Proposal Template for Machine Learning Course

XXX*
Department of Computer Science
University of Iowa
Iowa City, IA 52242
xxx@uiowa.edu

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

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

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 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.

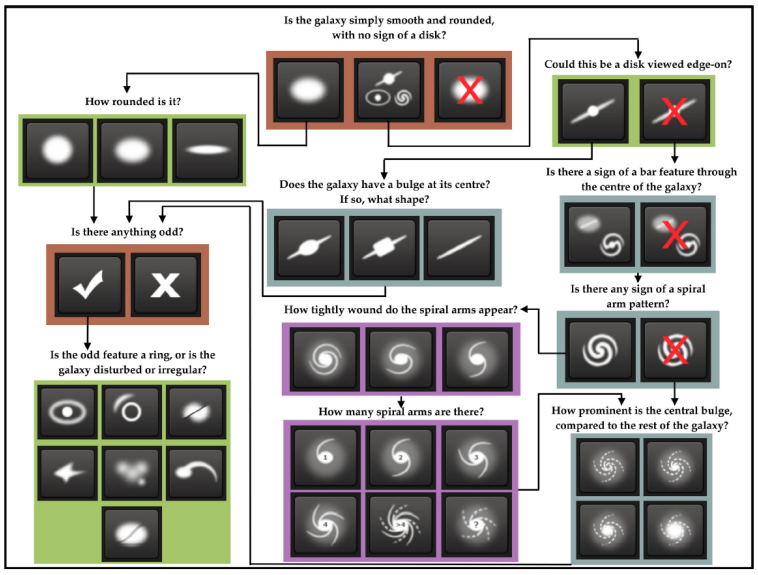

Figure 1: The Galaxy Zoo 2 decision tree. Image from Willett et al. (2013).

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.

One example of the simple classification approach was done by Kuminski & Shamir (2016), who, rather uniquely, made use of the "super clean" galaxies from the Galaxy Zoo 1 catalog (Lintott et al., 2008) to classify 900 000 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 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 winning model created 16 transformations for each image through the use of rotations and translations and trained the model on all 16 at once using convolutional layers and pooling.

2.1 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

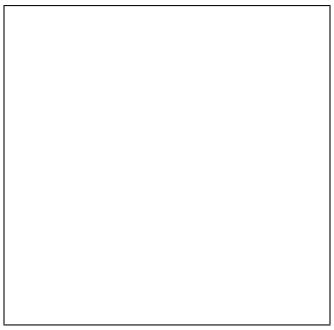

Figure 2: Sample figure caption.

2.2 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

3 The Proposed Work

Describe your proposed work in this section.

4 Plan

Describe your plan for the project. What data you are going to use to evaluate your methods? What are the baselines that you want to compare? How will you develop your methods? A timeline with important milestones is always perferred.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

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.

References

Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441

Kuminski, E., & Shamir, L. 2016, ApJS, 223, 20

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179

Szegedy, C., Liu, W., Jia, Y., et al. 2014, arXiv:1409.4842

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835