Deep Learning Classification of Galaxy Morphologies

Princeton Joseph
Georgia State University
33 Gilmer St SE Atlanta, GA
pjoseph11@student.gsu.edu

Abstract

Galaxy morphology is the study of the physical structure, shape, and appearance of galaxies. Observatories across the world collect large amounts of images and data that must be processed. The Sloan Digital Sky Survey (SDSS) for example, has been in operation for over eight years and has collected images of over 930,000 galaxies [1]. This study aims to implement a deep learning approach to classify galaxy shapes using a pre-trained network.

1. Introduction

The first classification system was introduced by astronomer Edwin Hubble [2]. Galaxies were grouped into four categories: spirals, barred spirals, ellipticals and irregulars (Figure 1). This system had its limitations and over time, additional categories were added. The shape of a galaxy is important information for astronomers to provide insight to how it may have formed and evolved over time. For example: spiral galaxies are thought to have formed from a combination of gravitational collapse, rotation, and accretion of gas and dust; elliptical galaxies are thought to be the result of a merger of two or more spiral galaxies. Observatories and telescopes across the world collect a large amount of data that must be processed. Currently, opensource projects like Galaxy Zoo rely on volunteers to help classify space objects. In the Galaxy Zoo database ~10^5 participants have helped classify nearly one million images [3]. The purpose of this investigation is to determine if a deep learning approach to galaxy morphology performs as well or better than current methods using humans.

1.1. Related Works: Data Collection

Images are collected directly from the SDSS database. The specific catalogue chosen for this project is the "Third Reference Catalog of Bright Galaxies" (RC3). This data set contains 1,244 images of unclassified galaxies. Images are published in full resolution at 1024 x 768 pixels. Thumbnails are provided at 119 x 119 pixels. The images are cropped such that the galaxies are already centered.

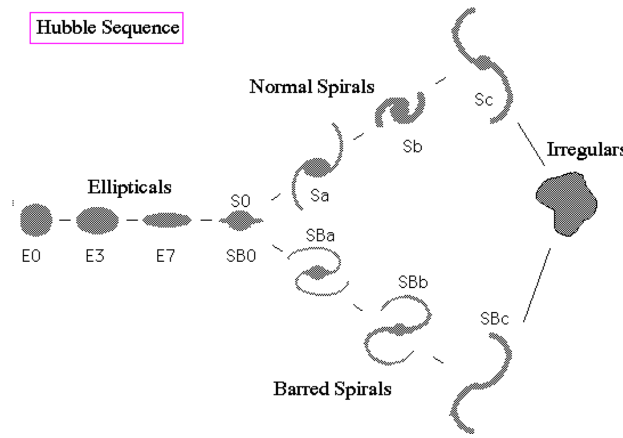

Figure 1: Edwin Hubble's original categories

1.2. Related Works: Image Categories

We chose five categories: Elliptical E0, E3, E6, Spiral Barred, and Spiral Normal. (Figure 2)

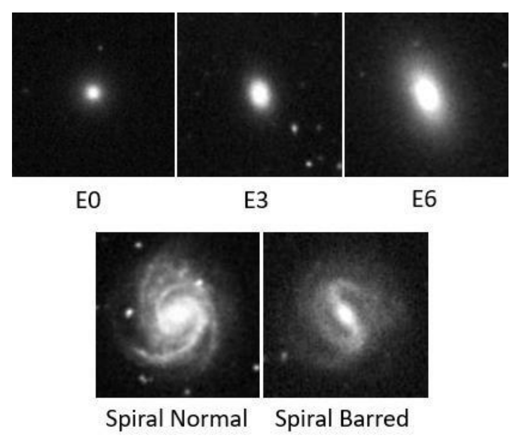

Figure 2: Five categories used in the project

The classification of elliptical shape refers to the amount of elongation. Where 'b' and 'a' are the semi-minor and semi-major axis. (Figure 3)

$$E = 10 \times (1 - \frac{b}{a})$$

Figure 3: Formula used to classify the category of elliptical galaxy

Spiral normal refers to the shape where the arms of a galaxy start and extend from the center. Spiral barred refers to the shape where the arms of a galaxy do not start from the center but rather extend from central bar shape. Astronomers estimate 77% of all galaxies to be spirals and of that two thirds to be barred spirals [4].

1.3. Model and Methods

This project uses the ResNet50 model from TorchVision that was trained on ImageNet database. We use this model as a feature extractor. The last fully-connected layer is replaced by a SoftMax classifier and this allows us to use the pre-trained network as a feature extractor.

We chose a batch size of 256 and to train the model over 40 epochs.

The full dataset consists of 1,244 imagers and are sorted randomly in three folders: train, val, and test. From the original dataset each new category consists of 70%, 10%, and 20% correspondingly.

The folder structure acts as the truth labels as well. (Figure 4)

- test
  - E0
  - E3
  - E6
  - Spiral
  - SpiralB
- train
  - E0
  - E3
  - E6
  - Spiral
  - SpiralB
- val
  - E0
  - E3
  - E6
  - Spiral
  - SpiralB

Figure 4: Dataset structure

1.4. Experiments and Results

The model was run on Google Colabatory cloud-based computing platform. Total runtime was 15,014.65 seconds.

Eval Mode	Loss	Accuracy
Training	1.710809	0.4857
Validation	inf	0.4113
Test	-	0.46

Figure 5: Results of the model after train/val for 40 epochs and final test accuracy.

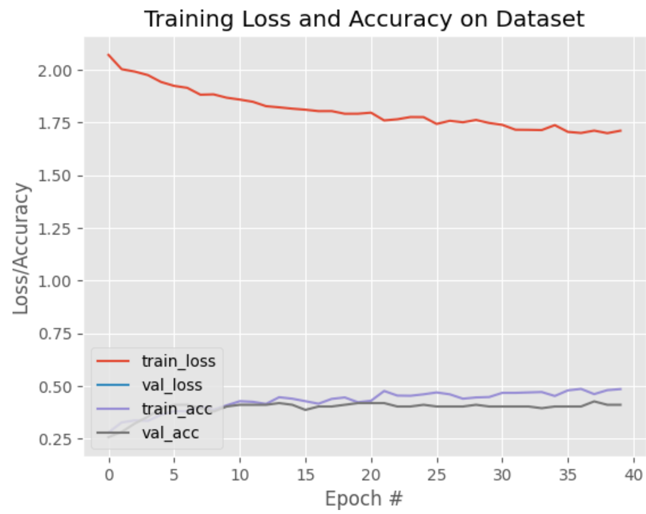

Figure 6: Training and Validation loss and accuracy over 40 epochs

1.5. Conclusions

The results of this experiment are not optimal. There are several sources of error that can be traced to the image dataset. Image processing like adjusting the contrast parameters can bring out more feature especially in the spiral galaxy categories. Additionally, many images contain other artifacts including stars, other galaxies, and noise inherent in the images from the telescope. (Figure 6) While the model was able to learn initially, after epoch 20 there was not much difference. The high validation loss also indicates the issues with the data.

Figure 6: Examples of noise, optical artifacts, and other objects in the image dataset respectively

Another factor is cropping the images. With elliptical galaxies a close crop can cut off the elongation and make different categories look the same. To improve the dataset images from an infrared telescope such as the James Webb rather than an optical one such as the Hubble Telescope, from which these images were collected from can provide more features for a network to learn from, however those images not yet available to the public.

References

[1] Sloan Digital Sky Survey. (n.d.). SDSS: The Sloan Digital Sky Survey. https://classic.sdss.org/
[2] Stovall, J. (n.d.). Lecture 11: The Lives and Deaths of Stars. University of Oregon.
[3] Khandelwal, A., Bhardwaj, A., Chaturvedi, P., & Varshney, L. R. (2019). Machine and Deep Learning applied to galaxy morphology - A comparative study. arXiv:1901.07047
[4] P. B. Eskridge; J. A. Frogel (1999). "What is the True Fraction of Barred Spiral Galaxies?". Astrophysics and Space Science. 269: 427–430. Bibcode:1999Ap&SS.269..427E. doi:10.1023/A:1017025820201. S2CID 189840251.

Appendix

[1] https://github.com/pjoseph11/CSC4851-Project