Classification of Galaxies Using Convolutional Neural Networks

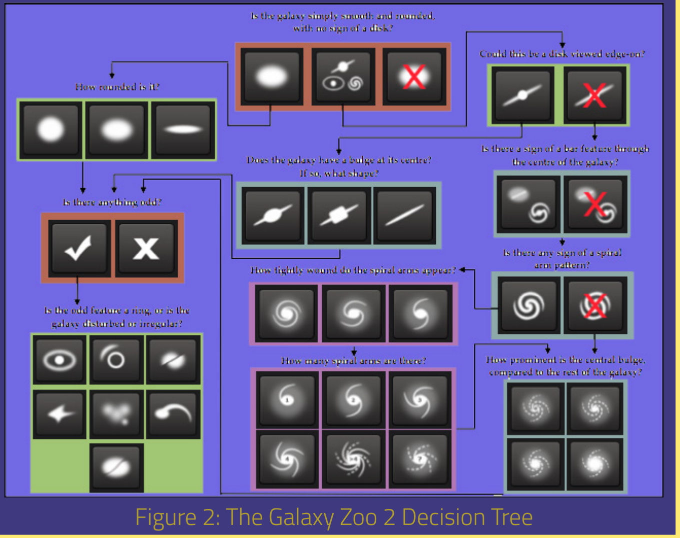
Alex Daniel

Introduction

The Sloan Digital Sky Survey (SDSS) generated a dataset of approximately 900,000 images of galaxies that needed classifying, figure 1. The brightest 250,000 of these were submitted to the Galaxy Zoo 2 citizen science project where volunteers were presented with an image then asked a series of questions outlined in figure 2. This process is obviously very labor intensive and as such a computational algorithm for rapidly classifying galaxies would be highly beneficial.

Using a 60,000 sub-sample of 424x424 images from the SDSS and the probabilities of each answer to the decision tree per galaxy a convolutional neural network (CNN) was created to follow the decision tree. To reduce the number of paths through the tree, the final two questions were neglected from this model.

Figure 1: An example SDSS CNN input image



Is the galaxy simply smooth and rounded, with no sign of a disk?
Could this be a disk viewed edge-on?
How rounded is it?
Does the galaxy have a bulge at its centre? If so, what shape?
Is there a sign of a bar feature through the centre of the galaxy?
Is there anything odd?
Is there any sign of a spiral arm pattern?
How tightly wound do the spiral arms appear?
Is the odd feature a ring, or is the galaxy disturbed or irregular?
How many spiral arms are there?
How prominent is the central bulge, compared to the rest of the galaxy?

Figure 2: The Galaxy Zoo 2 Decision Tree

Results

The network was optimised to deliver the minimum RMSE over the test image set with a reasonable training time per epoch. To increase the rate with which training could be carried out, an NVIDIA GeForce GTX960 GPU was used. This optimisation resulted in the network architecture shown in figure 3.

The network was trained both with and without augmentation via dividing each image into quarters. This method of augmentation produced over-fitting and as such the unaugmented waiting was settled upon.

This architecture produced a minimum RMSE of 0.02 when trained over 30 epochs taking a 30 minutes.

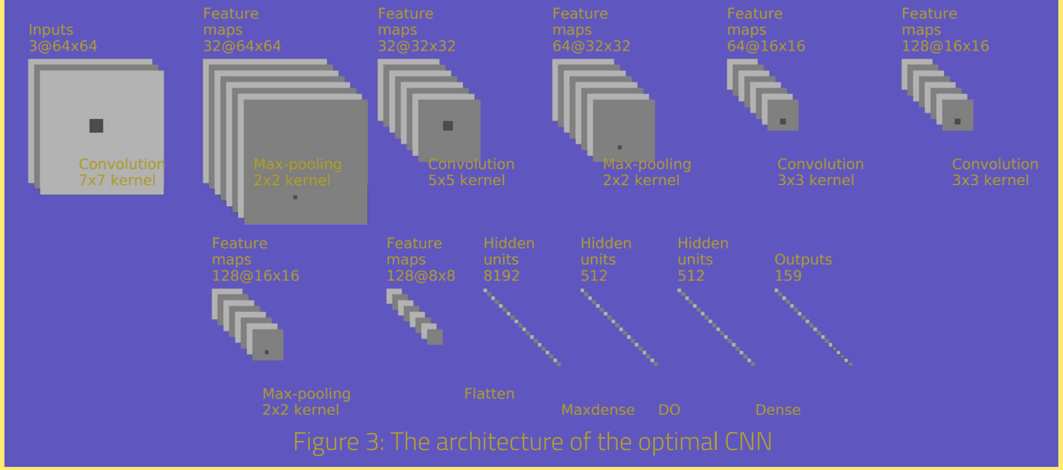


Figure 3: The architecture of the optimal CNN

Method

In the vast majority of images the target galaxy only occupied a small portion of the image and was always centered. To reduce the amount of data the CNN would have to process, the images were cropped and re-sampled to be 64x64. The data was then divided into a training and a test set with 95% of images being in the training set. The labels were processed to be the probability that a galaxy would follow each of the 159 routes through the decision tree.

Multiple different architectures of network and degrees of augmentation were experimented with and the output prediction accuracy expressed as a root mean square error as defined below.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(p_{\mathrm{predicted}} - p_{\mathrm{actual}}\right)^2}$$

Conclusion

Using a convorlutional neural network images from the Sloan Digital Sky Survey were assigned a probability of being in each of 159 categories representing routes through the decision tree used to classify the galaxies morphology during the citizen science project Galaxy Zoo 2. This CNN achieved a minimum root mean square error of 0.02.

Further developments would include the final two questions although these could be one independent CNN given all galaxies go through these questions so their answers do not depend on those higher in the tree.

EPSRC
The University of Nottingham
onbi