Design and Implementation of an Attention-Based Interpretable Model for Document Classification

Zuoxian Ye*1

1 School of Information Science and Technology, Peking University, Beijing, 100871, China

*Corresponding author’s e-mail: yezuoxian@pku.edu.cn

Abstract. With the rapid development of artificial intelligence and machine learning, deep learning technology has been successfully applied to various tasks in the field of natural language processing. However, the erroneous decisions produced by the deep learning model and the characteristics of its black box cause users to doubt its decisions. We urgently need the model to have the ability to make rationalized explanations of its decisions, that is, interpretability. In this paper, we propose an attention-based interpretable model for document classification, which aims to have a better model performance and at the same time have better interpretability. Experiments show that our classification model is close to the performance of other mainstream models. Also, the attribution score explanation provided by the interpretation method are effective.

1. Introduction

In recent years, with the rapid development of artificial intelligence and machine learning, deep learning [1] has been widely used in computer vision [2-3], natural language processing [4], speech recognition [5] and other fields. The deep neural network model not only surpasses the traditional machine learning model in many field tasks, but also becomes the best model at present, even comparable to human results in some real tasks [6]. Nowadays, people enjoy the convenience of deep learning technology through intelligent devices, and at the same time, deep learning technology also plays an important role in assisting people's decision-making in many fields.

Compared with traditional machine learning methods, deep neural network has achieved great success because of its ability to automatically extract features from data, carry out end-to-end learning, and do not need researchers to manually construct features. But there is also an unavoidable problem, that is, deep learning model may have unexpected errors. The researchers found that even if only slight changes were made to the input image of the model, the deep neural network model may get completely irrelevant prediction, and in most of the experiments on the counter samples, the deep neural network model shows poor robustness and is easy to be deceived by the counter samples. Also, most of the deep neural network models have millions of parameters to learn, and the huge number of parameters makes the model basically unable to be understood.

In this paper, we focus on the interpretability of model based on attention mechanism under text classification task, and propose an interpretable text classification model based on attention mechanism, which has good text classification performance and strong interpretability. Specifically, we first introduce the self-attention mechanism into bidirectional long short-term memory network, which
increases the model’s ability of modelling the relationship between words. Second, we introduce the attribution score to compute the importance of the input to model decision.

This paper uses many experiments and various evaluation metrics to verify the effectiveness of the model proposed in this paper. Also, we use logic rules to explain the decision that the model making.

2. Related Work

In recent years, researchers proposed many methods to improve the interpretability of deep learning model. The most direct interpretation method is to add the explanation structure into the model, which is also called the active interpretability intervention [7]. This method is based on designing the model architecture or modifying model training method in the process of model training. Bahdanau et al. add attention mechanism to seq2seq framework to improve the performance of the neural machine translate model, also regraded the attention weight as the soft align in translation [8]. Another interpretation method corresponding to the active intervention is the post-hoc interpretation method, which is also the mainstream interpretation method now. The method is suitable for explaining the model with large parameter and complex network structure. The most classic one is Local Interpretable Model-Agnostic Explanations (LIME) [9]. Also, some post-hoc interpretation method use visualization-based methods like Class Activation Mapping (CAM) [10], Grad-CAM [11] and Grad-CAM++ [12].

According to whether the interpretation method is related to the model, the interpretation method can be divided into model specific methods and model agnostic methods. The interpretation method of a specific model is limited to a specific model, so a specific model must be used for the task to obtain the interpretation of the method. If you want to get the decision tree interpretation of the model, you can only use the tree model for tasks. The interpretation method of the specific model related to the model corresponds to the interpretation method of the non-specific model, that is, the prediction and interpretation of the model are two independent parts respectively. This kind of method generally uses the post-hoc Interpretative method as the interpretation of the model.

3. The Attention-Based Interpretable Model

In this paper, we propose an attention-based interpretable model. We first divided our proposed methods into two parts. First is the classification module, and the second part is the interpretation module. The classification module takes the input $x$, and give the classification result $y$. After that, the interpretation module gives the explanation of the model decision. In this part, we will introduce the architecture of the model.

3.1. Classification module

As illustrated in Figure 1, based on Zhang's model [13], this paper proposed a self-attention based bidirectional long short-term memory model (SA-BiLSTM).

The model contains five layers. The first layer, the input layer takes the sentence as the input $x$, and then passed the sentence to the word embedding layer. The word embedding layer uses pre-trained word embedding map each word into fixed size word vector. The above process can be described by following,

$$v_i = Wx_i, \ i \in [1, n]$$

where $n$ refer to the number of words in the sentence, and $v_i$ refer to the vector of the word $i$.

The LSTM layer take the $v_i$ as input and computes the hidden state at time $t$. We use bidirectional LSTM, so we can get the forward hidden state $\overrightarrow{h_t}$, and the backward hidden state $\overleftarrow{h_t}$ at time $t$. The above process can be described by following,

$$\overrightarrow{h_t} = \text{LSTM}(v_i), \ i \in [1, n]$$

$$\overleftarrow{h_t} = \text{LSTM}(v_i), \ i \in [n, 1]$$
Figure 1 Architecture of SA-BiLSTM.

we use element-wise sum to combine the forward hidden state $\bar{h}_t$ and the backward hidden state $\bar{h}_t$:

$$h_t = [\bar{h}_t \oplus \bar{h}_t]$$ (4)

Inspire by the self-attention used in transformer [14], we use self-attention mechanism to compute the attention weight in self-attention layer. The self-attention layer first concatenates all output of the LSTM layer as metric $H$, and then multiply metric $H$ by trainable metrics $W^Q$, $W^K$, and $W^V$:

$$H = [h_1; h_2; \ldots; h_t], t \in [1, n]$$ (5)

$$Q = H \cdot W^Q$$ (6)

$$K = H \cdot W^K$$ (7)

$$V = H \cdot W^V$$ (8)

Then use $Q$, $K$ and $V$ to compute the self-attention layer output $H^*$:

$$H^* = \text{softmax}(QK^T \cdot d_K^{-1/2}) \cdot V$$ (9)

The output layer use MLP to compute the model decision $y$ according to $H^*$:

$$p(y|X) = \text{softmax}(Wh^* + b_h)$$ (10)

$$y = \text{argmax}_y p(y|X)$$ (11)

where $W_h$ and $b_h$ are trainable parameters.

3.2. Interpretation module

In interpretation module, we interpret the model by computing the attribution score. The attribution score measures the importance of the input feature to the model output. Inspire by ATTATTR proposed by Hao et al. [15], we propose attention and integrated gradient mixed attribution method (ATTIG), which can not only measure the importance of the input words also can measure the strength of the relationship between words.
Given the input $x$, the model’s output $y$ and the attention weight $A$ computed by model, the ATTIG can be described by following,

$$S_{ATTIG} = A \cdot f_0^1 \partial f_y(\alpha A) \cdot \partial A^{-1}$$  \hspace{1cm} (12)

The attribution score can be efficiently computed via Riemann approximation of the integration [16]. Specifically, we use m-step approximation from the zero attention matrix to the original attention weight $A$:

$$S_{ATTIG} = Am^{-1} \cdot \sum_{l=1}^{m} \partial f_y(m^{-1}A) \cdot \partial A^{-1}$$  \hspace{1cm} (13)

Finally, we maximum the ATTIG score by column as the attribution score of the input $x$:

$$s_k = \max(S_{i,k})_{i=1}^{n}, k = 1, ..., n$$  \hspace{1cm} (14)

4. Experimental Result and Analysis

In this paper, we try to clarify the performance of the classification model and the effectivity of the interpretation methods. We use classification task to verify the performance of the classification model and use interpretability task to verify the effectivity of the interpretation methods.

4.1. Datasets

We use Stanford Sentiment Treebank v2 (SST2) and THUCNews as our datasets. The SST2 dataset contains two categories, which distinguish a sentence as a positive or negative sentiment. THUCNews is generated by filtering historical data of Sina News RSS subscription channel from 2005 to 2011. The THUCNews dataset contains ten categories.

4.2. Compared methods and evaluate metrics

For the classification task, we use CNN [17], LSTM [18] and the Transformer [14] model as the compared methods. And we use the precision, recall and the F1 score as our evaluate metrics.

For the interpretability task, we use attention, GradientX [19], deepLIFT [20] and integrated gradient [16] as compared methods. And we use the difference of the precision, recall and the F1 score between before and after masking the max attribution score word.

4.3. Experimental result

The performance comparison of different methods of classification task in SST2 dataset and THUCNews dataset is demonstrated in Table 1.

|       | SST2    |         |         | THUCNews |         |         |
|-------|---------|---------|---------|----------|---------|---------|
| Model | Precision | Recall | F1 score | Precision | Recall | F1 score |
| CNN   | 78.33   | 78.33   | 78.33   | 88.46    | 88.32   | 88.33   |
| LSTM  | 78.14   | 78.05   | 78.04   | 89.77    | 89.69   | 89.69   |
| Transformers | 80.51 | 80.25   | 80.21   | 90.36    | 90.28   | 90.28   |
| SA-BiLSTM | **81.21** | **81.19** | **81.18** | **90.87** | **90.83** | **90.81** |

As shown in Table 1, our propose model SA-BiLSTM outperforms other methods in both SST2 dataset and THUCNews dataset.

The performance comparison of different methods of interpretability task in SST2 dataset and THUCNews dataset is demonstrated in Table 2.
Table 2  Experimental results of interpretability task.

| Model       | SST2 Precision | SST2 Recall | SST2 F1 score | THUCNews Precision | THUCNews Recall | THUCNews F1 score |
|-------------|----------------|-------------|---------------|--------------------|----------------|------------------|
| Attention   | -3.32          | -3.3        | -3.29         | -8.21              | -9.48          | -9.19            |
| GradientX   | -7.47          | -7.48       | -7.48         | -7.04              | -7.85          | -7.68            |
| DeepLIFT    | -7.33          | -7.32       | -7.31         | -7.37              | -8.2           | -8.02            |
| ATTIG       | **-7.88**      | **-7.87**   | **-7.86**     | **-8.42**          | **-9.38**      | **-9.18**        |

As shown in Table 2, our propose model ATTIG outperforms other methods in both SST2 dataset and THUCNews dataset.

5. Conclusion
Though its outstanding forecasting performance, the lack of interpretability brings many limitations in utilizing deep learning models in most realistic scenarios. In this paper, we propose an attention-based interpretable model. We divide the model into two modules, the classification module and the Interpretation module. The classification module takes the input x, and give the classification result y. The interpretation module gives the explanation of the model decision. We design various experiments to prove that our model has good performance and good interpretability.

Acknowledgments
The authors were supported by the National Key R&D Program of China 2018YFB1004100.

References
[1] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.
[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.
[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[4] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215.
[5] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 29(6), 82-97.
[6] Yonghui, W., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Macherey, K. (2016). Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.
[7] Zhang, Y., Tiño, P., Leonardis, A., & Tang, K. (2020). A Survey on Neural Network Interpretability. arXiv preprint arXiv:2012.14261.
[8] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
[10] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).
[11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[12] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 839-847). IEEE.

[13] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016, August). Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 207-212).

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

[15] Hao, Y., Dong, L., Wei, F., & Xu, K. (2020). Self-attention attribution: Interpreting information interactions inside transformer. arXiv preprint arXiv:2004.11207.

[16] Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International Conference on Machine Learning (pp. 3319-3328). PMLR.

[17] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

[18] Zhang, S., Zheng, D., Hu, X., & Yang, M. (2015, October). Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia conference on language, information and computation (pp. 73-78).

[19] Denil, M., Demiraj, A., & De Freitas, N. (2014). Extraction of salient sentences from labelled documents. arXiv preprint arXiv:1412.6815.

[20] Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In International Conference on Machine Learning (pp. 3145-3153). PMLR.