






A comparative study of convolutional neural networks for the detection of strong gravitational lensing

Daniel Magro ^{1,2}★ Kristian Zarb Adami ^{1,2,3} Andrea DeMarco ^{1,2} Simone Riggi ²
and Eva Sciacca ²

¹*Institute of Space Sciences and Astronomy, University of Malta, Msida MSD2080, Malta*

²*Istituto Nazionale di Astrofisica, Osservatorio Astrofisico di Catania, Via S. Sofia 78, I-95123 Catania, Italy*

³*Department of Astrophysics, University of Oxford, Oxford OX1 2JD, UK*

Accepted 2021 June 1. Received 2021 April 19; in original form 2020 December 4

ABSTRACT

As we enter the era of large-scale imaging surveys with the upcoming telescopes such as the Large Synoptic Survey Telescope (LSST) and the Square Kilometre Array (SKA), it is envisaged that the number of known strong gravitational lensing systems will increase dramatically. However, these events are still very rare and require the efficient processing of millions of images. In order to tackle this image processing problem, we present machine learning techniques and apply them to the gravitational lens finding challenge. The convolutional neural networks (CNNs) presented here have been reimplemented within a new, modular, and extendable framework, Lens EXtractor CaTania University of Malta (LEXACTUM). We report an area under the curve (AUC) of 0.9343 and 0.9870, and an execution time of 0.0061 and 0.0594 s per image, for the Space and Ground data sets, respectively, showing that the results obtained by CNNs are very competitive with conventional methods (such as visual inspection and arc finders) for detecting gravitational lenses.

Key words: gravitational lensing: strong – methods: data analysis – techniques: image processing – surveys – cosmology: observations.

1 INTRODUCTION

Strong gravitational lensed systems are unique systems in which a background galaxy and a foreground galaxy or cluster of galaxies are sufficiently well aligned so that the gravitational field of the foreground system lenses the background galaxies. Whilst these lensing systems hold a rich source of information of the gravitational field distribution of the foreground system and can be used to map dark matter distribution within the cluster, they are rare to come by. As a matter of fact, Kochanek et al. (1999) state that the number of known gravitational lenses was 47. The ‘CfA–Arizona Space Telescope Lens Survey (CASTLeS)’ website,¹ at the time of writing, lists 100 multiply imaged systems, 92 of which Kochanek et al. (1999) claim they are confident are lenses. Furthermore, the Cosmic Lens All-Sky Survey (CLASS; Myers et al. 2003), Sloan Lens ACS (SLACS; Bolton et al. 2006), Herschel Astrophysical Terahertz Large Area Survey (H-ATLAS; Negrello et al. 2016), and Strong Lensing Legacy Survey (SL2S; More et al. 2012) surveys have also contributed to the discovery of gravitational lenses.

Traditional methods for detecting these strongly lensed systems were based on visual inspection and this paper aims to address the automation of this detection. With experiments such as the Square Kilometre Array (SKA)² (Blake et al. 2004), Large Synoptic Survey Telescope (LSST; LSST Science Collaboration et al. 2009), Dark

Energy Survey (DES)³ (The Dark Energy Survey Collaboration 2005), Kilo-Degree Survey (KiDS)⁴ (de Jong et al. 2013), *Euclid*⁵ (Laureijs et al. 2011), and the *Nancy Grace Roman Space Telescope* (Dressler et al. 2012) coming online soon, thousands of these lensed systems are expected to be found and an efficient image processing technique is required in order to process the large amount of scientific images that will be produced by either of these facilities.

In order to study the detection efficiency of strongly lensed systems, the ‘Gravitational Lens Finding Challenge 1.0’⁶ was launched in 2019 (Metcalf et al. 2019). The challenge consists of 100 000 objects, the aim being to detect whether each one is a gravitational lensed system or not. Many machine learning techniques are presented by Metcalf et al. (2019), and this work aims to compare the techniques described in that paper with newer machine learning techniques, primarily convolutional neural networks (CNNs).

In the next section, we describe the framework developed and its features, followed by a description of the various methods implemented within it to tackle this problem. After this, we describe the data set provided for the challenge, and what additional techniques were utilized to ‘expand’ on this data set. We then go on to describe what metrics are presented by our framework, and how methods are evaluated, and compare the performances achieved with those achieved in other works. We conclude the work by describing further

* E-mail: daniel.magro.15@um.edu.mt

¹<https://www.cfa.harvard.edu/castles/>

²<https://www.skatelescope.org/>

³<https://www.darkenergysurvey.org/>

⁴<http://www.astro-wise.org/projects/KIDS/>

⁵<http://sci.esa.int/euclid/>

⁶http://metcalf1.difa.unibo.it/blf-portal/gg_challenge.html

improvements that can be implemented in order to make gravitational lensing detection methods more efficient and more accurate.

1.1 LEXACTUM

The framework developed in this work has been named Lens EXtractor CaTania University of Malta (LEXACTUM). The first of its main features are the image augmentation techniques described in Section 2.2 which can be toggled on or off to train for a greater number of epochs without overfitting. Another feature is the modularity of the code, allowing for the rather easy development of new models, with very easy integration of new models into the pipeline. Other features include the ability to set parameters from the command line. Examples of such parameters are the data set path, whether to train a model or load one from disk, the name of the model (to train or load), the number of epochs to train for, the batch size, and whether to use image augmentation during training or not. LEXACTUM also uses a custom ‘Data Generator’, which loads and pre-processes images in batches with the CPU, while the GPU can train on the last batch of images. Apart from image augmentation during training, all images are normalized using ZScale (NOAO 1997). Like other components, the normalization method can be easily swapped out for other techniques. Furthermore, LEXACTUM comes with a ‘results’ package, which scores the trained models and calculates several metrics, described in detail in Section 3. Moreover, LEXACTUM saves trained models to disk, and also provides functionality for loading trained models. Finally, the ‘visualize features’ component was created, which allows for the viewing of the feature maps at every convolutional layer that a trained model is ‘looking at’ during execution.

All of the architectures described in Section 1.2.5 were implemented in LEXACTUM. All of these models were then trained from scratch, on both the Space and Ground data sets, using ZScale normalization and image augmentation, for a varying number of epochs. As a starting point, all models were trained for 5 or 10, 25 or 50, 100, and 250 epochs. After that, if, say, a particular model obtained promising results, and did not seem to be overfitting (judging by the loss and accuracy of the validation set) after 250 epochs, it would then be further trained for 500, or even 1000 epochs.

1.2 Literature review

1.2.1 Conventional methods

The methods described in this subsection are not implemented in LEXACTUM, and are only presented to give a broad view of what other methods exist for tackling this problem.

1.2.2 Visual inspection

Hartley et al. (2017) go about this problem by visually inspecting and labelling each of the 100 000 images in each of the two data sets. Using their tool, BIGEYE, Hartley et al. (2017) claim that they can label around 2500 or 5000 images an hour. The final score achieved by this solution was 0.804 for the Space set and 0.889 for the Ground set (Metcalf et al. 2019). The score metric used is discussed in Section 3.

1.2.3 Arc finders

Arc finders, such as ARCFINDER (Alard 2006) and YATTALENSLITE (Sonnenfeld et al. 2018), attempt to detect lensing by looking for elongated structures, which are indicative of lensing. ARCFINDER

achieves a score of 0.66 on the Space set, whereas YATTALENSLITE achieves a score of 0.76 on the Space set and 0.82 on the Ground set (Metcalf et al. 2019).

1.2.4 Machine learning (pre-selected features)

Such methods normally involve the creation of a feature space of features deemed to be relevant by an expert. The classification is then determined by a boundary, specified either by intuition or trial-and-error. Hartley et al. (2017) attempted to solve this challenge with MANCHESTER-SVM, a Support-Vector Machine (SVM; Vapnik 1979) based solution that achieved a score of 0.81 on the Space set and 0.93 on the Ground set. Avestruz et al. (2019), on the other hand, use a Histogram of Oriented Gradients (HOG; Dalal & Triggs 2005) based approach in their solution, ALL, which scored 0.73 on the Space set and 0.84 on the Ground set (Metcalf et al. 2019).

1.2.5 Convolutional neural networks

Convolutional neural networks (CNNs) have shown to achieve very good results for both detection and recognition tasks in images and videos, among other applications. A CNN is a neural network that contains a convolutional layer. A convolutional layer ‘slides’ a kernel (also referred to as a filter) over the input image, or the output from the previous convolutional layer, and computes the output as the convolution of the pixels the ‘sliding window’ is currently over and the kernel. Each convolutional layer has a number of filters, each of which can be described as a pattern detector. The earlier layers extract geometric features, such as edges and corners, whereas deeper layers start to extract more sophisticated features, and are more capable of detecting objects such as eyes or noses (LeCun et al. 1989).

The need for convolutional layers in CNNs arises from the limitations of traditional fully connected layers when dealing with images. One such limitation is that, for a 101×101 pixel image, one layer would have more than 10 000 neurons, meaning one fully connected layer will thus have more than 100 million weights to be learnt. To put this value into perspective, from the CNNs implemented in this work, the total number of weights ranges from around 100 000 to around 6 million, for the entirety of each network. One further limitation of fully connected layers when dealing with 2D images, or 3D images when using images with more than one channel, is that when the images are flattened, most of the spatial correlation between pixels is lost. These ‘local correlations’ are very important for the recognition of low-level features, such as edges (LeCun et al. 1998).

All the techniques mentioned in this subsection are CNN based, and have been implemented in LEXACTUM. They have been implemented in PYTHON⁷ using KERAS,⁸ a high-level Application Programming Interface (API) for TENSORFLOW.⁹ All the source code and trained models mentioned in the results section are available on the GitHub repository <https://github.com/DanielMagro97/LEXACTUM>.

⁷<https://www.python.org/>

⁸<https://keras.io/>

⁹<https://www.tensorflow.org/>

1.2.6 Centre for Astrophysics and Supercomputing (CAS) Swinburne

This model was based on AlexNet (Krizhevsky, Sutskever & Hinton 2017). The input image is first passed through three convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation function and a max pooling layer. The output from the last max pool was put into two successive fully connected layers, each followed by a dropout layer (Jacobs et al. 2017; Metcalf et al. 2019). This model is discussed in further detail in Appendix A1.

1.2.7 LASTRO EPFL

This model follows a somewhat similar architecture to that described in Section 1.2.6, however has significantly more layers. This model starts off with three blocks, each block consisting of two consecutive convolutional layers, followed by a max pooling layer and a batch normalization layer. The third block is followed by a dropout layer to reduce the possibility of overfitting. Another pair of convolutional layers are added, each followed by a dropout layer. The last layer's output is passed to a triple of fully connected layers, which finally connect to a fully connected layer with a single neuron and a sigmoid activation (Schaefer et al. 2018; Metcalf et al. 2019). This model is discussed in further detail in Appendix A2.

1.2.8 CMU DeepLens

Like the previously described models, this model is CNN based, however it is made up of 'ResNet blocks'. A ResNet is a network where the input is passed through a series of convolutional layers, and the output is the addition of the original input and the output of the last convolution layer. This 'shortcut connection' from the input of the block to the end of it tackles the 'vanishing gradient problem', as it provides a 'faster' route for the gradients from back propagation to reach the earlier layers.

In the CMU DeepLens model, two different types of 'ResNet blocks' are used, one which keeps the original resolution of the image, and another which downsamples the image. In each case, the input of the 'ResNet block' goes through three convolutional layers, and is summed with the original input to the block to create the aforementioned 'shortcut connection'.

The CMU DeepLens model starts by passing the input image to a convolutional layer, followed by five groups of three successive ResNet blocks. The output from the last block is passed through an average pooling layer, and the model's prediction is computed by a fully connected layer with one neuron and a sigmoid activation (Lanusse et al. 2018; Metcalf et al. 2019). This model is discussed in further detail in Appendix A3.

1.2.9 WSI-Net

The WSI-Net model described in this paper was originally used to first detect tumours in breast scans, and then classify them. The same model was used up to the point of detection, to detect the presence of a lens in an image. The original paper does not specify hyperparameter values, those presented in this paper are those found to produce the best results, empirically.

The model starts with a convolutional layer, followed by two ResNet blocks. These ResNet blocks used are the same as those described in Section 1.2.8. This is followed by two blocks, each block consisting of a convolutional layer, a batch normalization layer, and a ReLU activation. A max pooling layer is added on next, followed by

two fully connected layers, the latter with one neuron and a sigmoid activation that determines the final classification (Ni et al. 2019). This model is discussed in further detail in Appendix A4.

1.2.10 LensFlow

In this model, the first operation carried out on the input image is an average pool. This is followed by a triple of 'convolutional layer+max pool' pairs. The last max pool layer is fed into a fully connected layer. During training, this layer is followed by a dropout layer to reduce overfitting. The final output is obtained from a fully connected layer with one neuron and a sigmoid activation (Pourrahmani, Nayyeri & Cooray 2018). This architecture is discussed in further detail in Appendix A5.

1.2.11 LensFinder

The LensFinder model has a relatively simplistic architecture, when compared to some of the solutions presented in this paper, however holds its weight with the score it obtains. The paper does not state specific values for the hyperparameters of each layer in the model, the values presented here are what were found to work best, empirically. The model starts with two blocks of 'convolutional layer, max pooling layer, and ReLU activation'. This is connected to a fully connected layer, which in turn connects to the final fully connected layer. In the original paper, a softmax activation is used, however since this is a binary classification problem, only one neuron is used in this layer, and a sigmoid activation is used instead (Pearson, Pennock & Robinson 2018). This model is discussed in further detail in Appendix A6.

2 METHODOLOGY

2.1 The data sets

Two separate labelled data sets of optical images were provided for this challenge, each with 100 000 simulated images. The first, called the 'Space' data set, is made up of single band (single channel) images that mimic data from a satellite survey such as *Euclid* (Metcalf et al. 2019). The 'Ground' data set, on the other hand, was simulated such that it mimics a ground-based survey, such as KiDS (de Jong et al. 2013), where each image has four channels of data in the 'bands (I, G, R, U)' (Metcalf et al. 2019): infrared (806 nm), green (464 nm), red (658 nm), ultraviolet (365 nm) (Binney & Merrifield 1998). For each data set, 20 000 of the 100 000 images were provided for training, whereas the other 80 000 were intended for evaluating and scoring the model. The data sets can be downloaded from http://metcalf1.difa.unibo.it/blf-portal/gg_challenge.html, for this work, 'Space set 1' and 'Ground set 1' were used. Fig. 1 shows an example of an image containing gravitational lensing, and another which does not, from the Space set. Similarly, Fig. 2 shows the two cases from the Ground set.

2.2 Image augmentation

20 000 training examples are provided for this challenge. In order to add diversity to the training set, and allow the model to train for a greater number of epochs without overfitting, image augmentation is employed. Image augmentation defines a set of transformations that can be applied to an image before it is passed to the neural network for training. It is important to note that this technique is only utilized for the training set, and not for validation or evaluation.

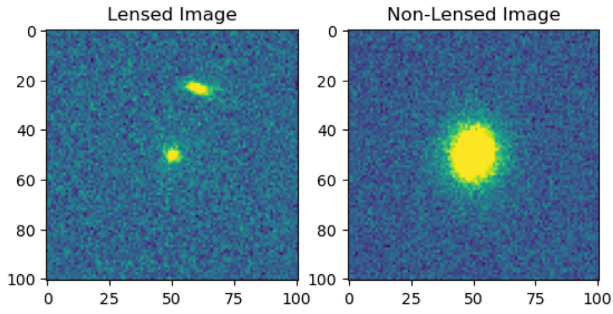


Figure 1. The image on the left is a random lensed image from the Space set, whereas the image on the right does not contain lensing. Reproduced from Metcalf et al. (2019).

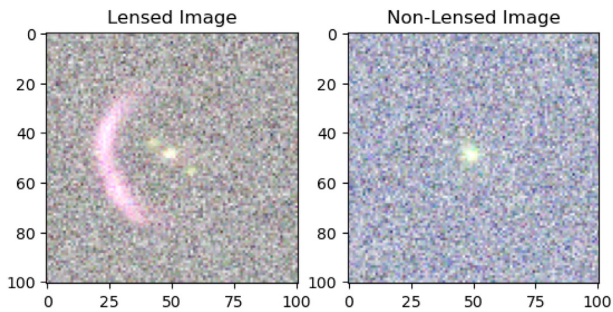


Figure 2. The image on the left is a random lensed image from the Ground set, whereas the image on the right does not contain lensing. Reproduced from Metcalf et al. (2019).

The image augmentation component utilizes the ‘IMGAUG’ library¹⁰ to define nine different transformations, of which a random amount are picked to be applied to the image. The transformations defined are the following:

- (i) a vertical or horizontal flipping of the image;
- (ii) a 90°, 180°, or 270° rotation of the image;
- (iii) a translation of [−10 per cent, 10 per cent] of the image along the *X* and/or *Y* axes;
- (iv) a scaling of [0.75, 1] of the image along the *X* and/or *Y* axes;
- (v) a shearing of [−20 per cent, 20 per cent] of the image along the *X* and/or *Y* axes.

CAS Swinburne, LASTRO EPFL, CMU DeepLens, and WSI-Net (Jacobs et al. 2017; Lanusse et al. 2018; Schaefer et al. 2018; Ni et al. 2019) all utilize image augmentation during training, however the techniques used are generally limited to flips and rotations. One feature of this framework is that it offers those transformations, along with the other three mentioned previously, as a standard to any architecture added to it.

3 RESULTS

The metrics used in the paper by Metcalf et al. (2019) were the area under the curve (AUC), TPR_0 , and TPR_{10} .

The true positive rate (TPR) is the rate of instances correctly labelled as positive. The false positive rate (FPR) is the rate of instances incorrectly labelled as positive, and thus are actually negative. The receiver operating characteristic (ROC) is a plot of

the TPR against the FPR at various thresholds. Such a plot illustrates the performance of the model, where a curve which is close to the $TPR=FPR$ diagonal would represent a model that is as effective as a coin flip, and a curve that very steeply approaches the value of $TPR=1$ represents a model that can achieve a high TPR without labelling many false positives. The area under the ROC (AUROC), or more simply the AUC, is a convenient method of quantitatively comparing ROCs.

The TPR_0 is the highest TPR achievable by the model, while keeping the FPR equal to 0. Given the difficulty in achieving a TPR_0 that is not 0, the TPR_{10} is defined, which is similarly the highest TPR achievable, while not classifying more than 10 false positives (Metcalf et al. 2019).

The final metric that was recorded for this paper was the average execution time of each model. This was obtained by recording the length of time it took for the already trained model to evaluate the test set. This was then divided by the number of images in the test set to obtain the average execution time for one image. The execution times for the same model trained for different numbers of epochs were averaged out, as they are still the same model. Furthermore, any times where the time was significantly different than the rest (outliers) were ignored, and not included in the average.

3.1 Space set results

Results obtained on the Space data set are shown in Table 1. The best TPR achieved was 0.8738 by CMU DeepLens when trained for just 25 epochs. The best FPR achieved was 0.0042, by LASTRO EPFL when trained for 5 epochs. The best AUC was 0.9343, by CMU DeepLens when trained for 500 epochs. In Metcalf et al. (2019)’s paper, the best AUC for the Space set was 0.93 by LASTRO EPFL, whereas the implementation of CMU DeepLens scored 0.92. The best TPR_0 was 0.2411, by CAS Swinburne when trained for 50 epochs. In Metcalf et al. (2019)’s paper, the best TPR_0 for the Space set was 0.22, by CMU DeepLens. The best TPR_{10} was 0.4211, by WSI Net when trained for 250 epochs. In Metcalf et al. (2019)’s paper, the best TPR_{10} for the Space set was 0.36, by GAMOCLASS, another CNN-based solution. This is a very interesting finding, as a ResNet-based network that was not included in the Metcalf et al. (2019) paper achieved a significantly higher score than that in the paper.

3.2 Ground set results

Results obtained on the Ground data set are shown in Table 2. The best TPR achieved was 0.9333, by CMU DeepLens when trained for 100 epochs. The best FPR achieved was 0.0232, by CMU DeepLens when trained for 25 epochs. The best AUC was 0.9870, by CMU DeepLens when trained for 150 epochs. In Metcalf et al. (2019)’s paper, the best AUC for the Ground set was 0.98, also by CMU DeepLens. The best TPR_0 was 0.6046, by CMU DeepLens when trained for 50 epochs. In Metcalf et al. (2019)’s paper, the best TPR_0 for the Ground set was 0.22, by MANCHESTER-SVM. The best TPR_{10} was 0.7042, again by CMU DeepLens when trained for 150 epochs. In Metcalf et al. (2019)’s paper, the best TPR_{10} for the Ground set was 0.45, also by CMU DeepLens. This is another very significant improvement, using essentially the same network as specified in the original paper. The only differences are the usage of slightly different image augmentation techniques, which possibly allowed our model to train for more epochs without overfitting. As we trained for up to 250 epochs, we were able to find more optimal weights at 150

¹⁰<https://pypi.org/project/imgaug/>

Table 1. This table shows the TPR, FPR, AUC, TPR₀, TPR₁₀, and average execution time for six different models, as described in Section 1.2.5, trained for a various number of epochs on the Space data set. Columns marked with an asterisk (*) indicate the score achieved by the model in Metcalf et al. (2019). Values in these columns marked in green indicate better performance compared to our implementations in LEXACTUM, whereas values in red indicate worse performance.

Model name	No. of training epochs	TPR	FPR	AUC	TPR ₀	TPR ₁₀	AUC*	TPR ₀ *	TPR ₁₀ *	Avg. execution time per image (s)
CAS Swinburne	5	0.5250	0.0603	0.8489	0.1531	0.1861				
	10	0.5517	0.1077	0.8171	0.1054	0.1509				
	25	0.7221	0.1178	0.8870	0.0000	0.2705				
	50	0.6252	0.0461	0.8894	0.2411	0.3000				
	75	0.6503	0.0474	0.8963	0.0000	0.3221				
	100	0.6604	0.0591	0.8915	0.0000	0.3016				
	500	0.6551	0.0295	0.9086	0.0000	0.3602	N/A			0.0124
LASTRO EPFL	5	0.3507	0.0042	0.8641	0.1539	0.2112				
	10	0.7302	0.3543	0.7825	0.1894	0.2455				
	50	0.6650	0.0287	0.9132	0.2107	0.3823				
	250	0.7937	0.0687	0.9322	0.0000	0.2268	0.93	0.00	0.08	0.0061
CMU DeepLens	5	0.6056	0.1539	0.7984	0.0000	0.1206				
	10	0.8268	0.2880	0.8710	0.0000	0.2309				
	25	0.8738	0.2726	0.9113	0.0000	0.0000				
	50	0.7570	0.0628	0.9243	0.0000	0.4073				
	100	0.8170	0.1321	0.9226	0.0000	0.0000				
	250	0.7592	0.0436	0.9291	0.0000	0.0000				
	500	0.7952	0.0626	0.9343	0.0000	0.0000				
	1000	0.8611	0.1634	0.9303	0.0000	0.0000	0.92	0.22	0.29	0.0061
WSI Net	5	0.7132	0.2955	0.7935	0.0000	0.0000				
	10	0.5437	0.0187	0.8867	0.1799	0.2934				
	50	0.7888	0.1194	0.9115	0.0000	0.0000				
	100	0.7348	0.0624	0.9069	0.0000	0.3976				
	250	0.7255	0.0531	0.9083	0.0000	0.4211	N/A			0.0055
LensFlow	5	0.6508	0.1520	0.8389	0.0728	0.1260				
	25	0.6431	0.0726	0.8799	0.1903	0.2704				
	100	0.6780	0.0636	0.8963	0.0000	0.3379				
	250	0.7384	0.0889	0.9046	0.0000	0.3632	N/A			0.0054
LensFinder	5	0.4915	0.1001	0.8038	0.0885	0.1056				
	25	0.6203	0.0663	0.8739	0.2103	0.2395				
	100	0.6912	0.0855	0.8857	0.0000	0.2721				
	250	0.7651	0.1062	0.9056	0.0000	0.3739	N/A			0.0197

epochs, whereas in the original work, the model was trained up to 120 epochs (Lanusse et al. 2018).

All the models described were trained and evaluated from scratch again, for both the Space and Ground data set, using a different split of the training, validation, and test sets (same ratio, different selection). This was done to evaluate the consistency of the results. It resulted that, for the Space set, between the two runs, the mean change between two runs of the same model with the same parameters was 0.96 per cent, with the greatest change between any two runs being 2.97 per cent. For the Ground set, the mean change was 0.39 per cent, with the greatest change being 2.99 per cent.

3.3 The importance of image augmentation

From Table 1, the results for CMU DeepLens when trained for 250 epochs with the Space data set using image augmentation are an AUC of 0.9291, TPR of 0.7592, and a FPR of 0.0436. To demonstrate the effectiveness of image augmentation, the same model was trained with the same data set, and parameters, only without image augmentation. The AUC obtained was 0.8800, the TPR obtained was 0.7103, and the FPR obtained was 0.1003. The accuracy of the model (without image augmentation) on the training

data during training can be seen rising epoch after epoch, and reaches 0.9996. On the other hand, the accuracy of the model on the validation set after 250 epochs was only 0.8156. The model obtains such a score as early as the 15th epoch, showing that the accuracy fails to improve and, thus, that the model is overfitting. When using image augmentation during training, after the same number of epochs, the model ‘only’ reaches an accuracy of 0.8989 on the training set, however manages a, relatively, impressive 0.8813 accuracy on the validation set. By the 15th epoch, this model has already achieved a validation accuracy of 0.8333, however manages to further improve on this, and as mentioned climbs to 0.88813.

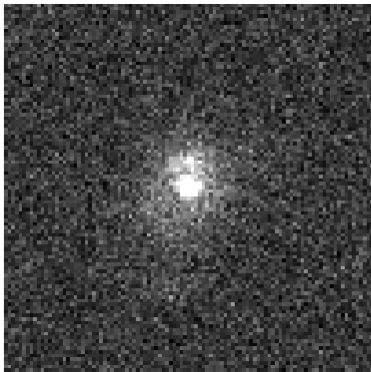
3.4 Visualizing and interpreting features extracted by convolutional layers

The ‘visualize features’ component makes it possible to visualize the outputs of each convolutional layer, for any chosen model given any image. Since it scored the highest, the ‘space.cmu.deeplens.500epochs.h5’ model was executed with a random image from the data set as an input, shown in Fig. 3. The features extracted by each convolutional layer were visualized and will be interpreted in this section.

Table 2. This table showing the TPR, FPR, AUC, TPR₀, TPR₁₀, and average execution time for six different models, as described in Section 1.2.5, trained for a various number of epochs on the Ground data set. Columns marked with an asterisk (*) indicate the score achieved by the model in Metcalf et al. (2019). Values in these columns marked in green indicate better performance compared to our implementations in LEXACTUM, whereas values in red indicate worse performance.

Model name	No. of training epochs	TPR	FPR	AUC	TPR ₀	TPR ₁₀	AUC*	TPR ₀ *	TPR ₁₀ *	Avg. execution time per image (s)
CAS Swinburne	10	0.8779	0.1077	0.9608	0.0000	0.0000				
	50	0.8995	0.0944	0.9720	0.0000	0.0000				
	100	0.8565	0.0406	0.9742	0.0000	0.0000				
	250	0.8726	0.0429	0.9758	0.0000	0.0000	0.96	0.02	0.08	0.0469
LASTRO EPFL	50	0.9073	0.0536	0.9824	0.0000	0.5133				
	100	0.9110	0.0482	0.9844	0.0000	0.5504				
	250	0.9197	0.0489	0.9862	0.0000	0.0000	0.97	0.07	0.11	0.0429
CMU DeepLens	25	0.7733	0.0232	0.9588	0.0000	0.3840				
	50	0.9138	0.0568	0.9825	0.6046	0.6827				
	75	0.9026	0.0550	0.9804	0.0000	0.6536				
	100	0.9333	0.0660	0.9851	0.0000	0.6673				
	150	0.9205	0.0445	0.9870	0.0000	0.7042				
	250	0.8593	0.0858	0.9570	0.0000	0.0000	0.98	0.09	0.45	0.0594
WSI Net	50	0.8560	0.0589	0.9620	0.0000	0.0000				
	100	0.8218	0.0301	0.9710	0.0000	0.5347				
	250	0.9127	0.0864	0.9742	0.0000	0.0000	N/A			0.0231
LensFlow	50	0.8784	0.0744	0.9708	0.0000	0.5101				
	100	0.8831	0.0738	0.9726	0.0000	0.5648				
	250	0.9006	0.0733	0.9758	0.0000	0.0000	N/A			0.0349
LensFinder	50	0.8556	0.0648	0.9665	0.0000	0.4442				
	100	0.8938	0.0805	0.9718	0.0000	0.5664				
	250	0.8997	0.0880	0.9671	0.0000	0.0000	N/A			0.0293

input image: imageEUC_VIS-100003.fits



conv2d

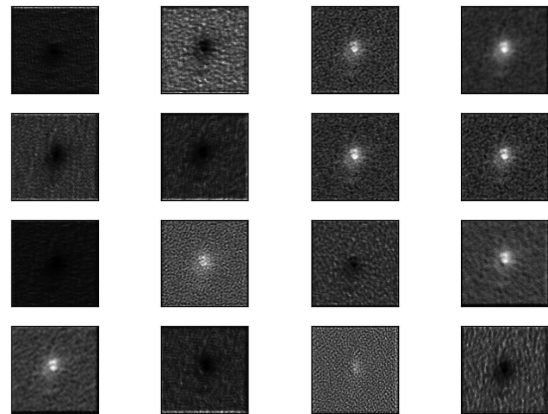


Figure 3. This is the input image, ‘imageEUC_VIS-100003.fits’, used to visualize the features extracted by the CMU DeepLens model that was trained for 500 epochs.

A sample of the features extracted by the first convolutional layer is shown in Fig. 4. At this stage the original image is still very clear in the extracted features, which is to be expected as at this stage the model is still in the process of extracting fine details from the image. For instance, the different features show the model’s efforts to emphasize certain details (that it has learnt are important and relevant) by changing the brightness, the separation from the foreground object to the background, and the emphasis on the boundary between them, to name a few.

Another sample of features extracted by the last convolutional layer of the first ‘three ResNet block’ is shown in Fig. 5. Similarly to the first convolutional stage, the original image is still quite visible

Figure 4. This is a visualization of the features extracted by the first convolutional layer of a CMU DeepLens model that was trained for 500 epochs.

in the features extracted by this layer. At this stage the model is still looking at fine details in the image.

In Fig. 6, a sample of features extracted from the last convolutional layer of the remaining ‘three ResNet blocks’ is shown. With each successive convolutional layer, the features extracted show less and less detail, with the features becoming increasingly difficult to interpret. Brownlee (2019) explains that this is due to the model extracting more abstract features in the deeper layers that show ‘more general concepts’ that make it easier for the model to make a classification.

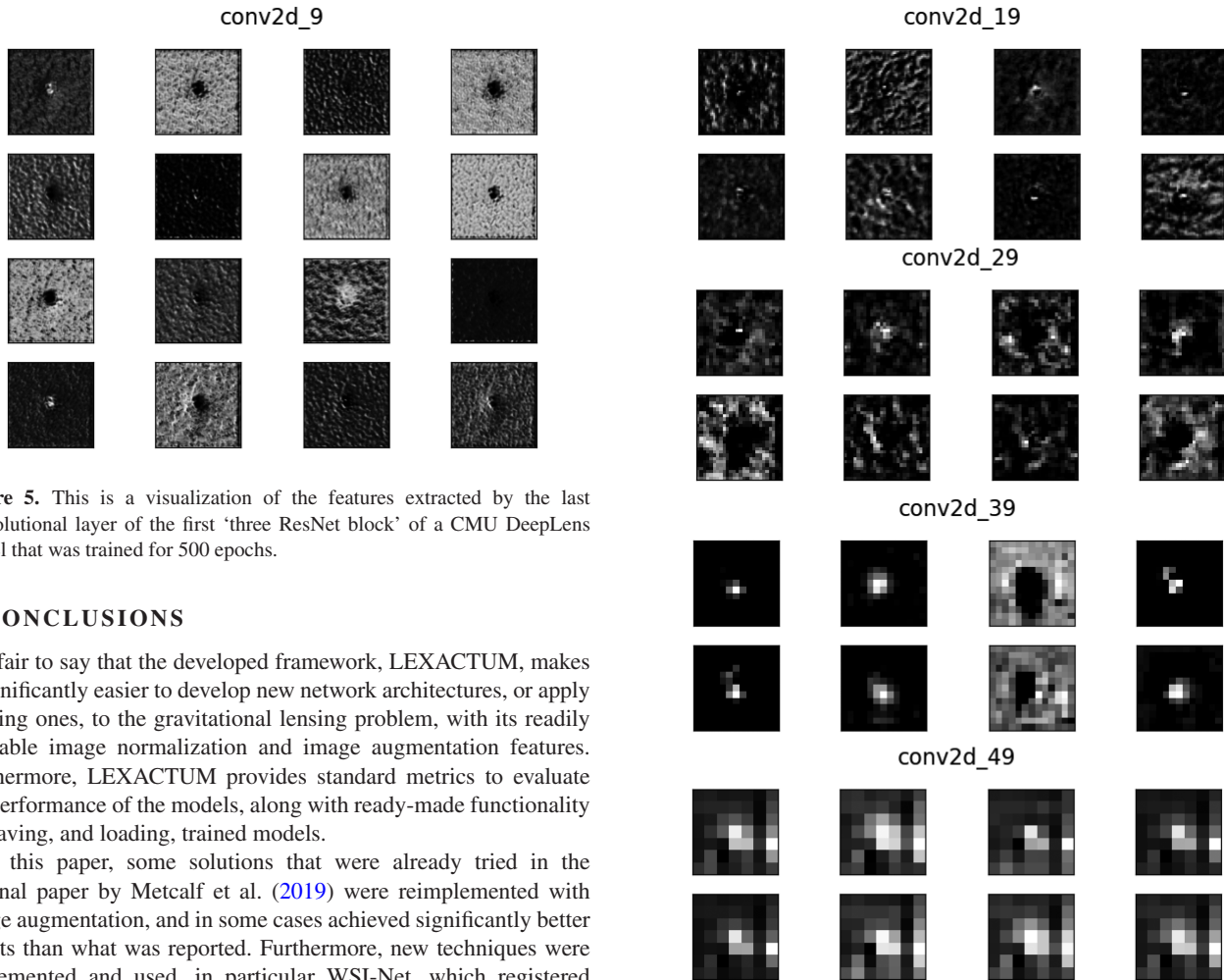


Figure 5. This is a visualization of the features extracted by the last convolutional layer of the first ‘three ResNet block’ of a CMU DeepLens model that was trained for 500 epochs.

4 CONCLUSIONS

It is fair to say that the developed framework, LEXACTUM, makes it significantly easier to develop new network architectures, or apply existing ones, to the gravitational lensing problem, with its readily available image normalization and image augmentation features. Furthermore, LEXACTUM provides standard metrics to evaluate the performance of the models, along with ready-made functionality for saving, and loading, trained models.

In this paper, some solutions that were already tried in the original paper by Metcalf et al. (2019) were reimplemented with image augmentation, and in some cases achieved significantly better results than what was reported. Furthermore, new techniques were implemented and used, in particular WSI-Net, which registered 17 per cent improvement in TPR_{10} over the winning solution in the original paper for the Space data set. A 56 per cent improvement in TPR_{10} was also registered over the winning solution for the Ground set by CMU DeepLens. CMU DeepLens also registered a very impressive 175 per cent improvement over the TPR_0 for the Ground set.

The work done here applies data pre-processing, in particular augmentation techniques, for extended training of models all the while avoiding overfitting the model to the training data. Furthermore, new techniques that were previously applied to other fields were applied to this problem, with the results obtained confirming the adaptability of CNNs. Ultimately, this work further proves the effectiveness of CNNs-based techniques for astronomical data problems.

4.1 Future work

It would be interesting to experiment with applying an elliptical Hough transform to the images as a pre-processing step, as this may make it easier for the models to locate the features that determine whether an image is classified as being lenses or not. Storkey et al. (2004) attempt to do something similar, however for their use case, they noted that it was only able to detect larger features. With this in mind, perhaps the output of the transform could be fed to the networks as an additional channel, rather than replacing the original image.

One other task that could be carried out to possibly maximize the performance of the existing models is to run hyperparameter

Figure 6. This is a visualization of the features extracted by the last convolutional layer of the remaining ‘three ResNet blocks’ of a CMU DeepLens model that was trained for 500 epochs.

optimization. A module could possibly be added to LEXACTUM that does this automatically with minimal configuration.

Further image augmentation techniques could also be tested, which would possibly allow the networks to train for an even greater number of epochs without overfitting.

Lastly, new network architectures can also be assessed. ResNet-based networks showed very promising results for this particular problem.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to the Osservatorio Astrofisico di Catania (OACT) and Istituto Nazionale di Astrofisica (INAF) for their financial support.

DATA AVAILABILITY

The data sets used for training and evaluating the CNNs are publicly available on the ‘Gravitational Lens Finding Challenge 1.0’ webpage, http://metcalf1.difa.unibo.it/blf-portal/gg_challenge.html. All the code written for LEXACTUM is also publicly available on the GitHub repository <https://github.com/DanielMagro97>

[/LEXACTUM](#), under the GNU General Public License v3.0.¹¹ The weights files for the trained models have also been made available on Zenodo <https://doi.org/10.5281/zenodo.4299924> and Google Drive <https://drive.google.com/drive/folders/1qn03htSDz-0aB6jRWbKmdk4QBz0epSmS?usp=sharing> (Magro et al. 2020).

REFERENCES

- Alard C., 2006, preprint (arXiv:astro-ph/0606757)
 Avestruz C., Li N., Zhu H., Lightman M., Collett T. E., Luo W., 2019, *ApJ*, 877, 58
 Binney J., Merrifield M., 1998, *Galactic Astronomy*. Princeton Univ. Press, Princeton, NJ
 Blake C., Abdalla F., Bridle S., Rawlings S., 2004, *New Astron. Rev.*, 48, 1063
 Bolton A. S., Burles S., Koopmans L. V. E., Treu T., Moustakas L. A., 2006, *ApJ*, 638, 703
 Brownlee J., 2019, How to Visualize Filters and Feature Maps in Convolutional Neural Networks. <https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/>
 Dalal N., Triggs B., 2005, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, Piscataway, NJ, p. 886
 de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, *Exp. Astron.*, 35, 25
 Dressler A. et al., 2012, preprint (arXiv:1210.7809)
 Hartley P., Flamary N., Jackson N., Tagore A. S., Metcalf R. B., 2017, *MNRAS*, 471, 3378
 Jacobs C., Glazebrook K., Collett T., More A., McCarthy C., 2017, *MNRAS*, 471, 167
 Kochanek C. S., Falco E. E., Impey C. D., Lehár J., McLeod B. A., Rix H.-W., 1999, in Holt S. S., ed., *AIP Conf. Proc. Vol. 470, The 9th Astrophysics Conference: After the Dark Ages, when Galaxies were Young (the Universe at $2 < Z < 5$)*. Am. Inst. Phys., New York, p. 163
 Krizhevsky A., Sutskever I., Hinton G. E., 2017, *Commun. ACM*, 60, 84
 Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, *MNRAS*, 473, 3895
 Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
 LeCun Y., Boser B., Denker J., Henderson D., Howard R., Hubbard W., Jackel L., 1989, *Adv. Neural Inf. Processing Syst.*, 2, 396
 LeCun Y., Bottou L., Bengio Y., Haffner P., 1998, *Proc. IEEE*, 86, 2278
 LSST Science Collaboration et al., 2009, preprint (arXiv:0912.0201)
 Magro D., Zarb Adami K., DeMarco A., Riggi S., Sciacca E., 2020, *LEXACTUM Trained Model Weights*
 Metcalf R. B. et al., 2019, *A&A*, 625, A119
 More A., Cabanac R., More S., Alard C., Limousin M., Kneib J.-P., Gavazzi R., Motta V., 2012, *ApJ*, 749, 38
 Myers S. T. et al., 2003, *MNRAS*, 341, 1

¹¹<https://www.gnu.org/licenses/gpl-3.0.html>

National Optical Astronomy Observatory, 1997, IRAF (Image Reduction and Analysis Facility) Display Help Page. <https://iraf.net/irafhelp.php?val=display>

- Negrello M. et al., 2016, *MNRAS*, 465, 3558
 Ni H., Liu H., Wang K., Wang X., Zhou X., Qian Y., 2019, in Suk H.-I., Liu M., Yan P., Lian C., eds, *Machine Learning in Medical Imaging*. Springer, Cham, Switzerland, p. 36
 Pearson J., Pennock C., Robinson T., 2018, *Emergent Sci.*, 2, 1
 Pourrahmani M., Nayyeri H., Cooray A., 2018, *ApJ*, 856, 68
 Schaefer C., Geiger M., Kuntzer T., Kneib J. P., 2018, *A&A*, 611, A2
 Sonnenfeld A. et al., 2018, *PASJ*, 70, S29
 Storkey A. J., Hambly N. C., Williams C. K. I., Mann R. G., 2004, *MNRAS*, 347, 36
 The Dark Energy Survey Collaboration, 2005, preprint (arXiv:astro-ph/0510346)
 Vapnik V., 1979, *Reconstruction of Dependences from Empirical Data*. Nauka, Moscow

APPENDIX A: OVERVIEW OF METHODS

A1 CAS Swinburne

This model was based on AlexNet (Krizhevsky, Sutskever & Hinton 2017). The input image is first passed through three convolutional layers, with kernel sizes of 11, 5, and 3, respectively, and 96, 128, and 256 feature maps, respectively. Each convolutional layer was followed by a ReLU activation function and a 3×3 max pooling layer. The output from the last max pool was put into two successive fully connected layers, with 1024 neurons each. A dropout layer with 0.5 probability was added after each fully connected layer (Jacobs et al. 2017). This architecture is shown graphically in Fig. A1.

A2 LASTRO EPFL

This model follows a somewhat similar architecture to that described in Section 1.2.6, however has significantly more layers. All layers in this model use a ReLU activation, unless specified otherwise. This model starts off with three blocks, where each block consists of two consecutive convolutional layers, followed by a max pooling layer and a batch normalization layer. The first block's convolutional layers use a kernel size of 4 and 3, respectively, with 16 features each. The convolutional layers in the second and third blocks all use a kernel size of 3, with the second block having 32 features, and the third having 64. As specified, all three blocks are followed by a max pooling and a batch normalization layer. After the third block, a dropout layer is added to reduce the possibility of overfitting. A convolutional layer with a kernel size of 3 and 128 features is added, followed by another dropout layer. This is followed by another convolutional layer of the same specifications, this time followed by a batch normalization layer and another dropout layer. The last

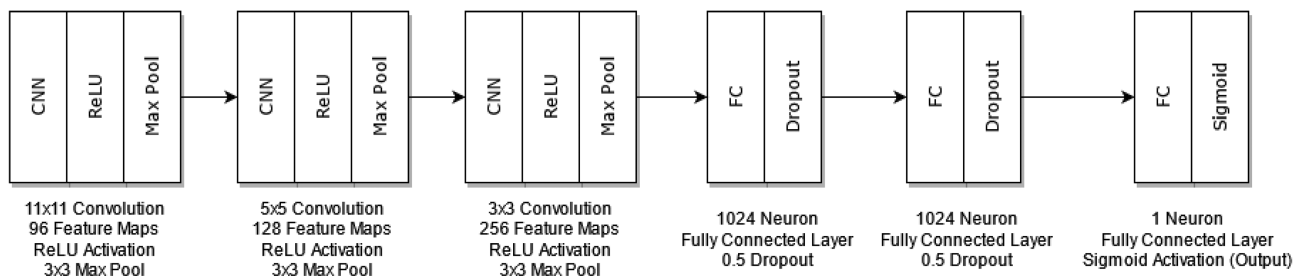


Figure A1. This is a graphical representation of the ‘CAS Swinburne’ model described in Section 1.2.6 and Appendix A1. Reproduced from Jacobs et al. (2017).

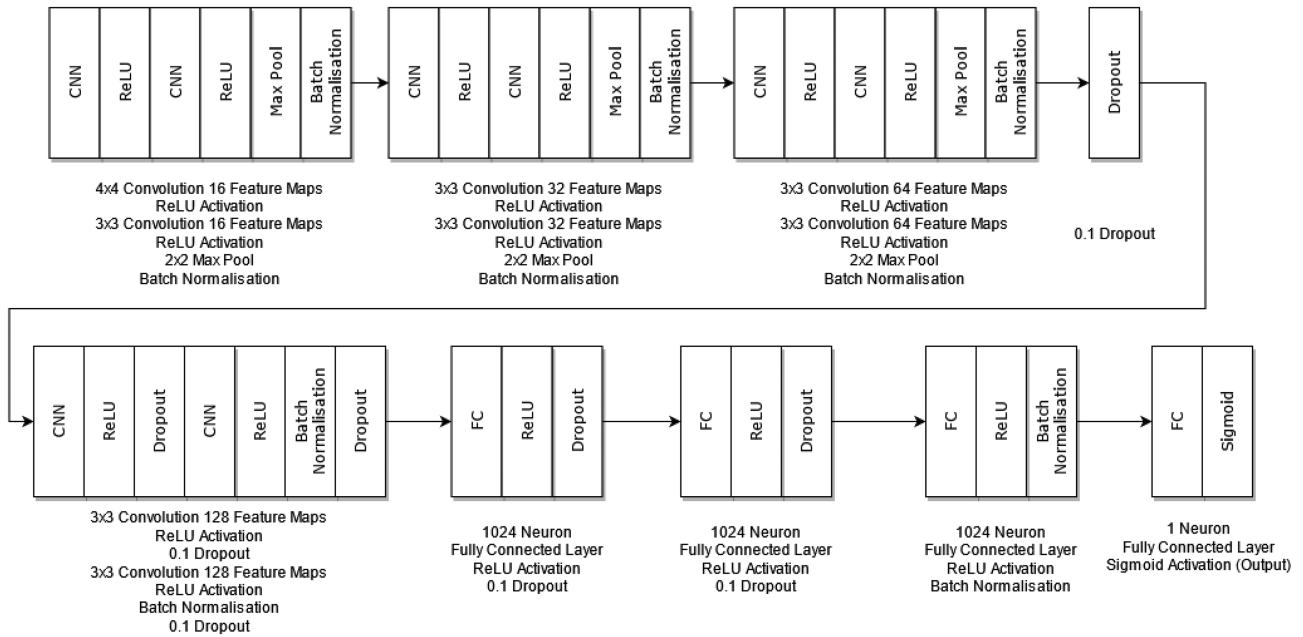


Figure A2. This is a graphical representation of the ‘lastro_epfl’ model described in Section 1.2.7 and Appendix A2. Reproduced from Schaefer et al. (2018).

layer’s output is flattened and passed to a triple of fully connected layers, with a dropout layer between each fully connected layer. Batch normalization is added after the last fully connected layer. The model’s output is obtained by passing the output of the last batch normalization to a fully connected layer, with a single neuron, with a sigmoid activation function (Schaefer et al. 2018). This architecture is shown graphically in Fig. A2.

A3 CMU DeepLens

As shown in Fig. A3, two different types of ‘ResNet blocks’ are used by this model, one which keeps the original resolution of the image, and another which downsamples the image by a factor of 2. In the case where the image is not downsampled, a copy of the input to the ResNet block is stored. The input is also passed through the triple of batch normalization, non-linearity (Exponential Linear Unit – ELU), and a convolutional layer three times. The result of these nine layers is summed with the original input to the ResNet block, and returned as the output. In the case where downsampling is employed, the input to the ResNet block first goes through batch normalization and non-linearity (ELU), after which a copy of the current tensor is stored for later use. This is followed by a convolutional layer with a stride of 2, and another two ‘batch normalization, ELU, and convolutional layer’ triples. The output of the last convolutional layer is summed with the aforementioned copy of the tensor at an earlier stage, after it has gone through a convolutional layer with stride 2, and returned as the block’s output.

The CMU DeepLens model is structured as follows. The image is first passed through a convolutional layer with a kernel size of 7, with 32 features, using an ELU activation function, and is followed by a batch normalization layer. This is then followed by three ResNet blocks, each with 32 features. This is followed by another four sets of ‘three ResNet blocks’. Each of these sets starts with a downsampling ResNet block, followed by two ‘regular’ ResNet blocks. The features of each ResNet block in each set are 64, 128, 256, and 512, respectively. The output from the last ResNet block is passed through an average pooling layer, and the model’s prediction

is finally computed by a fully connected layer with one neuron and a sigmoid activation (Lanusse et al. 2018). This architecture is shown in Fig. A4.

A4 WSI-Net

The first layer is a convolutional layer with a kernel size of 7, 32 features, and an ELU activation. This is then followed by two ResNet blocks of 32 and 64 features, respectively. These ResNet blocks used are the same as those described in Section 1.2.8. Following the two ResNet blocks is another convolutional layer with a kernel size of 1, 32 features, and an ELU activation. This is followed by a batch normalization layer, and a ReLU activation. A convolutional layer with kernel size 5, 32 features, and an ELU activation is used next, again followed by a batch normalization layer as well as a ReLU activation. A max pooling layer is added on next, followed by a fully connected layer with 256 neurons. The final classification is obtained by another fully connected layer with one neuron and a sigmoid activation (Ni et al. 2019). This architecture is shown in Fig. A5.

A5 LensFlow

In this model, the first operation carried out on the input image is an average pool with a pool size of 3×3 and a stride of 3. This is followed by a convolutional layer with a kernel size of 5 and 16 features, and a max pooling layer with a pool size of 2 and a stride of 2. This is again followed with another two ‘convolutional layer+max pool’ pairs, where the convolutional layers have a kernel size of 5 and 25 features, and a kernel size of 4 and 36 features, respectively, and both max pools have a pool size of 2 and a stride of 2. The last max pool layer is fed into a fully connected layer with 128 neurons and a ReLU activation. During training, this layer is followed by a dropout layer with 0.5 probability. The final output is obtained from a fully connected layer with one neuron and a sigmoid activation (Pourrahmani, Nayyeri & Cooray 2018). This architecture is shown in Fig. A6.

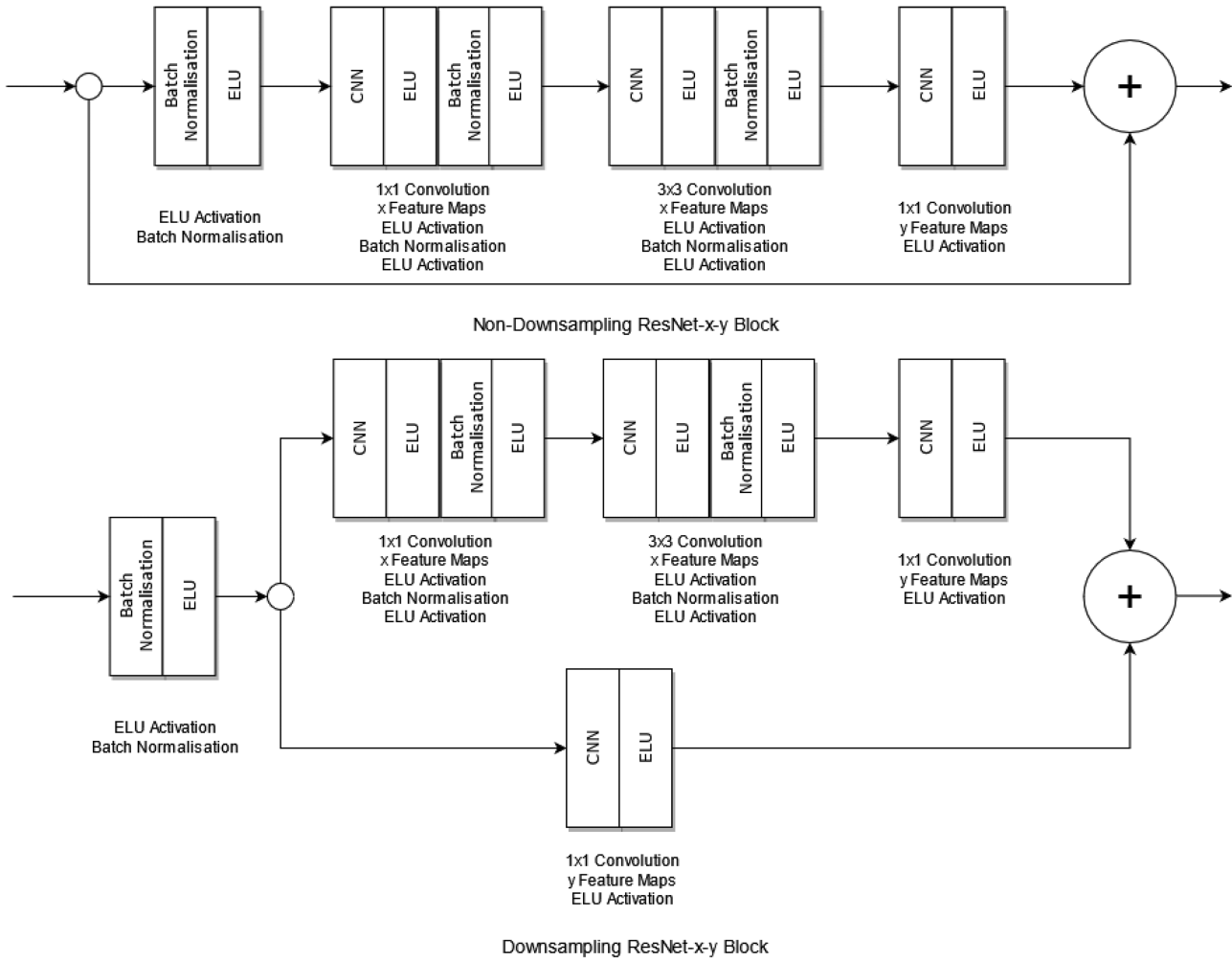


Figure A3. This is a graphical representation of the two types of ‘ResNet blocks’ used by the ‘CMU DeepLens’ model described in Section 1.2.8 and Appendix A3. Reproduced from Lanusse et al. (2018).

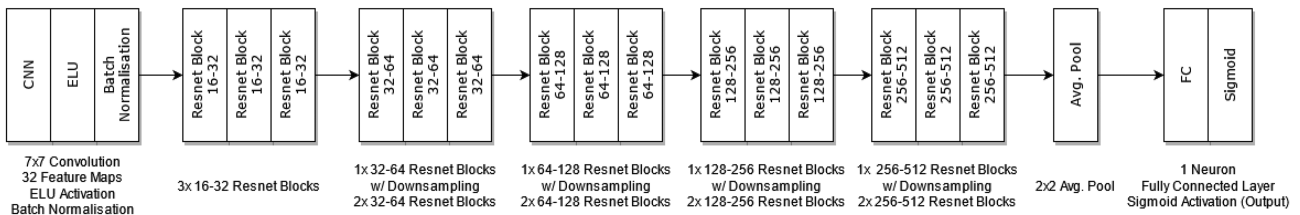


Figure A4. This is a graphical representation of the ‘CMU DeepLens’ model described in Section 1.2.8 and Appendix A3. Reproduced from Lanusse et al. (2018).

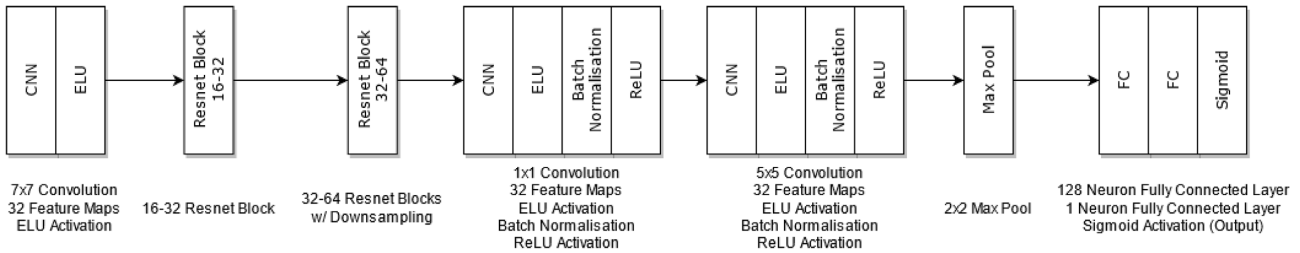


Figure A5. This is a graphical representation of the ‘WSI-Net’ model described in Section 1.2.9 and Appendix A4. Reproduced from Ni et al. (2019).

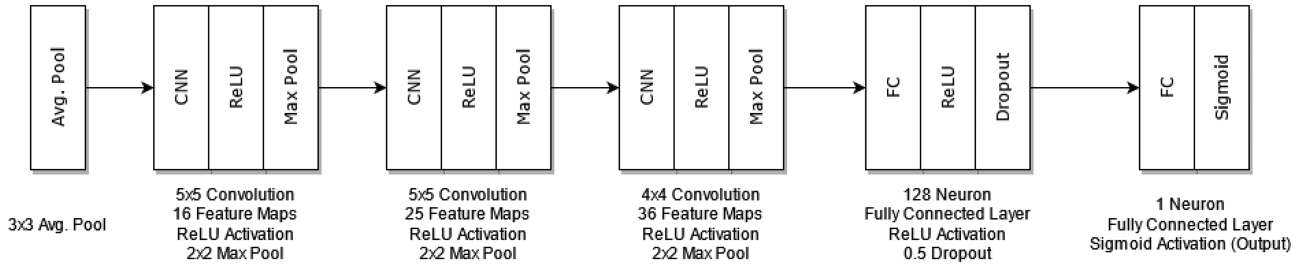


Figure A6. This is a graphical representation of the ‘LensFlow’ model described in Section 1.2.10 and Appendix A5. Reproduced from Pourrahmani et al. (2018).

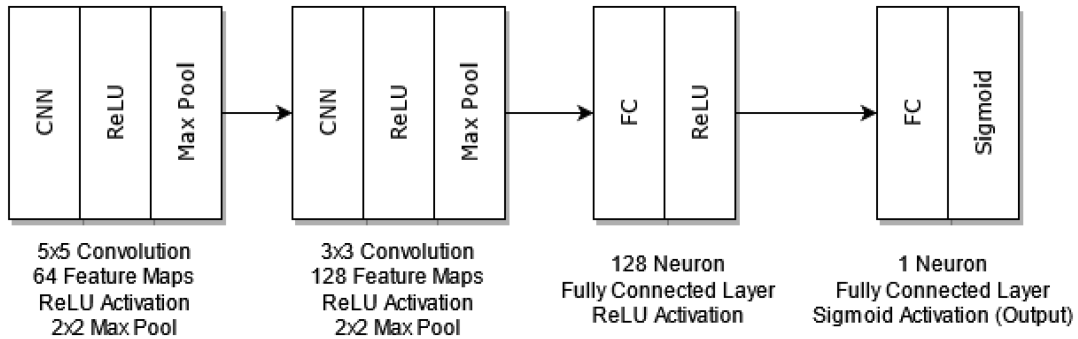


Figure A7. This is a graphical representation of the ‘LensFinder’ model described in Section 1.2.11 and Appendix A6. Reproduced from Pearson et al. (2018).

A6 LensFinder

The LensFinder model has a relatively simplistic architecture, when compared to some of the solutions presented in this paper, however holds its weight with the score it obtains. The paper does not state specific values for the hyperparameters of each layer in the model, the values presented here are what were found to work best, empirically. The model starts with a convolutional layer, with a kernel size of 5 and 64 features. A ReLU activation function is used. The result is fed into a max pooling layer. The output is then passed into another convolutional layer with a kernel size of 3, and 128 features. Here

again, a ReLU activation is used and the output goes into a max pooling layer. This is connected to a fully connected layer with 128 neurons, and a ReLU activation. This output is connected to the final fully connected layer. In the original paper, a softmax activation is used, however since this is a binary classification problem, only one neuron is used in this layer, and a sigmoid activation is used instead (Pearson, Pennock & Robinson 2018). This architecture is displayed in Fig. A7.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.