Automatic Document Classification Using Convolutional Neural Network

Xingping Sun¹,*, Yibing Li¹, Hongwei Kang² and Yong Shen²

¹ School of Software, Yunnan University, Kunming, China
² Key Laboratory for Software Engineering of Yunnan Province Republic of China, Yunnan University, Kunming, Yunnan, China

*Corresponding author e-mail: xpsun@mail.ynu.edu.cn

Abstract. Official document classification is an integral essential part of daily archives administration. The traditional manual document classification method is time-consuming and labour-intensive, and the classification effect cannot be fully guaranteed. With the popularity of computers and the development of machine learning, the convolutional neural network model is becoming more and more mature, and the CNN model is suitable for the problems encountered in the current official document classification. This paper proposed a model based on convolutional neural network to solve the problem of official document classification. To train and test the model, we manually classify the dataset into ten categories according to the classification of college archives entities. And it was trained on a dataset with size of 6765. And on the testing dataset with size of 676, it reached an accuracy of 90%. And for comparison, we also trained a LSTM model and a GRU model (they are both popular in the natural language processing field), and results showed that the automatic official document classification method based on convolutional neural network can improve the efficiency of traditional official document classification. Also, it provides a new way of thinking for the solution of the official document classification problem.

1. Introduction
In the current network environment, the management of official documents requires the archives management personnel to have both the knowledge and experience of secretarial and archival management, as well as the ability computer applications and practical skills. Commonly, older staff are lack of computer skills; while younger staff have relatively insufficient professional file management experience and standardized work ability, and the construction of the file management team in the network environment requires a breaking-in period [1].

Researches have shown that convolutional neural networks perform outstandingly in the natural language processing task [2]. Among them, a paper proposed a model to classify English text at character level [3] inspired us. It treats every letter and some certain symbols as ids in which way the text can be transformed into an id sequence. What we need to handle in our case is Chinese text, so we treated every character and symbol as id.

LSTM is getting more adherents in recent years. LSTM is an effective chain-link neural network widely used in NLP such as language model, machine language translation and speech recognition. Especially in the research of textual emotional analysis [4], LSTM has a great advantage in combining...
the characteristics of contextual semantics. The GRU model is an improvement of the LSTM model. Compared with LSTM, the GRU model is simpler, the calculation is faster, and the training effect is good.

2. Related work

2.1. Convolutional neural network

In 1962, biologists Hubel and Wiese studied the cat's brain visual cortex and inspired the generation and development of CNN. Convolutional neural network is a kind of feedforward neural network. The essence is to find a set of the most suitable information mapping relationship from input to output in a large parameter set by means of back propagation, that is, to train a suitable network to achieve classification, identification and other purposes. In the CNN model, the network extracts the local features of the input matrix by the movement of the convolution kernel, extracts the local features of the input information, and integrates the local information in a higher layer to obtain global information. And in this process, the features extracted by each convolution kernel are shared across the network. Of course, in order to obtain more different feature sets, the convolutional layer will have multiple convolution kernels to generate different features.

The convolutional neural network includes a convolutional layer and a pooling layer. The function of the convolutional layer is to locally perceive the input data, extract the main features, and filter by using multiple convolution kernels to distinguish the data features to the greatest extent. The role of the pooling layer is to compress the output of the convolutional layer. Under the premise of maintaining the main features of the data, the data is reduced, the dimension is reduced, the amount of calculation is reduced, and over-fitting is effectively avoided, thereby improving the generalization ability of the model. The commonly used pooling methods have the largest pooling and average pooling.

![Figure 1. Convolution and pooling.](image)

2.2. Recurrent neural network

The Recurrent Neural Network was first proposed by Goller [5]. The RNN is a neural network with memory feedback, which gives the corresponding output according to the input of the time series [6]. The RNN cycle process is to input the output of the network at time \( t \) to the network at time \( t+1 \), thereby affecting the final result. However, in practical applications, it is found that RNN will appear gradient disappearance when processing long sequence data [7], which makes the network forget the memory of longer time and only focuses on the content of the final stage of memory learning. In order to solve this problem, many new models have been proposed, including long and short memory neural networks and GRU neural networks.

2.2.1. Long short-term memory (LSTM)

The LSTM consists of an input layer, an implicit layer, and an output layer. Unlike a traditional neural network, the hidden layer includes a storage module that can store information for a long time. The memory cells in the memory module have self-joining, adaptive, and logic gate control flow information capabilities [8]. In the LSTM structure, there is a memory cell and three gate structures including input gates, forget gates, and output gates to control changes in the overall structure. Wherein, the input gate controls the data transfer to the memory cells.
at the moment, the forgetting gate controls the internal circulation of the memory cells and the
forgetting of the data, and the output gate determines the data transfer of the memory cells to other
parts of the network. This gating mechanism of LSTM is to allow information to pass selectively, so
that memory cells have the ability to preserve long-distance dependent information, and prevent
internal gradients from external interference during training. Each memory cell has a self-circulating
linear unit called a constant error conveyor, which allows the error to propagate internally at a constant
value, avoiding gradient disappearance and gradient explosion problems [9].

2.2.2. Gated Recurrent Unit (GRU). The Gated Recurrent Unit is an improvement from the LSTM
model. Although the LSTM neural network solves the long-range dependence defect of the traditional
RNN network, the structure is complex and the training time is long. The GRU neural network was
proposed to solve these problems. In the GRU network structure, the LSTM forgetting gate and the
input gate are integrated into an update gate, and the update gate is used to control the extent to which
the data at the previous moment is brought into the current state. The more data is brought in at the
moment. The reset gate is used to control the degree of ignoring the status information of the previous
moment. The smaller the value of the reset gate, the more data is ignored at the previous moment.

3. Data pre-processing
In the data preprocessing process, we first count the number of occurrences of each character and
symbol in the training data set and sort it, then select the top 5000 characters. And assign an ID to each
of them to form a vocabulary map.

Then we defined a max length every text can have. In our case, the max length is 300. For texts
whose length is longer than 300, we cut it from the top. And for shorter texts, we padded them with id
‘0’ at the end of them. (‘0’ means space in the vocabulary map). These fixed length texts are projected
into id sequences according to the vocabulary map we built in the first preprocessing step which
represents the data vector of every text.

4. Network structure
There are seven layers in our network, including embedded layer, convolution layer, pooling layer,
two-layer fully connected layer, and leakage layer and nonlinear mapping layer between two layers of
fully connected layers. The parameters of each layer are obtained by network training. The following
is a detailed description of each layer's structural functions and parameters.

4.1. Embedding layer
What the embedding layer does here is to express the characters into vectors that are closer to their
syntax. The matrix is updated during training. By observing the similarities among these vectors of
high-dimension, the relationships of vectors become more clearly. The embedding dimension is set to
64 in our model. The matrix inputted into this layer is of size n*300. By multiplying it with a matrix,
we can get the output matrix of size n*300*64. That is to say, every preprocessed vector is converted
into a higher-dimension matrix in this layer.

4.2. Convolutional layer
In the convolution layer, we convolved the input data with all the feature metrics respectively through
different trainable convolution kernel, and added bias, and then formed the output matrix, so as to
achieve the purpose of data feature extraction. The hyper-parameters that need to be set in this layer
are the convolution kernel size, the number of convolution kernels, and the step size. The size of the
convolution kernel determines the area of each local perceptual field. We set the kernel size to 5. The
number of convolution kernels determines how many feature maps we can get through the operation,
we set it to 256. The size of the step represents the distance the window slides after one convolute
operation. Here, we choose the default value.
4.3. Pooling layer
At the pooling layer, we use the maximum pooling method. We use the maximum pooling convolution kernel to perform the pooling operation on the matrix outputted by the former convolutional layer. That is, the maximum value of the feature points in the range of the maximum pooling convolution kernel is selected to form the output vector. This layer can avoid over-fitting effectively.

\[ Y = \text{max}_p \text{ool}(Z) \]  
\[ y_{r,i} = \max_{j=1}^{m} z_{r+(c-1)r+j,i} \]  

Figure 2. Max pooling

4.4. Fully connected Layer
Connected after the convolutional layer and the pooling layer, the purpose of the fully connected layer is to map the features learned by the network into the markup space of the sample. The fully connected layer converts the two-dimensional feature matrix of the convolution output into a vector. In our structure, we choose to connect two fully connected layers, and the number of neurons in the first fully connected layer is 256. The network can learn the relationship among the features extracted by convolution and pooling in the first fully connected layer. What the network obtains through the second fully connected layer is the probability of the input data belongs to the ten categories.

4.5. Dropout layer
Between the two layers of fully connected layers, we did dropout and nonlinear mapping processing to get better results. The processing of the dropout layer is another operation to avoid over-fitting on the network after the previous convolution and pooling. In this layer, given a certain probability, every neuron is dropped at this probability. The processing of dropout also reduces the interdependence of neurons.
4.6. Nonlinear mapping function
The purpose of adding a nonlinear mapping function in the network structure is to increase the nonlinear factors in the network, and transform the linear model we train into a network with more expressive capabilities. At this layer, we have chosen the Rectified Linear Units function. If the input value is negative, the output is 0, otherwise the original value is maintained.

\[ RELU(x) = \max(0, x) \]  

4.7. Classification layer
This second fully connected layer obtains the probability of the data belonging to the ten categories, and then judges the category of the data according to the probability.

5. Experiment

5.1. Environment
This experiment is done under the window operating system, the CPU uses intel core i7, the memory is 4G.

5.2. Data set
The goodness of the model is verified by classifying data in the domain of college official documents. According to the classification of archives entities in higher education institutions, college official documents are divided into accounting, product production and technology development, publishing, party group, administration, capital construction, teaching, scientific research, foreign affairs, equipment and equipment. The dataset includes 6,765 official documents, consisting of 762 accounting classes, 359 production and technology development categories, 230 publishing categories, 1012 party groups, 1038 administrative categories, 316 capital construction classes, 1118 teaching
classes, 619 scientific research classes, 1,121 foreign affairs classes, and 190 instruments and equipment.

5.3. Experimental results

As shown in the table as follow, we use the above network structure to verify on the dataset. For comparison, we replaced the convolutional layer and the pooling layer in the network structure with one LSTM layer and two layers of GRU layers respectively. The table shows the time these three models took and the accuracy they can reach on the testing dataset. The experimental results are shown in Table I.

Table 1. Comparison of time and accuracy.

| Model | Time(s) | Accuracy |
|-------|---------|----------|
| LSTM  | 2       | 0.77     |
| GRU   | 2       | 0.82     |
| CNN   | 1       | 0.90     |

From the experimental results, we can see that the CNN model has obvious advantages over the LSTM model and the GRU model both in accuracy and in time.

Table 2. Comparison of category accuracy.

| Model | Accounting | Publishing | Party group | Administration | Capital construction | Teaching | Scientific research | Foreign affairs | Equipment |
|-------|------------|------------|-------------|----------------|----------------------|----------|---------------------|----------------|-----------|
| LSTM  | 0.82       | 0.44       | 0.90        | 0.77           | 0.56                 | 0.98     | 0.58                | 0.89           | 0.86      |
| GRU   | 0.89       | 0.43       | 0.88        | 0.93           | 0.70                 | 0.82     | 0.68                | 0.88           | 1.00      |
| CNN   | 0.89       | 0.78       | 0.94        | 0.83           | 1.00                 | 0.98     | 0.87                | 0.93           | 1.00      |

Acknowledgments

This work is funded by the Open Foundation of Key Laboratory of Software Engineering of Yunnan Province under Grant No. 2015SE204. The dataset is provided by Kunming Medical University.

References

[1] Lei Wen. Problems in the management of official documents in colleges and universities and countermeasures [J]. File Management, 2013(05):41-42.
[2] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
[3] Goller C, Kuchler A. Learning task -dependent distributed representations by backpropagation through structure[C]/Neural Networks, 1996., IEEE International Conference on. IEEE, 1996, 1: 347 -352.
[4] Jun Liang,Yumei Chai,Huibin Yuan,Gao Minglei,Hongying Yan Emotional Analysis Based on Polar Transfer and LSTM Recursive Network[J] Chinese Journal of Information, 2015, 29(05): 152-159.
[5] Zhang X, Zhao J, Lecun Y. Character-level Convolutional Networks for Text Classification[J]. 2015:649-657.
[6] Huifeng Zhu. Speech recognition based on the fusion of DNN and RNN acoustic models[A]. China Information Society Professional Voice Information Committee. The 13th National Human-Machine Voice Communication Conference (NCMMSC2015) Proceedings [C]
[7] Hochreiter S, Bengio Y, Frasconi P, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies[J]. 2001.
[8] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to Diagnose with LSTM Recurrent Neural Networks," Computer Science, 2015.
[9] Fan Shu,Dan Qu,Wenlin Zhang,Lili zhou,Wu Guo.Low Resource Speech Recognition Method Based on Long and Short Time Memory Network[J].Journal of Xi'an Jiaotong University,2017,51(10):120-127.