Multi-classification application of Chinese news text based on deep learning

Siyu Xin*
Berkshire school, Sheffield, MA, USA

*Corresponding author e-mail: hzhe_wjdi@163.com

Abstract. As the core of Internet text processing and text mining, text categorization has become a key research issue in the field of natural language processing. The traditional methods mainly focus on shallow machine learning. With the rapid development of deep learning technology, its image recognition and speech A huge research breakthrough in the field of identification, the feature learning ability of the depth model is further proved. This paper is based on the THUCNews data set, combined with word embedding technology, using TextCNN and RNN to achieve multi-classification of text. The experimental results show that the accuracy of TextCNN on the test set reaches 96.04%, and the precision, recall and f1-score of all types exceed 0.9, and the accuracy of RNN on the test set reaches 94.22%, and various types of Precision, recall and f1-score, in addition to the home category, are over 0.9. Comparing the two models, it can be seen that RNN is not very satisfactory except for the classification in the home. Other categories are not much different from CNN, and further adjustment parameters can be used to achieve better results.

1. Introduction
Since the beginning of the 21st century, with the prosperous development of Internet industry, the big data era of rapid information transfer is coming. In mass data information, the expression of information is diversified, such as text, image, voice and so on. Among them, as a significant information carrying method, the text has the characteristics of less capacity occupancy and more convenient storage and management compared with image, voice and other carriers. In recent years, deep learning has made good achievements in the field of image and voice, and the application in the field of natural language processing has also become a mainstream direction. In the aspect of text classification technology, because the traditional text representation is highly dimensional and sparse, the feature expression ability is very weak. The traditional machine learning classification algorithm needs a lot of energy in feature engineering. Therefore, the solution of the natural language processing problem based on deep learning technology has become the mainstream research direction at present.

2. Basic Theory

2.1. Word Embedding
Word Embedding is an important concept in natural language processing (NLP). Conceptually, it refers that a high-dimensional space with the number of all words as a dimension into a much lower
dimensional continuous vector space, and each word or phrase is mapped to a vector in the real number field.

The methods of word embedding include artificial neural network [1], dimension reduction of word co-occurrence matrix [2][3][4], probability model [5], explicit expression of word context [6] and so on.

In the underlying input, the method of word embedding to express phrases greatly improves the effect of syntactic analyzer [7], text sentiment analysis and so on in NLP. Bengio uses a three-layer neural network to build a linguistic model as follows:

![Figure 1. Schematic diagram of linguistic model](image)

2.2. TextCNN Network Model

The TextCNN (Convolutional Neural Networks for Sentence Classification) can be seen as the representation of n-grams, and convolution kernels with the three feature sizes proposed by Yoon Kim can be considered to be 3-gram, 4-gram and 5-gram. The whole model structure is as follows. Firstly, the convolution kernels with different sizes (3, 4, 5) are used to extract the features and carry out maximum pooling, and the features extracted from convolution kernels with different sizes are spliced as the feature vectors input into softmax.

![Figure 2. Model architecture of sample sentence with two channels.](image)
2.3. Long Short-Term Memory
Long short-term memory (LSTM) is a time recurrent neural network, which is suitable for processing and predicting important events with relatively long interval and delay in time series. LSTM is proposed to solve the problem of "vanishing gradient" in the recurrent neural network (RNN), and it is a special recurrent neural network.

The internal structure of LSTM neurons is as follows:

![Figure 3. LSTM network model diagram.](image)

3. THUCNews Dataset
This data set is a subset of the THUCNews news text classification data set provided by the Tsinghua NLP group. This training uses 10 categories (sports, finance, real estate, home, education, technology, fashion, current politics, games and entertainment). There are 6500 items in each category and a total of 65000 news data. The data set division is shown as the following table:

| Data    | Shape         | Data    | Shape         |
|---------|---------------|---------|---------------|
| x_train | [50000, 600]  | y_train | [50000, 10]   |
| x_val   | [5000, 600]   | y_val   | [5000, 10]    |
| x_test  | [10000, 600]  | y_test  | [10000, 10]   |

4. Experimental Design
4.1. TextCNN Parameter Setting and Model Construction
According to the introduction in Chapter 2, textCNN is used and our model structure is shown in the following figure:
4.2. RNN Parameter Setting and Model Construction

After designing the textCNN network model, we also design the RNN network to construct the text multi-classification of Chinese news. The structure of the RNN model is as follows:

5. Experimental Data Analysis

The experimental analysis is mainly composed of the following two aspects: The first aspect is the analysis of textCNN network training and the test results data, and the second aspect is the analysis of RNN network training and the test results data.

First of all, the training process of textCNN network in the training set is as follows:
Figure 6. Training process diagram

The optimal effect on the verification set is 94.12%, and it has been stopped after only three iterations.

The accuracy rate and error are shown in the figure

![Figure 7. Training accuracy and error diagram](image)

The accuracy rate on the test set is 96.04%, and each precision, recall and f1-score exceed 0.9.

The confusion matrix is shown in the figure:

![Figure 8. TextCNN confusion matrix diagram](image)

The training process of the RNN network in the training set is similar to the textCNN network:

The optimal effect on the verification set is 91.42%. It has been stopped after 8 iterations and the speed is much slower than that of CNN.
The accuracy rate and error are shown in the figure:

![Figure 9. Training accuracy and error diagram](image)

The accuracy rate on the test set is 94.22%, and each precision, recall and f1-score exceed 0.9 except for the category of home.

The confusion matrix is shown in the figure:

![Figure 10. RNN confusion matrix diagram](image)

6. Conclusion
This study is the multi-classification of Chinese news text based on the deep learning algorithm. The experimental results show that convolutional neural network and recurrent neural network all have an excellent recognition effect on Chinese text. Moreover, we use word embedding in the input layer to carry out vector representation to the Chinese character, and the vector dimension between the words is greatly reduced compared with the previous traditional one-hot encoding. By calculating the Euclidean distance or the cosine distance of the two low-dimensional vectors, the semantic relation between the words can be excavated, so that the program has better recognition effect. At the same time, based on Python flask web framework and neural network model weights, we have made visualization of model inference, which has provided the basis for future product applications excavated by text classification and automatic semantics.

References
[1] Greff K, Srivastava R K, Koutnik J, et al. LSTM: A Search Space Odyssey [J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 28 (10):2222-2232.
[2] Gers F A, Schmidhuber J. LSTM recurrent networks learn simple context-free and context-sensitive languages [J]. IEEE Trans Neural Netw, 2001, 12 (6):1333-1340.
[3] Gers F A, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM [C]// International Conference on Artificial Neural Networks. 2002.
[4] Lan Wenfei, Xu Wei, Wang Tao. Chinese news text classification based on convolutional neural network[J]. Journal of South-Central University for Nationalities (Natural Sciences), 2018, 37 (1):138-143.
[5] Greff K, Srivastava R K, Koutnik J, et al. LSTM: A Search Space Odyssey [J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 28 (10):2222-2232.
[6] Gers F A, Schmidhuber J. LSTM recurrent networks learn simple context-free and context-sensitive languages [J]. IEEE Trans Neural Netw, 2001, 12 (6):1333-1340.

[7] Gers F A, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM[C]// International Conference on Artificial Neural Networks. 2002.