Draft version March 31, 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 Kaggle 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 classifications using third-party images.

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 in 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), which isn't limited to night time observing. 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.

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

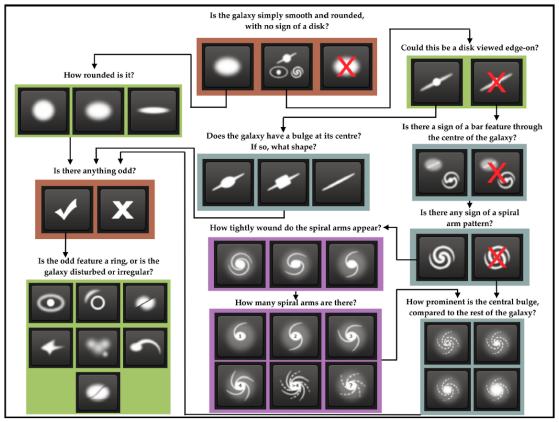

Figure 1. The Galaxy Zoo 2 decision tree. Image from Willett *et al.* (2013).

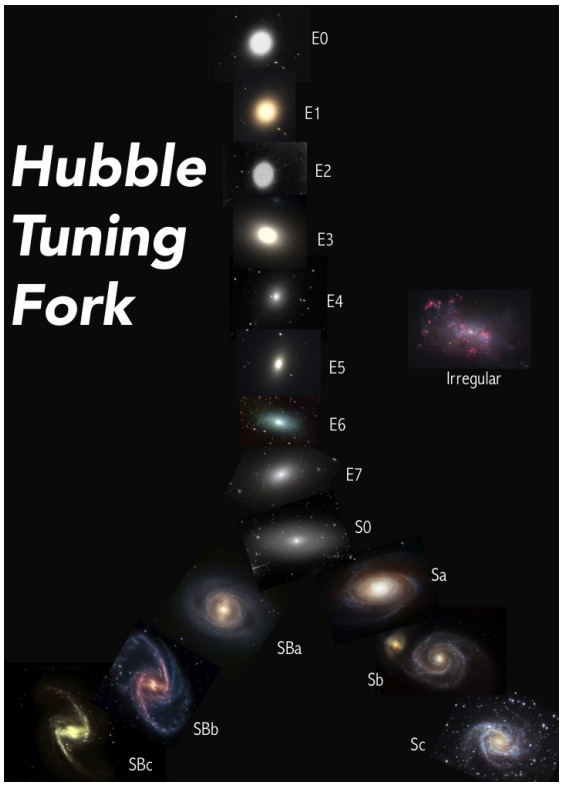

Figure 2. Hubble's tuning fork model. From http://ay17-chusic.blogspot.com/2015/10/20-hubble-tuning-fork.html

scheme. We will develop a machine learning system to 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 and irregular galaxies, and often work with much smaller datasets (see Dieleman *et al.* 2015 for a discussion). While the top methods can achieve $\sim$ 95% when separating ellipticals and spirals, they tend to perform much worse when the number of categories increases (de la Calleja and Fuentes 2004).

One example of the simple classification approach was done by Kuminski and Shamir (2016), who, rather uniquely, made use of the "super clean" galaxies from the Galaxy Zoo 1 catalog (Lintott *et al.* 2008) to classify 3 000 000 galaxies into spirals and ellipticals. They made use of an algorithm that extracted 2885 numerical descriptions from each image (... not really sure how they did their classifying)

Gauthier *et al.* (2016), students in Prof. Ng's machine learning class at Stanford, recently looked at several machine learning methods for classifying galaxies using the GZ2 dataset. While acknowledging the difficulty of directly classifying to the Hubble types, they sought to bridge the gap by modeling certain features, such as "roundness" and "diskiness". They utilized the GZ2 decision tree to assign each galaxy to one of five categories: disc, spiral, elliptical, round, and other. In their preprocessing stage, images were cropped to reduce the file size, as well as reduce the number of nearby sources contaiminating the images. The galaxies were then rotated to align the principle axis, before proceeding with a background subtraction.

To further reduce the dimensionality of the problem, the authors applied principal component analysis (PCA), selecting the top 125 components to maintain $>$ 99% of the variance. To classify the galaxies, they utilized a support vector machine (SVM) with a radial basis function (RBF) kernel, a decision tree, random forest, k-nearest neighbors, and an AdaBoost classifier, determing the classification accuracy using 10-fold cross validation. Overall, random forest produce the best results, achieving 67% accuracy. The poor success rate lead them to look into predicting probabilities (regression) rather than directly modeling the classes, similar to the Galaxy Zoo Kaggle challenge. They achieved better results in this regard, attaining $\sim$ 95% accuracy.

Overall, the biggest source of error was misclassifying spiral galaxies into the "other" category, which they attributed to the faintness (low signal-to-noise) of the spiral arms in many images. In addition, examining Figure 3 in their paper and comparing the original image with the 125 PC image, it appears that their method may hinder extracting the spiral arms by smoothing the disk and making classification more difficult. While this may be necessary for more traditional machine learning methods, deep learning can deal directly with the large feature space.

The Galaxy Zoo Kaggle challenge showed the power of convolutional neural networks (CNNs) when it comes to galaxy classification. Rather than relying on domain knowledge, the models had to learn to identify features on their own and were able to successfully reproduce the probabililty distributions of the citizen scientists. The processing pipeline for the top performing model (Dieleman *et al.* 2015) is schematically illustrated in Figure 3, which we shall examine next.

The winning algorithm was an ensemble method, av-

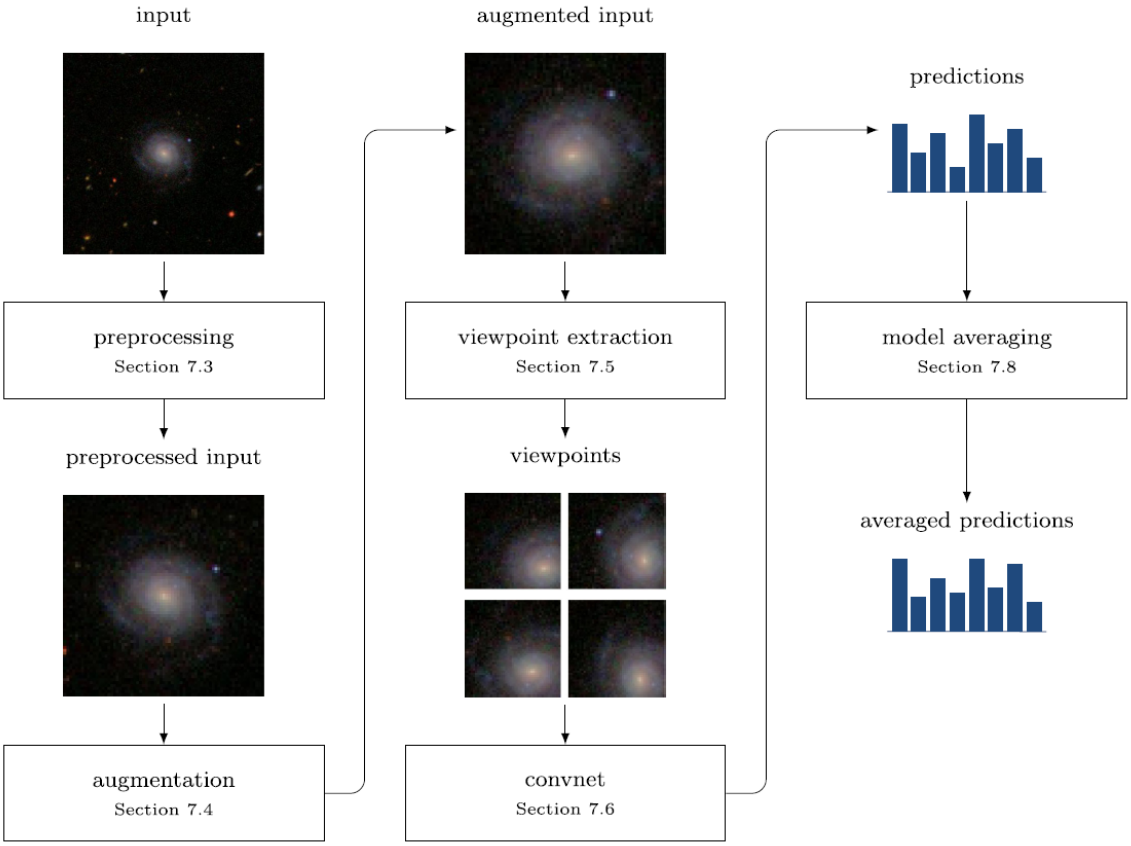

Figure 3. Processing pipeline for the top model in the Galaxy Zoo Kaggle competition. From Dieleman *et al.* (2015).

eraging the results of many different CNNs. In the preprocessing stage, the image was cropped and rescale several times down to 69×69 pixel images, which occasionally removed part of the galaxy. In some of their models, they used SExtractor to estimate the position and size of the galaxy, allowing them to center and rescale the galaxies to a standardized size. In addition, gray-scaling was examined, although this lead to worse results.

Due to the limited sample size, the number of images was increased by performing random pertubations, such as rotating, translating, scaling, and flipping, as well adjusting the color brightness on demand so that the model was never trained on the exact image twice. In addition, the number of images was increased by rotating, flipping and cropping each image into 16 different, but overlapping, 45×45 pixel images respresenting different viewpoints. Each of the 16 images were then passed together through the CNN, which performed several convolutions and pooling before concatenating the results and passing through a couple fully connected layers to output the final categorical probabilities. The probabilities were then averaged over 17 models.

Overall, the model did quite well, achieving $\sim$ 99% accuracy. It stuggled most with the larger angular sized galaxies (more nearby), as well as those that were not radially symmetric.

3. APPROACH

As dicussed earlier, the existing systems from the Galaxy Zoo Kaggle challenge do an excellent job of replicating the voting patterns 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 Galaxy Zoo datasets to classify new imagery from other sources using the popular Hubble Tuning Fork scheme. While this can be done to some extent using the kaggle models, it requires cross-correlating expertly annoted images to find the optimal probability cutoffs to transform the probability distributions to Hubble T-types, adding an additional layer of complexity that the machine wasn't required to learn. We will develop a mapping between the two classification schemes and develop such a machine learning system to directly classify images.

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

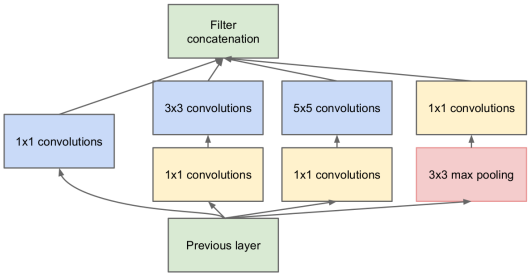

Figure 4. Inception block. The top image recognition CNNs in recent years use many inception blocks in their networks. From Szegedy *et al.* (2014).

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. This will favor the more nearby galaxies, whose properties the top performing model in the Kaggle competition had a harder time predicting accurately, hopefully, leading to an improvement in that regard. In addition, it would serve as a more interactive tool that could serve as a complement to the galaxy classification lab in *Stars, Galaxies, and the Universe*.

Based on prior work, the best approach to galaxy classification appears to be a Deep Convolutional Neural Network. The top image recognition CNNs in recent years have used the inception model (Szegedy *et al.* 2014) as a building block in their networks (Figure 4). We will follow this trend.

Since many of the images used in Galaxy Zoo 2 had poor consensus 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, edge-on disk, 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.

4. PLAN

- Map GZ to HTF
- Build CNN – test on GZ1 data
- Test on HTF types; fine-tune parameters
- Test against HTF types from other sources

There are two major bottlenecks to our task: completing the project within a month and computational resources.

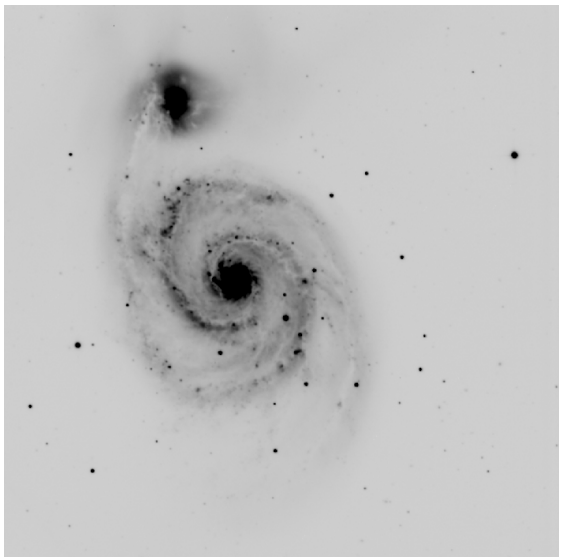

Figure 5. Unfiltered image of the Whirlpool Galaxy (Sb), taken with the Iowa Robotic Observatory.

CNNs and inherently computationally expensive and can take a long time to train, especially when using a large number of images. To cut down on computational time, GPUs are

These contraints can be contrasted with the winning Kaggle model, which tested around 100 different network schemes and averaged 17 of them to get the final probabilities. (70 hours to train)

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

K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. C. Keel, T. Melvin, R. C. Nichol, M. J. Raddick, K. Schawinski, R. J. Simpson, R. A. Skibba, A. M. Smith, and D. Thomas, MNRAS **435**, 2835 (2013), arXiv:1308.3496.

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

J. de la Calleja and O. Fuentes, MNRAS **349**, 87 (2004).

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.

A. Gauthier, A. Jain, and E. Noordeh, (2016).

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ArXiv e-prints (2014), arXiv:1409.4842 [cs.CV]