ORIGINAL ARTICLE

Fine-grained Multimodal Sentiment Analysis Based on Gating and Attention Mechanism

Yingxue Sun, Junbo Gao*, Yuanhao Guo

School of Information and Engineering, Shanghai Maritime University, Shanghai 201306, China. E-mail: jbgao@shmtu.edu.cn

Abstract: In recent years, more and more people express their feelings through both images and texts, boosting the growth of multimodal data. Multimodal data contains richer semantics and is more conducive to judging the real emotions of people. To fully learn the features of every single modality and integrate modal information, this paper proposes a fine-grained multimodal sentiment analysis method FCLAG based on gating and attention mechanism. First, the method is carried out from the character level and the word level in the text aspect. CNN is used to extract more fine-grained emotional information from characters, and the attention mechanism is used to improve the expressiveness of the keywords. In terms of images, a gating mechanism is added to control the flow of image information between networks. The images and text vectors represent the original data collectively. Then the bidirectional LSTM is used to complete further learning, which enhances the information interaction capability between the modalities. Finally, put the multimodal feature expression into the classifier. This method is verified on a self-built image and text dataset. The experimental results show that compared with other sentiment classification models, this method has greater improvement in accuracy and F1 score and it can effectively improve the performance of multimodal sentiment analysis.

Keywords: Multimodal Sentiment Analysis; Fine-grained; Attention Mechanism; Gating Mechanism; Late-fusion

1. Introduction

Sentiment analysis is a research hotspot in the field of NLP, mainly for learning various unstructured languages in texts, digging out useful information from these texts, and judging users' current preferences and emotional attitudes. Early research on text sentiment classification mainly used machine learning[1,2]; in recent years, deep language learning networks have played a vital role in various research methods, and more and more researchers use deep network learning text features.

However, with the popularization of the Internet and the introduction of new social media, the data generated by users is no longer in a single text form. More and more social users tend to use images and short texts such as multimodal content to express their opinions. The information of each modal may have different goals when being individually judged. For example, the emotion expressed by the text "Hah, I feel so happy" is positive, but the user's image expresses negative emotions. If the final label is negative at this time, text information alone cannot achieve the goal of negative emotion learning.

For multimodal sentiment analysis tasks, the core lies in how to make better use of the information of each modal to learn the interaction value between different

doi: 10.18686/esta.v7i4.166

modalities; the design of the information fusion mechanism between different modalities in the later period is also an important research point for multimodal sentiment classification. Although the single-modal deep learning model has made great progress, considering the above similarities and differences, multimodal sentiment analysis as a new field still needs in-depth research.

In order to overcome the deficiency of text information learning granularity in existing research, the poor feature learning ability caused by the noisy information of the image itself, and the relatively simple feature learning method between heterogeneous modalities, this paper proposes a fine-grained multimodal sentiment analysis method, which is based on gating and attention mechanism. The main contributions of this paper are as follows.

The text representation uses both word-level and character-level vectors. The word-level vectors learn context information through LSTM, adding the attention mechanism to assign different contributions of words. The character-level vectors learn the local optimal features of the text through CNN convolution, deeply mine the text fine-grained semantic emotions to enhance the ability to express effective features.

The image uses a pre-trained model to learn the image vector through migration and fine-tuning. To make the image better retain the original important information during the network learning process, a gating mechanism is added to control the transmission of the image information flow in the network.

The fusion image and text vector are put into the bidirectional LSTM together to enhance the interactive learning ability between the modalities. Finally, put the vectors into the classification layer to complete sentiment detection.

2. Related works

This section reviews the work done in the field of sentiment analysis from three aspects: sentiment analysis methods based on text, image, and multimodal data respectively.

2.1 Sentiment analysis based on text

In the past few years, user social information sentiment analysis research has been mainly carried out on one side: text information or image information. Text-level sentiment analysis started early and has now developed to a higher level. These models can predict author sentiment from text. Early language models mainly used the bag-of-words model[3] or the N-Gram method[4] to obtain text representations based on word frequency or co-occurrence window statistics, and then calculate text emotions; in addition, there are a lot of researches by extracting emotional keywords and designing a keyword library for text emotion detection[5,6]. These traditional methods have a simple structure and fast model learning speed, but they ignore the context and the syntactic dependence between words and fail to learn the information brought by the sentence syntax.

With the development of text learning research, text representation methods that incorporate contextual semantics, such as GloVe, word2vec, and Bert[7–11], have begun to be applied to text sentimental analysis. The text representation matrix generated by these methods contains contextual semantics. Words with a high co-occurrence rate have similar vector representations. Bert gives words dynamic representation to fully learn the different meanings in different contexts. These state-of-art language models combined with machine learning or deep learning methods have achieved good results[12–16]. The classifier proposed in the literature[13] consists of multiple trees constructed systematically by pseudo-randomly selecting subsets of components of the feature vector. The literature[14] proposed that the attention mechanism can concentrate on different parts of a sentence when different aspects are taken as input. In the paper[15], a neural network-based sequence model is proposed to classify certain sentences into three types according to the number of objects appearing in a sentence. Then, each group of sentences is entered into a one- dimensional convolutional neural network to categorize emotions.

2.2 Sentiment analysis based on image

Image sentiment analysis started late, with relatively few methods, and image sentiment analysis is more complicated than text sentiment analysis[17–21]. In the paper[17], they used SentiWordNet thesaurus to extract numerical values from accompanying textual metadata, performed a discriminative feature analysis based on information-theoretic methods, and applied machine learning techniques to predict the sentiment of image. In

the paper[18], they constructed more than 3,000 Adjective Noun Pairs (ANP) and proposed SentiBank, a novel visual concept detector library that can be used to detect the presence of 1,200 ANPs in an image.

2.3 Sentiment analysis based on multimodal data

With the rapid development of social networks, the frequency of image data and text data appearing at the same time is increasing. Multimodal sentiment analysis[22–28] has begun to enter people's sight. Literature[22] proposes to use LSTM to encode graphic data separately and capture the internal information of the single-mode separately. Literature[24] proposes to use a pre-trained model to learn an image sentiment classification model FCNN, and the bidirectional LSTM trains the text sentiment classification model WBLSTM, and then the decision-level fusion is used to obtain the classification results. The network proposed in paper[26] consists of two attention layers and a bidirectional gated recurrent neural network (BiGRU).

The research of multimodal sentiment analysis is still in its infancy, with relatively few research methods and immature research results. Among them, how to fully learn the classification characteristics of each modal and establish a multimodal information fusion mechanism is a problem in current research[29]. Therefore, this paper uses character-level and word-level feature fusion as the original text representation, uses the attention mechanism to strengthen word contributions, uses gating functions to control the output of image vectors, and then uses each modal tensor as a joint of multimodal data for expression. The bidirectional LSTM is proposed to learn the data as a whole to enhance the multimodal information interaction. Finally, classify the vectors to obtain the emotional category of the multimodal data.

3. Methods

This paper proposes a fine-grained multimodal sentiment analysis model FCLAG based on gating and attention mechanism, which predicts the emotional state of users according to the text and image information posted on social platforms. The complete model framework is shown in **Figure 1.** Each sample is defined as $S(T_i, M_i)$, in which T_i refers to the text information of the i-th sample, and M_i refers to the image information of the i-th sample. First, the text is composed of character-level and word-level, which are respectively input to the text representation layer for vector feature representation; the image vector representation layer is learned and represented by the pre-trained model. The emotional characteristics of the text and the image are obtained respectively. The late-fusion strategy is adopted to obtain the multimodal joint feature representation. Finally, put the vectors into a fully connected layer to predict the label.

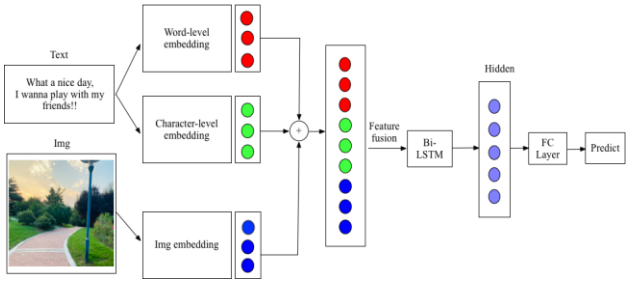


Figure 1. The structure of the FCLAG model.

3.1 Text feature extraction network

The model FCLAG extracts text features at the character level and text level at the same time. The text feature extraction network structure is shown in **Figure. 2.**

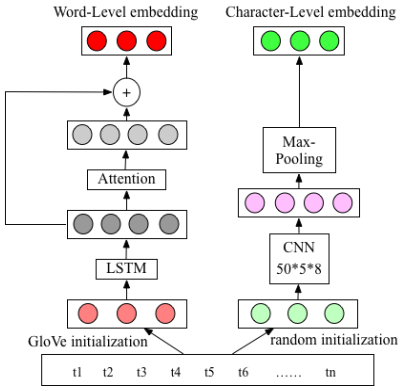


Figure 2. The structure of the text feature extractor.

The text with length n is represented as $T = (t_1, t_2, \ldots, t_n)$, each word t_i gets its vector representation w_i through the word vector embedding matrix GLoVe, and the word-level vector of the entire text is represented as $(w_1, w_2, \ldots, w_n)$. Put them into the LSTM network for context learning. The learning calculation process of the LSTM is as follows:

$$F_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$I_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$C't = \tanh(Wc \cdot [ht-1, xt] + bc) \quad (3)$$

$$C_t = F_t * C_{t-1} + I_i * C'_t \quad (4)$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = O_t * \tanh(C_t) \quad (6)$$

Where $[\cdot]$ denotes the multiplication of two matrices, $\sigma()$ represents the activation function of sigmoid, * represents the multiplication of corresponding vectors, W_f, W_i, W_c, Wo, b_f, b_i, b_c, and b_o are parameters to be learned; the values of F_t and I_t are 0-1, which determine the forgetting rate of the previous hidden layer and update rate of the candidate information; C_t represents the information of the current neuron, h_t represents the output of the text through the LSTM network.

The output of the hidden layer contains context information. Furthermore, in order to highlight the importance of keywords in the whole text, attention weighted calculation is carried out as follows:

$$Y_{it} = \tanh(W_a \cdot h_{it}) \quad (7)$$

$$A_{it} = \exp(Y_{it}{}^T C_w) / \sum \exp(Y_{it}{}^T C_w) \quad (8)$$

$$T_i = \sum A_{it} \cdot h_{it} \quad (9)$$

Where W_a and C_w are parameter vectors or matrices to be learned; the calculated attention weights are assigned to the corresponding words.

Character-level text representation can extract more fine-grained semantic features of the text. The model initializes the character-level vector randomly. The character-level vector in the text is represented as Z_i, $Z_i \in R^{n \cdot z \cdot d}$ where n is the number of words, z is the number of letters contained in the word w_i, and d is the letter dimension. Perform convolution operation on the character-level vector and calculated as follows:

$$C_i = f([W, Z_{i:i+h-1}] + b) \quad (10)$$

$$P_i = \max(c_1, c_2, \ldots, c_{n-h+1}) \quad (11)$$

Where W is the convolution kernel, h is the size of the convolution kernel, and b is the bias term. For the result of each convolution kernel $C_i = (c_1, c_2, \ldots, c_{n-h+1})$, we use maximum pooling to obtain the local optimal feature Pi.

The model extracts character-level text features from different levels by using multiple convolution kernels to convolve character-level vectors. Finally, character-level vectors are represented as Z_i, $Z_i = (P_1, P_2, .., P_n)$, n is the number of convolution kernels.

3.2 Image feature extraction network

According to the current dataset size, this model adopts the transfer learning method to improve the performance of the image classification task, which reduces the cost of learning and obtains more dense vector representation. First, perform data enhancement operation on the image, crop the image from the center into 244*244, and convert it into a vector form of[3, 244]. Then

put a batch of image data into ResNet18 for learning. Fine-tune and train the network by adding a dropout layer and a 50-dimensional fully connected layer. At this time, the image feature vector is expressed as $M_i = (m_1,m_2,...,m_n)$.

Since the image information has a lot of redundant information for the final sentiment detection, this information may lead to the wrong learning of the detection classification or other over-fitting behaviors. To solve this problem and improve the quality of the image classifier, a highway gate is added after fine-tuning to control the circulation of image information between layers, and the calculation formula is as follows:

$$H = M \cdot W_H \tag{12}$$

$$T = \sigma(M \cdot W_T) \tag{13}$$

$$M = H \cdot T + (1-T) \cdot M \tag{14}$$

Where W_H and W_T are parameter vectors to be learned, H represents the transformed data, and T represents the information flow probability.

The gate controls the training intensity of the model on the image. Some image features remain in their original state and are directly input to the next layer of the network, and the rest of the image features are transformed. The gating function of this layer controls information flow, which enhances the ability to protect the original data characteristics. Additionally, the output does not change the matrix size.

3.3 Multimodal data fusion

The text vectors obtained are represented as $W_i = (W_{i1},W_{i2},\ldots,W_{in})$, $Zi = (Z_{i1}, Z_{i2},\ldots, Z_{in})$, and the image vectors are represented as $M_i = \{M_{i1}, M_{i2},\ldots, M_{in}\}$. To further explore the connection between images and texts, making the following splicing:

$$S_i = W_i \oplus Z_i \oplus M_i \tag{15}$$

Where S_i represents the splicing results dimension, W_i represents the word dimension, Z_i represents the character dimension, and M_i represents the image dimension.

First, align the text and the image vectors according to the sequence number, and concatenate the vectors from the innermost dimension. At this time, the simple splicing of the vector cannot reflect the interaction between the two modes, so the vector is put into bidirectional LSTM. In the bidirectional LSTM, the hidden states obtained by forwarding LSTM and reversing LSTM respectively.

$$\overrightarrow{h_{it}} = \mathrm{LSTM}(\overrightarrow{h_{i(t-1)}}, S_{it}) \tag{16}$$

$$\overleftarrow{h_{it}} = \mathrm{LSTM}(\overleftarrow{h_{i(t-1)}}, S_{it}) \tag{17}$$

The combination $\overrightarrow{h_{it}}$ and $\overleftarrow{h_{it}}$ is $h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$ as the hidden state output at time t of data i.

At this time, the fusion vector not only includes the contextual information of the text but also includes various features extracted from the visual angle of the image. Moreover, the model learns the interaction between images and text. The vector output from this model contains various aspects of semantics and the characteristic sentiment.

3.4 Loss optimization

In each iteration of model training, the NLLLOSS() is used to calculate the error between the predicted value and the true value. The formula is as follows:

$$Loss(y, y') = -\sum_N \sum_C y \cdot y' \tag{18}$$

Where N represents the total number of samples, and C represents the number of categories. The loss function calculates the sum of the training losses of all samples in a batch, the optimizer uses Adam, the learning rate is set to 0.001, and the model parameters are optimized according to the loss value.

4. Experiments

In this section, we will introduce the source of the dataset, the dataset annotation standards, the design of each part of the parameter, the specific experimental process, the comparison results with the state-of-art models, and the verification of the validity of each part of the model.

4.1 Experimental data

The experimental data of this model include existing datasets on the Internet, and graphic information crawled from the Twitter social network platform and Weibo social network platform through crawlers.

First, perform basic data filtering on all data pairs, including deleting dirty data such as text length less than

3 and image information loss excessively; then manually label each item in the dataset, and keep the same number of text and image labels. They are marked by three groups of people. After processing, 1312 positive data pairs and 1039 negative data pairs are retained. In order to ensure a good fit for the training model, the dataset is divided proportionally, the number of training sets is 1971 and the number of validation sets is 380.

4.2 Text preprocessing

Since the text information contains a lot of dirty data, including punctuation, special symbols, tag links, etc., it is necessary to preprocess the text information first to remove the above content, so that the text is completely converted into a form composed of words.

4.3 Parameters setting

The word-level vector embedding adopts the GLoVe word vector model, the dimension is 100, the hidden layer output of the LSTM is set to 100 dimensions, and the characters are initially randomized to an 8-dimensional vector, subject to a uniform distribution of (-0.001, 0.001). The convolutional layer uses a 5*8 map, and the channel is set to 50 dimensions. In the fine-tuning process of image net, the fully connected layer is set to 50 dimensions, and the dropout is set to 0.3. The gating functions H(x), T(x) are all set to 50 dimensions; the hidden layer of bidirectional LSTM is set to 100 dimensions, and the number of classifications is 2.

The experimental results of this article are evaluated by accuracy and F1-score.

4.4 Experimental results

The model FCLAG is compared with the following advanced models:

(1) SVM[30]: proposed to use SVM classifier to learn the early generated multimodal vectors and complete classification;

(2) Random forest[13]: proposed to use the random forest to learn feature vectors and complete classification;

(3) FCNN + WBLSTM[24]: a pre-trained model is proposed to classify images FCNN, bidirectional LSTM extracts text context information to complete text classification WBLSTM, and finally uses decision level fusion to complete classification;

(4) MCNN[25]: CNN is proposed to learn the features of text and image respectively, and input into a fully connected layer to complete classification after splicing;

(5) CATF-LSTM[27]: proposed that after splicing multimodal vectors, the multi-layer attention mechanism is used to learn vector features, and the output vector of the last attention layer and the hidden layer vector of the previous layer are spliced together to input the full connection layer to complete the classification;

(6) FCLAG: the model proposed in this paper.

The results of multimodal sentimental analysis models are compared, as shown in **Table 1**:

Table 1. Experimental results of different model

	Accuracy	**F1-score**
SVM	0.6605	0.5609
RandomForest	0.6368	0.5655
FWB	0.8314	0.8169
MCNN	0.8530	0.8354
CATF-LSTM	0.8552	0.8374
Proposal model	0.8734	0.8512

Comparing the above experimental results, it can be seen that the method proposed in this article has achieved better results no matter in accuracy or F1-score. Compared with the results of traditional machine learning classification methods, the last four deep learning classification methods improve on accuracy by at least 20% and by at least 25% on F1-score. Both indicate that the deep learning method optimizing the parameters of each layer by calculating the prediction loss can better learn data characteristics and improve the accuracy of data classification in each category. FWB method uses decision-level classification after learning the respective vectors of images and texts. The model fusion formula is $P = \lambda P_t + (1-\lambda)P_m$, where λ =0.5, which makes the model consider that all modes have the same importance in the decision-making process. But in the social infor-

mation published by actual users, the text decision results and the image decision results have different importance. The latter three late fusion method classifications have proved this. The accuracy of these models has been improved to a certain extent. Among these, FCLAG proposed in this paper has the largest improvement, 4.2% on accuracy, and 3.43% on F1-score. This proves that the model using bidirectional LSTM to learn features at the fusion level can better learn the internal information and interaction information between modes.

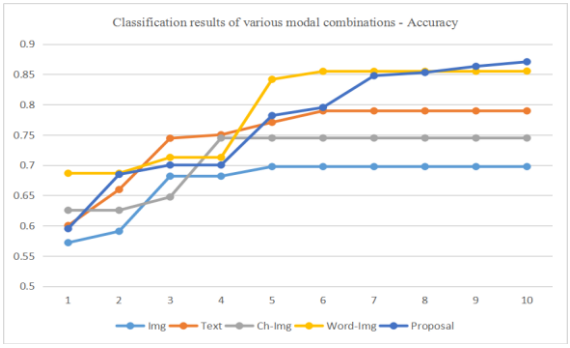
Compared with MCNN, the model proposed in this article has an increase of 2.04% on accuracy, an increase of 1.58% on F1-score, an increase of 1.82% on accuracy, and an increase of 1.38% on F1-score compared with CATF-LSTM, which proves that using both character and text level embedding, and gating function for controlling the output of image information flow based on CNN, not only can make the model pay attention to the context information but also learn the local optimal features and keep important information. These data all prove that FCLAG has more excellent performance of multimodal sentiment analysis.

4.5 Analysis of experimental results

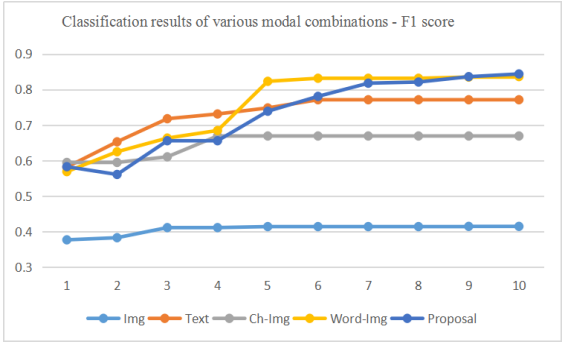
Multimodal fusion is used to consider multiple aspects of information, and integrate this information to complete sentimental classification. In order to intuitively reflect the effectiveness of multimodal information integration classification, the F1 value and accuracy value change line are drawn. The data is shown in **Table 2**, and the change of each epoch in the training process is shown in Figure. 3.

Table 2. Experimental results of the fusion of various modes

	Accuracy	**F1-score**
Img	0.6976	0.4151
Text	0.7896	0.7712
Ch-Img	0.7448	0.6693
Word-Img	0.8554	0.8363
Proposal	0.8707	0.8437



(a) Classification results of various modal combinations - Accuracy



(b) Classification results of various modal combinations - F1-score

Figure 3. Comparison of results of different combinations.

In the above chart, Img means that only image information is used for emotion classification, text means that only text information is used for emotion classification, Ch-Img means that character-level text representation and image information are fused for emotion classification, and Word-Img means that word-level text representation and image information are fused for sentiment classification. Comparing all classification results represented by broken lines, it can be concluded that the multimodal fusion model FCLAG proposed in this paper has been significantly improved both in accuracy and F1-score.

First, analyze the results of single-modal sentiment classification. Compared with the results of image classification, FCLAG has an increase of 17.31% on accuracy and an increase of 42.86% on F1-score. Compared with the results of text classification, FCLAG has an increase of 8.11% on accuracy and an increase of 7.25% on F1-score, which can prove that in various social information, text and image have different decision weights in determining the contribution of users' emotions. Image information is more abstract. Different users use different images to express their emotions. Online learning the image features are slow, so the classification accuracy is low, which will have a greater impact on the model classification results; the text information content is simple, to a certain extent, the model can learn the commonality of the data in the same category more quickly. However, the learning ability of the model in the later stage encounters a bottleneck, because the mation brought by a single-mode is limited, and the model cannot learn more useful information for classification under the current dataset.

Compared with Ch-Img and Word-Img, the accuracy of this model is increased by 12.59% and 1.53%, respectively, and the F1-score is increased by 17.44% and 0.74% respectively. Ch-Img has low indicators. However, compared with using images alone for emotion detection, there is an increase of 4.72% on accuracy and 25.42% on F1-score. This shows that whether the model separately combines character-level text or word-level text with image information for emotion detection, modal fusion will enhance the model's ability to learn favorable features for classification. Compared with the Ch-Img model, the result of the Word-Img model is improved by 11.06% on accuracy and 16.7% on F1-score, indicating that the coherence of word-level features in text expression is more conducive to learning features than character-level features. The emotional color carried by the words themselves has important decision-making properties in emotional classification.

At the same time, compared with the Word-Img model, the improvement in accuracy and F1-score of FCLAG proves that the integration of character-level representation can learn text context in a more granular manner. Gain, mine the useful information in the text.

This model uses the attention mechanism to assign different weights to words in the word-level text representation. The purpose is to learn the different contributions of different words in sentiment classification; at the same time, considering the presence of redundant information in the image, over-fitting problems that may occur in the learning process, this model adds a layer of gating mechanism by using the sigmoid function to set the transformation probability in the range of 0 to 1 for controlling the output of the pre-trained model, that is, whether to further learn the current vector. The experimental results are shown in **Table 3**.

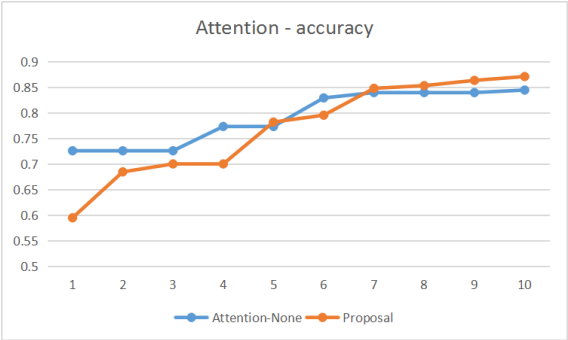
Table 3. Effectiveness verification results of attention and gate accuracy

	Accuracy	**F1-score**
Gate-None	0.7733	0.7453
Attention-None	0.8443	0.8282
Proposal	0.8707	0.8437

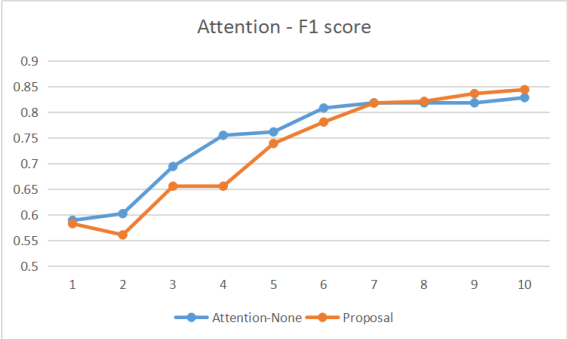
The results of the above table show that the attention mechanism contributes 2.64% to accuracy and 1.55% to F1-score, which proves that attention takes the different contributions of words in classification into account and assigns different weights to words, so its performance is significantly better than simple vector representation method.

The gating mechanism contributes 9.74% to the

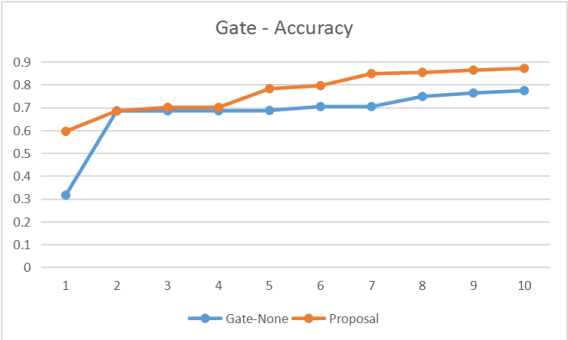
model's accuracy and 9.84% to the F1-score. When there is no gate, the training data achieves an accuracy of 98%, but the accuracy of the test data is always around 77%. It shows that there is an over-fitting phenomenon in network learning. The model proposed in this paper with gating mechanism not only ensures that the network can learn useful image feature vectors but also retains the original important information of the image. That is, it does not cause the wrong learning of the image vector due to the excessive number of network layers of nonlinear mapping, and reduces the risk of overfitting. So that the classification accuracy of the model is further improved. Figure. 4 visualizes the effectiveness of the attention mechanism and gating mechanism in the model learning process, quantified by accuracy and F1-score.



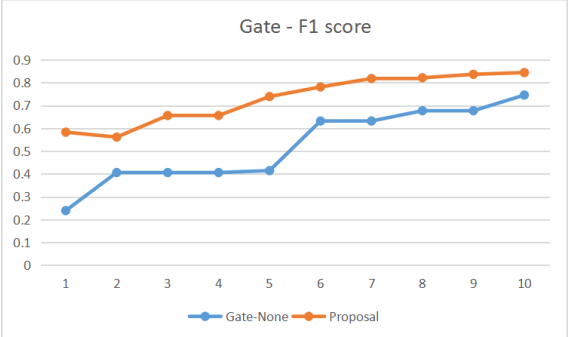
（a）Effect of attention mechanism on accuracy



（b）Effect of attention mechanism on F1-score



（c）Effect of gating mechanism on accuracy



（d）Effect of gating mechanism on F1-score

Figure 4. Learning process of FCLAG.

Among them, Attention-None means that the attention mechanism is not added, and Gate-None means that the gating mechanism is not added. From the above figures, we can learn more intuitively that whether it is in F1-score or accuracy, even if FCLAG with the attention mechanism is behind the model Attention-None learning effect in the early stage, by adjusting and optimizing the parameters of each layer in the later stage, the indicators can exceed the model Attention-None in the seventh epoch, indicating that the model with attention mechanism has a better ability to learn keywords. The FCLAG with the gating mechanism is significantly better than the Gate-None model at the initial training stage, indicating that the addition of the gate control mechanism makes the characteristics change within a certain probability. The model can learn richer image features and reduce the risk of overfitting. It also shows that improving the accuracy of image classification can clearly improve the accuracy of the whole model. In summary, whether it is a single-mode model or a combination of different modes, as well as the addition of various mechanisms, the model FCLAG proposed in this article has a better sentimental learning ability compared with the above advanced models.

5. Conclusion

This paper proposes a fine-grained multimodal sentiment analysis method based on gating and attention mechanism. First, the model learns each modal vector. Character-level and word-level are used to enhance text semantics. Use CNN convolutional neural network to extract character-level local optimal features, and use LSTM and attention mechanism to distribute different weights on word-level vectors. Both two ways representations enrich semantic information from multiple levels. Learn the image by transferring and fine-tuning. And FCLAG has a gating mechanism to control the output of the image stream. In the fusion stage, FCLAG adopts bidirectional LSTM for modal late-fusion. Bidirectional learning of vectors enhances the interactive learning ability between modalities. The experimental results show that the multimodal learning method proposed in this paper can obtain higher classification accuracy and F1 value. This also shows that when judging the user's social information emotional tendency, mining the interactive information between image and text can complete the sentiment detection task better.

However, there are still many aspects that need to be considered. First, how to accurately locate the important pixels in the image in order to obtain useful information for classification is a problem to be considered. Second, the attention mechanism has a limited range because it is calculated from a single aspect. In the future study, the above issues will be further researched.

References

1. Ye Q, Lin B, Li YJ. Sentiment classification for Chinese reviews: A comparison between SVM and semantic approaches. 2005 International Conference on Machine Learning and Cybernetics. IEEE 2005.
2. Montoyo A, Martinez-Barco P, Balahur A. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decision Support Systems 2012; 53(4): 675–679.
3. Rui W, Xing K, Jia Y. BOWL: Bag of word clusters text representation using word embeddings. Knowledge Science, Engineering and Management 2016; 3–14.
4. Wang H, He J, Zhang X, *et al*. A short text classification method based on N-Gram and CNN. Chinese Journal of Electronics 2020; 29(2): 248–254.
5. Yu Y, Chen K, Shou L, *et al*. Sentiment analysis of

user reviews based on keyword and key sentence extraction. Computer Science 2019; 46(10): 19–26.
6. Alessa A, Faezipour M, Alhassan Z. Text classification of flu-related tweets using fasttext with sentiment and keyword features. 2018 IEEE International Conference on Healthcare Informatics (ICHI). IEEE 2018.
7. Bai X, Chen F, Zhan S. A study on sentiment computing and classification of Sina Weibo with word2vec. 2014 IEEE International Congress on Big Data. IEEE 2014.
8. Alshari EM, Azman A, Doraisamy S, *et al*. Effective method for sentiment lexical dictionary enrichment based on Word2Vec for sentiment analysis. 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP). IEEE 2018.
9. Sharma Y, Agrawal G, Jain P, *et al*. Vector representation of words for sentiment analysis using GloVe. 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT). IEEE 2018.
10. Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019; 1: 380–385.
11. Gao Z, Feng A, Song X, *et al*. Target-dependent sentiment classification with BERT. IEEE Access 2019; 7: 154290–154299.
12. Socher R, Perelygin A, Wu J, *et al*. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing 2013; 1631–1642.
13. Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998; 20(8): 832–844.
14. Wang Y, Huang M, Zhu X, *et al*. Attention-based LSTM for aspect-level sentiment classification. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing 2016; 606–615.
15. Chen T, Xu R, He Y, *et al*. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications 2017; 72: 221–230.
16. Chen Y, Yuan J, You Q, *et al*. Twitter sentiment analysis via Bi-sense emoji embedding and attention-based LSTM. Proceedings of the 26th ACM International Conference on Multimedia 2018; 117–125.
17. Siersdorfer S, Minack E, Deng F, *et al*. Analyzing and predicting sentiment of images on the social web. Proceedings of the 18th ACM International Conference on Multimedia 2010; 715–718.
18. Borth D, Ji R, Chen T, *et al*. Large-scale visual sentiment ontology and detectors using adjective noun pairs. Proceedings of the 21st ACM International Conference on Multimedia 2013; 223–232.
19. Xu C, Cetintas S, Lee KC, *et al*. Visual sentiment prediction with deep convolutional neural networks.
20. Yang J, She D, Sun M, *et al*. Visual sentiment prediction based on automatic discovery of affective regions. IEEE Transactions on Multimedia 2018; 20(9): 2513–2525.
21. Srivastava RK, Greff K, Schmidhuber J. Highway networks. Computer Science 2015.
22. Liu Q, Zhang D, Wu L, *et al*. Multi-modal sentiment analysis with context-augmented LSTM. Computer Science 2019; 46(11): 181–185.
23. Lin M, Meng Z. Multimodal sentiment analysis based on attention neural network. Computer Science 2020, 47(11A): 508–514, 548.
24. Liao Y, Wang J, Liu T, *et al*. Joint visual-textual approach for microblog sentiment analysis. Computer Engineering and Design 2019; 40(4): 1099–1105.
25. Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing 2015; 2539–2544.
26. Kim T, Lee B. Multi-attention multimodal sentiment analysis. Proceedings of the 2020 International Conference on Multimedia Retrieval 2020; 436–441.
27. Poria S, Cambria E, Hazarika D, *et al*. Multi-level multiple attentions for contextual multimodal sentiment analysis. 2017 IEEE International Conference on Data Mining (ICDM) 2017.
28. Majumder N, Hazarika D, Gelbukh A, *et al*. Multi-modal sentiment analysis using hierarchical fusion with context modeling. Knowledge-Based Systems 2018; 161: 124–133.
29. Zhang Y, Rong L, Song D, *et al*. A survey on multimodal sentiment analysis. Pattern Recognition and Artificial Intelligence 2020; 33(5): 426–438.
30. Zadeh A, Zellers R, Pincus E, *et al*. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems 2016; 31(6): 82–88.