Deep Learning Redshift Estimation

Matthew Ferguson

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of Master of Science in Computer Engineering.

Dr. Creed Jones

Dr. Tim Talty

Dr. Paul Plassmann

December 20th, 2024

# Abstract

We analyze () from the Sloan Digital Sky Survey (SDSS) and

In this study, we develop a convolutional neural network to estimate redshift of galaxies. Flood fill, a technique commonly used in image processing, is applied to segment galaxies from the background and surrounding noise. The segmented regions are then characterized by their color histograms, which serve as features for a deep learning model aimed at estimating spectroscopic redshift.

Our approach leverages the simplicity of flood fill for segmentation, followed by the richness of color histograms to encapsulate the galaxies' photometric characteristics. The model’s performance suggests that combining classical image segmentation techniques with modern machine learning methods can provide valuable insights into galaxy classification and feature extraction, particularly with redshift-related phenomena.

# Abstract for general audience

There is a large amount of astronomy data and every year surveys create even more data. There is too much for humans to analyze manually. With the assistance of artificial intelligence and machine learning we can analyze this data and discover more about nature.

Here we show a technique to predict the redshift of galaxies by using machine learning.

We obtained 200,000 image sets centered on galaxies from the Sloan Digital Sky Survey. Each image set contains multiple bands of the same location in a different segment of the electromagnetic spectrum ultraviolet, green, red, infrared, etc. We cut the galaxies out from the rest of the image using the flood fill algorithm. These cutouts were used to train a deep learning model to predict the redshift of galaxies. We achieved an overall

# Acknowledgements

# Contents

[Abstract 2](#_Toc178357480)

[Acknowledgements 2](#_Toc178357481)

[Table of Contents 2](#_Toc178357482)

[List of Figures and Tables 2](#_Toc178357483)

[Abbreviations 2](#_Toc178357484)

[Introduction 2](#_Toc178357485)

[Literature Review 2](#_Toc178357486)

[Methodology 2](#_Toc178357487)

[Results 2](#_Toc178357488)

[Discussion 2](#_Toc178357489)

[Conclusion 2](#_Toc178357490)

[References 2](#_Toc178357491)

# List of Figures and Tables

# Abbreviations

# Introduction

Background

Astronomy has benefited from improvements in survey technology. Many large surveys have been conducted and more continue to be conducted. Surveys have created an enormous amount of astronomy data which will continue to grow more rapidly. Currently, creation of astronomy data is on the order of 100 petabytes per year [1]. Analysis of this data increasingly requires automation. Machine learning is a technology that provides methods to efficiently analyze large sets of astronomy data.

Spectroscopic Machine learning excels at classification and regression tasks for large datasets.

Try to describe the ramifications of an error in ‘classification or regression’. What is the burden on this system to work correctly, what performance metrics are necessary to advance the field. How often do I have to get it right to be contributing to the field, estimate of the required performance to be effective in this field.

The sheer volume of data means that human analysis is not feasible. Given that we arguably live in the golden age of astronomy, where data mapping the universe is available in a magnitude above previous eras, contributions which automate analysis of analyzing astronomy data is a valuable contribution to human knowledge.

## Aims

The aim of this paper is to develop a machine learning method to analyze a large set of astronomy data. The specific application we have selected is to estimate the redshift of galaxies using deep learning.

This work is in fulfilment of a master’s thesis, and we have set out to validate existing work in the intersection of machine learning and astronomy.

We also aim to provide value to those approaching the intersection of astronomy or machine learning for the first time by summarizing the existing state of research.

Objectives:

Review the state of machine learning and astronomy

**List contributions**

# Literature Review

Our literature review centralizes information at the intersection of astronomy and machine learning. We will summarize machine learning concepts, and discuss the state of astronomy research across several dimensions:

* Data types in astronomy
* Prominent astronomy data surveys
* Domains of machine learning tasks in astronomy
* Previous works that applies machine learning to astronomy

This literature review aids us in understanding the current state of affairs in ML and astronomy and how the author or reader might themselves expand on previous contributions to provide value to this field.

## Machine Learning

Machine learning models are broadly categorized based on the way models learn from data. These categories are:

* Supervised Learning – models trained on labeled data
* Unsupervised Learning – models trained on unlabeled data.
* Semi-Supervised Learning – models trained
* Self-Supervised Learning – models trained on unlabeled data which create labels
* Reinforcement Learning – models trained by over generations by an environment

Supervised learning requires pairs of labels and events. Events contain the features used as a model input, and the set of labels is the outcome of every event. Upon completion of training, supervised models accept unseen events and predict the corresponding label.

We have a large amount of redshift labels for provided by SDSS so we will take advantage of these labels by using a supervised learning model. Supervised models come in several types. The taxonomy of models given byPer Ref[1912.02934 (arxiv.org)](https://arxiv.org/pdf/1912.02934), the primary taxonomy of supervised machine learning models are:

* Classification – models which predict a categorical variable.
* Regression – models which predict a continuous variable.
* Clustering – models which group examples by proximity in some feature space
* Forecasting
* Generation and Reconstruction
* Discovery
* Insight

## Astronomy Data

Per ref, the data types of astronomy are:

* Spectroscopy
* Photometry
  + Images
* Light Curves
* Time Series
* Catalogues
* Simulations

Spectroscopy provides a continuous magnitude response Spectroscopy measures intensity over a range of wavelengths. Spectroscopy provides measures the electromagnetic radiation of a source and can be used to derive the atomic and molecular composition of the source along with its redshift.

Photometry is the precise measurement and study of visible light. It quantifies light

Photometry is the measurement of brightness of an object through a filter.

Images are a 2-dimensional representation of an object. We represent images in computing as matrices, where the values at each location in the matrix is an intensity (or set of intensities such as with RGB images). Astronomy images represent celestial objects. The intensity value of these images can be not only in the band visual light but at specific wavelength bands of light. These bands may be beyond the range of human vision such as with ultraviolet, or radio images. These bands may be subsets of visual light such as red or green, and there may be bands of images that represent both visible and invisible light.

Light Curve

Time Series

Catalogue

Simulation

A chart with black text and black dots

Description automatically generated with medium confidence

Classification and regression are basic machine learning tasks which scale well. A model can predict the morphology or redshift of a galaxy.

## Data Surveys

SDSS

## Astronomy Problems

## Astronomy Models

**Previous Work**

Traditional statistical methods were applied such as principal component analysis for morphological classification of spiral galaxies [Whitmore, 1984], or for quasar detection [Francis et al., 1992]. PCA is a standard statistical technique in astronomy analysis today.

There are significant previous works in ML and astronomy. Beginning in the early 1990’s astronomers began taking advantage of large labelled datasets with machine learning. Decision Trees were used for star-galaxy separation [Weir et al., 1995] and for classification of galaxy morphology [Kriessler et al., 1998, Owens et al., 1996]. Random forests gained popularity in the 2000’s and models estimating photometric redshift emerged[Carrasco Kind and Brunner, 2013]. Boosted decision trees, such as AdaBoost, have emerged more recently for both photometric redshifts [Hoyle et al., 2015a] and for star-galaxy separation.

**Rotation-invariant convolutional neural networks for galaxy morphology prediction** (2015) by Dieleman et al.:

<https://academic.oup.com/mnras/article/450/2/1441/979677>

This study examines the feasibility of using convolutional neural networks (CNNs) to classify galaxy morphology. They point out that the Sloan Digital Sky Survey (SDSS) has a readily available and large cohort of images of galaxies and that the Galaxy Zoo project has used crowdsourcing to provide many labelled galaxies. Using the labels provided by Galaxy Zoo, the best CNN was trained with an accuracy of 99% in predicting the label a human would apply to most galaxy morphology Galaxy Zoo questions. This study was highly cited and the earliest I found in deep learning astronomy applications.

A diagram of a diagram of a diagram

Description automatically generated

The architecture of the best model is shown above. This study is focused on morphology alone, rather than redshift. However, it sets a foundation for subsequent studies in deep learning astronomy tasks, and morphology classification is perhaps another avenue for research progress.

**Photometric redshifts from SDSS images using a Convolutional Neural Network**

(2018) by Pasquet et al.

[[1806.06607] Photometric redshifts from SDSS images using a Convolutional Neural Network (arxiv.org)](https://arxiv.org/abs/1806.06607)

A French team developed a Deep CNN classifier to estimate photometric (not spectrographic which is several orders of magnitude more accurate than photometric redshift and readily available as a label in SDSS) redshifts and the associated probability distribution functions (PDF) for galaxies in the Main Galaxy Sample of the SDSS for redshifts of z<0.4. Interestingly, the researchers state they use only information present in the images and no feature extraction which is a point which must be understood better. Do they use just features in the image or is there meaning that they perform a kind of convolution that has no intermediate features. Input data is 64x64 images centered on the spectroscopic targets and the galactic reddening value on the line of sight. A deeper understanding of these inputs should be obtained during my research. The researchers are pleased with their performance metrics and predictive power are best yet obtained for the time of publication. The researchers conclude that they are fundamentally limited by the signal to noise ratio contained in SDSS images, and that their method will scale better with improved measurements from upcoming surveys.

**Photometric Redshift Estimation with a Convolutional Neural Network: NetZ**

(2021) by Schuldt et al.

[[2011.12312] Photometric Redshift Estimation with a Convolutional Neural Network: NetZ (arxiv.org)](https://arxiv.org/abs/2011.12312)

A German team developed A CNN “NetZ” to estimate photometric redshift which was trained on data from the Hyper Suprime-Cam Subaru Strategic Program (HSC SSP) in five different filters. The team used images of galaxies and their photometry in contrast to previous methods which only used photometry. The range of redshifts for performance is for 0<z<4, and performed well in the high z range especially on luminous red galaxies. The team publishes 34 million predictions and sees value in upcoming surveys that provide billions of high-quality images for future work.

**Photometric redshift estimation via deep learning**

By A. D’Isanto and Polsterer

[Photometric redshift estimation via deep learning - Generalized and pre-classification-less, image based, fully probabilistic redshifts | Astronomy & Astrophysics (A&A) (aanda.org)](https://www.aanda.org/articles/aa/full_html/2018/01/aa31326-17/aa31326-17.html)

D’Isanto and Polsterer developed a probabilistic photometric redshift . The team developed a modified version of a deep CNN which was combined with a mixture density network. Their estimates are expressed as Gaussian mixture models to represent the PDF in redshift space. The resulting model can make predictions independent of image type (galaxy, quasar, start) which represents an improvement over other results the team concludes. A better inspection of performance metrics is warranted on this one I believe.

**A Deep Learning Approach for Characterizing Major Galaxy Mergers**

(2021) by Koppula et al.

[2102.05182.pdf (arxiv.org)](https://arxiv.org/pdf/2102.05182.pdf)

Abstract:

“Fine-grained estimation of galaxy merger stages from observations is a key problem useful for validation of our current theoretical understanding of galaxy formation. To this end, we demonstrate a CNN-based regression model that is able to predict, for the first time, using a single image, the merger stage relative to the first perigee passage with a median error of 38.3 million years (Myrs) over a period of 400 Myrs. This model uses no specific dynamical modeling and learns only from simulated merger events. We show that our model provides reasonable estimates on real observations, approximately matching prior estimates provided by detailed dynamical modeling. We provide a preliminary interpretability analysis of our models, and demonstrate first steps toward calibrated uncertainty estimation.”

**The PAU Survey: Photometric redshift estimation in deep wide fields**

(2023) by Navarro-Girones et al.

[[2312.07581] The PAU Survey: Photometric redshift estimation in deep wide fields (arxiv.org)](https://arxiv.org/abs/2312.07581)

This paper presents the application of deep learning techniques to estimate photometric redshifts using multi-band photometry. Photo-z was estimated across 40 narrow bands of the PAUS and the broad bands of the CFHTLEns and KiDS.

**Deep learning for galaxy surface brightness profile fitting**

(2018) by D Tuccillo et al.

[Deep learning for galaxy surface brightness profile fitting | Monthly Notices of the Royal Astronomical Society | Oxford Academic (oup.com)](https://academic.oup.com/mnras/article/475/1/894/4725057)

Abstract:

“Numerous ongoing and future large area surveys (e.g. Dark Energy Survey, EUCLID, Large Synoptic Survey Telescope, Wide Field Infrared Survey Telescope) will increase by several orders of magnitude the volume of data that can be exploited for galaxy morphology studies. The full potential of these surveys can be unlocked only with the development of automated, fast, and reliable analysis methods. In this paper, we present DeepLeGATo, a new method for 2-D photometric galaxy profile modelling, based on convolutional neural networks. Our code is trained and validated on analytic profiles (HST/CANDELS F160W filter) and it is able to retrieve the full set of parameters of one-component Sérsic models: total magnitude, effective radius, Sérsic index, and axis ratio. We show detailed comparisons between our code and GALFIT. On simulated data, our method is more accurate than GALFIT and ∼3000 time faster on GPU (∼50 times when running on the same CPU). On real data, DeepLeGATo trained on simulations behaves similarly to GALFIT on isolated galaxies. With a fast domain adaptation step made with the 0.1–0.8 per cent the size of the training set, our code is easily capable to reproduce the results obtained with GALFIT even on crowded regions. DeepLeGATo does not require any human intervention beyond the training step, rendering it much automated than traditional profiling methods. The development of this method for more complex models (two-component galaxies, variable point spread function, dense sky regions) could constitute a fundamental tool in the era of big data in astronomy.”

# Methodology

Our methodology has several parts:

Data Exploration – What data do we leverage and what have we learned about it

Preprocessing – How is data prepared for modelling

Segmentation – How are galaxies isolated from the background and neighboring objects

Feature Extraction – How is convolution configured

Training – WHAT UP BITCHES

Implementation

## Data Exploration

Labels and identifiers were obtained from SDSS via CasJobs using the query written by Pasquet et al and provided in appendix A. This query returns ~500,000 labelled galaxies. For every galaxy object there is a corresponding spectroscopic redshift value.

We plot the distribution of these galaxies in the sky along with their labelled redshift:

Diagram

Description automatically generated

We can see that the section of galaxies selected represents a broad region. Of the galaxies selected for our model we show the distribution of redshift values below:

Chart, histogram

Description automatically generated

We can see that there is a significant amount of redshift values between 0 and 0.2, with

Images

SDSS is a multi-band imaging and spectroscopic redshift survey using a dedicated 2-5 meter telescope at Apache Point Observatory in New Mexico.

It provides photometric observations in UGRIZ bands. We take data from data release 12 (DR12, Alam et al. 2015). The SDSS CasJob website is used to obtain DR12 data. 100,000 labels were obtained from SDSS which were classified as galaxies with redshifts less than 1.0.

To obtain all of the images used for our study we leverage the SDSS API. We obtain 10,000 samples to use for training and test.

Preprocessing

Segmentation

Feature Extraction

# Results

Compare to other work what are my performance results

# Discussion

My unique contribution is the use of flood fill segmentation, highlight this as a key novelty in methodology. Highlighted in the abstract. Reference or two for flood fill.

# Conclusion

Suggestions for future work, think very carefully about the conclusions.

# Appendices

# References

[1][The State of Data in Astronomy (dataiku.com)](https://blog.dataiku.com/the-state-of-data-in-astronomy" \l ":~:text=With%20some%20of%20these%20telescope,and%20unified%20access%20a%20nightmare.) [Data in Observational Astronomy | SpringerLink](https://link.springer.com/chapter/10.1007/978-3-031-29937-7_2)