Common Python Data Analysis Method Based on Deep Learning

Guilian Feng*

School of Physics and Electronic Information Engineering, Qinghai Nationalities University, Xining 810000, Qinghai, China

* Corresponding author e-mail: fengguilian@qhmu.edu.cn

Abstract. In recent years, with the continuous development and progress of machine learning technology, especially the successful application of deep learning algorithm in the field of image and big data processing, how to efficiently develop and apply machine learning technology has become a hot issue. This paper focuses on deep learning-based common Python data analysis methods. In this paper, a traffic flow prediction and analysis model based on convolutional neural network is established according to the traffic flow characteristics of the first ring expressway in this city. According to the characteristics of urban expressway traffic flow and the spatial correlation of traffic flow, the traffic flow data in the spatial and temporal dimension are combined in the form of two-dimensional matrix as the network input, and the CAFRE framework based on deep learning is used to design the expressway traffic flow prediction model based on convolutional neural network. The model takes into account the historical traffic flow of the predicted sections and the traffic flow of the upstream and downstream sections, and restricts the model structure by capturing the road range with high spatial correlation.

Keywords: Deep Learning, Python Language, Common Data, Data Analysis

1. Introduction
With the increase of information in social activities, the amount of data that people need to deal with every day is doubling, and the meaning and coverage of data are getting richer and richer. With so much information coming from the real world pushing the limits of the human brain, and the need to make use of it and create value, there is an urgent need for technology to improve the readability of information. In the field of Internet, although more powerful hardware facilities can be constantly updated and maintained to cope with the large amount of data generated by customers, existing software models still need to be continuously optimized and improved to meet customers’ demands more efficiently [1]. In the field of the Internet of Things, a large number of high-precision sensors gradually cover every corner of the real world, collect a large amount of information in a more real-time and comprehensive way, and depict the real environment through rich metadata. These data and information often involve different fields and are abstract and obscure, which makes it difficult for even experts to interpret the data quickly. People hope to interpret the data in a more intuitive way and
quickly establish an appropriate data model to perceive the real world [2]. In the financial field, market information rapidly changing, in order to capture the change rule of financial markets, the quantitative data analysts need to analyze a large number of market transactions, and in the process of trial and error to find the effective law and establish the financial model, but due to reasons such as policy adjustment and market volatility, the financial model of effective life cycles are getting shorter, need to constantly adjust and modify the model to adapt to the new economic laws, instant screen the optimal strategy. Therefore, there is an urgent need for an intuitive and visual way to combine policy logic to simplify the process of modeling, debugging, and validation.

Machine learning is an important research direction in the field of artificial intelligence. Deep learning is the second wave of machine learning technology. In addition to the development of academia, domestic and foreign IT giants have also started to get involved in this field, launching a large number of development frameworks and platforms for deep learning modeling and application. These deep learning frameworks inherit the traditional modeling and development approaches in the field of machine learning. Text programming language is used to describe algorithm logic and generate model through designing special API interface or defining DSL (domain-specific language) language [3]. You can use tools such as Netscope, Tensorboard to provide a wireframe of the internal structure or a pixel diagram of the parameters for the generated model. Or using the Microsoft platform for the machine learning (AzureMachineLearningStudio), based on the algorithm of precast assembling data processing module and data pipeline [4].

Based on the characteristics of spatiotemporal data, this paper proposes a set of traffic demand forecasting framework based on deep learning, and verifies its advantages in terms of forecasting accuracy and efficiency. At the same time, the effects of different data preprocessing methods on the experimental results are compared, and reasonable suggestions for selection are given.

2. Data Analysis Based on Deep Learning

2.1 Overview of Convolutional Neural Networks

Convolutional neural network is one of the most representative neural networks in deep learning at present, and it is a deep feed forward neural network with excellent performance in image processing [5]. For the commonly used convolutional neural network structure, the convolutional layer is alternated with the pooling (subsampling) layer to continuously extract and compress features and gradually obtain high-level features. The full connection layer completes the nonlinear transformation and processing according to the high level features; the output layer is the classification and recognition result [6].

2.1.1 Convolution layer. Using convolution operation to extract and transform features is the basis of realizing convolutional neural network. By setting the convolution kernel (filter matrix, usually 3x3 or 5x5) of the specified size, "sliding" on the input image or the middle layer feature map according to the step size, feature extraction is carried out in small regions one by one, and the target feature map is obtained. The calculation formula is as follows:

$$X^l_k = f\left(\sum_{i \in M} (X_{i,l-1} * K^l_{i,k}) + b^l_k\right)$$

(1)

In general, the activation function is needed to process the data in the feature graph after convolution calculation. Currently commonly used activation functions include Sigmoid function, Tanh function, ReLU function, LReLU (Leaky ReLU) function, etc. [7].

The nonlinear function is introduced as the activation function so that the input of each layer node is no longer just a linear combination of the output of the upper layer. The addition of activation function can improve the nonlinear expression ability of the neural network model, and the output of each layer can approach any function to solve the linear indivisibility problems encountered by the neural network in classification and recognition [8].
2.1.2 Pooling layer. In order to effectively reduce the size of the feature matrix after convolution, features are further compressed and important information is retained. Feature subsampling is often performed between convolution, that is, pooling operation. Pooling pays more attention to the feature of region and unit block rather than single pixel, which is beneficial to keep the translation invariance of feature. Pooling can reduce the number of parameters and speed up the efficiency of calculation. At the same time, it is beneficial to prevent the occurrence of overfitting problem. The pooling calculation is shown below the formula.

\[ X^l_k = f[\beta^{l-1}_k \text{down}(X^{l-1}_k) + b^l_k] \] (2)

Pooling operation is relatively simple and can be divided into maximum pooling and average pooling according to different value methods. There are only two super parameters involved in pooling calculation: tree pool size and step size. In practical applications, maximum pooling tends to retain texture and contour features, while average pooling has advantages in retaining overall data features and background [9].

2.1.3 Full connection layer. When convolutional neural network is used for tasks such as classification and recognition, it is usually necessary to add a full connection layer as a "classifier" at the end of the network model to integrate the final output results. Each neuron in this layer is connected with all neurons in the previous layer, and each connection is equivalent to a weight operator to calculate the contribution of the upper layer neuron to the output. Each neuron in the full connection layer is calculated by a large number of connection weights, which also ensures its ability to measure the overall features of the image.

It is worth noting that the output of the last full connection layer corresponds to a special activation function, the Softmax activation function. This function can map multiple neuron outputs to the interval (0,1) in the form of probability, and facilitate the completion of multiple classification tasks.

2.1.4 Loss function. In deep learning, Loss Function (L) is the basis of back propagation in network training. The process of minimizing the loss function is also the process of network optimization and continuous improvement of model capability. In the convolutional neural network training, the output prediction result is usually compared with the label value or the input, and the corresponding expression is constructed for optimization according to the different image processing tasks. Commonly used simple loss functions include L1 (mean absolute error, MAE) loss function, L2 (mean square error, MSE) loss function, root mean square error (RMSE) loss function and cross entropy loss function, etc. [10].

2.2 Data Analysis Model Based on Convolutional Neural Network

2.2.1 Structural design. In this paper, Caffe, the general framework of convolutional neural networks, is used to train and build neural networks. Therefore, the structure of the network is based on the structure of Caffe framework. The Caffe framework is generally structured as layers, each layer representing a different meaning and computing process. In this section, a prediction model based on convolutional neural network is designed according to the general functions of Caffe framework and the demand of traffic flow prediction. Considering the problem of gradient diffusion and computational complexity, the convolutional neural network prediction model in this paper contains three convolutional layers.

Each convolution layer in the network is followed by a pooling layer, and the convolution layer and the pooling layer are correspondingly connected. After the third pooling layer, all the data output by the pooling layer are expanded into a feature vector to form a full connection layer. The correction function is used after the full connection layer to correct the data, and finally a full connection layer is used to output the data.
2.2.2 Parameter construction. The convolutional neural network designed in this paper uses three convolutional layers, namely Conv1, Conv2 and Conv3. After the convolutional calculation of each layer, the ReLU layer is used for correction, and then the sub-sampling layer is used for pooling calculation.

The three convolutional layers with the same structure are superimposed successively. In addition to the Conv1 layer, which uses a 5×5 convolutional kernel with an edge filling size of 2, the other two convolutional layers all use a 3×3 small convolutional kernel with an edge filling size of 1.

In the actual process of convolution calculation, in order to consider the depth control of the network and the correspondence of the transmission of feature graphs between layers of different structures, the feature graphs of adjacent two layers are not one-to-one in most cases.

The convolutional neural network designed in this paper uses the pooling layer to carry out the pooling calculation after the convolution calculation of each convolutional layer, and adopts the maximum pooling operation. The three pooling layers are POOLL, POOL2 and POOL3, and the pooling size of 2×2 is taken for each layer, and the step size is S=2. In this design, there is no overlapping area in the pooling process, and the maximum value in the pooling area is taken as the neuron output value of the pooling layer, which is in favor of keeping the traffic flow information on the expressway to the maximum extent.

After pooling the convolution result of each convolution layer, the size of the feature graph output by the convolution layer on the line and column is halved, and the data volume is reduced to 1/4 of the input data. For example, the size of output feature graph of convolutional layer Conv1 is 16×16. After using a 2×2 convolutional kernel with a step size of 2 and pooling calculation in the Pool layer, the size of output feature graph becomes 8×8.

According to the above hypothesis, when the convolutional layer depth of the first layer is designed to be 6 and the convolutional layer depth of the third layer is designed to be 16, the full-connection layer FC4 after the third pooling layer fully expands the pooling result of the previous layer POOL3 layer into a feature vector, and a feature vector based on 1500 classifications is designed to be obtained.

Considering the actual demand, in the traffic flow prediction model, aiming at the traffic flow 10min after the predicted section, this paper hopes that the output of the model is an actual value based on the predicted results, so the output feature vector of FC4 layer cannot be directly used. It is necessary to add a full-connection layer FC5 after FC4 layer to integrate the feature vectors of FC4 layer, and solve the value with the highest probability among the 1500 classification as one of the final output feature values.

The advantage of ReLU function is its fast computation speed, which can alleviate the problem of gradient disappearance to a certain extent. The significance of the existence of this layer is to prevent the gradient of the feature image from disappearing too quickly after the convolution calculation. Actual experiments have proved that the existence of ReLU layer is of positive significance to the improvement of model prediction accuracy.

3. Model Implementation and Training

3.1 Model Implementation
The model platform implemented in this paper is shown below
- Platform: Python3.5
- Operating system: Windows 10 64-bit
- Processor: Inter Core i7
- Memory: DDR4 16GB
- Graphics card: NVIDIA RTX2080TI
This paper uses Python3.5 platform to implement the designed model.

3.2 Model Training
Based on the traffic flow data of the whole section of the first ring expressway in the city in 20 days last month, the data are divided into working days and holidays. The traffic flow time interval of the original data is 10 min, and there are 144 time periods every day. One expressway is divided into 64 sections. The size and form of the input data were determined as $16 \times 16$ in the previous chapter, and the traffic flow data of all sections were traversed when the input value of the network was programmed.

The number of workday training data of the forecast model is 68481, and the number of forecast data is 15394. The amount of training data on weekends and holidays is 50,963, and the predicted amount of data is 8,314, which is relatively large.

Using Python 3.5 platform, the convolutional neural network structure is trained by writing scripts.

4. Model Analysis Results

4.1 Working Day Traffic Flow

Table 1. The predicted value of weekday traffic flow compared with the actual value

|       | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|-------|---|---|----|----|----|----|----|
| Actual value | 172 | 967 | 403 | 579 | 615 | 708 | 652 |
| CNN model    | 201 | 986 | 637 | 721 | 634 | 746 | 743 |

Figure 1. The predicted value of weekday traffic flow compared with the actual value

As shown in Table 1 and Figure 1, the traffic flow prediction model based on convolutional neural network has a good prediction effect on the variation trend of weekday traffic flow, especially in the period when road traffic flow grows to the morning peak, and the variation trend of the predicted value is basically the same as the actual value. Peak but neural network model for traffic flow of the maximum and minimum value of low traffic flow prediction, traffic flow forecast will appear a certain error, prediction model of traffic flow data in detail to the actual traffic flow on the reduction of inaccurate, prediction of traffic flow time change a slightly hysteresis curve relative to the measured values.

4.2 Holiday Traffic
Innovation Basic

Acknowledgement

good show by matrix, calibrated, completed and flow algorithms the In

5. traffic always fluctuation actual forecast

Conclusions

As shown in Figure 2, when the convolutional neural network-based prediction model is used to forecast the traffic flow data in holidays, the predicted value of the model is more gentle than the actual value of the traffic flow. In general holidays, the road traffic flow is relatively large, and the fluctuation range of traffic flow is relatively small compared with weekdays, and the traffic flow is always maintained at a relatively saturated level. Similar to the weekday traffic flow prediction model, the holiday traffic flow prediction model has a good performance in reflecting the macro trend of traffic flow, but a little worse in reflecting the change of traffic flow details.

5. Conclusions

In recent years, with the large-scale popularization of big data analysis and processing technology and the rapid development of machine learning algorithms such as deep neural network, people use these algorithms and data processing technology to achieve a variety of applications. In this paper, a traffic flow sequence prediction model is established based on the theory of convolutional neural network, and the convolutional neural network is constructed layer by layer and fine-tuned parameters are completed with the help of Caffe framework. When the convolutional neural network model is calibrated, the spatio-temporal two-dimensional traffic flow data is considered as the unified input matrix, and the size and spatio-temporal span of the input data of the prediction model are determined by using the conclusions of the spatial correlation analysis of traffic flow. The experimental results show that the traffic flow prediction model based on convolutional neural network has a relatively good performance in the prediction results and can predict the change trend of general traffic flow.

Acknowledgement

Basic Research Projects in Qinghai Province “Research on the Development Strategy of the Innovation Cluster Construction in the Cultural Industry of Regong in Qinghai Province” (2020-ZJ-606)

References

[1] Yuanyuan, Guo, Shaoyan, et al. Creating the Learning Situation to Promote Student Deep Learning: Data Analysis and Application Case. AIP Conference Proceedings, 2017, 1839(1):1-5.

[2] Mishra B K, Barik R K, Priyadarshini R, et al. An Investigation Into the Efficacy of Deep Learning Tools for Big Data Analysis in Health Care. International Journal of Grid and High Performance Computing, 2018, 10(3):1-13.
[3] Ferdinand, Dhombres, Jean, et al. Formal Medical Knowledge Representation Supports Deep Learning Algorithms, Bioinformatics Pipelines, Genomics Data Analysis, and Big Data Processes. Yearbook of medical informatics, 2019, 28(1):152-155.

[4] Nykaza E T, Boedihardjo A P, Wang Z, et al. Deep learning for unsupervised feature extraction in audio signals: Monaural source separation. Journal of the Acoustical Society of America, 2016, 140(4):3424-3424.

[5] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading Text in the Wild with Convolutional Neural Networks. International Journal of Computer Vision, 2016, 116(1):1-20.

[6] Tajbakhsh N, Shin J Y, Gurudu S R, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?. IEEE Transactions on Medical Imaging, 2016, 35(5):1299-1312.

[7] Pereira S, Pinto A, Alves V, et al. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. IEEE Transactions on Medical Imaging, 2016, 35(5):1240-1251.

[8] Selim A, Elgharib M, Doyle L. Painting style transfer for head portraits using convolutional neural networks. Acm Transactions on Graphics, 2016, 35(4):1-18.

[9] Korfiatis P, Kline T L, Lachance D H, et al. Residual Deep Convolutional Neural Network Predicts MGMT Methylation Status. Journal of Digital Imaging, 2017, 30(5):622-628.

[10] Cai N, Su Z, Lin Z, et al. Blind inpainting using the fully convolutional neural network. Visual Computer International Journal of Computer Graphics, 2017, 33(2):249-261.