Mem. S.A.It. Vol. 94, 121

Memorie della

Harvesting the Lyman alpha forest with convolutional neural networks

Ting-Yun Cheng[1], Ryan J. Cooke[1] and Gwen Rudie[2]

[1] Centre for Extragalactic Astronomy, Durham University, South Road, Durham DH1 3LE, UK
[2] The Observatories of the Carnegie Institution for Science, 813 Santa Barbara Street, Pasadena, CA 91101, USA
e-mail: ting-yun.cheng@durham.ac.uk

Received: 28-11-2022; Accepted: 23-02-2023

Abstract. We develop an algorithm using a convolutional neural network (CNN) to identify low H I column density Lyα absorption systems ($\log N_{\rm HI}/{\rm cm}^{-2} < 17$) in the Ly$\alpha$ forest, and predict their physical properties, such as their H I column density, redshift, and Doppler width. The CNN models provides state-of-the-art predictions to 15 observed Keck/HIRES spectra of quasars at redshift $z \sim 2.5 - 2.9$, in < 3 minutes per spectrum. We demonstrate that CNNs can be used to analyse the enormous number of data available with current and future facilities, and thereby greatly increase the statistics of Lyα absorbers.

Key words. methods: data analysis – galaxies: high-redshift – quasars: absorption lines – intergalactic medium

1. Introduction

The Lyman-α (Lyα) forest are the absorption of photons emitted by a background quasar at the redshifted Lyα transition (rest-frame wavelength=1215.67Å). The Lyα absorbers with low H I column density dominate the Lyα forest and can be used to probe the distribution and evolution of the baryonic matter, structure formation, and constrain cosmological parameters (see reviews: Meiksin 2009). Conventionally, these absorption lines are manually fit with Voigt profiles (e.g. Rudie et al. 2012, hereafter R12) which requires many human hours. To overcome big data problems from future surveys and facilities, this work (Cheng et al. 2022), for the first time, applies a convolutional neural network (CNN) to efficiently identify Lyα forest systems ($N_{\rm HI} < 10^{17}$ cm^{-2}) and extract their physical properties, including the redshift, Doppler width, and H I column density.

2. Methodology

Our training data are simulated quasar spectra generated using packages in the PYIGM software. The generated spectra represent a typical quasar at redshift $z = 3$ and are convolved with an instrumental $v_{\rm FWHM} = 7$ km s^{-1}. The velocity per pixel of these spectra is set to 2.5 km s^{-1} pixel^{-1}. These choices reflects the typical properties of high resolution quasar spectra in current observatory archives. Additional noises (S/N $\simeq$ 10) are added into the training spectra to stablise predictions for

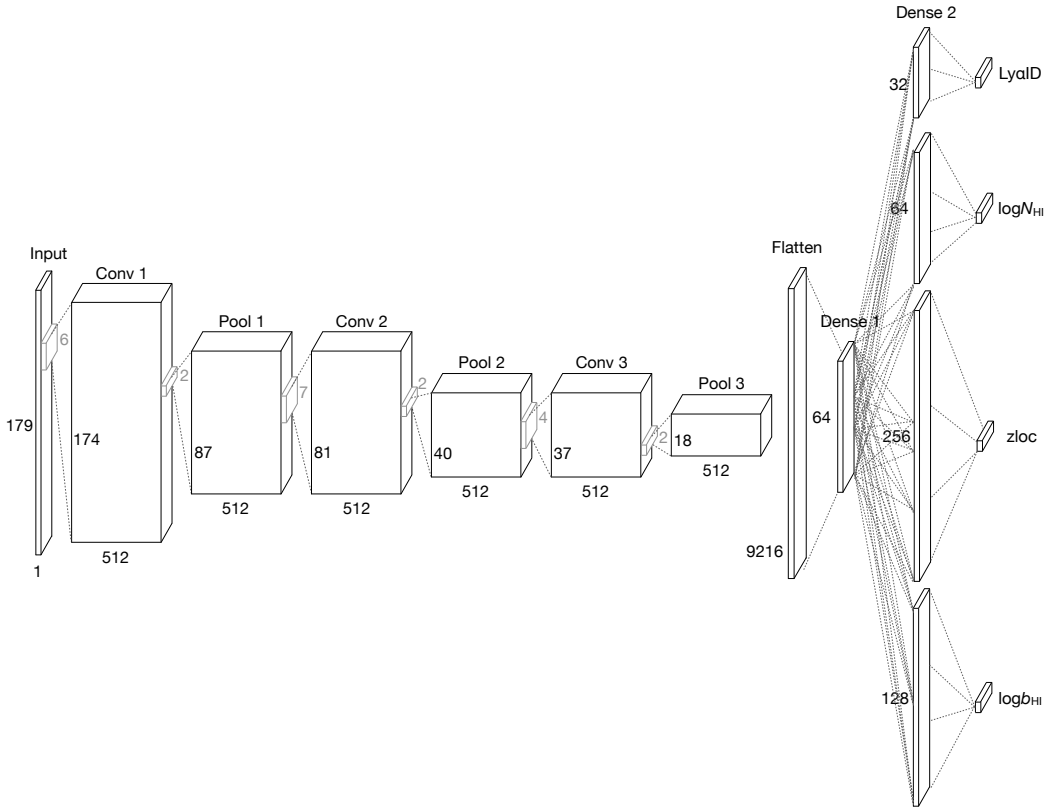

Fig. 1. The CNN architecture contains three 1-dimensional convolutional layers with pooling layers following each, one dense layer to connect each component, and four dense layers for four target outputs. The values of relevant hyperparameters are listed in Table 1.

spectra with different noise levels. We employ multi-task learning by training with and predicting four outputs (labels), also see Fig. 1:

- LyαID: it is set to a value of 1 if a Lyα absorber exists in this pixel, and 0 if not;
- log$N_{\rm HI}$: H I column density (in units of cm^{-2}) of the corresponding Lyα absorber on a logarithmic scale;
- zloc: the relative location of the centre of an absorption feature (in units of pixels). A pixel centred on an absorption feature is set to 0, and negative and positive values to pixels at the left and right, respectively;
- log$b_{\rm HI}$: Doppler width of the corresponding Lyα absorber on a logarithmic scale (km s^{-1}).

We apply similar training strategies to that adopted by Parks et al. (2018) to 'scan' through a spectrum with a fixed-size window (ws) and a 1 pixel step size. Fig. 1 shows our CNN architecture.

3. Results and Conclusion

We validate our CNN model by 15 quasar Keck/HIRES spectra observed and reduced by R12. Two metrics are considered in this work: (1) the root mean square error (RMSE) and (2) mean absolute error (MAE), to assess the 'accuracy' of the CNN predictions. The RMSE is strongly impacted by the outliers due to the square of the residual. Hence, we will focus

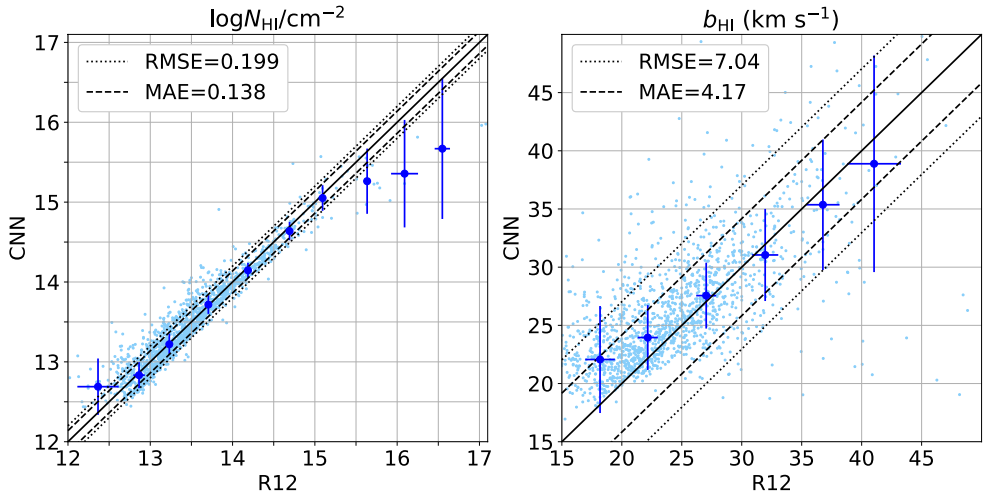

Fig. 2. Comparisons between the CNN and R12 values of $\log N_{\rm HI}/{\rm cm}^{-2}$ (left) and $b_{\rm HI}$ (right). The black solid line shows a one-to-one relation. Dark blue datapoints show the median values of CNN and R12 within different bins of R12. The x-axis error bar is defined by the median value of the estimation errors of the datapoints provided in R12, while the y-axis error bar presents the MAE of each bin.

Table 1. Hyperparameters selected using Bayesian Optimisation (GPyOpt)

	Hyperparameters	Value
Data Input	window size (*ws*)	179
	central pixels (*cnpix*)	1
CNN	L2	0.0
Architecture	dropout	0.1
	conv_filter_1	512
	conv_filter_2	512
	conv_filter_3	512
	conv_kernel_1	6
	conv_kernel_2	7
	conv_kernel_3	4
	dense_1	64
	dense_2_ID	32
	dense_2_N	64
	dense_2_z	256
	dense_2_b	128

on the MAE, which is more resilient to outliers than the RMSE, in this paper. Around 78 per cent of the ML-classified Lyα systems are matched with R12, and the MAE of $\Delta \log N_{\rm HI}/{\rm cm}^{-2}$, $\Delta z_{\rm HI}$, and $\Delta b_{\rm HI}$ are 0.14 dex, 2.7×10^{-5}, and 4.2 km s^{-1}, respectively. The comparison of column density and Doppler width between our CNN and R12 are shown in Fig. 2. This result validates the possibility to apply a CNN model with our approach to analyse the enormous quantity of data, in particular characterising low H I column density Lyα absorption systems ($\log N_{\rm HI}/{\rm cm}^{-2} < 17$), in the current archives (e.g. VLT/UVES, Keck/HIRES) and that will be obtained with new facilities (e.g. WEAVE, 4MOST, etc).

Acknowledgements. TYC thanks the comments provided by anonymous referee. TYC acknowledges the support of STFC grant ST/T000244/1 and Royal Society grant RF/ERE/210326, hosted at Durham University, and the support by Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 & The Alan Turing Institute.

References

Cheng, T.-Y., Cooke, R. J., & Rudie, G. 2022, MNRAS, 517, 755
Meiksin, A. A. 2009, Reviews of Modern Physics, 81, 1405
Parks, D., Prochaska, J. X., Dong, S., & Cai, Z. 2018, MNRAS, 476, 1151
Rudie, G. C., Steidel, C. C., Trainor, R. F., et al. 2012, ApJ, 750, 67