	34.33	58.07	45.39	49.88	48.66	30.74
Cleaned + Lemmatised	1000	2	41.63	60.74	56.02	*56.07*	46.33	28.10
Cleaned + Lemmatised	1000	3	45.10	62.35	59.63	**57.36**	45.93	26.94
Cleaned + Lemmatised	1000	4	47.89	58.73	57.31	53.41	50.72	23.93
Cleaned + Lemmatised	1000	5	51.46	38.65	55.34	36.90	43.11	22.22
Cleaned + Lemmatised + POS	1000	1	35.39	59.31	46.48	50.98	50.10	30.96
Cleaned + Lemmatised + POS	1000	2	42.49	62.99	55.75	*56.86*	50.31	26.81
Cleaned + Lemmatised + POS	1000	3	45.99	64.05	59.49	**57.92**	50.95	25.94
Cleaned + Lemmatised + POS	1000	4	48.56	58.54	58.16	53.65	46.72	25.70
Cleaned + Lemmatised + POS	1000	5	53.23	38.36	55.79	36.04	44.50	21.52

Table 3.5: ML-Decoder test results metrics based on the top 2000 frequent words (w) found in the different vocabulary sets ($\mathcal{V}$) with varying frequency count (f_{ML}) and COCO Objs labels.

Vocabulary	w	f_{ML}	Acc	P	R	F	P/label	R/label
COCO Objs (pretrained) (Ridnik et al., 2021c)	80	N/A	75.98	86.82	83.68	83.41	51.77	46.51
COCO Objs (trained)	80	N/A	75.61	85.18	84.93	83.21	42.25	47.31
Cleaned	2000	1	33.61	57.71	44.36	49.15	46.50	25.51
Cleaned	2000	2	41.33	59.72	56.39	*55.84*	43.49	25.77
Cleaned	2000	3	45.03	64.82	57.15	**57.01**	49.69	23.52
Cleaned	2000	4	48.04	59.18	57.23	53.50	48.78	23.09
Cleaned	2000	5	51.54	38.61	54.78	36.37	50.26	21.89
Cleaned + Lemmatised	2000	1	32.68	58.87	42.08	48.04	46.19	19.75
Cleaned + Lemmatised	2000	2	40.91	63.62	52.54	*55.29*	48.45	19.83
Cleaned + Lemmatised	2000	3	44.94	64.85	57.01	**57.03**	50.13	21.17
Cleaned + Lemmatised	2000	4	47.10	58.36	58.34	53.65	45.89	23.85
Cleaned + Lemmatised	2000	5	51.45	36.36	53.86	33.82	45.12	18.33
Cleaned + Lemmatised + POS	2000	1	33.64	58.03	44.27	49.11	43.77	20.35
Cleaned + Lemmatised + POS	2000	2	41.65	62.08	55.07	*56.08*	46.25	20.91
Cleaned + Lemmatised + POS	2000	3	45.34	63.32	58.69	**57.30**	47.44	22.61
Cleaned + Lemmatised + POS	2000	4	48.42	59.65	56.46	53.55	48.63	20.99
Cleaned + Lemmatised + POS	2000	5	51.84	37.07	55.07	35.15	49.42	20.29

Table 3.6: ML-Decoder test results metrics based on the top 3000 frequent words (w) found in the different vocabulary sets ($\mathcal{V}$) with varying frequency count (f_{ML}) and COCO Objs labels.

Vocabulary	w	f_{ML}	Acc	P	R	F	P/label	R/label
COCO Objs (pretrained) (Ridnik et al., 2021c)	80	N/A	75.98	86.82	83.68	83.41	51.77	46.51
COCO Objs (trained)	80	N/A	75.61	85.18	84.93	83.21	42.25	47.31
Cleaned	3000	1	32.81	58.37	42.57	48.22	46.33	18.90
Cleaned	3000	2	41.09	61.71	54.28	*55.58*	46.87	21.03
Cleaned	3000	3	45.07	63.33	58.06	**57.05**	49.26	22.33
Cleaned	3000	4	47.46	58.90	56.51	53.01	47.12	23.11
Cleaned	3000	5	51.27	37.84	55.28	36.02	37.76	21.88
Cleaned + Lemmatised	3000	1	32.34	57.49	42.28	47.71	44.10	15.63
Cleaned + Lemmatised	3000	2	40.46	61.65	53.21	*54.96*	44.37	19.67
Cleaned + Lemmatised	3000	3	44.93	64.88	57.23	**57.14**	46.05	20.48
Cleaned + Lemmatised	3000	4	46.74	57.46	58.27	53.22	42.40	23.12
Cleaned + Lemmatised	3000	5	50.45	36.76	53.66	34.51	40.53	18.85
Cleaned + Lemmatised + POS	3000	1	33.30	57.58	43.79	48.73	42.07	16.34
Cleaned + Lemmatised + POS	3000	2	41.19	62.00	54.24	*55.59*	47.13	18.52
Cleaned + Lemmatised + POS	3000	3	45.09	61.68	59.85	**57.26**	39.82	22.22
Cleaned + Lemmatised + POS	3000	4	47.45	58.03	57.88	53.36	42.81	22.60
Cleaned + Lemmatised + POS	3000	5	52.71	37.49	54.92	34.80	47.30	17.93

3.6 Summary

This chapter presented the methodology adopted to study the role of keywords in image captioning via the the proposed KENGIC framework. This chapter also details the experiments and reports results of a purposely developed multi-label model used for predicting keywords relevant to images. The following chapter presents the results of this study and discusses the main outcomes.

4 Results and Discussion

This chapter reports and discusses the quantitative results of this study. A qualitative analysis was also conducted to get deeper insights into how KENGIC generates captions when using different keyword sets and how these affect the evaluation metrics. To complement with this analyses, this chapter also presents insights from human evaluation.

4.1 Optimisation

Hyper-parameter optimisation (HPO) was carried out to find the optimal set of hyper-parameters. The optimisation was carried out using human defined keywords to provide the ideal parameters for the system. Since no statistical significance was noted when evaluating the quality of the five human ground-truth captions as discussed in Section 4.3, the optimisation was carried out by a grid search on HK-n keywords set extracted from HC-0 by varying the following parameters as follows: $n \in \{3, 4\}$ (n-gram used to construct knowledge graphs), $n_2 \in \{3, 4\}$ (n_2-gram size used to calculate the probability of the generated captions), $h \in \{1, 2\}$ (number of hops), $\text{fn}_i \mid 1 \leq i \leq 4$ (cost function), while the values of k (number of parents), e_f (minimum edge frequency count between n-grams in graphs) and o_p (optimal paths considered during path traversal) were manually set to 5 and q_n (maximum number of paths considered during path traversal) was set to 150.

4.2 Validation

KENGIC was optimised on the HK-n keywords extracted from HC-0 as reported in Section 4.1. This results in a total of 32 configurations as tabulated in Table 4.1. Due to the time complexity of high-ordered graphs, the optimisation was carried out on a random sample chosen from the validation set. Based on a population size of 5000, a sample size of 357 estimates the population results with a 95% confidence level and 5% margin of error. Therefore, a sample size of 500 images was chosen for the validation process. Since

image caption generators are generally optimised on the CIDEr metric (Anderson et al., 2018; Li and Jiang, 2019; Yang et al., 2020; Yao et al., 2017; Yu et al., 2020), the hyperparameters ($n = 3, h = 1, n_2 = 3,$ fn$_i = 4$) which maximised the CIDEr score were selected. This configuration led to a CIDEr and SPICE scores of 72.3 and 16.5, respectively. Apart from CIDEr, this configuration maximised ROUGE-L, whilst the other metrics scored best when using larger graphs ($h = 2$) and $n_2 = \{3, 4\}$.

4.3 Testing

When analysing the level of statistical significance between the evaluated human captions using the one-way ANOVA (Girden, 1992), no significant difference was observed when considering both CIDEr ($F(4, 4995) = 1.09, p > 0.05$) and SPICE ($F(4, 4995) = 0.54, p > 0.05$). The quality of the human captions reached a maximum CIDEr and SPICE scores of 89.3 and 21.2 respectively, which are lower than those obtained by current state-of-the-art image caption generators when evaluated on five captions, with the exception of the SPICE metric which is very comparable. Two conventional benchmark encoder-decoder based models (Mind's Eye (Chen and Lawrence Zitnick, 2015) and NeuralTalk (Karpathy and Fei-Fei, 2017)) have been included in Table 4.2 for comparison purposes alongside three other state-of-the-art models, namely the attention-based Up-Down (Anderson et al., 2018) model, the scene-graph based auto-encoder with knowledge distillation (SGAE-KD (Yang et al., 2020)) and the Multimodal Transformer (MT)-based generator (Yu et al., 2020). As illustrated in Fig. 4.1, a relatively high variation was found when analysing the score distributions of HC-0. This was most particularly observed in CIDEr score as its mean and standard deviation were 89.3 and 66.6 respectively. This confirms the high degree of linguistic variation and the difficulty in assessing captions, especially in cases when humans can take different perspectives in describing images.

Table 4.1: Evaluation metrics computed for different hyper-parameters based on the HK-n keywords extracted for 500 images from the validation set. Metrics are sorted according to CIDEr score in descending order.

Rank	n	h	n_2	fn_i	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr ↓	SPICE
1	3	1	3	4	50.6	34.5	22.5	13.7	18.9	**38.2**	**72.3**	16.5
2	3	1	3	3	52.4	35.7	23.1	14.1	19.3	38.1	71.7	16.8
3	4	2	3	3	54.3	35.9	23.9	15.4	19.7	37.3	66.5	16.9
4	3	1	3	2	43.1	29.1	18.8	11.9	17.8	36.8	65.7	15.4
5	3	2	3	2	51.7	34.3	21.9	13.8	18.5	36.7	65.2	16.2
6	3	1	4	3	40.2	28.4	19.1	12.7	17.1	36	64.8	15.1
7	4	2	3	4	47.4	32.1	21.3	13.5	18.4	36.6	64.2	15.5
8	3	2	3	4	**57.6**	**38.8**	24.6	15	**20.3**	37.3	63.9	**17.7**
9	3	2	4	4	55.5	37.9	**24.8**	**15.7**	19.5	37	63.8	16.9
10	3	1	4	4	37.9	26.9	18.1	12.1	16.9	36	63.5	14.8
11	3	2	4	2	52.7	35.5	22.9	14.3	18.5	36.3	62.5	16.4
12	4	2	4	3	48.1	33.2	22.5	14.9	18.6	36.8	62.2	15.6
13	4	2	4	4	43	29.7	19.9	12.8	17.7	35.7	61.3	14.7
14	3	1	3	1	41.4	27.7	17.6	10.7	17.1	35.3	60.9	14.5
15	3	1	4	2	35	24.7	16.6	11	16.4	35.3	60.5	14.4
16	4	2	3	2	38.3	25.9	17.6	11.3	16.9	34.9	59.2	14.4
17	4	2	4	2	38.6	26.8	18.5	12.3	16.8	34.7	57.9	14.3
18	3	2	4	3	52.1	34.7	22	13.7	19.9	36.2	56.4	17
19	3	1	4	1	31.5	22	14.8	10.1	15.6	34.1	56.2	13.4
20	4	1	3	4	28.9	20	13.7	8.9	15.6	32.6	53.6	13.3
21	4	1	3	3	28.9	20	13.6	8.8	15.5	32.3	53.2	13.7
22	3	2	4	1	49.7	33.1	21	13	17.7	34.2	51.8	14.7
23	4	2	3	1	33.3	21.9	14.4	9.1	15.5	31.9	51	12.7
24	4	1	4	3	25.8	18.2	12.6	8.1	14.9	31.7	50.5	12.9
25	4	2	4	1	32.9	22.3	14.8	9.7	15.3	32	50.1	12.7
26	4	1	3	2	23.1	16.1	11	7.2	14.7	31.4	49.9	12.9
27	4	1	4	4	24.2	17.1	11.7	7.5	14.8	31.5	49.6	12.7
28	4	1	4	2	21.5	15.1	10.7	7.1	14.3	31	48	12.4
29	3	2	3	1	49.8	31.9	19.3	11.3	17.8	33.2	47.8	14.6
30	4	1	4	1	17.2	12.1	8.5	5.7	13.2	29	43.2	11.2
31	3	2	3	3	39.7	25.8	15.8	9.5	20.2	32.9	41.7	16.8
32	4	1	3	1	16	11	7.4	4.8	12.9	28	41.2	10.9

Table 4.2: Baseline metrics in percentage of benchmark state-of-the art image caption generators when evaluated on five ground-truth captions and the metrics computed for each human ground-truth caption (HC) when compared with the other four human captions as found in the test set.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Mind's Eye (Chen and Lawrence Zitnick, 2015)	-	-	-	18.8	19.6	-	-	-
NeuralTalk (Karpathy and Fei-Fei, 2017)	62.5	45.0	32.1	23.0	19.5	-	66.0	-
Up-Down (Anderson et al., 2018)	77.2	-	-	36.2	27.0	54.9	113.5	20.3
SGAE-KD (Yang et al., 2020)	78.2	-	-	37.3	28.1	57.4	117.1	21.3
MT (Yu et al., 2020)	77.3	-	-	37.4	28.7	57.4	119.6	-
HC-0	63.6	44.1	29.7	19.9	24.4	47.3	89.3	21.2
HC-1	62.9	43.4	29	19.2	24.1	46.6	87.9	21
HC-2	63.1	43.7	29.5	19.7	24.1	46.5	87.8	21.1
HC-3	62.4	43	28.9	19.3	24.1	46.6	86.6	20.9
HC-4	62.7	43.1	28.8	19.2	24	46.4	87.4	21

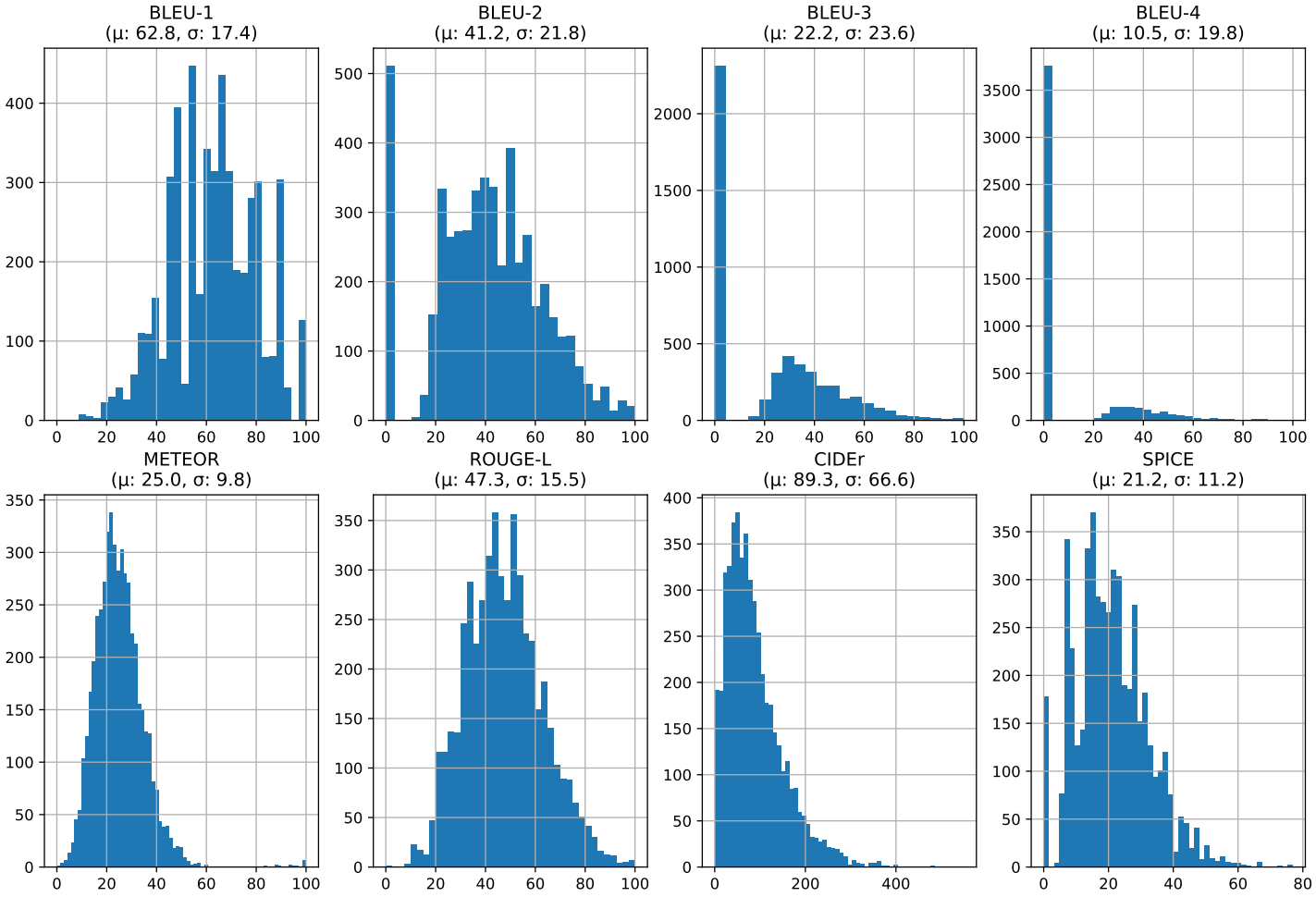


Figure 4.1: Frequency distributions of the results metrics of captions HC-0 when evaluated against the other four human ground-truth captions.

Table 4.3: POS analysis of HK-f keywords sets.

		Averages (Average % w.r.t to # of keywords)			
Keywords	Num. of keywords	Nouns	Attributes	Prepositions	Verbs
HK-f1	18.09	8.83 (48.84)	3.84 (21.26)	0.56 (3.07)	4.03 (22.32)
HK-f2	5.46	3.24 (61.17)	0.87 (15.15)	0.09 (1.58)	1.04 (18.26)
HK-f3	2.77	1.95 (74.88)	0.28 (8.3)	0.02 (0.62)	0.4 (12.51)
HK-f4	1.78	1.41 (83.91)	0.1 (4.15)	0.01 (0.31)	0.18 (8.12)
HK-f5	1.29	1.15 (91.33)	0.03 (1.63)	0 (0.12)	0.07 (4.42)

4.3.1 Human Keywords

The evaluation of the captions generated by KENGIC based on human extracted keywords is presented in Table 4.5. The first section of the table consists of captions which were generated based on keywords extracted from one human authored caption (i.e., HC-0), while the second section presents the evaluation of the captions generated based on the salient words found across all ground-truth captions (i.e., HK-fi). It was clear that the captions generated based on the non-composite keywords set are highly comparable and significantly better than those produced by the composite keywords. Although not very significant, HK-napv obtained the highest scores, except on ROUGE-L, CIDEr and SPICE which peaked when using HK-na. On the other hand, the composite keyword sets (i.e., HK-(na), HK-(nv)) were found to restrict the generation process and lead to the lowest scored captions.

Table 4.5 also shows the metrics of the captions generated when using frequency-based human keywords (HK-fi). When considering the best performing set of salient keywords (i.e., HK-f2), the metrics improved substantially over HK across all metrics, especially CIDEr score which increased from 68.3 to 112.3 when compared to HK-na. Surprisingly, HK-f1 obtained the highest SPICE score of 33.7 despite using an average of 18.09 keywords which makes it harder to construct sentences as noted in the other metrics. The evaluation metrics revealed that captions composed of just frequent keywords (kw) obtained considerably high scores. In fact, HK-f2(kw) even exceeded the quality of HC-0 in terms of CIDEr. These results confirm that metrics give an important weight to the mentioned keywords and pay less attention to the sentence structure and the order of the used words. This observation raises important questions on how captions are currently being evaluated. To get deeper insights about the extracted frequency-based keywords, a POS tag analysis was carried out to analyse the type of keywords which were commonly used for the captioning process. This analyis is tabulated in Table 4.3.

When considering the best salient keyword set (i.e., HK-f2), it was found that the average number of used keywords was 3.24 and these were predominantly nouns (61.17%),

Table 4.4: POS tag distribution of HK-f2 keyword set according to low, medium and high CIDEr metrics. The table tabulates the average (μ) number of keywords per each POS tag as well as the percentage of each POS tag with respect to the total number of keywords in brackets.

	CIDEr Scores		
	Low	Medium	High
	μ (% w.r.t # of keywords) of HK-f2		
Num. of keywords	4.18	5.53	6.62
Nouns ("boat")	2.75 (68.29)	3.33 (61.06)	3.56 (54.32)
Attributes ("wooden")	0.55 (12.56)	0.84 (14.8)	1.23 (18.44)
Prepositions ("on")	0.07 (1.52)	0.09 (1.53)	0.12 (1.73)
Verbs ("standing")	0.66 (14.45)	1.05 (18.86)	1.39 (20.84)

followed by verbs (18.26%), attributes (15.15%), and prepositions (1.58%). This conforms to the analysis of HK given that HK-n provided a good supporting baseline when compared to the other keyword sets. Adding attributes and verbs with nouns in the keyword set incrementally improved the metrics, while no effect was observed when introducing prepositions. A clear indication of why the latter had no effect is due to their infrequent usage (1.58% in HK-f2 keywords). The POS tags were further analysed according to the CIDEr metric as evaluated based on HK-f2 keywords as tabulated in Table 4.4. Captions were organised in three categories (low, medium high) in accordance with the three quartiles. It was confirmed that captions with high CIDEr scores had more keywords (μ=6.62), especially nouns ($\mu = 3.56$), followed by verbs ($\mu = 1.39$) and attributes ($\mu = 1.23$). This analysis confirmed that better captions can be generated through richer and diverse vocabularies. This is evident as 68.29% of the keywords used in low category consisted of nouns, while the high category decreased to 54.32%, as attributes and verbs were increased from 12.56% to 18.44% and 14.45% to 20.84%, respectively. On the other hand, less pronounced increase was noted in prepositions which therefore confirms their insignificant effect owing to their infrequent usage.

Table 4.5: Metrics computed based on human keywords (HK) extracted from HC-0 and compared against HC-$\{1-4\}$.

Keywords	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
HK-n	50	33.8	21.7	13.4	18.4	37.6	66.4	16.3
HK-na	54.3	36.7	23.4	14.4	19.3	**38**	**68.3**	**16.8**
HK-nap	54.6	36.7	23.3	14.3	19.4	**38**	67.6	**16.8**
HK-napv	**55.5**	**38**	**24.7**	**15.5**	**20.3**	37.7	65.2	16.3
HK-(na)	26.1	17.7	11.3	7	12.2	25	42.1	10.3
HK-(nv)	18.2	12.3	7.9	4.9	10.4	20.7	35.3	8.3
HK-f1(kw)	***85.6***	*14.7*	*2.3*	*0*	*24.8*	*23.7*	*53*	*17.6*
HK-f1	59.3	44.1	29.8	19.2	**30.8**	38.9	25.6	**33.7**
HK-f2(kw)	*55.4*	*15.3*	*3.6*	*0.9*	*24.2*	*32.8*	*91*	*16.1*
HK-f2	76.1	**57.9**	**41.4**	**28.6**	28.2	**47.6**	**112.3**	25.8
HK-f3(kw)	*10.8*	*3.7*	*1*	*0.3*	*16.9*	*27.4*	*56*	*11.7*
HK-f3	35.2	27.4	20.3	14.9	20.4	39.8	91.9	18
HK-f4(kw)	*0.6*	*0.2*	*0.1*	*0*	*9.7*	*18.7*	*34.8*	*8*
HK-f4	4	3.2	2.4	1.8	11.3	24.5	50.7	10.8
HK-f5(kw)	*0*	*0*	*0*	*0*	*4*	*9.5*	*17.5*	*4.5*
HK-f5	0	0	0	0	4.4	10.7	20.6	5.1

4.3.2 Predicted Keywords

The quality of the generated captions based on the predicted keyword sets is presented in Table 4.6. These results were compared against four captions to provide fair comparison with results generated based on human extracted keywords as well as with five captions.

When compared to HK-f2, the quality of the generated captions when using the predicted Objs keywords set (i.e., objects detected using Faster R-CNN (Ren et al., 2015)) decreased substantially. For instance, B-4 decreased from 19.9 to 1.8, while CIDEr and SPICE dropped from 89.3 and 21.2 to 22.9 and 6.5, respectively. Apart from the detection quality, one major reason behind the low scores when compared to both HC and HK based captions is the limited vocabulary set size of the used detector (i.e., 80 objects). This is much smaller when compared to the 3034 unique nouns found in the five human authored captions of the testing images. Out of these nouns, only 64 (2.1%) were found in the vocabulary set of the detector's output. Considerable improvements were recorded with the introduction of the two-word splitting (Objs+sp). In fact, improvements of more than 250% were recorded on the BLEU scores. On the other hand, an improvement of 73% was observed on METEOR, 60% on ROUGE-L, while CIDEr and SPICE recorded an increase of 59% and 63%, respectively. This conforms to the previous finding that composite human defined words (i.e., HK-(na), HK-(nv)) constrain KENGIC with limited and less commonly used keywords. Objs+sp+multi which was intended to handle multiple same objects by grouping them into their corresponding plural form ended up being less effective in all metrics, except in SPICE score where no changes were recorded.

The quality of the generated captions based on the ML-objs keywords set (i.e., the output of the ML-Decoder when trained to predict COCO objects) was found to be highly comparable with that obtained when using the Objs keyword set. The highest absolute difference was noted on the CIDEr metric (1.4), while the mean absolute difference was equal to 0.5. This confirmed the level of comparability between the outputs of the Faster R-CNN (Ren et al., 2015) and ML-Decoder (Ridnik et al., 2021c). Overall, it was found that ML-C-2K-f2[1] scored best on all BLEU, METEOR and SPICE scores, while ML-CL-1K-f2 and ML-C-1K-f2 recorded a slight improvement on ROUGE-L and CIDEr, respectively. As was revealed in Section 4.3.1, keywords with a frequency of 2 recorded the highest metrics. This was found to be consistent with the results of the ML-Decoder as the highest metrics were recorded based on ML-C-1K-f2 (CIDEr), ML-CL-1K-f2 (BLEU-4, ROUGE-L) and ML-C-2K-f2 (BLEU-{1-4}, SPICE). No significant difference was noted between the different types of keyword used (i.e., C: Cleaned, L: Lemmatised, P: POS). Although

[1] ML-C-2K-f2: Keyword set generated by the multi-label (ML) model based on 2000 (2K) cleaned (C) words which occur at least twice (f2) in corresponding human captions.

the difference is not significant, the cleaned set achieved the highest scores except for ROUGE-L. Since the ML-Decoder was found to predict the different sets with similar accuracy, the cleaned keyword set turns out to be a better fit for image captioning as it is less restrictive and therefore provides more diversity. Similar observations were noted when evaluating the captions against the full set of ground-truth captions. These results are found in brackets in Table 4.6. Overall, there was a slight improvement when comparing the generated captions with five human captions, except for SPICE score, where a slight decrease was noted in all configurations.

4.4 Comparison with State-of-the-Art Methods

The results generated by KENGIC based on ML-C-1K-f2 and ML-C-2K-f2 were juxtaposed with current state-of-the-art and benchmark image caption generators as shown in Table 4.7. The results are grouped in two sections to distinguish between models which are trained in the paired and unpaired setting. As expected, KENGIC falls short when compared with paired image caption generators since it does not make use of any end-to-end training on image-caption pairs. Therefore, for a fair assessment, KENGIC was compared with unpaired image caption generators. It was found that KENGIC performance is very close to that of current state-of-the-art unpaired generators, and in some metrics it even surpasses current benchmark models. Overall, the model was found to be on par with the current top two best performing unpaired captioning models (i.e., Graph-Align (Gu et al., 2019) and SCS (Ben et al., 2022)). Despite its simplicity, KENGIC based on the ML-C-2K-f2 keywords set achieved the highest METEOR and SPICE scores of 22.6 and 18.5 respectively. On the other hand, the use of ML-C-1K-f2 keywords ranked the model second in terms of CIDEr (69.8) when compared to the more complex SCS (Ben et al., 2022) model that is based on adversarial training.

Table 4.6: Caption metrics based on predicted keywords as evaluated on the testing set. Metrics in brackets were generated based on five ground-truth captions, while the others were computed against HC-$\{0-4\}$ for comparison purposes with previous results.

Keywords	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Objs	10.7 (11.6)	6.7 (7.5)	3.6 (4.1)	1.8 (2.1)	7.8 (8.2)	17.9 (18.6)	22.9 (23.7)	6.5 (5.8)
Objs+sp	**38.1** (**40.3**)	**24.1** (**26.3**)	**13.3** (**14.8**)	**7** (**7.9**)	**13.5** (**14.2**)	**28.7** (**29.8**)	**36.4** (**37.5**)	**10.6** (9.6)
Objs+sp+multi	34.8 (37)	21.4 (23.4)	11.6 (13)	5.9 (6.6)	12.6 (13.2)	27.2 (28.4)	34.7 (35.6)	**10.6** (**9.7**)
ML-objs	10.2 (11.2)	6.5 (7.3)	3.5 (4.1)	1.7 (2)	8.2 (8.7)	18.6 (19.4)	24.3 (25.2)	7.2 (6.4)
ML-C-1K-f2	60.1 (63.6)	40.8 (44.7)	25.7 (28.9)	15.7 (18)	21 (22)	38.7 (40.3)	**68.1** (**69.8**)	19.6 (18.3)
ML-C-1K-f3	32.6 (34.7)	22.8 (25.2)	14.9 (17)	9.3 (10.9)	16 (16.9)	32.9 (34.4)	56.5 (58)	15.7 (14.1)
ML-CL-1K-f2	61.3 (64.8)	41.7 (45.7)	26.5 (29.7)	**16.1** (18.4)	21.2 (22.3)	**38.8** (**40.5**)	67.3 (68.8)	19.7 (18.4)
ML-CL-1K-f3	33.6 (35.9)	23.3 (25.7)	15 (17.1)	9.2 (10.8)	16.2 (17.1)	33.1 (34.7)	55.7 (57.2)	15.7 (14.1)
ML-CLP-1K-f2	58.9 (62.2)	40.1 (43.9)	25.4 (28.6)	15.4 (17.7)	20.6 (21.6)	38.3 (40.1)	67.4 (69.1)	19.4 (18)
ML-CLP-1K-f3	30.8 (32.8)	21.6 (23.8)	14.2 (16.2)	8.9 (10.4)	15.7 (16.5)	32.3 (33.8)	54.7 (56.1)	15.5 (13.9)
ML-C-2K-f2	**62.7** (**66.3**)	**42.4** (**46.5**)	**26.7** (**30.1**)	**16.1** (**18.6**)	**21.5** (**22.6**)	38.7 (40.4)	66.2 (67.8)	**19.8** (**18.5**)
ML-C-2K-f3	29.2 (31.2)	20.6 (22.8)	13.6 (15.5)	8.5 (10)	15.5 (16.4)	32.1 (33.6)	55.1 (56.4)	15.4 (13.8)
ML-CL-2K-f2	58.2 (61.6)	39.7 (43.4)	25.2 (28.4)	15.3 (17.6)	20.6 (21.7)	38.3 (40)	67.1 (68.9)	19.1 (17.7)
ML-CL-2K-f3	30.3 (32.4)	21.2 (23.5)	13.9 (16)	8.6 (10.2)	15.8 (16.6)	32.8 (34.3)	54.8 (56.3)	15.4 (13.8)
ML-CLP-2K-f2	59.6 (63)	40.7 (44.6)	25.9 (29)	15.8 (18.1)	20.8 (21.9)	38.3 (40)	67.8 (69.2)	19.5 (18.3)
ML-CLP-2K-f3	31.4 (33.5)	21.9 (24.2)	14.3 (16.3)	8.9 (10.3)	15.8 (16.6)	32.6 (34.1)	55.1 (56.4)	15.5 (13.9)
ML-C-3K-f2	60.7 (64.3)	41.2 (45.3)	26 (29.4)	15.8 (18.3)	21.1 (22.2)	38.7 (**40.5**)	67.7 (69.4)	19.7 (18.4)
ML-C-3K-f3	31.6 (33.8)	22.1 (24.5)	14.4 (16.5)	9 (10.5)	15.9 (16.7)	32.8 (34.2)	55.9 (57.4)	15.8 (14.2)
ML-CL-3K-f2	60 (63.4)	41 (44.9)	26 (29.3)	15.9 (18.3)	21 (22.1)	38.6 (40.3)	67.6 (69.2)	19.3 (18)
ML-CL-3K-f3	29.9 (32)	20.7 (23)	13.5 (15.3)	8.3 (9.6)	15.6 (16.4)	32.4 (33.8)	54.4 (55.8)	15.4 (13.8)
ML-CLP-3K-f2	58.4 (61.9)	39.8 (43.7)	25.2 (28.5)	15.4 (17.8)	20.6 (21.6)	38.2 (39.8)	67.4 (69)	19.3 (18)
ML-CLP-3K-f3	34.4 (36.7)	23.8 (26.4)	15.5 (17.8)	9.7 (11.3)	16.2 (17.1)	33.6 (35.1)	56.7 (58.4)	15.9 (14.3)

Table 4.7: KENGIC results metrics in percentages compared with paired and unpaired state-of-the-art benchmark models sorted by CIDEr in descending order per each group. First ranked metrics are listed in bold while second ranked metrics are italicised.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr ↓	SPICE
Paired								
Mind's Eye (Chen and Lawrence Zitnick, 2015)	-	-	-	18.8	19.6	-	-	-
NeuralTalk (Karpathy and Fei-Fei, 2017)	62.5	45.0	32.1	23.0	19.5	-	66.0	-
Up-Down (Anderson et al., 2018)	77.2	-	-	36.2	27.0	54.9	113.5	20.3
SGAE-KD (Yang et al., 2020)	78.2	-	-	37.3	28.1	57.4	117.1	21.3
MT (Yu et al., 2020)	77.3	-	-	37.4	28.7	57.4	119.6	-
Unpaired								
Language-Pivoting (Gu et al., 2018b)	46.2	24.0	11.2	5.4	13.2	-	17.7	-
Adverserial+Reconstruction (Feng et al., 2019)	58.9	40.3	27.0	18.6	17.9	43.1	54.9	11.1
USGAE (Yang et al., 2020)	60.8	-	-	17.1	19.1	43.8	55.1	12.8
Multimodal Embeddings (Laina et al., 2019)	-	-	-	19.3	20.2	45.0	61.8	12.9
IGGAN (Cao et al., 2020)	-	-	-	21.9	46.5	21.1	64.0	14.5
KENGIC (ML-C-2K-f2)	*66.3*	46.5	30.1	18.6	**22.6**	40.4	67.8	**18.5**
Graph-Align (Gu et al., 2019)	**67.1**	*47.8*	*32.3*	*21.5*	20.9	*47.2*	69.5	15.0
KENGIC (ML-C-1K-f2)	63.6	44.7	28.9	18	22.0	40.3	*69.8*	*18.3*
SCS (Ben et al., 2022)	**67.1**	**47.9**	**33.4**	**22.8**	*21.4*	**47.7**	**74.7**	15.1

4.5 Qualitative Analysis

To complement the quantitative analysis, a qualitative analysis was conducted to get deeper insights into the presented metrics. This was performed on 200 randomly sampled images from the 495 images which were captioned by all keyword setups, including the human extracted keywords (HK), the predicted objects by the Faster R-CNN, and by the ML-Decoder (ML-objs, ML-C-1K-f2 and ML-C-2K-f2). The low intersection is due to empty keyword sets, especially in HK-f4 and HK-f5 which had 1252 and 1963 images without corresponding keywords, respectively.

When conditioning KENGIC on human extracted keywords (HK) it was found that in simple scenarios it was unable to construct meaningful and relevant captions. For example, for Fig. 4.2(a), the model found it hard to combine the "player" and "woman" as one entity and ended up generating the captions by mentioning both entities independently. The model was found to hallucinate when using extracted keywords. This is shown, for example, in Fig. 4.2(b) when using HK-{n,na,nap}. However, with the introduction of further keywords (HK-napv, HK-f2) captions were more relevant, as confirmed by the captions based on HK-napv and HK-f2 of the same example (i.e., Fig. 4.2(b)). Incorrect captions were also generated due to incorrectly used words found in the ground-truth captions. For example, the *ocean waves* found in Fig. 4.2(c) were incorrectly perceived as "snow" and therefore lead to inconsistent captions. This analysis also revealed that specific and more restrictive keywords were not always mentioned in the generated captions as no paths were found to connect the set of keywords. For this reason, captions ended up being incomplete as shown for example in Figure 4.2(d). Considering the lack of fluency of the captions generated based on HK-f1, their corresponding SPICE metrics were found to be considerably high and in most cases they exceeded captions with better structure like in Fig. 4.2(e) and (f) that were generated based on HK-f2. However, despite the large number of keywords of HK-f1, surprisingly, this set still was effective in some cases as illustrated in Fig 4.2(g) and (h).

Fig. 4.3 presents examples of good quality captions which were generated based on human extracted keywords. For instance, Fig. 4.3(c) and (d) shows how both images were correctly described using HK-na and HK-napv respectively. Interestingly, in spite of the increased precision, it was found that by the introduction of human extracted keywords from one reference caption, the generated captions can lack the generalisation aspect when compared to the remaining four captions. In such cases, the candidate captions were considerably penalised as shown in Fig. 4.3(a). In this case, HK-napv scored the least CIDEr score when compared to other keyword sets, despite its accuracy. This is further noted in Fig. 4.3(b), where HK-f2 achieved better scores, in spite of its lack of detail when

compared to HK-na/napv. This confirms that current evaluation metrics prioritise generic captions over precise and better quality captions. In some occasions, the introduction of further keywords was found to produce more generic captions as no direct paths were found during the construction of sentences as illustrated in Fig. 4.3(e), while in others it increased the specificity as shown in Fig. 4.3(f).

Similar trends were observed when analysing the captions which where generated based on the predicted keywords. Incorrectly captioned images are illustrated in Fig. 4.4, while good quality captioned images are found in Fig 4.5. Overall, it was found that the Objs and ML-Objs keyword sets were very similar, while the ML-C-1K-f2 keywords were found to be very relevant and accurate for the captions. Similar results were observed for the ML-C-2K-f2. From this analysis, it was found that in some cases, captions mentioned keywords more than once, while verbs were not used in their correct form as shown in Fig. 4.4(a). Furthermore, the synonymous nature of the predicted keywords proved to be a problem for the overall generation process. As shown in Fig. 4.4(b), the keywords of ML-C-1K-f2 are: {“blue”, fly”, “airplane”, “sky”, “plane”}. This resulted in a caption which mention neither the “plane” nor the “airplane”. Presumably, this problem is due to the conflicting keywords. Since KENGIC attempts to generate long captions, this was found to hinder the captioning processes in some occasions. For example, the caption generated for Fig. 4.4(c) based on ML-C-1K-f2 hallucinated after correctly linking the predicted keyword set by a “dog jumping up to catch a frisbee in his mouth”. Some captions also lacked important details. For example, all the captions for Fig. 4.4(d) failed to mention the girl that is sitting under the man who is riding the skateboard. Caption quality was also found to be affected due to incorrectly predicted keywords as found in Fig. 4.4(e). Finally, captions were generated incorrectly due to the complexity of such images and the degree of reasoning they require as shown, for example, in Fig. 4.4(f).

On the other hand, KENGIC was also found to be very effective in captioning images using the predicted keywords as illustrated in Fig 4.5. Generic images like for example, Fig 4.5(a) and (b) were captioned with fluent and accurate descriptions. Less common images were also found to be described with relevant captions. For instance, Fig 4.5(c), given the predicted keyword set: {“skiing”, “walk”, “snow”, “woman”, “street”, “people”, “ski”}, KENGIC was able to combine the keywords and generate the caption: “People walk down a snow covered street and a woman skiing in the snow on a ski”. However, despite the relevancy of this caption, CIDEr score was found to be relatively low. This confirms again the misalignment between caption quality and the evaluated metrics. Another uncommon combination is Fig 4.5(d). In this case, KENGIC correctly linked “leash”, “dog”, and “skateboard” with the caption: “dog wearing a leash on a skateboard”. Surprisingly, this analysis also revealed that the ML-Decoder was able to guess the quantity of objects

in images. This is shown in Fig 4.5(e), where the model correctly counted “three” buses but used the word “bus” instead of “buses” in the captions. Most probably, this is because “buses” is less popular in the vocabulary. Finally, KENGIC was also found to be effective in complex scenarios, like in Fig 4.5(f), where with the help of the predicted keywords (i.e., “school”, “mirror”, “bus”, “view”), managed to construct the sentence of a “view of a mirror in a school bus”. Less precise captions were generated due to the implicit use of spatial prepositions. As can be found in Fig. 4.6, the used spatial prepositions were not applied in accordance with the corresponding images. Therefore, this led to captions with prepositions that were not grounded in images. For instance, Fig 4.6(a) was captioned by a “large blue truck pulling a trailer on a log”. In this case, the spatial preposition “with” would have been a better fit instead of the preposition “on”. Moreover, the preposition “in front of” is more suitable than “with” in the caption “black dog with a book shelf” for Fig 4.6(b); whilst the preposition “behind” is better than “on” when describing Fig. 4.6(c) by a “black bear on a log”. Furthermore, the prepositions “in front of” or “near” would make the caption “small boat in the ocean with a sail on the island” more relevant to Fig. 4.6(d). This confirms the usefulness of explicit spatial relation detection in image caption generation which is going to be addressed in the following chapter.

a (id:11035)

HC-0: A woman tennis player is about to return the ball.

HK-n/na/nap: player, woman, ball, tennis
Caption: **Woman holding a tennis player with a ball.**
B-1: 87.5, **B-2**: 35.36, **B-3**: 0, **B-4**: 0, **M**: 23.63, **R-L**: 43.57, **C**: 90.89, **S**:19.05

HK-napv: player, woman, ball, tennis, return
Caption: **Woman holding a tennis player about to hit a ball up to return.**
B-1: 46.15, **B-2**: 19.61, **B-3**: 0, **B-4**: 0, **M**: 20.61, **R-L**: 35.62, **C**: 45.32, **S**:30

HK-f2: ball, tennis, player, playing, woman
Caption: **Woman playing tennis at a ball player.**
B-1: 74.30, **B-2**: 46.34, **B-3**: 33.39, **B-4**: 0.01, **M**: 28.34, **R-L**: 39.52, **C**: 129.20, **S**:28.57

b (id:7386)

HC-0: Students wait for a bus to come pick them up.

HK-n/na/nap: bus, students
Caption: **Bus full of students.**
B-1: 18.39, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 7.03, **R-L**: 15.72, **C**: 23.59, **S**:10.53

HK-napv: bus, students, wait, come, pick
Caption: Bus stopped to pick.
B-1: 9.20, **B-2**:0, **B-3**: 0, **B-4**: 0, **M**: 6.8, **R-L**: 15.72, **C**: 23.92 **S**:10.53 (id: 7386)

HK-f2: outside, kits, school, standing, young, bus
Caption: Young kids standing.
B-1: 18.89, **B-2**:18.89, **B-3**: 18.89 **B-4**: 0.60, **M**: 20.64, **R-L**: 38.85, **C**: 58.71 **S**:30.00

c (id:20178)

HC-0: **A man that is holding a surfboard in the snow.**

HK-n/na/nap: surfboard, snow, man
Caption: **Man on a surfboard in the snow.**
B-1: 53.68, **B-2**: 25.93, **B-3**: 0, **B-4**: 0, **M**: 21.15, **R-L**: 48.90, **C**: 116.14, **S**:11.76

HK-napv: surfboard, snow, man, holding
Caption: **Man holding a surfboard in the snow on a tennis court holding.**
B-1: 58.33, **B-2**:32.57, **B-3**: 21.97, **B-4**: 0, **M**: 29.21, **R-L**: 48.88, **C**: 110.50, **S**:9.09

HK-f2: holding, standing, man, surfboard, beach
Caption: **Man holding a surfboard on the beach standing on a tennis court holding.**
B-1: 69.23, **B-2**:53.71, **B-3**: 37.43, **B-4**: 0, **M**: 36.98, **R-L**: 56.39, **C**: 156.83, **S**:17.39

d (id:30202)

HC-0: The dog is playing catch with his master.

HK-n/na/nap: dog, master
Caption: **Dog with a.**
B-1: 9.02, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 9.97, **R-L**: 30.58, **C**: 28.88, **S**:11.76

HK-napv: playing, master, dog, catch
Caption: **Dog playing with a.**
B-1: 14.33, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 9.70, **R-L**: 28.77, **C**: 26.40, **S**: 0

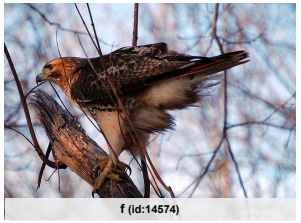
HK-f2: frisbee, dog, moth, cath, brown
Caption: Dog with a frisbee in its mouth.
B-1: 64.41, **B-2**:56.81, **B-3**: 45.95, **B-4**: 0.01, **M**: 23.91, **R-L**: 57.01, **C**: 127.32, **S**:36.36

e (id:40062)

HC-0: A giraffe walks on the tundra tree-lined park.

HK-f1: scanty, grassy, walking, lined, making, single, trees, tundra, walks, giraffe, way, area, brown, field, tree, park, ground, juvenile, standing, grass
Caption: **Brown grass near a tree lined area of a field with a single giraffe standing next to trees in the park on a grassy ground and a walking in a field with a large.**
B-1: 50, **B-2**: 36.93, **B-3**: 23.38, **B-4**: 0, **M**: 31.27, **R-L**: 34.27, **C**: 0.05, **S**: 35.90

HK-f2: grass, giraffe, trees, walking
Caption: Giraffe walking in the grass next to trees.
B-1: 75, **B-2**: 46.29, **B-3**: 32.93, **B-4**: 0.01, **M**: 22.01, **R-L**: 50, **C**: 136.53, **S**: 17.39



f (id:14574)

HC-0: A bird on a branch in a tree.

HK-f1: ruffling, feathers, photo, perched, close, black, top, branch, bird, brown, tree, sitting, small
Caption: **Black and white photo of a small bird sitting on a tree branch with its feathers on a brown and green bird perched on the top in a small bird is sitting close to a.**
B-1: 54.29, **B-2**: 47.28, **B-3**: 37.84, **B-4**: 26.7, **M**: 37.77, **R-L**: 40.53, **C**: 0.08, **S**: 55.56

HK-f2: tree, branch, perched, bird
Caption: Bird perched on a tree branch.
B-1: 71.65, **B-2**: 71.65, **B-3**: 71.65, **B-4**: 71.65, **M**: 36.13, **R-L**: 71.76, **C**: 311.2, **S**: 40

g (id:28500)

HC-0: A group of people flying kites over a beach.

HK-f1: kite, bikes, kites, wind, bikers, people, lower, beach, attached, sky, group, red, flying, assortment, lobster, shaped
Caption: **People are flying a kite in the sky on a beach with a group of people on bikes and a red kite that is on the beach flying kites shaped.**
B-1: 50, **B-2**: 29.36, **B-3**: 14.55, **B-4**: 0, **M**: 27.16, **R-L**: 28.40, **C**: 0.22, **S**: 34.48

HK-f2: people, kites, flying
Caption: People flying kites.
B-1: 13.53, **B-2**: 13.53, **B-3**: 13.53, **B-4**: 0.43, **M**: 23.41, **R-L**: 45.86, **C**: 101.35, **S**: 25

h (id:8696)

HC-0: A small white rusted boat floating across a body of water.

HK-f1: rusted, floating, across, ocean, boat, body, middle, older, trees, background, rusty, open, large, water, old, white, small
Caption: **Small white boat floating in the middle of a large body of water.**
B-1: 84.62, **B-2**: 75.11, **B-3**: 63.53, **B-4**: 52.66, **M**: 33.83, **R-L**: 53.43, **C**: 154.68, **S**: 51.61

HK-f2: white, ocean, small, boat, body, water
Caption: Small white boat in the ocean on a body of water near a body.
B-1: 85.71, **B-2**: 62.90, **B-3**: 32.06, **B-4**: 0, **M**: 32.41, **R-L**: 51.55, **C**: 103.06, **S**: 50

Figure 4.2: Incorrect captions based on human extracted keywords (top row) and good quality captions based on HK-f1 (bottom row).

a (id:28106)

HC-0: A beach crowded with people and many different colored umbrellas.

HK-n: people, umbrellas, beach
Caption: People on the beach while holding umbrellas.
B-1: 53.68, **B-2**: 36.67, **B-3**: 27.24, **B-4**: 0, **M**: 18.73, **R-L**: 45.61, **C**: 107.01, **S**: 28.57

HK-na/nap: many, umbrellas, different, people, beach
Caption: People on a beach filled with many different.
B-1: 55.16, **B-2**: 26.37, **B-3**: 0, **B-4**: 0, **M**: 17.23, **R-L**: 32.68, **C**: 48.77, **S**: 22.22

HK-napv: different, umbrellas, beach, colored, people, crowded, many
Caption: **People in a crowded beach filled with many different colored umbrellas.**
B-1: 45.45, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 24.77, **R-L**: 30.55, **C**: 64.86, **S**: 25.00

HK-f2: umbrellas, people, colorful, beach
Caption: People on a beach on the colorful umbrellas.
B-1: 77.22, **B-2**:54.04, **B-3**: 0, **B-4**: 0, **M**: 26.20, **R-L**: 45.56, **C**: 146.65, **S**:38.10

d (id:38049)

HC-0: A giraffe grazing in pen next to a tree trunk.

HK-na/napv: next, tree, giraffe, pen, trunk, grazing
Caption: **Giraffe grazing in a pen next to a tree trunk.**
B-1: 70, **B-2**: 55.78, **B-3**: 42.69, **B-4**: 32.47, **M**: 19.90, **R-L**: 40.45, **C**: 100.41, **S**:14.81

HK-f2: tree, grass, giraffe, next, grazing
Caption: **Giraffe grazing in the grass next to a tree.**
B-1: 88.89, **B-2**:66.67, **B-3**: 50.26, **B-4**: 38.14, **M**: 23.12, **R-L**: 48.41, **C**: 144.11, **S**:30.77

b (id:39655)

HC-0: A cake with icing and cartoon cake toppers.

HK-na/napv: cartoon, icing, cake, toppers
Caption: **Cake decorated with icing.**
B-1: 16.73, **B-2**: 11.16, **B-3**: 0, **B-4**: 0, **M**: 14.86, **R-L**: 36.90, **C**: 24.12, **S**:8.33

HK-f2: table, birthday
Caption: Table at a birthday.
B-1: 16.73, **B-2**:11.16, **B-3**: 0, **B-4**: 0, **M**: 15.07, **R-L**: 26.52, **C**: 66.10, **S**:15.38

e (id:23367)

HC-0: A boy with glasses playing a game with a Nintendo Wii controller.

HK-n/na/nap: game, glasses, controller, wii, boy
Caption: **Boy with a nintendo wii game controller.**
B-1: 74.30, **B-2**: 56.75, **B-3**: 38.22, **B-4**: 0.01, **M**: 30.84, **R-L**: 65.87, **C**: 128.36, **S**: 38.10

HK-napv: controller, nintendo, glasses, playing, game, boy, wii
Caption: **Boy playing a game.**
B-1: 36.79, **B-2**: 30.04, **B-3**: 25.51, **B-4**: 0, **M**: 27.06, **R-L**: 57.55, **C**: 125.22, **S**: 23.53

HK-f2: video, controller, boy, glasses, young, playing, wii, game
Caption: **Young boy holding a video game controller playing a wii video game.**
B-1: 75.0, **B-2**: 58.39, **B-3**: 40.85, **B-4**: 29.50, **M**: 41.74, **R-L**: 62.24, **C**: 167.17, **S**:16.0

c (id:13383)

HC-0: A cat in between two cars in a parking lot.

HK-na/napv: parking, cat, lot, cars
Caption: **Cars in a parking lot with a cat.**
B-1: 87.5, **B-2**: 70.71, **B-3**: 55.03, **B-4**: 42.73, **M**: 22.61, **R-L**: 36.53, **C**: 112.29, **S**:25.81

HK-f2: cars, two, white, cat, parked, parking, lot, grey
Caption: Two white black and grey cat with a lot of a parked cars next to a black and white on a parking.
B-1: 54.55, **B-2**: 32.23, **B-3**: 0, **B-4**: 0, **M**: 27.64, **R-L**: 32.24, **C**: 26.19, **S**:29.41

f (id:23942)

HC-0: A cat that is standing on a laptop

HK-na: laptop, cat
Caption: Cat using a laptop.
B-1: 27.59, **B-2**: 18.39, **B-3**: 0, **B-4**: 0, **M**: 18.61, **R-L**: 47.16, **C**: 101.70, **S**:25.00

HK-napv: laptop, cat, standing
Caption: **Cat standing on a laptop.**
B-1: 54.88, **B-2**:54.88, **B-3**: 54.88, **B-4**: 54.88, **M**: 34.92, **R-L**: 67.93, **C**: 254.34, **S**:33.33

HK-f2: white, cat, standing, laptop, black, top
Caption: Black and white cat sitting on top of a laptop that is standing in a black and white
B-1: 72.22, **B-2**:65.18, **B-3**: 57.07, **B-4**: 47.18, **M**: 43.52, **R-L**: 55.17, **C**: 90.67, **S**:57.14

Figure 4.3: Correct captions based on human extracted keywords.

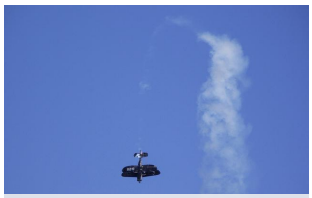
a (id:8607)

HC-0: A cat playing with a toy on a sofa.

Objs: couch, cat
Caption: Cat on a couch.
B-1: 22.31, **B-2**: 18.22, **B-3**: 15.47, **B-4**: 0, **M**: 21, **R-L**: 53.04, **C**: 71.30, **S**: 21.05

ML-Objs: cat, couch, banana
Caption: Cat on a couch holding a banana
B-1: 55.84, **B-2**: 34.82, **B-3**: 25.09, **B-4**: 0, **M**: 27.85, **R-L**: 45.61, **C**: 104.21, **S**: 30

ML-C-1K-f2: white, cat, banana, play, couch, toy, black
Caption: **Black and white cat on a couch with a toy banana and a black and white play.**
B-1: 52.94, **B-2**: 40.67, **B-3**: 32.10, **B-4**: 22.05, **M**: 33.58, **R-L**: 44.58, **C**: 89.99, **S**: 35.71



b (id:30864)

HC-0: The airplane is flying near a cloud in the sky.

Objs/ML-objs: airplane
Caption: Airplane.
B-1: 0.09, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 8.54, **R-L**: 19.49, **C**: 17.16, **S**: 22.22

ML-C-1K-f2: blue, fly, airplane, sky, plane
Caption: **Blue sky in the.**
B-1: 27.59, **B-2**: 18.39, **B-3**: 0, **B-4**: 0, **M**: 13.54, **R-L**: 31.44, **C**: 55.72, **S**: 16.67

c (id:30202)

HC-0: The dog is playing catch with his master.

Objs/ML-objs: dog, frisbee
Caption: Dog catching a frisbee.
B-1: 21.49, **B-2**: 14.33, **B-3**: 0, **B-4**: 0, **M**: 23.20, **R-L**: 43.16, **C**: 39.32/39.42, **S**: 22.22

ML-C-1K-f2: mouth, catch, dog, frisbee, jumping
Caption: **Dog jumping up to catch a frisbee in his mouth and a man is trying to catch.**
B-1: 58.82, **B-2**: 46.97, **B-3**: 38.89, **B-4**: 30.28, **M**: 34.82, **R-L**: 56.89, **C**: 115.08, **S**: 36.36

d (id:22093)

HC-0: A girl hitching a ride with her Dad on a skateboard.

Objs: person, handbag, skateboard
Caption: **Person on a skateboard.**
B-1: 27.59, **B-2**: 26.01, **B-3**: 23.18, **B-4**: 0, **M**: 15.27, **R-L**: 39.78, **C**: 88.53, **S**: 9.09

ML-Objs: handbag, person, skateboard, backpack
Caption: **Person on a skateboard.**
B-1: 27.59, **B-2**: 26.01, **B-3**: 23.18, **B-4**: 0, **M**: 15.27, **R-L**: 39.78, **C**: 88.53, **S**: 9.09

ML-C-1K-f2: man, skateboard
Caption: **Man riding a skateboard.**
B-1: 36.79, **B-2**: 36.79, **B-3**: 36.79, **B-4**: 36.79, **M**: 30.89, **R-L**: 57.55, **C**: 171.75, **S**: 19.05

e (id:30609)

HC-0: A shot of a tennis play in mid shot.

Objs/ML-Objs: person, tennis racket
Caption: Person swinging a tennis.
B-1: 14.33, **B-2**: 11.70, **B-3**: 0, **B-4**: 0, **M**: 16.24, **R-L**: 28.77, **C**: 45.58/45.57, **S**: 0

Objs+sp: tennis, racket, person
Caption: Person holding a tennis racket.
B-1: 17.97, **B-2**: 14.21, **B-3**: 0, **B-4**: 0, **M**: 15.8, **R-L**: 27.17, **C**: 57.11, **S**: 11.76

ML-C-1K-f2: ball, tennis, woman, player, serve
Caption: **Woman holding a tennis player about to serve a tennis ball.**
B-1: 72.73, **B-2**: 66.06, **B-3**: 62.35, **B-4**: 54.91, **M**: 38.07, **R-L**: 71.29, **C**: 244.85, **S**: 42.11

f (id:29756)

HC-0: A group of men on bikes hitching a ride on the back of a bus.

Objs: person, bus, bicycle
Caption: Person riding a bicycle with a bus.
B-1: 37.15, **B-2**: 23.17, **B-3**: 0, **B-4**: 0, **M**: 11.6, **R-L**: 39.52, **C**: 41.87, **S**: 17.39

ML-Objs: bus, person, bicycle, backpack
Caption: **Person with a backpack and a bus on a bicycle.**
B-1: 50, **B-2**: 23.57, **B-3**: 0, **B-4**: 0, **M**: 14.49, **R-L**: 34.01, **C**: 38.89, **S**: 14.81

ML-C-1K-f2: bike, bicycle, pink, people, bus, group
Caption: **People in a pink bus with a bicycle on a bike at a group.**
B-1: 53.33, **B-2**: 27.60, **B-3**: 0, **B-4**: 0, **M**: 22.83, **R-L**: 31.65, **C**: 46.44, **S**: 34.48

Figure 4.4: Incorrect captions based on predicted keywords.

a (id:7121)

HC-0: The top of a building with a clock on it.

Objs/Objs+sp+multi/ML-Objs: clock
Caption: Clock.
B-1: 0.09, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 7.62, **R-L**: 19.49, **C**: 29.61/29.73/29.54, **S**: 8.7

ML-C-1K-f2: tower, top, clock, building
Caption: **Clock tower at the top of a building.**
B-1: 62.5, **B-2**: 42.26, **B-3**: 30.99, **B-4**: 0, **M**: 27.18, **R-L**: 30.70, **C**: 90.2, **S**: 14.29

d (id:13910)

HC-0:A dog that is on a black skateboard.

Objs/ML-Objs: dog, person, skateboard
Caption: **Person with a dog on a skateboard**.
B-1: 74.30, **B-2**: 46.34, **B-3**: 33.39, **B-4**: 0.01, **M**: 21.76, **R-L**: 45.61, **C**: 120.41/119.81, **S**: 30

Objs+sp+multi: people, skateboard, dog
Caption: People on a skateboard with a dog.
B-1: 61.92, **B-2**: 42.30, **B-3**: 31.42, **B-4**: 0.01, **M**: 18.86, **R-L**: 34.21, **C**: 113.74, **S**: 20

ML-C-1K-f2: leash, dog, skateboard
Caption: **Dog wearing a leash on a skateboard.**
B-1: 61.92, **B-2**: 42.30, **B-3**: 31.42, **B-4**: 0.01, **M**: 21.20, **R-L**: 45.61, **C**: 73.34, **S**: 22.22

b (id:28074)

HC-0: A group of people at the beach with umbrellas.

Objs: person, backpack, umbrella
Caption: Person holding an umbrella with a backpack.
B-1: 32.21, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 17.92, **R-L**: 20.1, **C**: 14.1, **S**: 8.7

Objs+sp+multi: people, backpack, umbrella
Caption: People with an umbrella with a backpack.
B-1: 32.21, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 10.72, **R-L**: 24.45, **C**: 24.59, **S**: 17.39

ML-Objs: umbrella, handbag, person, backpack
Caption: Person in a.
B-1: 13.53, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 9.01, **R-L**: 24.06, **C**: 5.13, **S**: 10.53

ML-C-1K-f2: people, beach, umbrella
Caption: **People on a beach under an umbrella.**
B-1: 53.68, **B-2**: 36.67, **B-3**: 27.24, **B-4**: 0, **M**: 21.76, **R-L**: 52.59, **C**: 67.19, **S**: 34.78

e (id:21711)

HC-0: The picture of three buses on a lot.

Objs/ML-objs: bus
Caption: Bus.
B-1: 0.9, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 6.09, **R-L**: 11.53, **C**: 1.9, **S**: 11.11

Objs+sp+multi: buses
Caption: Buses.
B-1: 0.9, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 6.88, **R-L**: 19.49, **C**: 36.06, **S**: 11.11

ML-C-1K-f2: three, decker, double, park, bus, green
Caption: **Three double decker bus.**
B-1: 36.79, **B-2**: 36.79, **B-3**: 29.2, **B-4**: 0.01, **M**: 31.75, **R-L**: 34.4, **C**: 102.82, **S**: 30

c (id:20078)

HC-0: A group of people riding skis down a snow covered street.

Objs: skis, person, car, backpack
Caption: Person in a car with a backpack and a person on skis.
B-1: 41.67, **B-2**: 19.46, **B-3**: 0, **B-4**: 0, **M**: 13.37, **R-L**: 19.55, **C**: 48.32, **S**: 14.29

Objs+sp+multi: skis, cars, backpack, people
Caption: People ski in a backpack on a cars.
B-1: 66.19, **B-2**: 28.89, **B-3**: 0, **B-4**: 0, **M**: 14.56, **R-L**: 40.94, **C**: 17.55, **S**: 23.08

ML-Objs: truck, skis, car, backpack, person
Caption: Person with a backpack and a truck in a car of a person on skis.
B-1: 40, **B-2**: 16.9, **B-3**: 0, **B-4**: 0, **M**: 12.88, **R-L**: 26.18, **C**: 35.23, **S**: 13.33

ML-C-1K-f2: skiing, walk, snow, woman, street, people, ski
Caption: **People walk down a snow covered street and a woman skiing in the snow on a ski.**
B-1: 58.82, **B-2**: 42.87, **B-3**: 33.25, **B-4**: 26.92, **M**: 30.32, **R-L**: 44.97, **C**: 51.71, **S**: 26.67

f (id: 36003)

HC-0: The reflection of a bus in a vehicle.

Objs/Objs+sp+multi: person, bus
Caption: Person at a bus.
B-1: 18.39, **B-2**: 15.02, **B-3**: 0, **B-4**: 0, **M**: 12.91, **R-L**: 31.44, **C**: 57.79/57.95, **S**: 13.33

ML-Objs: bus
Caption: Bus.
B-1: 0.09, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 8.54, **R-L**: 19.49, **C**: 32.33, **S**: 15.38

ML-C-1K-f2: school, mirror, bus, view
Caption: **View of a mirror in a school bus.**
B-1: 87.5, **B-2**: 79.06, **B-3**: 47.05, **B-4**: 0.01, **M**: 32.48, **R-L**: 58.21, **C**: 187.16, **S**: 33.33

Figure 4.5: Correct captions based on predicted keywords.

a (id:4414)

HC-0: A blue and silver truck with logs trees and wires.

Objs/Objs+sp+multi: car, truck
Caption: Truck and a car.
B-1: 18.39, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 9.7, **R-L**: 15.72, **C**: 46.92/47.04 **S**: 11.11

ML-Objs: truck
Caption: Truck.
B-1: 0.09, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 7.62, **R-L**: 19.49, **C**: 39.28, **S**: 11.76

ML-C-1K-f2: trailer, log, blue, large, truck
Caption: **Large blue truck pulling a trailer on a log.**
B-1: 55.56, **B-2**: 37.27, **B-3**: 27.07, **B-4**: 0, **M**: 21.86, **R-L**: 35.67, **C**: 67.23, **S**: 38.10

b (id:4564)

HC-0: A dog is laying* in a chair in front of a book shelf.

Objs: book, dog
Caption: Book with a dog.
B-1: 28.65, **B-2**: 16.54, **B-3**: 0, **B-4**: 0, **M**: 11.62, **R-L**: 28.77, **C**: 38.89, **S**: 21.05

Objs+sp+multi: books, dog
Caption: Books and a dog.
B-1: 21.49, **B-2**: 14.33, **B-3**: 0, **B-4**: 0, **M**: 11.58, **R-L**: 28.77, **C**: 38.01, **S**: 22.22

ML-Objs: book, bed, dog
Caption: Dog on a bed reading a book.
B-1: 53.68, **B-2**: 36.67, **B-3**: 0, **B-4**: 0, **M**: 14.82, **R-L**: 45.61, **C**: 41.23, **S**: 20

ML-C-1K-f2: dog, shelf, black, book
Caption: **Black dog with a book shelf.**
B-1: 60.65, **B-2**: 46.98, **B-3**: 32.23, **B-4**: 0.01, **M**: 23.51, **R-L**: 49.35, **C**: 80.49, **S**: 38.10

c (id:26160)

HC-0: A bear that is laying* down in the dirt.

Objs/ML-objs: bear
Caption: Bear.
B-1: 0.09, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 7.23, **R-L**: 19.49, **C**: 31.14/31.59, **S**: 0

Objs/ML-objs: bears
Caption: Bears.
B-1: 0, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 0, **R-L**: 0, **C**:0, **S**: 11.1

ML-C-1K-f2: log, bear, black
Caption: **Black bear on a log.**
B-1: 43.90, **B-2**: 24.54, **B-3**: 0, **B-4**: 0, **M**: 13.21, **R-L**: 40.76, **C**: 57.01, **S**: 19.05

d (id:35170)

HC-0: A white sail boat sailing towards island.

Objs/Objs+sp+multi: boat/boats
Caption: Boat./Boats.
B-1: 0, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 0, **R-L**: 0, **C**: 0, **S**: 0

ML-objs: person, boat
Caption: Person on a boat.
B-1: 11.16, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 3.43, **R-L**: 13.26, **C**: 0.04, **S**: 0

ML-C-1K-f2: sail, ocean, boat, island, small
Caption: S**mall boat in the ocean with a sail on the island.**
B-1: 63.64, **B-2**: 35.68, **B-3**: 0, **B-4**: 0, **M**: 14.21, **R-L**: 39.15, **C**: 60.12, **S**: 6.67

* “laying” was incorrectly used instead of “lying” in the human caption.

Figure 4.6: Captions generated based on predicted keywords with incorrect use of spatial relations.

4.6 Human Evaluation

A human evaluation was carried out to analyse the quality of the generated captions. For this analysis, 500 captioned test images were chosen randomly. Inspired by the evaluation strategy of Mitchell et al. (2012) and Tanti (2019), the evaluators were instructed to assess the accuracy and fluency of the generated captions. The accuracy of the caption refers to how relevant the caption is to the image content, while fluency corresponds to how well the caption is written, irrespective of its accuracy. In this study, ten Maltese graduate evaluators with English bilingual proficiency were given 55 images at a time and were first trained over a few examples and then asked to rate both the accuracy and fluency of image captions on a five-point Likert scale which includes (1) Strongly Disagree, (2) Disagree, (3) Neutral, (4) Agree, and (5) Strongly Agree, as illustrated in Fig. 4.7. The evaluators were given captions that were generated from HK-f2, Objs+sp, ML-C-1K-f2, ML-C-2K-f2 keyword sets, and a human written caption (HC-0) to provide a sufficient upper bound. This results in a total of 2750 captions (55 images $\times$ 5 captions $\times$ 10 evaluators). These captions were non-identifiable and shuffled each time they were presented to the evaluators. Each set of 55 images was composed of 70% (35) unique images per evaluator, 20% (10) images common to all evaluators, to compute the inter-rate agreement, and 10% (5) images which an evaluator would have already seen, to calculate the intra-rater agreement.

The intra- and inter-rater agreements were calculated over all captions to evaluate the reliability of the collected evaluations. This was carried out by calculating both the percentage agreement and Cohen's kappa statistic (κ) (Cohen, 1960). The latter was specifically computed to also take into account the random chance agreement between the specified ratings. The overall mean intra-rater percentage agreements for the correctness and fluency were 59.6% and 64.4% respectively, while the mean kappa scores were 0.45 for correctness and 0.49 for fluency. According to Cohen (1960), this reflects a moderate intra-agreement and therefore confirms that the evaluators were generally not consistent in their own evaluations. The overall intra-rater agreements are listed in Table 4.8, whilst the intra-agreements per each configuration are tabulated in Table 4.9. Interestingly, the evaluators were least consistent when rating the accuracy of Objs-sp ($\kappa = 0.26$) and the fluency of ML-C-1K-f2 ($\kappa = 0.28$). On the other hand, the highest recorded intra-agreement was on ML-C-2K-f2's accuracy ($\kappa = 0.49$) and HC-0's fluency ($\kappa = 0.61$).

Unsurprisingly, the inter-rater agreements were much lower. The overall inter-rater percentage agreements, when computed on the rated accuracy and fluency were 40.6% and 44.2% respectively, while the kappa scores were 0.21 and 0.19, respectively which confirms a slight agreement between the evaluators. This shows that assessing image

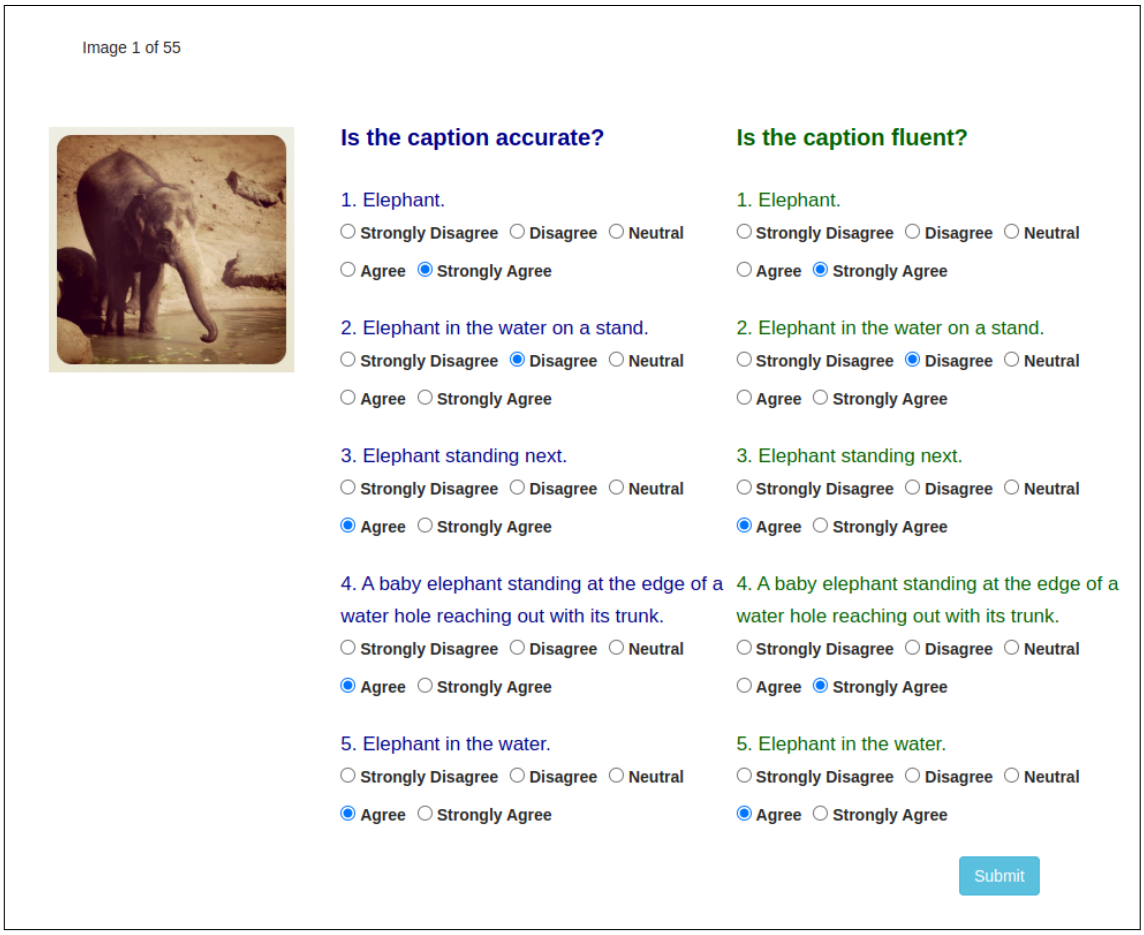


Figure 4.7: Screenshot of the human evaluation interface with corresponding annotations per each caption.

Table 4.8: Evaluator's percentage and Cohen's kappa intra-rater agreements for both accuracy and fluency.

	Accuracy		Fluency	
Evaluator	Percentage Agreement	Cohen's kappa Coeffiencient	Percentage Agreement	Cohen's kappa Coefficient
1	0.72	0.58	0.80	0.68
2	0.64	0.53	0.52	0.32
3	0.52	0.33	0.60	0.37
4	0.60	0.51	0.56	0.41
5	0.60	0.45	0.52	0.37
6	0.52	0.37	0.60	0.40
7	0.32	0.06	0.72	0.62
8	0.76	0.67	0.68	0.56
9	0.64	0.48	0.72	0.61
10	0.64	0.53	0.72	0.60

Table 4.9: Evaluator's percentage and Cohen's kappa intra- and inter- (in brackets) rater agreements per each configuration.

	Accuracy		Fluency	
Model	Percentage Agreement	Cohen's kappa Coeffiencient	Percentage Agreement	Cohen's kappa Coefficient
HC-0	0.64 (0.48)	0.37 (0.16)	0.82 (0.51)	0.61 (0.04)
HK-f2	0.64 (0.35)	0.48 (0.17)	0.62 (0.40)	0.43 (0.23)
Objs-sp	0.50 (0.51)	0.26 (0.34)	0.64 (0.53)	0.44 (0.27)
ML-C-1K-f2	0.56 (0.37)	0.30 (0.20)	0.54 (0.40)	0.28 (0.24)
ML-C-2K-f2	0.64 (0.32)	0.49 (0.16)	0.64 (0.36)	0.38 (0.18)

Table 4.10: Human Evaluation results for each configuration as compared to the CIDEr metric on 550 images. Results include the median and (mean, standard deviation) in brackets for the human rated accuracy and fluency, and the CIDEr metric.

Model	Accuracy	Fluency	CIDEr
HC-0	5 (4.47, 0.91)	5 (4.58, 0.89)	0.71 (0.86, 0.64)
HK-f2	**3 (3.14, 1.38)**	2 (2.77, 1.56)	**0.94 (1.07, 0.67)**
Objs-sp	2 (2.82, 1.59)	**4 (3.44, 1.63)**	0.26 (0.34, 0.38)
ML-C-1K-f2	3 (2.95, 1.40)	2 (2.84, 1.60)	0.54 (0.66, 0.53)
ML-C-2K-f2	3 (2.94, 1.38)	2.5 (2.86, 1.60)	0.56 (0.64, 0.44)

captions is very subjective. The inter-rater agreements per each model are listed in brackets in Table 4.9.

The human evaluation results are presented alongside the CIDEr metric for each keyword set in Table 4.10. The human ratings show consistent findings with those found from the automatic metrics. This is also evident from the comparison between the human ratings and the CIDEr score. Captions which were generated based on human extracted keywords (i.e., HK-f2) were rated with better accuracy ($\mu = 3.14$) when compared to captions based on predicted keywords. This is not surprising as these captions are composed of keywords extracted from human captions and therefore benefit from high accuracy. However, it was found that HK-f2 ranked last in terms of fluency ($\mu = 2.77$), while Objs+sp was found to be the most fluent ($\mu = 3.44$) most probably due to its simple and generic keyword set. It was also confirmed that ML-C-1K-f2 and ML-C-2K-f2 generate similar quality captions both in terms of accuracy ($\mu = \{2.95, 2.94\}$) and fluency ($\mu = \{2.84, 2.86\}$) and both are not far off from the overall accuracy of HK-f2 ($\mu = 3.14$). In contrast to the CIDEr metric, the human evaluated ratings recorded higher standard deviations especially in the rated fluency.

To study the level of statistical significance between the evaluated models, a one-

way ANOVA (Girden, 1992) on the 550 ratings (for both accuracy and fluency) was conducted. This analysis confirmed that there is a statistical significant difference between the evaluated captions (accuracy: $F(5, 544) = 141.32, p < 0.001$; fluency: $F(5, 544) = 146.55.06, p < 0.001$). Due to this statistical significance, the post hoc Tukey's Honest Significant Difference (HSD) test (Tukey, 1949), was used to calculate the level of significance between each pairwise combination. As reported in Table 4.11, when considering the human rated accuracy and taking $p < 0.05$, a significant difference was noted between HC-0 and the remaining sets, and between HK-f2 and Objs-sp. On the other hand, when considering the rated fluency, it was observed that there was no statistical significance between HK-f2 and ML-C-1K-f2, HK-f2 and ML-C-2K-f2, and between ML-C-1K-f2 and ML-C-2K-f2. In contrast, as tabulated in Table 4.11, all caption sets were found to differ significantly when being evaluated on CIDEr score, except ML-C-1K-f2 and ML-C-2K-f2.

Given this inconsistency between the analysed metrics, a correlation analysis between the human assigned ratings and CIDEr was performed. Pearson's correlation coefficients were computed over the human evaluated captions. The Person's correlation (r) between the human rated accuracy and fluency was 0.61 ($p < 0.0001$). This confirms that there is moderate positive correlation between the two human ratings. In contrast, the Pearson's correlation between the human rated accuracy and CIDEr was 0.20 ($p < 0.0001$) (weak correlation), and between fluency and CIDEr was 0.06 ($p < 0.05$) (negligible correlation).

It was found that captions rated with high accuracy and fluency, can score low in the automatic metrics. This is particularly shown in the captions generated by ML-C-1K-f2 in Fig. 4.8 . For example, although the caption produced by ML-C-1K-f2 describes correctly the content of Fig. 4.8(a), it was rated with a low CIDEr score of 14.82. On the other hand, although the caption of ML-C-1K-f2 describes accurately Fig. 4.8(b), it did not benefit from high scores. This is particularly noted in CIDEr score, which is close to that of the hallucinated caption generated by ML-C-2K-f2. Conversely, it was found that, in some occasions, the human ratings were overly generous when assessing captions that lack details. For example, the one-word captions generated by Obj-sp in Fig 4.8(a) and (b) were rated high in both accuracy and fluency, despite their lack of quality. This analysis, not only revealed the lack of correlation there is between automatic evaluation metrics and human ratings, but it also confirmed the difficulty in obtaining reliable human evaluation assessments.

a (id:26222)

HC-0: A group of people riding skis on a snowy slope.
Accuracy: 5, **Fluency**: 5

HK-f2: on, snowy, skiers, riding
Caption: People riding skis down a.
Accuracy: 4, **Fluency**: 2
B-1: 54.88, **B-2**: 38.80, **B-3**: 30.20, **B-4**: 0, **M**: 25.06, **R-L**: 59.08, **C**: 107.22 **S**: 23.53

Obj-sp: person
Caption: Person.
Accuracy: 5, **Fluency**: 5
B-1: 0, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 1.75, **R-L**: 0, **C**: 0, **S**: 0

ML-C-1K-f2: snow, people, cross, country, ski, mountain.
Caption: **People on a snow covered mountain at a cross country ski in the snow.**
Accuracy: 5, **Fluency:** 5
B-1: 50, **B-2**: 27.74, **B-3**: 0, **B-4**: 0, **M**: 16.40, **R-L**: 28.68, **C**: 14.82, **S**: 32

ML-C-2K-f2: mountain, snow, skier, people, skiing, cross, snowy, country
Caption: People cross country skiing.
Accuracy: 5, **Fluency**: 5
B-1: 9.19, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 9.67, **R-L**: 15.72, **C**: 7.6, **S**: 21

b (id:5797)

HC-0: A person is standing on a snowy hill overlooking a valley.
Accuracy: 5, **Fluency**: 5

HK-f2: man, mountain, lift, ski
Caption: Man standing on a snowy mountain with a person that is in a ski lift.
Accuracy: 5, **Fluency**: 5
B-1: 87.67, **B-2**: 70.37, **B-3**: 53.41, **B-4**: 39.92, **M**: 36.03, **R-L**: 46, **C**: 115.14, **S**: 35.71

Objs-sp: person
Caption: Person.
Accuracy: 4, **Fluency**: 4
B-1: 0, **B-2**: 0, **B-3**: 0, **B-4**: 0, **M**: 6.56, **R-L**: 17.48, **C**: 8, **S**: 10

ML-C-1K-f2: lift, ski, snow, man, stand, person, slope, skier, mountain
Caption: **Man on a snow covered ski slope.**
Accuracy: 5, **Fluency**: 5
B-1: 61.92, **B-2**: 29.91 **B-3**: 0, **B-4**: 0, **M**: 21.47, **R-L**: 52.70, **C**: 40.76, **S**: 16

ML-C-2K-f2: snow, mountain, person, motorcycle, man
Caption: Man on a motorcycle down a snow covered mountain.
Accuracy: 1, **Fluency**: 5
B-1: 55.56, **B-2**: 26.35 **B-3**: 0, **B-4**: 0, **M**: 21.37, **R-L**: 41.71, **C**: 33.91, **S**: 16

Figure 4.8: Human evaluated captions for two testing images with their corresponding evaluation metrics.

4.7 Summary

Both quantitative and qualitative analysis confirmed the efficacy of KENGIC for image caption generation. Good quality benchmark results were obtained when using human extracted nouns from single ground-truth captions as can be found in Figure 4.3. However, generally it was found that the overall performance degrades as prepositions and verbs were added. Presumably, the performance was reduced as more constraints extracted from single captions restricted the search while traversing the constructed graphs. This assumption was confirmed when using the extracted human salient keywords. When having generic and frequently used keywords, the proposed model was found robust in combining salient keywords. This led to the generation of high-quality captions. This chapter also presented a human evaluation study on the quality of the generated captions. Despite the low intra- and inter-rater agreement, the human analysis confirmed that captions generated based on human extracted keywords were marginally more accurate than those based on predicted keywords. On the other hand, the captions based solely on object labels were rated with the highest fluency.

Table 4.11: Tukey's HSD ($p < 0.05$) for pairwise comparison between the human evaluated accuracy and fluency, and CIDEr score based on 550 captions.

		Mean Difference			p-Adjusted			Reject		
Model 1	Model 2	Accuracy	Fluency	CIDEr	Accuracy	Fluency	CIDEr	Accuracy	Fluency	CIDEr
HC-0	HK-f2	-1.34	-1.81	0.21	0	0	0	T	T	T
HC-0	ML-C-1K-f2	-1.52	-1.75	-0.20	0	0	0	T	T	T
HC-0	ML-C-2K-f2	-1.53	-1.73	-0.22	0	0	0	T	T	T
HC-0	Objs-sp	-1.65	-1.44	-0.52	0	0	0	T	T	T
HK-f2	ML-C-1K-f2	-0.19	0.06	-0.41	0.15	0.96	0	**F**	**F**	T
HK-f2	ML-C-2K-f2	-0.19	0.08	-0.43	0.13	0.89	0	**F**	**F**	T
HK-f2	Objs-sp	-0.32	0.66	-0.73	0	0	0	T	T	T
ML-C-1K-f2	ML-C-2K-f2	-0.01	0.02	-0.02	1	1	0.98	**F**	**F**	**F**
ML-C-1K-f2	Objs-sp	-0.13	0.60	-0.32	0.501	0	0	T	**F**	T
ML-C-2K-f2	Objs-sp	-0.12	0.58	-0.31	0.57	0	0	T	**F**	**F**

5 Spatial Relation Detection in KENGIC

As discussed in the previous chapter (refer to 4.3), KENGIC made use of prepositions implicitly and without being able to ground such prepositions in images. Apart from the limited use of spatial relations in the text corpus[1], learning to infer spatial relations from images needs to consider the corresponding trajector and landmark objects. Therefore, predicting spatial relations between two objects directly from images in a multi-label way is difficult. To better handle the use of spatial relations in the generated captions, KENGIC was extended by a post-processing module designed to validate any spatial prepositions used in the generated captions. Captions were first processed to extract triplets in the form of (trajector, relation, landmark). These objects were then grounded in images by localising the objects in bounding boxes using the off-the-shelf Faster R-CNN (Ren et al., 2015) object detector. According to the literature, the Random Forest (RF) classifier was found to be the best single label model in spatial relation detection (Muscat and Belz, 2017b). Therefore, a Random Forest (RF) classifier was trained to predict prepositions between objects using corresponding labels and geometric features extracted from the two detected bounding boxes. Relations found in the extracted triplets which did not match with the detected prepositions were replaced by the output of the implemented spatial relation detector.

5.1 Spatial Role Labelling

To validate and correct the used spatial prepositions in the generated captions, triplets in the form of (trajector, relation, landmark) were extracted. This task, which is commonly referred to as Spatial Role Labelling (SpRL), is a sub-task in NLP which aims to extract spatial relations from natural language (Kordjamshidi et al., 2011). Based on the assumption

[1]The overall percentage of prepositions used in the optimal human extracted keywords (i.e., HK-f2) was 1.58.

that the generated captions are not produced with complex dependencies, spatial triplets were extracted by a simple developed SpRL tool which assumes captions with ordered triplets and prepositions that are always found between two objects. For instance, for a caption like “a person is on a boat that is parked in a field”, the following triplets are extracted: (“person”, “on”, “boat”) and (“boat”, “in”, “field”). This process was evaluated on ViSen Prepositions (Ramisa et al., 2015) and SpRL-2013 (Kolomiyets et al., 2013) datasets, and by human evaluation.

5.1.1 Visen Prepositions Dataset

The ViSen Prepositions dataset (Ramisa et al., 2015) was specifically proposed for spatial relation detection. It consists of triplets extracted from COCO and Flickr30k captions together with their object bounding boxes. The COCO subset which is split into training (8029) and test (3431) was used in this study since it is compatible with the previous work. This dataset contains triplets with both the original labels as found in COCO captions as well as their highlevel object labels. Spatial triplets were automatically extracted using the transition-based dependency parser of Chen and Manning (2014) as implemented in Stanford CoreNLP (Manning et al., 2014). Dependencies reflecting prepositional dependencies were extracted given both the governor and the dependent entities overlap with the objects mentioned in the descriptions and given both objects have corresponding bounding boxes in COCO dataset. When evaluating the extraction of triplets from the ViSen test dataset having original object labels, the calculated precision and recall were equal to 28% and 49% respectively, which resulted in an F-score of 36%.

5.1.2 SpRL-2013 Dataset

The SpRL-2013 dataset (Kolomiyets et al., 2013) is a dataset which is specifically used in spatial role labelling. This XML based dataset contains text with annotated tags which locate trajectors, landmarks and spatial prepositions in the text. Compared to the image captions featured in the ViSen dataset, the sentences found in this dataset are more complex in terms of dependencies between their mentioned objects. The text was parsed and triplets were extracted according to the XML tags. This resulted in a total of 876 triplets. As expected, when compared to the evaluation conducted on Visen Dataset, the precision and recall dropped to 7% and 22% respectively, while the F-score decreased from 36% to 11%.

5.1.3 Human Evaluation

To further assess the quality of the developed spatial role labelling approach, four human evaluators were recruited to evaluate the extracted spatial triplets. The evaluators were given 50 randomly sampled captions with corresponding extracted spatial triplets from the ViSen test split. The evaluators were asked to specify whether each extracted triplet complies with the caption (i.e., the evaluators had to ask the question of whether the given trajector *t* is related to the landmark *l* with the spatial relation *r* for the given caption). A value of 1 had to be given for correct triplets, whilst incorrect triplets had to be assigned a value of 0. The given captions were distributed into: 70% (35) unique for each evaluator, 20% (10) were evaluated by each evaluator, and 10% (5) of the captions were duplicates for each evaluator. This was intended to calculate both inter- and intra-rater agreements. The mean precision was equal to 52% with a 95% intra-agreement and 85% inter-rater agreement. The corresponding Cohen's kappa coefficients were 0.89 and 0.71 for the intra- and inter-rater agreements respectively. With such level of agreement, this evaluation confirmed once again that the spatial role labelling task is a complex task which requires a much more sophisticated approach.

5.2 Detection

After the extraction of spatial triplets, object labels were grounded in images. To maximise the number of detectable objects, a Faster R-CNN trained on Open Images V4 (Kuznetsova et al., 2020) with a pre-trained Inception Resnet V2 (Szegedy et al., 2017) was used. In cases when multiple objects were present for a given extracted object label, the closest pair in terms of the distance between the two bounding box edges was considered. In contrast to the 80 detectable objects of the COCO dataset, Open Images has a total of 601 detectable entities. Out of these, 404 (67%) labels are found in COCO vocabulary. On the other hand, the ViSen COCO split (with original labels) was used to train the spatial relation detector. This dataset was chosen primarily because it is based on COCO dataset and owing to its diversified object vocabulary set. The dataset has a total of 328 distinct objects which are all part of the Open Images detectable entities. To reduce the ambiguity in the training data, (a) synonym prepositions were mapped according to Cambridge dictionary[2] as follows: {"by": "near", "beside":"next to", "inside of": "inside", "onto": "on", "underneath": "under", "beneath": "under", "outside": "outside of", "out of":"outside of"}, (b) non spatial relations including: "with", "of", "at", "as", "for", "from", "to", "about", "off", "past", "across", "down", and "before" were eliminated, and (c) prepositions which occur

[2]https://dictionary.cambridge.org/

Table 5.1: Distribution of spatial relations found in ViSen (COCO split) after pre-processing.

#	Spatial Relation	Count	Ratio (%)
1	on	4333	70.49
2	in	1062	17.28
3	near	335	5.45
5	next to	146	2.38
5	under	117	1.90
6	over	39	0.63
7	behind	36	0.59
8	inside	27	0.44
9	above	17	0.28
10	in front of	15	0.24
11	between	11	0.18
12	around	9	0.15

less than five times were not considered. This resulted in a total of 12 spatial relations distributed across 6147 instances and 259 object labels, as tabulated in Table 5.1. As can be noted from the table, the distribution of the dataset is highly skewed with prepositions "on" and "in" dominating the entire dataset. The dataset was split into development (80%) and test (20%) sets. The former split was further divided with the same ratio into train and validation splits. Splitting was performed through stratified sampling to keep consistent distributions between each split.

5.3 Features

The spatial triplets were represented with both linguistic and geometric features extracted from corresponding bounding boxes. Labels of the two objects were represented using the pre-trained 50-dimensional global vectors (GloVe) (Pennington et al., 2014) (F_0, F_1), while the spatial orientation between the two objects was encoded by a set of 13 geometric features (refer to Fig. 5.1, as proposed in Muscat and Belz (2017b), as follows:

- $F_{\{2,3\}}$: The areas of the two bounding boxes enclosing objects $obj_{\{0,1\}}$ and normalised by the total area of the image.
- F_4: Ratio of obj_0 area with respect to the area of object obj_1.
- F_5: Euclidean distance as computed between the two centroids of the bounding boxes normalised by the image diagonal.
- F_6: The overlapping area over the two bounding boxes normalised by the area of the smaller bounding box.

- F_7: Euclidean distance between the two centroids divided by half the sum of square root of the two areas (an approximate average width/height of the two bounding boxes). This feature accounts for the space between bounding boxes.
- F_8: The cardinal position of obj_0 in relation to obj_1 dependent on the angle between the two centroids which is represented by one of the four cardinal directions: north, south, east and west.
- F_{9-12}: Given the distance of the left margin between the image and obj_0's left edge is a_0 and to the right edge is b_0, and for obj_1 same measures are represented by a_1 and b_1 respectively, $F_9 = (a_1 - a_0)/(b_0 - a_0)$; $F_{10} = (b_1 - a_0)/(b_0 - a_0)$. Similarly, F_{11} and F_{12} are computed with respect to the image's bottom edge and the bounding boxes' horizontal edges respectively. F_{9-12} provide information on (a) the amount of bounding box overlap in x and y directions, (b) extent of free space in between the trajector and landmark, and (c) whether the trajector is to the right, left, top or bottom of the landmark. As such they are correlated with some other features. In addition, an earlier study of Muscat and Belz (2017b) shows that F_{12} partly acts as a proxy to depth and is useful in predicting *behind/in front of* relations.
- $F_{\{13,14\}}$: Aspect ratio of width to height of each bounding box, obj_0 and obj_1.

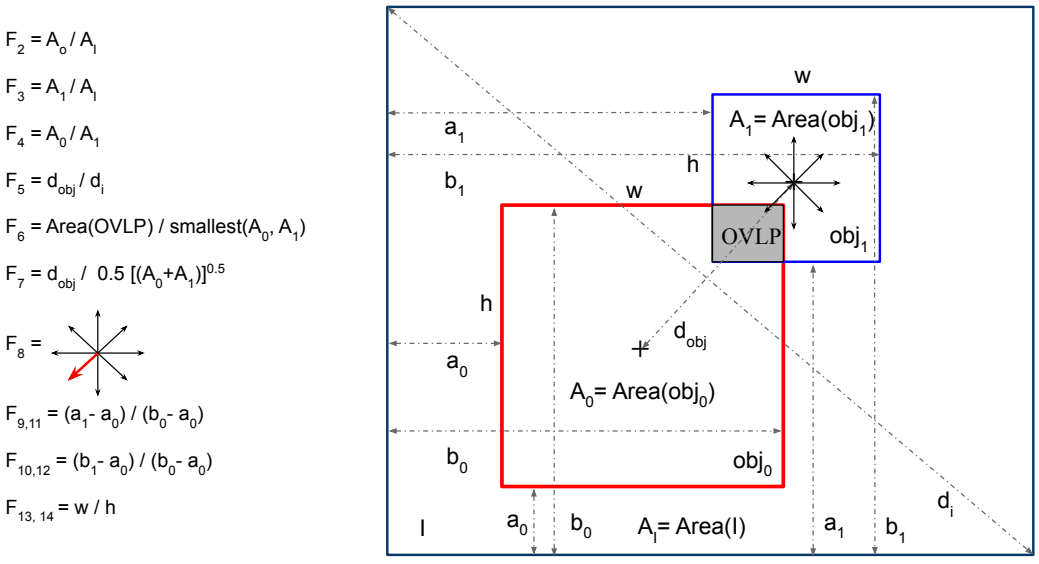


Figure 5.1: Geometric features as proposed in Muscat and Belz (2017a).

Table 5.2: Spatial relation prediction results metrics on Visen (COCO) split.

Split	Accuracy	Precision	Recall	F-Score
Validation	0.85	0.27	0.26	0.26
Testing	0.83	0.26	0.23	0.24
Testing (bal)	0.64	0.26	0.25	0.25

5.4 Model and Results

In a previous study (Muscat and Belz, 2017b), the RF (Breiman, 2001) classifier which is an ensemble classifier composed of multiple decision trees, was found to be the best performing single-label model for predicting spatial relations based on the features discussed in Section 5.3. Therefore, a RF classifier was optimised on the validation set by varying the number of estimators ($1-200$), maximum tree depth ($1-50$), maximum number of features ($1-113$), minimum number of samples to split ($2-40$), and the minimum number of samples required to output a leaf node ($1-40$) on a logarithmic scale. This resulted in a total of 81000 combinations. The best performing configuration in terms of F-score reached a score of 0.36 and an accuracy of 0.79 with a maximum depth of 9, maximum of 40 features, minimum of 15 samples to split and a minimum of 3 samples to output as leaf nodes. The full results on the validation and test sets are tabulated in Table 5.2. It was found that when training and testing the spatial relation detector on balanced sets no significant difference was noted in the metrics when compared to the non-balanced configuration. A confusion matrix based on the non-balanced configuration is depicted in Fig. 5.2. This analysis confirmed that less popular and more spatially constrained prepositions, such as “above”, ‘around” and “between” were incorrectly classified with generic and near synonym prepositions. For example, instances with the prepositions “above” were classified with “on”, while “around” was confused with “near”. As expected, the spatial relation detector was most notably accurate in predicting the prepositions “in” and “on” given their generic aspect and their popularity in the dataset. On the other hand, despite being infrequent, the preposition “under” was predicted correctly 58% of the time. On the other hand, the detector struggled to find the distinction between “near” and other more spatially constrained relations such as “next to”. This points out the need for (1) more representative and less skewed spatial relation dataset to better understand the distinction between spatial relations, (2) more features to encode the spatial triplets, and (3) multi-label models that can address the polysemous nature of spatial relations.

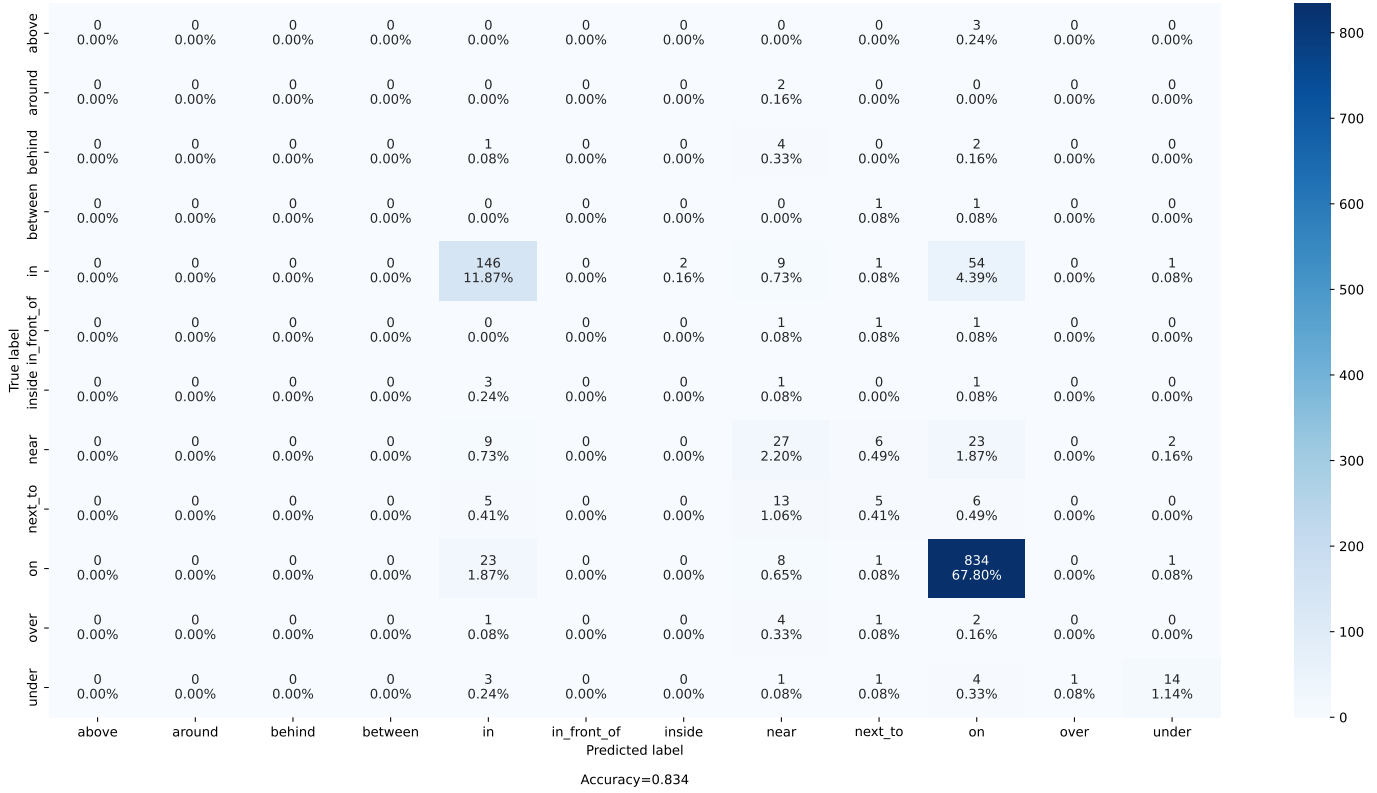


Figure 5.2: Confusion matrix based on the non-balanced test configuration.

5.5 Model Integration

Following the generation of image captions, the spatial relation detection module was used to post-process the captions to validate or correct any used prepositions as discussed in the previous sections. The effect of this module was analysed on the captions which were generated based on ML-C-1K-f2 (since it recorded the highest CIDEr score). The corresponding results metrics are reported in Table 5.3. The results are organised in two groups. The first group reports only the metrics for the captions which were changed after the post-processing, while the second group reports the overall computed metrics on the full testing set with the changed captions. Captions were modified using the spatial relation (SR) detection model as described in Section 5.4 when trained on the full and balanced training data. Metrics computed on the pre-processed captions and percentage differences are tabulated for comparison purposes. When using the SR detection module trained on the full training set, 131 out of the 5000 test captions were modified, while the balanced version altered 145 captions. Overall, it was found that with the introduction of the SR modification, caption quality decreased marginally across all metrics. The largest decrease was noted when using the balanced SR model on BLEU and CIDEr scores. For instance, BLEU-4 decreased from 20 to 16.7 (i.e., $\text{-}16.5\%$) and CIDEr dropped from 81.8 to 76.2 (i.e., $\text{-}8\%$). Metrics were less decreased when using the SR model trained on the full training set (non-balanced). Given the small percentage of captions that were modified (2.62% by the non-balanced SR model and 2.9% by the balanced model), a slight decrease was noted in the overall metrics, while no changes were observed in METEOR and ROUGE-L when using the non-balanced SR model. This confirmed that the developed SR model was not effective for enhancing the quality of the generated captions. To understand the main cause of this outcome, a qualitative analysis was conducted. This analysis was performed on 50 random sampled captions which were modified using the non-balanced SR model. As illustrated in Fig. 5.3, all the modified captions scored lower values in all metrics, except in Fig. 5.3(d), where ROUGE-L increased from 50.41 to 57.55, despite introducing correct spatial relations as shown in Fig. 5.3(d-f). This confirmed the fact that captions with spatial relations that are not grounded in images can score better metrics than captions which are grounded. This was also observed in SPICE score which takes into consideration the relations between entities as shown in Fig. 5.3(e) where SPICE dropped from 46.67 to 40 after changing the relation (man, in, dog) to (man, near, dog). A possible reason why this is happening is because the swapped prepositions are commonly used in the ground-truth captions but used in different contexts. For example, when modifying the caption from a “cat in a chair” to “cat on a chair” in Fig. 5.3(d), the metrics got lower (except ROUGE-L), despite the fact that both prepositions can be used.

In this case, this happened as the preposition "in" was frequently used with "sleeping/laying in" in the five ground-truth captions. The five corresponding human captions are as follows: (1) "A calico kitty sleeping *in* an orange chair", (2) "A calico cat sleeps on a red desk chair", (3) "A fluffy cat laying[3] *in* an orange chair", (4) "A calico cat sleeping on an orange office chair", and (5) "A cat laying[3] down and resting *in* a chair on a hardwood floor." On the other hand, from this analysis it was found that the SR module in some cases modified non-spatial prepositions with incorrect spatial relations, as shown in Fig. 5.3(a,c). The incorrect prediction of prepositions shows the limitations of the features used and the dataset. For instance, the relationship between the "man" and "dog" as found in Fig. 5.3(c) could possibly be resolved better with the introduction of depth features. In that case, depth estimates of objects can give higher weight to the preposition "behind" instead of "on" when bounding boxes overlap with each other. Furthermore, this analysis showed that multiple prepositions can be applicable for a given context as illustrated in Fig. 5.3(b). Although the preposition "near" ranked second with a probability of 0.3 it was replaced by the preposition "behind" which ranked first with a probability of 0.31. In such case, linking the "zebra" with "tree" using both "near" and "behind" would enhance the precision of the caption. This analysis therefore confirms that more accurate captions can be generated by using multi spatial relation detection and with the introduction of added features, such as depth estimates. A study which focuses solely on multi spatial relation detection in images using depth estimates is presented in the following Chapter.

[3] "laying" was used incorrectly instead of "lying" in the human caption.

Table 5.3: Evaluation metrics computed on the testing set for the changed captions and for all captions in COCO dataset. Results for the original generated captions (pre) are included for comparison together with their corresponding percentage difference. The results for the changed captions are presented based on an SR model which was trained on both balanced (bal) and non-balanced training sets.

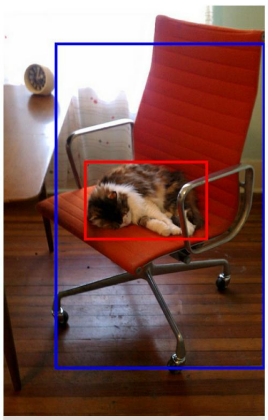
	Num. of captions	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
changed-pre (bal)	131 (145)	69.2 (69)	49.7 (49.9)	31.8 (32.5)	18.7 (20)	26.8 (26.8)	43.4 (44)	81.8 (82.5)	23.0 (23.2)
changed-post (bal)	131 (145)	66.7 (64.6)	46.7 (45.1)	28.1 (27.9)	15.8 (16.7)	26.3 (26)	42.5 (42.2)	76.2 (75.9)	22.7 (22.7)
% difference	-	-3.6 (-6.4)	-6 (-9.6)	-11.6 (-14.2)	-15.5 (-16.5)	-1.9 (-3)	-2.1 (-4.1)	-6.8 (-8)	-1.3 (-2.2)
all-pre (bal)	5000	63.6 (63.6)	44.7 (44.7)	28.9 (28.9)	18.0 (18.0)	22.0 (22.0)	40.3 (40.3)	69.8 (69.8)	18.3 (18.3)
all-post (bal)	5000	63.5 (63.1)	44.6 (44.2)	28.8 (28.5)	17.9 (17.7)	22.0 (21.9)	40.3 (40.2)	69.6 (69.0)	18.3 (18.0)
% difference	-	-0.2 (-0.8)	-0.2 (-1.1)	-0.3 (-1.4)	-0.6 (-1.7)	0 (-0.5)	0 (-0.2)	-0.3 (-1.1)	0 (-1.6)

a (id:14224)

HK-C-1K-f2: brick, building, window, large, clock
Caption: Large clock **on** a brick building **with** a window.

SR triplets: (clock, **on**, brick), (building, **with**, window)
SR detection: (building, {**on** (0.97), in (0.01), next to (0.01)}, window)
Caption+SR: Large clock on a brick building **on** a window.

Pre (Post): B-1: 89.48 (79.57), **B-2**: 83.70 (73.06), **B-3**: 71.02 (58.94), **B-4**: 57.18 (49.71), **M**: 37.10 (37.10), **R-L**: 64.99 (64.99), **C**: 226.70 (213.43), **S**: 27.78 (27.78)

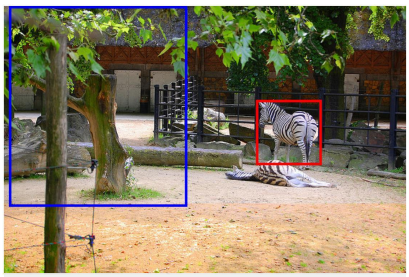


d (id:7560)

HK-C-1K-f2: chair, cat
Caption: Cat in a chair.

SR triplets: (cat, **in**, chair)
SR detection: (cat, {**on** (0.74), in (0.19), near (0.03)}, chair)
Caption+SR: Cat **on** a chair.

Pre (Post): B-1: 36.79 (36.79), **B-2**: 30.04 (30.04), **B-3**: 25.51 (0), **B-4**: 0 (0), **M**: 19.39 (18.21), **R-L**: 50.41 (**57.55**), **C**: 100.74 (97.41), **S**: 19.05 (19.05)



b (id:28174)

HK-C-1K-f2: zebra, stand, two, tree, zoo
Caption: Two zebra stand **near** a tree in a zoo.

SR triplets: (zebra, **near**, tree), (tree, **in**, zoo)
SR detection: (zebra, {**behind** (0.31), near (0.3), in (0.27)}, tree)
Caption+SR: Two zebra stand behind a tree in a zoo.

Pre (Post): B-1: 66.67 (55.56), **B-2**: 50 (37.27), **B-3** 32.93 (27.07), **B-4**: 0 (0), **M**:19.67 (18.11), **R-L**: 37.14 (37.14), **C**: 94.76 (90.84), **S**:16 (16)

e (id:21762)

HK-C-1K-f2: man, dog, tree, christmas, hat, small
Caption: Man in a small dog wearing a hat and a christmas tree.

SR triplets: (man, **in**, dog), (hat, wearing, christmas)
SR detection: (man, {**near** (0.31), on (0.3), in (0.18)},dog)
Caption+SR: Man **near** a small dog wearing a hat and a christmas tree.

Pre (Post): B-1: 100 (91.67), **B-2**: 90.45 (76.38) , **B-3**: 68.91 (55.93), **B-4**: 0.01 (0.01), **M**: 31.97 (28.36), **R-L**: 53.51 (46.82), **C**: 174.45 (156.88), **S**: 46.67 (40)

c (id:32013)

HK-C-1K-f2: man, dog, woman, next, sit, grass
Caption: Man **with** a dog on a grass and a woman sit next.
SR triplets: (man, with, dog), (man, on, grass), (man, on, woman), (man, on, sit)
SR detection: (man, {**on** (0.47), near (0.31), next to (0.1)}, tree)
Caption+SR: Man on a dog on a grass and a woman sit next.

Pre (Post): B-1: 83.33 (75), **B-2**: 55.05 (45.23), **B-3**: 39.28 (27.35), **B-4**: 0.01 (0), **M**: 37.05 (33.38), **R-L**: 43.26 (43.26), **C**: 88.16 (76.59), **S**: 38.46 (30.77)

f (id:15066)

HK-C-1K-f2: boy, ball, young, field, soccer, play
Caption: Young boy in a field play soccer **on** a ball.

SR triplets: (boy, in, field), (boy, on, play), (boy, on soccer), (boy, on, ball)
SR detection: (boy, {**near** (0.52), in (0.33), next to (0.06)}, ball)
Caption+SR: Young boy in a field play soccer **near** a ball.

Pre (Post): B-1: 100 (90), **B-2**: 57.74 (54.77), **B-3**: 0 (0), **B-4**: 0 (0), **M**: 31.49 (31.10), **R-L**: 50 (50), **C**: 116.84 (114.17), **S**: 24.24 (24.24)

Figure 5.3: A sample of captions which were modified based on SR detection together with their corresponding extracted SR triplets and detections. The top three prepositions for each detection are listed with their probabilities in brackets. Evaluation metrics for pre/post processed captions are also listed.

6 Multi Spatial Relation Detection

This chapter presents content published in the following journal article:

Birmingham, B. and Muscat, A. Multi spatial relation detection in images. *Spatial Cognition & Computation*, 1-35, 2021.

6.1 Overview

Detecting the spatial relationship between objects plays a very important role in vision and language understanding. In fact, it was evident that captions generated by KENGIC can benefit from the explicit use of spatial relation detection. However, as shown in the qualitative analyses of the previous chapter, prediction of spatial relations is not straightforward when considering the multi-label nature and ambiguity of the problem. For this reason, this chapter describes research carried out in Birmingham and Muscat (2022) to study the spatial relation prediction from a multi-label perspective. This was specifically intended to provide insights on how prediction of prepositions can be improved in image caption generation. Several models, including a k-Nearest Neighbour multi-label model which was also published in:

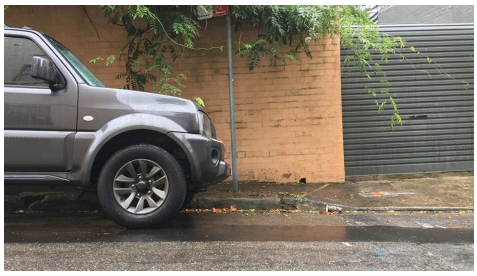
Birmingham, B. and Muscat A. Clustering-based model for predicting multi-spatial relations in images. In *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics, ICINCO2019 - Volume 2, Prague, Czech Republic*, pages 147-156, 2019,

were used and developed to investigate the problem of multi-spatial relation detection based on (a) label embeddings, (b) geometric features extracted from object bounding boxes, and (c) depth features. The addition of 'depth' was originally proposed and studied in single-label spatial relation detection in the following publication:

Birmingham, B., Muscat, A., and Belz, A. Adding the third dimension to spatial relation detection in 2D images. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands*, pages 146-151, 2018.

6.2 Motivation

Spatial language plays an important role in instruction based natural language communication, for example when instructing a robot in a household environment to accomplish a specific task (e.g., "put the clothes *in* the laundry bin", "stack the books *on top of* the desk *in the right* corner", etc.) (Fasola and Matarić, 2012). The use of spatial language is also important when interacting with autonomous vehicles, especially when instructing them to perform a particular task, for example to park either *in front of* or *inside* a garage. It is also useful when self-driving cars are designed to provide textual explanations for their actions (e.g., "the car is going to move to *the right* lane because the car *in front* is slowing down") (Kim et al., 2018), and also when their control systems can accept short-term human textual advice to influence their vehicle controller during navigation (e.g., "if you see a child *on* the sidewalk, slow down.") (Kim et al., 2019).



(a) Instructing an autonomous vehicle to park "behind" and "close to" the street post (Chesterton, 2017).

(b) A visually impaired person crossing a busy road while two vehicles are coming "in front of" and "close" to him from his "left" direction (Uwamariya, 2019).

Figure 6.1: The use of multiple prepositions in two different scenarios.

Assistive technology for the visually impaired needs to be highly verbose and descriptive to portray their visual surrounding with audible natural language descriptions, especially in circumstances where a good level of spatial knowledge and cognition is required. One such example is in cases where navigational assistance is required to help the visually impaired navigate from one place to another. In this context, the use of spatial relations is essential to describe how the objects are related to each other (allocentric), as well as to how the same person is spatially related to the surrounding objects (egocentric) (Ball

et al., 2009). To accurately describe the relationship between objects and to simultaneously construct a detailed cognitive map (Tversky, 1993) for the visually impaired and for autonomous robots (Thrun, 2008), the use of multiple spatial relations is crucial to help them both understand and navigate in unfamiliar environments with a high degree of coordination. As an example, if an autonomous vehicle needs to be parked as shown in Fig. 6.1(a), prepositions "behind" and "close to" should be equally understood and interpreted by its navigational system. Also, as shown in Fig. 6.1(b), an assistive technology in the depicted scenario should not only detect that the visually impaired person is going to cross a busy road, but most importantly, it should alert that person not to cross the road as there are two vehicles coming "in front of" and "close" to him from his "left" direction.

The examples demonstrate the importance of using either a specific preposition or multiple prepositions in contexts that require precision. As discussed in the previous chapter, the explicit use of spatial prepositions is very useful in automatic image captioning, especially in cases involving unusual setups where spatial descriptions are necessary to provide exact and relevant captions. For example, it is unusual to encounter scenarios like boat "in" a tree, or ship "in" a bottle as illustrated in Fig 6.2 in common training image captioning datasets. Therefore, image caption generators should be trained to reason about the spatial setup of the main image entities.

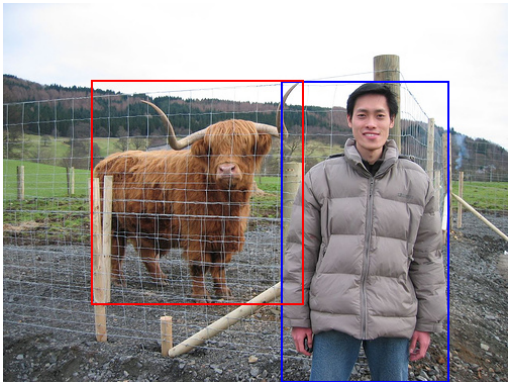
(a) boat "in" tree (Chesterton, 2017).

(b) ship "in" a bottle (Scott, 2013).

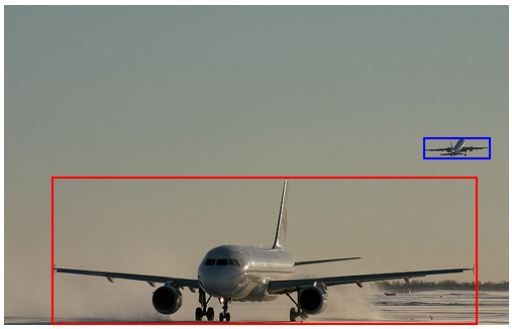
Figure 6.2: Unusual scenarios which require precise spatial descriptions.

Moreover, to effectively describe the visual content where the use of multi spatial relations is useful to provide rich descriptions as shown in Fig 6.1, this chapter addresses the problem of detecting multiple spatial relations using pattern recognition methods. Although the selection of spatial prepositions is typically a straightforward task for human beings, the machine interpretation of spatial relationships is known to be a difficult and complex problem in NLP (Tellex et al., 2020). Detecting spatial relations between objects is difficult for two primary reasons. On one hand, there is the data collection problem since annotators are not always consistent when choosing prepositions (Muscat and Belz,

2017b), possibly because of near-synonym prepositions that are semantically equivalent (e.g., “by” and “close”) and on the other hand spatial relations overlap, sometimes even if they are antonyms, as when the frame of reference changes. Some examples where multi spatial relations are used to describe the spatial relationship between two objects enclosed in bounding boxes (BBs) are given in Figure 6.3.



(a) A person (in blue BB) standing “in front of” and “near” a cow (in red BB) - image 2008_007586.jpg from Pascal VOC 2008 dataset (Everingham et al., 2010).



(b) An aeroplane (in blue BB) flying “above”, “behind”, and “far from” the aeroplane (in red BB) - image 2008_007096.jpg from Pascal VOC 2008 dataset (Everingham et al., 2010).

Figure 6.3: Multi spatial relations in images.

6.3 Problem Definition

The standard way to detect spatial relations is to combine extracted feature vectors with single class labels by learning a function $f : X \rightarrow Y$, where X and Y represent the instances by their corresponding feature vectors, and label spaces respectively. By assuming that each feature vector $x_i \in X$ belongs to one class label $y_i \in Y$, the training feature vectors combined with their corresponding class labels found in set $D = \{(x_i, y_i) \,|\, 1 \leq i \leq m\}$ are used to automatically learn the semantic relationship between each $x_i \in X$ and $y_i \in Y$, $\forall\, m$ instances (Zhang and Zhou, 2014). A typical feature vector x_i includes multiple features $f_i \in x_i$ to describe that given instance x_i. Spatial relations can therefore be represented using a combination of linguistic and geometric features which can include for example the object class labels, the areas of the two bounding boxes normalised by the image size and the distance between the two centroids of the two bounding boxes. Based on these spatial features, the function $f : X \rightarrow Y$ is trained to predict the sets of spatial relations $\{$“in front of” ,“near”$\}$ and $\{$“above” ,“behind”, “far from”$\}$ when given the object labels and the bounding boxes of Figures 6.3(a) and (b) respectively.

To address the multi-label nature of preposition selection, the spatial relation detection can be cast as a multi-label classification problem (Tsoumakas and Katakis, 2007), by assigning a set of appropriate labels (prepositions) for each instance. For a given space $X = \Re^d$ denoting each d-dimensional feature vector per instance, and $Y = \{y_1, y_2, \dots y_q\}$ which represents the label space with q distinct class labels for each and every $x_i \in X$, multi-label learning aims to infer a function $h : X \rightarrow 2^q$ from the multi-label training data D which is represented by (x_i, y_i), where x_i is a d-dimensional feature vector, while $y_i \subseteq Y$ is the corresponding set of associated labels. Finally, to predict the set of class labels $h(x_i) \subseteq Y$ for a given unseen $x_i \in X$, the learned multi-label model $h(\cdot)$ is applied.

The simplest way to handle multi-label classification problems is by decomposing the same problem into several classification problems. This means that one binary classifier is trained for each class label and therefore used to predict whether a given instance is attributed to that label or not. This approach is known as binary relevance (BR) learning. This kind of approach has been criticized for not taking into account the hidden information that can be found in the label space, i.e., information about the interdependencies between the labels. Given that the presence or absence of the different class labels has to be predicted simultaneously, the exploitation of the dependency between classes can be crucial in multi-label learning (Dembczyński et al., 2012). In fact, a good multi-label model internally models the dependency between classes.

In this chapter, several multi-label models are developed. The models are evaluated with automatic metrics, as well as with human evaluations. A quantitative analysis is carried out to compare the performance of the various multi-label models and to assess the collected human evaluations. In addition, to get an insight into the rankings of a single-label classifier, the standard way of predicting prepositions, the multi-label annotations are compared to the output of a single-label Random Forest classifier (RF). Furthermore, the human evaluations, which are independent of the ground truth labels inform on the quality of the original dataset human annotations. Finally, a qualitative analysis is carried out to highlight errors in the predictions and discuss possible causes, informing directions for future work.

6.4 Dataset

The French SpatialVOC2K dataset (Belz et al., 2018) was used in this study. Objects in this dataset are annotated with textual labels and corresponding bounding boxes, while the relationship between objects is encoded as sets of prepositions (multi-label). This dataset was collected by instructing French native speakers to specify the single preposition (in-

putted as a free text entry) that best describes the spatial relationship between a given pair of objects. In a second step, the annotators had to select all relevant prepositions from a list of prepositions, such that each preposition fits in the given context. The labels in this dataset can therefore be considered complete, and hence are suitable to conduct the multi-label experiments. The dataset has a total of 21 unique prepositions which are distributed across a total of 5320 object pair annotations out of which 80 annotations are duplicates. These duplicates, which were used to measure the intra- and inter-annotator agreements in Belz et al. (2018) were then grouped by taking the union of all spatial relations that were selected for a given object pair. This leads to a total of 5240 unique object pair combinations from a total of 20 object categories in 1554 images. The dataset has an average label cardinality (the average number of prepositions per object pair) of 2.16 and follows the distribution tabulated in Table 6.1. The total number of prepositions used in the experiments was reduced to 17 after eliminating prepositions: *à côté* ("beside"), *au-dessous de* ("below") and *en travers de* ("across") which were recorded once and replacing *près* ("near"), which was also recorded once, by *près de* ("near")[1]. The dataset was split into development (80%) and test set (20%), while the development set was further sub-divided into training and validation sets in the same ratio. Each dataset split was assured to have a similar class distribution to that tabulated in Table 6.1.

Table 6.1: Distribution of preposition set sizes.

Spatial Relations Set Size	Frequency
1	1117
2	2351
3	1597
4	166
5	8
6	1

6.5 Features

Following previous works (Muscat and Belz, 2017b; Ramisa et al., 2015), this study considered both linguistic and geometric features as input to the models, as in Section 5.3. To address the limitations of this feature set as discussed in Section 5.3, additional ob-

[1]Note that *près de* ("near") was recorded 2856 times.

ject depth estimations were included. For the purpose of the reported experiments, the following six discrete sets of features: (a) Language Label Encodings (LE), (b) Language Indicator Vector (IV), (c) Word Embeddings (GloVe), (d) Word Embeddings (Word2vec), (e) Geometric Features (GF), and (f) Depth Features (DF) are defined. In the experiments, combinations of the above features were used to train the models. For example, {IV, GF} or {GloVe, GF, DF}. The following section details the used depth features.

6.5.1 Depth Features (DF)

This section contains content from the following publication:

Birmingham, B., Muscat, A., and Belz, A. Adding the third dimension to spatial relation detection in 2D images. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands*, pages 146-151, 2018.

6.5.1.1 MonoDepth Features

In order to automatically infer the depth for image objects, *monoDepth*[2] (Godard et al., 2017), which is a CNN-based method trained on stereo image pairs was used. This model maps images to depth maps consisting of pixels having values that represent the estimated distance from the viewer. More specifically, the monodepth-cityscapes model trained on the Cityscapes dataset (Cordts et al., 2016) was used. Figure 6.4 shows an image from the dataset alongside the depth map generated for it by the monodepth-cityscapes model. The more towards the dark blue end of the colour spectrum an area is, the further away it is from the viewer, and the more towards the bright yellow end, the closer. The model produces an impressively accurate rendering of the depths of the two trees, car, person, and road. Once a depth map is generated for a given image, depth values were obtained for the pixel grids in the bounding boxes of the pair of objects under consideration. The following features were computed for each bounding box:

- **Average (AVG) depth**: simply the average depth value within each object bounding box.
- **Radially weighted average (RWA) depth**: starting from the central pixel(s), each pixel is weighted by the inverse proportion to its distance from the centre and a weighted average is computed.

[2] https://github.com/mrharicot/monodepth

Figure 6.4: Example of SpatialVOC2K image and depth map generated by monoDepth.

Looking at the example in Figure 6.4, the average depth value is much less in the red person bounding box (BB) than in the blue car BB, making "*person in front of car*" a possibility. The RWA is also less for the person BB, but the difference is less pronounced than would be the case if all of the car was further away than the person, thus making "*person next to car*" an alternative possibility.

6.5.1.2 Manually Estimated Depth Feature

Human depth estimates of BB level for $1,554$ images and $3,642$ objects were collected as follows. Participants were shown an image with all relevant objects surrounded by BBs. Their task was to assign a number out of 100 to each bounding box, indicating the average depth of (just) the object, where 100 is the maximum distance. The annotators were trained and mentored for some time before starting annotations proper. Three participants in total contributed to the annotations. Depth values were normalised to range from 0 to 1 for each image.

Pearson's correlation coefficients were computed over all object BBs between the human estimated object depths and the corresponding AVG and RWA figures. Pearson's correlation (r) between human and AVG depth values was 0.535 $(p < 0.0001)$, and between human and RWA was 0.523 $(p < 0.0001)$. The correlation between AVG and RWA was 0.995 $(p < 0.0001)$. MonoDepth and human-estimated average object depths were converted to categorical values (foreground, background, neutral), and the percentage agreements were 60.8% for the average depth and 60.3% for RWA. When both automatic and manual depth estimations were used alongside the linguistic and geometric features in the experiments published in Birmingham et al. (2018), it was confirmed that depth enhances the prediction accuracy in all single-label scenarios tested. However, it was confirmed that automatically computed depth is still some way off manually an-

notated toplines, and for this reason only manual depth estimations were used for the multi-spatial relation detection study. Therefore, estimated depth for each object in the interval between 0 (foreground) and 100 (background) was used. For this study, these values were normalised between 0 and 1 and represented by $\{F_d \mid 15 \leq d \leq 16\}$[3]. Furthermore, the depth difference F_{17} between the two objects was used to reflect the depth order between the two objects obj_0 and obj_1.

6.5.2 Preprocessing

Features vary in magnitude (e.g., in the training set, the label encoded object value varies between 0 and 19, while the depth of an object varies between 0 and 1). This impacts the model performance as features with high magnitude can dominate the models during training, as well as slow down the machine learning process. Therefore, as part of the data pre-processing stage, each feature was normalised to zero mean and unit standard deviation.

6.6 Evaluation Metrics

Label cardinality is used to calculate the average number of predicted prepositions per instance and is a global metric useful to determine whether the model is under or over generating output labels (i.e., prepositions). It is compared to the training set's cardinality, which is 2.16. The label cardinality (LCard) for dataset D is computed as follows:

$$LCard(D) = \frac{1}{n}\sum_{i=1}^{n}|Y_i|, \tag{6.1}$$

where Y_i is the set of spatial relations for the i^{th} instance and n is the total number of instances in the pertaining set D.

In addition to the Label Cardinality, example-based metrics including accuracy (Acc) (Refer to Eqn. 3.2), precision (P) (Refer to Eqn. 3.3), recall (R) (Refer to Eqn. 3.4) and F-score (F) (Refer to Eqn. 3.4) (Zhang and Zhou, 2014), were used to evaluate the multi-label prediction problem as described in Section 3.5.4.1.

[3]The full feature set can be found in Section 5.3

6.7 Models

This section describes the various multi-label models developed for predicting multi-spatial relations between pairs of objects which take as input combinations of linguistic, geometric and depth features. The models implemented are the k-Means (kM-C) and the Agglomerative Hierarchical (A-HC) based clustering models, the Nearest Neighbour (NN) model, which is a distance-based classification method, and the Multi-Label Neural network (ML-NN). A Random Forest (RF) single-label classifier was also implemented to compare results to the single-label classification problem. The following sections describe each model and its development.

6.7.1 Nearest Neighbour (NN)

This method uses the normalised training data as discussed in Section 6.5 to predict prepositions by finding the closest instance x_j in set X. As the training data is considered exhaustive, this approach permits the selection of the most similar instance based on Euclidean distance, so that the corresponding spatial relations are taken as the predicted prepositions $h(x_i)$. Formally, the predicted prepositions for the unseen instance x_u are the prepositions belonging to the spatial relation instance x_j such that:

$$x_j = \arg\min_{x_i} |x_u - x_i| \tag{6.2}$$

From the results generated on the validation set, which are presented in Table 6.2, it can be noted that when the spatial relations are only represented by the linguistic features, the highest average accuracy rate obtained is 0.216 when using the GloVe feature vector, even though the label cardinality exceeds the value of 10. The accuracy rate increased to 0.377 when using the geometric features only, and when the geometric features were combined with the GloVe feature vector, the accuracy increased to 0.438. When combining the full feature set (i.e., GloVe+GF+DF), the accuracy increased further to 0.442 and the label cardinality dropped to 2.153.

6.7.2 k-Means Clustering (kM-C)

This subsection presents the k-Means Clustering (kM-C) based method which was published in Birmingham and Muscat (2019). In this model, disjointed clusters composed of features characterised by low intra-variability and high inter-variance in comparison to other cluster members are formed. Unseen instances are then assigned to the closest clusters based on euclidean distance. The probability distribution of the prepositions

Table 6.2: Evaluation metrics computed on the validation set for each feature vector for the NN model.

Features	LCard(Val)	Acc	P	R	F
LE	10.293	0.215	0.219	0.954	0.356
IV	10.335	0.215	0.219	0.957	0.357
W2V	10.317	0.215	0.220	0.956	0.357
GloVe	10.285	0.216	0.220	0.954	0.358
GF	2.163	0.377	0.483	0.485	0.484
GloVe+GF	2.167	0.438	0.560	0.554	0.557
GloVe+GF+DF	**2.153**	**0.442**	**0.566**	**0.558**	**0.562**

linked with the corresponding training instances found within the assigned clusters is computed. The spatial relations that exceed a predefined threshold are then considered as part of the predicted set for a given unseen instance. The approach is designed to group similarly oriented spatial relations based on their linguistic and visual properties. By making use of the *k*-means clustering algorithm (Pedregosa et al., 2011) and without taking into consideration the ground-truth preposition sets, the scaled feature vectors having zero mean and unit variance were grouped into *k* distinct clusters. The probability distribution of prepositions across each cluster was exploited for both the classification of unseen instances as well as for preposition similarity.

6.7.2.1 Model

The developed model is based on k-means clustering which aims to partition the instance space X into k disjointed and non-hierarchical clusters represented by set C (Jain et al., 1999). The method is designed to iteratively assign each $x_i \in X$ into one of the available clusters defined in set C in a 2-stepped approach until a terminating condition is met. Starting from an initial set of k centroids represented by the randomly initialised means $M^{(t)} = \{m_1^{(t)}, m_2^{(t)}, \ldots, m_k^{(t)}\}$ at time-step t, each having a dimension $|x_i|$, the first step requires the assignment of each instance x_i to the closest cluster centroid based on Euclidean distance. This is calculated between $x_i \in X$ and $m_i \in M$, such that each cluster $c^{(t)} \in \{C_c^{(t)} \mid 1 \leq c \leq k\}$ is composed of:

$$\{x_i : ||x_i - m_c^{(t)}||^2 \leq ||x_i - m_j^{(t)}||^2 \ \forall_j, 1 \leq j \leq k\}, \tag{6.3}$$

where each x_i is assigned to only one cluster $c^{(t)}$ irrespective of any instances which might fit in multiple clusters. The algorithm continues by updating each cluster mean found in

set M by:

$$m_c^{(t+1)} = \frac{1}{|c^{(t)}|} \sum_{x_i \in c^{(t)}} x_i. \tag{6.4}$$

These two steps are repeated until either the centroids and instances stabilise (i.e., centroids stop changing their position and instances keep consistent cluster membership), or until a number of iterations are performed. Given the non-deterministic nature of this method and since it does not guarantee a global optimum, initial centroid seeds are initialised via the k-means++ algorithm (Arthur and Vassilvitskii, 2007) to speed up convergence. Furthermore, the method was executed for 1000 consecutive runs and each run was allowed to perform 300 iterations. This was performed to increase the likelihood of finding the centroids that best minimise the within-cluster variance.

Once the set of data points X are clustered into the final k clusters, multi spatial relation detection is implemented by first computing the preposition likelihoods $P(P \mid C)$ for each spatial preposition $p_i \in P$ over each cluster $c_i \in C$. Preposition likelihoods are then normalised with respect to the maximum likelihood found per each cluster c_i, such that the dominant prepositions found in each cluster have a likelihood equal to 1 given that:

$$P(p_i \mid c_j) = \frac{P(p_i \mid c_j)}{\arg\max_{p_i} P(p_i \mid c_j)}. \tag{6.5}$$

6.7.2.2 Classification

The multi spatial relation set for a given unseen object pair represented by x_i is predicted by a two-stepped approach. The first step is to find the closest cluster C_m represented by its mean m that minimises the L^2 norm distance among all cluster means by:

$$m = \arg\min_{m \in M} \{||x_i - m||^2\}. \tag{6.6}$$

The second step is to extract the prepositions belonging to the closest cluster C_m which have a likelihood ratio that exceeds a specified threshold t. Mathematically, the predicted spatial relations $h(x_i)$ are denoted by:

$$h(x_i) = \{p_i : P(p_i \mid C_m) \geq t\}. \tag{6.7}$$

The training phase and the details for optimising the hyper-parameters k and t of the presented model are discussed in subsection 6.7.2.4.

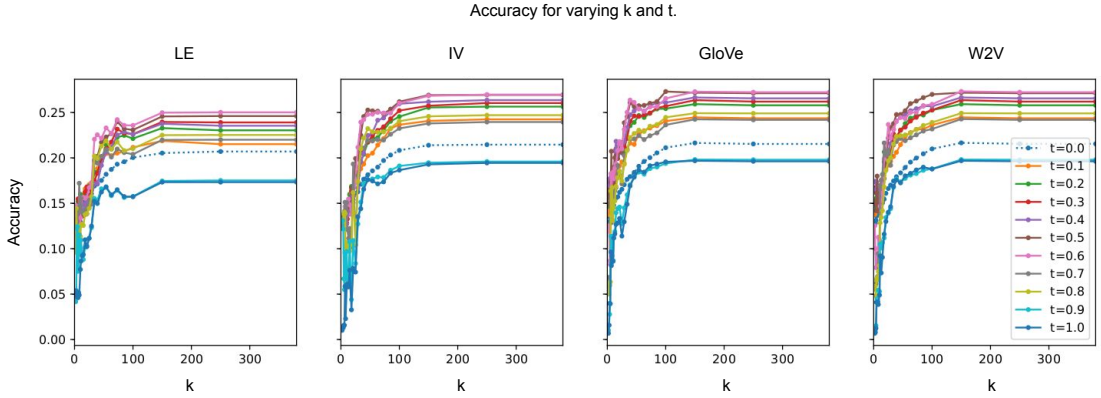


Figure 6.5: Accuracies computed on the validation set for varying clusters (k) and thresholds (t) based on linguistic features including Label Encoding (LE), Indicator Vector (IV), GloVe and Word2Vec (W2V) word embeddings.

6.7.2.3 Distance Metric

To get deeper insights into how prepositions are related to each other, the clustering-based model offers a way to compute the similarity between each preposition $p_i \in P$. By representing how each preposition p_i is clustered through its distribution over each cluster $c_i \in C$, spatial prepositions can be compared via a distribution distance metric. Given that the prepositions $p_{\{i,j\}}$ are represented by the probability distributions $P(C \mid p_{\{i,j\}})$, prepositions were compared via the histogram intersection method which computes the distance metric $d(p_i, p_j)$ as follows:

$$d(p_i, p_j) = \sum_{c_k \in C} min(P(c_k \mid p_i), P(c_k \mid p_j)) \tag{6.8}$$

6.7.2.4 Optimisation

The above metrics were computed under various k and t values to gain insight into how the clustering-based model performs when using both linguistic and visual features. The first experiment was carried out to evaluate the model based solely on linguistic properties. This was intended to identify the language feature set that best represents the object labels whilst also maximising the discussed evaluation metrics. Figure 6.5, shows the accuracies obtained when predicting spatial relations for the instances found in the validation set based on each linguistic feature set. The plots show how the accuracy varies with the different number of clusters (k) and thresholds (t). The accuracy peaks when approaching the 100^{th} cluster for all varied thresholds, and the top two performing

thresholds were 0.5 and 0.6 for each configuration. Furthermore, it was evident that the Indicator Vector (IV) feature set marginally improves on the Label Encoding (LE), while the GloVe and Word2Vec (W2V) slightly outperform the IV. The overall accuracies for each feature set computed across all k and t values (i.e., total of 342 per each feature set) were analysed. Table 6.3 shows that the highest accuracy recorded was 0.273 for both GloVe and W2V embeddings, while the highest accuracy mean (0.195) and median (0.198) were obtained when using GloVe features. For this reason, the GloVe feature set was used for the following experiments in conjunction with both geometric and depth features.

Table 6.3: Overall statistics per each linguistic feature set.

Features	Mean	Median	Min	Max
LE	0.172	0.166	0.042	0.250
IV	0.180	0.180	0.132	0.270
W2V	0.187	0.196	0.153	0.273
GloVe	**0.195**	**0.198**	**0.164**	**0.273**

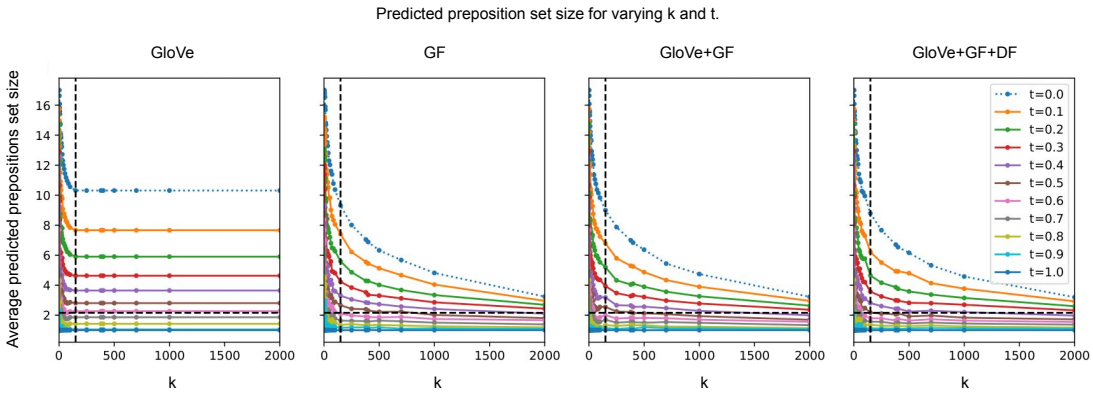


Figure 6.6: Average predicted preposition set sizes generated for the validation set for varying clusters (k) and thresholds (t) based on a combination of linguistic and visual features. The plots show the dataset's average prepositions set size (i.e., 2.16) and the region where cluster stabalise (i.e., $@k = 150$) with the horizontal and vertical dashed lines respectively.

The hyper-parameters k and t were both optimised with respect to the corresponding average predicted preposition set size as obtained on the validation set. As illustrated in Figure 6.6, the model was assessed in terms of how many prepositions are generated for a given unseen instance when represented by a combination of linguistic and visual features. This was performed for varying values of k and t. The plots show that when the model was parameterised with thresholds of 0.5 and 0.6, it gave an average preposition set size that is very comparable to the overall dataset's label cardinality (i.e., 2.16 and which

Table 6.4: Evaluation metrics computed on the validation set for each feature vector.

Features	k,t	LCard(Val)	Acc	P	R	F
GloVe	150, 0.6	2.265	0.273	0.356	0.385	0.343
GF	150, 0.6	2.079	0.211	0.276	0.307	0.269
GloVe+GF	150, 0.6	2.004	0.270	0.367	0.353	0.336
GloVe+GF+DF	**150, 0.5**	**2.167**	**0.283**	**0.370**	**0.388**	**0.353**

is marked by the horizontal dashed line in the respective plots), given that the number of clusters (k) falls within the stable region (i.e., within the elbow curve which is represented by the vertical dashed line in the plots). Therefore, the number of optimal clusters for each configuration was set to 150, while a threshold $t = 0.6$ was used when the model was based on: {*GloVe*, *GF*, *GloVe+GF*} sets, and t was set to 0.5 when the model used the combined feature set composed of: {*Glove+GF+DF*}.

The remaining evaluation metrics associated with the respective chosen hyper-parameters are tabulated in Table 6.4. The table shows that the linguistic features highly influence the spatial relation detection. The accuracy obtained when using linguistic features only was 0.273. The accuracy decreased to 0.211 when spatial relations were predicted based on their geometric features. When both feature sets were combined (i.e, GloVe+GF), the average precision (AP) increased by 3.1%, over that obtained when using GloVe features alone, while the average recall (AR) decreased by 8.3% which resulted in a loss of 1.1% in accuracy. However, when adding the depth features together with the linguistic and geometric properties (i.e., GloVe+GF+DF), the average accuracy (Acc) increased by 3.7% and reached the highest recorded accuracy of 0.283, thus confirming the effectiveness of the added depth features.

The final model was trained on the full development set with $k = 150$ for all feature sets. The likelihood threshold t was set to 0.5 when trained on the complete feature set, otherwise t was set to 0.6. The model was evaluated on the testing set for 50 times to compute the average metrics which are reported in Table 6.7.

6.7.3 Agglomerative Hierarchical Clustering (A-HC)

Hierarchical Clustering is a method which organises the training data into a hierarchy of clusters (Johnson, 1967). The agglomerative hierarchical clustering approach (Day and Edelsbrunner, 1984) is designed to merge instances in a bottom-up approach. The algorithm starts by treating each instance as a singleton cluster. The process continues by merging the two closest clusters iteratively to form one single cluster until all clusters are merged together into one composite cluster containing all instances. Throughout the

whole process, a distance matrix containing all distances between clusters is maintained. After each merge, the distance matrix is updated to reflect the distances between the newly created cluster in relation to all other remaining clusters and instances. Clusters can be merged via different linkage methods which include the Single-link, Average-link, Complete-link, Weighted-link, Centroid-link, Median-link and Ward-link:

In the Single-link (S) approach, the distance between the newly formed cluster composed of u and v is the minimum distance for all points in clusters u and v computed by:

$$d(u,v) = min(dist(u[i], v[j])), \tag{6.9}$$

for all points i and j in clusters u and v respectively.

In Complete-link (C), the distance between two clusters is the largest distance between clusters u and v which is computed as follows:

$$d(u,v) = max(dist(u[i], v[j])), \tag{6.10}$$

for all points i in cluster u and j in cluster v.

The distance in Average-link (A) is specified by the average distance found between the two clusters computed by:

$$\sum_{ij} \frac{d(u[i], v[j])}{|u||v|}, \tag{6.11}$$

for all points i and j where $|u|$ and $|v|$ are the cardinality of clusters u and v respectively.

When using the weighted-linkage (W) method, the distance between the resultant cluster and the other remaining clusters is found by computing the arithmetic mean of the distances between each cluster as follows:

$$d(u,v) = (dist(s,v) + dist(t,v))/2, \tag{6.12}$$

where cluster u is formed with clusters s and t, and v is a remaining cluster in the forest.

The centroid (C) linkage calculates the distance between two clusters by calculating the Euclidean distance between the cluster centroids as follows:

$$dist(s,t) = ||c_s - c_t||_2, \tag{6.13}$$

where c_s and c_t are the centroids of clusters s and t respectively.

Like the centroid linkage, the Median (Med) link takes into account the average of the two cluster centroids as a distance measure.

Finally, the Ward (W) linkage method uses the Ward variance minimization algorithm to calculate the distance between the two centroids u and v by the following equation:

$$dist(u,v) = \sqrt{\frac{|v|+|s|}{T}\,d(v,s)^2 + \frac{|v|+|t|}{T}\,d(v,t)^2 - \frac{|v|}{T}\,d(s,t)^2}, \tag{6.14}$$

where u is the newly joined cluster consisting of clusters s and t, and v is an unused cluster in the forest $T = |v| + |s| + |t|$.

The result of this merging process is a tree-based representation, i.e., a dendrogram of varying height (h) which depends on the instance's scaled vector space, as well as the linkage (l) method that is used. Once the dendrogram is created based on the feature vector of spatial relations, the classification of unseen instances is carried out by first cutting the tree at the most appropriate cut-off point (c). Once the tree is cut, the closest instance based on Euclidean distance is found from the dendrogram. The corresponding cluster is found from the dendrogram so that the probability distribution of the prepositions found in that cluster $P(p_i)$ can be calculated. The prepositions that exceed a specified threshold (th) are considered as the predicted ones.

The experiments carried out on the validation set for this model considered all linkage methods and analysed the predicted accuracy for varying threshold values in the interval between 0 and 1 in steps of 0.1. The model was also evaluated using different feature setups, linkage methods and 10 varying dendrogram cut-off points in relation to each corresponding depth (d). From the results tabulated in Table 6.5, it can be seen that the GloVe feature vector, which obtained an accuracy of 0.216, again outperformed the other linguistic features by 0.5%, given that all other linguistic features obtained an accuracy of 0.215. In this case, when using the geometric features only, the best accuracy was 0.263 and when combined with the GloVe feature vector it improved by 34.2% to an accuracy of 0.353. The highest accuracy achieved on the validation set was 0.377 when using the combination of linguistic, geometric and depth features. In the latter case, the average cardinality was equal to 4.5. All hyper-parameters are tabulated in the same table.

6.7.4 Multi-label Neural Network (ML-NN)

A Multi-label Neural Network (ML-NN) (Bishop, 1995) is a network of inter-connected neurons designed to output the probabilities of multi-labels from an output neuron layer. Specifically, this model is based on an input layer of size equal to the number of features

Table 6.5: Evaluation metrics computed on the validation set for each feature vector after hyper-parameter optimisation for the A-HC model. Headers l, d, c and th are the linkage type, tree depth, tree cut-off point and threshold values respectively.

Features	l	d	c	th	Clusters	LCard(Val)	Acc	P	R	F
LE	S	0.451	0.1	0	156*	10.293	0.215	0.219	0.954	0.356
IV	S	9.664	1	0	156*	10.335	0.215	0.219	0.957	0.357
W2V	S	37.500	3.8	0	156*	10.317	0.215	0.220	0.956	0.357
GloVe	S	13.071	1.4	0	156*	10.285	0.216	0.220	0.954	0.358
GF	A	17.885	1.8	0	629	7.626	0.263	0.281	0.865	0.425
GloVe+GF	A	22.678	2.3	0	1179	5.253	0.353	0.391	0.807	0.526
GloVe+GF+DF	**A**	**22.745**	**2.3**	**0**	**1516**	**4.500**	**0.377**	**0.433**	**0.763**	**0.553**

* All linguistic features were best clustered into singleton clusters where each cluster corresponds to individual object pair. This shows that clustering based solely on linguistic features was ineffective.

plus an additional neuron for the bias, and a set of hidden layers that can be of varying height ($h : 5 - 500$) and depth ($d : 1 - 40$). In this implementation, both the input and hidden layers are composed of ReLU activation units. The model has an output layer which is composed of sigmoid activation units to reflect the probability of each preposition. The number of output neurons is equal to the number of spatial relations which in these experiments was set to 17. The Adam optimisation algorithm (Kingma and Ba, 2015), which is an extension to the stochastic gradient descent (Robbins and Monro, 1951), was used to iteratively update the network weights. The learning phase of the model was left to execute for 1000 epochs; however, early stopping with patience 20 was used to make sure that the model is trained without overfitting, by stopping at the optimal epoch (e) and by using the optimal batch size ($b : 32, 64, 128$). This was carried out by minimising the validation loss on the validation set for 5 consecutive times so that the results are averaged over multiple runs. Furthermore, since the output layer is composed of sigmoid units which output values that vary between 0 and 1, the binary cross entropy loss was used as the loss function to train the network, as shown below:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)), \tag{6.15}$$

where y is whether the corresponding preposition is 1 or 0 when represented by the hot-encoding vector and $p(y)$ is the predicted probability of that instance being 1, for all N points.

The average results generated by ML-NN when computed over 50 repetitions on the

Table 6.6: The average evaluation metrics of the ML-NN when computed on the validation set after hyper-parameter optimisation for each feature vector. The labels d and h correspond to the depth and height of the neural network, while labels b and e refer to the batch size and number of epochs respectively.

Features	d	h	b	e	LCard(Val)	Acc	P	R	F
LE	5	500	128	63	1.149	0.294	0.492	0.318	0.386
GloVe	10	40	128	55	1.242	0.336	0.519	0.365	0.429
IV	10	150	64	26	1.264	0.340	0.518	0.372	0.433
W2V	1	150	64	12	1.479	0.355	0.524	0.404	0.456
GF	5	200	32	13	1.651	0.461	0.626	0.521	0.569
GF+LE	5	500	128	13	1.717	0.479	0.631	0.545	0.585
GF+W2V	5	150	64	16	1.784	0.516	0.677	0.592	0.632
GF+IV	5	500	64	6	1.809	0.525	0.687	0.609	0.645
GF+GloVe	5	200	64	10	1.793	0.531	0.692	0.608	0.647
GF+LE+DF	5	500	64	8	1.832	0.537	0.682	0.616	0.648
GF+W2V+DF	5	300	128	16	1.829	0.564	0.726	0.641	0.681
GF+IV+DF	5	150	64	10	1.874	0.571	0.722	0.655	0.687
GF+GloVe+DF	**5**	**300**	**32**	**8**	**1.853**	**0.573**	**0.732**	**0.653**	**0.690**

validation set are tabulated in Table 6.6. It shows that the W2V linguistic feature set obtained the highest accuracy of 0.355, when compared to the other linguistic features: LE (0.294), GloVe (0.336) and IV (0.340). When using geometric features only, the accuracy increased to 0.461 and when combined to linguistic features, it reached the highest accuracy when paired with the GloVe feature vector (i.e., 0.531). The highest accuracy achieved among all models was that of 0.573 when adding the depth features to linguistic (GloVe) and geometric features. All hyper-parameters are tabulated in the same table. From the table, it can also be seen that, compared to the other models, the multi-label neural network generates fewer prepositions. In fact, in the best configuration setup, the average label cardinality was 1.853 when having a network composed of 5 hidden layers, each with 300 neurons, and trained with a batch size of 32 for 8 epochs.

6.7.5 Single-Label Random Forest (RF) Classifier

To get an insight into how the rankings of a single-label classifier relate to the multi-labels, the results from a single-label Random Forest (RF) classifier, which was found to be the

best classifier in a comparative study (Muscat and Belz, 2017b), are analysed. For the purpose of the analysis, the k top-ranked prepositions generated by the RF model are considered as the set of "multi-labels", and since the dataset's cardinality is 2.16, $k = 2$ and $k = 3$ are expected to yield the relevant multi-label output. In the results section, the results for $k = 1$ to 4 are tabulated to justify this assumption. Since k is a constant, it is expected, that in some cases, the RF will either under- or over-generate.

The Random Forest (Breiman, 2001) is an ensemble classifier composed of multiple decision trees (Quinlan, 1986), where each decision tree is used to fit each training subsample, through replacement, using an averaging mechanism to enhance the predictive accuracy, as well as to avoid over-fitting. To train the RF model, the multi-label instances in each split were expanded into separate single-label instances and the feature vector, F, was set to the concatenation of $\{GF, GloVE, DF\}$. The model was fine tuned by varying the number of estimators (1-200), maximum tree depth (1-50), maximum number of features (1-116) from features F, minimum number of samples to split (2-40) and the minimum number of samples required to output a leaf node (1-40) on a logarithmic scale, resulting in a total number of $65{,}610$ combinations. The best performing configuration on the validation set achieved an accuracy rate of 0.386 with 61 estimators, a maximum depth of 13, maximum number of features equal to 23, minimum of 11 samples to split and a minimum of one sample to output a leaf node.

6.8 Results and Discussion

The results presented in this section are generated from training all models on the development set and evaluated on the test set. This process is repeated 50 times and the computed average metrics are tabulated in Table 6.7.

From the table, it can be noted that although the k-means clustering approach was trained to generate an average number of prepositions close to the dataset's label cardinality (i.e., 2.16), it ended up being the least accurate when predicting multiple spatial relations. The computed average accuracy and F-score recorded were 0.288 and 0.35 respectively. The agglomorative hierarchical clustering predicted spatial relations with an average label cardinality of 4.6 and achieved an average accuracy of 0.377. It is clear that the A-HC is over generating prepositions.

With an average label cardinality equal to the dataset's overall cardinality (i.e., 2.16), the Nearest Neighbour obtained an average accuracy of 0.47. Furthermore, the Multi-label Neural Network performed better (accuracy of 0.56) than all the other multi-label models, despite obtaining an average label cardinality of 1.85 which reflects its under

generation and scoring less in recall (0.638) when compared to the recall rate obtained by the A-HC model (0.767). The ML-NN; however, enjoys the highest precision (0.715), followed by NN (0.587), which turns out to be a good overall simple baseline model.

The RF model is a single-label classifier that ranks the prepositions in order of preference, and as such it is not a multi-label model. However, it gives some insights into its relative merits. Table 6.7 gives results for recall@k, for $k = 1$ to 4. The accuracy increases as k increases till $k = 3$ and then decreases, reflecting under and over generation of labels. At $k = 3$ the accuracy is 0.585 and recall 0.804, higher than ML-NN; however, precision (0.646) is less than in ML-NN. On the other hand, selecting $k = 3$ is like fitting the RF model to the dataset and strictly speaking requires another evaluation.

Table 6.7: Average metrics computed on the testing set when trained on the full development set based on the full feature vector: GloVe+GF+DF.

Model	LCard(Test)	Acc	P	R	F
kM-C	2.322	0.288	0.356	0.396	0.350
A-HC	4.626	0.377	0.427	**0.767**	0.548
NN	2.157	0.472	0.587	0.582	0.584
ML-NN	1.847	**0.560**	**0.715**	0.638	**0.674**
RF-1	1.000	0.380	0.848	0.380	0.509
RF-2	2.000	0.569	0.755	0.646	0.678
RF-3	3.000	0.585	0.646	0.804	0.699
RF-4	4.000	0.514	0.537	0.882	0.653

6.8.1 Human Evaluation

Human evaluations (HE) were also collected to assess both the correctness of the predicted preposition set by each model, and also to investigate the reliability of the dataset's ground truth labels. In this data collection exercise, eight French native speaking individuals were instructed to describe the spatial relationship between a given pair of objects depicted in an image by selecting all relevant prepositions out of all distinct spatial relations that were predicted by all models (P_M) and the corresponding dataset's ground truth (P_{GT}). In other words, the evaluators had to select all applicable prepositions from the union of all preposition sets ($P_{union} = \{P_m \mid m \in M\} \cup P_{GT}$). Each union set was listed together with the corresponding image containing the two objects enclosed in bounding boxes marked in blue and red to match the corresponding object labels. The evaluators were instructed to choose all prepositions (P_{HE}) that can be used to describe the spatial

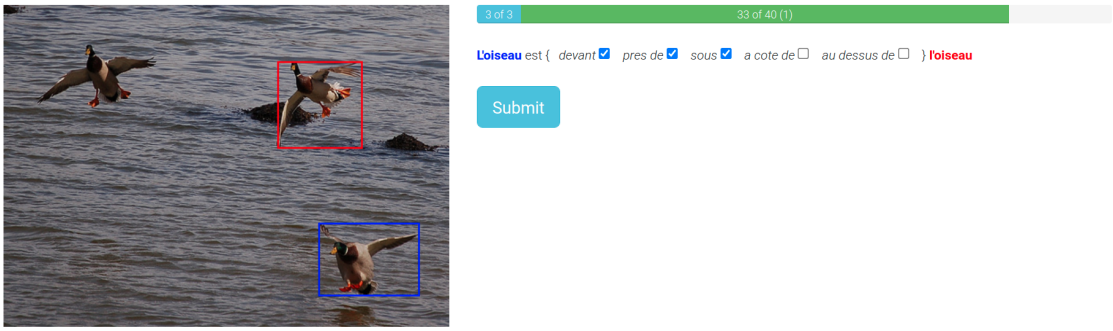


Figure 6.7: Human evaluation exercise: In this case, the French evaluator selected the prepositions *devant* ("in front of"), *près de* ("near") and *sous* ("under") as relevant spatial relations for the relationship between the two birds as found in image 2008_008268.jpg, Pascal VOC 2008 dataset (Everingham et al., 2010).

relationship between the two depicted objects by ticking all checkboxes found on the right of each preposition, as illustrated in Fig. 6.7, to simultaneously assess the quality of each model and the corresponding ground truth. The metrics described in Section 6.6 for the predicted and ground truth sets were computed with respect to the evaluated set (i.e., P_{HE}) rather than the actual dataset's ground truth labels Y. In this evaluation exercise, evaluators had to rate 40 multi-label predictions, in which 28 (70%) were unique for all evaluators, 8 (20%) duplicate across all evaluators which were used to compute the inter-rater agreement, and 4 (10%) duplicates for the same evaluator to compute the intra-rater agreement. From this evaluation, 5 out of 8 evaluators completed the full exercise and a total of 275 evaluations were collected as listed in Table 6.9. The mean and standard deviation of the label cardinality of the collected human evaluated set of prepositions were found to be equal to 2.244 and 1.129 respectively.

The average intra- and inter-rater percentage agreements were computed to assess the reliability and quality of the evaluations. The percentage intra-rater agreement was computed by calculating the average agreement over each pair of evaluations (total of 4 pairs per evaluator). This was carried out by first translating each P_{union} into one-hot encoding vector P_v, where each bit reflects whether each preposition was selected (1) or not (0) in accordance with P_{HE}. The disagreement between each pair (P_{v1}, P_{v2}) was obtained by identifying the non-matching bits as $P_d = P_{v1} \oplus P_{v2}$. The percentage agreement between the pair of evaluations was then calculated as:

$$Agg = 1 - \frac{1}{|P_d|} \sum_{i=1}^{|P_d|} P_d[i], \tag{6.16}$$

The mean of each evaluator's set of duplicate pairs were averaged to report the percentage intra-rater agreement which was found to be equal to 94%. Similarly, the average inter-rater agreement was calculated by computing the overall average pairwise agreement between the five evaluations that were collected for eight different images. The inter-rater agreement was found to be equal to 74%.

Since these calculated percentage agreements do not take into account the random chance agreement, they are prone to overestimate the level of reliability (McHugh, 2012). Hence, Cohen's kappa statistic (κ) (Cohen, 1960) which is one of the most widespread measures for inter-rater reliability (Delgado and Tibau, 2019) is computed. Given that this coefficient is not applicable to multi-label classification, the kappa agreement was calculated for each evaluated label as conducted by Bobicev and Sokolova (2017). This means that a one-hot indicator vector of size equal to the number of evaluations was used per label to denote in which evaluation that same label was used. For example, if the spatial relations between three different object pairs were described by {[*dans* ("in"), *près de* ("near")]; [*dans* ("in")]; [*derrière* ("behind"), *près de* ("near")]} by Evaluator A and {[*dans* ("in")]; [*près de* ("near")]; [*derrière* ("behind")]} by Evaluator B respectively, the following vectors are created:

Evaluator A: { *dans*: [1, 1, 0], *près de*: [1, 0, 1], *derrière*: [0, 0, 1] }
Evaluator B: { *dans*: [1, 0, 0], *près de*: [0, 1, 0], *derrière*: [0, 0, 1] }

The above vectors indicate that the preposition *dans* ("in") was used in the first two evaluations by Evaluator A ($[\mathbf{1}, \mathbf{1}, 0]$), while Evaluator B used it only in the first evaluation ($[\mathbf{1}, 0, 0]$). Based on these occurrences, a confusion matrix was then created for each label to reflect the agreement count between the two evaluations as shown in Table 6.8. For example, since the preposition *dans* ("in") was used by both evaluators in the first evaluation, cell ($A = 1, B = 1$) for column *dans* ("in") is set to 1. Similarly, since *derrière* ("behind") was only used in the third evaluation by both evaluators, cell ($A = 1, B = 1$) for column *derrière* ("behind") is set to 1, cell ($A = 0, B = 0$) is set to 2 and the rest with zeros.

The observed (p_o) and expected (p_e) probabilities are then calculated for each label. For example, the observed probability for the preposition *dans* ("in") is calculated by adding the number of times there is an agreement between the two evaluations i.e., $count(0,0) + count(1,1) = 2$ and divided by the total number of evaluations, which in this case is 3. Therefore, $p_o = \frac{2}{3} = 0.67$. On the other hand, the hypothetical probability of chance agreement p_e is computed by calculating the probability of not having the preposition *dans* ("in") by $p_0 = [(count(0,0) + count(0,1))/3] \times [(count(0,0) + count(1,0))/3] = \frac{1}{3} \times \frac{2}{3} = 0.22$. Similarly, p_1 is calculated by $\frac{2}{3} \times \frac{1}{3} = 0.22$. The overall expected probability

Table 6.8: Agreement count between the two evaluations (A, B) for the prepositions *dans* (“in”), *près de* (“near”), and *derrière* (“behind”).

		A					
		dans		*près de*		*derrière*	
		0	**1**	**0**	**1**	**0**	**1**
B	**0**	1	1	0	2	2	0
	1	0	1	1	0	0	1

is then calculated by: $p_e = p_0 + p_1 = 0.44$. The kappa score (κ) for the preposition *dans* (“in”) is then computed by:

$$\kappa = \frac{p_o - pe}{1 - pe} = \frac{0.67 - 0.44}{1 - 0.44} = 0.4 \tag{6.17}$$

This approach was applied to calculate the average intra- and inter-rater agreement by averaging over the number of selected labels. The intra-rater agreement was calculated per label for each duplicate evaluation and averaged over each label and five evaluators. Furthermore, the inter-rater agreement was computed by averaging the pairwise inter-rater agreement for each label across eight evaluations over five evaluators. The average intra-rater agreement was equal to 0.82, which according to Landis and Koch (1977), reflects an almost perfect agreement. On the other hand, the average inter-rater agreement was equal to 0.52 which indicates a moderate agreement and thus confirming the inconsistency between evaluators when it comes to choosing spatial relations. Both the percentage and the kappa values were taken as an indication that the human evaluations are of good quality given that the selection of spatial prepositions is an ambigous task (Retz-Schmidt, 1988) and the kappa score is an overly conservative measure (Gottschall, 2008; Murphy and Ciszewska-Carr, 2005; Muscat and Belz, 2017b). A summary of intra- and inter-rater agreements is tabulated in Tables 6.9 and 6.10 respectively.

As reported in Table 6.11, the human evaluation confirmed again that the kM-C approach is the least accurate method, followed by A-HC, NN and ML-NN. On the other hand, in contrast to the previous results, the human evaluations showed that the ML-NN achieved an accuracy of 0.502 and hence slightly better than the RF-3 (0.490) and RF-2 (0.485) by a margin of 2.45% and 3.51% respectively. It is interesting to note that both ML-NN and RF-3 were rated less by the evaluators than the ground truth. This observation was also noted when humans were asked to evaluate machine generated single-label spatial prepositions by Muscat and Belz (2017b) and Dobnik (2009). On the other hand, the evaluators in Dobnik (2009) often rated higher machine generated motion related

Table 6.9: Evaluators' percent and Cohen's kappa intra-rater agreements together with the corresponding number of evaluations per evaluator. The intra-rater agreement for evaluators 2, 3 and 7 was not computed since they did not complete the full evaluation exercise.

Evaluator	No. of evaluations	Percentage Agreement	Cohen's Kappa Statistic
1	41	0.97	0.88
2	14	-	-
3	12	-	-
4	80	1.00	1.00
5	40	0.93	0.67
6	40	0.84	0.64
7	8	-	-
8	40	0.98	0.93
Average	(Total=275) 34.38	0.94	0.82

Table 6.10: Inter-rater agreements for each evaluation using percentage agreement (a) and for each pair of evaluators using Cohen's kappa score (b). Note that evaluators 2, 3 and 7 are not included since they did not complete the full evaluation exercise.

Evaluation	Percentage Agreement
1	0.82
2	0.67
3	0.95
4	0.68
5	0.68
6	0.68
7	0.60
8	0.83
Average	0.74

(a)

Evaluator Pairs	Cohen's Kappa Statistic
1,4	0.54
1,5	0.56
1,6	0.51
1,8	0.76
4,5	0.62
4,6	0.50
4,8	0.47
5,6	0.41
5,8	0.55
6,8	0.32
Average	0.52

(b)

Table 6.11: Metrics (Mean, Std) computed over 275 French human evaluations for the multi-label models. Results are compared to the the RF model and to the evaluated dataset's ground truth (*GT*) which was considered as an additional model during the human evaluation process.

Model	LCard	Acc	P	R	F
GT	*2.222, 0.832*	*0.572, 0.289*	*0.715, 0.310*	*0.729, 0.315*	*0.681, 0.266*
kM-C	2.105, 1.323	0.299, 0.340	0.387, 0.398	0.404, 0.323	0.367, 0.368
A-HC	4.509, 2.812	0.379, 0.268	0.439, 0.314	**0.753, 0.356**	0.497, 0.279
NN	2.160, 0.785	0.431, 0.324	0.571, 0.374	0.558, 0.372	0.530, 0.331
ML-NN	1.964, 0.931	**0.502, 0.346**	**0.652, 0.460**	0.600, 0.379	**0.590, 0.346**
RF-1	1.000, 0.000	0.352, 0.323	0.698, 0.460	0.352, 0.323	0.442, 0.342
RF-2	2.000, 0.000	0.485, 0.300	0.645, 0.343	0.618, 0.344	0.596, 0.294
RF-3	3.000, 0.000	0.490, 0.267	0.559, 0.296	0.766, 0.315	0.612, 0.264
RF-4	4.000, 0.000	0.462, 0.228	0.490, 0.242	0.878, 0.246	0.598, 0.225

terms (movement in the four cardinal directions in mobile robot communications). These differences may be due to whether the task is strict, as in the multi-label case where all plausible prepositions are desired, or whether the detection of just one single preposition out of a set of equally acceptable ones is enough.

The human evaluations also revealed that the ground truth is not complete in terms of multi-label as its computed recall was found to be equal to 0.729. Also, the evaluation confirms human inconsistencies or disagreements when choosing spatial relations, as the precision of the ground truth preposition sets, was found to be equal to 0.715, and thus resulting in an overall accuracy of 0.572. The high standard deviation recorded across all measures for each evaluated model is probably related to the missing labels in the ground truth. Furthermore, on the basis of the human evaluation, the ML-NN and RF predict prepositions with an accuracy rate and F-score that are close to the evaluated ground truth. This result shows that despite the fact that theoretically the RF is not taking into consideration the dependency between classes, when predicting the top k prepositions, the RF compares well to the ML-NN model. This also raises the question about how much of the label dependency the ML-NN is making use of. Furthermore, Table 6.12 presents the recall per spatial relation as recorded by each model based on the full feature set. The results are presented together with the number and ratio of training, validation and testing instances that were used in these experiments.

The table shows that six prepositions: *au dessus de* ("above"), *le long de* ("along"), *autour de* ("around"), *par delà* ("beyond"), *aucun* ("none"), and *à l'exterieur de* ("outside of") were

best recalled using the k-means clustering approach despite its lowest average recall rate as presented in Table 6.7. The preposition *en face de* ("opposite") was best recalled (40%) by the Hierarchical Clustering approach. The greatest margin achieved by the kM-C was noted during the prediction of the preposition *le long de* ("along"). The model predicted the preposition with a recall rate of 0.81 and thus outperforms the second best multi-label model (i.e, A-HC) with a gain of 200%. Overall, the A-HC achieved a mean average recall of 0.57 followed by kM-C (0.52), NN (0.42) and ML-NN (0.41). As noted earlier, the ML-NN under generates and therefore does not do so well in recall. On the other hand, the mean weighted average recall results reveal that the A-HC predicts prepositions with a weighted average recall of 0.77, followed by the ML-NN (0.62), NN (0.58), and kM-C (0.39). For comparison, the RF-$\{2, 3\}$ achieved an average recall rate of 0.45 and 0.57 respectively and a weighted average recall of 0.64 and 0.81.

To study the level of significance between the developed models, a one-way ANOVA (Girden, 1992) on the 275 accuracy rates that where computed on the French human evaluations for each model (i.e., 8 models) and for the dataset's ground truth was first carried out. The analysis confirmed a significant difference among the predictive models and ground truth set (GT), $F(8, 266) = 21.89, p < 0.001$. Due to this statistical significance, the Tukey's HSD test (Tukey, 1949), was used to calculate the level of significance between each pairwise combination. When taking $p < 0.05$, as shown in Table 6.13, there was no significant difference between the accuracies obtained from the A-HC to those calculated on the NN and RF-1. Also, the analysis showed that the k-means clustering is not significantly different from the RF-1, while the ML-NN does not statistically differ from NN and RF-$\{2-4\}$. Furthermore, the Tukey's test confirmed that there was no significant difference between the NN and all Random Forest models (i.e., RF-$\{1-4\}$). The reported analysis also points out that there was a statistical difference between RF-1 model and all other RF models (i.e., RF-$\{2-4\}$) but no significant difference between the pairwise combinations of RF-2, RF-3 and RF-4. When comparing the evaluated dataset's ground truth with all models it was found that the accuracies of the ground truth significantly differ from all models except the ML-NN.

Table 6.12: Average recall per spatial relation (SR) when trained on the full feature set and computed on the testing set's ground-truth. Models are organised in Multi-Label and Single-Label categories. Each preposition is combined with the corresponding number and probabilities (Prob) of instances which were used during training, validation and testing.

				Average Recall							
				Multi-Label				Single-Label			
French SR (English SR)	Training instances (Prob)	Validation instances (Prob)	Testing instances (Prob)	ML-NN	NN	kM-C	A-HC	RF-1	RF-2	RF-3	RF-4
au dessus de ("above")	88 (0.01)	30 (0.02)	30 (0.01)	0.03	0.23	**0.44**	0.30	0.03	0.07	0.16	0.20
contre ("against")	462 (0.06)	126 (0.07)	142 (0.06)	0.31	0.43	0.27	**0.87**	0.11	0.57	0.69	0.89
le long de ("along")	59 (0.01)	15 (0.008)	11 (0.005)	0.03	0.09	**0.81**	0.27	0.00	0.00	0.00	0.10
autour de ("around")	34 (0.005)	2 (0.001)	6 (0.003)	0.55	0.50	**1.00**	0.50	0.79	0.89	0.91	0.94
au niveau de ("at the level of")	725 (0.10)	183 (0.10)	247 (0.11)	0.56	0.54	0.59	**0.78**	0.07	0.31	0.63	0.83
derrière ("behind")	858 (0.12)	217 (0.12)	251 (0.11)	0.64	0.61	0.19	**0.80**	0.49	0.71	0.82	0.87
par delà ("beyond")	29 (0.005)	8 (0.004)	10 (0.004)	0.00	0.20	**0.29**	0.20	0.00	0.00	0.16	0.29
loin de ("far from")	306 (0.04)	74 (0.04)	96 (0.04)	**0.66**	0.49	0.37	0.57	0.15	0.79	0.90	0.94
dans ("in")	50 (0.007)	9 (0.005)	15 (0.007)	0.36	0.47	**0.69**	0.67	0.25	0.45	0.76	0.87
devant ("in front of")	880 (0.12)	210 (0.11)	282 (0.12)	0.61	0.47	0.20	**0.72**	0.47	0.73	0.82	0.89
près de ("*near*")	1840 (0.25)	447 (0.24)	570 (0.25)	0.82	0.72	0.30	**0.87**	0.64	0.89	0.98	1.00
à côté de ("next to")	1124 (0.15)	280 (0.15)	369 (0.16)	0.60	0.59	0.49	**0.78**	0.15	0.52	0.88	0.95
aucun ("none")	16 (0.002)	5 (0.003)	7 (0.003)	0.00	0.00	**0.34**	0.17	0.00	0.00	0.01	0.02
sur ("on")	292 (0.04)	86 (0.04)	69 (0.04)	0.82	0.72	0.83	**0.87**	0.77	0.83	0.86	0.88
en face de ("opposite")	219 (0.03)	52 (0.03)	62 (0.03)	0.14	0.23	0.39	**0.40**	0.07	0.11	0.22	0.37
à l'exterieur de ("outside of")	31 (0.004)	12 (0.007)	8 (0.004)	0.20	0.13	**0.84**	0.13	0.00	0.10	0.17	0.37
sous ("under")	336 (0.05)	95 (0.05)	102 (0.05)	0.69	0.68	0.75	**0.76**	0.69	0.70	0.74	0.81
Mean Average Recall				0.41	0.42	0.52	**0.57**	0.28	0.45	0.57	0.66
Mean Weighted Average Recall				0.62	0.58	0.39	**0.77**	0.38	0.64	0.81	0.88

6.8.2 Qualitative Analysis

A qualitative analysis was conducted to gain further insight into where and why the models were failing. For this study, 100 distinct object pairs were randomly sampled from the 275 French evaluated instances. The predicted prepositions per model, the ground truth annotations and the evaluators' selections were compared to observe salient patterns. Fig. 6.8 depicts nine of the 100 instances representative of the observed patterns.

As anticipated, humans do not always agree when selecting spatial relations. This is mostly seen in the selection of near synonyms (Fig. 6.8(d) and (f)) and when choosing the frame of reference (Fig. 6.8(a) and (b)). In the case of near synonyms it is also likely that humans differ in the definition of spatial bounds (related to acceptability levels in template models), for example in the case of *près de* ("near") and *à côté de* ("next to"), as shown in Fig. 6.8(g) and (i).

In general, as also noted in the quantitative analysis, the ML-NN performed better than the NN, kM-C and A-HC models, at the cost of fine-grained and more spatially constrained relations. The ML-NN can be thought of as a conservative model with a higher tendency towards generating less spatially constrained prepositions, while opting for the more fine-grained spatial relations when absolutely necessary. This observation is in conformity with the highest recorded multi-label precision reported in the quantitative analysis. Surprisingly, despite being the overall least accurate, the kM-C model correctly predicts some of the less frequent spatial relations that other models struggle with, including for example, the relations *autour de* ("around"), *en face de* ("opposite"), and *le long de* ("along") as seen in Fig. 6.8(e) and (f), indicating that clustering models merit further investigation.

As expected, the models fail to select a more appropriate frame of reference (other than the camera view point) probably because there are no features that indicate the position or orientation of the objects (Fig. 6.8(a) and (b)) and possibly also because of the presence of *"person"* as one of the objects.

It was also evident that the models lacked accuracy because of near-synonym ambiguity. For example in some configurations humans were selective when dealing with prepositions sets like $\{$*à côté de* ("next to"), *près de* ("near"), *le long de* ("along")$\}$ and $\{$*loin de* ("far from"), *par delà* ("beyond")$\}$, while the models find it hard to comprehend the subtle differences among such prepositions. Likewise, the models found it difficult to discriminate between the prepositions *en face de* ("opposite") and *devant* ("in front of") as shown in Fig. 6.8(d), where object pose may also help in these situations. When dealing with antonyms, most of the time, the A-HC model chooses both *devant* ("in front of") and *derrière* ("behind"). The RF-3 (single-label model) sometimes chooses both in the first

Figure 6.8: A subset from the 100 human evaluations (HE) that were used for the qualitative analysis. Each sub-figure shows the pair of objects enclosed in bounding boxes. The ground-truth (GT) and the human evaluated (HE) prepositions are listed in the top part, the predictions of the multi-label models (kM-C, A-HC, NN, ML-NN) in the middle part, and the single-label (RF-3) model in the bottom part. Note that the prepositions are shown translated in English. The original French terms are included in the referring text.

three rankings. One would expect a single-label model to separate antonyms by a large distance in the rankings list. Spatial relations *dans* ("in") and *à l'exterieur de* ("outside of") are confused by the ML-NN and k-MC when the visible part of the object inside is partially outside and the bounding box only encloses the outside part; however, in the case of

the RF model it seems that language features help. The ML-NN and RF output *dans* (“in”) and *sur* (“on”) simultaneously (*sur* (“on”) is the correct one) when the subject is smaller than the object. Probably *sur* (“on”) is learnt from the language and *dans* (“in”) is learnt from the geometric features. When multiple spatial relations are required to accurately describe configurations in perspective and oblique views, mostly by (*à côté de* (“next to”) and *derrière* (“behind”)) all models fail to predict *derrière* (“behind”). An example is given in Fig. 6.8(g). This omission is possibly due to the close proximity and the marginal overlap between the two bounding boxes.

Results show that all models were prone to the dataset’s linguistic bias. In uncommon contexts, as shown for example in Fig. 6.8(h), the models were influenced by how normally a *“person”* is depicted in the presence of a *“horse”* (as in the majority of images in the training data). For another example, no model predicted *au dessus de* (“above”) in the context of Fig. 6.8(i), due to the bias problem and the lack of reasoning on spatial configurations. The lack of reasoning was further shown in Fig. 6.8(j), since for obvious reasons, all models owing to their limited features and lack of commonsense and world knowledge reasoning, were unable to correctly understand that the actual *“car”* is being reflected in the mirror, which in turn depicts a completely different geometric setup of that context.

Finally, the RF-3 model, which is inherently a single-label model, fails in contexts where fewer than three prepositions are suitable, a situation that this model cannot handle by design. In Fig. 6.8(c), the RF-3 over generates prepositions that are not suitable.

In summary, the analysis sheds light on promising directions this research could take. In particular, such directions include what other features are needed to discriminate between near synonyms and resolving the appropriate frame of reference, as well as methods to integrate common sense and world knowledge models.

6.9 Multi Spatial Relations in KENGIC

Given the availability of a compatible multi-label Spatial Relation (SR) dataset, the detection of multi SRs can be potentially used to further improve the quality of the generated captions by KENGIC. Multiple SRs can provide finer and more spatially descriptive captions. This can be achieved either by (a) using the most likely SRs instead of the implicitly generated relations generated by the n-gram graph, or (b) by using the set of predicted prepositions to validate the implicitly generated SRs. While the former could be used to provide rich and more spatially descriptive captions, the latter approach could provide a broader set to validate and correct SRs in KENGIC.

6.10 Summary

In this chapter, the usefulness of predicting multi and overlapping spatial relations in images was explored and a number of multi-label classification models were developed and analysed quantitatively and qualitatively. The results were also compared and contrasted with a single-label classifier which outputs a preposition list of constant length. Finally, the findings in the analysis serve to inform future work.

The accuracy, precision, recall and F-score were used in the quantitative analysis. The accuracy was computed as the intersection over union which penalises both over and under generation of labels. Overall, the multi-label neural network (ML-NN) performed better than all other multi-label models in terms of accuracy, precision and F-score. The Nearest Neighbour model provided a good baseline, while the clustering methods discriminated better in favour of the less common and difficult cases. The single-label Random Forest Classifiers (RF-2,3) scored better than the ML-NN when evaluated in terms of accuracy, recall and F-score on dataset's ground-truth, but the ML-NN outperformed the RF classifiers in terms of accuracy on the independent French evaluations. In addition, the ML-NN enjoys higher precision than the RF models, throughout. This is an indication that the ML-NN generalises better whilst the RF tends to overfit more to the ground truth, not least to the dataset's cardinality, since the latter is chosen manually. Furthermore, the qualitative analysis reveals examples, where the RF over-generates prepositions that are not correct (as evaluated by humans). When considering the human evaluation accuracy of all models and the ground truth, no significant difference was found between the ML-NN and RF and between the ML-NN and the GT, but the RF and GT were significantly different. Coupled with the results from the qualitative analysis, it can be concluded that the ML-NN has an advantage over the single-label RF model. This study also shows that the multi-preposition labels in the SpatialVOC2K dataset (Belz et al., 2018) used in the experiments are probably not always complete, which makes both training and evaluation of such machine learning models even harder. For these cases, the use of outlier detection and data imputation mechanisms may be appropriate. The qualitative analysis indicates that the inclusion of other features, in particular object position and orientation, and extrinsic knowledge models would potentially help in providing better discrimination between near synonyms and in addition resolve issues related to the frame of reference selection problem. Finally, this chapter concludes by projecting how multi spatial relation detection can be used in KENGIC to improve the quality of the generated captions.

Table 6.13: Tukey's HSD ($p < 0.05$) for pairwise comparison between the accuracies computed on each model and the dataset's ground truth based on the 275 evaluated instances by French native speaking individuals.

Model 1	Model 2	Mean Difference	P-Adjusted	Reject
A-HC	GT	0.193	0.001	T
A-HC	kM-C	-0.080	0.050	T
A-HC	ML-NN	0.123	0.001	T
A-HC	**NN**	**0.052**	**0.520**	**F**
A-HC	**RF-1**	**-0.027**	**0.900**	**F**
A-HC	RF-2	0.106	0.001	T
A-HC	RF-3	0.111	0.001	T
A-HC	RF-4	0.083	0.035	T
GT	kM-C	-0.273	0.001	T
GT	**ML-NN**	**-0.071**	**0.129**	**F**
GT	NN	-0.141	0.001	T
GT	RF-1	-0.220	0.001	T
GT	RF-2	-0.087	0.019	T
GT	RF-3	-0.082	0.036	T
GT	RF-4	-0.110	0.001	T
kM-C	ML-NN	0.202	0.001	T
kM-C	NN	0.131	0.001	T
kM-C	**RF-1**	0.053	0.506	**F**
kM-C	RF-2	0.185	0.001	T
kM-C	RF-3	0.190	0.001	T
kM-C	RF-4	0.162	0.001	T
ML-NN	**NN**	-0.071	0.130	**F**
ML-NN	RF-1	-0.150	0.001	T
ML-NN	**RF-2**	-0.017	0.900	**F**
ML-NN	**RF-3**	-0.012	0.900	**F**
ML-NN	**RF-4**	-0.040	0.808	**F**
NN	**RF-1**	-0.079	0.054	**F**
NN	**RF-2**	0.054	0.474	**F**
NN	**RF-3**	0.059	0.346	**F**
NN	**RF-4**	0.031	0.900	**F**
RF-1	RF-2	0.1328	0.001	T
RF-1	RF-3	0.1378	0.001	T
RF-1	RF-4	0.1097	0.001	T
RF-2	**RF-3**	0.005	0.900	**F**
RF-2	**RF-4**	-0.023	0.900	**F**
RF-3	**RF-4**	-0.028	0.900	**F**

7 Conclusions and Future Work

The use of keywords in a novel Keyword-driven and N-Gram Graph-based Image Captioning (KENGIC) framework was studied in this PhD programme. Inspired by the way how the human brain fires neurons while processing and interpreting the visual world, this PhD hypothesised that images can be automatically described given a set of visual keywords and linked through other intermediary words using an n-gram graph-based approach. This was proposed (a) as an alternative approach to models in image caption generation, (b) to reduce the dependency on large-scale paired image-caption datasets, and (c) to provide answers for the following research questions:

1. *Can image caption generation be cast as a graph search problem through a keyword-based n-gram graph?*
   This question was set to investigate whether image descriptions can be grounded in image keywords by using an n-gram graph-based data structure. This research confirmed that n-gram graphs can serve as intermediary representations that can efficaciously combine the vision and language domains in image captioning. This approach showed how the task of image captioning can be modularised in a two-step approach. The first step is to detect visual keywords from images to construct n-gram graphs that link the detected keywords through other intermediary words. The second step is to traverse the graph in a breadth-first approach to search for the best captions that describe the given query images based on a given set of criteria. This approach led to the generation of quality captions and hence confirmed that image captioning can be cast as a graph search problem through keyword-based n-gram graphs.

   a) *What is the role of image keywords in KENGIC?*
      This question was set to examine which type of keywords are most important for the generation of n-gram graphs. When constructing n-gram graphs based on human defined words, it was confirmed that nouns were the most important keywords for constructing n-gram graphs, as evaluated on automatic met-

rics. Adding attributes with nouns was found to slightly enhance the quality of the generated captions, but the use of finer keywords such as prepositions and verbs led to inferior captions as these impose further complexity in the graph generation and traversal. This was exacerbated when constructing graphs using composite keywords composed of nouns and attributes, or nouns and verbs due to their added constraint while traversing the graphs.

b) *What visual detectors are required?*
This research investigated both object detection and the prediction of visual keywords in a multi-label approach as detectors for KENGIC. Keywords consisting of detected objects led to accurate and fluent captions for images having few objects. However, more complex images were often described with high verbosity and non fluent captions due to larger keyword sets consisting of background objects. For this reason, a multi-label model was trained to predict the most visually relevant keywords including nouns, attributes and verbs. This keyword set provided less ambiguity and images were generally captioned better. Therefore, this research confirmed that multi-label visual detectors that are trained to detect salient visual keywords can provide pertinent keywords for KENGIC. Furthermore, given that spatial relations are difficult to be predicted from images, spatial relation detection was applied to enhance the use of spatial prepositions.

c) *What is the quality of the generated captions?*
Both quantitative and qualitative analysis showed that KENGIC generates quality image captions. As expected, given that KENGIC is not trained end-to-end on image caption pairs, it fell short when compared with models trained in the paired setting. On the other hand, KENGIC reached similar performance with current leading state-of-the-art image caption generators that are trained in the unpaired setting. Despite this performance, the qualitative analysis unveiled that (a) spatial prepositions were not always grounded in images, and (b) captions are generally penalised when they lack mentioning the frequently used human words, despite being relevant and accurate captions. The latter was not surprising since evaluation metrics generally measure the maximum overlapping n-gram sequences found between the candidate and reference captions (BLEU, METEOR, ROUGE), while CIDEr and SPICE consider all captions in their final metric and therefore, benefit from commonly used words across captions. De-

spite the lack of reliable human evaluation, the generated captions, were in general, rated well above average both in terms of accuracy and fluency.

2. *How does the selection of keywords affect the evaluation performance in image caption generation as measured by current automatic metrics?*
   Since image captions are generally evaluated by automatic metrics, this research questioned how keywords affect the evaluation performance based on such metrics. The preliminary work based on human extracted keywords confirmed that metrics influence the choice of words in image captions, in such a way that rich keywords are penalised over generic and less specific words. This was particularly evident when high quality captions composed of keywords extracted from human authored captions did not overlap with the remaining set of human captions. On the contrary, when using frequently used human keywords extracted from all ground-truth captions, the computed scores of the generated captions exceeded the human baseline. This concludes that the current popular evaluation metrics are highly biased in choosing frequent keywords and pay less attention to the structure of the generated captions. This claim was further substantiated given the fact that captions composed solely of frequent keywords were rated higher than the corresponding human authored captions. Both findings supported the crucial need for more robust automatic evaluation metrics.

3. *How does spatial relation detection contribute in automatic image captioning?*
   Given that the majority of current image captioning models generate captions without the explicit use of spatial relation detectors, this study aimed to investigate the role of spatial relations in image captioning. For this purpose, a spatial relation detector was developed and integrated in KENGIC framework. It was found that generally, the explicit detection of spatial relations based on geometric and linguistic features was not enough to enhance the quality of the relations generated by the n-gram graph. Apart from the limitations of the used spatial role labelling and relation detector which were later addressed in Chapter 6, captions mentioning frequently used prepositions benefited from higher scores, irrespective of their spatial grounding. Furthermore, it was also confirmed that generic prepositions were preferred over more spatially constrained prepositions. This led to the study of multi-spatial relation detection for image captioning to further increase the specificity of the generated captions.

7.1 Limitations

Despite the promising results obtained by the proposed approach, the implemented KENGIC framework has the following limitations which affected the overall caption quality:

1. **Visual Keywords**: Since the multi-label model was trained to predict visual keywords relevant to images, synonym words were found to considerably affect the generation process. In such cases, KENGIC ended up, either mentioning synonym words in same captions, or else lacking from including important image aspects which were detected by synonymous words. A more robust multi-label model could therefore, provide better quality image captions.
2. **Visual Relationship Detection (VRD)**: The semantic relationship between objects was limited only to Spatial Relation (SR) detection. However, this could be further extended to the more general VRD problem (Lu et al., 2016) that includes actions (e.g., "kick"), prepositions (e.g., "with"), verbs (e.g., "contain"), comparatives (e.g., "larger than"), and prepositional phrases (e.g., "stand on"). Furthermore, the employed SR detection model was based on simple and naïve spatial role labelling approach which in turn, affected the accuracy of the extraction of SRs from the generated captions. More sophisticated techniques can be introduced for better extraction and to enhance the spatial relation integration in KENGIC. The integrated SR detection model was found to reduce the quality of the generated captions due to its limited feature set as, well as, its highly skewed training data. Although these limitations were specifically studied with the introduction of depth features and the use of better quality dataset for multi spatial relation detection, the aforementioned limitations were not integrated in KENGIC due to the disparity between the two datasets.
3. **Hyper-parameter Tuning**: KENGIC was fine-tuned for the generation of captions based on human extracted keywords. Due to time and hardware availability constraints, the same hyper-parameters were used for the experiments based on predicted keywords. These hyper-parameters could possibly not be the optimal parameters for all the tested models.
4. **Evaluation**: Given that this research showed major limitations in the current standard automatic evaluation metrics, it weakens its quantitative analysis. Furthermore, the human evaluation was not found to be very reliable and the reported analysis is not very conclusive, given the low intra- and inter- annotator agreements. This calls for a possible revision of the human evaluation exercise. In particular two suggestions are put forward; (a) evaluators are recruited from a pool of linguists who are native language speakers and (b) to further increase the reliability of the evaluation exercise,

rather than using a Likert-scale that spans over five ratings to assess the correctness and fluency of captions, a binary decision is used to assess different criteria such as whether captions have (i) hallucinations, (ii) major and (iii) minor grammatical mistakes. This would better control the evaluation process, provide consistent quality and reliable assessments whilst reducing the subjectivity of the human evaluators.

7.2 Future Work

The proposed KENGIC framework paved the way for an alternative research direction in image caption generation. In contrast to current state-of-the-art image caption generators, this has provided a mechanism where each sub-module can be extended and improved independently. Furthermore, this research opened the door for the following research questions:

1. *How can image captions be automatically better evaluated?*
   This research confirmed that most of the current dominating evaluation metrics are not going to help the research community in the long-term. The n-gram based metrics (i.e., BLEU, METEOR, ROUGE and CIDEr) are not reliable and they are even showing that the generated image captions exceed human level quality, despite their lack of accuracy and fluency. This calls for robust and well developed metrics which go beyond calculating the n-gram overlap between the generated and corresponding human captions. For example, SPICE makes use of scene graphs to assess the quality of image captions.
2. *What is the quality of an unsupervised KENGIC?*
   This research investigated KENGIC when having images and corresponding text corpus from the same domain. It would be worth studying the quality of the generated captions given that the two modalities are not from the same domain. This would further help in reducing the dependency of datasets consisting of images and corresponding textual data. KENGIC could be further extended by the web-retrieval framework described in Birmingham and Muscat (2017) to fetch for visually related keywords from the Web. In addition, this would further reduce the need for labelled visual keywords relevant to query images in KENGIC, while opening the opportunity for a life-long learning paradigm in image captioning.
3. *Can KENGIC be used in Visual Question Answering (VQA)?*
   Given that KENGIC is modularised and based on n-gram graphs, this framework could also be explored in VQA to provide answers for visual questions.

7.3 Final Remarks

This PhD studied how image captions can be generated from a set of visual keywords using a novel graph-based approach. The experiments of this study confirmed the potential of this method and how this research field can benefit from the findings of this work. The encouraging results and the new insights that have been presented to narrow the research gaps in this field, could further help the AI research community moving towards Alan Turing's vision who hoped that one day, "machines will eventually compete with men in all purely intellectual fields" (Turing, 1950).

References

Christmas tree on a boat. URL https://www.lovethispic.com/image/138607/christmas-tree-on-a-boat. [Online; accessed June 10, 2022].

Aditya, S., Yang, Y., Baral, C., Aloimonos, Y., and Fermüller, C. Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 173:33–45, 2017. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2017.12.004.

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Aisopos, F., Papadakis, G., and Varvarigou, T. Sentiment analysis of social media content using n-gram graphs. In *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media*, WSM '11, page 9–14, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450309899. doi: 10.1145/2072609.2072614. URL https://doi.org/10.1145/2072609.2072614.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1098. URL https://aclanthology.org/D17-1098.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1247–III–1255. JMLR.org, 2013. URL http://dl.acm.org/citation.cfm?id=3042817.3043076.

Arthur, D. and Vassilvitskii, S. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5. URL http://dl.acm.org/citation.cfm?id=1283383.1283494.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *Journal of machine learning research*, 3:1–48, 2002.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Ball, K., Smith, D., Ellison, A., and Schenk, T. Both egocentric and allocentric cues support spatial priming in visual search.

Neuropsychologia, 47(6):1585 – 1591, 2009. ISSN 0028-3932. doi: https://doi.org/10.1016/j.neuropsychologia.2008.11.017. Perception and Action.

Banerjee, S. and Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W05-0909.

Barnard, K. and Forsyth, D. Learning the semantics of words and pictures. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 408–415 vol.2, 2001. doi: 10.1109/ICCV.2001.937654.

Belz, A., Muscat, A., Aberton, M., and Benjelloun, S. Describing spatial relationships between objects in images in English and French. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 104–113, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2816. URL https://aclanthology.org/W15-2816.

Belz, A., Muscat, A., Birmingham, B., Levacher, J., Pain, J., and Quinquenel, A. Effect of data annotation, feature selection and model choice on spatial description generation in French. In *Proceedings of the 9th International Natural Language Generation conference*, pages 237–241, Edinburgh, UK, September 5-8 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-6639. URL https://aclanthology.org/W16-6639.

Belz, A., Muscat, A., Anguill, P., Sow, M., Vincent, G., and Zinessabah, Y. SpatialVOC2K: A multilingual dataset of images with annotations and features for spatial relations between objects. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 140–145, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6516. URL https://aclanthology.org/W18-6516.

Ben, H., Pan, Y., Li, Y., Yao, T., Hong, R., Wang, M., and Mei, T. Unpaired image captioning with semantic-constrained self-learning. *IEEE Transactions on Multimedia*, 24:904–916, 2022. doi: 10.1109/TMM.2021.3060948.

Ben-Younes, H., Cadene, R., Thome, N., and Cord, M. Block: Bilinear superdiagonal fusion for visual question answering and ship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109, 2019.

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1171–1179, Cambridge, MA, USA, 2015. MIT Press.

Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, March 1994. ISSN 1045-9227. doi: 10.1109/72.279181.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2 2016. ISSN 1076-9757. doi: 10.1613/jair.4900.

Bickel, S., Haider, P., and Scheffer, T. Predicting sentences using n-gram language models. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 193–200, 2005.

Birmingham, B. and Muscat, A. The use of object labels and spatial prepositions as keywords in a web-retrieval-based image caption generation system. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 11–20, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2002. URL https://aclanthology.org/W17-2002.

Birmingham, B. and Muscat, A. Clustering-based model for predicting multi-spatial relations in images. In *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics, ICINCO 2019 - Volume 2, Prague, Czech Republic, July 29-31, 2019*, pages 147–156, 2019. doi: 10.5220/0008123601470156. URL https://doi.org/10.5220/0008123601470156.

Birmingham, B. and Muscat, A. Multi spatial relation detection in images. *Spatial Cognition & Computation*, 22(3-4):293–327, 2022. doi: 10.1080/13875868.2021.1957897.

Birmingham, B., Muscat, A., and Belz, A. Adding the third dimension to spatial relation detection in 2d images. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 146–151, 2018. URL https://aclanthology.info/papers/W18-6517/w18-6517.

Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., USA, 1995. ISBN 0198538642.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Bobicev, V. and Sokolova, M. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6{_}015. URL https://doi.org/10.26615/978-954-452-049-6{_}015.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130401. URL https://doi.org/10.1145/130385.130401.

Bowerman, M., Levinson, S. C., and Levinson, S. *Language acquisition and conceptual development*. Number 3. Cambridge University Press, 2001.

Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.

Cangelosi, A., Coventry, K. R., Rajapakse, R., Joyce, D., Bacon, A., Richards, L., and Newstead, S. N. Grounding language in perception: A connectionist model of spatial terms and vague quantifiers. In Cangelosi, A., Bugmann, G., and Borisyuk, R., editors, *Modeling Language, Cognition and Action: Proceedings of the 9th Neural Computation and Psychology Workshop*, pages 47–56. World Scientific, Singapore, 2005.

Cao, S., An, G., Zheng, Z., and Ruan, Q. Interactions guided generative adversarial network for unsupervised image captioning. *Neurocomputing*, 417:419–431, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.08.019.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58451-1. doi: 10.1007/978-3-030-58452-8{_}13. URL https://doi.org/10.1007/978-3-030-58452-8{_}13.

Carlson-Radvansky, L. A. and Logan, G. D. The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37:411–437, 1997.

Carlson-Radvansky, L. A. and Radvansky, G. A. The influence of functional relations on spatial term selection. *Psychological Science*, 7(1):56–60, January 1996.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

Chen, D. and Manning, C. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1082. URL https://www.aclweb.org/anthology/D14-1082.

Chen, S., Jin, Q., Wang, P., and Wu, Q. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Chen, T., Wang, Z., Li, G., and Lin, L. Recurrent attentional reinforcement learning for multi-label image recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

Chen, X., Zhang, M., Wang, Z., Zuo, L., Li, B., and Yang, Y. Leveraging unpaired out-of-domain data for image captioning. *Pattern Recognition Letters*, 132:132 – 140, 2020b. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2018.12.018. Multiple-Task Learning for Big Data (MTL4BD).

Chen, X. and Lawrence Zitnick, C. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2422–2431, 2015.

Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Chesterton, A. "An image of a car parked close to a street post and a garage." How close can you legally park to a driveway or corner?, 14 Mar 2017, 2017. URL https://www.carsguide.com.au/car-advice/how-close-can-you-legally-park-to-a-driveway-or-corner-53696. [Online; accessed June 8, 2020].

Chisholm, M. C. Learning decision rules by randomized iterative local search. In *Proceedings ICML-02*, pages 75–82. Morgan Kaufmann, 2002.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014a. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics, 2014b. doi: 10.3115/v1/D14-1179. URL http://aclweb.org/anthology/D14-1179.

Chomsky, N. Language and nature. *Mind*, 104(413):1–61, 1995.

Clark, P., Porter, B., and Works, B. P. Km–the knowledge machine 2.0: Users manual. *Department of Computer Science, University of Texas at Austin*, 2(5), 2004.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Cortes, C. and Vapnik, V. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

Coventry, K. R. and Garrod, S. C. *Saying, seeing and acting: The psychological semantics of spatial prepositions*. Psychology Press, London, 2004.

Coventry, K. R., Prat-Sala, M., and Richards, L. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of Memory and Language*, 44(3):376 – 398, 2001. ISSN 0749-596X. doi: https://doi.org/10.1006/jmla.2000.2742.

Coventry, K. R., Cangelosi, A., Rajapakse, R., Bacon, A., Newstead, S., Joyce, D., and Richards, L. V. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Freksa, C., Knauff, M., Krieg-Brückner, B., Nebel, B., and Barkowsky, T., editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, pages 98–110, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32255-9.

Dai, B., Fidler, S., Urtasun, R., and Lin, D. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017a.

Dai, B., Zhang, Y., and Lin, D. Detecting visual relationships with deep relational networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308, Los Alamitos, CA, USA, July 2017b. IEEE Computer Society. doi: 10.1109/CVPR.2017.352. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.352.

Dai, J., Li, Y., He, K., and Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.

Dai, Z., Cai, B., Lin, Y., and Chen, J. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1610, June 2021.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

Day, W. H. and Edelsbrunner, H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

Delgado, R. and Tibau, X. A. Why cohen's kappa should be avoided as performance measure in classification. *PLOS ONE*, 14(9):1–26, 09 2019. doi: 10.1371/journal.pone.0222916.

Dembczyński, K., Cheng, W., and Hüllermeier, E. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 279–286, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

Denkowski, M. and Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-3348.

Dey, A., Jenamani, M., and Thakkar, J. J. Senti-n-gram: An n-gram lexicon for sentiment analysis. *Expert Systems with Applications*, 103:92–105, 2018.

Ding, X., Li, B., Xiong, W., Guo, W., Hu, W., and Wang, B. Multi-instance multi-label learning combining hierarchical context and its application to image annotation. *IEEE Transactions on Multimedia*, 18(8):1616–1627, 2016. doi: 10.1109/TMM.2016.2572000.

Dobnik, S. *Teaching mobile robots to use spatial words*. PhD thesis, University of Oxford, 2009.

Dobnik, S. and Kelleher, J. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third Workshop on Vision and Language*, pages 33–37, Dublin, Ireland, 2014. Dublin City University and the Association for Computational Linguistics. doi: 10.3115/v1/W14-5405. URL http://www.aclweb.org/anthology/W14-5405.

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, 2017.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

Dunbar, R. Why only humans have language. 2009.

Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993. cited By 1072.

Duygulu, P., Barnard, K., Freitas, J. F. G. d., and Forsyth, D. A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, page 97–112, Berlin, Heidelberg, 2002. Springer-Verlag. ISBN 3540437487.

Elliott, D. and Keller, F. Image description using visual dependency representations. In *Conference on Emprical Methods in Natural Language Processing*, volume 13, pages 1292–1302, 2013.

Ellis, N. C. Emergentism, connectionism and language learning. *Language learning*, 48(4):631–664, 1998.

Elman, J. L. Language as a dynamical system. *Mind as motion: Explorations in the dynamics of cognition*, pages 195–223, 1995.

Erhan, D., Szegedy, C., Toshev, A., and Anguelov, D. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick, C. L., and Zweig, G. From captions to visual concepts and back. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1473–1482, 2015.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. Every picture tells a story: Generating sentences from images. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15561-1.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1184. URL https://www.aclweb.org/anthology/N15-1184.

Fasola, J. and Matarić, M. J. Using spatial language to guide and instruct robots in household environments. In *AAAI Fall Symposium: Robots Learning Interactively from Human Teachers*, Arlington, VA, Nov 2012. URL "http://robotics.usc.edu/publications/785/".

Fedus, W., Goodfellow, I. J., and Dai, A. M. Maskgan: Better text generation via filling in the _____. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=ByOExmWAb.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2009.167.

Feng, Y., Ma, L., Liu, W., and Luo, J. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Friedman, A. Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of experimental psychology: General*, 108(3):316, 1979.

Fukushima, K. and Miyake, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

Gao, L., Song, J., Nie, F., Yan, Y., Sebe, N., and Shen, H. T. Optimal graph learning with partial tags and multiple features for image and video annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4371–4379, 2015. doi: 10.1109/CVPR.2015.7299066.

Gao, L., Wang, B., and Wang, W. Image captioning with scene-graph based semantic concepts. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, ICMLC 2018, page 225–229, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450363532. doi: 10.1145/3195106.3195114. URL https:

//doi.org/10.1145/3195106.3195114.

Gatt, A. and Reiter, E. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, page 90–93, USA, 2009. Association for Computational Linguistics.

Ghamrawi, N. and McCallum, A. Collective multi-label classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, page 195–200, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931406. doi: 10.1145/1099554.1099591. URL https://doi.org/10.1145/1099554.1099591.

Ghosh, S., Burachas, G., Ray, A., and Ziskind, A. Generating natural language explanations for visual question answering using scene graphs and visual attention. *arXiv preprint arXiv:1902.05715*, 2019.

Giannakopoulos, G., Karkaletsis, V., Vouros, G., and Stamatopoulos, P. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, 5(3), oct 2008. ISSN 1550-4875. doi: 10.1145/1410358.1410359.

Gilberto Mateos Ortiz, L., Wolff, C., and Lapata, M. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1174. URL https://aclanthology.org/N15-1174.

Girden, E. R. *ANOVA: Repeated measures*. Number 84. Sage Publications, Inc., Newbury Park, CA, 1992.

Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Godard, C., Mac Aodha, O., and Brostow, G. J. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

Gong, Y., Jia, Y., Leung, T., Toshev, A., and Ioffe, S. Deep convolutional ranking for multilabel image annotation. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.4894.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Gottschall, J. *Literature, science, and a new humanities*. Cognitive Studies in Literature Performance. Palgrave Macmillan, New York, 2008.

Graves, A. and Schmidhuber, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2005.06.042. IJCNN 2005.

Graves, A., Wayne, G., and Danihelka, I. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 13–23, Genoa, Italy, 2006.

Gu, J., Wang, G., Cai, J., and Chen, T. An empirical study of language cnn for image captioning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1231–1240, 2017. doi: 10.1109/ICCV.2017.138.

Gu, J., Cai, J., Wang, G., and Chen, T. Stack-captioning: Coarse-to-fine learning for image captioning. In *AAAI*, 2018a.

Gu, J., Joty, S., Cai, J., and Wang, G. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018b.

Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., and Wang, G. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10323–10332, 2019.

Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 309–316, 2009. doi: 10.1109/ICCV.2009.5459266.

Guo, Y. and Gu, S. Multi-label classification using conditional dependency networks. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, page 1300–1305. AAAI Press, 2011. ISBN 9781577355144.

Gupta, A., Verma, Y., and Jawahar, C. Choosing linguistics over vision to describe images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2012.

Hamid Amiri, S. and Jamzad, M. Efficient multi-modal fusion on supergraph for scalable image annotation. *Pattern Recogn.*, 48(7):2241–2253, jul 2015. ISSN 0031-3203. doi: 10.1016/j.patcog.2015.01.015.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. Simultaneous detection and segmentation. In *European conference on computer vision*, pages 297–312. Springer, 2014.

Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

Harzallah, H., Jurie, F., and Schmid, C. Combining efficient object localization and image classification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 237–244, 2009. doi: 10.1109/ICCV.2009.5459257.

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., Chomsky, N., and Lewontin, R. C. The mystery of language evolution. *Frontiers in Psychology*, 5, 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2014.00401.

He, K., Zhang, X., Ren, S., and Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017a.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017b. doi: 10.1109/ICCV.2017.322.

Hecht-Nielsen, R. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.

Hède, P., Moëllic, P.-A., Bourgeoys, J., Joint, M., and Thomas, C. Automatic generation of natural language descriptions for images. In *Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, RIAO '04, pages 306–313, Paris, France, France, 2004. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. ISBN 905450-09-6. URL http://dl.acm.org/citation.cfm?id=2816272.2816300.

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018.

Herdade, S., Kappeler, A., Boakye, K., and Soares, J. *Image Captioning: Transforming Objects into Words*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Herskovits, A. Language, spatial cognition, and vision. In Stock, O., editor, *Spatial and temporal reasoning*, chapter 6, pages 155–202. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

Herskovits, A. On the spatial uses of prepositions. In *Proceedings of the 18th Annual Meeting on Association for Computational Linguistics*, ACL '80, page 1–5, USA, 1980. Association for Computational Linguistics. doi: 10.3115/981436.981438. URL https://doi.org/10.3115/981436.981438.

Herzig, R., Bar, A., Xu, H., Chechik, G., Darrell, T., and Globerson, A. Learning canonical representations for scene graph to image generation. *ArXiv*, abs/1912.07414, 2019.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hodosh, M., Young, P., and Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Hokamp, C. and Liu, Q. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1141. URL https://aclanthology.org/P17-1141.

Holtzman, A., Buys, J., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *International Conference on Learning Representations*, 2020.

Huang, F., Li, Z., Wei, H., Zhang, C., and Ma, H. Boost image captioning with knowledge reasoning. *Machine Learning*, 109: 1–20, 12 2020. doi: 10.1007/s10994-020-05919-y.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.

Hürlimann, M. and Bos, J. Combining lexical and spatial knowledge to predict spatial relations between objects in images. In *Proceedings of the 5th Workshop on Vision and Language*, pages 10–18, 2016.

Jaimes, A. and Chang, S.-F. Conceptual framework for indexing visual information at multiple levels. In *Electronic Imaging*, pages 2–15. International Society for Optics and Photonics, 1999.

Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2407–2415, 2015.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015.

Johnson, J., Gupta, A., and Fei-Fei, L. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

Jozefowicz, R., Zaremba, W., and Sutskever, I. An empirical exploration of recurrent network architectures. In *International conference on machine learning*, pages 2342–2350. PMLR, 2015.

Kalchbrenner, N. and Blunsom, P. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709. Association for Computational Linguistics, 2013. URL http://aclweb.org/anthology/D13-1176.

Kamarainen, J.-K., Kyrki, V., and Kalviainen, H. Invariance properties of gabor filter-based features-overview and applications. *IEEE Transactions on image processing*, 15(5):1088–1099, 2006.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017. doi: 10.1109/TPAMI.2016.2598339.

Karpathy, A., Joulin, A., and Li, F. F. F. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

Kelleher, J. D. and Kruijff, G. M. A context-dependent algorithm for generating locative expressions in physically situated environments. In Wilcock, G., Jokinen, K., Mellish, C., and Reiter, E., editors, *Proceedings of the Tenth European Workshop on Natural Language Generation, ENLG 2005*, Aberdeen, UK, 2005. ACL. URL https://www.aclweb.org/anthology/W05-1607/.

Kelleher, J. D., Ross, R. J., Sloan, C., and Namee, B. M. The effect of occlusion on the semantics of projective spatial terms: a case study in grounding language in perception. *Cognitive Processing*, 12(1):95–108, Feb 2011.

Kemmerer, D. *Cognitive neuroscience of language*. Psychology Press, 2014.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–578, Munich, Germany, 2018.

Kim, J., Misu, T., Chen, Y. T., Tawari, A., and Canny, J. Grounding human-to-vehicle advice for self-driving vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10591–10599, Long Beach, CA, USA, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR, Conference Track Proceedings*, San Diego, CA, USA, 2015. URL http://arxiv.org/abs/1412.6980.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. Unifying visual-semantic embeddings with multimodal neural language models. In *Advances in Neural Information Processing Systems Deep Learning Workshop*, 2015.

Klein, D. and Manning, C. D. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075150. URL https://www.aclweb.org/anthology/P03-1054.

Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT. URL http://mt-archive.info/MTS-2005-Koehn.pdf.

Kojima, A., Tamura, T., and Fukunaga, K. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, Nov 2002. ISSN 1573-1405. doi: 10.1023/A:1020346032608.

Kolesnikov, A., Kuznetsova, A., Lampert, C., and Ferrari, V. Detecting visual relationships using box attention. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1749–1753, Los Alamitos, CA, USA, oct 2019. IEEE Computer Society. doi: 10.1109/ICCVW.2019.00217. URL https://doi.ieeecomputersociety.org/10.1109/ICCVW.2019.00217.

Kolomiyets, O., Kordjamshidi, P., Moens, M.-F., and Bethard, S. SemEval-2013 task 3: Spatial role labeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 255–262, Atlanta, Georgia, USA, June 2013. Association for Computational

Linguistics. URL https://aclanthology.org/S13-2044.

Kordjamshidi, P. and Moens, M. F. Global machine learning for spatial ontology population. *Journal of Web Semantics*, 30: 3 – 21, 2015. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2014.06.001. Semantic Search.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process.*, 8(3), dec 2011. ISSN 1550-4875. doi: 10.1145/2050104.2050105.

Kouloumpis, E., Wilson, T., and Moore, J. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 538–541, 2011.

Kreiman, G., Koch, C., and Fried, I. Imagery neurons in the human brain. *Nature*, 408(6810):357–361, 2000.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, May 2017. ISSN 1573-1405. doi: 10.1007/s11263-016-0981-7.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kuhl, P. K. Brain mechanisms in early language acquisition. *Neuron*, 67(5):713–727, 2010. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2010.08.038.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 35(12):2891–2903, dec 2013a. ISSN 1939-3539. doi: 10.1109/TPAMI.2012.162.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013b.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 359–368. Association for Computational Linguistics, 2012.

Laina, I., Rupprecht, C., and Navab, N. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7414–7424, 2019.

Lampert, C. H., Blaschko, M. B., and Hofmann, T. Beyond sliding windows: Object localization by efficient subwindow search. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587586.

Lanchantin, J., Wang, T., Ordonez, V., and Qi, Y. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16478–16488, June 2021.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420.

Lebret, R., Pinheiro, P. O., and Collobert, R. Phrase-based image captioning. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

Lee, K.-H., Palangi, H., Chen, X., Hu, H., and Gao, J. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators, 2019.

Lei, C., Liu, D., and Li, W. Social diffusion analysis with common-interest model for image annotation. *IEEE Transactions on Multimedia*, 18(4):687–701, 2016. doi: 10.1109/TMM.2015.2477277.

Lenneberg, E. H. The biological foundations of language. *Hospital Practice*, 2(12):59–67, 1967. doi: 10.1080/21548331.1967.11707799.

Li, G., Zhu, L., Liu, P., and Yang, Y. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8928–8937, 2019a.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics, 2011.

Li, X., Wang, L., and Sung, E. Multilabel svm active learning for image classification. In *2004 International Conference on Image Processing, 2004. ICIP '04.*, volume 4, pages 2207–2210 Vol. 4, 2004. doi: 10.1109/ICIP.2004.1421535.

Li, X. and Jiang, S. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, 21(8): 2117–2130, 2019.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

Li, Y., Ouyang, W., Wang, X., and Tang, X. Vip-cnn: Visual phrase guided convolutional neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7244–7253, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.766. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.766.

Li, Y., Yu, J., Zhan, Y., and Chen, Z. Relationship graph learning network for visual relationship detection. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, MMAsia '20, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383080. doi: 10.1145/3444685.3446312. URL https://doi.org/10.1145/3444685.3446312.

Li, Y., Ouyang, W., Zhou, B., Shi, J., Zhang, C., and Wang, X. Factorizable net: An efficient subgraph-based framework for scene graph generation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 346–363, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01246-5.

Li, Z., Wang, R., Chen, K., Utiyama, M., Sumita, E., Zhang, Z., and Zhao, H. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*, 2019b.

Liang, X., Lee, L., and Xing, E. P. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 848–857, 2017.

Lienhart, R. and Maydt, J. An extended set of haar-like features for rapid object detection. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I, 2002. doi: 10.1109/ICIP.2002.1038171.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Lin, D., Fidler, S., Kong, C., and Urtasun, R. Generating multi-sentence natural language descriptions of indoor scenes. In *British Machine Vision Conference*, 2015.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common

objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017a.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017b.

Liu, A., Xu, N., Zhang, H., Nie, W., Su, Y., and Zhang, Y. Multi-level policy and reward reinforcement learning for image captioning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 821–827. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/114. URL https://doi.org/10.24963/ijcai.2018/114.

Liu, F., Xiang, T., Hospedales, T. M., Yang, W., and Sun, C. Semantic regularisation for recurrent image annotation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4160–4168, 2017a. doi: 10.1109/CVPR.2017.443.

Liu, F., Gao, M., Zhang, T., and Zou, Y. Exploring semantic relationships for image captioning without parallel data. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 439–448. IEEE, 2019a.

Liu, H. and Singh, P. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004. ISSN 1358-3948. doi: http://dx.doi.org/10.1023/B:BTTJ.0000047600.45421.6d.

Liu, S., Demirel, M. F., and Liang, Y. *N-Gram Graph: Simple Unsupervised Representation for Graphs, with Applications to Molecules*. Curran Associates Inc., Red Hook, NY, USA, 2019b.

Liu, S., Zhang, L., Yang, X., Su, H., and Zhu, J. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., and Murphy, K. Improved image captioning via policy gradient optimization of spider. In *ICCV*, pages 873–881. IEEE Computer Society, 2017b.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

Logan, G. D. and Sadler, D. D. *A computational analysis of the apprehension of spatial relations*, pages 493–529. The MIT Press, Cambridge, MA, US, 1996. ISBN 0-262-02403-9 (Hardcover).

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, nov 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000029664.99615.94.

Lu, C., Krishna, R., Bernstein, M., and Fei-Fei, L. Visual relationship detection with language priors. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 852–869, Amsterdam, The Netherlands, 2016. Springer International Publishing. ISBN 978-3-319-46448-0.

Lu, J., Xiong, C., Parikh, D., and Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250, 2017. doi: 10.1109/CVPR.2017.345.

Lu, J., Yang, J., Batra, D., and Parikh, D. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ma, L., Lu, Z., Shang, L., and Li, H. Multimodal convolutional neural networks for matching image and sentence. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2623–2631, Dec 2015. doi: 10.1109/ICCV.2015.301.

MacWhinney, B. Models of the emergence of language. *Annual review of psychology*, 49(1):199–227, 1998.

Makadia, A., Pavlovic, V., and Kumar, S. A new baseline for image annotation. In Forsyth, D., Torr, P., and Zisserman, A., editors, *Computer Vision – ECCV 2008*, pages 316–329, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN

978-3-540-88690-7.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-5010. URL https://aclanthology.org/P14-5010.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*, 2015.

Marcheggiani, D. and Titov, I. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1159. URL https://www.aclweb.org/anthology/D17-1159.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017.

Maron, O. and Lozano-Pérez, T. A framework for multiple-instance learning. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1998. URL https://proceedings.neurips.cc/paper/1997/file/82965d4ed8150294d4330ace00821d77-Paper.pdf.

Martinez, G. C., Cangelosi, A., and Coventry, K. R. A hybrid neural network and virtual reality system for spatial language processing. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, volume 1, pages 16–21, Washington, DC, USA, July 2001. doi: 10.1109/IJCNN.2001.938984.

Mason, R. and Charniak, E. Nonparametric method for data-driven image captioning. In *ACL (2)*, pages 592–598, 2014.

McCloskey, M. Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2(6):387–395, 1991. doi: 10.1111/j.1467-9280.1991.tb00173.x.

McHugh, M. L. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

Meyering, D. Tampa riverwalklighted boat parade, 2020. URL https://tampabaydatenightguide.com/wp-content/uploads/sites/2/2020/11/tampa-riverwalk-boat-parade-1024x684.jpg. [Online; accessed June 10, 2022].

Mikolov, T. and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. Recurrent neural network based language model. In Kobayashi, T., Hirose, K., and Nakamura, S., editors, *INTERSPEECH*, pages 1045–1048. ISCA, 2010. URL http://dblp.uni-trier.de/db/conf/interspeech/interspeech2010.html#MikolovKBCK10.

Milewski, V. S. J., Moens, M.-F., and Calixto, I. Are scene graphs good enough to improve image captioning? In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 504–515, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.aacl-main.50.

Miller, G. A. *WordNet: An electronic lexical database*. MIT press, 1998.

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., and Daumé III, H. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 747–756, Avignon, France, 2012. Association for Computational Linguistics. ISBN 978-1-937284-19-0. URL http://dl.acm.org/citation.cfm?id=2380816.2380907.

Murphy, E. and Ciszewska-Carr, J. Sources of difference in reliability: Identifying sources of difference in reliability in content analysis of online asynchronous discussions. *International Review of Research in Open and Distance Learning*, 6

(2):108–119, 2005. doi: 10.19173/irrodl.v6i2.233.

Muscat, A. and Belz, A. Learning to generate descriptions of visual data anchored in spatial relations. *IEEE Computational Intelligence Magazine*, 12(3):29–42, Aug 2017a. ISSN 1556-603X. doi: 10.1109/MCI.2017.2708559.

Muscat, A. and Belz, A. Generating descriptions of spatial relations between objects in images. In *ENLG 2015 - Proceedings of the 15th European Workshop on Natural Language Generation, 10-11 September 2015, University of Brighton, Brighton, UK*, pages 100–104, 2015. URL http://aclweb.org/anthology/W/W15/W15-4717.pdf.

Muscat, A. and Belz, A. Learning to generate descriptions of visual data anchored in spatial relations. *IEEE Computational Intelligence Magazine*, 12(3):29–42, Aug 2017b.

Muscat, A., Belz, A., and Birmingham, B. Exploring different preposition sets, models and feature sets in automatic generation of spatial image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, Berlin, Germany, August. Assosciation for Computational Linguistics*, pages 65–69, 2016.

Najibi, M., Rastegari, M., and Davis, L. S. G-cnn: an iterative grid based object detector. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2369–2377, 2016.

Nenkova, A. and Vanderwende, L. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*, 2005.

Nikolaus, M., Abdou, M., Lamm, M., Aralikatte, R., and Elliott, D. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1009. URL https://www.aclweb.org/anthology/K19-1009.

Oliva, A. and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, may 2001. ISSN 0920-5691. doi: 10.1023/A:1011139631724.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011.

Ordonez, V., Han, X., Kuznetsova, P., Kulkarni, G., Mitchell, M., Yamaguchi, K., Stratos, K., Goyal, A., Dodge, J., Mensch, A., et al. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision*, pages 1–14, 2015.

Papert, S. A. The summer vision project. 1966.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL http://www.aclweb.org/anthology/P02-1040.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. PMLR, 2013.

Patterson, G., Xu, C., Su, H., and Hays, J. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

Pauls, A. and Klein, D. Faster and smaller n-gram language models. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 258–267, 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), pages 1532–1543, Stroudsburg, PA, USA, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Perez-Cruz, F. Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE International Symposium on Information Theory*, pages 1666–1670, 2008. doi: 10.1109/ISIT.2008.4595271.

Peters, A. M. *The units of language acquisition*, volume 1. CUP Archive, 1983.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015.

Plummer, B. A., Mallya, A., Cervantes, C. M., Hockenmaier, J., and Lazebnik, S. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1928–1937, 2017.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., and Carin, L. Variational autoencoder for deep learning of images, labels and captions. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2360–2368, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL http://dl.acm.org/citation.cfm?id=3157096.3157360.

Quinlan, J. R. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Rahgooy, T., Manzoor, U., and Kordjamshidi, P. Visually guided spatial relation extraction from text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 788–794, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2124. URL https://www.aclweb.org/anthology/N18-2124.

Ramisa, A., Wang, J., Lu, Y., Dellandrea, E., Moreno-Noguer, F., and Gaizauskas, R. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/D15-1022.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. Sequence level training with recurrent neural networks. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06732.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 139–147, 2010.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. *Machine learning*, 85(3): 333–359, 2011.

Redmon, J. and Farhadi, A. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

Redmon, J. and Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Regier, T. and Carlson, L. A. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology General*, 130(2):273–298, 2001.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 91–99, Montreal, Canada, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969239.2969250.

Ren, Z., Wang, X., Zhang, N., Lv, X., and Li, L. Deep reinforcement learning-based image captioning with embedding reward. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1151–1159, 2017. doi: 10.1109/CVPR.2017.128. URL https://doi.org/10.1109/CVPR.2017.128.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. Self-critical sequence training for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, 2017.

Retz-Schmidt, G. Various views on spatial prepositions. *AI magazine*, 9(2):95–95, 1988.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021a.

Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., and Friedman, I. Tresnet: High performance gpu-dedicated architecture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1400–1409, January 2021b.

Ridnik, T., Sharir, G., Ben-Cohen, A., Ben-Baruch, E., and Noy, A. Ml-decoder: Scalable and versatile classification head. *arXiv preprint arXiv:2111.12933*, 2021c.

Robbins, H. and Monro, S. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Rosenfeld, A. Iterative methods in image analysis. *Pattern Recognition*, 10(3):181–187, June 1978.

Sadeghi, M. A. and Farhadi, A. Recognition using visual phrases. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1745–1752, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 9781457703942. doi: 10.1109/CVPR.2011.5995711. URL https://doi.org/10.1109/CVPR.2011.5995711.

Schrijver, A. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

Schuster, M. and Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11): 2673–2681, Nov 1997. ISSN 1053-587X. doi: 10.1109/78.650093.

Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., and Manning, C. D. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2812. URL https://www.aclweb.org/anthology/W15-2812.

Scott, B. I still see you boat!, 2013. URL https://external-preview.redd.it/vmVrdDTbwm8DZNHKqnnZ8n{_}MkM9eeSsDLbTg47p8HyY.jpg?width=960&crop=smart&auto=webp&s=b1af0aae22cfcf9e7c4b1c811230f64f95fca02b. [Online; accessed June 10, 2022].

Shatford, S. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly*, 6(3):39–62, 1986.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Skinner, B. F. Cognitive science and behaviourism. *British Journal of psychology*, 76(3):291–301, 1985.

Socher, R. and Fei-Fei, L. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 966–973. IEEE, 2010.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

Song, J., Gao, L., Nie, F., Shen, H. T., Yan, Y., and Sebe, N. Optimized graph learning using partial tags and multiple features for image and video annotation. *IEEE Transactions on Image Processing*, 25(11):4999–5011, 2016. doi: 10.1109/TIP.2016.2601260.

Speer, R., Chin, J., and Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press, 2017.

Stork, D. G. *HAL's Legacy: 2001's Computer as Dream and Reality*. Mit Press, 1997.

Sukhbaatar, S., szlam, a., Weston, J., and Fergus, R. End-to-end memory networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28, pages 2440–2448. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

Sutton, R. S. and Barto, A. G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4278–4284. AAAI Press, 2017.

Tanti, M. *On Architectures for Including Visual Information in Neural Language Models for Image Description*. PhD thesis, University of Malta, 2019.

Tanti, M., Gatt, A., and Camilleri, K. P. Where to put the image in an image caption generator. *Natural Language Engineering*, 24(3):467–489, 2018. doi: 10.1017/S1351324918000098.

Tellex, S., Gopalan, N., Kress-Gazit, H., and Matuszek, C. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55, 2020. doi: 10.1146/annurev-control-101119-071628.

Terry, R., Laura, C., and Bryce, C. *Attention in Spatial Language: Bridging Geometry and Function*, chapter 9, pages 191–204. Oxford University Press, Oxford, 2005.

Thelen, E. and Smith, L. B. *A dynamic systems approach to the development of cognition and action*. MIT press, 1996.

Thrun, S. Robotics and cognitive approaches to spatial mapping, chapter simultaneous localization and mapping, 2008.

Tomasello, M. The social-pragmatic theory of word learning. *Pragmatics*, 10(4):401–413, 2000.

Tomasello, M. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, 2003. ISBN 9780674010307. URL http://www.jstor.org/stable/j.ctv26070v8.

Tripathi, A., Srivastava, S., Lall, B., and Chaudhury, S. Using scene graphs for detecting visual relationships. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10074–10081, Los Alamitos, CA, USA, jan 2021a. IEEE Computer Society. doi: 10.1109/ICPR48806.2021.9412337. URL https://doi.ieeecomputersociety.org/10.1109/ICPR48806.2021.9412337.

Tripathi, S., Nguyen, K., Guha, T., Du, B., and Nguyen, T. Q. Sg2caps: Revisiting scene graphs for image captioning. *ArXiv*, abs/2102.04990, 2021b.

Tromp, E. and Pechenizkiy, M. Graph-based n-gram language identification on short texts. In *Proc. 20th Machine Learning*

conference of Belgium and The Netherlands, pages 27–34, 2011.

Tsoumakas, G. and Katakis, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, January 2007. ISSN 1548-3924. doi: 10.4018/jdwm.2007070101.

Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114, 1949. ISSN 0006341X, 15410420. doi: 10.2307/3001913.

Turing, A. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

Tversky, B. Cognitive maps, cognitive collages, and spatial mental models. In Frank, A. U. and Campari, I., editors, *Spatial Information Theory A Theoretical Basis for GIS*, pages 14–24, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg. ISBN 978-3-540-47966-6.

Ushiku, Y., Yamaguchi, M., Mukuta, Y., and Harada, T. Common subspace for model and similarity: Phrase learning for caption generation from images. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2668–2676, Dec 2015. doi: 10.1109/ICCV.2015.306.

Uwamariya, N. “An image of a blind man crossing the road." How do visually impaired people live in their surroundings?, 3 Oct 2019, 2019. URL https://www.kigalitoday.com/ubuzima/urusobe-rw-ubuzima/article/abafite-ubumuga-bwo-kutabona-babana-bate-n-ibibakikije. [Online; accessed June 8, 2020].

Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H.-J. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, 2001. doi: 10.1109/83.892448.

Van de Sande, K. E., Uijlings, J. R., Gevers, T., and Smeulders, A. W. Segmentation as selective search for object recognition. In *2011 international conference on computer vision*, pages 1879–1886. IEEE, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Vedantam, R., Zitnick, L., and Parikh, D. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087. URL http://dx.doi.org/10.1109/CVPR.2015.7299087.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph Attention Networks. *International Conference on Learning Representations*, 2018. accepted as poster.

Vetter, H. J. and Howell, R. W. Theories of language acquisition. *Journal of psycholinguistic research*, 1(1):31–64, 1971.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. pages 3156–3164, 2015. doi: 10.1109/CVPR.2015.7298935.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., and Wierstra, D. Matching networks for one shot learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3630–3638. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.

Vygotsky, L. Interaction between learning and development. *Readings on the development of children*, 23(3):34–41, 1978.

Wang, D., Beck, D., and Cohn, T. On the role of scene graphs in image captioning. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 29–34, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6405. URL https://www.aclweb.org/anthology/D19-6405.

Wang, H., Huang, H., and Ding, C. Image annotation using multi-label correlated green's function. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2029–2034, 2009. doi: 10.1109/ICCV.2009.5459447.

Wang, H., Huang, H., and Ding, C. Image annotation using bi-relational graph of images and semantic labels. In *CVPR 2011*,

pages 793–800, 2011. doi: 10.1109/CVPR.2011.5995379.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Wang, J. K., Yan, F., Aker, A., and Gaizauskas, R. A poodle or a dog? evaluating automatic image annotation using human descriptions at different levels of granularity. In *Proceedings of the Third Workshop on Vision and Language*, pages 38–45, 2014.

Wang, Q. and Chan, A. B. Describing like humans: On diversity in image captioning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4190–4198, 2019. doi: 10.1109/CVPR.2019.00432.

Wang, S., Wang, R., Yao, Z., Shan, S., and Chen, X. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1508–1517, 2020.

Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00061. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00061.

Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., Zhao, Y., and Yan, S. Hcp: A flexible cnn framework for multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1901–1907, 2016. doi: 10.1109/TPAMI.2015.2491929.

Wu, Q., Shen, C., Liu, L., Dick, A., and Van Den Hengel, A. What value do explicit high level concepts have in vision to language problems? In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–212, 2016. doi: 10.1109/CVPR.2016.29.

Xiong, C., Merity, S., and Socher, R. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2397–2406. JMLR.org, 2016.

Xu, D., Zhu, Y., Choy, C., and Fei-Fei, L. Scene graph generation by iterative message passing. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Xu, J., Jagadeesh, V., and Manjunath, B. S. Multi-label learning with fused multimodal bi-relational graph. *IEEE Transactions on Multimedia*, 16(2):403–412, 2014. doi: 10.1109/TMM.2013.2291218.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2048–2057. JMLR.org, 2015.

Xu, N., Liu, A.-A., Liu, J., Nie, W., and Su, Y. Scene graph captioner: Image captioning based on structural visual representation. *Journal of Visual Communication and Image Representation*, 58:477 – 485, 2019. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2018.12.027.

Xu, Y., Wei, H., Lin, M., Deng, Y., Sheng, K., Zhang, M., Tang, F., Dong, W., Huang, F., and Xu, C. Transformers in computational visual media: A survey. *Computational Visual Media*, 8(1):33–62, 2022.

Xue, X., Zhang, W., Zhang, J., Wu, B., Fan, J., and Lu, Y. Correlative multi-label multi-instance image annotation. In *2011 International Conference on Computer Vision*, pages 651–658, 2011. doi: 10.1109/ICCV.2011.6126300.

Yagcioglu, S., Erdem, E., Erdem, A., and Cakıcı, R. A distributed representation based query expansion approach for image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2015.

Yan, F. and Mikolajczyk, K. Deep correlation for matching images and text. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3441–3450, June 2015. doi: 10.1109/CVPR.2015.7298966.

Yang, J., Lu, J., Lee, S., Batra, D., and Parikh, D. Graph r-cnn for scene graph generation. In *Proceedings of the European*

Conference on Computer Vision (ECCV), September 2018.

Yang, S.-J., Jiang, Y., and Zhou, Z.-H. Multi-instance multi-label learning with weak label. In *IJCAI*, 2013.

Yang, X., Tang, K., Zhang, H., and Cai, J. Auto-encoding scene graphs for image captioning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10677–10686, 2019.

Yang, X., Zhang, H., and Cai, J. Auto-encoding and distilling scene graphs for image captioning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, dec 2020. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.3042192.

Yang, Y., Teo, C. L., Daumé III, H., and Aloimonos, Y. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454. Association for Computational Linguistics, 2011.

Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., and Salakhutdinov, R. R. Review networks for caption generation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2369–2377, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL http://dl.acm.org/citation.cfm?id=3157096.3157361.

Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T. Boosting image captioning with attributes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4904–4912, 2017.

Yao, T., Pan, Y., Li, Y., and Mei, T. Exploring visual relationship for image captioning. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 711–727, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01264-9.

Yatskar, M., Vanderwende, L., and Zettlemoyer, L. See no evil, say no evil: Description generation from densely labeled images. In *Proc. 3rd Joint Conference on Lexical and Computational Semantics*, pages 110–120, Dublin, Ireland, August 23-24, 2014.

Yeh, C.-K., Wu, W.-C., Ko, W.-J., and Wang, Y.-C. F. Learning deep latent spaces for multi-label classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 2838–2844. AAAI Press, 2017.

Yoo, D., Park, S., Lee, J.-Y., Paek, A. S., and So Kweon, I. Attentionnet: Aggregating weak directions for accurate object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2659–2667, 2015.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yu, J., Li, J., Yu, Z., and Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4467–4480, 2020. doi: 10.1109/TCSVT.2019.2947482.

Yu, R., Li, A., Morariu, V. I., and Davis, L. S. Visual relationship detection with internal and external linguistic knowledge distillation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1068–1076, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.121. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.121.

Yu, W.-J., Chen, Z.-D., Luo, X., Liu, W., and Xu, X.-S. Delta: A deep dual-stream network for multi-label image classification. *Pattern Recognition*, 91:322–331, 2019. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2019.03.006.

Yun, S., Choi, J., Yoo, Y., Yun, K., and Choi, J. Y. Action-decision networks for visual tracking with deep reinforcement learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1349–1358, July 2017. doi: 10.1109/CVPR.2017.148.

Zellers, R., Yatskar, M., Thomson, S., and Choi, Y. Neural motifs: Scene graph parsing with global context. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.

Zhang, J., Wu, Q., Shen, C., Zhang, J., and Lu, J. Multi-label image classification with regional latent semantic dependencies. *IEEE Transactions on Multimedia*, 20(10):2801–2813, 2018a. doi: 10.1109/TMM.2018.2812605.

Zhang, M. and Zhou, Z. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, Aug 2014. ISSN 1041-4347. doi: 10.1109/TKDE.2013.39.

Zhang, M.-L., Li, Y.-K., Liu, X.-Y., and Geng, X. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202, 2018b.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, June 2021a.

Zhang, Y., Shi, X., Mi, S., and Yang, X. Image captioning with transformer and knowledge graph. *Pattern Recognition Letters*, 143:43 – 49, 2021b. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2020.12.020.

Zhong, Y., Wang, L., Chen, J., Yu, D., and Li, Y. Comprehensive image captioning via scene graph decomposition. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, pages 211–229, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58568-6.

Zhou, C., Sun, C., Liu, Z., and Lau, F. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.

Zhou, L., Xu, C., Koch, P. A., and Corso, J. J. Image caption generation with text-conditional semantic attention. *CoRR*, abs/1606.04621, 2016.

Zhou, Y., Sun, Y., and Honavar, V. Improving image captioning by leveraging knowledge graphs. In *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019, pages 283–293, United States, March 2019. Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/WACV.2019.00036. 19th IEEE Winter Conference on Applications of Computer Vision, WACV 2019 ; Conference date: 07-01-2019 Through 11-01-2019.

Zhu, F., Li, H., Ouyang, W., Yu, N., and Wang, X. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017a.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017b.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=gZ9hCDWe6ke.