Title 1: Medical Information System

Supervisor: Abela Charlie

Subject Area: Machine Learning, Web Intelligence, Generative AI

Project Description: Demand for clinical decision support systems in medicine and self-diagnostic symptom checkers has substantially increased in recent years. Existing platforms rely on knowledge bases manually compiled through a labor-intensive process or automatically derived using simple pairwise statistics.

In this dissertation we'd like to investigate how to automatically or semi-automatically, learn a knowledge graph for healthcare by leveraging on multiple sources such as medical journals and medical imagery, through the use of NLP and Generative AI.

A possible use case for this knowledge graph is to serve as a tool for healthcare students to be able to learn about a medical condition by navigating through comprehensive information about the condition.

Example: Student is interested to learn more about "pulmonary edema". The tool will potentially present: known causes, symptoms, treatment (all information that can be extracted from different online sources), as well as images (x-rays or otherwise) that show how this condition manifests itself.

The system can be accessed both through a web interface as well as a mobile interface.

Recommended Resources:

Pre-requisite: ICS2205, ICS3206

Title 2: Automatic Bug Triage

Supervisor: Abela Charlie

Subject Area: Graph Neural Networks and Natural Language Processing

Project Description: Software development companies use bug repositories such as JIRA to manage software bugs. Developers and users create a report about the bug/s that is sent to the bug repository. This report describes the bug details and features related to the submitted bugs. The information includes reference to the component or product, versioning, priority, severity, assignee, bug creation date etc.

Bug reports should be resolved in a timely manner so that the quality of the software is maintained. Bug triage is the process of prioritising bugs based on their severity, frequency, and risk and to assign these to the appropriate develops for validation and resolution [1]. Performing this process manually can be laborious and therefore many automatic techniques have been used over the last decade.

Recent research on automated bug triage has focused on the use of a graph neural networks (GNN) since they allow for rich relational information to be learnt [2, 3].

In this project, the aim is to investigate the feasibility of graph-based models for bug triaging. A JIRA dataset which includes bug reports and other related information will be made available. The objectives include:

1. pre-processing of the bug related data
2. representing this data as a graph
3. the deployment and evaluation of different graph-based models

Recommended Resources: [1] I. Alazzam, A. Aleroud, Z. Al Latifah and G. Karabatis, "Automatic Bug Triage in Software Systems Using Graph Neighbourhood Relations for Feature Augmentation," in IEEE Transactions on Computational Social Systems, vol. 7, no. 5, pp. 1288-1303, Oct. 2020

[2] S. F. A. Zaidi and C. -G. Lee, "Learning Graph Representation of Bug Reports to Triage Bugs using Graph Convolution Network," 2021 International Conference on Information Networking (ICOIN), 2021, pp. 504-507

[3] S. Fang, Y. -s. Tan, T. Zhang, Z. Xu and H. Liu, "Effective Prediction of Bug-Fixing Priority via Weighted Graph Convolutional Networks," in IEEE Transactions on Reliability, vol. 70, no. 2, pp. 563-574, June 2021

Pre-requisite: ICS2205 and ICS3206

Title 3: Link Prediction using Graph Neural Networks

Supervisor: Abela Charlie

Subject Area: Machine Learning Graph Neural Networks

Project Description: Link prediction methods try to predict the likelihood of a future connection between two nodes in a given network. This approach can be used in biological networks to infer protein-protein interactions or to suggest possible friends to a user in an online social network. Due to the enormous amounts of data that is collected today, there is a need for scalable approaches to this problem.

One of the main issues with real-world networks however is their extreme class skewness. For instance, in a social network such as Facebook, although the number of users exceeds 2 billion, the average user will in general only connect to about 100 nodes in the social graph. This leads to a significantly high, class imbalance between existing links and missing links. If an algorithm has to consider all the missing links to perform link prediction, this would be computationally infeasible.

A long-term goal for link prediction could be that of eventually replacing search with some reliable intelligent assistant (similar to Cortana or Siri) that would provide recommendations based on the user's preferences and behavior.

The overall aim of this dissertation is to investigate the performance of both supervised as well as unsupervised link prediction approaches (separately or combined) using a number of networks (such as Facebook, DBLP or DrugBank). A dashboard style link-prediction module can be developed to analyze the performance of the algorithms and to visualize the results.

Recommended Resources: 1. Darcy Davis, Ryan Lichtenwalter, and Nitesh V. Chawla. "Supervised methods for multi-relational link prediction". In: Social Network Analysis and Mining (2012), pp. 127–141. issn: 1869-5450.

2. Mohammad Al Hasan and Mohammed J. Zaki. "A Survey of Link Prediction in Social Networks". In: Social Network Data Analytics. Ed. by Charu C Aggarwal. Springer US, 2011, pp. 243–275.

3. Yang Yang, Ryan N. Lichtenwalter, and Nitesh V. Chawla. "Evaluating link prediction methods". In: Knowledge and Information Systems (2014). issn: 0219-1377.

4. Farrugia Lizzy, Azzopardi M. Lilian, Debattista, Jeremy, Abela Charlie. "Predicting Drug-Drug Interactions Using Knowledge Graphs, submitted to the 10th International Conference on Artificial Intelligence and Applications (AI 2024)

Pre-requisite: ICS2205 and ICS3206

Title 4: WYSIWYT: A rapid development platform for ML pipelines.

Supervisor: Bajada Josef

Subject Area: Machine Learning

Project Description: Building a machine learning pipeline involves copying and pasting a lot of boilerplate code, which is often similar, but tedious to get right due to the requirement of having matching input sizes and configuration parameters. Taking inspiration from visual tools such as Tensorflow Playground, Visual Blocks for ML, and Blockly, together with data pipeline tools such as Azure Data Factory, Apache Nifi, and Apache AirFlow, this project will explore ideas to guide a Machine Learning engineer to build a machine learning pipeline through a low-code/no-code visual user interface. The WYSIWYT (What you see is what you train) platform will automatically generate the python code, compatible with modern APIs such as Pytorch or Keras, that works out of the box. The user will be able to drag and drop datasets into the WYSIWYT tool, through which it will automatically detect the columns and give the option to the user to select which columns to use as features and join data based on foreign keys (generating the necessary dataframe manipulation code automatically). Applying preprocessing such as dimensionality reduction, feature normalization/standardization, and imputation will also be possible. The user should be given the option to determine how the model will be validated (train/test/validation split ratios, k-fold cross validation etc.). The user will then be able to select a model architecture, such as a fully connected neural network, a CNN, an RNN or a Transformer, giving the user the option to configure layer sizes, activation functions and other hyperparameters specific to the chosen model architecture. Tools such as the Jinja Templating Engine can be used to generate the necessary python code from templates. The user should then be able to press a button to execute the pipeline and see the results, displaying the standard metrics such as Accuracy/Precision/Recall/F1 Score and confusion matrices for classification tasks, mean squared error/mean absolute error for regression tasks. While the focus will be on supervised ML pipelines, unsupervised ML tasks can also be considered.

Recommended Resources: García-Peñalvo, Francisco, Andrea Vázquez-Ingelmo, Alicia García-Holgado, Jesús Sampedro-Gómez, Antonio Sánchez-Puente, Víctor Vicente-Palacios, P. Ignacio Dorado-Díaz, and Pedro L. Sánchez. "KoopaML: a graphical platform for building machine learning pipelines adapted to health professionals." (2023).

Frank, E. et al. (2009). Weka-A Machine Learning Workbench for Data Mining. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA.

https://playground.tensorflow.org/

https://visualblocks.withgoogle.com/

https://developers.google.com/blockly

https://jinja.palletsprojects.com/en/3.1.x/

Pre-requisite: Python, Machine Learning

Title 5: Next-Basket Prediction System

Supervisor: Bajada Josef

Subject Area: Machine Learning, Recommender Systems

Project Description: Next-basket Prediction involves predicting the items in a user's next set of purchases, referred to as a basket. This is especially useful in recurrent retail settings such as groceries and supermarkets. Predicting the items a user is likely to buy next can provide various opportunities, such as personalized offers and automatic basket recommendations.

In this project we will explore various techniques to predict the items a user is likely to buy next from the previous history of the user and other users with similar buyi ng patterns. We will investigate techniques such as Recurrent Neural Networks (RNNs), Collaborative Filtering and Time-Series Analysis. We will have a special focus on grocery shopping, since this has a high element of repeat item purchasing. We will analyse one of more datasets such as the Instacart dataset or the Acquire Valued Shoppers Challenge dataset, and determine which Machine Learning techniques are most effective for this problem.

Recommended Resources: Mozhdeh Ariannezhad, Sami Jullien, Ming Li, Min Fang, Sebastian Schelter, and Maarten de Rijke. 2022. "ReCANet: A Repeat Consumption-Aware Neural Network for Next Basket Recommendation in Grocery Shopping." In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 1240–1250. https://doi.org/10.1145/3477495.3531708

Ming Li, Sami Jullien, Mozhdeh Ariannezhad, and Maarten de Rijke. 2023. "A Next Basket Recommendation Reality Check." ACM Trans. Inf. Syst. 41, 4, Article 116 (October 2023), 29 pages. https://doi.org/10.1145/3587153

Hu, Haoji, et al. "Modeling personalized item frequency information for next-basket recommendation." Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. 2020.

Faggioli, G., Polato, M., & Aiolli, F. (2020, July). Recency aware collaborative filtering for next basket recommendation. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (pp. 80-87).

Pre-requisite: Machine Learning, Python

Title 6: Reinforcement Learning for Air Traffic Control in Busy Airports

Supervisor: Bajada Josef

Subject Area: Reinforcement Learning

Project Description: Air Traffic control is a challenging task for humans, as air traffic i ncreases and the success of the task depends on the operators being continuously alert and aware of all the paths and possible actions that could be taken by each aircraft, taking into consideration its speed, direction, altitude, fuel levels, target waypoint, and aircraft characteristics. Having an AI-assistant suggesting or even taking over part of the air-traffic control tasks can alleviate this pressure and decrease the risk of human error.

In this project we will use a high-fidelity simulation, such as the BlueSky Open Air Traffic Simulator, to generate realistic scenarios, and investigate the effect of applying AI techn iques, focusing on but not limited to reinforcement learning, to guarantee collision-free outcomes and assist aircraft in reaching their target destinations.

Recommended Resources: https://www.eurocontrol.int/sites/default/files/2019-10/tim-oct-2019-modelling-simulation-08-hoekstra.pdf

Hoekstra, J. M., & Ellerbroek, J. (2016, June). Bluesky ATC simulator project: an open data and open source approach. In Proceedings of the 7th international conference on research in air transportation (Vol. 131, p. 132). Washington, DC, USA: FAA/Eurocontrol.

Brittain, M., & Wei, P. (2019). Autonomous air traffic controller: A deep multi-agent reinforcement learning approach. arXiv preprint arXiv:1905.01303.

Wang, Zhuang, et al. "Review of deep reinforcement learning approaches for conflict resolution in air traffic control." Aerospace 9.6 (2022): 294.

Deniz, S., & Wang, Z. (2022). A multi-agent reinforcement learning approach to traffic control at future urban air mobility intersections. In AIAA SCITECH 2022 Forum (p. 1509).

Z. Mahboubi and M. J. Kochenderfer, "Continous time autonomous air traffic control for non-towered airports," 2015 54th IEEE Conference on Decision and Control (CDC), Osaka, Japan, 2015, pp. 3433-3438.

Passerini, A., & Schiex, T. (2022, August). Automating the resolution of flight conflicts: Deep reinforcement learning in service of air traffic controllers. In PAIS 2022: 11th Con ference on Prestigious Applications of Artificial Intelligence, 25 July 2022, Vienna, Austria (co-located with IJCAI-ECAI 2022) (Vol. 351, p. 72). IOS Press.

Ghosh, S., Laguna, S., Lim, S. H., Wynter, L., & Poonawala, H. (2021). A Deep Ensemble Method for Multi-Agent Reinforcement Learning: A Case Study on Air Traffic Control. Proceedings of the International Conference on Automated Planning and Scheduling, 31(1), 468-476.

Pre-requisite: Reinforcement Learning, Machine Learning, Python

Title 7: A Multilingual Approach to the Digitialisation of Maltese Historic Manuscripts

Supervisor: Borg Claudia

Subject Area: NLP, Handwriting Recognition

Project Description: Handwriting recognition can be seen as one of the first image to text tasks - a research area that began in the early 80s. However, whilst clean handwriting is easy to automatically transcribe and digitalise, older handwritten documents present a more challenging task.

This project will specifically look at old manuscripts written in Maltese. The main aim of this thesis is to collaborate with the Library at the University of Malta and build a model that is able to assist librarians in the digitalisation process of manuscripts in its archive.

Specifically, we will look at improving character recognition models by supplementing knowledge from the Maltese BERT-based language model. Experiments will look at possibly taking a multilingual approach into the character recognition model but then specifically improve the performance through the language model.

Recommended Resources: https://aclanthology.org/2020.emnlp-main.478.pdf

https://aclanthology.org/2021.wnut-1.31.pdf

Pre-requisite: ICS2203/ARI2203, NLP, ML

Title 8: Exploring Maltese Language Models in the domain of Mental Health

Supervisor: Borg Claudia

Subject Area: NLP, LLMs, Chatbot

Project Description: This project will first explore the current state of the art for Chatbots in Mental Health in English. The research will then layout a path forward for transposing this capability to the Maltese language. The project will aim towards creating the necessary building blocks for a chatbot in the mental health domain that is capable of conversing in Maltese. The output of the chatbot will be evaluated by at least one mental health professional. The main focus of the project is the creation of the datasets and the building and training of the models. The evaluation will be carried out with mental health professionals rather than potential end users.

Recommended Resources: https://www.jmir.org/2023/1/e46448/PDF

https://mhealth.jmir.org/2023/1/e44838/PDF

Pre-requisite: ICS2203/ARI2203, NLP, ML.

Title 9: Maltese Spell and Grammar Correction

Supervisor: Borg Claudia

Subject Area: NLP

Project Description: This project will look at our continued efforts to improve our spellchecking engine (currently works-in-progress) for Maltese. This is not a simple task primarily due to the lack of spell checking data and because a huge number of documents/articles, etc., are not written in 100% correctly written Maltese. Moreover, there is a stronger issue of grammatical agreement in Maltese - *Jien taqbeż should be corrected to Jien naqbeż so that the verb can agree with the person, or Int/Hija taqbeż so that the person can agree with the verb. These cases highlight the difficulty that spell checking for Maltese is challenging and why a simple wordlist approach will not work.

The thesis will explore approaches used in Machine Translation and how these are deployed to error correction. Moreover, how these models can be supplemented further with additional data and BERT-based language models such as BERTu.

Recommended Resources:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9597183

https://ieeexplore.ieee.org/document/8228160

Pre-requisite: ICS2203/ARI2203, NLP, ML.

Title 10: Virtual Reality Swim Training Environment with Generative AI Coaches

Supervisor: Camilleri Vanessa

Subject Area: VR, Sport, Virtual Coaching

Project Description: Create a VR environment for swim training. Users will use their arms and legs (tracked by controllers) to simulate swimming strokes. The VR environment will provide real-time visual feedback on the user's stroke technique, similar to what a swimmer sees underwater. Utilise generative AI to create virtual coaches that provide real-time personalised feedback, adapting their guidance based on the swimmer's performance and learning style. Implement interactive elements like virtual competitors or underwater challenges to enhance engagement.

Analyse the ethical implications of AI-generated coaches and ensure their guidance is unbiased, transparent, and adheres to established coaching principles. Explore how to balance personalised feedback with the importance of human connection in sports coaching.

Recommended Resources:

Pre-requisite:

Title 11: Stroke Style Generation and Optimisation using Generative Adversarial Networks (GANs)

Supervisor: Camilleri Vanessa

Subject Area: Sport, GANs

Project Description: Train a GAN model on a swimming (sport) dataset to generate realistic swim stroke sequences for different styles. Use the GAN to optimise stroke mechanics by exploring variations within the generated sequences, identifying patterns associated with improved efficiency or speed.

Analyse the theoretical underpinnings of GANs for generating and manipulating sports motion data. Investigate the ethical considerations surrounding the use of GANs in sports, such as the potential for creating unrealistic or unfair performance advantages. Discuss the implications of AI-generated stroke styles on the future of swimming technique and training.

Recommended Resources:

Pre-requisite:

Title 12: Ethical Framework for AI in Sports Performance Enhancement

Supervisor: Camilleri Vanessa

Subject Area: Ethics, AI Analytics, Sport

Project Description: Design and implement an AI-powered tool that assesses the ethical implications of different AI interventions in swim training (e.g., bio-mechanical feedback, VR environments, generative AI coaches). This tool could serve as a checklist or decision-making aid for coaches and athletes. Conduct a comprehensive review of ethical frameworks relevant to AI in sports. Analyse the potential benefits and risks of AI-driven performance enhancement, considering issues like fairness, equity, and the potential for misuse. Propose an ethics framework specific to swimming, incorporating insights from stakeholders in the swimming community.

Recommended Resources:

Pre-requisite:

Title 13: LLM Rescoring Methods for Maltese ASR

Supervisor: De Marco Andrea

Subject Area: Speech and Language Processing

Project Description: This project will investigate the use of LLMs fine-tuned for Maltese, to generate N-best rescoring for Maltese ASR systems. ASR systems typically improve once the acoustic model is augmented with a language model e.g. n-gram based language models. However, for low-resource languages such as Maltese, word error rates remain relatively high. We can attempt to improve on this process by utilising LLMs for Maltese (for which quite a voluminous amount of data is available), in order to build better language models to boost ASR word error rates. The concept will be to utilise a wide-context LLM to assess the top-N candidates proposed by the ASR decoder, and utilising the power of a Maltese LLM model to pick the best candidate. This architecture (and various LLM models) will be assesses and compared to a baseline Maltese ASR system, as well as one with a standard language model.

Recommended Resources:

Pre-requisite: Good knowledge of ML toolkits (PyTorch or Tensorflow) and Speech/NLP concepts.

Title 14: Maltese Speech Meeting Summarization

Supervisor: De Marco Andrea

Subject Area: Speech and Language Processing

Project Description: The aim of this project is to build a Maltese speech summarization system for a multi-speaker scenario. The pipeline will be able to identify different speakers in a sound track, transcribe each intervention in the track, and be able to tag salient points or provide a brief summary of the meeting. This will require the use of various subsystems for the Maltese language, including a fine-tuned LLM and ASR system for the Maltese language, as well as other non-language specific modules for speaker separation.

Recommended Resources:

Pre-requisite: Good knowledge of ML toolkits (PyTorch or Tensorflow) and Speech/NLP concepts.

Title 15: Assessing TTS Architectures for Low-Resource Multi-Speaker Corpora

Supervisor: De Marco Andrea

Subject Area: Speech and Language Processing

Project Description: The aim of this project will be to assess various state of the art text to speech (TTS) systems and their performance when adapted/fine-tuned for a low-resource language (Maltese). In particular the focus will be to look into the use of low-quantity and multi-speaker training data setups, as opposed to the more traditional single-speaker training sets used for TTS systems. Possibly, enhancements to the TTS model architectures, or the use of neural voice conversion systems will be evaluated.

Recommended Resources:

Pre-requisite: Good knowledge of ML toolkits (PyTorch or Tensorflow) and Speech/NLP concepts.

Title 16: Virtual Reality Mirror Therapy

Supervisor: Dingli Alexiei

Subject Area: VR Avatars, VR Environment, Interactive Features, Real-Time Feedback and Analysis

Project Description: Virtual Reality Mirror Therapy (VRMT) utilises immersive VR technology to create a simulated environment where patients can engage with a virtual representation of themselves. This innovative approach combines traditional mirror therapy techniques with the latest VR technology, enabling personalised and controlled therapeutic sessions.

Key Components:

VR Environment Creation: Tailored virtual scenarios are developed to reflect patients' personal experiences and body image concerns. These scenarios can include customisable avatars that closely mimic the patient's appearance and can be adjusted to simulate body transformation or other relevant visual cues.

Interactive Features: The therapy includes interactive elements where patients can engage with their virtual representation. These activities can involve challenging negative thoughts about body image and promoting a healthier perception of one's body .

Real-Time Feedback and Adaptation: Advanced algorithms monitor the patient's responses (e.g., physiological and emotional reactions) and adapt the virtual environment in real-time to optimise the therapeutic effect. This might include adjusting the difficulty level of tasks or changing the environment to ensure the patient remains engaged but not overwhelmed.

System Infrastructure:

The VR systems used may include commercially available VR headsets, which have made VR technology more accessible and affordable.

Recommended Resources: Online

Pre-requisite: Unity Development, Machine Learning

Title 17: INSIGHT - Intelligent Navigation System for Integrated Guidance Haptic Technologies

Supervisor: Dingli Alexiei

Subject Area: Human-Computer Interaction and Large Language Models

Project Description: This project aims to explore and develop new user interfaces on mobile devices for visually impaired individuals by leveraging multimodal large language models (LLMs). The system will integrate visual, auditory, and tactile feedback to interpret the surrounding environment, recognize faces, and assist with navigation. By processing real-time data from the mobile device's camera and other sensors, the application will provide spoken feedback, vibrations, and other tactile cues to help the user understand their immediate surroundings and interact more effectively with the world.

Recommended Resources: Various online sources

Pre-requisite: Familiarity with computer vision and natural language processing is an asset.

Understanding of mobile application development.

Basic knowledge of user interface design principles.

Title 18: Mind to Message: Textual Decoding of EEG Patterns

Supervisor: Dingli Alexiei

Subject Area: BCI

Project Description: A number of individuals suffer from impairments limiting their ability to communicate with others. With communication being such an important part of our lives, these individuals are significantly disadvantaged when compared to their peers. This project aims to develop a Brain Computer Interface (BCI) capable of translating EEG patterns into text through the use of machine learning techniques applied to data gathered by commercially available EEG bands. Additionally, part of the aim is to create a generalisable BCI which requires minimal calibration when used by individuals not part of the original training set.

This project will require an in depth look at current literature covering EEG-based BCIs and current methods used to improve the generalisability of BCIs.

Data will be collected from a number of participants performing tasks such as thinking of words or phrases. This data will then need to be cleaned and pre-processed before being used to train any model.

A model such as a transformer will be developed to decode the EEG signals. The model will be evaluated on its accuracy and generalisability.

By the end of this project a BCI that can accurately translate EEG signals into text for an individual, regardless of whether they were part of the training set should be produced, opening up new possibilities for communication.

Recommended Resources: Available Online

Title 19: Deciphering Human Population Structure and Migration Patterns from Genomic Variants

Supervisor: Galea Ingrid

Subject Area: Bioinformatics

Project Description: This project aims to analyze genetic variations across different human populations represented in the 1000 Genomes Project. By examining allele frequencies and genetic markers, this project will identify patterns that reveal historical migration paths and the structure of various populations. Techniques such as Principal Component Analysis (PCA), clustering, and other statistical methods will be used to interpret genetic data and infer relationships between populations.

Recommended Resources: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/

Pre-requisite: Proficiency in using Python for data analysis

Title 20: Investigating the Impact and Distribution of Rare Genetic Variants in Human Populations

Supervisor: Galea Ingrid

Subject Area: Bioinformatics

Project Description: This project aims to explore the distribution and potential functional impacts of rare genetic variants across various human populations, utilizing the Variant Call Format (VCF) files from the 1000 Genomes Project. This project will analyze the frequency and types of rare variants, examine their distribution across different ethnic groups, and potentially investigate associations with phenotypic traits or susceptibility to diseases. Techniques such as filtering, statistical analysis, and bioinformatic tools will be used to identify and categorize these variants.

Recommended Resources: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/

Pre-requisite: Proficiency in using Python for data analysis

Title 21: Comparative Study of Linkage Disequilibrium Patterns Across Human Populations

Supervisor: Galea Ingrid

Subject Area: Bioinformatics

Project Description: This project will explore and compare the patterns of linkage disequilibrium (LD) across different human populations using the VCF files from the 1000 Genomes Project. It will examine how linkage disequilibrium varies geographically and ethnically, providing insights into genetic diversity, population history, and the mechanism of genetic recombination. Techniques such as calculating LD statistics, generating LD plots, and statistical analysis to interpret geographical or population-specific differences will be used.

Recommended Resources: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/

Pre-requisite: Proficiency in using Python for data analysis

Title 22: An Investigation of Context-Free Language Learning Techniques

Supervisor: Guillaumier Kristian

Subject Area: Machine Learning, Grammatical Inference, Search Algorithms, Heuristics, Algorithm Design

Project Description: Grammatical inference is the task of learning a formal grammar from strings which belong to a language and strings which do not belong to a language. This inference task has many real-world applications including robotics, data mining, structural pattern recognition, speech recognition, and bioinformatics.

In this project, we are concerned with surveying and implementing techniques used in the inference of context-free languages (CFL) from structured data. The primary aims are to:

- Survey the literature regarding current state-of-the-art algorithms.
- Investigate practical applications of CFL inference.
- Build a simple framework consisting of foundational data structures and CFL learning algorithms.
- Verify the behaviour of our implementations on Omphalos-style problems.

Recommended Resources: Material and tutorials will be provided to the student to support the FYP

Pre-requisite: Data Structures and Algorithms 1 & 2, Machine Learning 1 & 2

Title 23: Extracting DFAs from trained Neural Networks

Supervisor: Guillaumier Kristian

Subject Area: Machine Learning, Grammatical Inference, Search Algorithms, Heuristics, Algorithm Design

Project Description: Grammatical inference is the task of learning a formal grammar from strings that belong to the language and strings that do not. In this project, we will focus on the identification of regular languages (as Deterministic Finite State Automata, DFAs) from training sets consisting of both positive and negative examples. This inference task has many real-world applications including robotics, data mining, structural pattern recognition, speech recognition, and bioinformatics.

While state merging algorithms are one of the most powerful and successful approaches in DFA learning, there has been increasing interest in exploring connectionist approaches to the problem (typically RNNs, LSTMs, and transformers). In this FYP, we aim to study how DFAs representations can be extracted from trained neural networks, what their advantages and limitations are, and compare their behaviour to state merging approaches.

Recommended Resources: Material and tutorials will be provided

Pre-requisite: Data Structures and Algorithms 1 & 2, Machine Learning 1 & 2

Title 24: Grammatical Inference Applications in Real-World Applications

Supervisor: Guillaumier Kristian

Subject Area: Machine Learning, Grammatical Inference, Search Algorithms, Heuristics, Algorithm Design

Project Description: Grammatical inference is the task of learning a formal grammar from strings which belong to a language and strings which do not belong to a language. This inference task has many real-world applications including robotics, data mining, structural pattern recognition, speech recognition, and bioinformatics.

In this project, we are concerned with surveying grammatical inference approaches used in such real-world applications. The primary aims are:

- To focus on DFA learning.

- Surveying the literature including current state-of-the-art techniques.

- Implement algorithms and apply them to one or more practical applications and study their behaviour.

Recommended Resources: Material and tutorials will be provided to the student to support the FYP

Pre-requisite: Data Structures and Algorithms 1 & 2, Machine Learning 1 & 2

Title 25: Leveraging Tensor Neural Networks Towards Hyperspectral Data Classification

Supervisor: Makantasis Konstantinos

Subject Area: Remote Sensing, Machine learning

Project Description: An increasing number of emerging applications in data science and engineering are based on multidimensional and structurally rich data. A very common example of such data is images captured by hyperspectral cameras installed on satellites. The irregularities, however, of high-dimensional data often compromise the effectiveness of standard machine learning algorithms. This final-year project aims to investigate the effectiveness of tensor-based neural networks on the classification of hyperspectral data. Tensor-based neural network models impose tensor decomposition on their parameters, thereby offering two core advantages compared to typical machine learning methods. First, they handle inputs as multilinear arrays, bypassing the need for inputs in vector format, and can thus fully exploit the structural information along every data dimension. Moreover, the number of the models' parameters that need to be estimated during training is substantially reduced, making them very efficient for small sample setting problems where the number of labelled examples is limited. The implemented tensor-based neural network will be evaluated and compared against state-of-the-art in terms of computational efficiency, sample complexity, and recognition accuracy using publicly available hyperspectral datasets (Indian Pines, Salinas, Pavia Centre, Pavia University, Kennedy Space Center and Botswana). The outputs of this study will be i) a thorough evaluation of tensor-based neural networks in the framework of hyperspectral data classification and ii) an open-access code repository for downloading the implemented algorithms.

Recommended Resources: [1] Li, Xiaoshan, et al. "Tucker tensor regression and neuroimaging analysis." Statistics in Biosciences 10.3 (2018): 520-545.

[2] Makantasis, Konstantinos, et al. "Tensor-based nonlinear classifier for high-order data analysis." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[3] Makantasis, Konstantinos, et al. "Tensor-based classification models for hyperspectral data analysis." IEEE Transactions on Geoscience and Remote Sensing 56.12 (2018): 6884-6898.

[4] Makantasis, Konstantinos, et al. "Deep supervised learning for hyperspectral data classification through convolutional neural networks." 2015 IEEE international geoscience and remote sensing symposium (IGARSS). IEEE, 2015.

Pre-requisite: Python

Title 26: Semi-supervised learning on tensor embeddings for hyperspectral data classification

Supervisor: Makantasis Konstantinos

Subject Area: Remote Sensing, Machine learning

Project Description: Hyperspectral data classification is one of the fundamental problems in remote sensing. Several algorithms based on supervised machine learning have been proposed to address it. The performance, however, of the proposed algorithms is inherently dependent on the amount and quality of annotated data. Due to recent advances in hyperspectral imaging and autonomous (unmanned) aerial vehicles, collecting new hyperspectral data is an easy task. Annotating those data, however, is a tedious, time-consuming and costly task requiring the in-situ presence of human experts. One way to loosen the requirement of a large amount of annotated data is to shift to semi-supervised learning combined with highly sample-efficient tensor-based neural network embeddings. This project aims to reframe hyperspectral data classification by investigating the degree to which knowledge from unlabelled data can boost the performance of classifiers. Towards this direction, this study will leverage tools from the semi-supervised learning paradigm. Semi-supervised concepts will be implemented to inject knowledge about the distribution of unlabelled data points into the classification models. The implemented algorithms will be evaluated against fully supervised learning models regarding prediction accuracy, sample complexity and computational efficiency using publicly available hyperspectral datasets (Indian Pines, Salinas, Pavia Centre, Pavia University, Kennedy Space Center and Botswana). The outputs of this study will be i) a thorough evaluation of semi-supervised algorithms in the framework of hyperspectral data classification and ii) an open-access code repository for downloading the implemented methodologies.

Recommended Resources: [1] Georgoulas, Ioannis, et al. "Graph-based semi-supervised learning with tensor embeddings for hyperspectral data classification." IEEE Access (2023).

[2] Makantasis, Konstantinos, et al. "Tensor-based nonlinear classifier for high-order data analysis." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[3] Makantasis, Konstantinos, et al. "Tensor-based classification models for hyperspectral data analysis." IEEE Transactions on Geoscience and Remote Sensing 56.12 (2018): 6884-6898.

Pre-requisite: Python

Title 27: Learning representations of hyperspectral data for pixel-wise image classification

Supervisor: Makantasis Konstantinos

Subject Area: Remote Sensing, Machine learning

Project Description: Conventional representation learning methods excel in optimising encoders for discriminative tasks. However, in scenarios where only a few labelled samples are available, like in hyperspectral data classification, they struggle to eliminate the inductive bias when transferring from source to target classes. This is a by-product (and inherent limitation) of their underlying optimisation process that involves training a model to maximise class separation without optimising for within-class cohesion. This project aims to investigate the effectiveness of the recently proposed Silhouette Distance loss as the representation learning objective on pixel-wise hyperspectral image classification tasks. Minimising Silhouette Distance loss aims to enhance the quality of learned representations by emphasising both the cohesion and separation of representation clusters for each class. The developed methodologies will be tested against state-of-the-art models in terms of prediction accuracy, sample complexity and computational efficiency using publicly available remote sensing corpora (Indian Pines, Salinas, Pavia Centre, Pavia University, Kennedy Space Center and Botswana). The outputs of this study will be i) a thorough evaluation of Silhouette Distance loss representation learning algorithms for pixel-wise hyperspectral data classification and ii) an open-access code repository for downloading the implemented methodologies.

Recommended Resources: [1] Jaiswal, Ashish, et al. "A survey on contrastive self-supervised learning." Technologies 9.1 (2020): 2.

[2] Khosla, Prannay, et al. "Supervised contrastive learning." Advances in neural information processing systems 33 (2020): 18661-18673.

Pre-requisite: Python

Title 28: VWLE … the virtual classroom

Supervisor: Montebello Matthew

Subject Area: VR, AI in Education

Project Description: Implementing a virtual classroom in VR working on numerous projects that have been implementing the foundation platform to employ the medium effectively to simulate a learning environment.

Recommended Resources:

Pre-requisite:

Title 29: ECHO – Enhanced Cloning for Higher Online education

Supervisor: Montebello Matthew

Subject Area: Generative AI, AI in Education

Project Description: The ECHO project is intended to develop next generation e-learning content employing cutting-edge R&D aimed at revolutionizing the e-learning industry by automating the creation of high-quality educational content. Leveraging state-of-the-art artificial intelligence (AI) and deep audio-visual cloning technology, this project seeks to generate e-learning courses from scratch, complete with simulated real lecturer faces and voices that promises to reshape the future of online education. The incorporation of deep audio-visual technology is a pivotal element, aiming to simulate a real lecturer's facial expressions and voice. The enhanced learner engagement and immersion will provide highly effective e-learning content complemented with facilities to generate smart educational programmes through the employment of natural language processing and machine learning.

Recommended Resources:

Pre-requisite:

Title 30: WAVE - Water-Rescue training and Artificial Intelligence within a Virtual Reality Environment

Supervisor: Montebello Matthew

Subject Area: VR, AI in Training, Generative AI.

Project Description: The project WAVE initiative is a cutting-edge training solution developed to enhance the skills of water-rescue personnel. By leveraging advancements in Virtual Reality (VR) and Artificial Intelligence (AI), this project aims to provide realistic, immersive training experiences that are crucial in preparing rescuers for the complexities of real-life water emergencies. The key opportunity identified by this project lies in addressing the existing gaps in traditional water-rescue training methods. Current training programs often lack the variability and intensity of real-world scenarios, limiting the preparedness of rescue personnel. Additionally, there is a need for more targeted and adaptive training methods that can respond to the unique challenges presented by different water-rescue situations.

Recommended Resources:

Pre-requisite:

Title 31: A Multimodal AI Approach to Analysing Gender in Maltese Online News

Supervisor: Seychell Dylan

Subject Area: AI Media Analysis

Project Description: This project explores how artificial intelligence can be used to analyse gender representation in Maltese online news. Using a combination of computer vision and natural language processing techniques, the project examines both visual and textual portrayals of men and women.

Computer vision models will identify and classify people in news images, analysing factors like gender distribution and pose to reveal potential biases. Meanwhile, natural language processing techniques will quantify how often men and women are mentioned in news articles and assess the associated sentiment.

By comparing the findings from both analyses, this project aims to understand how gender is represented in Maltese online news comprehensively. The insights generated will be valuable for media organisations, policymakers, and the public, contributing to a more informed and equitable media landscape in Malta.

Recommended Resources: Datasets, access to experts and journalists, and any relevant literature will be provided by the supervisor.

Pre-requisite: The student needs to follow the following courses:

- ARI3129 - Advanced Computer Vision for AI

- ARI3900 - Ethics and Artificial Intelligence

Title 32: AI Methods for Assessing Crime Reporting

Supervisor: Seychell Dylan

Subject Area: AI Media Analysis

Project Description: This project aims to explore and experiment with automated methods for extracting crime data from Maltese online news sources. Using web scraping, natural language processing (NLP) and Computer Vision (CV) techniques, the project will develop a system to identify and categorise crime-related articles, extracting key information such as crime types, locations, and dates. The project will investigate simple image similarity techniques to analyse images accompanying news articles, potentially revealing patterns in visual representation.

The aim is to create a tool that measures and visualises the overlap between media-reported crimes and those documented in official police reports. This will shed light on potential biases and gaps in crime reporting. The project's output will provide insights into the complexities of crime reporting and contribute to a better understanding of public perception versus official statistics.

Recommended Resources: Datasets, access to experts and journalists, and any relevant literature will be provided by the supervisor.

Pre-requisite: The student needs to follow the following courses:

- ARI3129 - Advanced Computer Vision for AI

- ARI3900 - Ethics and Artificial Intelligence

Title 33: Accurate Name Extraction from News Video Graphics

Supervisor: Seychell Dylan

Subject Area: AI Media Analysis

Project Description: This project tackles the challenge of extracting names from the diverse graphical elements (captions, lower thirds) used in news videos. Due to variations in fonts, styles, and placements across different news channels, a robust machine learning system is needed.

The project's key objectives are classifying graphic types, accurately locating name regions, extracting text using OCR (Optical Character Recognition), and identifying names through NER (Named Entity Recognition). This involves developing a machine learning model to categorise graphics, pinpoint regions of interest, apply OCR to extract text, and then use NER to identify the specific name. Error handling and validation mechanisms will be implemented to improve accuracy. An important part of this project will be designing and creating a dataset that enables machine learning methods to carry out such tasks and evaluate them accordingly.

This project aims to address a real-world problem with practical applications in media monitoring, news analysis, and information retrieval.

Recommended Resources: Datasets, access to experts and journalists, and any relevant literature will be provided by the supervisor.

Pre-requisite: The student needs to follow the ARI3129 - Advanced Computer Vision for AI study unit