Jan, Albert

Discovering Rare Neutral-Carbon Absorbers with a Deep Neural Network to Probe Star-Forming Environments of the Early Universe

Abstract

Big Bang cosmology presents a widely accepted model explaining the formation and evolution of the universe. It explains the existence of Cosmic Microwave Background Radiation (CMBR), the oldest electromagnetic radiation in the universe, and predicts that it cools over time at a rate governed by the expansion of the universe. This theory can be tested by using the excitation levels of certain interstellar gases, such as neutral-carbon (C I), in galaxies of varying distances from Earth to make cosmic microwave background temperature measurements at various points in time after the universe's formation. C I can be detected via its two absorption lines at rest-frame wavelengths of ~1560 and ~1656 Å in the spectra of distant quasars, extremely luminous galactic nuclei whose light shines through and is absorbed by interstellar gases while travelling to Earth. C I is often found in the same molecular clouds as other cold gases, which can be used to probe star forming environments in the early universe via their absorption lines as well. C I content is also correlated with interstellar dust content, which can be used to investigate how galaxies chemically evolve over time. However, the current sample of C I absorbers is too small to address the mysteries that they could potentially uncover about the universe; besides the rarity of interstellar C I, this problem can be attributed to subpar traditional methods of detecting absorption features, which are slow and require frequent human interference. Thus, this research modified a convolutional neural network (CNN) architecture, with the author coding and engineering the final architecture, to detect C I absorbers in the SDSS-12 dataset and probe their physical properties. C I absorbers are a subset of Mg II absorbers, so traditional methods were adopted to narrow a dataset of 41,894 Mg II absorbers down to 2,036 possible C I absorber candidates. After thorough inspection, 113 of them were deemed as C I absorbers, which were used to evaluate the CNN's performance. 20,000 spectra, 10,000 with artificially inserted C I absorption lines and 10,000 without, were created and served as the training dataset for the CNN. A testing set of 4,000 spectra, also split half-half, was used to evaluate the CNN's performance after training. The final model achieved respective training and testing accuracies of 99.81% and 96.30% on the artificially generated spectra, and it successfully detected 92.92% of the 113 C I absorbers discovered with traditional methods. Additionally, the average column densities and depletions of several metals in the 113 C I host galaxies were measured to probe their physical composition. The depletions of these metals are in broad agreement to that of the Milky Way disks, implying that C I host galaxies are also relatively dusty systems, supporting previous studies that classify C I absorbers as excellent probes of the cold, dense interstellar environments from which stars form.

Discovering Rare Neutral-Carbon Absorbers with a Deep Neural Network to Probe Star-Forming Environments of the Early Universe

Introduction

The Big Bang Theory is the prevailing model explaining the universe's origins, with a vital piece of supporting evidence being the existence of Cosmic Microwave Background Radiation (CMBR): faint electromagnetic radiation dating back to events of the early universe. While the CMBR temperature (T_{CMB}) today has been measured to be roughly uniform across the observable universe as expected (Mather *et al.*, 1999), standard Big Bang cosmology also predicts the CMBR to cool over time at a rate governed by the expansion of the universe. This prediction can be rigorously tested by using the excitation of atomic or molecular gases within distant galaxies at varying distances from Earth to measure T_{CMB} at varying points in time; since light travels at a finite speed, the farther one looks into space, the farther one sees back in time. These gases can be detected by analyzing the spectra of quasars, which are the extremely luminous nuclei of certain distant galaxies. As a quasar's light travels through interstellar clouds within galaxies on the way to Earth, various gases absorb unique wavelengths of the light, giving rise to absorption lines in the quasar's spectrum that can be used to deduce the composition of the gas clouds. However, few species detected are actually suitable for measuring T_{CMB}, one of which is atomic neutral-carbon (C I), whose excitation levels have been used to measure past values of T_{CMB} (Ge, Bechtold, & Black, 1997; Songalia *et al.*, 1994). Besides being an excellent cosmic thermometer, C I is also often found in the same molecular clouds as other cold, neutral gases, such as H_2 (Lizst, 1981), from which stars form (Glover & Clark, 2016). Thus, if C I absorption lines are detected, absorption lines for these other gases will likely be present and representative of the composition of star forming environments in the early universe; the galaxies being analyzed are often so distant that one would be studying them as they were several billion years in the past (Ledoux, Noterdaeme, Petitjean, & Srianand, 2015). Finally, C I content can be correlated with interstellar dust content, whose absorption and scattering of quasar light can also be measured in galaxies of varying distances, and therefore varying points in time, to determine how dust content in galaxies chemically evolved over time (Heintz *et al.*, 2019a). However, the current sample of quasar spectra containing C I absorption lines, which are referred to as C I absorbers, is still too small to address the questions that they could potentially be used to answer about the universe. This is not only because C I absorption lines are rare, weak, and difficult to detect at exceptionally far distances (Ge, Bechtold, & Black, 1997), but also due to the fact that traditional computational methods for detecting absorption features are extremely time-consuming and require frequent human intervention (Zhao *et al.*, 2019). Thus, this research aims to use artificial intelligence, namely deep learning algorithms, to detect C I absorbers in the Sloan Digital Sky Survey (SDSS) Data

Release (DR) 12 quasar spectra dataset, as well as probe their physical composition to assess previous implications made about their properties.

The Redshift and Cosmic Microwave Background Radiation Temperature Relation

According to the Big Bang theory, the universe cooled to a point where electrons and protons could finally bind to each other and form neutral atoms approximately 300,000 years after its formation. This allowed for photons to freely flow throughout space ever since, which are now known as the CMBR (Durrer, 2015). As the universe expands, due to the Doppler Effect, CMB photons undergo an increase in wavelength, also known as redshift, and consequently lose energy over time. Friedmann cosmology, which relies on the Big Bang model to describe the expansion of the universe, can be used to derive a linear relationship between the CMBR temperature (T_{CMB}) and its redshift (Peebles, 1993), as shown in Equation 1:

$$T_{CMB}(z) = T^0_{CMB}(1+z) \quad (1)$$

where T^0_{CMB} is the CMBR temperature today, which has been measured to be 2.725 ± 0.002 K (Mather *et al.*, 1999), and z is a dimensionless quantity used to denote the redshift of the light source (CMBR photons in this case) as defined in Equation 2:

$$z = \frac{\lambda_{obs} - \lambda_{rest}}{\lambda_{rest}} = \frac{\Delta\lambda}{\lambda_{rest}} \quad (2)$$

where λ_{obs} is the observed wavelength of the source, and λ_{rest} is the wavelength of the source when it is at rest. Light sources move away from Earth at speeds directly proportional to their distance (Hubble, 1929), and a source appears to be more redshifted the faster it is moving away from the observer. This provides some intuition to Equation 1: the more redshifted the CMB photon is observed to be from Earth (the greater the value of z), the farther back in time one is seeing when analyzing it, resulting in higher T_{CMB} measurements. However, the linear relationship must be rigorously tested as means of evaluating the Big Bang model.

In space, the CMBR is able to excite certain interstellar atomic or molecular species. When the relative population ratios of a species' energy levels are in equilibrium with the CMB photons, its excitation temperature is equal to the upper-bound of T_{CMB} at that redshift (Noterdaeme, Petitjean, Srianand, Ledoux, & López, 2010). The strength of the two C I lines, which occur at rest frame wavelengths of approximately 1560.31 Å and 1656.93 Å, can be used to derive the relative populations of its energy levels (Morton, York, & Jenkins, 1988). For example, Ge, Bechtold, & Black (1997) used C I absorption in quasar spectra to measure T_{CMB} to be 11.6 ± 1.0 K at a redshift of z = 1.97. C I is one of the only detected species suitable for measuring CMBR temperatures, as its fine-structure levels have the very small energy separation levels.

Using C I to Probe the Early Universe

Besides being an excellent cosmic thermometer, C I also serves as a useful probe of cold gas in the interstellar medium (ISM), the vast space between star systems of a galaxy composed of gases and dust that vary in temperature and composition. Since stars form from dense molecular clouds containing cold gas and dust, studying the physical and chemical states of these regions at high redshifts will provide vital clues to mechanisms of star formation in the early universe (Ledoux, Noterdaeme, Petitjean, & Srianand, 2015). While cold, neutral, and therefore well-shielded gases in the ISM of distant galaxies are usually difficult to detect, C I has a relatively low ionization potential, so it is often found in the same regions as them. Thus, targeting C I absorbers is an efficient way to search for absorption lines of other cold gases, such as H_2, which can be used to investigate the environments from which stars formed billions of years ago. Damped Lyman-α systems (DLAs), absorbers that contain relatively high column densities of neutral hydrogen, are also often targeted to study interstellar gas; unlike C I absorbers, however, DLAs only excel at probing warm neutral gas in the universe (Petitjean *et al.*, 2000).

Furthermore, previous studies suggest that C I absorbers can be targeted to study interstellar dust content. Heintz *et al.* (2019a) showed that C I content is positively correlated with dust extinction, a phenomenon where interstellar dust scatters light of shorter wavelengths, causing objects to appear redder than expected. The degree of dust extinction can be studied at galaxies of various redshifts to investigate how dust content in galaxies chemically evolves over time. Heintz *et al.* (2019b) suggested that C I content is also correlated with the strength of a rare dust extinction feature that occurs at a rest frame wavelength of 2175 Å, which has been studied in different galaxies, including the Milky Way, to track dust and metal content in the universe.

These claims can, in part, be assessed and expanded on by analyzing gaseous metal abundances in a larger sample of C I absorbers. Morton (1975) showed that several interstellar gaseous refractory elements, such as Fe, can be depleted due to being coalesced in dust grains, while a few non-refractory elements, such as Zn, are generally undepleted (York & Zura, 1982). Comparing the abundances of depleted elements to that of Zn in C I absorbers provides a general sense of the prevalence and nature of dust in their host galaxies, which can be compared with other systems such as DLAs and even parts of the Milky Way to draw further conclusions.

Sloan Digital Sky Survey Dataset

The Sloan Digital Sky Survey (SDSS; York *et al.*, 2000) contains a sample of over 500,000 quasar spectra that can be searched for C I absorbers. Each SDSS quasar spectrum contains 4,373 flux data points for wavelengths from 3,800 Å to 10,400 Å. When the wavelength is plotted against flux, significant visual dips indicate absorption lines. However, when the redshift of an intervening galaxy

containing the species of interest is unknown, the observed wavelength of absorption is also unknown, significantly complicating the algorithms typically used to detect spectral features. Traditional methods of detecting absorption features are extremely time consuming, often taking months to survey the entire SDSS dataset while requiring frequent human intervention. As a part of the process, spectra are typically fit to a continuum, the accuracy of which plays a significant role in the algorithm's ability to detect absorption lines. However, there is an unavoidable tradeoff between computational speed and accuracy. Spectra can be quickly fitted with a smoothing spline function, but this process often yields inaccurate results that require frequent human intervention. The more reliable alternative, which involves implementing Principal Component Analysis (PCA), takes many hours to produce a fit for just a single spectrum, which is simply unfeasible for such large datasets. Thus, a computationally fast but still reliable method for detecting absorption features would be extremely useful in this field.

Convolutional Neural Networks

Machine learning is a branch of artificial intelligence involving the use of algorithms and statistical models that allow a computer to make decisions without explicit human instruction, relying on data to recognize patterns. Deep learning is a subfield of machine learning based on multi-layer artificial neural networks that are loosely modeled on the human brain. A convolutional neural network (CNN) is a class of deep neural networks that excels in analyzing images (Gu *et al.*, 2015). Take, for example, the task of image classification, where a CNN takes in an image and classifies it as one of the possible categories given, perhaps as either a dog or a human. One could manually define the features that the CNN should look for in the image in order to make an accurate classification (e.g. the number of visible legs would be a good distinguishing feature between dogs and humans), but defining features in terms of image pixels is exceptionally difficult, if not impossible. So, in essence, a CNN is a model that trains itself on many samples to learn which features it must detect in visual data to perform best in a given task. Thus, CNNs are highly compatible with Astronomy, a field that almost exclusively relies on quantifiable, visual data to make conclusions.

The structure of a CNN includes an input layer, an output layer, and multiple layers in between them, which are referred to as hidden layers. Each layer is composed of individual units called neurons, and the outputs for each layer are fed as inputs into the next layer, until the final (output) layer is reached. The input layer consists of the original data; in the case of an image, each neuron of the input layer may represent the color value of a single pixel. There are multiple different types of hidden layers in CNNs, the most important of which is the convolutional layer. The neurons of a convolutional layer are called *filters*. Each filter is a fixed-length matrix of values referred to as *weights* and slides across the input data matrix, computing the dot-product, also known as a convolution, of the filter and the portion of the input

matrix beneath it at every location as it moves across. The result of each convolution is modified by other operations detailed below before finally being added to an output matrix, called a *feature map*. Therefore, a convolutional layer comprised of *n* filters will output *n* feature maps. Assuming that the input matrix is one-dimensional, the feature map produced by the k^{th} filter in a convolutional layer is formally defined in Equation 3:[1]

$$A_i^k = \sigma(\left(W^k \cdot s_i\right) + b_k) \quad (3)$$

where A_i^k is the i^{th} value of the k^{th} feature map in convolutional layer *A*, σ is a non-linear function that transforms the input to an output (e.g. the hyperbolic tangent function, $\sigma(x) = tanh(x)$), W^k represents the corresponding filter matrix, s_i represents the corresponding portion of the input matrix being convolved with W^k, and b_k represents a bias value that is added to the convolved result. Note that s_i must have the same length as W^k. The weights of a filter can be set to values such that convolving the filter with a portion of the input data yields a higher value if a certain feature, such as the edge of an object, is present in the data; this may end up increasing the value of a certain output neuron, making the CNN more likely to classify the data as whichever category that output neuron represents. However, the CNN must learn the optimal values of its weights by modifying their values such that it minimizes the loss function, a measure of error in a neural network. The loss is calculated by comparing the model's prediction of the training data with the ground truth of the training data, which can be thought of as an answer key. In the case of binary classification, where the CNN outputs either a 0 or 1 (two different categories) for each prediction, the loss function being minimized is typically the binary cross-entropy cost function, which is given in Equation 4:

$$L\left(w_1, w_{2,} w_3, \ldots w_{n-1}, w_n\right) = -\frac{1}{m}\sum_{i=1}^{m} y_i \, log(a_i) + (1 - y_i) \, log(1 - a_i) \quad (4)$$

where the loss *L* is a function of *n* total weights in the CNN, a_i is the i^{th} value in the model output (either a 0 or 1 in binary classification), y_i is the ground truth of that prediction (0 or 1), and *m* is the total number of outputs. As the gradient of a multivariable function points to the function's direction of steepest ascent, the gradient of $L(w)$ is repeatedly calculated using back-propagation algorithms in order to update the weight values such that they move in the opposite direction of the gradient and thus decrease the loss function (Rumelhart *et al.*, 1986). This method is known as *gradient-descent* and is summarized in Equation 5:

$$w_i \rightarrow w_i' = w_i - \eta \frac{\partial L}{\partial w_i} \quad (5)$$

[1] See Lecun, Bengio, & Hinton (2015), for more information on the equations and definitions introduced in this section.

where w_i represents the original value of some weight in the CNN, w_i' represents the updated value of w_i, and η represents the learning rate parameter, which is set before training begins.

Besides hidden convolution layers, the presence of pooling (or subsampling) layers and dropout layers are also key to the success of a CNN. A pooling layer decreases the number of parameters in the CNN in order to reduce training time. For example, *max pooling* groups multiple adjacent values from the previous layer and only outputs the highest value to the next layer. A dropout layer randomly sets the values of neurons in that layer to zero, which is meant to reduce overfitting during the training process by forcing neurons whose values may not have affected the output neurons as much to play a more important role in the CNN. In fully connected layers, every neuron is connected to every neuron in the next layer. Figure 1 shows the CNN architecture of LeNet-5 (Lecun, Bottou, & Bengio, 1998), a well-known model used for image classification.

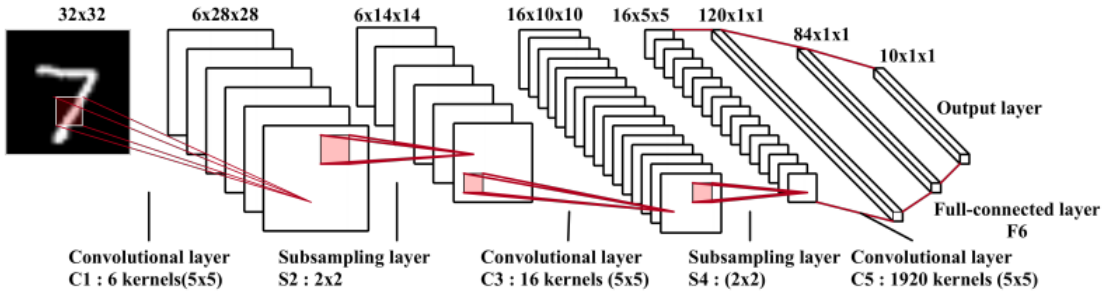

Figure 1. Le-Net 5 CNN architecture used to classify handwritten digits. A 32x32 image of a handwritten "7" serves as the input data to the CNN, which contains three convolutional layers with varying numbers of 5x5 kernels, and two subsampling layers. Each convolutional layer can be thought of as a set of feature detectors, and each subsampling layer reduces the dimensionality of the input data, thus reducing training time. The data is eventually flattened into one dimension before the CNN spits out one of ten possible outputs (digits). Figure adapted from Gu *et al.* (2015).

Purpose

Zhao *et al*. (2019) developed a CNN that classified approximately 50,000 quasar spectra based on the presence of narrow Magnesium (Mg) II absorption lines in only nine seconds on a 1 GeForce Tesla M40 Graphics Processing Unit (GPU), achieving an overall accuracy of 94.1%. The overwhelming success of this study suggested that CNNs could be used to identify spectral features of varying strengths, possibly replacing time-consuming traditional detection methods that require frequent human intervention. Although analyses of a large enough C I absorber sample could potentially lead to highly influential findings in various subfields of cosmology, few have been detected thus far. For example, Ledoux *et al.* (2015) was only able to detect 66 C I absorbers in a systematic search of 41,696 C I absorbers. Therefore, the purpose of this project is to develop a convolutional neural network that detects C I absorbers in the Sloan Digital Sky Survey Data Release 12 quasar spectra dataset, as well as derive their average metal abundances to draw conclusions on the properties of C I host galaxies.

Methodology

Overview

Traditional detection methods were first used to discover C I absorbers from a dataset of 41,894 Mg II absorbers within SDSS DR12. When the dataset was narrowed down to 2,036 possible C I absorber candidates, the author inspected each candidate one by one and deemed 113 of the spectra as real C I absorbers; the detections were verified by a mentor. Then, artificial spectra with and without inserted C I absorption lines were generated to train a CNN. The line insertion code was adopted from a previous project, but the author independently experimented to determine the most effective parameters specific to this research. The final CNN architecture was developed and coded by the author; the data-processing pipeline was also coded by the author.[2] A mentor's code was adopted for smaller, miscellaneous portions of the CNN (e.g. feeding data into the model). The trained CNN was assessed by its ability to re-detect the 113 C I absorbers. Mean composite spectra were generated to measure average abundances of various metals among the 113 C I host galaxies via their absorption lines. The author inspected each spectrum individually to determine which ones were suited for being incorporated in the composite spectrum. Here, experimentation also incentivized the author to modify various parameters in the code to improve the accuracy of the composite spectra.

Searching Dataset

It was originally planned to conduct a blind survey of the Sloan Digital Sky Survey (SDSS) DR12 dataset, which includes nearly 500,000 quasar spectra, for C I absorbers. However, it was later concluded that the searching window for a spectrum, which ranges from ~3800 to ~10400 Å, was far too large for even a trained neural network to consistently detect C I absorption lines, as they are narrow, weak, and difficult to distinguish from the heavy noise in the data. Instead, the dataset in which C I absorbers were searched for was restricted to only the 41,894 SDSS-12 quasar spectra containing Mg II absorption lines detected by Zhao *et al*. (2019). As C I absorbers are a subset of Mg II absorbers, the wavelengths at which Mg II absorption lines were already observed can be used to derive the redshift of their host galaxies, thus significantly reducing the searching window. In other words, if Mg II absorption is found to occur at a redshift of z in one of the spectra, it is known that a C I λ1560 absorption line, if one exists, would be observed at a wavelength $\lambda_{obs} \approx 1560.31\ (1 + z)$, as explained earlier. Thus, the searching window could be reduced from the whole spectrum to just the region around $\lambda \approx 1560.31\ (1 +$

[2] Code for the starting CNN architecture was provided by a mentor. However, after completely revamping the data-processing method later on into the project, the author replaced the original model with a new architecture better suited for the input data size.

z), causing any existing absorption features to stand out considerably more while drastically reducing computation time; the same goes for the detecting the C I λ1656 absorption line**.**

Using Traditional Methods to Discover C I Absorbers

Traditional methods for detecting spectral features were applied to discover new C I absorbers in the Mg II absorber dataset within SDSS DR12, so the CNN's ability to re-detect them could be evaluated later on. The process began with a candidate pool of 41,894 spectra containing Mg II absorption lines, which were loaded on Python to be analyzed and eventually narrowed down to a smaller subset of C I absorbers. Spectra where the redshifts (*z*) of observed Mg II absorption lines were too high or too low to observe potential C I λ1560 or λ1656 absorption lines at the same *z* were filtered out, leaving only spectra where Mg II absorption occurred at ~1.43 < z < ~2.67. Code was also written to eliminate a spectrum if a potential λ1560 line was within 40 Å from possibly interfering Lyman-alpha emission lines, hydrogen spectral lines occurring at a rest-frame wavelengths of ~1215 Å that commonly arise at the source quasar. This narrowed down the sample to 12,766 C I spectra. Next, each remaining spectrum was fit to a continuum using the locally weighted scatterplot smoothing (LOWESS) algorithm, a nonparametric curve-fitting regression method. Flux value data points where the residual (the flux value at that point minus the flux value predicted by the curve at that wavelength) was greater than the 3 standard deviation (σ) error, calculated from the local S/N (signal-to-noise ratio), were given weightings of zero in the fitting. Starting from the curve fitting, this process was repeated three times so that the curve fitting would not be influenced by outliers, such as absorption and emission lines. Then, each spectrum was normalized; that is, each flux value was divided by the value predicted by the curve at that point. Extensive experimentation was conducted by the author to determine the most effective parameter values (e.g. the number of times standard deviation filtering was applied).[3] Using a weighted least-squares method, Gaussian curves were fit to the data around potential C I λ1560 and λ1656 absorption features. If the peaks of both curves in a spectrum had significance levels (error in σ, abbreviated as SL) of at least 2σ, or if at least one peak had a SL of at least 3σ, the spectrum was deemed as a C I absorber candidate. The rest of the spectra were filtered out, leaving only 2,036 remaining C I candidates. From there, each spectrum was closely inspected, one by one, by the author to eliminate spectra where the SL was high at λ1560 and λ1656 due to factors other than C I absorption, such as bad continuum fitting around those regions or nearby absorption lines that influenced the Gaussian curve fitting. It was finally determined that at least 113 of the spectra contained real C I absorption lines, and they were officially deemed as newly detected C I absorbers. Detections made by the author were verified by a mentor.

[3] Initial continuum fitting and normalization code were provided by a mentor, but there were many areas to be improved. The author suggested these changes to a mentor, who added his own insights or approved of them.

An example of a detected C I absorber is shown in Figure 2.

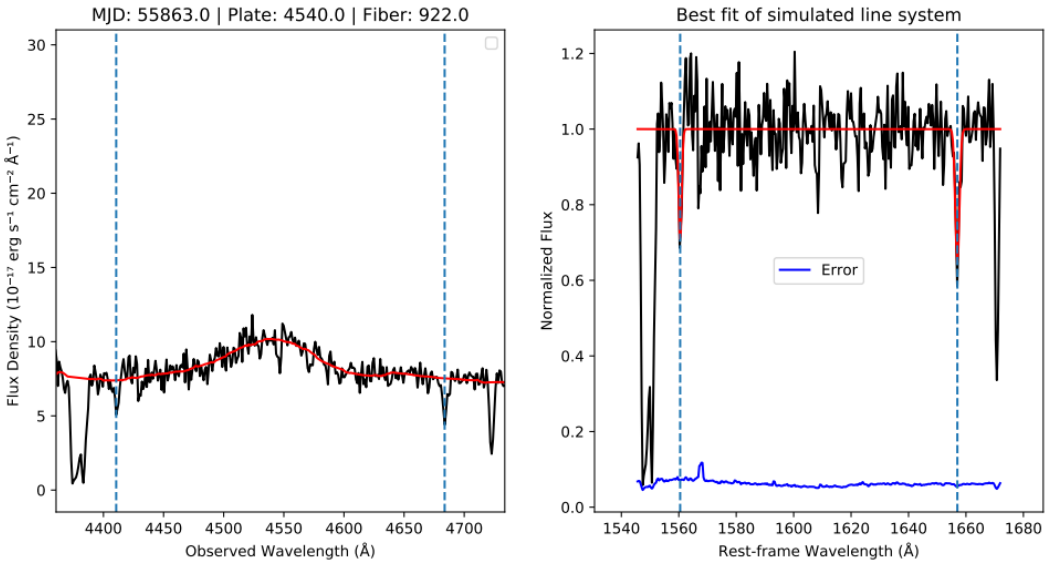

Figure 2. An example of a C I absorber detected by traditional methods and confirmed after visual inspection. The left plot shows a portion of the unmodified spectrum, identified by its Modified Julian Date (MJD), plate, and fiber, towards a quasar located at a right ascension (RA) of ~8.86 and a declination (dec) of ~9.67. The red line represents the continuum fit. The two dashed blue lines mark the C I λ1560, 1656 lines. Since Mg II absorption was known to occur at $z = 2.246$, the blue lines mark the observed wavelengths of the two C I lines at $\lambda = 1560.31(1 + 2.246)$ Å and $\lambda = 1656.93(1 + 2.246)$ Å. The right plot shows the spectrum after normalization. Gaussian profiles are fit to both lines with respective significance levels of 3.77 and 5.45. Graphic by author.

Generating Artificial Training Data

Deep neural networks typically require several thousand labeled training samples in order to reach respectable performance metrics, as was the case for this study. Since the existing C I sample was far too small to use as training data alone, it was necessary to create a large training dataset of spectra with and without artificially inserted C I absorption lines. 20,000 spectra were randomly selected from the Baryonic Oscillation Spectroscopic Survey (BOSS) quasar spectra dataset within SDSS DR12. The spectra were normalized and fit to a continuum using the same method described previously. Spectral regions where the S/N was greater than 3.0 were identified using the noise values provided by SDSS. Then, C I λ1656 and λ1560 lines were randomly inserted in these regions with SL greater than 3.4σ and 2.0σ (calculated from the local S/N) respectively, as the former line is ~1.7 times as strong as the weaker line according to their measured oscillator strengths (Morton, York, & Jenkins, 1988). The line insertion code was adopted from Zhao *et al*. (2019), but parameters, such as the wavelengths and SL of injected absorption lines, were modified. These spectra were deemed the positive sample, as they contained inserted C I lines. The remaining 10,000 spectra, which were deemed the negative sample, underwent the same process, except no C I lines were inserted. Out of the 20,000 training samples, 20% (4,000) of them were randomly selected to make up the validation set, while the remaining 80% (16,000) made up the

training set. The CNN does not train on the validation sample, but classifies them after every epoch (iteration) of training on the training set. If the training accuracy is significantly higher than the validation accuracy during training, the model is most likely overfitting on the training data. Additionally, 2,000 positive and 2,000 negative samples were generated as the testing set, which were classified by the CNN after training as means of evaluating the model.

Data Preprocessing

A data preprocessing pipeline was developed to modify the 24,000 spectra before they were fed into the CNN. Gaussian noise was added to each artificial spectrum to simulate real noise, but only data points on the rest-frame wavelength interval [1360.31 Å, 1856.93 Å], which is from 200 Å blue-ward of the λ1560 line to 200 Å red-ward of the λ1656 line, were kept and normalized. Originally, all the data in this ~500 Å wide window was simply sent to the CNN for training, but this mandated the insertion of many other artificial absorption lines commonly found across the window that the CNN had to learn to distinguish from C I lines (C IV λ1548, C IV λ1550, Fe II λ1608, Al II λ1670, etc.), resulting in a much more complicated training process. After multiple stages of experimenting with various pre-processing methods, it was decided to trim the interval further by combining two small, separate pieces of the larger window together: one with data points extremely close to the C I λ1560 line, and one with data points extremely close to the C I λ1656 line. The implementation is described as follows. The first data point at a wavelength greater or equal to 1560.31 Å was located, and only that data point and the closest 16 data points on each side were kept (approximately ± 6 Å from the λ1560 line). These 33 data points were concatenated with the 33 data points obtained from repeating the same process at 1656.93 Å. Thus, the pre-processing phase reduced spectra with 4,373 data points down to only 66 data points. By masking out all irrelevant data, the CNN no longer had to distinguish the two C I lines from other absorption lines in between or close to them on each spectrum, thereby simplifying the training process greatly. Note that Zhao *et al.* (2019) simply input the entire spectrum to the CNN when detecting Mg II absorbers, as their absorption lines are much stronger and easier to pick out even from >4,000 points; the specific pre-processing method proposed in this research was designed by the author and a mentor, and all code was written by the author. Positive samples were labeled ‘1’ and negative samples were labeled ‘0’ before being fed into the CNN. Figure 3 compares a normalized spectrum with artificially inserted C I λ1560, 1656 lines before and after the final preprocessing step, with a real detected absorber before and after the final preprocessing step.

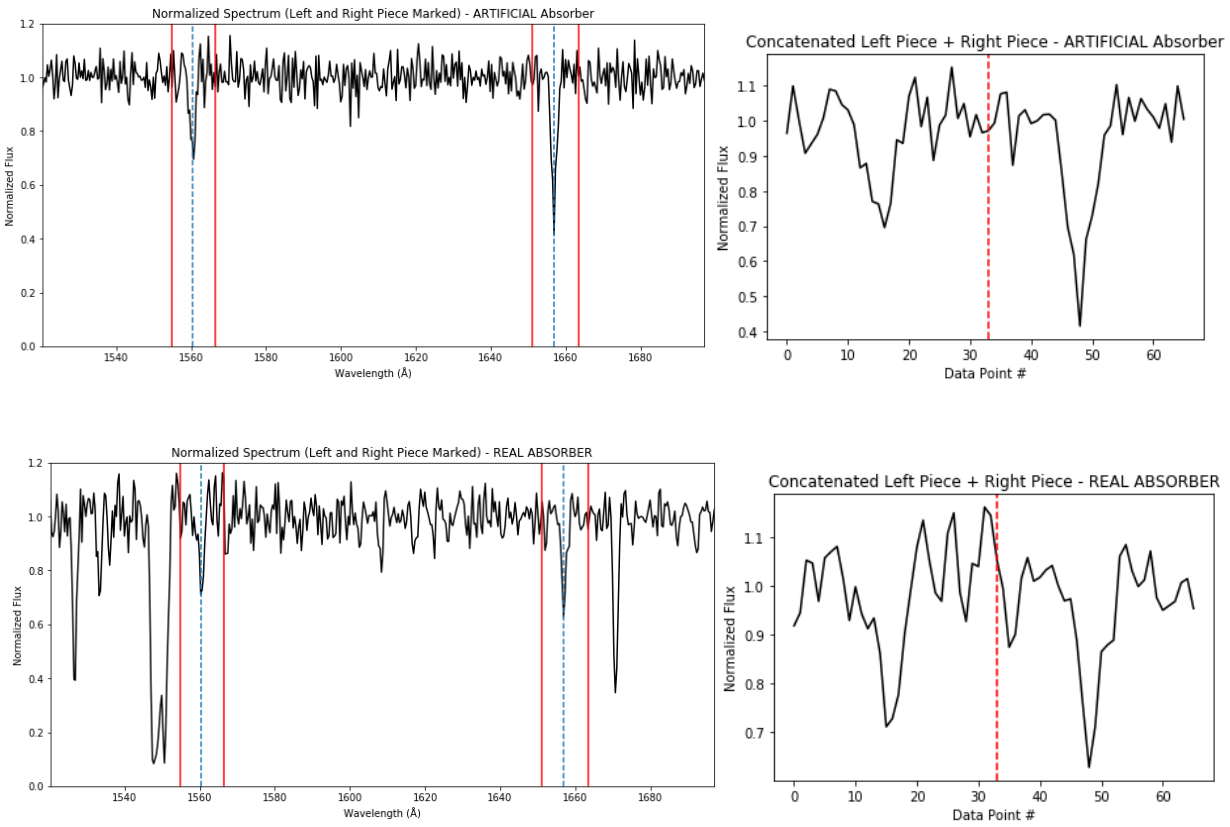

Figure 3. Comparison between a real and synthetic C I absorber. The top left plot shows a normalized spectrum with artificially inserted Gaussian noise and artificially inserted C I λ1560, 1656 lines. The bottom left plot shows the normalized spectrum for a real detected C I absorber. The dashed blue lines indicate the wavelengths of absorption. Notice that the real spectrum, unlike the synthetic one, contains many non-C I absorption lines, such as C IV λ1548, 1550, Si II λ1526, and Al II λ1670. However, since the final preprocessing step only keeps and concatenates the two pieces enclosed by the vertical red lines (33 data points per piece), the CNN no longer has to distinguish C I lines from other lines. So, the final data forms of the artificial and real absorbers, shown on the top and bottom right respectively, are very similar to each other. The dashed red line separates the two halves taken from the normalized spectrum. Graphic by author.

Development and Final Architecture of the CNN

The CNN model was implemented in the programming language Python 3.7.6 using the open-source neural network libraries TensorFlow and Keras, which include built-in functions that implement all of the algorithms and CNN layers described. All neural network training was done using an Nvidia GeForce RTX 2080 Ti Graphics Processing Unit (GPU). Various standard CNN architectures were repeatedly evaluated, tweaked, and compared with each other, including AlexNet (Kriezhevsky, Sustskever, & Hinton, 2012), GoogLeNet (Szegedy et al., 2014), ResNet (Kaiming, Zhang, Ren, & Sun, 2015), and VGGNet (Simonyan, Zisserman). The stochastic gradient descent optimization algorithm, which uses the standard gradient descent algorithm to update weights after being trained on a subset of the training sample, was used to train the CNN; the size of each subset is a parameter referred to as the batch size, which was set to 32. A mentor provided code for an initial CNN architecture. However, after

the data-processing method was revamped to include much fewer data points, the author independently developed his own CNN architecture best suited for the 1 x 66 input data. It included 2 convolutional layers with 8 and 16 filters respectively, and a dense layer with 1024 neurons. Batch normalization layers, which standardize the inputs to a layer, were added after each convolutional layer. The rectified linear unit (ReLU) activation function was applied to each hidden layer. Further experimentation conducted by the author incentivized more specific changes in the architecture as well.[4] For example, the filter (kernel) size in the first convolutional layer was reduced to 10, as trial and error revealed that it was the best matrix length to detect the relatively weak and narrow C I absorption lines. Since the CNN performed binary classification, the singular output neuron's value was equal to the predicted probability that the input spectrum was a C I absorber. Thus, the sigmoid activation function was applied to the output layer. The ReLU and sigmoid functions are given in Equations 6 and 7, respectively (Goyal, Goyal, & Lall, 2019):

$$R(x) = max(0, x) \quad (6) \qquad \sigma(x) = \frac{1}{1 + e^{-x}} \quad (7)$$

where x is the output value from the previous layer.

Evaluation Metrics

The performance of each CNN was evaluated by calculating its accuracy and recall when classifying the testing and training data, as well as the 113 real C I absorbers detected using traditional methods. The accuracy is defined as the ratio of correct predictions to all predictions, which provides insight on the overall performance of the CNN. The recall is defined as the ratio of correctly predicted positive predictions to the number of positive predictions, which provides insight on the CNN's ability to detect positive samples in the dataset.

Measuring Column Densities and Dust Depletion

This research also derived the average abundances of various metals among the 113 newly discovered C I absorbers to investigate their physical composition. Instead of going through the 113 samples one by one, all spectra were stacked to compute a mean composite spectrum for each metal absorption line that was measured. Using the redshift of Mg II absorption, each spectrum was shifted to the quasar rest-frame wavelength. Next, each spectrum was cut down to only include data points on the interval [λ_{obs} - 55 Å, λ_{obs} + 55 Å], where λ_{obs} is the rest-frame wavelength of absorption for the line being measured. Remaining data points, excluding those within the region of absorption, were used the fit the spectrum to a continuum with a smoothing spline function. Iterative standard deviation filtering was

[4] A mentor contributed code for feeding in data to the CNN and outputting the CNN's prediction.

performed to mask out outliers affecting the fit. Then, the fit was used to normalize the spectrum. Each spectrum was closely inspected by the author and removed if its continuum fit was deemed too inaccurate, or if data points nearby λ_{obs} were missing. Since each spectrum contained the same number of data points, a composite spectrum was computed by simply taking the median normalized flux value at each index out of the 113 (or slightly fewer, if any were removed) processed spectra after standard deviation filtering was applied to the data points at each index. A mean error value at each index j, E_j, of the composite spectrum was also computed via Equation 8 (Mas-Ribas *et al.*, 2017):

$$E_{j=}\frac{1}{\sqrt{\sum_i \frac{1}{e_{ij}^2}}} \quad (8)$$

where e_{ij} is the error value at index j for the i^{th} spectrum among the stack, obtained from the provided SDSS error data. This process was repeated at several wavelengths of absorption to generate composite spectra that would later be used to measure various Al, Cr, Fe, Mn, Si, Ti, and Zn absorption lines.
Figure 4 displays the generated mean composite spectrum centered at 1608 Å, in preparation of measuring the moderately weak Fe II λ1608 line.

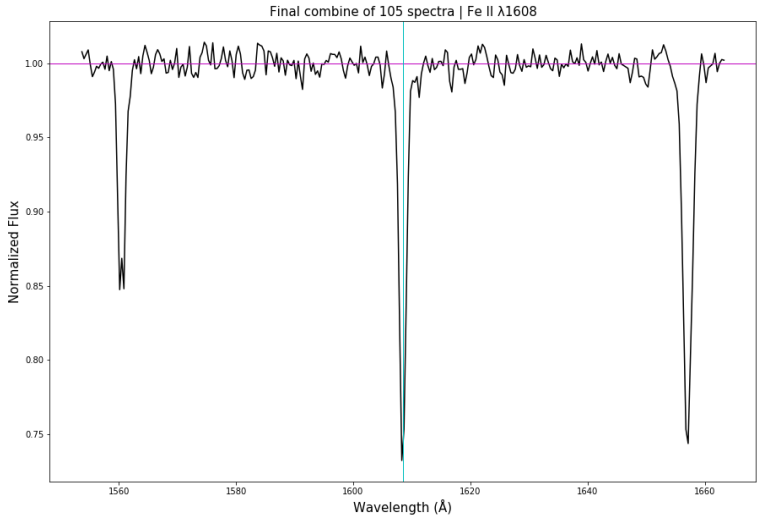

Figure 4. An example of a mean composite spectrum. It was generated from stacking 105 C I absorbers on the interval [1608 – 55 Å, 1608 + 55 Å]. Out of the original sample of 113 spectra, 8 were handpicked and removed for having heavily inaccurate continuum fits. The vertical cyan line marks the center of the composite Fe II λ1608 line, which is a good representation of the feature's average strength across all detected C I systems. Graphic by author.

For each stacked spectrum, a Gaussian profile was fit to the the data points at the absorption feature using the Non-Linear Least-Squares Minimization and Curve-Fitting (LMFIT) library in Python, which employs a χ^2 minimization algorithm. The profile's base was set to the continuum, and its center was given 1 Å of wiggle-room from the theoretical rest-frame wavelength of absorption. The parameters of

the fit were used to measure the stacked metal line's equivalent width (W_λ), a measure of an absorption feature's strength, using Equation 9:[5]

$$W_\lambda = \left| d\sigma\sqrt{2\pi} \right| \quad (9)$$

where d and σ are the depth and standard deviation of the profile, respectively. As an example, Figure 5 shows the solved Gaussian fit for the stacked Si II λ1808 line.

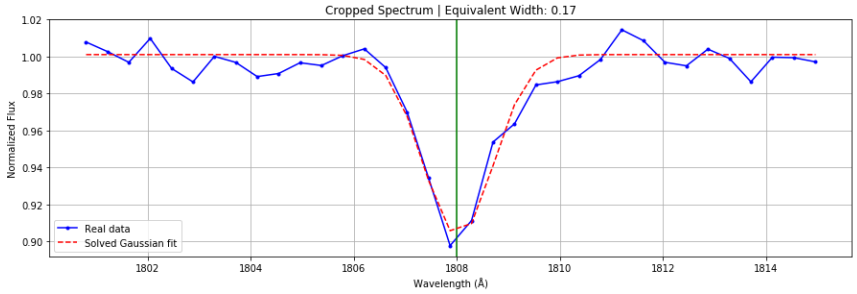

Figure 5. An example of a solved Gaussian fit to an absorption feature. The vertical green line marks the theoretical center of the Si II λ1808 line. The solution for the solved Gaussian fit is shown in red. The base of the profile is the continuum. The noticeable asymmetry of real data points covering the absorption feature, shown in blue, may be attributed to blending with another absorption feature on the right side of the feature; W_λ is measured to be 0.17 Å. Graphic by author.

The fits were created by the author and their accuracies were verified by a mentor.[6] From there, the equivalent widths of the absorption lines were used to measure their species' column densities using the apparent optical depth method (AODM; Savage & Sembach, 1991). The column density (N) of a species is equivalent to the number of absorbing atoms per cm^2 from the line of sight of the observer to the absorbing galaxy. In the case of weak (optically thin) lines, a linear relationship exists between the equivalent width of a line and the column density of its corresponding species, which is adopted in this project via Equation 10 (Morton, York & Jenkins, 1988):

$$N = 1.130 \times 10^{20} \frac{W_\lambda}{f\lambda^2} \quad (10)$$

where f is the oscillator strength of the line and λ is the rest-frame wavelength of absorption. However, as lines become stronger (optically thick), they become saturated: the $W_\lambda \propto N$ relationship evolves into a much slower growing $W_\lambda \propto \sqrt{lnN}$ relationship, and then a $W_\lambda \propto \sqrt{N}$ relationship. To identify saturated lines, the measured W_λ for Fe II λ2260, 2344, 2374, 2382, 2586, and 2600 were used to generate a curve

[5] On a wavelength versus intensity plot, the equivalent width of a spectral line is defined as the width of a rectangle with the same height and area as the spectral line. The well-known integral of a Gaussian function can be used to derive a general formula for any Gaussian profile's equivalent width.

[6] Initial code segments for generating mean composite spectra and Gaussian profiles were provided by a mentor. However, after experimenting, the author independently modified various parameters (e.g. the smoothing parameter for the smoothing spline function) to improve the measurements' accuracies.

of growth (COG), a theoretical plot of W_λ against N for a system.[7] The parameters for the COG calculated using the Fe II equivalent widths need not be elaborated here, but this research followed an identical method to York *et al*. (2006). Fe II λ2382, Fe II λ2600, and Al II λ1670 do not lie on the linear portion of the COG, indicating that their lines are saturated. Thus, their column density measurements using Equation 6 were deemed as extreme lower limits. While the COG can be used to estimate the column densities of other species as well, the AODM was still adopted for the sake of accuracy. This study also accounted for lines that were blended together (Zn II and Mg I at ~λ2026; Zn II and Cr II at ~λ2062) by following an identical method to York *et al*. (2006) as follows. $W_{\text{Cr II }\lambda2056}$ and $W_{\text{Cr II }\lambda2066}$ were averaged to estimate the value of $W_{\text{Cr II }\lambda2062}$, as the oscillator strength of the Cr II λ2062 is nearly equal to the average of the other two Cr II lines' strengths. Then, the calculated value of $W_{\text{Cr II }\lambda2062}$ was subtracted from the measured equivalent width at λ2062 to estimate $W_{\text{Zn II }\lambda2062}$. Similarly, the value of $W_{\text{Mg I }\lambda2582}$ was used to infer the value of $W_{\text{Mg I }\lambda2026}$, which was subtracted from the measured equivalent width at λ2026 to estimate $W_{\text{Zn II }\lambda2026}$. Errors for these values should be very small. The overall column density (N_{wm}) for a species was calculated by taking the weighted mean of all the measured column densities of unsaturated lines for that species (i.e. six different Fe II lines were measured, but only four of them were unsaturated and included when calculating N_{wm}(Fe)), using Equation 11 (Pomme & Keightley, 2015):

$$N(X)_{wm} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \quad (11)$$

where a species X has n measured unsaturated lines. Next, the abundances of metals relative to Zn (apparent or dust depletions), [X/Zn], were measured and are hereafter referred to as how they are defined in Equation 12 (e.g. Welty *et al.*, 1999):

$$\left[\frac{X}{Zn}\right] = log\left(\frac{N(X)}{N(Zn)}\right) - log\left(\frac{N(X)}{N(Zn)}\right)_{solar} \quad (12)$$

where $log\left(\frac{N(X)}{N(Zn)}\right)_{solar}$ is the abundance of a metal X relative to Zn in the sun's photosphere.[8] Recall that Zn is a non-refractory element that is relatively undepleted in comparison to other metal lines measured in this study, so comparing their column densities provides insight to the dust content of C I host galaxies. If the value of $log\left(\frac{N(X)}{N(Zn)}\right)$ is much smaller than the solar abundance of that element relative to Zn, then that gaseous metal is significantly more depleted due to being coalesced on dust grains, implying a dustier environment. The depletion patterns of Al, Si, Mn, Cr, Fe, Ni, and Ti in C I absorbers were also compared

[7] A mentor provided the implementation code for the COG. However, all equivalent width measurements were made by the author, and so all column density measurements were as well.
[8] Solar abundances are taken from Asplund *et al.* (2009).

to that of the Milky Way discs and halo, as well as that of Damped Lyman-α systems (Welty *et al.*, 1999; Nas *et al.*, 2017).

Results and Discussion

CNN Training & Testing Performance

After 50 epochs, the final CNN model reached a training accuracy of 99.82% and achieved a testing accuracy of 96.30%. The training accuracy and loss curves are shown in Figure 6.

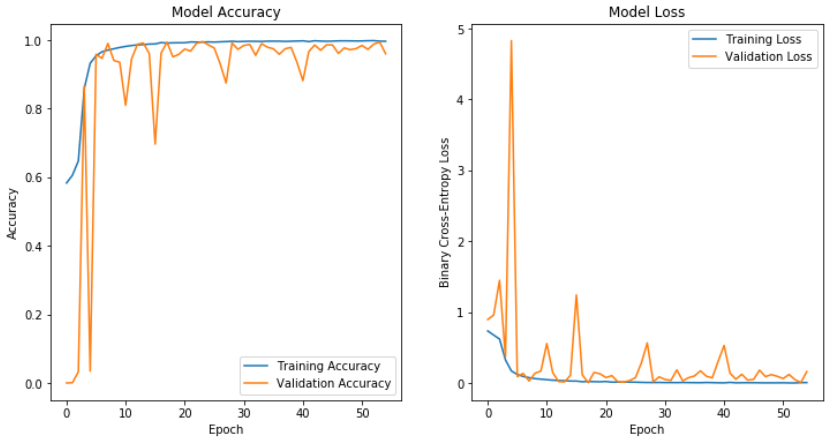

Figure 6. Training accuracy and loss curves over the course of 50 epochs. On the left plot, the training accuracy quickly converges towards 100% while the validation accuracy gradually converges over time. On the right plot, the training binary cross-entropy loss quickly converges towards 0 while the validation loss gradually converges. Graphic by author.

The CNN took approximately 14 minutes to finish training on the 20,000 spectra, whereas traditional detection methods would have taken far longer to analyze them.

CNN Performance on Newly Detected C I Absorbers

When applied to the 113 C I absorbers discovered using traditional detection methods in this project, all preprocessed using the same method as the artificial spectra, the CNN achieved a recall of 92.92%, successfully detecting 105 out of 113 samples. This shows that the CNN not only successfully classifies the artificially generated spectra, but also detects the large majority of positive samples. The CNN took under 1 second to classify all 113 C I absorbers.

Column Densities and Dust Depletions

Table 1 shows the measured column densities and depletions (relative abundances to Zn) of all metals in the stacked spectra. The measurements reported for the Zn II λ2026, 2062 lines have accounted for contamination with other lines. The column densities for individual lines were directly used to calculate the overall column densities reported.

Table 1. Results from measuring various stacked lines. Column densities for individual lines, overall column densities for metals, and metal abundances relative to Zn (apparent depletions) are shown. Oscillator strength (*f*) values are taken from Morton, York, & Jenkins (1988). All errors reported are ± 1σ. Lower bound column density measurements indicate that the line is saturated. The subscript [r] indicates that a species was removed from the metal's overall column density measurement due to its line either being saturated or too weak (W_λ < 3σ) to be considered a reliable detection. The subscript [c] indicates that the species' line is contaminated; the W_λ reported is the final value obtained after attempts to remove contamination, as described in the methodology. The subscript [i] indicates that the line's W_λ was inferred using the *f*-values of other lines of the same species. Table by author.

Species (*X*)	λ (Å)	*f*	W_λ (Å)	log*N*(*X*)	logN_{wm} (*X*)	[*X*/Zn]
Al II[r]	1670.79	1.740	1.03 ± 0.007	13.38 ± 0.003	13.22 ± 0.010	-1.57 ± 0.025
Al III	1854.72	0.559	0.31 ± 0.008	13.26 ± 0.011		
Al III	1862.79	0.278	0.19 ± 0.010	13.35 ± 0.022		
Cr II	2056.25	0.167	0.03 ± 0.007	12.68 ± 0.091	12.81 ± 0.049	-1.22 ± 0.054
Cr II[i]	2062.23	0.121	0.04 ± 0.005	12.89 ± 0.058		
Cr II	2066.16	0.0798	0.04 ± 0.013	13.12 ± 0.122		
Fe II	1608.45	0.062	0.40 ± 0.008	14.45 ± 0.009	14.33 ± 0.003	-1.56 ± 0.024
Fe II	2260.78	0.0028	0.08 ± 0.010	14.06 ± 0.006		
Fe II	2344.21	0.108	0.94 ± 0.012	14.25 ± 0.006		
Fe II	2374.46	0.0395	0.49 ± 0.008	14.40 ± 0.007		
Fe II[r]	2382.77	0.328	1.40 ± 0.010	>13.93		
Fe II	2586.65	0.0573	0.84 ± 0.011	14.39 ± 0.006		
Fe II[r]	2600.17	0.203	1.40 ± 0.018	>14.06		
Mg I[i]	2026.48	0.112	0.02 ± <0.0005	12.69 ± <0.0005	—	—
Mg I	2852.96	1.81	0.65 ± 0.012	12.70 ± 0.008		
Mn II	2576.88	0.288	0.09 ± 0.010	12.73 ± 0.046	12.65 ± 0.039	-1.18 ± 0.045
Mn II	2594.50	0.223	0.05 ± 0.010	12.58 ± 0.079		
Mn II	2606.46	0.158	0.03 ± 0.009	12.50 ± 0.114		
Ni II	1709.60	0.047	0.06 ± 0.012	13.69 ± 0.079	13.72 ± 0.035	-0.90 ± 0.041
Ni II	1741.55	0.0679	0.09 ± 0.010	13.69 ± 0.046		
Ni II	1751.92	0.04	0.07 ± 0.011	13.81 ± 0.063		
Si II	1808.00	0.0055	0.17 ± 0.011	15.03 ± 0.027	15.03 ± 0.027	-0.88 ± 0.036
Ti II[r]	3073.88	0.0573	0.05 ± 0.027	12.76 ± 0.188	12.55 ± 0.027	-0.79 ± 0.035
Ti II	3242.93	0.183	0.06 ± 0.004	12.55 ± 0.028		
Ti II	3384.74	0.284	0.12 ± 0.024	12.62 ± 0.079		
Zn II[c]	2026.14	0.412	0.13 ± 0.009	12.94 ± 0.029	12.96 ± 0.024	0
Zn II[c]	2062.66	0.202	0.08 ± 0.007	13.02 ± 0.039		

Figure 7 compares measurements of relative abundances to Zn, [*X*/Zn], for various species in C I host galaxies with those of the Milky Way cold disk, warm disk, and halo, measured by Welty *et al*. (1999), and ~27,000 DLA systems (Mas-Ribas *et al*., 2017). The measured average relative abundances in C I absorbers are in broad agreement with that of the MW disks. Al, Si, Mn, Cr, and Fe are consistently a few times more depleted in C I systems than they are in the MW warm disk, with the exception of Ni and Ti.

Al, Si, and Mn depletion levels closely follow the trend of the MW cold disk, but Ni and Ti are several times more depleted in C I host galaxies. These inconsistencies may be attributed to differences in the nature of dust or nucleosynthesis, considering that C I absorbers are found at high redshifts, with the average z of the sample being 1.89, while the MW is at $z = 0$. Regardless, C I absorbers are most similar to the MW disks in terms of dust depletion. All elements measured are also consistently several times more depleted than those in DLAs and the MW halo.

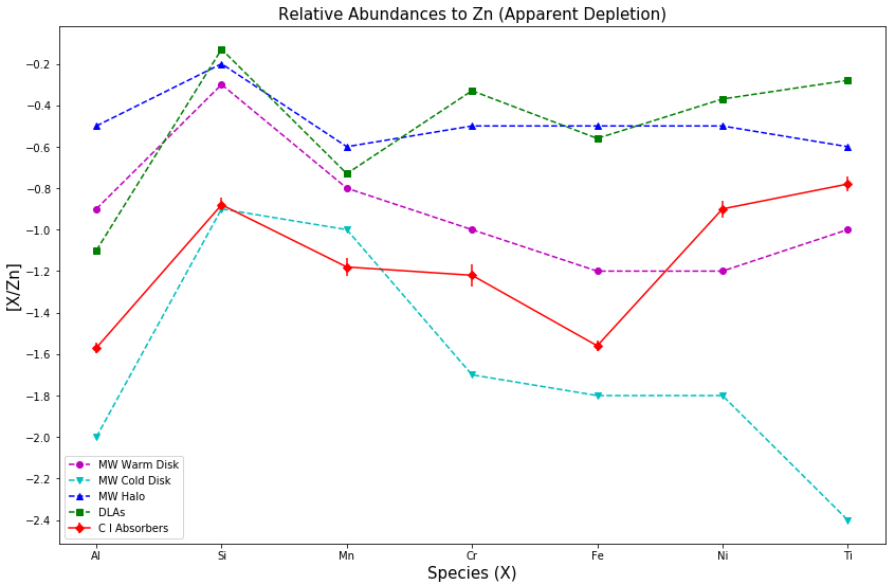

Figure 7. Comparison of depletion patterns. Relative abundances of various metals to Zn plotted against that of the Milky Way disks and halo (Welty *et al*., 1999), and DLAs (Mas-Ribas *et al.*, 2017). Error bars represent ± 1σ. The C I depletion pattern is in broad agreement with that of the Milky Way warm disk. This may suggest that C I absorbers are, in general, similar in composition and thus relatively dusty systems as well. Graphic by author.

Relation to Previous Literature

While depletions in C I host galaxies have not been measured in the past, this research supports previous literature that classified C I absorbers as excellent probes of the well-shielded, cold, neutral gas of the ISM (e.g. Ledoux, Noterdaeme, Petitjean, & Srianand, 2015). Since C I is easily photoionized by ultraviolet radiation, it is probable that any detected C I is well-shielded by dust in the ISM. This especially must be true considering that the mean redshift of C I absorption in the sample presented here is 1.89, while numerous studies have suggested that star formation in galaxies peaks at $z \approx 2$ (Madau & Dickinson, 2014). Therefore, star formation is very much prevalent in C I host galaxies, increasing the amount of local UV radiation and thus increasing the amount of dust necessary to shield C I from photoionization. So, it makes sense that the depletion patterns obtained in this analysis are in line with that of the relatively dusty MW disks. It is also no surprise that the metals have been measured to be several times more depleted than those found in DLAs, which have been shown to instead probe warm,

and thus not as deep nor shielded, neutral gas (Petitjean *et al.*, 2000). Therefore, it is heavily encouraged to conduct future analyses of this sample, which provides a rare opportunity to investigate star-forming environments.

Limitations

There were a number of limitations in this study that most likely worsened the performance of the CNN and decreased the accuracy of column density measurements. Perhaps the most significant limitation was the low resolution of quasar spectra taken from SDSS DR12. Since flux data exists at wavelengths between 3,800 Å and 10,400 Å and there are only 4,373 recorded data points per spectrum, data points are separated from each other by an average distance of ~1.5 Å. As a result, the Gaussian profiles that were fit to absorption features had significant room for error. In fact, absorption lines in high resolution spectra are typically fit with Voigt profiles, which more accurately model features when there are many data points. The high noise in the data also contributed to inaccuracies in the profile fitting and increased the 1σ error for line measurements, which were derived from the local S/N. Both of these limitations also complicated the CNN training process, mandating the generation of artificial noise in the synthetic training sample. As C I absorption lines are relatively weak, the CNN most likely also struggled to differentiate them from the heavy noise in the data, causing false negatives. Even when traditional methods were used to detect C I absorbers, quite a few of them were most likely missed due to limitations of the data; it can be assumed that the Gaussian fits for the weakest C I absorption lines did not exceed the 2σ or 3σ SL limit necessary to be deemed a C I absorber candidate by the computer program due to high local noise, and thus were unknowingly eliminated. Of the 2,036 candidates that were closely inspected by hand, a significant amount of real absorbers were also likely missed simply due to uncertainties. To avoid the risk of including too many false positives in the 113 C I absorber sample that was later analyzed, this research was extremely conservative in regard to classifying spectra as C I absorbers.

Future Research

This research can be expanded on in multiple aspects. In this study, the CNN only re-detected the 113 C I absorbers identified using traditional methods. There is still the rest of SDSS DR12, which includes ~40,000 Mg II absorbers, to be surveyed for missed C I absorbers. There are also ~50,000 Mg II absorbers recently discovered by Zhao *et al.* (2020, in preparation) yet to be surveyed, of which at least a few hundred of them are most likely undiscovered C I absorbers. The CNN could even search through the already surveyed SDSS DR7 quasar spectra dataset for new C I absorbers previously missed by Ledoux *et al.* (2015), whose blind search was only able to detect 66 C I absorbers out of 41,696 spectra using traditional detection methods.

In addition to measuring average metal column densities and dust depletion patterns, future studies can use the newly discovered C I absorbers to study the physical properties of cold, neutral gases whose absorption lines are typically difficult to detect due to being well-shielded in the ISM. These C I host galaxies are a good representation of typical environments for star-formation at high redshifts, as stars form from the dense, and thus cold, concentrations of interstellar gas and dust within molecular clouds. A CLOUDY (Ferland *et al.*, 2013) photoionization model could be constructed to derive average physical conditions of the C I absorbing clouds. The future is especially exciting in this regard, as previous studies have thoroughly investigated star formation rates at high redshifts, but not the specific physical properties of the galaxies in which such phenomena occurs (e.g. Marastron *et al*., 2010). Furthermore, significant improvements can be made on the artificial training data and CNN architecture itself. More time can be devoted to analyzing spectra where the CNN missed C I absorption lines. Finding common trends in the spectra that were misclassified by the CNN can provide a sense of direction when tweaking the model to account for them, reducing the amount of time spent aimlessly changing the network's parameters. The detected C I absorption lines can also be measured to determine the relative population ratios of C I energy levels in each spectrum, which can be ultimately used to measure CMBR temperatures at various redshifts in order to rigorously test the Big Bang theory. Perhaps the most significant implication of this research, however, is the potential to expand the application of CNNs to detect any species of interest via their absorption lines. The data-preprocessing pipeline presented here, in itself, serves as a proof of concept that CNNs can successfully identify any relatively weak spectral feature, as long as the redshift of absorption is known. For instance, future studies can apply a similar pre-processing method and an improved CNN architecture to detect rare Ca II absorbers, whose properties will place constraints on models for the existence of cool gas in extended galactic regions (Sardane, Turnshek, & Rao, 2015), or DLAs, which probe neutral gas evolution and the kinematics of galaxy disks. The application of CNNs in surveying quasar spectra will tremendously speed up and simplify the process of obtaining large absorber samples to study, opening new doors in the field of cosmology.

Summary

Traditional detection methods were used to detect 113 C I absorbers in the SDSS DR12 quasar spectra dataset. A convolutional neural network was developed to re-detect them. The final model achieved a training accuracy of 99.81% and a testing accuracy of 96.30% on the artificially generated spectra, and it successfully detected 92.92% of the 113 C I absorbers discovered using traditional methods in this project. A spectrum stacking technique was implemented to create mean composite spectra, which were used to measure the average column densities and depletions of several metals via their absorption lines. Depletion patterns of the C I absorbers are in agreement to those of the Milky Way disks, implying that C I host galaxies are relatively dusty systems, as predicted by previous literature.

Literature Cited

Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. (2009). The Chemical Composition of the Sun. *Annual Review of Astronomy and Astrophysics*, 47, 481-522. Retrieved November 8, 2020, from Astrophysics Data System database.

Durrer, R. (2015). The Cosmic Microwave Background: The history of its experimental investigation and its significance for cosmology. *Classical Quantum Gravity*, 32, 12. Retrieved July 20, 2020, from Astrophysics Data System database.

Ferland, G. J., Porter, R. L., van Hoof, P. A. M., Williams, R. J. R., Abel, N. P., Lykins, M. L., . . . Stancil, P. C. (2013). The 2013 Release of CLOUDY. *RevMexAA*, 49, 137-163. Retrieved November 8, 2020, from Astrophysics Data System database.

Ge, J., Bechtold, J., & Black, J. H. (1997). A New Measurement of the Cosmic Microwave Background Radiation Temperature at z = 1.97. *The Astrophysical Journal*, 474, 67-73. Retrieved July 20, 2020, from Astrophysics Data System database.

Glover, S. C. O., & Clark, P. C. (2016). Is atomic carbon a good tracer of molecular gas in metal-poor galaxies? *Monthly Notices of the Royal Astronomical Society*, 456, 3596-3609. Retrieved July 20, 2020, from Astrophysics Data System database.

Goyal, M., Goyal, R., & Lall, B. (2019). Learning Activation Functions: A new paradigm for understanding Neural Networks. *Proceedings of Machine Learning Research*, 101, 1-18. Retrieved November 8, 2020, from Astrophysics Data System database.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, Bing., . . . Chen T. (2015). Recent Advances in Convolutional Neural Networks. *Computing Research Repository*, 1512. Retrieved November 8, 2020, from Astrophysics Data System database.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770-778. Retrieved November 8, 2020, from IEEE XPlore database.

Heintz, K. E., Zafar, T., De Cia, A., Vergani, S. D., Jakobsson, P., Fynbo, J. P. U., . . . Andersen, C. (2019). On the dust properties of high-redshift molecular clouds and the connection to the 2175 Å extinction bump. *Monthly Notices of the Royal Astronomical Society*, 486, 2063-2074. Retrieved July 20, 2020, from Astrophysics Data System database.

Heintz, K.E., Ledoux, C., Fynbo, J. P. U., Jakobsson, P., Noterdaeme, P., Krogager, J.-K., . . . Kaper, L. (2019). Cold gas in the early Universe: Survey for neutral atomic-carbon in GRB host galaxies at $1 < z < 6$ from optical afterglow spectroscopy. *Astronomy & Astrophysics*, 621, A20. Retrieved July 20, 2020, from Astrophysics Data System database.

Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences of the United States of America*, 15, 168-173. Retrieved July 20, 2020, from Astrophysics Data System database.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional

Neural Networks. *Neural Information Processing Systems*, 25. Retrieved November 8, 2020, from ResearchGate database.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436-444.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, H. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE.* Retrieved November 8, 2020, from IEEE XPlore database.

Ledoux, C., Noterdaeme, P., Petitjean, P., & Srianand, R. (2015). Neutral atomic-carbon quasar absorption-line systems at z > 1.5. *Astronomy & Astrophysics*, 580, A8. Retrieved July 20, 2020, from Astrophysics Data System database.

Liszt, H. S. (1981). The carbon abundance in diffuse interstellar clouds. *Astrophysical Journal*, 246, 147-150. Retrieved July 20, 2020, from Astrophysics Data System database.

Madau, P., & Mark, D. (2014). Cosmic Star-Formation History. *Annual Review of Astronomy and Astrophysics*, 52, 415-486. Retrieved November 8, 2020, from Astrophysics Data System database.

Marastron, C., Pforr, K., Renzini, A., Daddi, E., Dickinson, M., Cimatti, A., & Tonini, C. (2010). Star formation rates and masses of z ~ 2 galaxies from multicolor photometry. *Monthly Notices of the Royal Astronomical Society*, 407, 830-845. Retrieved November 8, 2020, from Astrophysics Data System database.

Mather, J.C., Fixsen, D. J., Shafer, R. A., Mosier, C., & Wilkinson, D. T. (1999). Calibrator Design for the *COBE* Far Infrared Absolute Spectrophotometer (FIRAS). *The American Astronomical Society*, 512, 511-520. Retrieved July 20, 2020, from Astrophysics Data System database.

Mas-Ribas, L., Miralda-Escude, J., Perez-Rafols, I., Arinyo-i-Prats, A., Noterdaeme, P., Petitjean, P., . . . Ge, J. (2017). The Mean Metal-line Absorption Spectrum of Damped Lyα Systems in BOSS. *The Astrophysical Journal*, 846(36). Retrieved July 20, 2020, from Astrophysics Data System database.

Morton, D. C. (1975). Interstellar Absorption Lines in the Spectrum of Zeta Ophiuchi. *The Astrophysical Journal*, 197, 85-115. Retrieved November 8, 2020, from Astrophysics Data System database.

Morton, D. C., York, D. G., & Jenkins, E. B. (1988). A Search List of Lines for Quasi-Stellar Object Absorption Systems. *The Astrophysical Journal Supplement Series*, 68, 449-461. Retrieved November 8, 2020, from Astrophysics Data System database.

Noterdaeme, P., Ledoux, C., Petitjean, P., & Srianand, R. (2008). Molecular hydrogen in high-redshift damped Lyman-α systems: the VLT/UVES database. *Astronomy & Astrophysics*, 481, 327-336. Retrieved July 20, 2020, from Astrophysics Data System database.

Noterdaeme, P., Petitjean, P., Srianand, R., Ledoux, C., & Lopez, S. (2010). The evolution of the Cosmic Microwave Background Temperature Measurements of TCMB at high redshift from carbon monoxide excitation. *Astronomy & Astrophysics*, 526, L7. Retrieved July 20, 2020, from Astrophysics Data System database.

Peebles, P. J. E. (1993). *Principles of Physical Cosmology*. New Jersey: Princeton University Press.

Petitjean, P., Srianand, R., & Ledoux, C. (2000). Molecular hydrogen and the nature of damped Lyman-α systems. *Astronomy & Astrophysics*, 364, 26-30. Retrieved July 20, 2020, from Astrophysics Data System database.

Pomme, S., & Keightley, J. (2015). Determination of a reference value and its uncertainty through a power-moderated mean. *Metrologia*, 52(3). Retrieved November 9, 2020, from IOPscience database.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533-536. Retrieved July 20, 2020, from Astrophysics Data System database.

Sardane, G., Turnshek, D., & Rao, S. (2015). Ca II absorbers in the Sloan Digital Sky Survey: element abundances and dust. *Monthly Notices of the Royal Astronomical Society*, 452, 3192-3208. Retrieved November 8, 2020, from Astrophysics Data System database.

Savage, B. D., & Sembach, K. R. (1991). The Analysis of Apparent Optical Depth Profiles for Interstellar Absorption Lines. *The Astrophysical Journal*, 379, 245-259. Retrieved November 8, 2020, from Astrophysics Data System database.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Neural Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556.

Songalia, A., Cowie, L. L., Vogt, S., Keane, M., Wolfe, A. M., Hu, E. M., . . . Lanzetta, K. (1994). Measurement of the microwave background temperature at a redshift of 1.776. *Nature*, 371, 43-45. Retrieved November 8, 2020, from Astrophysics Data System database.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going Deeper with Convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 1-9. Retrieved November 8, 2020, from IEEE Xplore database.

Welty, D. E., Frisch, P. C., Sonneborn, G., & York, D. G. (1999). Interstellar Abundances in the Magellanic Clouds. II. The Line of Sight to SN 1987A in the Large Magellanic Cloud. *The Astrophysical Journal*, 512, 636-671. Retrieved November 8, 2020, from Astrophysics Data System database.

York, D. G., Adelman, J., Anderson Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., . . . Yasuda, N. (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120, 1579-1587. Retrieved November 8, 2020, from Astrophysics Data System database.

York, D. G., & Jura, M. (1982). Observations of Interstellar Zinc. *The Astrophysical Journal*, 254, 88-93. Retrieved November 8, 2020, from Astrophysics Data System database.

Zhao, Y., Ge, J., Yuan, X., Zhao, T., Wang, C., & Li., X. (2019). Identifying Mg II narrow absorption lines with deep learning. *Monthly Notices of the Royal Astronomical Society*, 487, 801-811. Retrieved July 20, 2020, from Astrophysics Data System database.