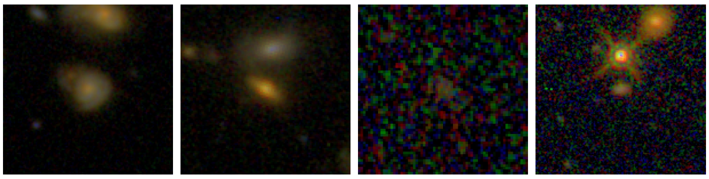
GALAXY CLASSIFICATION USING CUSTOM CONVOLUTIONAL NEURAL NETWORKS

A.CHENG, B.SIMMONS
DEPARTMENT OF PHYSICS, LANCASTER UNIVERSITY, LANCASTER, LA1 4AB, UK
S.A.CHENG@LANCASTER.AC.UK, B.SIMMONS@LANCASTER.AC.UK

1 - INTRODUCTION

For large sky surveys, it is useful to classiy galaxies from stars and other artifacts for further analysis. This process, done manually, can be laborious and time consuming. AI algorithms, such as CNNs, can help.

Convolutional neural network, aka CNNs, are a type of neural network often used for images. Using data from GalaxyZoo and images fom CANDELs survey, this project evaluates the possibility of using CNNs for galaxy classification.



(a) (b) (c) (d)

Figure 1. Typical images from the dataset: (a) spiral galaxy, (b) an elliptical galaxy, (c) other artifact, (d) a star.

2 - DATAPROCESSING

CNNs requires training and validation images. For training images, the data processing software TOPCAT was used to split images into their categories based on crowd sourced classifications of the images. Roughly 20% of the training image set was taken as validation images.

To increase the size of the dataset, images were flipped, rotated, and translated. A larger dataset reduces the possibility of overtraining, which makes the program more robust.

3 - CNN ARCHITECTURE

The following diagram shows the structure for our network's custom training algorithm:

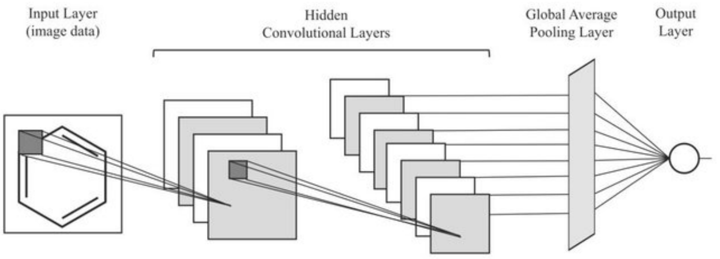
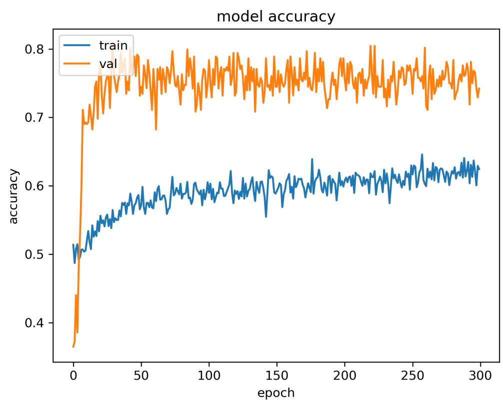


Figure 2. The input images are fed into the hidden layers which operates on the images (as a 2D matrix). Within leach layer, a filter (like small matrices) is applied to each image and undergoes pooling, which reduces the spacial size of the outputs. Pooled images are then flattened which reverts the images into usable output. In this case, '0' for galaxy or '1' for not-galaxy. Sample image sourced from [2].

A CNN requires many iterations of training to improve upon its results. Using a dataset of around 12,000 images per category, the program ran for 300 epochs (iterations) to ensure that the program is fully trained.

4 - OUTPUTS AND ACCURACY

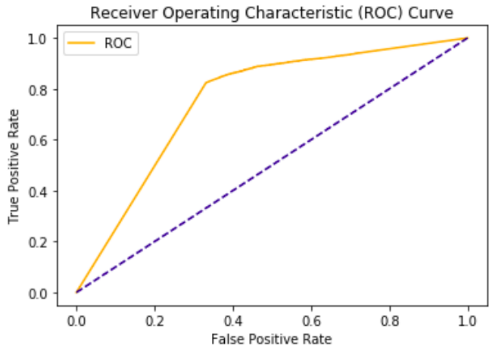
Figure 3. The training accuracy and the validation (i.e. prediction) accuracy respectively. The fluctuations are caused by noise in the data. The average prediction accuracy is at 78%. The gap between the two accuracy plots is due to heavy use of regularisation techniques.



To test the network's performance, the accuracy per epoch was recorded as the program executes.

Post analysis was required using ROC curves and histograms to see whether the predictions are correct as well as accurate.

Figure 4. The orange ROC curve lies within the top left corner, more angled than flat with the value for the area under the curve as 0.76. Where 1 is when all predictions are true positive.



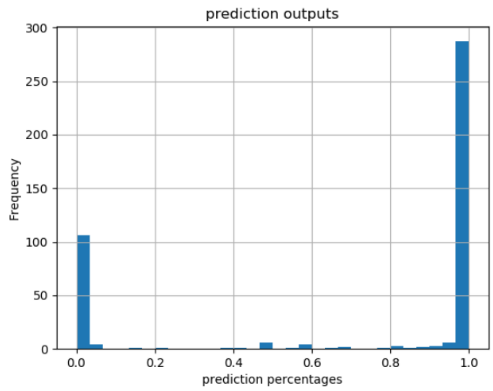


Figure 5. Histogram showing the prediction outputs, where the histogram at '0.0' means 'definitely a galaxy' and '1.0' means 'definitely not a galaxy', while '0.50' indicates complete uncertainty for the image's classification.

5 - CONCLUSIONS

Our CNN architecture is able to successfully classify 24,000 images between galaxies and non-galaxy images with a high validation accuracy of 78%.

Out of all the classified images, the rate of true positives from predictions from the data is high, with a value of 0.76.

Of the misclassified images, the network conservatively classified lower quality images of galaxies as non-galaxy images, producing a reliable set of galaxy images.

Images with uncertain classifications are often noisy or blank images, E.g. the figure 1(c), which has a prediction value of 0.46.

6 - FUTURE DEVELOPMENT

Further development can be done with the architecture to further increase the accuracy and the the ROC curve.

Improve the use of regularisation to increase the average training accuracy.

7 - REFERENCES

[1] Salman Khan & Hossein Rahmani et al. (2018) "A Guide to Convolutional Neural Networks for Computer Vision" Morgan & Claypool, pp. 43-68, doi.org/10.2200/S00822ED1V01Y201712COV015

[2] Rajarapollu, Prachi and Vijay Mankar (2017). “Bicubic Interpolation Algorithm Implemen-tation for Image Appearance Enhancement”. In:www.ijcst.com8, pp. 14, ISSN: 2229-4333 (print)