Data science project

Klay Matthew 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

Summarise your report.

Index Terms

Data science

I. Introduction

Motivate the problem and the general approaches you will use to solve it. Explain in enough detail so that a non-expert can understand why the subject-area problem is important and how you propose to solve it.

A. Motivation

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

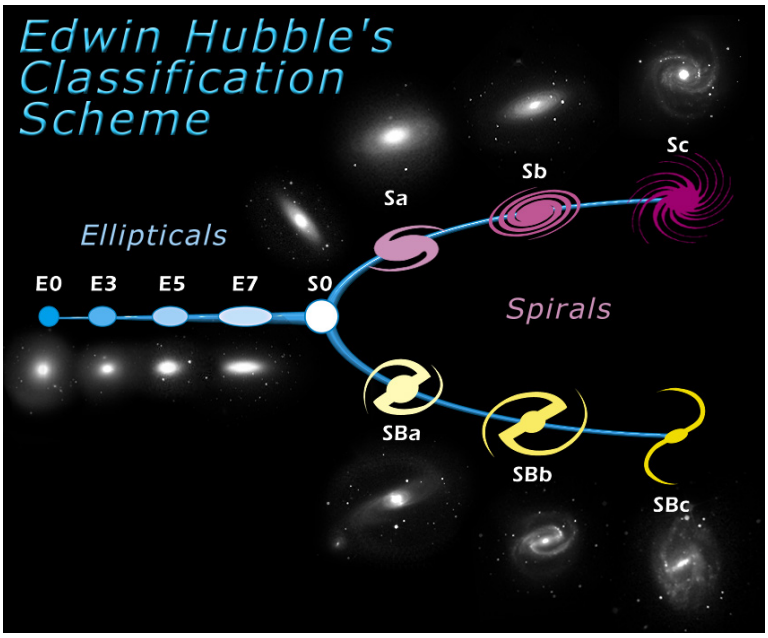

Fig. 1: Edwin Hubble's 'Tuning Fork' classification scheme. Galaxies are classified into two main categories, spiral and elliptical, with further sub-classifications representing minute changes [1].

There are many more differences between the two galaxy morphologies that are not as glaringly. 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. 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. Classifying galaxies based on their morphology therefore has a large impact on research in many fields of astronomy, including things from galactic dynamics to the age of the Universe!

Humans can easily distinguish between the two forms given the collected data is high enough resolution. 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 - a neural network trained to classify images into these categories would be extremely useful to many astronomers. A convolutional neural network (CNN) should be able to identify galaxies much faster than humans, with comparable accuracy.

II. Data

A. Data Source

The data we used for this project came from the Sloan 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. THE SDSS inages 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 volunteer votes, we only used the data entries that have high confidence classifications, that is, 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].

The data we got from SDSS are in Flexible Image Transport System (FITS) format. For each galaxy five FITS files are provided by SDSS, one for each band (u, g, r, i and z). Each FITS file contains a header part and a data part. We removed the header part and resized the images to 200×200 pixels. As we discussed in Section I, '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[1] as well as the

[1] is this the correct way to say it?

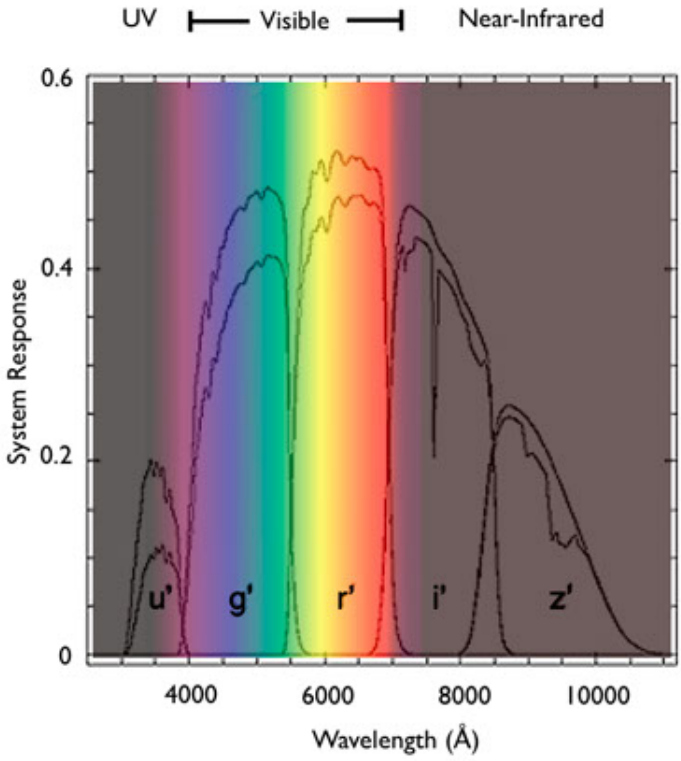

Fig. 2: The 'ugriz' filter shcematic with a spectrum overplotted

	Training	Validation	Testing	Total
Spiral				
Elliptical				

TABLE I: Numbers of data points in each class for training, validation and testing

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.

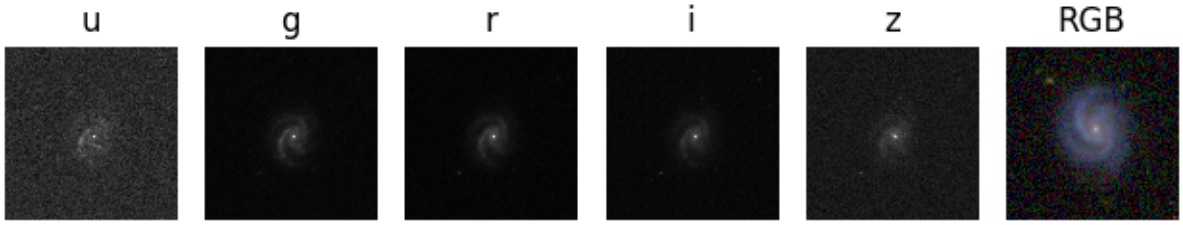

Fig. 3: The 'ugriz' images and the RGB image of the same galaxy

Then we split[2] the data into three sets for training, validation and testing. The number of each class for each set is shown in Table I.

[2]explain further more about how we split the data

III. Method

Describe the data science methods you applied, why you applied them, and how you applied them. Assume that your reader has similar background in data science methods as you do.

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.