An aspect sentiment analysis model based on adversarial training and multi-attention

Qing Wang, *Hailong* Chen[*], and *Xin* Zheng

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China

Abstract. Aiming at the disadvantages of the gradient vanishing and exploding of the Recurrent Neural Network in the traditional deep learning algorithm and the problem that the Convolutional Neural Network cannot obtain the global features of the classified text, a CNN(Convolutional Neural Network)-BiLSTM (Bidirectional Long Short-Term Memory) sentiment analysis method based on adversarial training and multi-layer attention is proposed to give full play to the ability of CNN to extract phrase-level features of text and the ability of BiLSTM to extract global structural information of text, and the multi-layer attention mechanism will assign higher weights to keywords, and the adversarial training can well solve the model instability problem of the current deep learning model. Using the public data set Laptop reviews and Restaurant Reviews from SemEval 2014 for verification, the results show that the accuracy of the model proposed in this paper is 1 and 1.9 percentage points higher than that of the original model on the two data sets. In contrast, the model is more efficient in aspect-level sentiment classification tasks.

Keywords: Sentiment analysis, Convolution neural network, Bidirectional long short-Term memory network, Attention mechanism, Adversarial training.

1 Introduction

In recent years, with the rapid development of Internet technology, massive amounts of video, text, pictures, and other data information are being generated all the time. The amount of text information is huge and the data is chaotic, so it is difficult to classify and organize manually. The work of judging their emotional trends is even more difficult to do manually, so how to achieve emotional polarity analysis through machines has become increasingly important. In 2003, Nasukawa et al. proposed the concept of sentiment analysis[1]. The purpose of sentiment analysis is to clarify people's views on a comment, an article, etc.

So far, there are three mainstream sentiment analysis methods: (1) sentiment analysis method based on sentiment dictionary[2], Lin Jianghao et al[3] proposed a method of constructing a domain sentiment dictionary based on word vectors, based on the semantic

[*] Corresponding author: hrbustchl@163.com

similarity calculation of specific domain corpus, makes the extracted sentiment features more domain-specific, and is not constrained by the range of candidate sentiment word sets. (2) Sentiment analysis method based on machine learning. In 2002, Pang[4] et al. used the machine learning method to perform sentiment analysis on film review data for the first time. In 2018, Fan Zhen[5] and others proposed a method based on a sentiment dictionary. The sentiment tendency of the comment text is calculated, and the weak annotation information of user ratings and the sentiment tendency based on the dictionary method is used to automatically mark the comment text. Finally, SVM is used to classify the sentiment of the comment text. (3) The method based on deep learning[6], in 2019 Wan Qibin[7] and others proposed a BiLSTM-based Attention-CNN hybrid neural network text classification method, introduces the attention mechanism[8], extracts the attention score of each value, in 2021 Hu Jiming[9] uses the CNN-BiLSTM-Attention model to classify the policy text, improve The effect and accuracy of the policy text classification. At present, deep learning has been widely used in sentiment analysis tasks at home and abroad, but there are still some problems: (1) Adversarial training is widely used in image processing, but less in text processing, especially in sentiment analysis. (2) The fusion model is rarely used in text sentiment analysis, and most of the existing research innovations are to improve and optimize a single neural network model, lacking the application of fusion network models.

The main contributions of this paper are as follows:

(1) Combined with CNN and BiLSTM, the ability of CNN to extract text phrase-level features and the ability of BiLSTM to extract text global structure information are fully utilized. (2) The perturbed word vector embedding layer is added to avoid overfitting and the appearance of adversarial samples, which is used to solve the model instability problem of the current deep learning model, improve the stability of the model, and enhance the generalization ability of the model.

2 Aspect sentiment analysis model combining adversarial training and multi-attention

This paper takes RAM as the basic model and combines the CNN layer based on the BiLSTM layer in the original model to give full play to the ability of CNN to extract text phrase-level features and BiLSTM to extract the ability of text global structure information, and the perturbed word vector embedding layer is added to prevent the occurrence of adversarial samples due to overfitting, which is used to solve the model instability problem of the original model, improve the stability of the model, and enhance the generalization ability of the model. The improved model includes an input layer, CNN layer, BiLSTM layer, position weighted memory layer, and recurrent multi-attention layer. The overall architecture is shown in Figure 1.

2.1 Input Layer

The input layer of the model consists of the word embedding[10] layer and the word embedding perturbation layer. The purpose of the word embedding perturbation layer is to increase the update of parameters in the training process. By using adversarial training to create a perturbation to the text input to the model, and then with the initial samples participate in the training of the model together, which has a good effect on preventing the model from overfitting and improving the generalization ability, In 2014, Goodfellow[11] first proposed the concept of adversarial training and applied it to computer vision, and in 2016, Miyato[12] et al. first applied adversarial training to text processing.

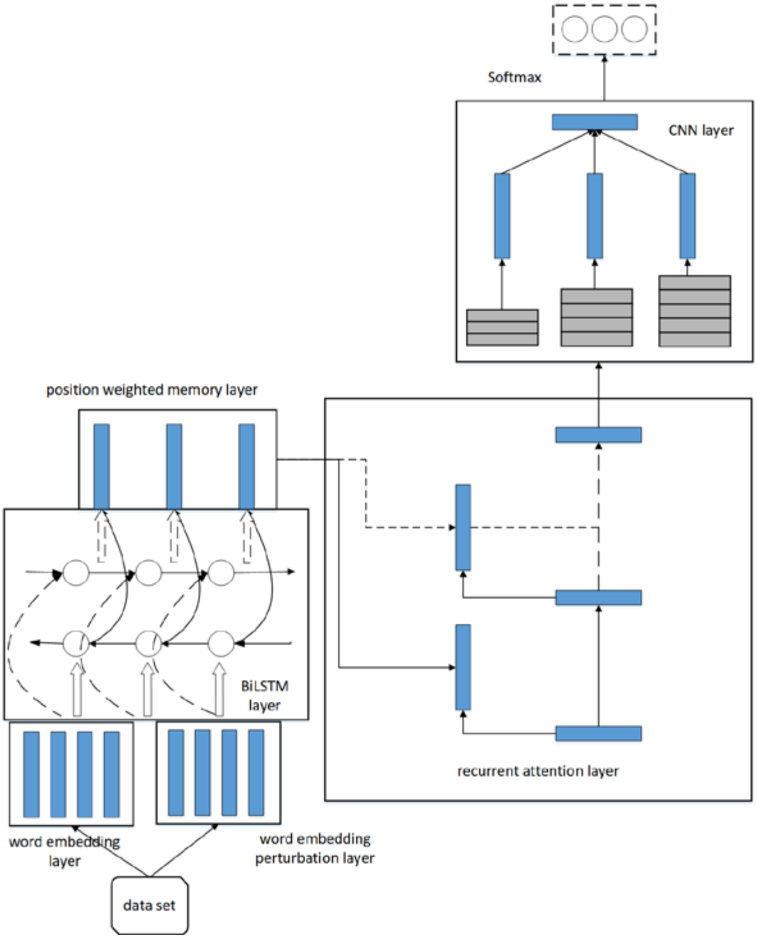
Softmax
CNN layer
position weighted memory layer
BiLSTM layer
recurrent attention layer
word embedding layer
word embedding perturbation layer
data set

Fig. 1. Overall model architecture.

In 2014, Szegedy first proposed the concept of adversarial samples at the ICLR conference, which is considered to be the pioneering work of adversarial samples[13].This paper adopts the continuous method to implement adversarial training. The current adversarial training methods are mainly divided into four types: FGSM and FGM methods, PGD methods, FreeAT methods and FreeLB methods[14]. We use the FGM method for adversarial training, which can be expressed as a minimum-maximization formula, such as formula (1) shown

$$\min_{\theta} E_{(x,y)\sim D}[\max_{\delta\in\varsigma} L(\theta, x+\delta, y)] \tag{1}$$

Among them, x is the input training sample, y is the label of the training sample, θ is the set of model parameters, D is the sample training set, δ is the confrontation disturbance, and L is the loss function of the neural network. Using the FGM method, the T words contained in each text are expressed as formula (2)

$$\left\{ w^{(t)} \mid t = 1, \ldots, T \right\} \tag{2}$$

The corresponding category is y , the word vector matrix is expressed as the formula (3)

$$V \in R^{(K+1)\times D} \tag{3}$$

K is the number of words in the vocabulary, D indicating the dimension of the word vector.

In the word vector embedding layer, $\mathrm{v_k}$ is the embedding of the i th word, and the perturbed word vector embedding layer uses the regularized embedding $\mathrm{v_k}^{`}$ to represent $\mathrm{v_k}$, the purpose is to convert the discrete vector input into a continuous vector input, as shown in formula (4)(5)

$$v_k^{`} = \frac{v_k - \sum_{j=1}^{K} f_j v_j}{\sqrt{Var(v)}} \tag{4}$$

$$Var = \sum_{j=1}^{K} f_j (v_j - \sum_{j=1}^{K} f_j v_j)^2 \tag{5}$$

f_i represents the word frequency of the i th word, and r_{adv} is obtained by the backpropagation gradient descent function and the L2 regularization constraint, as shown in equations (6) and (7), where x is the model input and $\bar{\theta}$ is the short text classifier parameter.

$$g = \nabla_x \log(y \mid x; \bar{\theta}) \tag{6}$$

$$r_{adv} = \frac{g}{\|g\|^2} \tag{7}$$

2.2 BiLSTM Layer

BiLSTM is an improved LSTM, which is a type of RNN. The forgetting gate f_t, memory gate i_t , and output gate o_t are all calculated from the hidden state of the previous moment h_{t-1} and the input of the current moment x_t . Finally, the current hidden state is obtained by calculation h_t [15].

Setting the hidden state output by forwarding LSTM at time t as $\overrightarrow{h}$, and the hidden state output by reverse LSTM as $\overleftarrow{h}$, the hidden state output by BiLSTM can be h is expressed as formula (8)

$$h = \overrightarrow{h} \oplus \overleftarrow{h} \tag{8}$$

So the vector matrix generated in this layer of work is H_t^*, as the following formula (9)(10)

$$H_t^* = \{h_1^*, h_2^*, \ldots, h_t^*, \ldots, h_T^*\} \tag{9}$$

$$h_t^* = (\overrightarrow{h_t}, \overleftarrow{h_t}) \tag{10}$$

2.3 Position weighted memory layer

In the original model RAM, the weight value of the t^th^ word in the sentence is calculated by the formula (11), where $m_{\max}$ is the maximum length of the input sentence.

$$\alpha_t = 1 - \frac{|m - \tau|}{m_{\max}} \tag{11}$$

The relative offset between each word in the sentence and the target word is expressed as (12)

$$\beta_t = \frac{|m - \tau|}{m_{\max}} \tag{12}$$

The final position weighted memory value is obtained as shown in formula (13) (14)

$$H_t = \{h_1, h_2, \ldots, h_t, \ldots h_T\} \tag{13}$$

$$h_t = (\alpha_t \cdot h_t, \beta_t) \tag{14}$$

2.4 Recurrent multi-attention layer

The original model makes multiple attentions[16] on the output of the position-weighted memory layer, and the words concerned by each attention are different. Finally, the weighted results of the attention are combined nonlinearly with the GRU network.

Where e_{t-1} is the result of the attention at the previous moment, x_t is the information of the current attention, m_t is the input of the attention layer, e_{t-1} is the result of the attention at the previous moment, first calculate the attention value of each input vector matrix, $[,\beta_\tau]$ represents the attention The final result of the layer output is related to the comment target entity, as shown in formula (15)

$$\mathrm{g_t} = W_t(m_t, e_{t-1}[,\beta_\tau] + b_t) \tag{15}$$

Next, normalize the attention value of each input vector matrix, as shown in Eq. (16)

$$\alpha_j = \frac{\exp(g_t)}{\sum_i \exp(g_i)} \tag{16}$$

At time t, the result of the attention at the previous moment x_t and the input e_{t-1} at the current moment is used as the input of the GRU layer, as shown in formula (17)

$$x_t = \sum_{j=1}^{T} \alpha_j^{\ t} m_t \tag{17}$$

2.5 CNN layer

The output of the recurrent multi-attention layer is used as the input of the CNN layer, it consists of input layer, convolution layer, pooling layer, fully connected layer and output layer[17], and the convolution operation is performed, and the set filter is used to achieve feature extraction, as shown in formula (18)

$$S_i = f(\omega \times X_{i:i+g-1} + b) \tag{18}$$

Among them, ω is the convolution kernel, g is the size of the convolution kernel, $X_{i:i+g}$ represents the sentence vector-matrix formed from the words i to $i+g-1$, and b is the bias vector. After the convolution layer, the feature matrix is obtained as formula (19)

$$S = [s_1, s_2, \ldots, s_{n-g+1}] \tag{19}$$

Then, through the pooling layer, the local feature matrix of the sentence passed through the convolution layer is down-sampled, and the maximum pooling technology Max Pooling is used to obtain the local optimal solution as shown in formula (20)

$$M = \max\{s_1, s_2, \ldots, s_{n-g+1}\} \tag{20}$$

Finally, the fully connected layer connects the M_i vector after the pooling layer into a vector Q as the input of BiLSTM, as shown in formula (21)

$$Q = \{M_1, M_2, \ldots, M_n\} \tag{21}$$

2.6 Output layer

Through the classification of the Softmax function, the probability distribution value of the category to which each text belongs is obtained, and the category with the largest value is the predicted category. Among them, N is the category set of sentiment analysis, which adopts three categories; D is the data set of training samples; y is a one-hot vector; $f^i(x;\theta)$ indicates the predicted sentiment distribution of the model, and λ is the weight of the regularization term, such as Formula (23) shows

$$L = \sum_{i \in N} \sum_{(x,y) \in D} y^{i} \log f^{i}(x;\theta) + \lambda \left\| \theta \right\|^{2} \tag{23}$$

3 Experiment and analysis

3.1 Data set

The two datasets used in the experiment are both from SemEval 2014 (Pontiki et al., 2014), which contain user comments on the restaurant domain and the laptop domain, and select the positive, negative and neutral data for the experiment, it is divided into training set, validation set and test set with a ratio of 8:1:1. The basic information of the data set division is shown in Table 1.

Table 1. Basic information of data set division.

	Positive		**Neutral**		**Negative**	
Restaurant reviews	Train	Test	Train	Test	Train	Test
	2159	730	800	195	632	196
Laptop reviews	Train	Test	Train	Test	Train	Test
	980	340	454	171	858	128

3.2 Experimental environment and parameter settings

Environment: OS: Windows10, memory: 8GB, CPU: Intel Core i7, python3.6, 64-bit, developed using the deep learning framework TensorFlow;

Model hyperparameters: the word embedding dimension is 300, the BiLSTM hidden layer size is 300, the dropout rate is 0.5, the convolution kernel size is 5, the number of convolution kernels is 200, the learning rate is 0.005, the number of classifications is 3, the attention layers is 3 and the l2 regularization parameter is 0.001.

Model training parameters: the batch size is 32, and the maximum number of training iterations is 500 and the against perturbation strength is 5.0.

3.3 Comparative experiment

We compare the single sentiment analysis models LSTM[18], BiLSTM[15], CNN[19] with the improved model in this paper. The experimental results are shown in Table 2.

Table 2. Comparison of experimental accuracy.

Model	**Laptop reviews**		**Restaurant reviews**	
	Acc	Macro-F1	Acc	Macro-F1
CNN	0.667	0.625	0.712	0.641
LSTM	0.725	0.686	0.781	0.662
BiLSTM	0.734	0.667	0.793	0.691
Our model	0.754	0.755	0.822	0.748

According to Table 2, the accuracy of our model is 8%, 3% and 2% higher than that of the CNN, LSTM and BiLSTM models on the dataset Laptop reviews. On the dataset Restaurant Reviews, it is improved by 11%, 4% and 3%.

The selection of attention layers is critical to the accuracy of the final sentiment analysis. Therefore, it is necessary to select an optimal attention layer for the sentiment classification task through experiments, the experimental results are shown in Table 3. Considering that two datasets are used for validation, so three layers of attention are applied to perform sentiment analysis on both datasets.

Table 3. The effect of the number of attention layers on accuracy.

data set	accuracy				
	one floor	two floors	three floors	four floors	five floors
Laptop	0.717	0.754	0.748	0.735	0.730
Restaurant	0.800	0.814	0.820	0.822	0.809

Next, the improved sentiment analysis models CNN-BiLSTM, CNN-BiLSTM-Attention[20], RAM[21] and the model proposed in this paper are used to conduct comparative experiments. The experimental results are shown in Table 4.

Table 4. Experimental results.

Model	Laptop reviews		Restaurant reviews	
	Acc	Macro-F1	Acc	Macro-F1
CNN-BiLSTM	0.739	0.688	0.793	0.673
CNN-BiLSTM-Attention	0.741	0.684	0.797	0.698
RAM	0.744	0.713	0.803	0.708
Our model	0.754	0.755	0.822	0.748

According to Table 4, the accuracy of our model is 1.5%, 1.3% and 1% higher than that of the CNN-BiLSTM, CNN-BiLSTM-Attention and RAM models on the dataset Laptop reviews. On the dataset Restaurant Reviews, it is improved by 2.9%, 2.5% and 1.9%.

To verify the effectiveness of the model in preventing overfitting, compare the convergence of the two models as shown in Figure 2 and Figure 3. It can be seen from this that improving the adversarial training in the model can prevent the model from overfitting, and by generating a misclassified counterattack model, the wrong samples are added to the training process so that the model has the robustness and relatively good generalization ability.

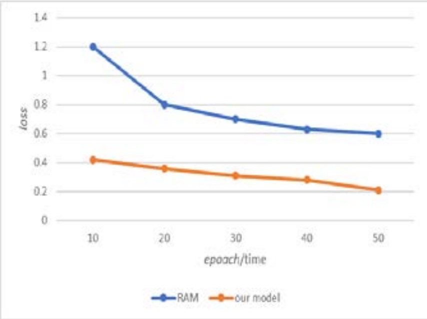

Fig. 2. Loss function value change curve.

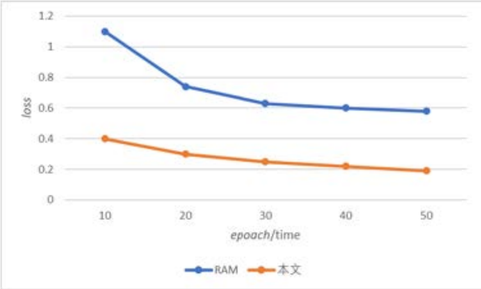

Fig. 3. Loss function value change curve.

4 Conclusion

This paper proposes a CNN-BiLSTM sentiment analysis model based on adversarial training and multi-attention. Incorporating adversarial training into the model can not only prevent overfitting from generating adversarial samples but also add wrong samples to training by generating a misclassified counterattack model. In the process, the model has the robustness and relatively good generalization ability, and it also achieves the purpose of fully mining the hidden semantic features of the text by integrating neural networks with different advantages. The linear combination can capture emotional features that are relatively far away. Compared with other sentiment analysis models, the effectiveness of the proposed model is verified and the accuracy of sentiment analysis is improved.

The model only verifies the effectiveness of the model on two datasets, and the datasets used are small in scale. Most of the experimental data are short texts and they are all English datasets. In the follow-up work, we will consider citing different large-scale Chinese or English datasets. Experiments are carried out on long text data in English. In addition, since the word embedding perturbation layer is added to the word embedding layer, the training time of the model is slightly higher than that of other sentiment classification models. We will optimize this problem in the follow-up work.

This work was financially supported by Special Foundation of Scientific and Technological Innovation for Young Scientists of Harbin, China (Grant No. 2017RAQXJ045). National Natural Science Foundation of China (NSFC) (Grant No. 61772160).

References

1. Nasukawa T, Yi J.Sentiment analysis: Capturing favorability using natural language processing[C]// International Conference on Knowledge Capture. DBLP, 2003.
2. Abdulmohsen Al-Thubaity,Qubayl Alqahtani,Abdulaziz Aljandal. Sentiment lexicon for sentiment analysis of Saudi dialect tweets[J]. Procedia Computer Science,2018,142.
3. Lin J, Zhou Y, Yang A, et al. Building of domain sentiment lexicon based on word2vec[J]. Journal of Shandong University (Engineering Science), 2018.
4. PANG B, LEE L, VAITHYANATHAN S, et al. Thumbs up? sentiment classification using machine learning techniques[C]. Empirical Methods in Natural Language Processing, Philadelphia, July 2002, 2002: 79-86.
5. FAN Z, GUO Y, ZHANG Z H, et al. Sentiment analysis of movie reviews based on dictionary and weak tagging information[J]. Journal of Computer Applications, 2018, 38(11): 3084-3088.
6. Jian Zhang, Shifei Ding, Nan Zhang. An overview on probability undirected graphs and their applications in image processing[J]. Neurocomputing,2018.
7. WAN Q B, DONG F M, SUN S F. Text Classification Method Based on BiLSTM-Attention-CNN HybridNeural Network[J]. Computer Applications and Software, 2020,37(9): 94-98, 201.
8. Usama M, Ahmad B, Yang J, et al. Equipping recur-rent neural network with CNN-style attention mechanisms for sentiment analysis of network reviews[J]. Computer Communications, 2019, 148.
9. Hu J M, FU W L, QIAN W, et al.Research on Pol-icyText Classification Model Based on Topic Model and Attention Mechanism[J]. Information studies: Theory and Application,2021,44(07):159-165.

10. Li S, Pan R, Luo H, et al. Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling[J]. Knowledge-Based Systems, 2021, 218(4):106827.
11. GOODFELLOW I J, SHLENS J, SZEGEDY C.Explaining and harnessing adversarial examples[G]//Proceedings of the International Conference on MachineLearning. Lille, France: International Machine Learning Society, 2015: 1-13.
12. MIYATA T，DAI A M， GOODFELLOW I. Adversarial training methods for semi-supervised text classification[G]//Proceedings of the International Conference on Learning Representations. Toulon, France: International Machine Learning Society, 2017:1-11.
13. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. Computer Science, 2013.
14. ZHANG X H.Research on Text Representation and Text Classification Method Based on Adversarial Training [D].Beijing Jiaotong University,2020.
15. BAI J, LI F, JI D H. Attention based BiLSTM-CNNChinese microblogging position detection model[J]. Computer Applications and Software, 2018, 35(3): 266-274.
16. Aishan Wumaier, WEI W L, Zaokere Kaddeer. Sentiment analysis based on bilstm+attention in sports fie-ld[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English), 2020, 37(2):142-149.
17. LUO F, WANG H F. Chinese text sentiment classification by RNN-CNN[J]. Acta Scientiarum NaturaliumUniversitatis Pekinensis, 2018, 54(3): 459-465.
18. YU W, ZHOU W N. Sentiment analysis of commodity reviews based on LSTM[J]. Computer Systems and Applications, 2018, 27(8): 159-163.
19. REN M, GAN G. Text emotion classification based on bidirectional LSTM model [J]. Computer Engineering and Design, 2018,39(07):2064-2068.
20. WANG L Y, LIU C H*, CAI D B, et al. Text Sentiment Analysis Based on CNN-BiLSTM Network and attention Model[J]. Journal of Wuhan Institute of Technology,2019,(04):386-391.
21. Peng C, Sun Z, Bing L, et al. Recurrent Attention Network on Memory for Aspect Sentiment Analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.