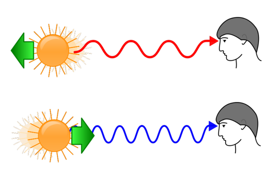
Deep Learning for Photo-Z reconstruction

Rim Shayakhmetov & Engelbert Mephu Nguifo
University Clermont Auvergne
LIMOS - CNRS UMR 6158 - Institut d'Informatique - ISIMA

github.com/shayakhmetov/redshift_prediction

1 Photometric redshift

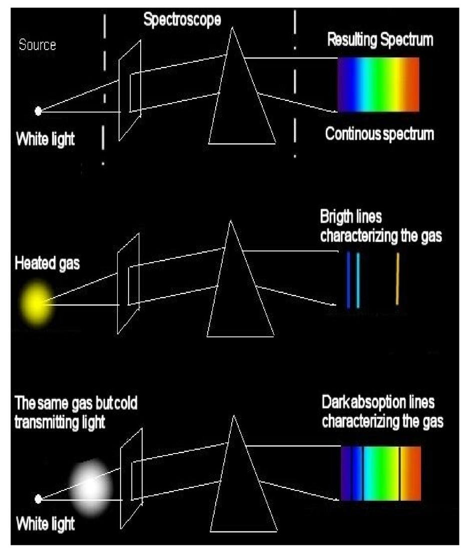
1.1 What is redshift

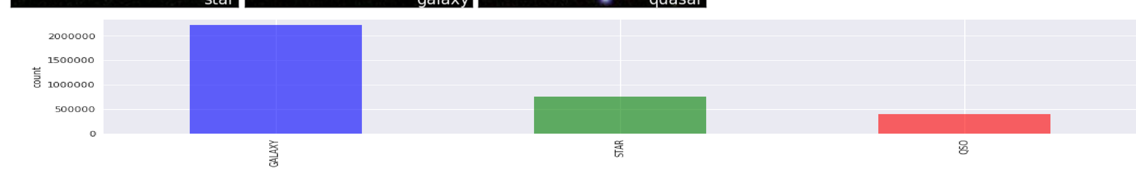


Redshift is an increase in wavelength of an object appearing moving away from an observer. $1 + z = \frac{\lambda_{observed}}{\lambda_{emitted}}$, where z-value characterize the shift and λ is a wavelength.

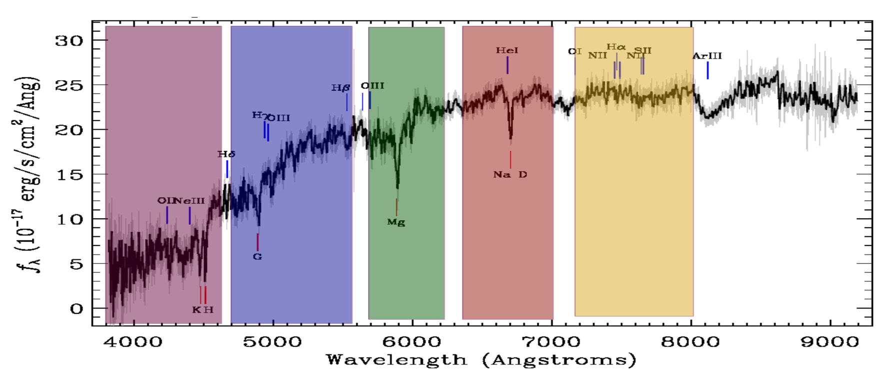
1.2 Spectroscopic redshift

Spectroscopy based on the dispersed light from an object. Redshift is easily identified from the spectrum of an object, but expensive and slow. SDSS survey Data Release 13 provides spectroscopic data for 3.3 million objects, which are divided into three spectroscopic classes: stars, galaxies, quasars:





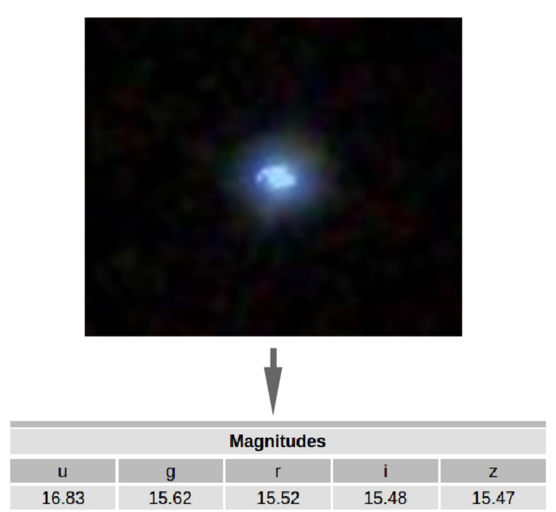
1.3 Photometry



Photometry provides a cheaper and faster way to measure intensity, but only in the several intervals of wavelengths.

1.4 Magnitudes

In astronomy magnitude is a measure of brightness. There are many different methods to convert an image to the magnitude measurements: PSF magnitudes, galaxy model magnitudes, Petrosian magnitudes, etc. Apart from these, the differences among bands is valuable and called colors.



2 Machine Learning

2.1 Supervised learning

The goal is to find a function f s.t. $z = f(magnitudes, colors)$. The quality metrics are usually based on the normalized error $\Delta z_{norm} = \frac{z_{photo} - z_{spec}}{1 + z_{spec}}$. Most common are the $mean(\Delta z_{norm})$, $\sigma(\Delta z_{norm})$, $\sqrt{mean(\Delta z^2_{norm})}$, and the percentage of outliers defined by $|\Delta z_{norm}| > 0.15$ (or 3σ)

One can find current research results in the table below:

	class	std(Δz_{norm})	bias(Δz_{norm})	$\mid \Delta z_{norm} \mid > 0.15$
[2]	galaxies	0.041	-0.003	0.99%
[3]	galaxies	$\sigma_{68} = 0.03$	-0.001	1.56%
[9]	galaxies	$\sigma_{68} = 0.0248$	0.0008	0.73%
[13]	quasars	0.15	0.032	$> 0.3 : 6.53\%$
[7]	galaxies	0.0490	-0.0081	7.6%
[10]	galaxies	0.024	0.0	1.51%
[6]	galaxies	0.0205	0.00005	4.11%

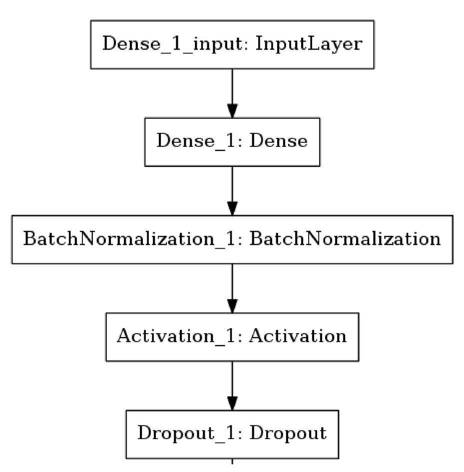
2.2 Data

Data comes from the SDSS survey containing 300 million clean photometric measurements in 5 filters, 3.3 million objects with clean spectra. In our work we first use 105 photometric features, including magnitudes, errors, colors, and difference between PSF and cModel magnitudes.

2.3 Gradient Boosting

As boosting techniques showed [9] superior performance, we explore more advanced regularized gradient boosting algorithm, implemented via XGBoost library [11]. For hyperparameters tuning Bayesian optimization with random initialization was conducted. 32 most important features increase XGBoost performance, which include colors from fiber and PSF magnitudes, and differences between PSF and cModel magntides.

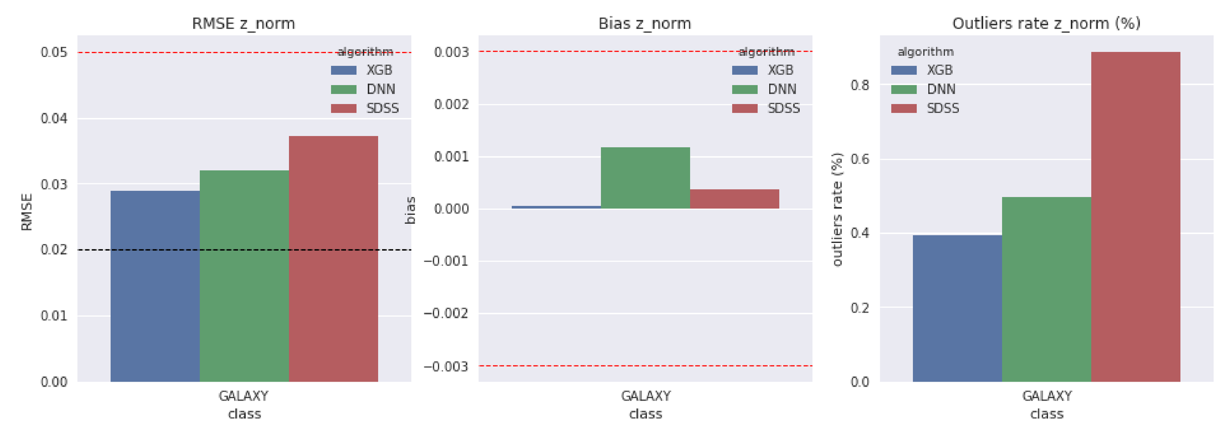
2.4 Deep Neural Networks



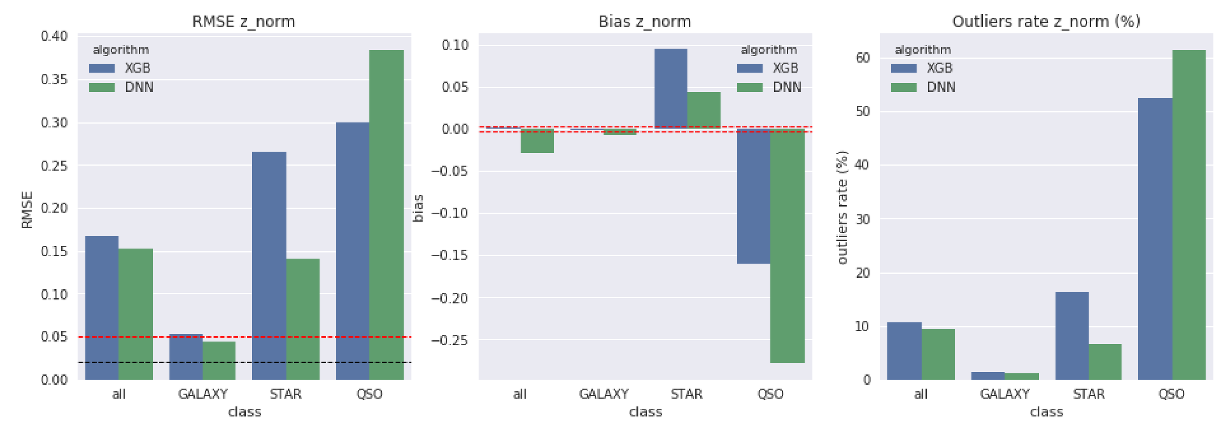
For developing deep learning architectures we use Keras library [12]. Training deep neural networks requires several techniques for improving gradient flow and predictions that we are using: L1/L2 regularization, dropout technique, and batch normalization. One can see the building block of the deep neural netwrok, consisting of above-mentioned techniques. The best architecture was chosen based on the Bayesian optimization on number of layers, regularization parameters, etc.

2.5 Experimental results

Considering only galaxies for training and predictions, we can get better results than current SDSS default predictions in the database. The DNN for galaxies consisted of 5 layers of above-mentioned blocks.



The results for considering all objects in the SDSS spectroscopic data (galaxies, stars, and quasars) show that a deep feed forward network can on average outperform boosting technique. The best DNN for all objects consisted of 10 layers.



2.6 Different deep learning architectures

Current research is being conducted for a search of different deep learning architectures, that will provide better predictions, and could utilize a huge amount of unlabeled data.

2.6.1 Unsupervised pre-training - autoencoders

Autoencoders - self-supervised learning for encoding information (dimensionality reduction). The idea is to match the input and the output, producing compact representation in the middle of the network (code). Stacking autoencoders (layer-wise construction of autoencoders) and denoising autoencoders (reconstructing noisy input) are excpected to leverage big unlabeled data and photometric errors.

2.6.2 Convolutional neural networks

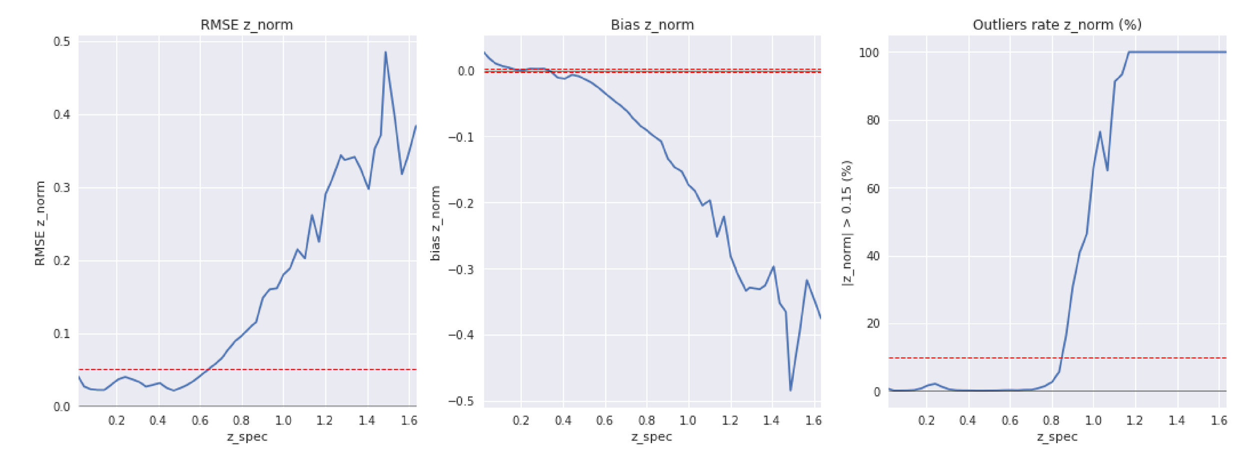
Convolutional neural networks (CNNs) are designed to process images, and several authors [3] [4] have observed their applications, and found that CNNs can achieve state-of-the-art quality of predictions. The advantage is an automatic feature extraction from images for objects.

2.6.3 Residual neural networks

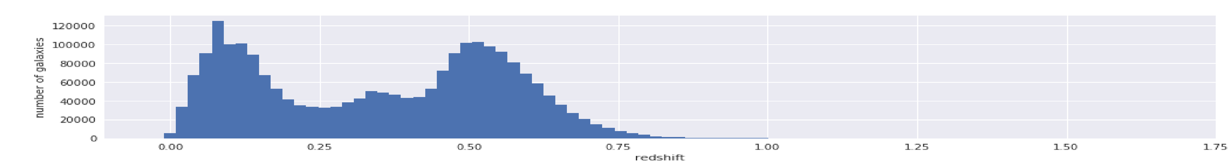
Current state-of-the-art convolutional networks besides proper initialization, batch normalization and other techniques use residual connections - shortcuts for the gradient to flow. This helped to train deep convolution layers from tens to hundreds of layers (ResNet). Adding residual connections can help to train deep neural networks for more than ten layers.

2.7 Meeting the requirements

Although, the quality of predicted galaxy redshift meets the requirements on average, for the farthest galaxies the quality decreases substantially.



One can observe the coverage of the SDSS redshifts for galaxies, which is correlated with the increase of errors.



Acknowledgements

Special thanks to the Galactica project, for providing us with large scale highly distributed computing and storage resources. The first author is also partially supported by the French LabEx project IMobS3, under the mobility program.

The work uses SDSS data. 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/`.

References

[1] LSST Science Collaborations and LSST Project, LSST Science Book, Version 2.0, (2009)
[2] Ben Hoyle et al., Mon.Not.Roy.Astron.Soc. 449 (2015) no.2, 1275-1283 (2015).
[3] Ben Hoyle, Astronomy and Computing, Volume 16, July 2016, Pages 34–40, arXiv:1504.07255v2 (2016).
[4] A. D'Isanto, *Uncertain photometric redshifts with deep learning Methods*, arXiv:1703.01979 (2016).
[5] R. Beck et al., Mon Not R Astron Soc 468 (4): 4323-4339 (2017).
[6] R. Beck et al., MNRAS 460, 1371–1381 (2016).
[7] R. Beck et al., Astronomy and Computing, Volume 19, p. 34-44 (2017).
[8] *SDSS Data Release 13*, `http://www.sdss.org/dr13/`
[9] R. Zitlau et al., Mon.Not.Roy.Astron.Soc. 460 (2016) no.3, 3152-3162, (2016).
[10] Ben Hoyle et al., Mon.Not.Roy.Astron.Soc. 450 no.1, 305-316 (2015)
[11] *XGBoost library*, `http://xgboost.readthedocs.io/`
[12] *Keras library* `https://keras.io/`
[13] M. Brescia et al., The Astrophysical Journal, 772, no.2, 140 (2013).