

Searching for Rare Quasar Ca II Absorbers in the Early Universe

Iona Xia

Abstract

The universe is still largely a mystery to scientists. Quasars are one such mystery, whose emission spectra produce absorption lines such as Ca II when passing through gas of distant galaxies. This data helps astronomers understand more about interstellar gas, dust, and galaxy and star formation and evolution (including my Milky Way). However, these current absorber databases are extremely limited, and traditional methods make them hard to detect. Thus, I seek to discover more Ca II absorbers by developing deep neural networks, which are more accurate and faster. I first found absorbers traditionally to produce a test set. I cropped, normalized, and handpicked through thousands of spectra and discovered 256 Ca II original absorbers. To obtain large training sets, I generated tens of thousands of artificial samples by inserting Ca II lines at corresponding wavelengths in real spectra. I preprocessed the data and created neural network models after testing different hyperparameter configurations. Overall, my accuracy for absorber detection is 95% (Ca II), 15 times higher than traditional methods, and I added significant amounts of new absorbers to the current dataset for Ca II, completing my goal. As for challenges, I concluded that most false negatives are due to noise and weak lines. Furthermore, my discovered absorbers agree with statistical tests of previous studies. In the future, I plan to discover more absorbers using my models and run statistical studies on them.

1 Introduction

1.1 Background

In order to understand galaxy formation and evolution, dust properties, and element abundances and history, quasar absorption lines have become essential due to their information



Figure 1: An example of a quasar, a celestial body in the center of galaxies.

on the evolution of gas and dust components in the universe and their usage in detecting cloud structures and physical processes (Zych, et al. 2008). Quasars are a massive celestial object that is at the center of galaxies, and they emit a lot of energy. Quasar absorption lines are formed when quasar light from billions of years ago emits energy in a wide spectrum of electromagnetic radiation that passes through different sorts of atoms and ions in the gas clouds between the quasar and the Earth. Each element imprints a different type of line which can then be

observed from the best telescopes on Earth.

Unlike most other studies involving quasar absorption lines, the analysis of calcium II (Ca II) absorbers is concentrated on quasars at lower redshifts. Because of this, it is able to provide information about recent absorbers and host galaxies. As the Universe is constantly

expanding, galaxies are moving away from us. Therefore, the lights from these galaxies are getting longer and longer, and this is also called redshifted. When galaxies are located at lower redshift, they are more recent galaxies. Ca II absorbers can be found from a redshift of around 1.4 all the way to the present epoch, or about the past 9 billion years; while the more popular Mg II absorbers can only be accessed at a redshift of 0.4 and above. In fact, Ca II absorbers are the only absorbers that can be accessed at lower redshifts (Sardane, et al. 2014). Ca II absorbers are also considered to be one of the dustiest absorbers, which is important in tracing out paths in which the Universe recycles materials and where they originally came from. This allows Ca II absorbers to be often associated with star formation (Zych, et al. 2008). Therefore, Ca II absorbers would enable detailed studies of gas and dust components of interstellar clouds, their properties, and physical processes to better understand star formation. Nevertheless, due to calcium being highly refractive, it is among the most depleted in the universe, and thus Ca II absorbers have not been well studied or understood (Sardane, et al. 2014).

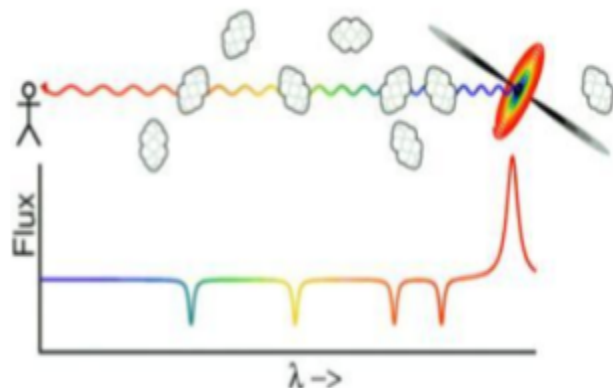


Figure 2: A diagram of how quasar absorption lines are created and picked up on Earth. The quasar light from far away quasars pass through the gas of galaxies, and then get picked up by telescopes.

1.2 Motivations

Many of the crucial questions involving Ca II absorbers are unanswered due to only small sample sets available. Generally, when studying Ca II absorbers, the doublet at 3934 Å and 3969 Å is used. Currently, the largest dataset of Ca II absorbers only has 435 absorbers (Sardane, et al. 2014). In addition, traditional methods of detecting these absorbers are slow and require lots of human interference and effort, making it unideal for discoveries.

1.3 Project Goal

My project aims at using deep learning methods to identify more Ca II absorbers in an efficient, automatic, and more accurate way. Previous studies on Mg II absorbers have proven to be successful (Zhao, et al. 2019), though there are differences between Mg II absorbers and Ca II absorbers. My goal is to explore if a neural network approach can achieve a high accuracy in classifying quasar spectra with true Calcium II absorbers so as to develop a larger dataset for future research.

2 Methods

2.1 The Traditional Method and the Test Set

Before building a neural network, it was necessary to find absorbers the traditional way in order to provide a test dataset. I used a total of 78,000 Mg II absorbers from Sloan Digital Sky Survey's Data Releases (DR) 7 and 12, released in 2009 (Abazajian et al. 2009) and 2015 (Alam

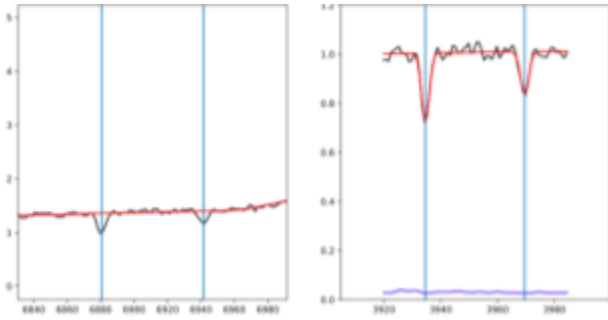


Figure 3: An example of a real Ca II absorber that was discovered. The left shows the absorber on the observer frame, and the right shows the absorber on the rest frame.

et al. 2015), respectively. Mg II absorbers are relatively common absorbers that can be employed as references to find Ca II absorbers as the latter are fairly weak and cannot easily be located by themselves (Zych et al. 2008). I cropped the spectrum of each Mg II absorber to locate where the Ca II absorber can possibly be found, and then I fit it to a double line profile and normalized it. Next, I measured the equivalent width (EW) using the signal-to-noise ratio (SNR), and removed any absorber with $\text{SNR} < 2$, as it contained too much noise. From there, I

handpicked through about 4,000 candidate absorbers multiple times and found a total of around 500 absorbers, of which about 250 overlaps with samples found in Sardane et al. (2014), leaving 256 completely newly discovered absorbers (Figure 1).

2.2 Artificial Data and Preprocessing

After forming the test set, I generated artificial spectra data as my training data. As the amount of natural absorbers is small, it is impossible to train with just real data. To produce the simulated data, I used quasar spectra from DR12 with a specified redshift (z), and then inserted an artificial absorption line at $3934(1+z)$ Å and another at $3969(1+z)$ Å, for the two Ca II lines.

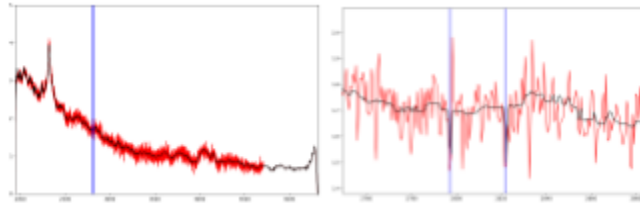


Figure 4: An example of the artificial spectra is created, with the whole spectra on the left, and a close up of the Ca II absorbers that have been added on the right.

For the lines to simulate the real absorbers, I set the EWs to be weaker and matched with the distribution of the real data by setting a minimum EW and a maximum EW (Figure 2). The negative samples are just the original spectra. I created a total of 86,000 positive data and 86,000 negative data this way.

I then preprocessed the data for both the training and testing sets, with the testing set consisting of the 500 absorbers discovered and 500 randomly selected negative samples.

2.3 Deep Learning Method

After that, I constructed a deep convolutional neural network model and conducted model training and inferencing with the open-source Python libraries Keras and TensorFlow. When creating my model, I tested with extensive sets of hyperparameters including the number, type, and size of layers, optimizer, number of epochs, learning rate, dropout rate, etc. until I felt confident with the results. My model has 5 convolutional layers and 2 fully connected layers. I also used a stochastic gradient descent, and a relu activation function, with sigmoid for output, as well as a learning rate of 0.01.

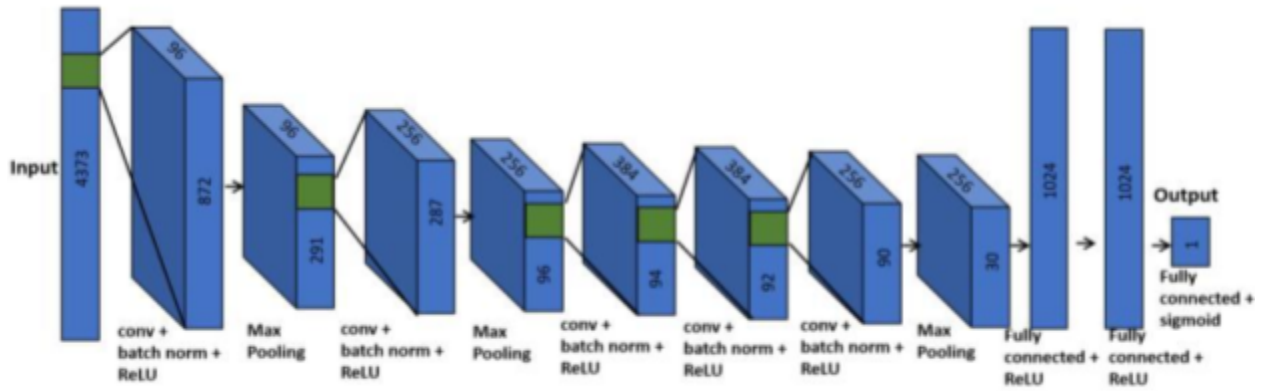


Figure 5: A diagram of my neural network model.

3 Results

My current results are composed of the following:

Accuracy: $\frac{\text{correct predictions}}{\text{total predictions}}$

Precision: $\frac{\text{true positives}}{\text{all positives}}$

Recall Rate: $\frac{\text{true positives}}{\text{all true}}$

F1 score: $\frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$

Accuracy	0.9213973799
Precision	0.932735426
recall	0.9082969432
f1	0.9203539823

Figure 6: The results from training for my neural network.

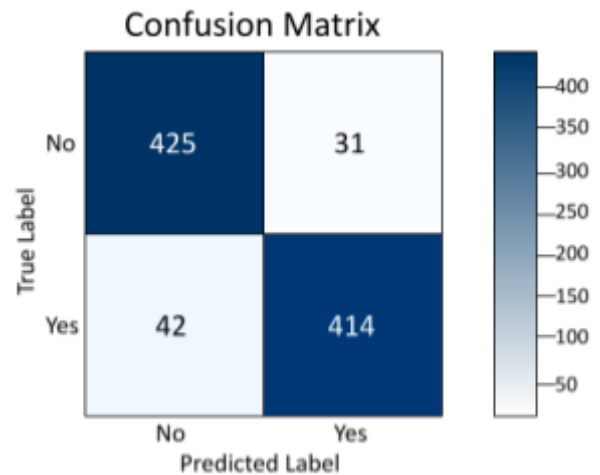


Figure 7: The confusion matrix of my best training model. The majority of the absorbers were correctly predicted.

With my current model, on the testing set, my best accuracy of testing is 95%. The traditional method had an extremely low accuracy due to its inclination to pick out spectrums with a lot of noise. Out of the 4000 absorbers that it picked, it had a 11.4% accuracy. This leads a 95% accuracy to be an extremely good improvement. takes about 15 minutes for the computer to train the model with 154,000 samples over 15 epochs, and less than 0.5 seconds for it to predict for 1000 samples. On the other hand, for the traditional method,

it takes up to an hour to fit the spectra for around 900 samples. It then takes around 5-6 hours to handpick through all of them. This means that the machine learning method is much faster.

4 Discussion and Future Research

As the recall rate is relatively lower than the precision, indicating that the model is more likely to get positive samples wrong than false samples, it is possible that the training data is still a little too strong, leading some weaker samples to be classified as false absorbers, or high signal noises.

When looking at the false negatives, it is clear that the model struggles to find extremely weak absorbers and signal noises, although it is still able to find most weak absorbers. In addition, statistical analysis on our detected absorbers mainly agree with results. The equivalent widths, or basically the strength of the absorber, seems to match with published results, although the model is able to detect weaker absorbers than traditional methods. Redshift distributions mainly match, except published measurements are able to find absorbers with redshift < 0.4 . My model is still unable to do that due to the use of the restframe.

In the future, I plan to improve the neural network by trying it on the observer frame in order to find absorbers with redshift < 0.4 , as well as involving noise analysis in the neural network to reduce false positives. I can also analyze many properties of these quasars, as well as studying properties such as dust reddening, star formation, and physical conditions.

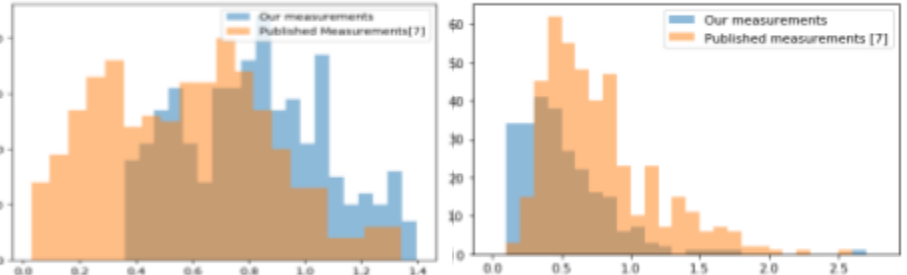


Figure 8: Statistical analyses of both the redshift (left) and the equivalent width (right). They are similar to published results.

5 References

1. K. Abazajian, et al., “The Seventh Data Release of the Sloan Digital Sky Survey”, *Astrophys. J. Suppl.* 182:543-558, 2009
 2. S. Alam, et al., “The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III”, *Astrophys. J. Suppl.* 219:12, 2015
 3. G. M. Sardane, D. A. Turnshek, S. M. Rao, “Ca II Absorbers in the Sloan Digital Sky Survey: Statistics”, *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 444:1747-1758, 2014
 4. B. J. Zych, et al., “Dust Depletion, Chemical Uniformity and Environment of CaII H&K Quasar Absorbers”, *MNRAS*, 392:1429-1450, 2008
 5. Y. Zhao, et al., “Identifying Mg II Narrow Absorption Lines with Deep Learning”, *MNRAS*, 487: 801-811, 2019
- Figure 1: <https://science.sciencemag.org/content/309/5731/75/tab-article-info>
- Figure 2: <http://www.astro.ucla.edu/~wright/Lyman-alpha-forest.html>