Using deep Residual Networks to search for galaxy-Ly α emitter lens candidates based on spectroscopic selection

Rui Li,[1,2,3,4]★ Yiping Shu,[5,6]★ Jianlin Su [7] Haicheng Feng,[1,2,3,4] Guobao Zhang,[1,2,3,4] Jiancheng Wang[1,2,3,4]★ and Hongtao Liu [1,2,3,4]

[1] *Yunnan Observatories, Chinese Academy of Sciences, 396 Yangfangwang, Guandu District, Kunming, 650216, P. R. China*
[2] *University of Chinese Academy of Sciences, Beijing, 100049, P. R. China*
[3] *Center for Astronomical Mega-Science, Chinese Academy of Sciences, 20A Datun Road, Chaoyang District, Beijing, 100012, P. R. China*
[4] *Key Laboratory for the Structure and Evolution of Celestial Objects, Chinese Academy of Sciences, 396 Yangfangwang, Guandu District, Kunming, 650216, P. R. China*
[5] *Purple Mountain Observatory, Chinese Academy of Sciences, 2 West Beijing Road, Nanjing, Jiangsu, 210008, China*
[6] *Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[7] *School of Mathematics, Sun Yat-sen University, Guangzhou, China*

ABSTRACT

More than 100 galaxy-scale strong gravitational lens systems have been found by searching for the emission lines coming from galaxies with redshifts higher than the lens galaxies. Based on this spectroscopic-selection method, we introduce the deep Residual Networks (ResNet; a kind of deep Convolutional Neural Networks) to search for the galaxy-Ly α emitter (LAE) lens candidates by recognizing the Ly α emission lines coming from high- redshift galaxies ($2 < z < 3$) in the spectra of early-type galaxies (ETGs) at middle redshift ($z \sim 0.5$). The spectra of the ETGs come from the Data Release 12 (DR12) of the Baryon Oscillation Spectroscopic Survey (BOSS) of the Sloan Digital Sky Survey III (SDSS-III). In this paper, we first build a 28 layers ResNet model, and then artificially synthesize 150 000 training spectra, including 140 000 spectra without Ly α lines and 10 000 ones with Ly α lines, to train the networks. After 20 training epochs, we obtain a near-perfect test accuracy at 0.995 4. The corresponding loss is 0.002 8 and the completeness is 93.6 per cent. We finally apply our ResNet model to our predictive data with 174 known lens candidates. We obtain 1232 hits including 161 of the 174 known candidates (92.5 per cent discovery rate). Apart from the hits found in other works, our ResNet model also find 536 new hits. We then perform several subsequent selections on these 536 hits and present five most believable lens candidates.

Key words: gravitational lensing: strong – galaxies: structure.

1 INTRODUCTION

Galaxy-scale strong gravitational lens, which can provide very tight constraints on the projected mass of the lens galaxies, is not only an available probe to study the structures, formation, and evolution of galaxies (e.g. Koopmans et al. 2006; Auger et al. 2010; Bolton et al. 2012; Sonnenfeld et al. 2013), but also a crucial technique to study cosmology (Suyu et al. 2013). Additionally, gravitational lenses could act as natural telescopes and give us magnified views of the background objects, which could help us to study the celestial bodies in the distant universe (e.g. Impellizzeri et al. 2008; Swinbank et al. 2009; Treu et al. 2015).

Several methods have been used to search for the galaxy-scale strong gravitational lenses. Among them, an efficient one is searching for the arc-like morphological features of strong lens systems (Lenzen, Schindler & Scherzer 2004; Horesh et al. 2005; Seidel & Bartelmann 2007; Kubo & Dell'Antonio 2008; More et al. 2012). In order to find strong lens systems in a great deal of galaxy images., Gavazzi et al. (2014) subtracted the light of the central galaxies using multiband images and then searched for the arc-like shapes in the residual images. Another modified work was done by Joseph et al. (2014), who successfully subtracted the light of the foreground galaxies with Principal Component Analysis (PCA), and then searched for gravitational lens features in the residual images. They pointed out that the PCA-based galaxy subtraction algorithm

performed better than the traditional model fitting method (Gavazzi et al. 2014).

Besides the method of morphological recognition, Warren et al. (1996) proposed a spectroscopic-selection technique to search for galaxy–galaxy strong lens systems. This technique identifies lens candidates by looking for compound spectra. Let us use the identified lensing system SDSSJ 1110 + 2808 (Shu et al. 2016b) as an example to clarify the compound spectra. The foreground galaxy of SDSSJ 1110 + 2808 is an early-type galaxy (ETG, $z \sim 0.6073$). Between 3800 Å and 5000 Å, there is no emission line in the spectrum. However, near 4150 Å, we find a Ly α line which comes from a Ly α emitter (LAE). The LAE is at the redshift of $z = 2.3999$ and lies close to the line of sight of the foreground ETG. The light from the LAE is deflected by the foreground ETG, and then several images of the background LAE emerge around the foreground galaxy. Further detailed description of the spectroscopic-selection technique can be found in Bolton et al. (2004). In general, this technique identifies lens candidates by searching for the spectra with emission lines from background sources.

Several galaxy-scale strong lens surveys have been made with the above spectroscopic-selection technique, such as the Sloan Lens ACS Survey (SLACS; Bolton et al. 2006, 2008), the Sloan WFC Edge-on Late-type Lens Survey (SWELLS; Treu et al. 2011), the Baryon Oscillation Spectroscopic Survey (BOSS) Emission-Line Lens Survey (BELLS; Bolton et al. 2012), the SLACS for the Masses (S4TM; Shu et al. 2015), and the BELLS for the GALaxy-Lyα EmitteR sYstems Survey (BELLS GALLERY; Shu et al. 2016b). More than 100 galaxy–galaxy strong gravitational lenses have been found in these surveys. A lot of significant works about the structures, formation, and evolution of ETGs have been done with these lens systems (Koopmans et al. 2006; Auger et al. 2009; Bolton et al. 2012; Shu et al. 2015; Li et al. 2018). Recently, using a similar technique, Meyer et al. (2017) have found nine secure QSO-Galaxy lens candidates, presenting the potential application of the spectroscopic-selection technique in QSO Lens searching. So far, this spectroscopic-selection technique is one of the most efficient methods. However, the speed of this traditional method is not fast enough, especially in the coming age of the big data in astronomy.

Recently, the Convolutional Neural Networks (CNNs), a kind of machine learning algorithm, have been applied in astronomy. The first application of the CNNs is the spectral classification of the data release 10 (DR10) of the Sloan Digital Sky Survey (SDSS). Hála (2014) applied his CNN model on more than 60 000 spectra of SDSS and yielded a success rate of nearly 95 per cent. Their work conclusively proved the great potential of the CNNs in astrophysics. Later, the CNNs have been applied in some other works, such as the morphological classification of SDSS galaxies (Dieleman, Willett & Dambre 2015), the photometric redshift estimates of SDSS galaxies (Hoyle 2016) and the star/galaxy classification (Kim & Brunner 2017). In the field of gravitational lenses, Petrillo et al. (2017) and Lanusse et al. (2018) have used the CNNs to recognize galaxy-scale strong gravitational lenses based on morphological classification. In Petrillo et al. (2017), they found 761 strong lens candidates in the Kilo Degree Survey (KiDS) and presented 56 most reliable candidates after manual selection.

The CNNs have not been applied for searching gravitational lens candidates based on the spectroscopic-selection technique. In this paper, we will use the deep Residual Networks (ResNet; one popular model of CNNs) to search for the galaxy-LAE lens candidates in DR12 of SDSS by recognizing the high-redshift ($2 < z < 3$) Lyα emission lines. LAEs are a kind of young, low-mass galaxies with an extremely high star formation rate. They are important components of the early universe, and critical for studying the formation and evolution of galaxies. Benefited from the magnifying effect, galaxy-LAE lenses can provide us good opportunities to study the LAEs themselves. More importantly, galaxy-LAE lenses can help us to exploit the small-scale dark substructures around the lens galaxies and constrain the slope and normalization of the substructure-mass function (Shu et al. 2016a).

The paper is organized as follows. In Section 2, we provide a brief description of Machine Learning, CNNs, and ResNet. In Section 3, we talk about how to use ResNet to recognize emission lines that are not from the foreground galaxies themselves. In Section 4, we present five new gravitational lens candidates found by our ResNet model. In Section 5, we discuss the advantages and the possible improvement of our strategy.

2 MACHINE LEARNING, CNNS, AND RESNET

In this section, we provide a brief introduction of Machine Learning, CNNs, and ResNet.

2.1 The Machine Learning and Deep learning

Machine Learning was coined in 1959 by Samue (1959). It is a field of computer science that gives computers the ability to learn from train data and then make predictions on predictive data. Depending on whether there are given 'labels' for the learning systems, Machine Learning is typically classified into two classes, supervised learning and unsupervised learning. For supervised learning, the train data as well as the 'labels' for the train data are given to the learning algorithm. Then the learning algorithm would 'learn' a general rule that maps inputs to outputs. For unsupervised learning, no 'labels' are given, the learning algorithm has to discovery hidden patterns in the train data by itself. The Machine Learning algorithm could be designed to achieve various tasks, such as classification, regression, clustering, density estimation, and so on. The task in our work is in fact a classification task which can be tackled in a supervised way. The train data in our work could be divided into two classes and given different 'labels'. Then the learning algorithm could 'learn' from the train data and their 'labels' then produce a model that assigns the inputs to one of the two classes.

Deep learning is a kind of Machine Learning with multiple layers to extract the feature of the inputs. Each later layer uses the outputs of the previous layer as inputs. Most deep learning models are based on artificial neural networks (ANNs) which consist of many artificial neurons. Similar with biophysical neurons, the artificial neurons of ANNs can receive inputs, change their states according to the inputs, and then produce outputs depending on their states. If the artificial neurons of deep ANNs can only receive informations from the previous layer and output informations to the next layer, we call this kind of ANNs as deep feedforward neural networks. There are no cycles or loops in these kinds of networks. More detailed information about Deep learning can be found in Lanusse et al. (2018).

2.2 The CNNs

As one of the most popular models of deep learning, recent years, the CNNs have gained a lot of attentions in image recognition, speech recognition, motion analysis, nature language processing, as well as many other researching fields. CNNs are a kind of deep feedforward
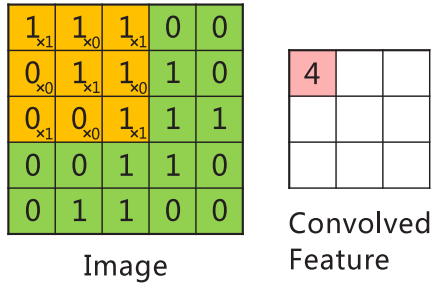
Figure 1. Illustration of the convolutional layer. A 5×5 image (green) is convolved by a 3×3 convolutional kernel (yellow). The result of this convolutional operation is a 3×3 feature map. The convolutional kernel moves on the image by 1 stride and then output the convolutional result to the feature map.

neural networks and formed by a stack of distinct layers. The layers mainly contain the convolutional layers, the pooling layers and the fully connected layers. Here are some important terminologies of CNNs used in this paper:

(i) *Convolutional layers:* The convolutional layers apply convolutional operations to the inputs and then pass the results to the next layers. Fig. 1 (from the internet, just for illustration) shows a simple convolutional operation. A 5×5 input image (green) is convolved by a 3×3 convolutional kernel (yellow, also known as convolutional neurons). The result of this convolutional operation is a 3×3 feature map. Simply speaking, the convolutional layers are used to extract different features of the inputs. Usually, there are several convolutional neurons in one feature map. The convolutional neurons are composed of weights and bias which could be written as $\{W, b\}$. The weights and bias could be trained by the train data.

(ii) *Pooling layers:* The pooling layers are introduced between two convolutional layers (local pooling layers) or before the fully connected layers (global pooling layers) to compress the feature maps and simplify the calculations. There are two common pooling methods, max pooling and average pooling. Fig. 2 (from the internet, just for illustration) shows a max pooling with 2×2 filters. Each filter could find the maximum value of the corresponding local area to represent this local area in order to form a new compressed feature map.

(iii) *Fully connected layers:* The fully connected layers connect every neuron in one layer to every neuron, in another layer to turn all the local feature maps into a global feature map.

(iv) *Dropout layers:* All of the parameters of the weights and bias would occur in the fully connected layers, which probably cause overfitting. One method to reduce the overfitting is to introduce the Dropout layers. During each training process, we can randomly 'drop out' some parameters of $\{W, b\}$ with probability $1 - p$ or keep the parameters with probability p. After a 'Dropout' operation, a reduced network is left and would be gained on data.

(v) *Activation function:* The Activation function defines the outputs of a neuron given inputs. A neuron receives signals and 'activate' them using Activation function, then output the activated signals to the next neuron. For CNNs, only non-linear Activation functions could be used, because if the Activation functions are linear, there were only linear transformations between two layers, which could reduce the capacity of the networks. Additionally, non-linear Activation functions can limit the outputs to be finite values (Gulcehre et al. 2016).
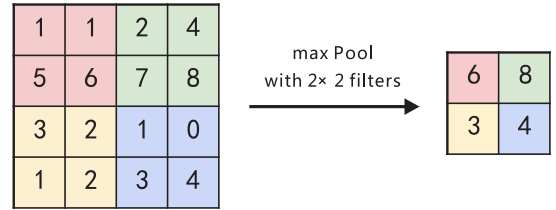
MAX POOLING

Figure 2. Illustration of the pooling layer. It is a max pooling layer with 2×2 filters. The filters will find the maximum values of the corresponding local areas and then use the maximum values to represent these local areas to form a new compressed feature map.

(vi) *Loss:* A Machine Learning model is trained by minimising a given loss that measure the deviation between the predicted values and the real values. The loss function could be defined by the user.

(vii) *Accuracy and completeness:* The accuracy and completeness are used to measure the capacity of a classification model. For example, we have totally m samples, among them, n samples are positive samples, the remain are negative. If i samples are classified as positive by our model, among them, j samples are real positive. Then the accuracy is defined as j/i, the completeness is defined as j/n.

(viii) *Epoch:* Epoch is a terminology of Keras[1] used in this work. Keras is a high-level neural networks application programming interface (API), and capable of running on top of TensorFlow,[2] CNTK,[3] or Theano.[4] For Keras, the train dataset would be divided into lots of batches, the networks are updated when a batch is processed. An epoch is an iteration over the entire dataset.

The entire training process of a CNN model is as follow: (1) Initialize the weights and bias $\{W, b\}$ of the CNN model. (2) The inputs are convolved by the convolutional layers, resulting in some feature maps. Then the feature maps are compressed by the pooling layers. Repeat this process if necessary. (3) At the end of the networks, the fully connected layers are used to obtain the predicted values. (4) Calculate the deviation between the predicted values and the input values. If the deviation is bigger than our expectation, update the weights and bias $\{W, b\}$ according to the deviation and retrain the networks. Fig. 3 (from the internet, just for illustration) shows a simple but complete CNN model with two convolutional layers, two pooling layers and one fully connected layer. The input image is convolved by three convolutional kernels, resulting in three feature maps. These three feature maps are then compressed by a pooling layer. Next, the outputs of the pooling layer are processed by another convolutional layer and pooling layer. Finally, a fully connected layer is used to turn the local feature maps into a global feature map and output the prectived values.

2.3 The ResNet

Many evidences show that the network depth is very important for the classification and regression tasks. Theoretically, the training results of the networks should become better as the layers become

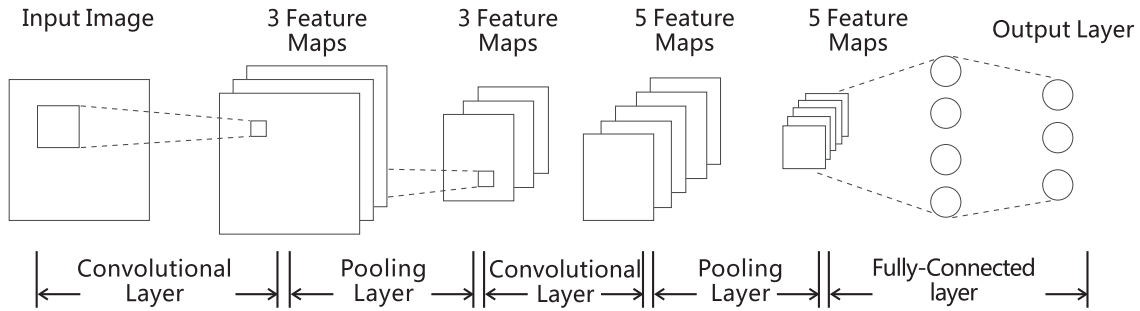[1] https://keras.io
[2] https://github.com/tensorflow/tensorflow
[3] https://www.microsoft.com/en-us/cognitive-toolkit/
[4] http://deeplearning.net/software/theano/

Figure 3. A simple but complete CNN model with two convolutional layers, two pooling layers, and one fully connected layers.
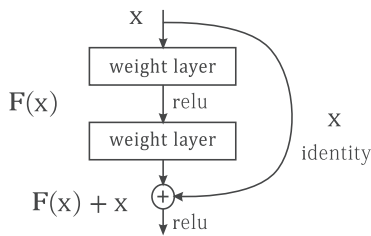

Figure 4. Residual learning: a building block with a shortcut connection (He et al. 2015). The weight layers are the assemble of several convolutional layers and other necessary layers. Relu (Xavier et al. 2011) is one kind of Activation Function.

deeper. However, for deep networks, a 'degradation' problem would occur when they start converging. With the increase of network depth, the accuracy gets saturated and then degrades rapidly (He et al. 2015), which is not caused by overfitting. That's why many popular CNN models have only a few layers (e.g. five layers in LeNet, seven layers in AlexNet; Lecun et al. 1998; Krizhevsky, Sutskever & Hinton 2012). In order to overcome the degradation of the accuracy of deep networks, He et al. (2015) proposed a new CNN model: the deep Residual Networks.

ResNet assumes that it is easier to optimize the residual mapping than to optimize the origin, unreferenced mapping. Denoting the desired underlying mapping as $H(x)$ which would be fitted by a few stacked layers. x is the inputs of the first layer. ResNet lets the stacked layers to fit the residual mapping of $F(x) := H(x) - x$. Then the original mapping $H(x)$ becomes $F(x) + x$, and it could be realized by the networks with 'shortcut connections' (see Fig. 4). The 'shortcut connections' simply perform the identity mapping, then skip one or more layers and add their outputs to the outputs of the stacked layers. In He et al. (2015), the authors explained that 'With the residual learning reformulation, if identity mappings are optimal, the solvers may simply drive the weights of the multiple nonlinear layers toward zero to approach identity mappings.'

In our study, we use Machine Learning to find gravitational lens candidates based on spectroscopic selection. Spectrum is one-dimensional sequence that has local correlations. We use one-dimensional CNNs to build the Machine Learning model because CNNs have good performance in distinguishing the inputs with local correlations. In order to prevent the 'degradation' problem of deep networks and reduce the training difficulty, we adopt the ResNet model to our study. In fact, before the final choice of ResNet, we have tried several 'plain' networks (e.g. LeNet; AlexNet), but they are difficult to converge in our study.

3 FIND EMISSION LINES WITH RESNET

So far, there are several gravitational lens surveys based on spectroscopic selection (e.g. the SLACS, BELLS, S4TM, and BELLS GALLERY). The SLACS, BELLS, and S4TM concentrated on searching for multiple emission lines (meanly the [O II] 3727 and [O III] 5007) from the galaxies with redshift higher than the foreground galaxies, while the BELLS GALLERY focused on only one Ly α line. We put forward that Machine Learning could be used to search for galaxy-scale gravitational lens candidates based on spectroscopic selection. As the first application, in this study we use ResNet to find some galaxy-LAE lens candidates in the BOSS spectra by searching for high redshift ($2 < z < 3$) Ly α lines in the spectra of middle redshift ($z \sim 0.5$) galaxies. However, in the band used in this work (3800 Å $< \lambda < 5000$ Å, see Section 3.2), the spectra with lower redshift [O II] 3727 or other emission lines could also be found by our networks. We will abandon these spectra using several other subsequent selections (see Section 4).

3.1 The BELLS GALLERY survey

The lens candidates of BELLS GALLERY are selected from the DR12 of the BOSS by searching for Ly α emission lines from high redshift LAEs, and then following up high-resolution *Hubble Space Telescope* (*HST*) imaging observations to confirm the lens nature. In Shu et al. (2016a), the first and most important step is selecting out the spectra with emission lines that are not from the target galaxies themselves. To do this, they confined their research to the observed wavelength range of 3600 Å $< \lambda < 4800$ Å (roughly $2 < z < 3$ for the Ly α lines) and used an error-weighted matched-filter with a Gaussian kernel to search for the Ly α lines in BOSS targets. This main step yielded 4982 hits. Then they performed several subsequent selections to reject the false hits and finally found 187 lens candidates with obvious Ly α emission lines. They presented 21 high quality BELLS GALLERY candidates in their paper. Among them 17 candidates have been confirmed as strong gravitational lenses by *HST* imaging observations. The goal of the BELLS GALLERY survey is exploiting the small-scale dark substructure around ETGs and constraining the slope and normalization of the substructure-mass function (Shu et al. 2016a).

BELLS GALLERY is a successful lens candidates searching work. Whereas, the first step of the search is complex and consume too much time. First, they have to obtain the accurate spectral flux errors across the observed spectra. Secondly, they need to fit each spectrum with standard ETGs spectroscopic templates to obtain the galaxy-subtracted residual spectrum. For millions of targets, these procession would consume a lot of time. In order to increase the

searching speed, we introduce ResNet to automatically select out the spectra with emission lines that are not from the targets themselves.

3.2 The predictive data and the train/test data

We aim to use ResNet to search for galaxy-LAE lens candidates from the DR12 of BOSS of SDSS-III. At beginning, we need to filter out all the ETG spectra from the database to build our predictive data. The BOSS targets have been classified as several classes, such as 'GALAXY', 'SKY', 'STAR'. Therefore, we can choose the 'GALAXY' class and reject the targets with sub-class of 'STARFORMING', 'BROADLINE', and 'AGN'. We then select the target with redshift of $z > 0.3$. These selections result in 1 241 799 recorders. For these recorders, we only use the spectra from 3800 Å to 5000 Å to build our predictive spectra. Note that we do not use the same wavelength range as BELLS GALLERY (3600 Å $<\lambda <$ 4800 Å), because the spectra in the wavelength from 3600Å to 3800 Å is very noisy, which could decrease the accuracy of our ResNet model.

Our predictive data includes 177 known lens candidates from Shu et al. (2016a). The remaining 10 candidates found in their work are eliminated because of the redshift limitation. Among the 177 candidates, there are 3 spectra with Ly α lines falling out of the wavelength of 3800 Å $<\lambda <$ 5000 Å. Therefore, only 174 candidates could be used to test our ResNet model.

Finding lens candidates in a great deal of galaxy spectra is a classification problem. Therefore, we use ResNet to build a classifier which can select out the spectra with emission lines not from the foreground galaxies themselves. Then, we use several other subsequent selections similar with previous work to reject the false hits. The quality of the train/test data is critical for the accuracy of the classifier and directly affect the efficiency of the selection. For the train/test data, we require: (1) the train/test spectra are not in the DR12 of the BOSS of SDSS-III, but look similar as the spectra in the database; (2) the number of the train/test spectra has to be large enough, usually larger than 10 000; (3) the train/test spectra have been classified into two groups, one group has no Ly α lines (labelled [0]), and the other has Ly α lines (labelled [1]); 4) for the real spectra in DR12, the spectra with Ly α lines are much less than the spectra without Ly α lines, which makes the data imbalance. We need a similar data imbalance in the train/test data.

So far, there are no more than 200 galaxy spectra with Ly α emission lines from high redshift, which is far from enough. Therefore, we use the following procedures to build an artificial train/test data set.

(i) We grid the redshift to small bins ($\delta z = 0.002\,5$), and at each redshift bin we perform a principal component analysis (PCA) to the spectra of ETGs (Chen et al. 2012; Shu et al. 2012). Then, we use the first six eigenvectors of the PCA to create some artificial galaxy spectra at the corresponding redshift. This procedure yields totally 150 000 artificial galaxy spectra.

(ii) We add the 'SKY' (from BOSS targets) and noise to the created spectra. Now the spectra look similar as real spectra. We then cut out the spectra ranged from 3800Å to 5000 Å.

(iii) In the rest-frame wavelength, we create 10 000 artificial Ly α emission lines using Gaussian profiles. The mean wavelength of the Gaussian profiles is at 1216Å and the full width at half-maximum (FWHM) is randomly obtained from a normal distribution with the mean value of 300 km s^{-1} and standard deviation of 200 km s^{-1}. We then limit the FWHM greater than 0.5 Å and the peak of the emission lines greater than 1.5×10^{-17} erg cm^2 Å^{-1}.

Finally, we randomly shift the emission lines to the wavelength of 3800 Å $<\lambda <$ 5000 Å.

(iv) We randomly select 10 000 spectra from the 150 000 artificial spectra and add the 10 000 artificial Ly α lines to them, then label them as [1]. The remaining spectra without emission lines are labelled as [0]. The above 1/15 selection is to overcome the data imbalance mentioned above.

Now, we have successfully created 150 000 train/test spectra. The PCA approach we performed in the first step could loss some information of the spectra. We point out that the lost information does not affect our training results. Because, first, the first 6 eigenvectors are enough to fit a real spectrum with an acceptable χ^2; secondly, most of the lost information is from noise, after adding the 'SKY' and noise in the second step, the created spectra look almost the same as real spectra. Fig. 5 shows two high signal-to-noise ratio (SNR) spectra with Ly α emission lines. The left is the real spectrum of SDSS J1110 + 2808 (one of BELLS GALLERY samples), the right is an artificial one with a synthetical Ly α line. They look very similar. Another point need to be mentioned is the Gaussian profile approximations for creating Ly α lines. As we will show in Section 4, the Ly α lines are in fact skewed. However, in this study, we just want to use ResNet to recognize the emission lines with any shapes. The shape judgement will be done by some other subsequent selections. Therefore, using skewed profiles to create Ly α lines is not needed. During the experimental process, we can efficiently find the emission lines with any shapes using Gaussian profile approximations.

3.3 Our ResNet model

Our ResNet model is implemented using the Keras with Tensorflow as its backend. Our model has 8 blocks, 28 convolutional layers, much deeper than plain CNNs. We add a 'max pooling' layer between any two blocks. A Dropout layer is employed before the fully connected layer to prevent overfitting. As mentioned before, the samples with label [1] are far less than the samples with label [0], making the data imbalance. Here, we also add another measurement – 'Focal Loss' – to overcome the effect of the data imbalance. The 'Focal Loss' can be written as

$$L_{fl} = \begin{cases} -(1-\hat{y})^y \log \hat{y} & \text{when } y = 1 \\ -\hat{y}^y \log(1-\hat{y}) & \text{when } y = 0, \end{cases} \quad (1)$$

where $y \in \{0, 1\}$ is the real label and $\hat{y}$ is the prediction value. The 'Focal Loss' is designed to improve the classification task with extreme data imbalance (Lin et al. 2017).

We then use the train/test data (train data/test data = 9: 1) to train our ResNet model. In the training process, we mask the location of [Ne V] 3347, [Ne V] 3427, [O II] 3727, and [Ne III] 3869 lines which may come from the ETGs themselves. The output of our model for each spectrum is the probability to be a lens candidate. We treat the samples with probabilities greater than 0.5 as lens candidates. Fig. 6 shows the accuracy and loss of the train/test data. We obtain a near-perfect test accuracy at 0.995 4, the corresponding loss is 0.002 8. Also, the completeness is 93.6 per cent. After successfully training, we apply the model to our predictive data (the emission lines of the ETGs also have been masked). We find all the 21 BELLS GALLERY lens candidates with the possibilities extremely close to 1. For all the 174 known lens candidates in our predictive spectra, we find 161. The discovery rate is 92.5 per cent. We do not call the discovery rate as completeness because the total number of the lens candidates in our predictive data are unknown.
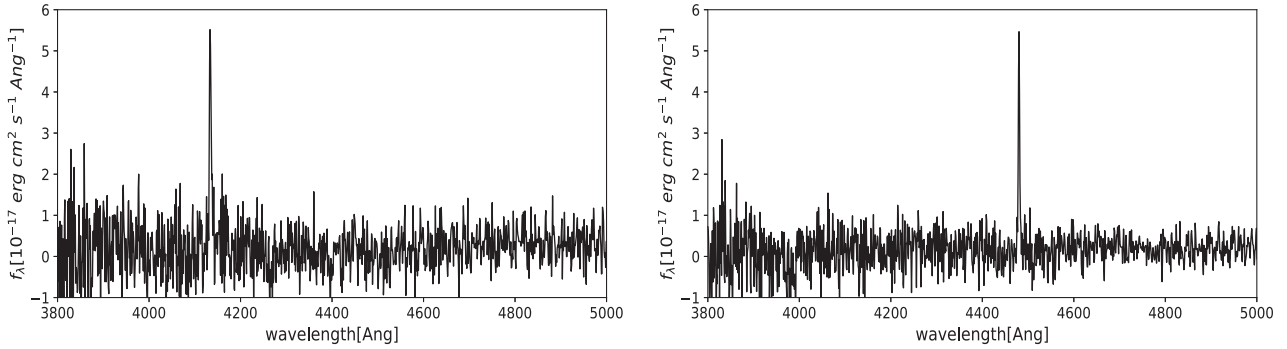
Figure 5. The left is the spectrum of SDSSJ 1110 + 2808 from BELLS GALLERY. The right is an artificial one. Both have high SNR Ly α lines.
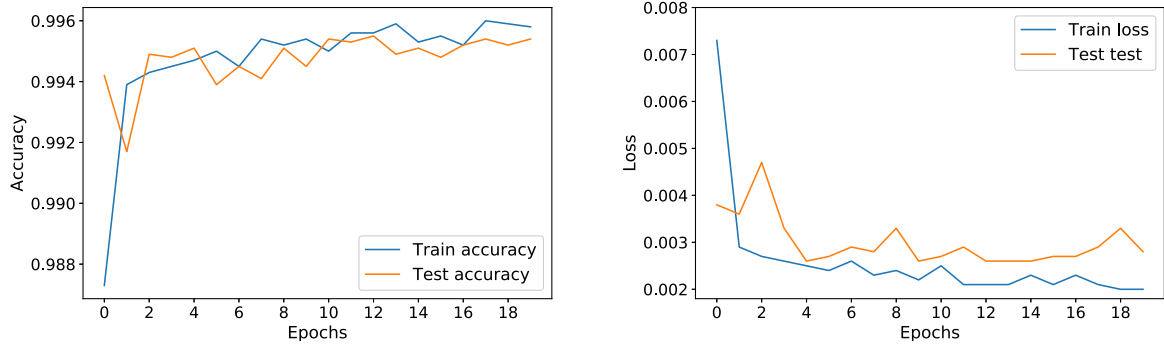

Figure 6. The left is the train/test accuracy and the right is the corresponding loss.

4 NEW STRONG GRAVITATIONAL LENS CANDIDATES

We find 1232 hits using our ResNet model, including 161 candidates found in Shu et al. (2016a). Apart from the existed hits of the 4982 hits in Shu et al. (2016a), we find 536 new hits. Next, we perform several subsequent selections used by Shu et al. (2016a) to remove the false hits. First of all, in order to ensure the high quality, we apply a visual inspection to the 536 hits and reject the samples with larger noise or emission lines of low peak flux (the peak is about less than $1 \times 10^{-17} \, \mathrm{erg \, cm^2 \, \mathring{A}^{-1}}$). For the left 124 hits, we remove the hits with significant numerical overdensities in both observed wavelength (associate with airglow features, see it in Shu et al. 2016a) and BOSS target-galaxy rest wavelength (associate with template-subtraction residuals). This step removes 77 spurious hits and left 47. One typical feature of high-redshift Ly α lines is the 'blue edge, red tail' profiles (see BELLS GALLERY samples). Therefore, for each spectrum, we use a skew normal profile plus a horizontal line to fit the emission line and the continuum. This fitting is achieved by non-linear least square method. We use four parameters (the Skewness, Mean Value, Standard Deviation, and Kurtosis) to describe the skew normal profiles and 1 parameter to describe the horizontal line. All these five parameters are varied during the fitting. The errors of the spectra are provided by SDSS. After this selection, 21 samples with skewness parameters bigger than 0 are left. We then remove 16 samples with low-redshift [O II] emission lines. In this step, for each sample, we suppose the line is [O II] and then calculate the relative positions of [O III], H α, and H β lines. If any of these positions has an emission line, we reject the sample. Finally, we obtain five galaxy-LAE lens candidates. The properties of these five candidates are shown in Table 1. We

also show the emission profiles and their best-fitting skew normal profiles in Fig. 7.

5 DISCUSSION AND IMPROVEMENT

Comparing with Shu et al. (2016a), we find an similar number of galaxy-LAE lens candidates (161 + 5) in the same database (DR12 of BOSS) using ResNet. The recognition speed of our strategy is much higher than that of theirs. After successfully training the model, we can finish the recognition process within just 1.5 h using the central processing unit (CPU) of a common laptop. By contrast, the method used in Shu et al. (2016a) needs about 10 h to finish a similar recognition. A graphics processing unit (GPU) with lots of Stream Processors can efficiently accelerate the convolutional operations. Although we have not tried the GPU computation, we would expect that the recognition speed could be even faster. This speed advantage would become important in the coming big data age of astronomy. We note that our recognition process does not include the training process. Although we have to spend several hours to train the networks in the beginning, our strategy is still a high efficient one. Once the model has been successfully trained, it can be used to any other predictive data without retraining.

This work is the first attempt to find lens candidates with ResNet based on spectroscopic selection. Our strategy still has a large improvement space. The first problem is the noise. During the repeating experiments, we find the spectra with too large noise could decrease the accuracy. The second problem is the subsequent selections. Our ResNet can successfully find the spectra with emission lines from different redshifts, but it can not judge whether they are Ly α lines from higher redshift or [O II] lines from lower redshift.

Table 1. Selected properties of the five galaxy-LAE lens candidate systems.

Target	Plate–MJD–Fibre	z_L	z_s	R.A.	Decl.	m_i	Ly α Flux	Skewness
SDSS J111749.50+014036.9	4731-55656-197	0.5840	2.1570	11:17:49.50	+01:40:36.95	19.35	30.4	2.17 ± 0.70
SDSS J121650.13+500700.2	6671-56388-589	0.6613	2.7894	12:16:50.13	+50:07:00.26	19.83	36.7	5.45 ± 1.60
SDSS J122502.89−000907.8	3847-55588-309	0.4868	2.4395	12:25:02.89	−00:09:07.85	19.61	12.3	4.03 ± 1.83
SDSS J142310.55+231928.1	6013-56074-66	0.4717	2.4099	14:23:10.55	+23:19:28.17	19.66	50.8	1.72 ± 0.43
SDSS J143428.11+470111.5	6736-56366-505	0.5070	2.3913	14:34:28.11	+47:01:11.53	19.25	24.2	2.44 ± 1.43

Note. Column 1 is the SDSS system name in terms of truncated J2000 R.A and Decl. in the format HHMMSS.ss± DDMMSS.s. Column 2 provides the plate–MJD–fibre of the spectra. Columns 3 and 4 are the redshifts of the foreground lenses and the background galaxies inferred from the BOSS spectra. Columns 5 and 6 are the coordinates in terms of truncated J2000 R.A and Decl. Column 7 is the BOSS-measured *i*-band apparent cmodel magnitude within the $1''$ fibre. Column 8 is the total apparent flux of the Ly α emission lines in units of 10^{-17} erg cm^{-2} s^{-1}. Column 9 is the skewness of the emission line profiles.
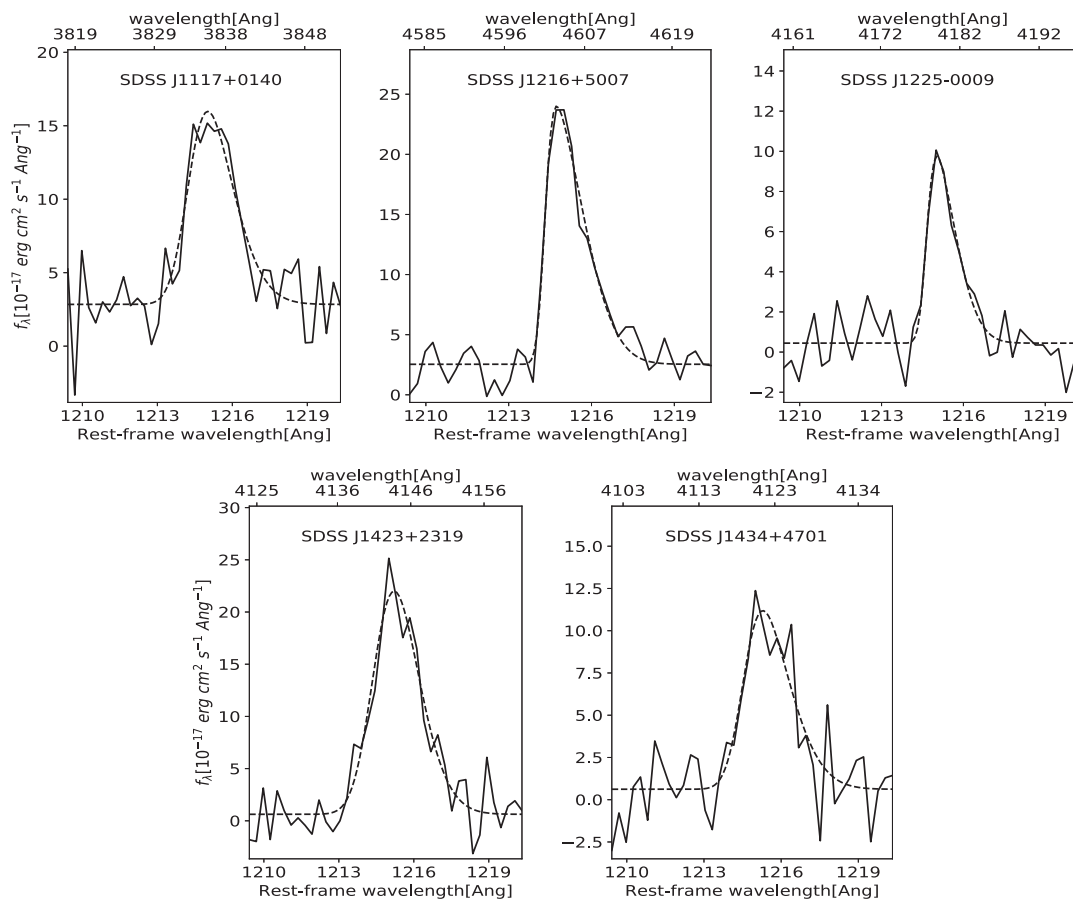

Figure 7. Zoomed-in views of the Ly α emission lines for the candidates ordered by decreasing apparent Ly α flux, de-redshifted to the rest frame of the LAEs. Dashed lines are the best-fitting continuum flux for the Ly α emission lines using skew normal distribution. Note the line profiles with 'blue edge' and 'red tail' are the characteristic of LAEs.

Therefore, after the ResNet selection, several subsequent selections are needed. If we want to use ResNet to directly find the emission lines with skewness bigger than 0, building the lines with skewed profile and classifying the train data to three classes is an approach, the spectra with no emission lines, the spectra with negative skewness lines and the spectra with positive skewness lines. However, this approach still needs subsequent selections and seems unnecessary in this work. An alternative method is using a regression ResNet model to give the redshifts of the source galaxies. This method requires the whole spectrum. We hope the regression model could learn the redshift of the source through the relative position of

multiple emission lines. However, to overcome the data imbalance, a regression model is more difficult than a classification model.

6 CONCLUSIONS

In this work, we apply the ResNet to search for galaxy-LAE lens candidates based on spectroscopic selection. We use 140 000 spectra without Ly α lines and 10 000 spectra with Ly α lines to train a 28 layers ResNet model and get a near-perfect test accuracy of 0.995 4 with the corresponding loss of 0.002 8. After apply

our ResNet model to the 1 241 799 predictive spectra from the DR12 of the BOSS of SDSS-III, we find 1232 hits. For all the 174 known candidates in our predictive spectra, we find 161 of them (92.5 per cent). Finally, we perform several subsequent selections to 536 new hits and find 5 new galaxy-LAE lens candidates. The most obvious advantage of our strategy is the recognition speed. Comparing with previous work (Shu et al. 2016a), we find a similar number of lens candidates in the same database with much less time. This work is the fist attempt to find lens candidates using Machine Learning. We will continue to improve this strategy and hope it could become better.

ACKNOWLEDGEMENTS

REFERENCES

Auger M. W., Treu T., Bolton A. S., Gavazzi R., Koopmans L. V. E., Marshall P. J., Bundy K., Moustakas L. A., 2009, ApJ, 705, 1099
Auger M. W., Treu T., Bolton A. S., Gavazzi R., Koopmans L. V. E., Marshall P. J., Moustakas L. A., Burles S., 2010, ApJ, 724, 511
Bolton A. S., Burles S., Koopmans L. V. E., Treu T., Gavazzi R., Moustakas L. A., Wayth R., Schlegel D. J., 2008, ApJ, 682, 964
Bolton A. S., Burles S., Koopmans L. V. E., Treu T., Moustakas L. A., 2006, ApJ, 638, 703
Bolton A. S., Burles S., Schlegel D. J., Eisenstein D. J., Brinkmann J., 2004, AJ, 127, 1860
Bolton A. S. et al., 2012, ApJ, 757, 82
Chen Y.-M. et al., 2012, MNRAS, 421, 314
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Gavazzi R., Marshall P. J., Treu T., Sonnenfeld A., 2014, ApJ, 785, 144
Gulcehre C., Moczulski M., Denil M., Bengio Y., 2016, preprint (arXiv: 1603.00391)
He K., Zhang X., Ren S., Sun J., 2015, preprint (arXiv:1512.03385)
Horesh A., Ofek E. O., Maoz D., Bartelmann M., Meneghetti M., Rix H.-W., 2005, ApJ, 633, 768
Hoyle B., 2016, A&C, 16, 34
Hála P., 2014, preprint (arXiv:1412.8341)
Impellizzeri C. M. V., McKean J. P., Castangia P., Roy A. L., Henkel C., Brunthaler A., Wucknitz O., 2008, Nature, 456, 927
Joseph R. et al., 2014, A&A, 566, A63
Kim E. J., Brunner R. J., 2017, MNRAS, 464, 4463
Koopmans L. V. E., Treu T., Bolton A. S., Burles S., Moustakas L. A., 2006, ApJ, 649, 599
Krizhevsky A., Sutskever I., Hinton G., 2012, in Pereira F., Burgs C. J. C., Bottou L., Weinberger K. Q., eds, ImageNet classification with deep convolutional neural networks, NIPS 25, p. 1097
Kubo J. M., Dell'Antonio I. P., 2008, MNRAS, 385, 918
Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, MNRAS, 473, 3895
Lecun Y., Bottou L., Bengio Y., Haffner P., 1998, Proc. IEEE, 86, 2278
Lenzen F., Schindler S., Scherzer O., 2004, A&A, 416, 391
Lin T.-Y., Goyal P., Girshick R., He K., Dollár P., 2017, preprint (arXiv: 1708.02002)
Li R.et al., 2018, MNRAS, 480, 431
Meyer R. A., Delubac T., Kneib J.-P., Courbin F., 2017, preprint (arXiv: 1711.01184)
More A., Cabanac R., More S., Alard C., Limousin M., Kneib J.-P., Gavazzi R., Motta V., 2012, ApJ, 749, 38
Petrillo C. E. et al., 2017, MNRAS, 472, 1129
Samue A. L., 1959, IBM J. Res. Dev., 3, 3
Seidel G., Bartelmann M., 2007, A&A, 472, 341
Shu Y., Bolton A. S., Schlegel D. J., Dawson K. S., Wake D. A., Brownstein J. R., Brinkmann J., Weaver B. A., 2012, AJ, 143, 90
Shu Y. et al., 2015, ApJ, 803, 71
Shu Y. et al., 2016a, ApJ, 824, 86
Shu Y. et al., 2016b, ApJ, 833, 264
Sonnenfeld A., Treu T., Gavazzi R., Suyu S. H., Marshall P. J., Auger M. W., Nipoti C., 2013, ApJ, 777, 98
Suyu S. H. et al., 2013, ApJ, 766, 70
Swinbank A. M. et al., 2009, MNRAS, 400, 1121
Treu T., Dutton A. A., Auger M. W., Marshall P. J., Bolton A. S., Brewer B. J., Koo D. C., Koopmans L. V. E., 2011, MNRAS, 417, 1601
Treu T. et al., 2015, ApJ, 812, 114
Warren S. J., Hewett P. C., Lewis G. F., Møller P., Iovino A., Shaver P. A., 1996, MNRAS, 278, 139
Xavier G., Antoine B., Bengio Y., 2011, in Geoffrey G. G., David B. D., Miroslav D., eds, Sparse Rectifier Neural Networks. AISTATS 15, p. 315

This paper has been typeset from a TEX/LATEX file prepared by the author.