Data science project

Klay Kulik*, Deepak Sharma† and Linxiao Wang†
*Department of Physics and Astronomy, †Department of Computer Science
Email: kkulik2@uwo.ca, dsharm55@uwo.ca, lwang739@uwo.ca

Abstract

In this project, we discuss the problem of galaxy morphological classification, where the input data contain five bands (u, g, r, i and z) instead of the typical three channel (RGB) images. We considered the binary case where a galaxy is either spiral or elliptical. The data we used came from the Sloan Digital Sky Survey and the Galaxy Zoo project. We first performed Logistic Regression as a baseline, which gave a testing accuracy of 58.9%. Then we built a Convolutional Neural Network with five convolution layers that gave an accuracy of 98.5% with around 18000 training data for each type. Obviously CNN gave a much more promising result comparing to Logistic regression for this problem, but further improvements should also be considered, for example, fine tuning the CNN. Also the accuracy of the 'true' labels should also be further discussed since this part of the data is collected by crowdsourcing.

I. Introduction

A. Motivation

Galaxies have morphology that can be split into two main categories: spiral and elliptical. The differences between them are fairly obvious when looking at images: spiral galaxies are disk shaped and have curved arms extending out from the centre, while elliptical galaxies are shaped like a three dimensional blob that is brighter in the middle.

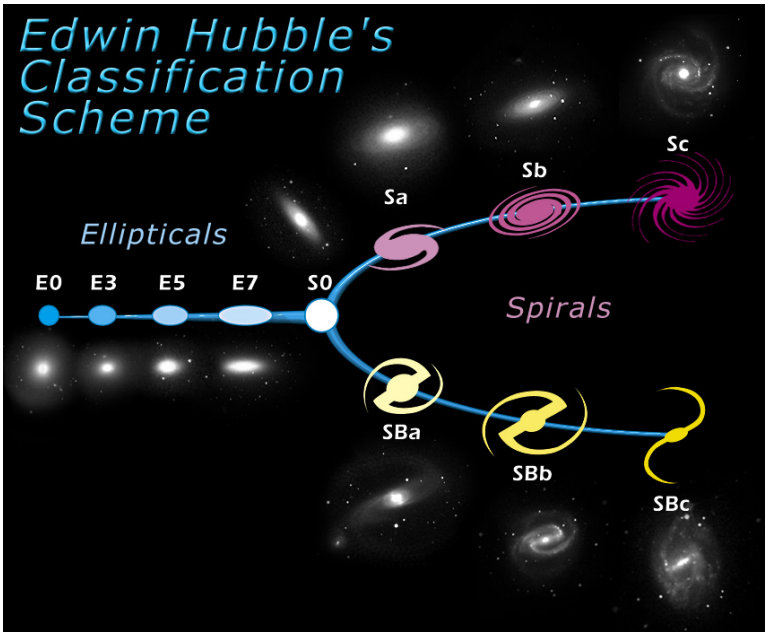

Fig. 1: Edwin Hubble's 'Tuning Fork' classification scheme. Galaxies are classified into two main categories, spiral and elliptical, with further sub-classifications representing minute changes not discussed in this report [1].

There are however many differences between the two galaxy morphologies that are not as glaringly obvious. For example, Spiral galaxies tend to be bluer than their elliptical cousins. Dense spiral arms are

the birthplace of new stars, which excite gas around them and cause it to glow a blue colour. Elliptical galaxies lack the gas and dust required to form new stars, and therefore tend to be redder.

Classifying galaxies based on their morphology has an enormous impact on research in many fields of astronomy, including things from galactic dynamics to actually determining the age of a galaxy! That last point may sound skeptical, however consider the event of galaxy collisions. When two galaxies of any type collide, they eventually form a large elliptical galaxy. Elliptical galaxies are much more common near the centre of galaxy clusters for this reason, and can be many times larger than their spiral counterparts. This is why scientists see more spiral galaxies as we look deeper into space - because it is analogous to looking further back in time where less spiral galaxies had time to collide and form ellipticals.

Humans can easily distinguish between the two morphologies by just looking at the galaxy images. However there is one major problem in this field of research that is rather unexpected - we have too much data available to look through and separate! The Sloan Digital Sky Survey is one of the newer collection surveys that is starting to monopolize the field. SDSS has a list of over 200,000,000 individual galaxies across more than one third of the night sky[2]!

B. Approach

This is where data science comes in! We first performed Logistic regression on the data, specifically on the average flux value of each image channel. This was to see if there was any obvious relationship between the colour of the galaxy and its morphology. We then trained a neural network to classify galaxy images into spiral or elliptical. If successful, a convolutional neural network (CNN) should be able to classify galaxies much faster than humans, with comparable accuracy.

II. Data

A. Data Source

The data we used for this project came from the Sloan Digital Sky Survey (SDSS) Data Release (DR)7 [3], and the Galaxy Zoo (GZ) project first data release [4]. SDSS is a major multi-spectral imaging and spectroscopic redshift survey using a dedicated wide-field 2.5 m telescope located at Apache Point Observatory (APO) near Sacramento Peak in Southern New Mexico. DR7 contains five-band photometry for 357 million distinct objects [3]. The GZ project collected simple morphological classifications of nearly 900,000 galaxies drawn from SDSS DR7, contributed by hundreds of thousands of volunteers [4].

We took the galaxy names and classifications from the GZ database, and searched SDSS for the corresponding images. For each galaxy five FITS files are provided by SDSS. The SDSS images are unique in that the data is stored in five bands (‘ugriz’) instead of the typical three channels (RGB). ‘ugriz’ channels are also absolute, rather than relative, meaning that instead of ranging from 0 to 255, the pixels have any integer greater than zero. This is important because

B. Pre-processing

Since the classifications in GZ are from crowdsourcing, we only used the data entries that have high confidence classifications. We choose galaxies with debiased probability (given in [4]) greater than 0.985 for spiral galaxies, and 0.926 for elliptical galaxies, respectively. We choose these thresholds to ensure that: (i) the galaxies used for training the neural network have highly accurate classifications; and (ii) the number of data for both classes in the training and test sets are balanced [5]. We obtained 19306 and 18811 galaxies for spiral and elliptical respectively.

The data we got from SDSS are in Flexible Image Transport System (FITS) format. Each FITS file contains a header part and a data part. We removed the header part and resized the images to 200×200

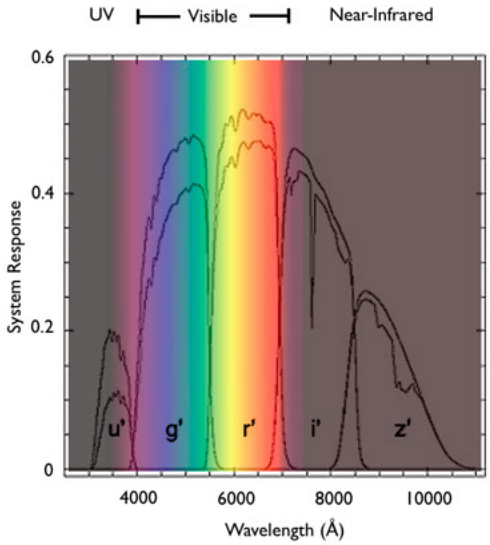

Fig. 2: The ‘ugriz’ filter shcematic with a spectrum overplotted

pixels. As we discussed in Section II-A, ‘ugriz’ covers a broader band than RGB images. Common ways to map ‘ugriz’ files to RGB images would take 3 or 4 channels (out of 5) of ‘ugriz’, and do a linear transformation on them, thus would definitely lose some information. In order to keep a more complete data for each galaxy we used ‘ugriz’ files rather than RGB images as the input to our classifier. Figure 3 shows the gray-scale images of the u, g, r, i and z band photometry of a spiral galaxy as well as the RGB image of the same galaxy. We can see that there is information on each band but the RGB image only uses 3 or 4 of them which would cause information loss.

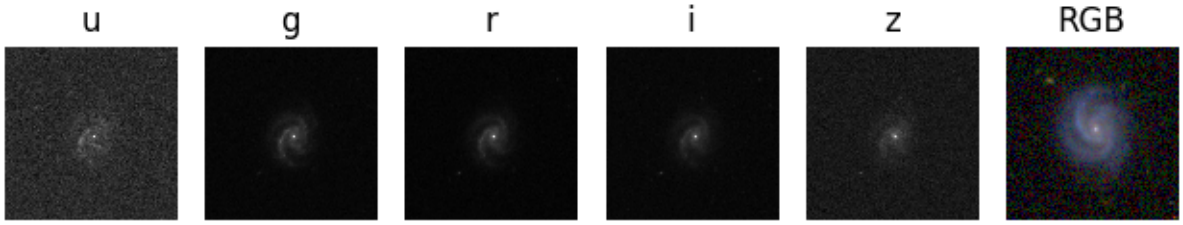

Fig. 3: The ’ugriz’ images and the RGB image of the same galaxy

III. Method

A. Logistic Regression

Baseline model

B. Deep Learning Model - CNN

Deep Learning has achieved significant results and a huge improvement in visual detection and recognition with a lot of categories. Raw data images are used by deep learning as input without the need of expert knowledge for optimization of segmentation parameter or feature design. We used open source software stacks for our project. The deep learning APIs used are Keras and Tensorflow. The proposed architecture of the deep network for the morphological classification is illustrated in detail

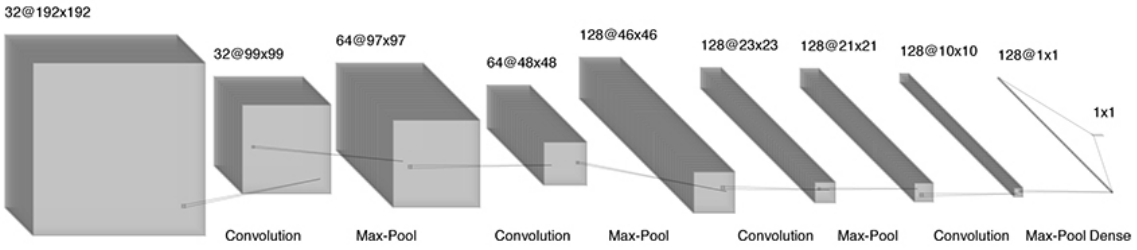

Fig. 4: Architecture of the CNN model

in Figure 4. It consists of 15 layers, made up of 5 main layers for features extraction, followed by two principle fully connected layers for classification. The first layer is the input layer. Every main layer is further made of one convolutional layer with the Rectified Linear Unit(ReLU) as the nonlinear activation function and a max pooling layer at the end for subsampling. The first fully connected layer has 128 neurons with ReLU activation function, while the last fully connected layer has one neuron and uses a sigmoid to obtain class memberships. Visualizing the feature extraction and classification layers in the proposed deep neural architecture will give a better understating. In the main layers, features are extracted and the patterns identified become more complex as we go deeper into the network. CNNs are generally used for image classification but they are also very useful for finding patterns in any data that can benefit from filters. Using a 5 channel raw input is not typical when employing CNNs but since the image data(RGB) for the galaxy is a subset of the wavelength range of the 5 channels a CNN is very well suited for this classification task. This becomes more apparent with the results(accuracy) of the model.

IV. Results

Describe the results of applying your methods.

V. Discussion

Explain the meaning of the results you obtained. E.g., did you find good/bad performance? Could your work be used in practice? What would help improve it?

References

[1] The Hubble Tuning Fork. [Online]. Available: https://skyserver.sdss.org/dr1/en/proj/advanced/galaxies/tuningfork.asp
[2] SDSS Scope. [Online]. Available: https://www.sdss.org/dr12/scope
[3] K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. A. Prieto, D. An, K. S. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall *et al.*, "The seventh data release of the sloan digital sky survey," *The Astrophysical Journal Supplement Series*, vol. 182, no. 2, p. 543, 2009.
[4] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. C. Nichol, M. J. Raddick *et al.*, "Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies," *Monthly Notices of the Royal Astronomical Society*, vol. 410, no. 1, pp. 166–178, 2010.
[5] A. Khan, E. Huerta, S. Wang, R. Gruendl, E. Jennings, and H. Zheng, "Deep learning at scale for the construction of galaxy catalogs in the dark energy survey," *Physics Letters B*, 2019.

Appendix A