Draft version March 30, 2017
Preprint typeset using LaTeX style AASTeX6 v. 1.0

HUBBLE'S TUNING FORK: A MACHINE LEARNING APPROACH

Brandon Bergerud[1] and Ossian Mogensen[2]

[1]Department of Physics and Astronomy, University of Iowa, Iowa City, IA 52242
[2]Department of Computer Science, University of Iowa, Iowa City, IA 52242

ABSTRACT

With the introduction of powerful telescopes such as the Hubble Space Telescope, vast quantities of high-fidelity imagery of remote galaxies have become available. Manual analysis of these images by experts has become infeasible, spawning citizen science projects such as Galaxy Zoo. However, the next generation of telescopes are expected to generate enormous volumes of data, going far beyond the capacity even of crowdsourced volunteers. In this study, we will extend the work done on automatic galaxy image classification in the Galaxy Zoo 2 challenge by developing a mapping between the various Galaxy Zoo "classification trees" and the popular Hubble Tuning Fork model. We will build a convolutional neural network to classify galaxies by leveraging the various crowdsourced Galaxy Zoo "gold standard" datasets. The model will be tested against expert-annotated datasets using the Tuning Fork scheme.

1. INTRODUCTION

The size and scope of astronomy datasets has increased dramatically in recent years. The introduction of telescopes such as the Hubble Space Telescope (HST) and projects like the Sloan Digital Sky Survey (SDSS) have given astronomers access to imagery of millions of celestial objects. Traditional methods of data analysis, manually inspecting and classifying celestial objects, have become untenable in the face of this embarrassment of riches of data.

Astronomers have successfully turned to citizen science projects such as Galaxy Zoo to leverage vast numbers of volunteers to help classify objects. The human visual system can, with little effort or training, provide image recognition capabilities that match or exceed the state of the art in computer image recognition.

With the dawn of a new generation of telescopes, astronomy is threatened to be deluged by a sea of data. The GAIA spacecraft will produce a 3D map of over 1 billion astronomical objects. The Thirty Meter Telescope (TMT) and the 40-meter European Extremely Large Telecope (E-ELT) will view the visible universe at unprecedented depth. The Large Synoptic Survey Telescope (LSST) is estimated to generate 15 TB of data each night as it surveys the entire sky. Even these vast sums of data pale in comparision to the 1 TB/s output expected from the monsuvian Square Kilometer Array (SKA). Such enormous sums of data are beyond the ability of crowdsourcing to handle: they can only be handled by leveraging supercomputers, sophisticated algorithms, and machine learning techniques.

The Galaxy Zoo Kaggle challenge was a competition in 2013 to produce a machine learning model that could replicate the classifications of citizen science volunteers on a dataset of 70000 galaxy images captured by HST. The top models performed very well in this challenge, but several questions remain. Can the galaxy classification scheme used by Galaxy Zoo be effectively mapped to astronomical classification schemes such as Hubble's Tuning Fork, or the more complex de Vaucouleurs system? Will machine learning models trained on the Galaxy Zoo dataset generalize well to other sources?

To answer these questions, we will develop a mapping system between the various Galaxy Zoo decision tree classification schemes and the Hubble Tuning Fork scheme. We will develop a machine learning system to

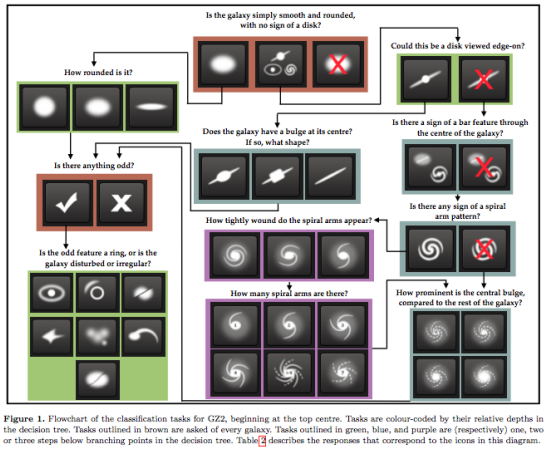

Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

Figure 1. The Galaxy Zoo 2 decision tree.

produce Tuning Fork classifications and train it on data from the Galaxy Zoo projects. We will then locate 3rd party datasets of expert-annotated galaxy images and test our system on these images. This project will investigate the generalizability of the Galaxy Zoo training data and the feasibility of mapping between the two galaxy classification schemes.

2. RELATED WORK

In the astronomical community, the few automated galaxy classification systems have relied on more tradition methods, focusing on aggressive feature extraction algorithms making use of domain knowledge (such as WND-CHARM) to identify relationships among galaxies. These, however, have tended to focus on the more narrow classification of spirals and ellipticals, occasionally including edge-on spirals (S0) and irregular galaxies.

Kuminski and Shamir (2016) made use of "super clean" galaxies from the Galaxy Zoo 1 catalog (Lintott *et al.* 2008) to classify galaxies into spirals and ellipticals. They made use of an algorithm that extracted 2885 numerical descriptions from each image.

The Galaxy Zoo Kaggle challenge showed the power of convolutional neural networks (CNNs) when it comes to image recognition. Participants were forbidden from making use of any domain knowledge when building their models (e.g. ellipticals tend to be red and spirals blue)

Dieleman *et al.* (2015)

3. APPROACH

As dicussed earlier, the existing systems from the Galaxy Zoo Kaggle challenge do an excellent job of replicating the classifications of citizen science volunteers on the Galaxy Zoo 2 dataset. However, it would be useful to develop an automated system based on the large annotated datasets in Galaxy Zoo projects to classify new imagery from other data sources using the popular Hubble Tuning Fork scheme. We will develop a mapping between the two classification schemes and develop such a machine learning system.

Our model will differ slightly from the format of the Kaggle challenge. The Kaggle Galaxy Zoo challenge formulated the problem as a regression on the class probabilities, defined as the ratio of citizen science volunteers that gave a given galaxy a certain classification. To match the structure of our gold standard Tuning Fork scheme data, we will instead treat this as a classification problem and select only those galaxies whose vote fractions are within our chosen threshold for each Hubble type.

4. PLAN

Based on prior work, the best approach to galaxy classification appears to be a Deep Convolutional Neural Network. The python module Keras, using a Tensorflow backend, will be used to build our CNN. We'll likely follow the basis of the inception model [cite], which has been the building block of the top image classifiers in recent years.

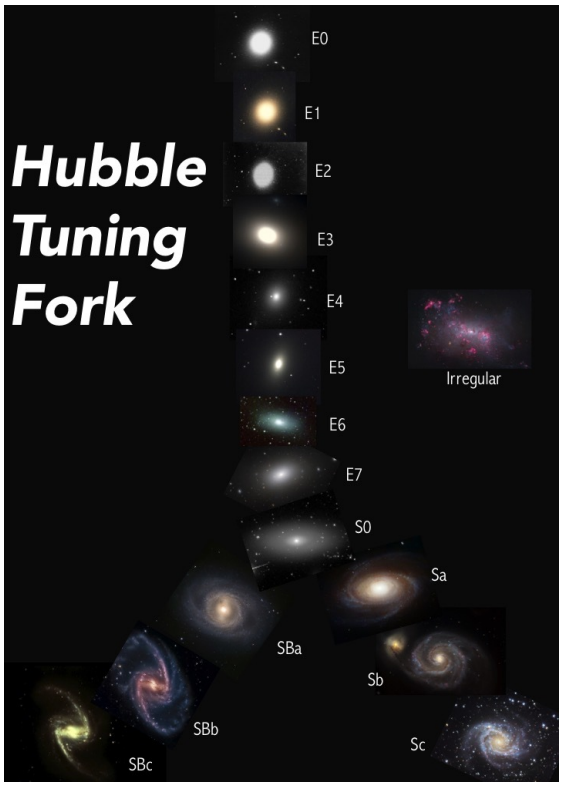

Figure 2. Hubble's tuning fork model. From http://ay17-chusic.blogspot.com/2015/10/20-hubble-tuning-fork.html

Since many of the images used in Galaxy Zoo 2 had significant disagreement among the citizen scientists, we will attempt to achieve better results by incorporating the results from Galazy Zoo 1, Galazy Zoo: Hubble, and Galaxy Zoo: CANDELS, and pruning the dataset. Galaxy Zoo 1, which is the largest of the datasets, will be mostly inadequate for classification purposes as it aimed at determining whether something was a spiral, elliptical, lenticular (S0), or a merging system (irregular). It will, however, provide a good dataset for initial testing to verify that we can separate the basic mophologies. The three remaining Galaxy Zoo projects asked similar questions, allowing for similar mapping schemes to the Hubble tuning fork.

With large datasets, CNNs become computationally expensive, and can take a very long time to run. We will attempt to gain access to GPUs on the cluster. If this proves infeasible, we can reduce our numbers of images by requiring higher thresholds, or reduce the computation costs by using grey-scaled images.

Finally, our model will evaluated on expert-annoted

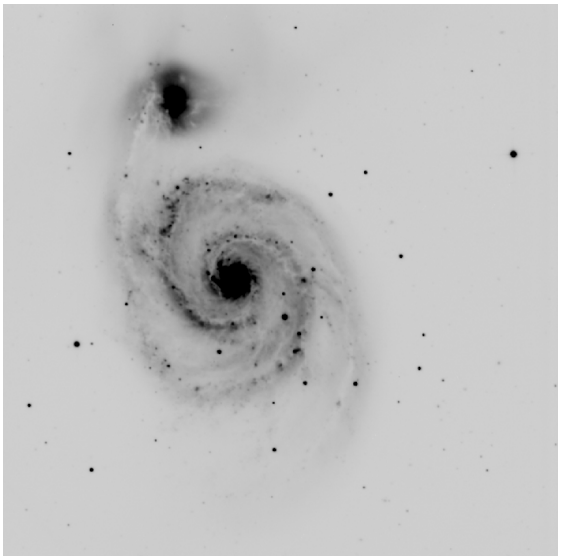

Figure 3. Unfiltered image of the Whirlpool Galaxy (Sb), taken with the Iowa Robotic Observatory..

datasets using images taken by other observatories, such as the Iowa Robotic Observatory.

We would like to acknowledge the work of the Galaxy Zoo team and the countless citizen volunteers in collecting and annotating the massive Galaxy Zoo dataset that makes this work possible.

REFERENCES

E. Kuminski and L. Shamir, ApJS **223**, 20 (2016), arXiv:1602.06854.

C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, MNRAS **389**, 1179 (2008), arXiv:0804.4483.

S. Dieleman, K. W. Willett, and J. Dambre, MNRAS **450**, 1441 (2015), arXiv:1503.07077 [astro-ph.IM]