ISSN 2616-5775 Vol. 6, Issue 6: 121-126, DOI: 10.25236/AJCIS.2023.060619

Multimodal emotion recognition based on fusion with residual connection

Ziheng Gao[1,*], Haoxia Guo[2]

[1]Faculty of Science, Guilin University of Technology, Guilin, 541000, China
[2]Faculty of Mechatronic Engineering, Lanzhou University of Technology, Lanzhou, 730000, China
[*]Corresponding author: gzh2317026357@163.com

Abstract: *To improve the accuracy of emotional identification, a multi-mode fusion emotional recognition method of two-way long and short-term memory network (Bi-LSTM), Multi-Head Attention and Residual Connection blended emotional identification methods. This method performs long-term memory through LSTM, then uses the Attention mechanism to screen out important information, and finally improves the ability of network information transmission through the residual connection. Through this method of concentrated verification, the accuracy rate of emotional classification reaches 61.7%. The experimental results show that compared with models such as CNN, CMN, BC-LSTM, the accuracy and F1 value of the proposed model are effectively improved.*

Keywords: *Multi-modal Emotion Recognition, LSTM, Attention mechanism, Residual connection*

1. Introduction

With the rapid development of the information age, social media has also developed rapidly. More and more people have begun to upload their views to share with others through video, audio and text. This has led to a rapid growth in social media data. Faced with numerous social media data, emotional analysis [1] can tap the emotional semantic information from a large amount of data, predict people's views and attitudes of things. Emotional analysis is to use various digital quantitative information and analyze emotions in combination with computer and artificial intelligence technology. The main steps of emotional analysis include the extraction of emotional characteristics, establishing the mapping relationship between emotional characteristics and emotional labels, and predicting emotional predictions and accurately judging the emotional state based on this prediction [2]. Among them, the extraction of emotional characteristics involves multiple modes. Inspired by humans accepting emotional information, these emotional characteristics include emotional characteristics of various modes such as emotional characteristics, voice emotional characteristics, posture emotional characteristics, and text emotional characteristics [3].

With the development of the times, information becomes diversified, and more and more modular data in forms such as text, pictures, videos and other forms. Single mode cannot integrate analysis of these data. It is difficult to deal with the relationship among the modes by analyzing them one by one. In some cases, it is difficult to judge the emotional state of the target. The accuracy of analysis is often low. The TextCNN proposed by Zhang et al. TextCNN is a text emotional analysis model based on the convolutional neural network Convolutional Neural Networks (CNNs). This model network structure is simple and calculates the emotional relationship between text through the introduction of phrase vector. However, this model is difficult to evaluate the important degree of each emotional characteristics, it is easy to ignore the important emotional characteristics in the text, resulting in low accuracy. AbdelwahabM [5] and others use deep learning networks to understand the emotional information contained in human voice signals. However, in view of the concealment and complexity of the emotional expression, the accuracy of the calculation is often large.

The above-mentioned emotional analysis research mostly adopts the integration of "Serial connection", and simply integrates different modal emotional characteristics. When the emotional expression is hidden and complex, this fusion method cannot handle the relationship between the modes, capture the poor accuracy of important emotional information, and cause information redundancy. In response to the above problems, this article proposes a multi-mode emotional identification model that combines Bi-LSTM and multi-head, it can solve the problem of low emotional recognition accuracy of a single modal, and can well extract the above semantic information. In addition, to solve the problem of information collapse, we have also introduced the idea of residual connection, which can improve the generalization of models.

2. Related Work

With the advancement of information technology, artificial intelligence gets more attention and human-computer interaction becomes more intelligent. Therefore, in order to study artificial intelligence in deeper levels, computer recognition and analysis of natural language and emotional information in it need to be analyzed. Information becomes diversified, and modal data in the form of text, pictures, and videos is becoming more and more. Single modes cannot integrate analysis of these data, analyzing the relationship between each modal state is difficult to analyze one by one, and the accuracy of analysis is often low. Gradually, multimodal emotion analysis has become a hot research topic.

Yoon et al. [6] use two Bi-LSTM network encoding analysis to analyze the interaction between the modal. Initially improve the accuracy of identification. Chen et al. [7] proposed a multi-modal emotional recognition algorithm based on the DR-Transformer model, which solved the problem that a single text modal information could not accurately determine the emotional polarity. Calculate through network or model and integrate calculation results, which greatly improves recognition accuracy. Wang et al. [8] responses to the problem of low-modal emotional recognition accuracy, they proposed a dual-mode emotional identification model algorithm based on Bi-LSTM-CNN. The accuracy of identification is greatly improved. Wu Peng et al. [4] proposed a model combined with emotional word vector and Bi-LSTMs. They recognize the negative emotion of netizens, and got a good effect by using bi-directional long-term and short-term memory model. Although the existing research solves the problem of modal emotion recognition well, they neglect the problem of information collapse. Therefore, to improve the generalization of the model, we introduce the idea of residual connection, thereby increasing the accuracy of emotional recognition.

3. Preliminary

3.1 Problem Definition

The task of this paper is to identify the emotional changes of the speaker during the conversation. We represent N speakers as $x_1, x_2, \dots, x_N$. U is used to represent the set of contextual statements that the speaker says in the conversation. $U = \{u_1, u_2, \dots, u_M\}$, M indicates that there are several sentences in the context of a conversation. $\alpha_p(p \in 1,2,\dots,M)$ is the set of contextual statements with emotional labels, represents a statement of N speaker. The collection U is represented as $U_1 \cup U_2 \cup \dots \cup U_N$, U_j is the β_j statement. $j \in (1,2,\dots,N)$, $\beta_p = (S_p^1, S_p^2, \dots S_p^{lp})$ is the total number of statements by a speaker in chronological order. Among them, S_p^i is the ith sentence of μ_p. l_p is the total number of words spoken by Speaker μ_p, $p \in \{1,2,\dots,N\}$.

This paper tries to get the emotional label of the speaker's current statement in the end. In a time period t, $t \in [1, M]$. We want to get the speaker's emotions at t. The context statements for $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{N}$ are $\mathrm{Y}_1, \mathrm{Y}_2, \dots, \mathrm{Y}_\mathrm{N}$. Set H to the dialog history context window size.

$$Y_\lambda = \{u_i | i \in [t-H, t-1], u_i \in U_\lambda, |Y_\lambda| \leq H\} \tag{1}$$

3.2 Audio Feature Extraction

In this paper, the feature extraction of sound signal is based on the depth learning method of time domain signal. This method uses one-dimensional vector to represent the original signal and network input. Then extracts automatically from the sound signal by encoder architecture. The encoder used is four layers Bi-LSTM and four layers linear. In order to analyze the importance of the sound signal, we use the attention mechanism to assign attention scores. Finally, the final feature vector is obtained by weighted average.

3.3 Visual Feature Extraction

The best indicators of a person's emotional state is a change in his or her facial expression. Therefore, this paper uses 3D-CNN to extract the speaker's facial features. Furthermore, the comprehension ability of the model to the utterance emotion is improved. 3D-CNN captures the subtle changes in facial expressions by extracting the deep meaning of human facial expressions. This network consists of convolution layer, pooling layer, and fully connected layer. We input the video frame vector to CNN, then convolve and pool it, and use the ReLU function to change it non-linearly. Finally, the obtained feature vector obtains a high -dimensional vector after the full connection layer, and this vector is used as a visual feature.

3.4 Word Embedding Layer

Since the computer cannot process the characters directly, this article is preprocessed first. Through the tokenizer method, the discourse text is divided into words, and obtain the mapping relationship between words and indexes. Then we input the data into Roberta's pre-training model to get a 300-dimension word embedding vector.

4. Methodology

4.1 Sequential context semantic information modeling

This paper uses Bi-LSTM to model the discourse context. Discourse text is composed of word sequences arranged in a specific order. Bi-LSTM is composed of LSTM stitching with positive and reverse order. Therefore, we can better capture two-way context semantic information. As shown in Fig. 1, LSTM consists of the input words of time, cell state, temporary cell state, hidden layer state, forgotten gate, memory gate and output gate.

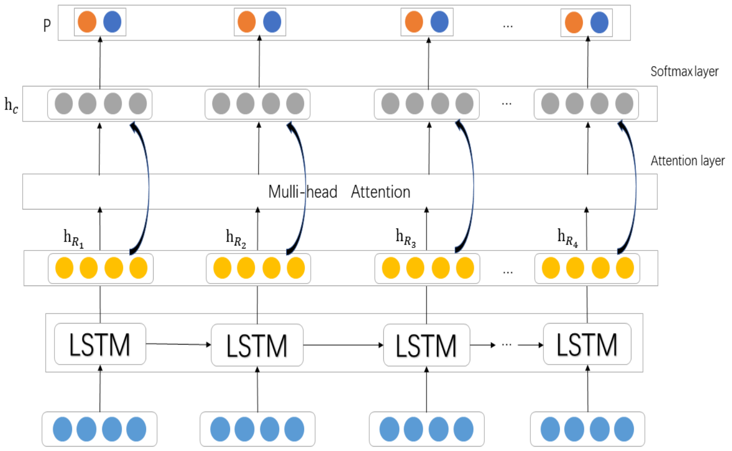


Figure 1: The network architecture diagram.

Calculation:

Calculate the Forgotten Gate: select the information to be forgotten, input the hidden state h_{t-1} at the previous moment and the current moment of the input word x_t, output the value of the Forgotten Gate f_t.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2}$$

Computational Memory Gate: select the information to be remembered, input the previous hidden layer state h_{t-1} and the current moment of the input word x_t, output value of memory gate i_t and temporary cell state $\widetilde{C}_t$.

$$i_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

$$\widetilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

Current cell state: Input Memory Gate i_t, Forgotten Gate f_t, and temporary cell state $\widetilde{C}_t$, and the last time cell state C_{t-1}, the output is the current time cell state C_t.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{5}$$

The output gate and the current state of the hidden layer are calculated as follows: the state h_{t-1} of the hidden layer at the previous time, the input word x_t at the current time and the current state C_t at the current time, and the output value o_t of the output gate and the state h_t of the hidden layer.

$$o_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_o\right) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

Finally, the hidden layer state sequence of the same length as the sentence is obtained $\{h_1, h_2, \dots, h_{n-1}\}$.

4.2 Attention layer

The basic idea of Attention mechanism is to break the traditional encoder-decoder structure in the encoding and decoding are dependent on the internal fixed-length vector constraints. It consists of query, value, and key. For short q, k, v. For feature U, using $[q, k, v] = U[X^q, X^k, X^v]$ to calculate q, k, v three matrices $q \in R^{T_q * i_q}$, $k \in R^{T_k * i_k}$ and $v \in R^{T_v * i_v}$. T_q, T_k, T_v represent the sequence length. q, k, v. i_q, i_k, i_v represent the dimensions of q, k, v.

The self-attention formula is defined as follows:

$$V = softmax\left(\frac{qk^t}{\sqrt{e_k}}\right)v \tag{8}$$

Where V represents the weight of value and e_k represents the feature dimension of value, and gets multiple attention:

$$M(A) = (A_1, \dots, A_h)X \tag{9}$$

Where $A_1, \dots, A_h$ represents the attention layer, h represents the number of layers, X represents a weight parameter.

4.3 Residual connection

The idea of residual connection originates from centralization. In a neural network system, the input data is transformed centrally, that is, the data is subtracted from the mean value. We use a residual connection layer with normalization to improve the information transfer capability of the model. The formula is defined as follows:

$$H = Norm\left(F + MultiHead(F)\right) \tag{10}$$

Where H represents the predicted value and F represents the feature extracted by Bi-LSTM.

5. Experimental Setting

5.1 Implementation Details

All the research in this paper is carried out in Nvidia 3060TI server with a total memory capacity of 8 GB. The programming language used for the experiment is Python 3.8.10, and the deep learning framework is PyTorch 1.13.1. Using Adam [10] as an optimization algorithm, the batch size was set to 32, the number of iterations was 60, the initial learning rate was 3e-4, and the attenuation coefficient of L2 weight was 5e-5.

5.2 Datasets Used

In this paper, IEMOCAP is used to test the algorithm model, and based on this data set to evaluate the model of good or bad degree. The classification of validation sets, training sets, and test sets in the data set, and the model evaluation indicators are shown in the following Table 1:

IEMOCAP: this dataset contains videos of binary conversational relationships with ten interlocutors, five of whom are male and five of whom are female. These video dialogue relationships are divided into five stages. Men and women dialogue are allocated at each stage. Each dialogue contains an emotional label (happy, neutral, sad, angry, frustrated or excited). This article uses the first four stages as a training set and a validation set, and the fifth stage as a test set.

Table 1: Breakdowns in the IEMOCAP dataset as well as the number of emotion categories and assessment metrics

Datasets	Utterance Count			Dialogue Count			Classes	Evaluation Metrics
	Train	Validation	Test	Train	Validation	Test		
IEMOCAP	5320	490	1623	108	12	31	6	Accurancy/f1

5.3 Baselines and State of the Art

CNN: Kim et al. proposed that CNN uses convolution kernels to extract the global semantic information of text, but its context modeling capability is poor.

bc-LSTM: Poria et al. proposed a bidirectional LSTM that captures contextual information in different directions, but it does not take into account the location relationship between the speaker and the conversation context.

CMN: Hazarika et al. proposed that CMN uses GRU to get many context semantic vector representations, which can be input into the memory network, and then the long-term context information can be modeled.

6. Results and Discussion

6.1 Comparison with State of Art and Baseline

Our model is compared with the baseline model. According to the experimental results, the performance of the proposed model is improved compared with the current work.

As shown in Table 2, in IEMOCAP dataset, our model accuracy rate is 61.7%, increased by 12.8% than CNN, increased by 6.5% than bc-LSTM, increased by 5.2% than CMN. Our model f1 is 61.0%, increased by 12.9% than CNN, increased by 6.1% than bc-LSTM, increased by 4.9% than CMN.

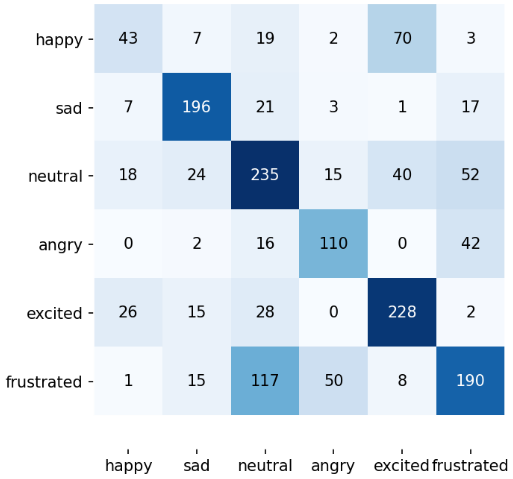
Compared with other baseline models, the proposed model achieves a certain degree of performance improvement under the IEMPCAP data set. The main reason is that our model combined with bidirectional LSTM, can extract positive. And negative semantic information, integrate attention mechanism, capture key semantic information. Then add residual connection, improve the ability of network information transmission.

Table 2: Comparison of model accuracy under the IEMPCAP dataset with other baseline models

Methods	IEMPCAP						
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Average
	Acc .F1	Acc .F1	Acc .F1	Acc .F1	Acc .F1	Acc .F1	Acc .F1
CNN	27.7 29.8	57.1 53.8	34.3 40.1	61.1 52.4	46.1 50.0	62.9 55.7	48.9 48.1
bc-LSTM	29.1 34.4	57.1 60.8	54.1 51.8	57.0 56.7	51.1 57.9	67.1 58.9	55.2 54.9
CMN	25.0 30.3	55.9 62.4	52.8 52.3	61.7 59.8	55.5 60.2	71. 1 60.6	56.5 56.1
Our Model	45.3 36.0	75.7 77.8	53.9 57.3	61.1 62.9	65.7 70.6	62.1 55.3	61.7 61.0

6.2 Analysis of the Experimental Result

By classifying the labels predicted by the model, the obfuscation matrix of the data set is obtained as shown in Figure 2. By analyzing the confusion matrix, we found that our model will mistakenly classify the "happy" class as "excited" class, classify the "frustrated" class as "neutral" class, and classify the "sad" class as "neutral" class. We think this is because there are fewer differences between these emotions. By expanding the size of the data set, we believe that the model can learn the differences between the two, and then get accurate results.



	happy	sad	neutral	angry	excited	frustrated
happy	43	7	19	2	70	3
sad	7	196	21	3	1	17
neutral	18	24	235	15	40	52
angry	0	2	16	110	0	42
excited	26	15	28	0	228	2
frustrated	1	15	117	50	8	190

Figure 2: Confusion matrix in IEMOCAP dataset

7. Conclusion

This article applies the above method to IEMOCAP dataset. First, after processing data pre-processing and entering BI-LSTM, incorporate the Attention mechanism and residual connection for emotional recognition. In the end, the difference between the accuracy of emotional recognition and the F1 score is displayed. The experimental results show that compared with CNN, CMN, BC-LSTM and other models. The accuracy rate and F1 score of this model are effectively improved. At the same time, the rationality of this model is tested by analyzing the confusion matrix.

References

[1] Xu Heng, Zhang Menglu, Zhong Zhen. The present situation of research in the field of comment mining and affective analysis in our country [J]. Journal of Jilin Normal University science, 2020. 41(03): 51-61.

[2] Chen Caihua. Tri-modal Mandarin emotion recognition based on speech, expression and gesture [J]. Control Engineering, 2020, 27(11): 2023-2029.

[3] Ying R K, Shou Y, Liu C. Prediction Model of Dow Jones Index Based on LSTM-Adaboost[C]//2021 International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE, 2021: 808-812.

[4] Shou Y, Meng T, Ai W, et al. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis[J]. Neurocomputing, 2022, 501: 629-639.

[5] Meng T, Shou Y, Ai W, et al. A Multi-Message Passing Framework Based on Heterogeneous Graphs in Conversational Emotion Recognition [J]. Available at SSRN 4353605, 2021.

[6] Yoon S, Byun S, K Jung. Multimodal speech emotion recognition using audio and text [J]. 2018 IEEE Spoken Language Technology Workshop (SLT), 2018: 112-118.

[7] Xie Z, Guan L. Multimodal Information Fusion of Audio Emotion Recognition Based on Kernel Entropy Component Analysis[J]. international journal of semantic computing, 2012.

[8] Wang Lanxin, Wang Weiya, Cheng Xin. Bi-modal emotion recognition model [J] based on Bi-LSTM-CNN. Computer Engineering and applications, 2022, 58(4): 192-197.

[9] Fu Z, Liu F, Wang H, et al. A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition[J]. 2021.

[10] Shou Y, Meng T, Ai W, et al. Object Detection in Medical Images Based on Hierarchical Transformer and Mask Mechanism[J]. Computational Intelligence and Neuroscience, 2022.