LSTM-CNN을 이용한 텍스트 분류

신성윤[1*] · 신광성[2] · 이현창[2]

[1]군산대학교 · [2]원광대학교

Text Classification Using LSTM-CNN

Seong-Yoon Shin[1*] · Kwang-Seong Shin[2] · Hyun-Chang Lee[2]

[1]Kunsan National University · [2]Wonkwang University

E-mail : s3397220@kunsan.ac.kr / waver0920@wku.ac.kr / hclglory@wku.ac.kr

요 약

본 논문은 딥 러닝 모델 융합, 즉 LSTM-CNN 모델을 기반으로 한 다중 레이블 뉴스 데이터 세트 THUCNews에서 텍스트 감정 분류 실험을 수행하는 방법을 제안한다. 비교 실험을 통해, 텍스트 정서 분류 실험에서 모델 융합 방법 LSTM-CNN의 정확도는 97.68 %에 이르렀으며 이는 랜덤 포레스트 알고리즘보다 28.12 % 증가했으며, RNN (Recurrent Neural Network) 모델에 비해 2.84 % 증가한 것이다.

ABSTRACT

This paper proposes a method based on deep learning model fusion, namely LSTM-CNN model, which performer text sentiment classification experiments in the multi-label news dataset THUCNews. Through comparison experiments, the accuracy of the model fusion method LSTM-CNN in text sentiment classification experiments reaches 97.68%, which is increased by 28.12% than the Random Forest algorithm, and increased by 2.84% compared with the Recurrent Neural Network model (RNN).

키워드

deep learning model fusion, LSTM-CNN, THUCNews, RNN

Ⅰ. Introduction

Natural language processing (NLP) has always been the focus of attention in the field of science and technology. For example, popular product recommendation systems, speech translation systems, intelligent robot question and answer systems, etc., are typical applications of natural language processing research [1, 2]. Natural language processing is mainly used in spell checking, network information extraction, text sentiment classification, machine translation, spoken dialogue, and complex question and answer systems [3].

* corresponding author

Ⅱ. Proposed Model

The model we proposed consists of an initial LSTM layer that receives the word embedding for each token in the experimental data after data preprocessing. The token it outputs not only stores the information of the initial token, but also stores any previous tokens. In other words, the LSTM layer is generating a new encoding for the original input. The output of the LSTM layer then inputs to the desired local feature convolution layer, and the output of the final convolution layer will be aggregated into a smaller dimension, ultimately outputting the classification label of the text sentiment. In the past experiments, the LSTM-CNN model was mostly used for the binary category

problem of the text. In this experiment, we mainly solved the multi-classification problem of the text. The structure of the LSTM-CNN model is shown in Fig. 1.

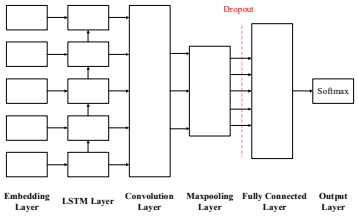

Fig. 1. LSTM-CNN model structure.

III. Experiments

The experimental data training has a batch size of 64, and each batch is trained in 100 batches, which are divided into 79 training batches. In the process of training, the best recording algorithm is used, that is, the performance of the current training result and the previous training result are checked and compared. If the performance of the training result is not improved after a certain number of training times, the training is ended early and the best training model is saved. Our analysis of the experimental results is as follows.

1). Compared with the traditional machine learning model, the deep learning-based neural network model has certain advantages in the classification effect of text sentiment classification.

2). Neural network model is generally better for text sentiment classification. The accuracy of LSTM-CNN based on model fusion is up to 97.68%, which is slightly better than the CNN model and better than LSTM model.

3). The best training accuracy was achieved when LSTM-CNN was trained about 300 times. Figs. 2 and 3 intuitively illustrate the fast convergence of the LSTM-CNN model, which is more efficient in obtaining accuracy.

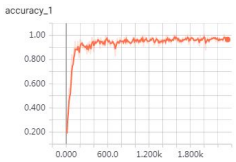

Fig. 2. Training accuracy curve of the LSTM-CNN model.

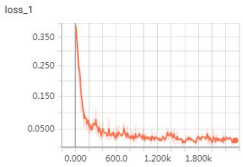

Fig. 4. Training loss rate graph of the LSTM-CNN model.

4). LSTM-CNN confusion matrix are shown in Fig. 4 intuitively illustrates that the classification of finance, science and technology, fashion, and game is the best, reaching 100%. But the lowest classes of home life classification, only 92%. This indicates that the text contents of the home life news are short, and the features extracted by the model are limited, resulting in a decline in the classification effect.

IV. Conclusion

This paper proposed a text sentiment classification method based on deep learning method model fusion. The method is compared with CNN, RNN, LSTM and two traditional text classification methods. A comparative experiment of text categorization was performed on the dataset THUCNews. The proposed method is to fuse the convolution neural network with the long-short term memory network to form the LSTM-CNN neural network model.

Acknowledgement

“This research is partially supported by Institute of Information and Telecommunication Technology of KNU.”

References

[1] Chetan Arora, Mehrdad Sabetzadeh, Lionel Briand, Frank Zimmer, "Automated Checking of Conformance to Requirements Templates Using Natural Language Processing", *IEEE Transactions on Software Engineering*, Vol.41, pp.944-968, 2015. DOI: 10.1109/TSE.2015.2428709.

[2] Mohd Ibrahim, Rodina Ahmad, "Class Diagram Extraction from Textual Requirements Using Natural Language Processing (NLP) Techniques",

2010 Second International Conference on Computer Research and Development, pp.200-204, 2010. DOI: 10.1109/ICCRD.2010.71.

[3] Sweta P. Lende, M. M. Raghuwanshi, "Question answering system on education acts using NLP techniques", *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, pp.1-6, 2016. DOI: 10.1109/STARTUP.2016.7583963.