Applying Deep Learning to the Galaxy Photo-z Problem

Samantha E. Thrush[1]⋆ and Robert J. Brunner[1,2,3,4]
[1]*Department of Astronomy, University of Illinois, Urbana, IL 61801 USA*
[2]*Department of Physics, University of Illinois, Urbana, IL 61801 USA*
[3]*Department of Statistics, University of Illinois, Champaign, IL 61820 USA*
[4]*National Center for Supercomputing Applications, Urbana, IL 61801 USA*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

When large data sets of galaxies are collected, their associated redshifts are usually collected by finding their photometric redshifts, which takes additional time for a survery to complete. However, if deep learning via deep convolutional neural networks (ConvNets) are utilized, photometric redshifts of galaxies can be predicted without the need for photometry to be completed by the associated survey. In this paper, we present an application of ConvNets on reduced and pixelated galaxy images and redshift data from the Sloan Digital Sky Survey (SDSS). We will show that ConvNets are able to accurately predict the redshift of a sample galaxy image in a way that is competitive with previous works on this subject.

Key words: methods: data analysis – techniques: image processing – methods: statistical – surveys – galaxies:statistics.

1 INTRODUCTION

Before the advent of photometric redshift techniques, finding a galaxy's redshift required the collection of the galaxy's spectra so as to ascertain its recession speed and thus its distance and redshift from the observer using Hubble's Law (Hubble 1929). Unfortunately, even in large scale surveys like the Sloan Digital Sky Survey, spectroscopy can only be done on a select number of targets at a time due to time constraints since spectra take approximately an hour to observe (nine spectroscopic plates are completed in a night). Additionally, due to the usage of an aluminum plate and optical fiber system, or a single-slit setup (which are used to prevent the spectra from being tainted by extraneous light) only a small subset of galaxies may be observed at a time, thus further reducing the number of objects a survey can collect spectroscopic data on.

In contrast, photometric images allow many more targets to be observed since each image taken requires a comparatively shorter exposure time (54.1 seconds for each photometric band of the Sloan Digital Sky Survey) (York 2000). In addition, since there is no need to block out extraneous light, photometry can be done without the use of aluminum observing plates or slits, which allows observers to observe all of the objects within a single frame. Because of this, the idea of only using an object's image to ascertain its redshift is extremely attractive.

TALK ABOUT THE HUGE AMOUNT OF DATA AND WHY PHOTOZS WOULD HELP (BY WHAT FACTOR)!!

Photometric redshift techniques did not come without pitfalls. The initial techniques (Baum 1962; Puschell et al. 1982; Koo 1985; Connolly et al. 1995) could only work on small subsets of galaxies with certain characteristics. Fortunately, the application of machine learning to the problem (citations needed) solved two main issues: it was then possible to find photometric redshifts of all types of galaxies, to varying degrees of success (Hildebrandt et al. 2010); and these techniques are particularly well equipped to work with very large amounts of data (citation needed). One such successful technique was deep learning with the use of neural networks. Although this technique needs large training sets in order to accurately predict redshifts, such techniques have achieved prediction accuracy of over 95% in previous works (Hoyle 2015).

In this paper, we will first review previous works, present integral themes, and take a foray into the underlying mathematics of this subject in Section 2. In Section 3, the Deep Learning code and the selection of the data set from the Sloan Digital Sky Survey (SDSS) will be discussed, as well as key ways this body of work differs from previous works. Section 4 will contain the results of the Deep Learning code, which will be subsequently discussed and analyzed. Finally, Section 5 will contain the conclusions of this paper as well as future goals.

2 BACKGROUND

Ascertaining the redshifts of galaxies using spectroscopic data has classically been a time-consuming pursuit due to the necessary exposure times for a spectroscopic image. Fortunately, with the advent of photometric redshift techniques, it is possible to ascertain the redshift of a galaxy only using photometric data, but this always comes at a cost, be it large training sets or a reduced number of candidates that the technique may be used upon. The project at hand is composed of many parts, so in order to organize the logic

⋆ thrush2@illinois.edu

behind all of the choices and decisions made, it is important to reflect upon previous research and important themes that will present themselves later in the paper. First, previous papers on photometric redshifts will be presented. Then, the Sloan Digital Sky survey and its bearing on the project will be discussed. Finally, we will consider previous projects that utilize Neural Networks for its applications to images and regressing galaxy images to find their redshifts.

2.1 Previous Works

Using an object's photometric image to find its redshift (henceforth known as finding an object's photo-z) was first successfully done in the paper "Photoelectric Magnitudes and Red-shifts" wherein the author created a makeshift spectrum by gathering the magnitudes of each galaxy at multiple bandpasses and creating plots of wavelength versus magnitude, which made a make-shift spectra. Unfortunately, the main downfall of his method was that it only worked for elliptical galaxies. Additionally, due to the fact that the magnitudes of the galaxies were measured with a photometer, this meant that Baum could only gather data on one galaxy at a time, thus further harming the feasibility of scaling up this technique to a large number of galaxies (Baum 1962).

Other early photo-z techniques included the use of color-shape plots, which worked in the following way: used lines of constant redshift lines that worked for all types of galaxies(Koo 1985), but each of these techniques only worked on a small subset of galaxies.

in a widespread manner until ***YEAR*** (citation needed). This was when ***Event*** happened. The main limiting factor to the advent of finding photo-z's was computational power; *** NEED TO EXPAND HERE***

2.2 Sloan Digital Sky Survey

The Sloan Digital Sky Survey was conceived for the purpose of collecting as much information about the sky as possible; it's initial objective was to collect images and spectra of started science measurements in April of 2000 (York 2000).

The Sloan Digital Sky Survey Data Release 12 (SDSS-DR12; Alam et al. 2015), which includes data from phases I-III (April 2000 through 14 July 2014), has served the astronomical community by providing photometric data on 14,555 square degrees of the sky (approximately one-third of the sky), resulting in the observation of approximately 470 million individual targets in five different bands (u, g, r, i, and z), with magnitudes brighter, or comparable to 22 (the limiting magnitude). Of these targets, only approximately 200 million are galaxies. In addition to photometric information, SDSS-DR12 also provides spectroscopic data on over 4.35 million targets, 2.4 million of which are galaxies. Due to the sheer number of photometrically and spectroscopically collected galaxies in this survey, SDSS-DR12 is a fertile ground for harvesting data for training a code for photometric redshifts.

2.3 Neural Networks and Photometric Redshifts

Deep learning is a very

Deep learning (LeCun et al. 2015), like its name implies is a more in depth kind of machine learning. In this kind of learning, images are passed in to a deep neural network, which is composed of artificial neurons organized in layers of nodes, so as to train the code. In order to be considered a deep neural network, there must be at least two hidden layers of nodes; that is, there must be at least two layers in between the input layer which recieves images and the output layer. Once these images are passed in, the code decomposes the images into sections, finds features that might be of importance, and tries to assign rules that allow the computer to correspond features to a numerical output. or each layer of the neural network, different features, and features of features are analyzed. They key to this technique's ability to learn is the fact that the weights of the node connections and the biases of the node, which sets a threshold that allows a neuron to be activated, are changed throughout the training of the code so as to minimize the cost function through gradient descent. In this way, the "rules" that govern the prediction of an image's redshift are changed.

Gradient descent in three dimensions can be compared to rolling a ball down a hill in order to find the lowest valley: by changing the x and y coordinates of the ball, the z coordinate may be minimized. While this is a pleasing metaphor, deep learning algorithms function with more than two features (thus more than three "dimensions" in this case), thus, gradient descent is much more complicated than the simple hill analogy. Instead, for more than three dimensions gradient descent is purely a

DEFINE THE COST FUNCTION

TALK ABOUT DIFFERENT APPLICATIONS OF DEEP NEURAL NETWORKS, THEN NARROW DOWN TO PHOTO-Z'S

There are many different ways of ascertaining a galaxy's photometric redshift using machine learning, as was shown in the PHAT Photo-z data trial (Hildebrandt et al. 2010). One method that showed promise, was the use of deep learning via a trained deep neural network (Collister & Lahav 2004). In this code setup, images are fed into a deep neural network that uses regression in order to predict an image's photometric redshift.

GIVE ALL OF THE FACTS ABOUT HOYLE (summary).

Other research has been done with neural networks and photometric redshifts (Hoyle 2015)

3 METHODOLOGY

3.1 Data Selection

The various band pass trials described below were completed on SDSS galaxy images. These galaxies and were selected with CasJobs (Li & Thakar 2008) in such a fashion so as to maximize fidelity of the data. This was achieved by selecting for galaxies with clean photometry (photo.clean = 1), no known issues with its spectra (spec.zWarning = 0), small redshift errors (spec.zErr < 0.1) and redshifts under 2 (spec.z < 2). In addition, the exponential scale fit radius and the de Vaucouleurs fit scale radius were required to be below 30 arc seconds (so as to exclude bad photometric images).

*** FLESH THIS OUT MORE, SAY HOW YOU THINK THIS GIVES A RANDOM ENOUGH DISTRIBUTION ***

100,000 galaxies selected from SDSS-DR12 were randomly selected from the subset of galaxies fitting the criteria above using CASJOBS. These galaxies were randomly selected, instead of taken from one area of the sky, because previous research (Cunha et al. 2012) has shown that variance of the spectroscopic measurements over the sky can lead to biased photo-z results if the data is collected from a small area or a collection of small areas (1 square degree or less at a time), thus facilitating a need for randomly collected data. These selection criteria were taken and modified from Edward Kim's work (used in Kim and Brunner 2016).

3.2 Data Processing

The collected images are sent through the following pipeline created by Edward Kim (cite paper)

WHY AM I GOING TO BE DOING BAND PASS RESULTS WITH LESS THAN 5 BANDS? NEED TO FIGURE OUT WHY EVERYONE ELSE DOES 5 BAND TRIALS, NOT LESS. MAYBE BECAUSE ITS LACKING IN CRUCIAL DATA? i MEAN, IT'S NATURAL TO THINK THAT DEEP LEARNING THRIVES ON MORE DETAILED DATA, NOT LESS, BUT UNDER WHAT PRECIDENT? – Possibly the Baum paper: there is a known correlation between the magnitudes at different band passes and a galaxy's redshift.

3.3 Neural Network Code

Through the use of a supervised neural network code created by Edward Kim (used in Kim and Brunner 2016), the data collected from SDSS

TALK ABOUT CPU VS. GPU, WHY GPU'S ARE BETTER.

3.4 Band Pass Trials

In order to ascertain the effectiveness of different band pass image combinations on the Neural Network's ability to assign an accurate redshift to the image,

4 RESULTS AND DISCUSSION

4.1 Five Band Pass Results

4.2 Four Band Pass Results

4.3 Three Band Pass Results

4.4 Two Band Pass Results

4.5 One Band Pass Results

4.6 Discussion

5 CONCLUSIONS

ACKNOWLEDGEMENTS

We acknowledge support from the National Science Foundation Grant No. AST-1313415. RJB acknowledges support as an Associate within the Center for Advanced Study at the University of Illinois.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

REFERENCES

Alam S., et al., 2015, ApJS, 219, 12
Baum W. A., 1962. p. 390, http://adsabs.harvard.edu/abs/1962IAUS...15..390B
Collister A. A., Lahav O., 2004, PASP, 116, 345
Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655
Cunha C. E., Huterer D., Busha M. T., Wechsler R. H., 2012, MNRAS, 423, 909
Hildebrandt H., et al., 2010, A&A, 523, A31
Hoyle B., 2015, arXiv:1504.07255 [astro-ph, physics:physics]
Hubble E., 1929, PNAS, 15, 168
Koo D. C., 1985, ApJ, 90, 418
LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436
Li N., Thakar A. R., 2008, Computing in Science Engineering, 10, 18
Puschell J. J., Owen F. N., Laing R. A., 1982, ApJ, 257, L57
York D. G., 2000, AJ, 120, 1579

This paper has been typeset from a TEX/LATEX file prepared by the author.