ISSN 2616-5775 Vol. 6, Issue 6: 127-132, DOI: 10.25236/AJCIS.2023.060620

Multimodal sentiment recognition based on Bi-LSTM and fusion mechanism

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

[1]Lanzhou University of Technology, Faculty of Mechatronic Engineering, Lanzhou, China, 73000
[2]Guilin University of Technology, Faculty of Science, Guilin, China, 541000
**Corresponding author: ghx18330885019@163.com*

Abstract: *The research of multimodal emotion recognition has important application value in artificial intelligence, human-computer interaction and other fields. With the development of deep learning, emotion recognition has been paid more and more attention by researchers. Existing research has solved the problem of unimodal emotion recognition, but neglected the research on the combination of bidirectional long and short neural networks and attention mechanisms. Based on this, we propose an emotion recognition model based on Bi-LSTM and multi-head attention mechanism, which combines the characteristics of LSTM for long-term memory and the advantage that the attention mechanism can quickly screen out more important information among many information, and further improves the accuracy of multimodal emotion recognition. Experimental results show that compared with CNN, CMN, BC-LSTM and other models, this model has better accuracy and f1-score.*

Keywords: *Attention mechanism, LSTM, multimodal emotion recognition*

1. Introduction

People usually express their emotions through the three modal characteristics of text, video, and audio. Therefore, how to design a multimodal algorithm is a challenging task. In addition, multimodal emotion recognition tasks have been widely used in the field of safe driving and intelligent recommendation. In many aspects of supervised learning, human emotions play a key role, but human emotions are difficult to identify. For example, human driving safety detection, voice expressions and other factors will affect safe driving. Combined with multimodal emotion recognition can make the recognition of human emotions more accurate. Multimodal emotion recognition is to extract multimodal emotion information through machines, establish models and make predictions, and finally reflect human emotions.

In the traditional emotion recognition method, the emotional information contained in a single pattern is limited. Husam [1] et al. only use acoustic information to build an emotional recognition model due to technical limitations, which lacks important information such as semantics, and cannot combine many factors. Baltrušaitis [2] et al. divides the research of multimodal emotion recognition into modal representation, alignment, fusion, and cooperative learning, and most of the current research focuses on the level of modal fusion. Bahdanau [3] et al. first applied attention mechanisms to the activities of machine translation. The main idea is to calculate the weights of the feature vectors and sum them, reflect the importance of different features through different weight numbers, and give greater weight to more important features. However, traditional emotion recognition models require a large number of features, which increases the training parameters and leads to the omission of some more critical information. In addition, the attention mechanism cannot learn the sequential relationships in sequences.

However, most of the existing multimodal emotion recognition research extracts more emotional features for classification, but important features may be overwritten by unimportant features, making key features missing, and resulting in models unable to quickly extract key features. With the gradual application of neural network models to multimodal emotion recognition, RNN models play an important role in processing sequence tasks. As a simple recurrent neural network, RNN also has the problem of gradient disappearance as the number of neural network layers increases, and can only be short-term memory. To solve these problems, we propose an emotion recognition model based on LSTM [4] and the multi-head attention mechanism, which combines the characteristics of LSTM for long-term memory and the advantage that the attention mechanism can quickly screen out more important information among many information, and further improves the accuracy of multimodal emotion recognition.

ISSN 2616-5775 Vol. 6, Issue 6: 127-132, DOI: 10.25236/AJCIS.2023.060620

2. Related Work

With the development of deep learning, the research of emotion recognition has attracted more and more attention from researchers and has become a key research object in the field of human-computer interaction. However, extracting useful contextual information from video, audio, and text features remains a challenging task.

In a recent study, Liu et al. [5] proposed LSTM-based multimodal emotion recognition, which uses the Bi-LSTM model to recognize speech and facial expressions respectively, overcoming the problem that the recognition accuracy of a single modal model depends on the emotion type. Hazarika [6] and others proposed a feature fusion method based on self-attention, encoding the text, and then using CNN for feature extraction, which effectively solved the problem of missing key features caused by too many features. Huang et al. [7] proposed a transformer model, in which the multi-attention mechanism generates multimodal emotional intermediate representations of the semantic feature space, combines the transformer with the LSTM model, and obtains the results through the fully connected layer, which further improves the model performance. Wang Lanxin et al. [8] combined Bi-LSTM and CNN with affixes to propose a Bi-LSTM-CNN model and performed multimodal emotion recognition, which improved the recognition accuracy. Although the existing research has solved the problem of modal emotion recognition, it ignores the study of the combination of bidirectional long and short neural networks and attention mechanisms. Therefore, this paper synthesizes the characteristics of LSTM for long-term memory and the advantages of an attention mechanism that can extract more important information, so as to improve the accuracy of emotion recognition.

3. Preliminary

3.1 Problem Definition

Since the main research content of this article is the mood changes before the multiple interlocutors, it is used $\alpha_1, \alpha_2, \dots, \alpha_P$ to represent the interchanger of the conversation, where P is the number of interlocutors, defines the interlocutor's set of contextual statements in the conversation $M, M = \{M = M_1, M_2, \dots, M_N\}$, There are several sentences that will be represented N as a conversation context [9]. $L_\lambda(\lambda \in 1, 2, \dots, N)$ is the context of the conversation with emotional labels, a statement that represents one of the interlocutors. Gather M can be expressed as $M_1 \cup M_2 \cup \dots \cup M_P$, and α_j is the discourses of M_j, $j \in (1, 2, \dots, P), \beta_\lambda = \left(W_\lambda^1, W_\lambda^2, \dots W_\lambda^{l_\lambda}\right)$ ss the number of statements of the last interlocutor sorted by time. Where, W_λ^i Indicates the η_λ first sentence spoken i, l_λ is the total number of conversations for the interlocutor $\eta_\lambda, \lambda \in \{1, 2, \dots P\}$.

The main research objective of this paper is to get the emotional label of the interlocutor's current conversation, which we need to get over a period of time $t, t \in [1, I], Y_1, Y_2, \dots Y_P$ is the context statement for $\alpha_1, \alpha_2, \dots, \alpha_P, G$ is the size of the history context window, A collection of contextual statements for the first interlocutor λ.The context history statement formula of M_λ is:

$$Q_\lambda = \{L_\lambda | i \in [t - G, t - 1], L_\lambda \in M_\lambda, |Y_\lambda| \leq G\} \quad (1)$$

3.2 Audio Feature Extraction

We use time-domain signals to extract feature information through end-to-end deep learning methods, and use an encoder architecture consisting of four Bi-LSTMs and four linear layers to automatically extract features from sound signals. In the encoder input process, it is divided into multi-segment sound signals, each sound signal is processed to obtain different output results, and attention is used to give different attention scores to sound signals with different degrees of importance, and finally weighted to obtain the final feature vector [10].

3.3 Visual Feature Extraction

Human facial expression changes are the best way to reflect human emotional conditions in video extraction features, so we use 3D-CNN to extract the facial features of the interlocutor, 3D-CNN can extract the deep characteristics of human expressions, accurately capture the subtle changes of human expressions to improve the model's ability to understand the emotions of dialogue sentences. 3D-CNN consists of three parts: convolutional layer, pooling layer and fully connected layer. We first input the

ISSN 2616-5775 Vol. 6, Issue 6: 127-132, DOI: 10.25236/AJCIS.2023.060620

video frame vector into CNN, then perform convolution and pooling operations, and use the ReLU function to perform nonlinear changes, and finally input the obtained feature vector to the fully connected layer to obtain a high-dimensional vector, which is the extracted visual features.

3.4 Word Embedding Layer

To obtain the feature vector containing the context semantic information, we preprocess the discourse text, use the Tokenizer method to segment the text to obtain the mapping relationship between the word and the index, and input it to the Roberta pre-training model to adjust and obtain a 300-dimensional word embedding vector.

4. Methodology

4.1 Bi-LSTM Layer

Words have a sequential order in the sentence, and Bi-LSTM is composed of forward LSTM and backward LSTM, which can better model and capture bidirectional semantic information. Therefore, we use the Bi-LSTM model to process the context semantic information. The LSTM model is composed of the input word of the moment, the cell state, the temporary cell state, the hidden layer state, the forgetting gate, the memory gate, and the output gate, and the working principle is as shown in Fig. 1:

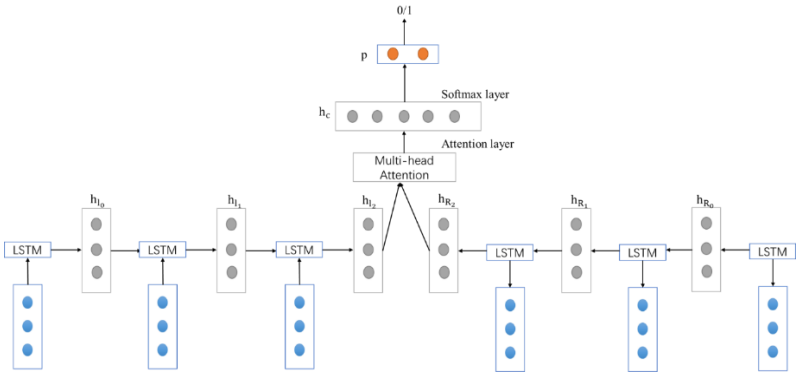

Figure 1: Bi-LSTM model diagram

Here's how it works:

Forget Gate: Select the past information, enter the hidden layer state of the previous moment h_{t-1} and the input word of the current moment x_t, and output the value of the forget gatef_t.

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

Memory gate: Select the information of memory, enter the hidden state of the previous moment h_{t-1} and the input word of the current moment x_t, and output the value of the memory gate i_t and the temporary cell state c_t.

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

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

Current cell state: Enter memory gate, forget gate and temporary cell state, and the previous moment cell state C_{t-1}, and output is the current moment cell state C_t.

$$c_t = f_t * C_{t-1} + i_t * c_t \tag{5}$$

Output gate and current hidden layer state: Enter the previous moment hidden layer state, the current moment input word x_t and the current cell state C_t and the output is the output gate value o_t and hidden layer state h_t.

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

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

Finally, we get the hidden layer state sequence $\{h_1, h_2, \ldots, h_{n-1}\}$.

ISSN 2616-5775 Vol. 6, Issue 6: 127-132, DOI: 10.25236/AJCIS.2023.060620

4.2 Attention layer

The essential idea of attention is to selectively filter out a small amount of important information from a large number of information and focus on these important information, ignoring unimportant information. The attention mechanism usually consists of three parts: Query, Key and Value, which we refer to as short. For one feature U, we pass $[q,k,v] = U[W^q, W^k, W^v]$ to calculate out three matrices about q,k,v.They are $q \in R^{B_q * e_q}, k \in R^{B_k * e_k}$ and $v \in R^{B_v * e_v}$. Where B_q, B_k, B_v represents the length of the sequence q,k,v. The dimensions of q,k,v are e_q, e_k, e_v. The expression formula for self-attention is:

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

A is the weight of value, and the dimension of the feature is. In this way, multiple self-attention layers are connected to get the multi-attention layer:

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

The output is represented by $A_1, \dots, A_h$,the attention layer, h representing the number of layers and representing the weight parameter.

5. Experimental Setting

5.1 Implementation Details

The experiment in this paper is carried out on NVIDIA's 1660Ti, memory capacity of 16G server, the experimental programming language is Python 3.8, the deep learning framework is Pytorch 1.8.1.

5.2 Datasets Used

The benchmark dataset for emotional dialogue recognition for the experiment of the algorithm is IEMOCAP, and the advantages and disadvantages of the algorithm are evaluated by this data set. The following table shows the division and model evaluation metrics of the validation, training, and test sets in the dataset as shown in Table 1:

Table 1: Classification and evaluation indicators of the IEMOCAP dataset

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

IEMOCAP: This dataset is a database collected by laboratories at the University of Southern California. Contains information from ten interlocutors, videos of binary relationship conversations for five men and five women. The videos are divided into five stages, each with a male and a female dialogue, and each sentence contains the label of happiness, neutrality, sadness, anger, frustration, or agitation. The first four phases serve as the training and validation sets, and the fifth phase is the test set.

5.3 Baselines and State of the Art

CNN: Kim et al. propose a CNN for a baseline model of conversational text classification, but it cannot use multimodal data and does not resolve inter-interlocutor or sentence context dependencies.

bc-L STM: Poria et al. propose a two-way LSTM combining forward LSTM and backward LSTM, which can capture semantic information from the historical context of the interlocutor and the current dialogue sentence, but it does not take into account the position relationship between the interlocutor and the context and the interaction relationship of the interlocutor.

CMN: The CMN proposed by Hazarika et al. uses GRU to obtain a rich contextual semantic vector representation and input it into its memory network. This allows the model to process contextual semantic features for a long time, but it does not take into account the detection of emotional changes in multiple interlocutors.

ISSN 2616-5775 Vol. 6, Issue 6: 127-132, DOI: 10.25236/AJCIS.2023.060620

6. Results and Discussion

6.1 Comparison with State of Art and Baseline

By comparing the model established in this paper with the existing baseline model, according to the experimental results, we can find that the model established in this paper has a certain performance improvement compared with the current existing research.

Table 2: Comparison of model accuracy under the IEMPCAP dataset

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	43.4 39.7	87.2 67.9	54.5 58.3	78.2 58.3	66.7 63.2	59.4 53.6	61.9 57.2

As shown in Table 2, in the IEMPCAP dataset, our model accuracy is 61.9%, which is 13% better than CNN, 6.7% better than bc-LSTM, and 5.4% better than CMN. The value of f1 in this paper is 57.2%, which is 9.1% better than CNN, 2.3% better than bc-LSTM, and 1.1% better than CMN.

Compared with other models, the accuracy and f1 value of the proposed model under the IEMPCAP dataset have been improved, the main reason is that the model construction is different, and the proposed model combines LSTM and attention mechanism, integrates the characteristics of LSTM for long-term memory and the advantage that the attention mechanism can quickly screen out more important information among many information, and further improves the accuracy of multimodal emotion recognition.

6.2 Analysis of the Experimental Result

We classify the labels predicted by the model to obtain the confusion matrix of the dataset as shown in Fig. 2, and after analysis, we find that the model will mistakenly classify the "happy" class as the "excited" class, the "frustrated" class as the "neutral" class, and the "sad" as "neutral", which may be caused by the small difference in these emotional performances. Combined with larger datasets, we think the model can learn the differences in confusing labels to get the right results.

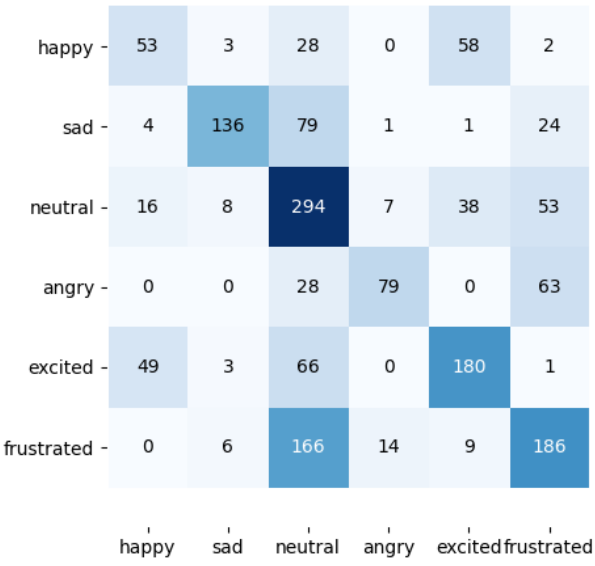

Figure 2: Confusion matrix

ISSN 2616-5775 Vol. 6, Issue 6: 127-132, DOI: 10.25236/AJCIS.2023.060620

7. Conclusion

On the IEMOCAP dataset, this paper first preprocesses the data, and then uses a model combining Bi-LSTM and attention mechanism for experimentation. Finally, we compare the f1 value and accuracy with the existing models as evaluation indicators, and the results show that compared with CNN, CMN, BC-LSTM and other models, our model has better accuracy and f1-score, and the rationality of the model is judged by analyzing the confusion matrix.

References

[1] 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.

[2] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423–443

[3] 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.

[4] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780. [doi: 10.1162/neco. 1997.9.8.1735]

[5] Liu Jingjing, Wu Xiaofeng. Multimodal emotion recognition and spatial annotation based on long short-term memory network [J]. Fudan Journal (Natural Science Edition), 2020, 59(5): 565-574.

[6] Hazarika D, Gorantla S, Poria S, Et Al. Self-attentive feature-level fusion for multimodal emotion detec tion[C]//Proceedings of the IEEE 1st Conference on Multi-media Information Processing and Retrieval, Miami, Apr 10-12, 2018. Piscataway: IEEE, 2018: 196-201.

[7] Huang J, Tao J H, Liu B, et al. Multimodal transfo rmer fusion for continuous emotion recognition [C]//Procee-dings of the IEEE 2020 International Conference on Acou stics, Speech and Signal Processing, Barcelona, May 4-8, 2020. Piscataway: IEEE, 2020: 3507-3511.

[8] Wang Lanxin, Wang Weiya, Cheng Xin. Bi-LSTM-CNN dual-modal emotion recognition model for speech and text [J]. Computer Engineering and Applications, 2022, 58(4): 192-197.)

[9] 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.

[10] 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.