Generalization of Convolutional Neural Networks for Searching for O-Star Spectra Using Generative Adversarial Networks

Zipeng Zheng and Bo Qiu*
School of Hebei University of Technology, Tianjin, China

*Corresponding author email: qiubo@hebut.edu.cn

Abstract. In the stellar spectral data released by LAMSOT, the O-star spectrum is very rare, and the total amount of O-star spectra that can be utilized is only 156. We recommend generating a simulated real spectrum to overcome the above limitations. Using the real O-star spectrum as the model spectral image, we propose a one-dimensional spectral generation confrontation network (1D SGAN) to create artificial spectra based on real data sets. We use a combination of real and artificial spectra to train a one-dimensional convolutional neural network (1D CNN) to create a classifier that classifies the stellar spectra into seven categories. We demonstrate that using the proposed balanced data set with 1D SGAN generated images improves the performance of the 1D CNN classifier compared to the same 1D CNN trained with only the original data set.

1. Introduction
The DR5 catalog released by LAMOST contains more than 8 million spectra, of which more than 2 million (including the Unknown type) are unable to measure the spectra of their parameters due to low signal-to-noise (SNR) ratio. Obviously this situation makes LAMOST’s output lower than expected in some sense. Once these low SNR spectra can be enhanced and resolved, and finally fully utilized, it is equivalent to increasing the effective output of LAMOST and increasing the effective observation range of the telescope, which is of great significance.

Among all the spectra released by LAMOST, the total number of O-star spectra with high signal-to-noise ratio and available for utilization is $156^{[6-8]}$. Compared with other types of stars, the number of o-stars is smaller. At present, scientists attribute the reasons to the following two reasons: First, the surface temperature of the O-star is higher, and the O-star is heavier. The internal reaction rate will make the gravitational field unable to be restrained, which makes the O-shaped star unable to exist stably. Second, the metal abundance of the O-type star in the Milky Way is relatively high, which makes the stellar Wind stronger and blows off the circumstellar disk. These two reasons make the number of O-stars rare relative to other types of stars$^{[1]}$.

The traditional template matching method cannot effectively classify all spectra due to the inherent low signal-to-noise ratio problem of one-dimensional spectra. This project innovatively applies the deep learning method to the processing of spectral data$^{[2]}$. It is precisely because of the hugeness of spectral data that it provides favorable conditions for the application of deep learning, which not only saves a lot of manpower input, but also greatly improves spectral classification. The accuracy of the astronomical data processing has expanded into a new world$^{[3]}$.

We propose a means of simulating the generated spectrum as a means of controlling the increase of data. The simulated data can be used to increase the number of available images, thus providing a
means of balancing the data set for neural network training. An ideal data simulation scheme can generate any number of composite images, simulate the characteristics of real images in any given class, and be sufficiently diverse to successfully train the network. In this paper, we propose to use the Generated Confrontation Network (GAN) to generate the O-star spectrum and modify the traditional two-dimensional generation confrontation network to one-dimensional generation against the network to adapt to the generation of the spectrum, thereby increasing the number of O-star spectra[9]. These composite images are used to balance and enhance the stellar spectral dataset to train convolutional neural networks (CNN) in stellar spectral classification[5].

2. Proposed Method

Based on the traditional two-dimensional GAN, we propose a 1D SGAN suitable for generating spectra, as shown in Figure 1. The resulting spectrum is combined with the actual spectrum to balance and extend the training data set to train the proposed CNN. And based on the traditional two-dimensional CNN, a 1D CNN suitable for one-dimensional spectral data is proposed, and its structure is shown in Figure 2, which is used to perform spectral classification work. These models are discussed in more detail below.

2.1. Generating O-star Spectra

The G generator and D discriminator form a 1D SGAN. The two neural networks compete with each other on the existing training data to improve the spectrum quality of the generated signal. 1D SGAN uses CNNs to generate O-star spectra for the G and D components of the model[4].

We first train generator G to model the potential probability distribution p_g of the output O-star spectra. By creating an artificial mapping G(Z, θ_g), we can input the noise variable p_Z(z) C is mapped to real spectral data. As shown in Fig.1, a 900-dimensional noise vector z is projected onto a spatially expanded convolutional representation with 1,024 feature maps. Due to the characteristics of the spectrum are extremely subtle, high signal-to-noise ratio data is very important for machine learning in spectral generation. Therefore, the projected and reshaped noise vector z is converted into O-Star spectra of 1×3700 pixels by three convolutional layers.

2.2. Classification of Stellar Spectra

After a certain number of iterative trainings, 1D SGAN can stably generate O-star spectra that can be utilized. We input the generated spectrum and the real spectrum into the 1D CNN to train a seven-classifier of the stellar spectrum.

The CNN used in this paper is based on lenet architecture. The spectral data is first calculated through convolution layer, then maximized pooling, and finally output through full connection layer. In this paper, we adjust the size of convolution kernel, the number of characteristic graphs and the number of network layers to adapt to the classification of one-dimensional spectral data.

| Layer | Type          | Maps | Kernel Size | Stride | Padding | Size    | Activation |
|-------|---------------|------|-------------|--------|---------|---------|------------|
| OUT   | -             | -    | -           | -      | -       | 1×7     | Softmax    |
| F9    | Fully Connected | -    | -           | -      | -       | 1×11500 | Relu       |
| S8    | Max-Pooling   | 64   | 1 ×2        | 1      | SAME    | 1×230   | -          |
| C7    | Convolution   | 64   | 1×2         | 1      | SAME    | 1×460   | Relu       |
| S6    | Max-Pooling   | 32   | 1 ×2        | 2      | SAME    | 1×461   | -          |
| C5    | Convolution   | 32   | 1×3         | 1      | SAME    | 1×922   | Relu       |
| S4    | Max-Pooling   | 16   | 1 ×2        | 1      | SAME    | 1×924   | -          |
| C3    | Convolution   | 16   | 1×2         | 1      | SAME    | 1×1848  | Relu       |
| S2    | Max-Pooling   | 8    | 1 ×2        | 1      | SAME    | 1×1849  | -          |
| C1    | Convolution   | 8    | 1×3         | 1      | SAME    | 1×3698  | Relu       |
| IN    | -             | 1    | -           | -      | -       | 1×3700  | -          |
3. Experiments
This section mainly shows the details of the implementation and the results we obtained. The experimental flow chart is shown in Figure 1.

3.1. Pretreatment
In spectral classification, the main feature we use is the continuum of the spectrum and its emission and absorption lines. The magnitude of the flow value is not the primary reference. Due to the different brightness of the stars, the observation conditions such as the weather conditions on the day of shooting and the atmospheric seeing will cause significant changes in the spectral flow value, and the inevitable noise factor should also be considered. In order to reduce the impact of these factors, we must first pre-process the spectrum, mainly including the elimination of abnormal points and the normalization of data.

We consider that in the UNKNOWN data released by LAMOST, the signal-to-noise ratio is relatively low, and there are abnormal points in many places. The existence of these abnormal points has a great impact on our work. Here, the abnormal points are removed by curve smoothing, and then normalized, which makes the network converge more quickly.

3.2. Divide the UNKNOWN Spectrum into 3 Categories
In Experimental parameters: The selected data is the spectral data of DR4 released by LAMOST. Among them, Galaxy, QSO and Star each take 3,000 samples as training samples, each take 500 samples as test samples, and the UNKNOWN type data of DR1-DR6 is selected as the data to be mined. The minimum batch size for training is 500, the number of training iterations is 1000, and the learning rate is $1 \times 10^{-4}$.

Experiment process:
(1) Three types of data training convolutional neural networks are loaded. After the training is completed, the classification accuracy rate is shown in Figure 2. In the end, the classification accuracy rate of Galaxy is 93.4%, the classification accuracy of QSO is 96.8%, and the classification accuracy rate of Star is 98.4%.
Figure 2. The accuracy curves of dividing into 3 types.

(2) After the training of the convolutional neural network is completed, the classification model is saved.

(3) The classification accuracy of the model meets the actual requirements, and the classification model completed by the training is loaded, and then the UNKNOWN spectrum is classified.

3.3. Divide the Star Spectrum of the Mining into 7 Categories

In Experimental parameters: The selected data is the spectral data of DR4 released by LAMOST. Among them, O, B, A, F, G, K, and M each take 1,000 samples as training samples, each take 400 samples as test samples. The minimum batch size for training is 500, the number of training iterations is 1500, and the learning rate is 1×10^{-4}.

Experiment process:

(1) Load seven types of data training convolutional neural networks. After the training is completed, the classification accuracy rate is shown in Figure 3, and the classification accuracy of each type is shown in Table 2.

Figure 3. The accuracy curves of dividing into 7 types.

Table 2. Classification accuracy of 7 types of stars.

| Star Type | O    | B    | A    | F    | G    | K    | M    | Average |
|-----------|------|------|------|------|------|------|------|---------|
| Accuracy  rate | 88.5% | 93.3% | 97.3% | 95.8% | 96.0% | 97.5% | 98.8% | 95.3%   |

(2) After the training of the convolutional neural network is completed, the classification model is saved.

(3) The classification accuracy of the model meets the actual requirements, and the classification model completed by the training is loaded, and then the Star spectrum classified in 3.2 is classified into seven categories.

4. Verification through Three Distance Measurement

In The template matching method is one of the commonly used methods in the automatic classification of stellar spectral. In order to verify the accuracy of the classification results, three distance metrics
between the template spectra and classified spectra are used to verify accuracy of results. They are the Euclidean distance, Cosine distance and Residual distribution metric. Further, in order to more intuitively observe the similarity distribution of the found spectra, 4,000 samples of B-M types were downloaded that have been classified by LAMOST DR6, and made the same similarity measure to prove the accuracy of the classification of the "UNKNOWN" type spectrum.

(1) Euclidean distance The spectral data is assumed to be a vector of n-dimensional European space, each of which can be represented as \( x(x_1, x_2, \ldots, x_n) \), where \( x_i (i = 1, 2, \ldots, n) \) is the spectrum \( x \) the i-dimensional coordinates. The Euclidean distance \( d(x, y) \) between the two spectra is defined as shown in equation (9).

\[
d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

The smaller the Euclidean distance between the two spectra, the two spectra are more similar.

(2) Cosine distance the cosine distance of the two spectra is calculated as follows

\[
d(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}
\]

This distance defines the magnitude of the cosine of the two spectral angles in n-dimensional space. When \( d = 1 \), the two spectra are completely consistent, and the smaller the value, the larger the distance.

(3) Residual distribution distance the residual distribution distance applied in this paper is as follows: First, calculate the residual spectrum \( X_{\eta_i} \) of the two spectra

\[
X_{\eta_i} = \frac{x_i - y_i}{x_i}
\]

Secondly, calculate the standard deviation of \( X_{\eta_i} \), take \( S \) as the residual distribution distance. The smaller the residual distribution distance, the high the similarity of the two.

The Euclidean distance is taken as the abscissa and the residual distribution distance as the ordinate, and then draw the scatter plot of “UNKNOWN” O-type spectra of known O-type spectra and classification as shown in Figure4. The distribution of find_O and real_O is almost the same. This can illustrate the credibility of the spectra found.

5. Conclusion

To solve the problem of imbalance in the number of different kinds of spectral data sets, we first proposed the use of Gan to simulate the generation of spectral data, so as to expand the number of fewer kinds of spectra. The use of balanced data sets can significantly improve the classification performance of CNN, which shows that the neural network trained by antagonism can better complete the generation of spectra. The completion of this work is helpful to the search for rare spectra provides astronomers with more valuable spectra.

We manually verify O-star spectra found, expanded the existing O-star spectral dataset, this confirms the feasibility of the method. There is still room for improvement in this method. How to effectively extend the method to the search work of other rare spectra, how to scientifically and effectively evaluate the quality of the generated spectrum remains to be further studied.

In the future, the spectral diversity will be analyzed and the quantitative standard will be established to guide the generation of the spectrum by the generation of the countermeasure network. The difference between spectral data and image data is that its type division is continuous. The work will measure the spectral parameters, convert the discrete class variables into continuous variables, obtain the information contained in the spectrum more accurately and interpret it better.
Acknowledgments
Supported by the National Natural Science Foundation of China, the Joint Research Fund in Astronomy (Grant No. U1931134)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{scatter_diagram.png}
\caption{Scatter diagram of similarity measure of seven kinds of stellar spectra.}
\end{figure}

References
\begin{enumerate}
\item Conti P. S., Garmany C. D., De Loore C., Vanbeveren D., 1983, ApJ, 274, 302
\item Diaz-Hernández R., Peregrina-Barreto H., Altamirano-Robles L., González-Bernal J. A., Ortiz-Esquível A. E., 2014, Exp. Astron., 38, 193
\item Fabbro S., Venn K. A., O"Briain T., Bialek S., Kielty C. L., Jahandar F., Monty S., 2018, MNRAS, 475, 2978
\item Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, Adv. Nat. Inf. Proc. Syst., 3, 2672
\item Kim E. J., Brunner R J., 2016, MNRAS, preprint (abs:1608.04369)
\item Liu C. et al., 2015, RAA, 15, 1137
\item Liu Z., Song L., Zhao W., 2016, MNRAS, 455, 4289
\item Liu Z. B., Song W. A., Zhang J., Zhao W. J., 2017, J. A&A, 38
\item Rui W., et al., 2018, MNRAS, 485, 2167
\end{enumerate}