Proposal Template for Machine Learning Course

Brandon Bergerud

Department of Astrophysics

University of Iowa

Iowa City, IA 52242  
 brandon-bergerud@uiowa.edu

Ossian Mogensen

Department of Computer Science

University of Iowa

Iowa City, IA 52242   
ossian-mogensen@uiowa.edu

# Abstract

With the introduction of powerful space-borne telescopes such as the Hubble, 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 volume of data is now growing 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 Galaxy Zoo 2 “classification tree” and the popular Hubble Tuning Fork model. We will build an automated system that classifies new images beyond the Galaxy Zoo 2 dataset based on the Hubble Tuning Fork classification scheme by leveraging the crowdsourced Galaxy Zoo 2 “gold standard” dataset to train a model and applying that model to new data. 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 space borne telescopes such as the Hubble and projects such as the Sloan Digital Sky Survey has given scientists 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.

In the subfield of galaxy classification, scientists have 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.

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 70 000 galaxy images captured by the Hubble. 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 the popular Hubble Tuning Fork classification scheme? 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 Galaxy Zoo 2 “decision tree” classification scheme and the Hubble Tuning Fork scheme. We will develop a machine learning system to produce Tuning Fork classifications and train it on the Galaxy Zoo 2 data. 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 2 training data and the feasibility of mapping between the two galaxy classification schemes.

# 2 Related Work

Brandon – wanna tackle this? You would know more here.

2.1 The Galaxy Zoo Challenge

Some explanation of the Galaxy Zoo background and other papers

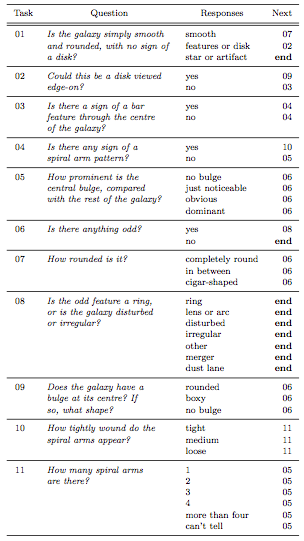


Figure 1: The Galaxy Zoo 2 decision tree

3 The Proposed Work

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 and using the Galaxy Zoo 2 “decision tree” classification scheme. However, it would be useful to develop an automated system based on the large annotated dataset in Galaxy Zoo 2 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 with the class defined as the highest probability class given by the citizen scientist classifications.

# 4 Plan

Based on prior work, the best approach to classify galaxy imagery is a Deep Convolutional Neural Network. The training set of ~70000 images should allow for the training of a Deep CNN from the ground up. If time permits, we also intend to investigate transfer learning approaches. We plan to develop a mapping between the Galaxy Zoo 2 classification scheme and the Hubble Tuning Fork. We will then build a Deep Convolutional Neural Network to classify galaxy images according to the Hubble Tuning Fork, using the mapped classifications from the Galaxy Zoo Kaggle challenge as gold standard data. This model will be evaluated on expert-annoted datasets that are available using the Hubble Tuning Fork classification scheme.

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 2 dataset that makes this work possible.

# References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to small (9 point) when listing the references.

1. Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.
2. Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.
3. Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* 15(7):5249-5262.

3