Deep Learning for Photometric Redshift Estimation in Astronomy

Rim Shayakhmetov
University Clermont Auvergne, CNRS, LIMOS
F-63000 Clermont-Ferrand, France
shayakhmetov.rim@gmail.com
rrshayakhmetov@edu.hse.ru

Engelbert Mephu Nguifo
University Clermont Auvergne, CNRS, LIMOS
F-63000 Clermont-Ferrand, France
engelbert.mephu_nguifo@uca.fr

***Abstract*—Photometric redshift estimation is an important problem in astrophysics. Accurate redshift predictions for all astronomical objects are still needed, as well as scalable algorithms to leverage big data in sky digital surveys. The goal of this work is to explore scalable deep learning architectures and algorithms, which can improve current state-of-the-art results.**

***Index Terms*—photometric redshift, machine learning, deep learning**

I. Introduction

Accurate photometric redshift estimation is crucial for upcoming multi-band sky surveys. Many empirical methods were proposed based on the large amount of data generated by the surveys. Majority of the methods are based on measured magnitudes and colors for objects. This is based upon the spectrum of an object's radiation having strong spectral lines that can be detected by the relatively crude filters. Apart from that, some authors [1] propose that using additional morphological features can help to improve predictive power for extended galaxies.

Success in deep learning methods such as convolutional networks enabled use of raw sky images as an input instead of an output from the fixed pipeline for extracting different measures of magnitudes for an object. Several researchers have explored such techniques, and found that they can reach the accuracy of state-of-the-art methods using convolutional neural networks [2][3]. These methods can help to automate feature extraction step from images. Although, there is a trade-off as they require significantly more computational resources compared to the standard machine learning algorithms.

II. Difficulties that algorithms face

There are several main features of the redshift reconstruction problem, which algorithms should take into account:

- There is a lack of spectroscopic coverage of magnitudes-colors space in sky surveys [4].
- There is a mismatch between magnitude error distributions associated with spectroscopic and photometric data sets [4].
- The number of photometric observations outnumbers the number of spectroscopic observations (where we get true value of redshift). This fact provides an opportunity for semi-supervised learning as there is a huge amount of unlabeled data.
- Databases sizes grow quicker and bigger as new sky digital surveys appear. This fact set requirements for algorithms to scale horizontally as volume and velocity are the main components of big data in sky surveys.

Apart from these difficulties, most of the research is concentrated on separate class of objects such as galaxies or quasars. For example, template-ml predictions for SDSS DR12 and DR13 database were based only on photometrically identified galaxies [5], having low prediction power for other types of objects or misclassified cases. In addition, different ways to measure magnitudes are better for different types of objects (for example, for point-like objects like stars and quasars PSF magnitudes are the best measure, colors derived from model magnitudes are better for galaxies, etc. [7]).

III. Towards new methods with deep learning

The goal of our work is to develop an algorithm that can provide accurate redshift predictions for all photometric measurements. First, we explore boosting algorithms that provided state-of-the-art results according to [8]. We developed algorithms based on XGBoost library [10] for gradient boosting algorithm, as it provides out-of-the-box distributed training on multiple machines and supports many cloud dataflow systems.

Currently we are exploring how deep learning methods can improve predictions based on different measures of magnitudes. We propose several models that can improve photometric redshift predictions for all objects (galaxies, quasars, stars) in the SDSS Data Release 13 [7]. For building deep neural network architectures we use Keras library [11] on top of TensorFlow [12], which provides out-of-the-box scalability for multiple machines (CPUs and GPUs).

We focus on developping one single model for redshift prediction of all photometric observations in SDSS DR13 with available spectroscopic redshift for 3.3 million objects. We use different measures of magnitudes, derived colors, photometric errors, and differences of PSF magnitudes with composite model magnitudes for implicit galaxy separation. First, we consider only galaxies, and then show how other objects as quasars and stars influence the quality of redshift prediction. The loss function for algorithms to optimize is the root mean

squared of normalized errors (Δz_{norm}). In the table I one can find the results for training on galaxies that we obtained for the best gradient boosting model and deep neural network with 5 hidden layers (using regularization and batch normalization technique).

TABLE I
REDSHIFT PREDICTION FOR GALAXIES ONLY

	std(Δz_{norm})	bias(Δz_{norm})	$\mid \Delta z_{norm} \mid > 0.15$
XGBoost galaxies	0.0316	0.0001	0.58%
DNN galaxies	0.0333	0.0001	0.61%

We can observe that the results of training on galaxies are quite satisfactory for both models, although boosting performs slightly better. In the table II one can see current research results with the main metrics for the best reported models. In some papers instead of standard deviation the reported metric was a 69^{th} percentile of normalized errors ($\sigma_{68}(\Delta z_{norm})$), although, usually standard deviation, σ_{68}, and RMSE have similar values.

TABLE II
DIFFERENT REPORTED RESULTS IN THE LITERATURE

	std(Δz_{norm})	bias(Δz_{norm})	$\mid \Delta z_{norm} \mid > 0.15$
[1] galaxies	0.041	-0.003	0.99%
[2] galaxies	σ_{68} = 0.03	0.001	1.56%
[8] galaxies	σ_{68} = 0.0248	0.0008	0.73%
[13] quasars	0.15	0.032	> 0.3 : 6.53%
[6] galaxies	0.0490	0.0081	7.6%
[9] galaxies	0.024	0.0	1.51%
[5] galaxies (template algorithm for SDSS DR13)	0.0205	0.00005	4.11%

Next, in the table III we show how adding all other types of objects results into less accurate predictions. Although, the quality of predictions varies a little for galaxies, one can see that the biggest errors are made for quasars.

Deep neural network has shown better RMSE score and reduces outliers rate for stars by a half, compared to the boosting model. Although, the predictions for quasars have more than a half of catastrophic outliers, current work on pretraining deep neural networks with utilizing large amounts of unlabelled data and seek for other architectures are being investigated. The developed code is available online on GitHub [14].

ACKNOWLEDGEMENTS

Special thanks to the Galactica project, for providing us with large scale highly distributed computing and storage resources.

TABLE III
REDSHIFT PREDICTION FOR ALL OBJECTS

	std(Δz_{norm})	bias(Δz_{norm})	$\mid \Delta z_{norm} \mid > 0.15$
XGBoost overall	0.1689	0.0015	10.7%
XGBoost galaxies	0.0547	-0.001	1.3%
XGBoost stars	0.2668	0.0954	16.5%
XGBoost quasars	0.3003	-0.1607	52.3%
DNN overall	0.1481	-0.021	9.1%
DNN galaxies	0.0477	-0.008	1.3%
DNN stars	0.1575	0.0415	8%
DNN quasars	0.3546	-0.2182	54.5%

The first author is also partially supported by the French LabEx project IMobS3, under the mobility program.

REFERENCES

[1] Ben Hoyle et al., *Feature importance for machine learning redshifts applied to SDSS galaxies*, Mon.Not.Roy.Astron.Soc. 449 (2015) no.2, 1275-1283 arXiv:1410.4696v4 (2015).
[2] Ben Hoyle, *Measuring photometric redshifts using galaxy images and deep neural networks*, Astronomy and Computing, Volume 16, July 2016, Pages 3440, arXiv:1504.07255v2 (2016).
[3] A. DIsanto, *Uncertain photometric redshifts with deep learning Methods*, arXiv:1703.01979 (2016).
[4] R. Beck et al., *On the realistic validation of photometric redshifts, or why Teddy will never be Happy*, Mon Not R Astron Soc (2017) 468 (4): 4323-4339, arXiv:1701.08748v2 (2017).
[5] R. Beck et al., *Photometric redshifts for the SDSS Data Release 12*, MNRAS 460, 13711381 (2016), arXiv:1603.09708v2 (2016).
[6] R. Beck et al., *Photo-z-SQL: Integrated, flexible photometric redshift computation in a database*, Astronomy and Computing, Volume 19, p. 34-44 (2017).
[7] *SDSS Data Release 13*, `http://www.sdss.org/dr13/`
[8] R. Zitlau et al., *Stacking for machine learning redshifts applied to SDSS galaxies*, Mon.Not.Roy.Astron.Soc. 460 (2016) no.3, 3152-3162, arXiv:1602.06294v2 (2016).
[9] Ben Hoyle et al., *Data augmentation for machine learning redshifts applied to SDSS galaxies*, Mon.Not.Roy.Astron.Soc. 450 no.1, 305-316 arXiv:1501.06759 (2015)
[10] *XGBoost library*, `http://xgboost.readthedocs.io/`
[11] *Keras library* `https://keras.io/`
[12] *TensorFlow library* `https://www.tensorflow.org/`
[13] M. Brescia et al., *Photometric redshifts for Quasars in multi band Surveys*, The Astrophysical Journal, 772, no.2, 140 (2013).
[14] *The project repository*, `https://github.com/shayakhmetov/redshift_prediction`