Complex & Intelligent Systems (2023) 9:951–963
https://doi.org/10.1007/s40747-022-00841-3

ORIGINAL ARTICLE

A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context

Yurui Xu[1] · Hang Su[2] · Guijin Ma[1] · Xiaorui Liu[1]

Received: 6 August 2021 / Accepted: 25 July 2022 / Published online: 18 August 2022

Abstract

With regard to human–machine interaction, accurate emotion recognition is a challenging problem. In this paper, efforts were taken to explore the possibility to complete the feature abstraction and fusion by the homogeneous network component, and propose a dual-modal emotion recognition framework that is composed of a parallel convolution (Pconv) module and attention-based bidirectional long short-term memory (BLSTM) module. The Pconv module employs parallel methods to extract multidimensional social features and provides more effective representation capacity. Attention-based BLSTM module is utilized to strengthen key information extraction and maintain the relevance between information. Experiments conducted on the CH-SIMS dataset indicate that the recognition accuracy reaches 74.70% on audio data and 77.13% on text, while the accuracy of the dual-modal fusion model reaches 90.02%. Through experiments it proves the feasibility to process heterogeneous information within homogeneous network component, and demonstrates that attention-based BLSTM module would achieve best coordination with the feature fusion realized by Pconv module. This can give great flexibility for the modality expansion and architecture design.

Keywords Emotion recognition · Dual-modal · Pconv · BLSTM

Introduction

Emotion recognition is an important part of the interaction between people and machines [1], it has broad application prospects in the fields of distance education [2], psychological therapy [3], assisted driving [4], etc. Now considerable work about emotion recognition has been performed, but there still exists some obstacles toward accurate emotion recognition. First, the expression of human emotion varies with surroundings, social protocols and language. Second, humans tend to express emotion within multiple social behavior modalities, such as voice, tuning and facial expression [5]. Comparing with the social surrounding and protocol that involves significant randomness and cultural diversity, the multiple modalities of social information are deemed to have more potential of interpretability [6]. There are two reasons to account for this interpretability from the perspective of physiology. Firstly, human social behaviors usually come from specific underlying neuronal correlations. Several findings have revealed that the human social behavior modalities actually sharing the common emotional motivation [7]. Secondly, many studies has indicated that there exists certain cross-coupling between the various social modalities. Due to above findings and studies, it has demonstrated that learning from multimodal sources offers the possibility of capturing correspondences between modalities and gaining an in-depth understanding of human emotion. Therefore, multimodal fusion becomes a rising solution for accurate emotion recognition and has attracted increasing attention in recent years. Baltruaitis et al. [8] identified core challenges for multimodal learning, among which representation learning and fusion mechanisms stand in a fundamental position. For representation learning for social emotion, a universal

✉ Xiaorui Liu
liuxiaorui1979@163.com

Yurui Xu
xuyurui0428@foxmail.com

Hang Su
hang.su@polimi.it

Guijin Ma
Malaodi0110@163.com

[1] Automation School of Qingdao University, Institute of Future, Qingdao, China

[2] Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy

recognition model has not been realized in practice. Instead, the practical problem is to select the combination of social modalities for representation, and how to develop the feasible architect for learning.

With regard to the sensory social information, many modalities have been developed by now [9]. Although some contact sensory modalities, typified by EEG, can provide a unique representation space for emotion state, its acquisition mode has difficulties to apply in the practical social scene [10–12]. Therefore, video stream, audio signals and speech context are still supposed to be the ideal modalities due to their universality and accessibility. Qi et al. [13] noted the importance of task oriented HMI and stated that it is scenes and tasks to define the means of perception. Although it is evident that emotion recognition benefits from the combination of multiple social information [14], it has not been an easy task to fuse these modalities. Srivastava and Salakhutdinov [15] identify the desirable properties for multi-modal representations: similarity in the representation space should reflect the similarity of the corresponding concepts. To pick up the right information from the heterogeneous sources, decision-level or feature-level fusions are the typical options for the classifier or end-to-end framework. The decision-level fusion indicates that operator performs detection at the sensor level and then merges detections to generate result. In many cases, decision-level fusion is suboptimal. If some information is not embodied by all modalities, it will not achieve the full benefits of fusion [16]. By comparison, feature-level fusion, concatenating both sets of features for learning, is expected to be more promising solution [17]. Nguyen et al. [18] adopted feature-level fusion to combine video and audio for emotion recognition. Liu and Fu [19] applied multichannel EEG and textual feature fusion in the time domain to recognize different human emotions. Although the existing fusion methods and architectures greatly improve the ability to process multimodal information, they are mostly highly customized, only utilize limited social corpus and cannot be effectively adapted. When it comes to do with modality extension or single/multiple modality switch, it usually means the reconstitution of the general architecture. Recently, several studies have paid attention to the unified-modal process achieved for the audio–visual integration [20, 21]. Since these studies mostly employed the image-text pair as input, quiet a lot of work was done to investigate the alignment or cross-attention between modalities. This would weaken the study on the unification of component or architecture.

In this paper, our efforts are mainly taken to explore the fusion towards audio signal and context. Compared with the classical image-text pair, these two modalities are naturally temporal synchronous and complementary. On the basis of synchronous audio-text pair, it can focus on exploring the potential to process heterogeneous information within homogeneous network architecture. For this purpose, a novel emotion recognition framework is proposed. This framework is composed of a parallel convolution (Pconv) module and an attention-based bidirectional long short-term memory (BLSTM) module. The main contributions of this paper are as follows:

(a) We adopt the fusion on the feature-level, proposes a dual-modal emotion recognition framework which employs the unified component (Pconv module).
(b) The Pconv module is constructed by convolution levels and polling units in parallel. The experimental results demonstrate that Pconv module is both effective to realize the feature abstraction of single modality and fusion among different modalities. This proves the feasibility to process heterogeneous information within homogeneous component, and will bring great flexibility for the modality expansion and framework design.
(c) Through comparing with existing frameworks, it is found that the attention-based BLSTM module would achieve best synergy performance to coordinate with the Pconv module. This combination could achieve higher accuracy within less cost.

This paper is organized as follows. The following section reviews the related work in the field of multimodal emotion recognition. "Design of algorithm" presents the details of the proposed framework and illustrates it in detail. "Experiment" shows the performance of our dataset of this study and the experimental settings, displays our experimental results and the discussion. "Conclusions" concludes this paper and outlines further work.

Related work

Up to now, it has been widely realized that the algorithm framework based on neural network is one effective tool for the emotion representation and learning, and most impressive achievement in emotion recognition starts with the single model recognition. Wang et al. [22] used a bidirectional recurrent neural network (BI-RNN) for affective learning of facial emotion on only static images and obtained significant advancements in the classification. For one-dimensional inputs, it is relatively convenient to build straight neural architecture that achieve promising performance on the self-supervised learning tasks. Ancilin and Milton [23] improved the extraction of the mel-frequency cepstrum coefficient and finally achieved a good recognition effect on multiple datasets. However, it has been realized that single social modalities is not sufficient to represent the complete emotion state in the dynamitic social scenes. In order to further improve the accuracy of emotion recognition, multimodality

fusion, as a methodology, is presented and has attracted most attention by now.

For multimodal fusion, one of the major challenges is how to effectively integrate data from different sources and design moderate architecture to complete representation learning. According to the fusion level or location, fusion concept can be divided into three types: data-level, feature-level and decision-level [24]. As the development of the machine learning, feature-level and decision-level fusion (or called early and late fusion) have spawned a variety of studies. For example, Lu et al. [25] present ViBERT model to learning image content and nature language, and through co-attention transformer layer to interact. Li et al. [26] proposed VisualBERT which used for vision and language tasks, the core idea is reuse the self-attention mechanism to implicitly align elements. Chen et al. [27] introduce UNITER which a universal image-text representation, and through ablation study to find an optimal combination of pre-training tasks. Within the feature-level, we are able to explore the interaction between raw features across modalities, but it also need avoiding to potentially suppress the modality-specific interaction. Furthermore, the raw features represent different physical properties of the signals in the respective modalities, this also presents challenges to the design of fusion architecture. Wang et al. [28] used directional pairwise cross-modal attention for sentiment analysis. They showed positive potential to both construct feature abstraction and fusion with chiral attention layer. Xu et al. [29] used the attention mechanism to integrate speech and audio files at the feature layer and achieved the best performance on the IEMOCAP dataset. Singh and Dhall [30] used traditional video texture features and a bag-of-audio-words feature set (BOAW), which was generated by OpenXBow, to carry out feature layer fusion for emotion recognition.

Through above investigation, it shows that great progress has been made in multimodal emotion recognition in recent years. However, the representation and learning on each modality is much customized in existing frameworks. This limits the expansion or switch among modalities. In this paper, we would like to explore the possibility to complete the feature abstraction and fusion through the homogeneous network component, and propose our solution in place of concatenation to model the correlation between dual modalities.

Design of algorithm

As discussed above, one effective methodology that can adapt to different social signals and realize high-accuracy recognition are left undetermined. In this paper, we propose a novel approach based on the combination of parallel convolution module (Pconv), BLSTM and an attention mechanism for dual-modal emotion recognition, as shown in Fig. 1. The details of this framework are discussed in the following sections.

Preprocessing

For audio emotion recognition, its feature description includes temperament features, sound spectrum features and sound quality features. In this study, the emotional characteristics of audio were analyzed by the mel-frequency cepstrum coefficient (MFCC) [31]. The MFCC displays the exact channel shape in the short-term power spectrum envelope of audio. It also provides audio feature parameters in the frequency domain.

To determine and capture the relationships between audio and text in expressing emotion, the original signal was converted into MFCC format. We adapted the Hamming window to process audio signals because it can reduce the impact of energy leakage in the frequency domain caused by truncation. The Hamming window formula is shown below:

$$w(n) = 0.54 - 0.46\cos(2\pi n/L - 1), \quad 0 \le n < L - 1. \tag{1}$$

In the formula (1), L is the window total length, and n is the current state. It can be seen from the formula that its characteristics come through the cosine function, which reflects the intermediate data, and the data information on both sides is lost. When the window is moved again, only 1/3 or 1/2 of the window is moved. Thus, the data lost in the previous frame or the second frame is reflected again.

The text corpus needs data processing before emotion classification. This paper chose Chinese dataset that has a specific grammar structure and word segmentation. To process raw text data, Jieba segmentation was used to conduct word segmentation on the input text data. Then, the Baidu Stop Word Database was used to remove the stop words. The processed text information was converted into word vector form within word2vec [32] to construct a word vector dictionary. Compared with other embedding methods, such as one-hot encoding and bidirectional encoding, word2vec is a lightweight neural network that only includes an input layer, hidden layer and output layer. Considering the quantity of text data used in this study, CBOW was used as the default word2vec model.

Attention-based BLSTM connects with Pconv module

Audio and text data are both temporal information. That is, the state of a certain moment is related to the context. In this paper, we design an architecture in which an attention-based BLSTM connects with feature extraction (as shown in Fig. 1). The feature extraction module utilizes Pconv module

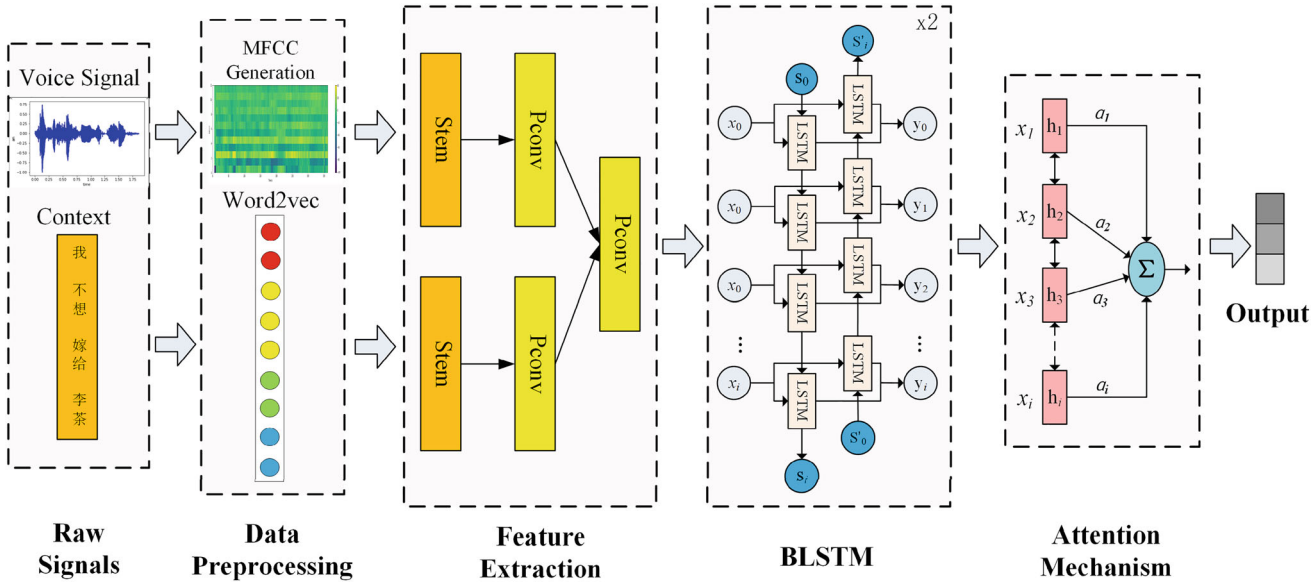

Fig. 1 Dual-modal emotion recognition framework

to obtain the overall feature from partial information aggregation, and then an attention-based BLSTM module is supposed to capture the correlation among temporal information. This structure design ensures that the relevance of the information and the network are lightweight. Although BLSTM can obtain contextual information, it has poor performance in long-term fitting. By considering the weight parameters of different elements, an attention mechanism can be added to highlight the part that needs more attention and suppress other useless information to quickly extract the important data features. Next, the details of the Pconv module, BLSTM and attention mechanism are explained.

Pconv module

This study splices the convolutional layer with the max-pooling layer to ensure that parallel convolutional cores can simultaneously extract multiple features. Additionally, to prevent weight attenuation, the shortcut connections method is adopted to fit the residual mapping of the previous layer. A depthwise separable convolution layer is added to the shortcut connections to form the Pconv module. This design can ensure that the size of the features before and after the shortcut connections remain consistent with the number of channels and reduce the number of framework parameters. The structure diagram of the Pconv module is shown in Fig. 2, and the experimental results are demonstrated in "Experiment".

The main idea of designing Pconv module is to determine how an optimal partial sparse module exists in a one-dimensional data network without disturbing self-continuity.

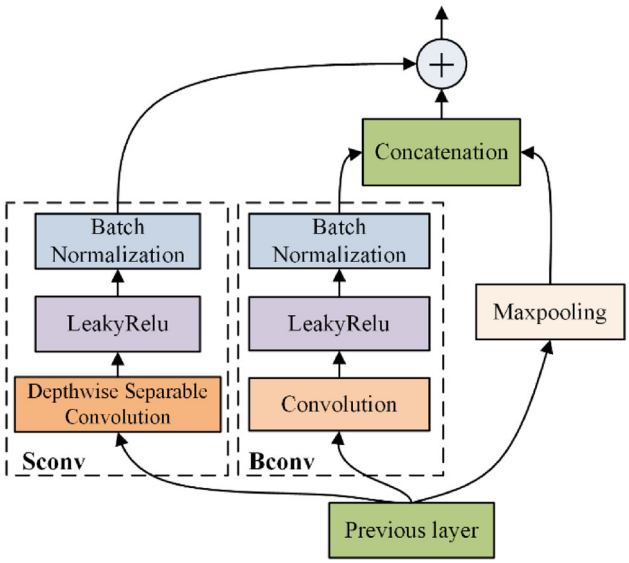

Fig. 2 The structure diagram of Pconv module

In the Pconv module, the input is fed into different convolutions and pooling simultaneously (as shown in the Fig. 2). Then, the parallel outputs will be merged by concatenation and addition operator. By appropriately designing the component and structure, the Pconv module will have potential to process diverse social modalities. Additionally, max-pooling removes redundant information from the previous layer.

For depthwise separable convolution, it is designed to split the multi-channel feature graph into single-channel feature graph, execute convolution calculation separately, and finally stack the results together. For traditional convolution process, input channel M and output channel N are supported. The

width and height of the input feature graph are G_{I}, and the width and height of the convolution kernel are F_{H}. Then the calculation times of the traditional convolution layer, C_{T}, is:

$$C_{\mathrm{T}} = G_{\mathrm{I}} \times G_{\mathrm{I}} \times M \times N \times F_{\mathrm{H}} \times F_{\mathrm{H}}. \tag{2}$$

Under the same conditions, the calculation times (C_{D}) of depthwise separable convolution layer is:

$$C_{\mathrm{D}} = G_{\mathrm{I}} \times G_{\mathrm{I}} \times M \times N \times F_{\mathrm{H}} \times F_{\mathrm{H}} + M \times N \times G_{\mathrm{I}} \times G_{\mathrm{I}}. \tag{3}$$

The ratio of deeply separable volume product to traditional convolution, P, is:

$$\begin{aligned} p &= \frac{C_{\mathrm{D}}}{C_{\mathrm{T}}} \\ &= \frac{G_{\mathrm{I}} \times G_{\mathrm{I}} \times M \times F_{\mathrm{H}} \times F_{\mathrm{H}} + M \times N \times G_{\mathrm{I}} \times G_{\mathrm{I}}}{G_{\mathrm{I}} \times G_{\mathrm{I}} \times M \times N \times F_{\mathrm{H}} \times F_{\mathrm{H}}} \\ &= \frac{1}{N} + \frac{1}{F_{\mathrm{H}} \times F_{\mathrm{H}}}. \end{aligned} \tag{4}$$

When $F_{\mathrm{H}} = 3$ and $N = 128$, the P value is 0.119. With the increase of the number of output channels and the size of the convolution kernel, the reduction of computation will be more obvious.

As shown in the Fig. 2, the Pconv module is divided into three parts. The first is max-pooling to provide an abstracted form of the representation. The second is the Bconv unit, which is made up of traditional convolution, LeakyReLU activation function and the batch normalization (BN) layer. The last is the Sconv unit, it is similar with Bconv in structure, but employs depthwise separable convolution instead of traditional convolution. At the end of the Pconv module, the output of max-pooling and Bconv merged by concatenation. The concatenated result is added with the output of the Sconv to form the output of Pconv. Through parallel structure, it can abstract the diverse feature of given modalities, from the different dimension. The concatenation and addition operator will further enhance this diversity, which is helpful for the following fusion and classification. Additionally, the design of the Pconv module can avoid gradient explosion and information redundancy.

BLSTM

After the Pconv module, the concatenated feature vectors are obtained continuously, which need to build their intercorrelation in the time series. To prevent the disappearance or explosion of gradients in temporal model training, the long short-term memory network (LSTM) [33] is utilized to accommodate the dynamic information over time. It has three gates to protect and control the state of the cell: input gate, forget gate and output gate. Its structure is shown in Fig. 3b. The formula of a single LSTM unit for time step t is given in the formula (5)–(10), where W is the weight matrix and b is the bias term:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{5}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{6}$$

$$\widetilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{7}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{8}$$

$$o_t = \sigma\left(W_O\left[h_{t-1}, x_t\right] + b_o\right) \tag{9}$$

$$h_t = o_t * \tanh(C_t). \tag{10}$$

Formula (5) determines what information we want to discard from the cell state, where h_{t-1} is the previous output state, x_t is the current state, σ represents the sigmoid function and f_t is the part expected to be forgotten. Formula (6) determines which values we want to update, and formula (7) then creates candidate vectors $\widetilde{C}_t$ to determine the updated information. The updated result, as shown in formula (8). C_t is measured by how much of each status value is updated. Formula (9) determines the part of the cell state to be output. In formula (10), C_t is the value obtained by formula (8), which is normalized between -1 and 1 through hyperbolic tangent function(tanh) and multiplied by the output of formula (9) to determine the final output.

LSTM can only take into account information prior to the current moment, but cannot take advantage of future moment information. To solve this problem, bidirectional long short-term memory network (BLSTM) is generated. BLSTM consists of two layers of LSTM superimposed from top to bottom. Based LSTM's capacity to make predictions based on current and previous information, this paper uses BLSTM to learn different forms of information. The working mechanism of BLSTM is shown in Fig. 3a. The first layer inputs information from left to right according to sequence, while the second layer is the initial input of sequence from right, and the output is determined by the state of the two layers. By linking several LSTM modules and building an internal status loop, BLSTM can store long-term subsequent information and understand the influence of context information on the state of a specific moment.

Attention mechanism

The BLSTM primarily realizes the memory and representation of temporal vectors. To maximize the contribution of the relevant vectors, the attention mechanism [34] is added after BLSTM, to form attention-based BLSTM module. The attention mechanism can not only improve the BLSTM output but

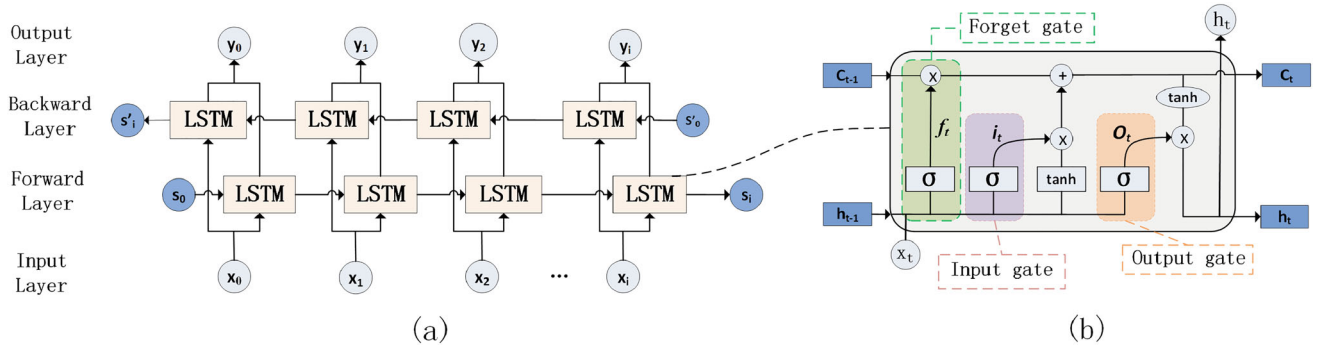

Fig. 3 The working mechanism of bidirectional long short-term memory networks

also be used to update storage units. The overall structure of attention mechanism is shown in Fig. 1. The attention layer is composed of three parts: query, key and value. The query is usually a single vector as input, while the key is multiple eigenvectors. The formula is described as formula (11)–(13).

$$e_i = f(Q, K_i) \tag{11}$$

$$\alpha_i = \exp(e_i) / \sum \exp(e_i) \tag{12}$$

$$\text{attention}\, e_x = \sum_i \alpha_i V_i \tag{13}$$

where Q, K, and V represent query, key and value, respectively. Formula (11) is a similar calculation function between key and query, and the calculation function of different attention mechanisms is inconsistent, as is the calculated degree of similarity. Then, the score which get from formula (11) was numerically converted in the formula (12). To get the final attention value, weighted summation of values based on weighting coefficients in the formula (13).

Fusion modeling and architecture design

Based on the above framework and module design, this study explores different feature fusion algorithms. By the fusion of different data modalities can realize complement of each other, thus ensure the robustness and increase recognition accuracy of emotion (the details are demonstrated in "Experiment"). Through comparison experiments, we selected the Pconv module for fusion as well. Figure 4 illustrates the model of the dual-modal emotion fusion recognition based on the attention-based BLSTM and Pconv modules. The preprocessed audio and text data are filled and then sent to their respective stem module which made up with fewer feature extraction to realize shallow feature extraction. For audio input, the processed MFCC was roughly extracted through the stem module, which included a convolution layer and a pooling layer. Additionally, to ensure the feasibility of audio and text information fusion, the dimensions and channels of the stem module output remain the same between audio and text model. Similarly, for text input, the processed information is sent into the stem module. The role of the embedding layer is to increase the correlation between features and improve the accuracy of the model. Adding BLSTM after the embedding layer can maintain sequence relationships between texts.

The design of this model not only ensures the lightweight nature of the network, but also enhance the relevance of the feature. Meanwhile, uses the parallel network structure can use fewer number of network layers to extract the deep emotional states. So that this model effectively prevents the gradient explosion caused by the deepening of network layers.

Experiment

Dataset selection

In this paper, we adopt the CH-SIMS [35] dataset that consists of 60 Chinese films, TV dramas and varieties showing unprocessed fragments with emotional expression. In addition, only Mandarin is considered. Compared with others datasets, this dataset is closer to life scenarios. This dataset contains 2,281 fragments whose average duration is 3.67 s, and the average text length is 15 words per sentence. There were five experts tagging the same segment at the same time, and the average tagging result of the five experts was used as the final emotional state. Emotion states of the dataset were classified as neutral, positive and negative whose labels were set as 0, 1, and 2, respectively. Then, the data in the dataset were divided into a training data, verification data and testing data at a ratio of 7:2:1.

Experimental setup

In this paper, the Keras framework was adopted for model building and training. The framework parameters were

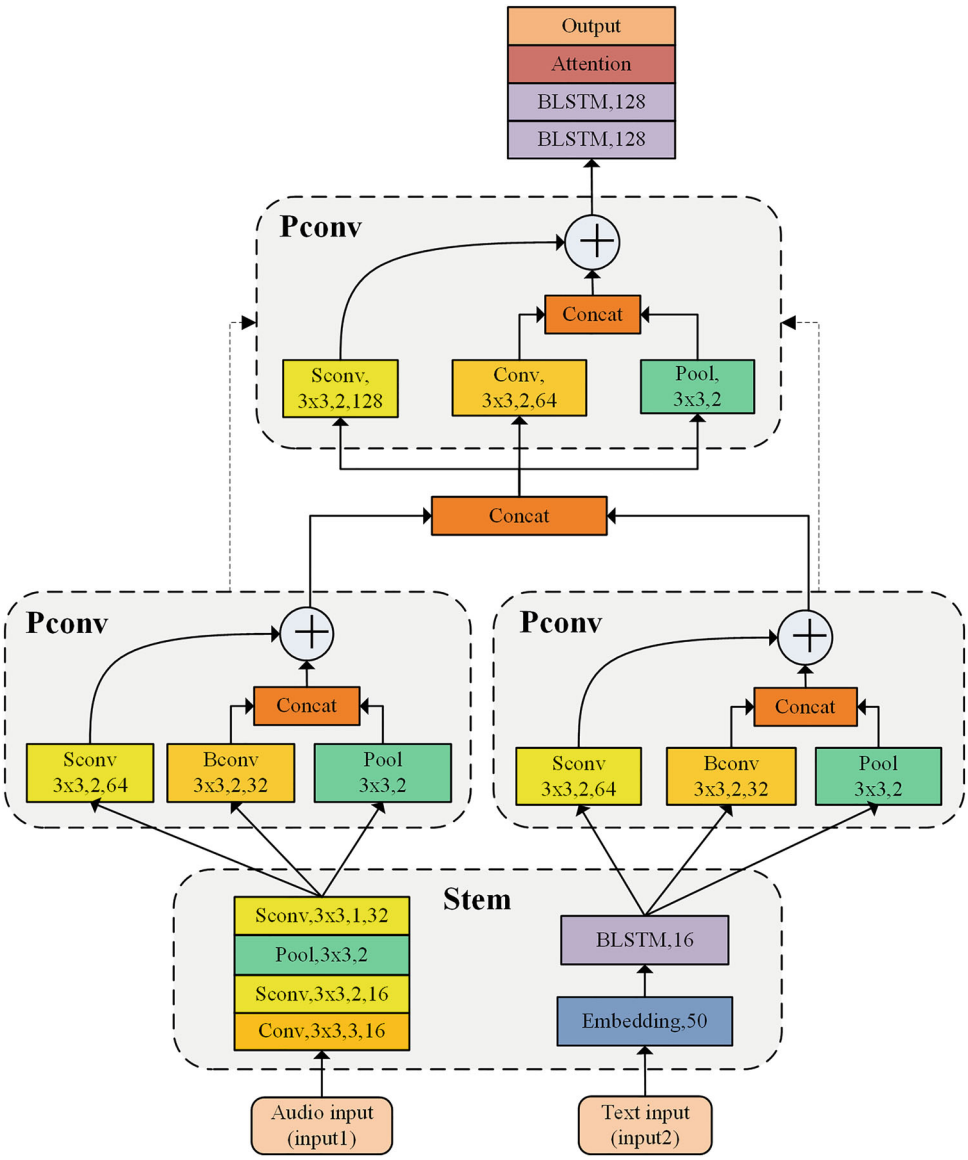

Fig. 4 Structure diagram of the fusion emotion recognition model

basically the same in both single-modal and dual-modal. The training strategy selected stochastic gradient descent (SGD) with the Nesterov acceleration optimizer to prevent the occurrence of partial optimal. In addition, LeakyReLU was used as the activation. Through test and comparison the finally detailed selection of training parameters is shown in Table 1.

In this paper, three evaluation indexes were used to evaluate the performance of emotion recognition. Accuracy: the percentage of the correct quantity predicted by the model in the total quantity; it is one of the most commonly used evaluation indexes. Prediction: the percentage of a sample identified as a positive category that is actually a positive category. $F1$-score: a statistic that is a weighted and harmonic average of precision and recall rate; it combines the results of precision and recall rate and is often used to evaluate the quality of models. The calculation formula of above indexes is shown in (14)–(17):

Table 1 Training relevant hyperparameters

Hyperparameters	Value
Batch size	32
Learning rate	0.0001
Vocab_dim	50
Maxlen	50
Epoch	5000
LeakyReLU	0.02
BLSTM nodes	128

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{14}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$F1\text{-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

where TP (true positive) refers to the number of positive classes correctly predicted to be positive classes. TN (true negative) is the number of negative classes correctly predicted to be negative classes. FP (false positive) is the number of negative classes that are wrongly predicted to be positive. FN (false negative) is the number of false predictions of a positive class as a negative class.

Experimental results and discussion

Single-modality classification

Tables 2 and 3 show the performance comparison of audio and text recognition using different methods on the CH-SIMS dataset. To verify the superiority of our proposed method, we took the most basic CNN (convolution neural network) to build feature extraction module, this module is called 'BCNN' model and shown in the Fig. 5. In the following experiments, it will be regarded as the baseline to evaluate the effectiveness of Pconv or other structure. The Pconv-based model is similar to the BCNN model in structure, but it replaces the traditional convolution layers in the BCNN with the depthwise separable convolution layer at the shortcut connection. Similarly, the Pconv + BLSTM (PB) model adds two layers of BLSTM based on the Pconv model. The Pconv + BLSTM + attention (PBA) model adds an attention mechanism based on the PB model, it is the final single model adopted in this study, the structure of the PBA model is shown in Fig. 4 (connections with dotted lines). It is the same as the dual-modal, except that there is no concatenation between different models. Thus, it can be verified that the dual-modal is superior to the single-modal model. To ensure the credibility of the comparative experiment, the structure of the shallow feature extraction module of these four models, with their parameters as well, would remain unchanged.

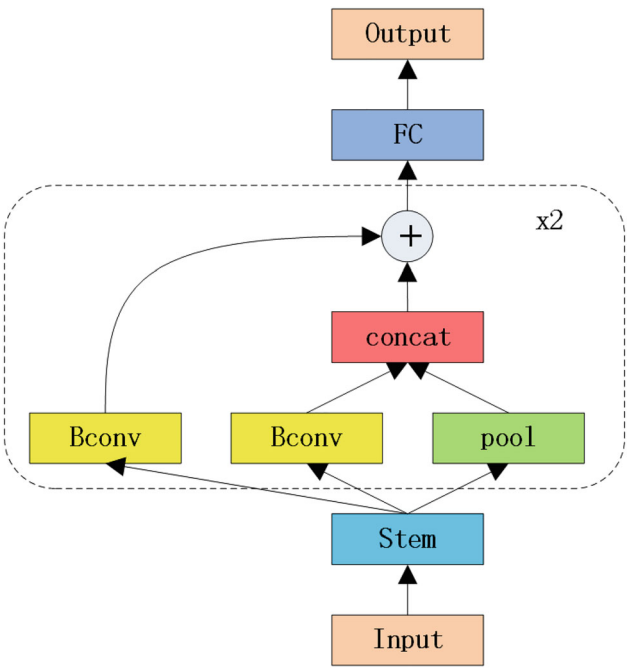

Fig. 5 The structure of BCNN model

Table 2 shows the performance with different methods of the audio model in the same scenario. In Table 2, we can clearly observe that the results of the PBA model adopted in this study are better than those of other models in terms of accuracy, prediction and $F1$-score. It achieved 74.70% accuracy, which was 10% higher than that of the other models. In addition, the result also reached 76% on prediction. The PB model achieved the best performance in $F1$-score at neutral and negative emotion, which achieved 78% and 82%, respectively. However, at positive emotion recognition, the model proposed in this paper reached 54%, which was the best of these methods.

Table 3 shows the performance with different methods of the text model in the same scenario. The PBA model achieved the best performance with accuracy, prediction and $F1$-score. The accuracy was 77.13%, which also improved by approximately 10%. It also achieved 78.6% prediction and the best performance on the $F1$-score for emotional status (neutral, positive and negative).

Through comparison in the Tables 2 and 3, it proved the superiority of PBA model for the emotion recognition of single modality. The proposed model (PBA) generally improved the $F1$-score in three emotion classifications. The recognition result gaps among these emotion classifications are lowest under the PBA condition. Similarly, in Tables 2 and 3, by comparing the PBA, PB models with the Pconv model, it can observe that the addition of BLSTM and the attention mechanism improved the accuracy of both the audio and text. In addition, we also employ some classic network for comparison, such as LeNet, AlexNet and VGG, the performance of these models is shown in Table 4. Through comparison it can be found that VGG has lowest accuracy and prediction, the main reason for it is that VGG's over-deep layered structure makes the over-fitting take place during training. The performance of LeNet and AlexNet are lower than that of proposed model in this paper. This shows that attention-based BLSTM connects with Pconv module can extract feature information

Table 2 The performance results of the audio model with different methods

Model	Accuracy	Prediction	$F1$-score		
			0	1	2
BCNN	0.6521	0.664	0.69	0.44	0.70
Pconv-based	0.6813	0.655	0.68	0.46	0.71
PB	0.7202	0.729	0.78	0.30	0.82
PBA	0.7470	0.760	0.75	0.54	0.78

Table 3 The performance results of the text model with different methods

Model	Accuracy	Prediction	$F1$-score		
			0	1	2
BCNN	0.6764	0.655	0.63	0.07	0.75
Pconv-based	0.7080	0.655	0.69	0.19	0.74
PB	0.7445	0.734	0.73	0.37	0.79
PBA	0.7713	0.786	0.84	0.57	0.86

Table 4 The comparison results of audio and text

Model	Audio		Text	
	Accuracy	Prediction	Accuracy	Prediction
LeNet	0.5600	0.625	0.5440	0.555
AlexNet	0.6057	0.659	0.6015	0.642
VGG	0.5092	0.585	0.4966	0.507
Proposed (PBA)	0.7470	0.760	0.7713	0.786

in fewer layers of the network, and improve the final recognition effect.

Fusion modality classification

Here it shows the results of dual-modal emotion recognition under different fusion methods. Moreover, the PBA model was adopted for both the audio model and text model. The partial structure of the fusion methods is shown in Fig. 6. The fusion method adopted in this study is the '1-layer Pconv', and its model structure is shown in Fig. 6b. '2-layer Pconv' represents data fusion before being input to the first Pconv module, as shown in Fig. 6a. Similarly, as shown in Fig. 6c–f, '2-layered BLSTM' represents data fusion before being input of the first BLSTM layer; '1-layered BLSTM' represents data fusion before the second BLSTM layer; 'attention' is data fusion before being input into the attention mechanism; 'FC' stands for the adoption of decision-level fusion, where data are fused before entering the full connection layer. In addition, the results are shown in Table 5.

Table 5 shows that the accuracy of the '1-layer Pconv' fusion method was 90.02% (increasing 7.5% compared with the other fusion methods). The prediction of '1-layer Pconv' also showed the better than the other fusion methods. The accuracy of the '2-layer Pconv' and the '1-layer Pconv' fusion methods all better than other, which proves that the Pconv module has great advantages in data fusion, and it can realizes the mutual complement of different sensory data. The $F1$-score of the '1-layer Pconv' fusion method on neutral and negative emotions was also higher than 90% and higher than others, which fully proves the reliability of this fusion method.

Comparison between single-modal and dual-modal

By comparing Tables 2, 3 and 4, it can be seen that dual-modal was higher than any single-modal of the accuracy. To compare the performance between the single-modal model and the dual-modal model. Here, we selected the PBA model of both the audio and text models and the '1-layer Pconv' fusion model for comparison. The PBA model processes audio and text data separately. The '1-layer Pconv' model processes audio and text data simultaneously. The model training was conducted on a 24-GPU server for 5000 rounds, and the curve comparison of accuracy and loss values is shown in Fig. 7. The figure proves that the accuracy of this fusion method was higher than that of the single-modal recognition model, the loss value decreased faster than that of the single model,

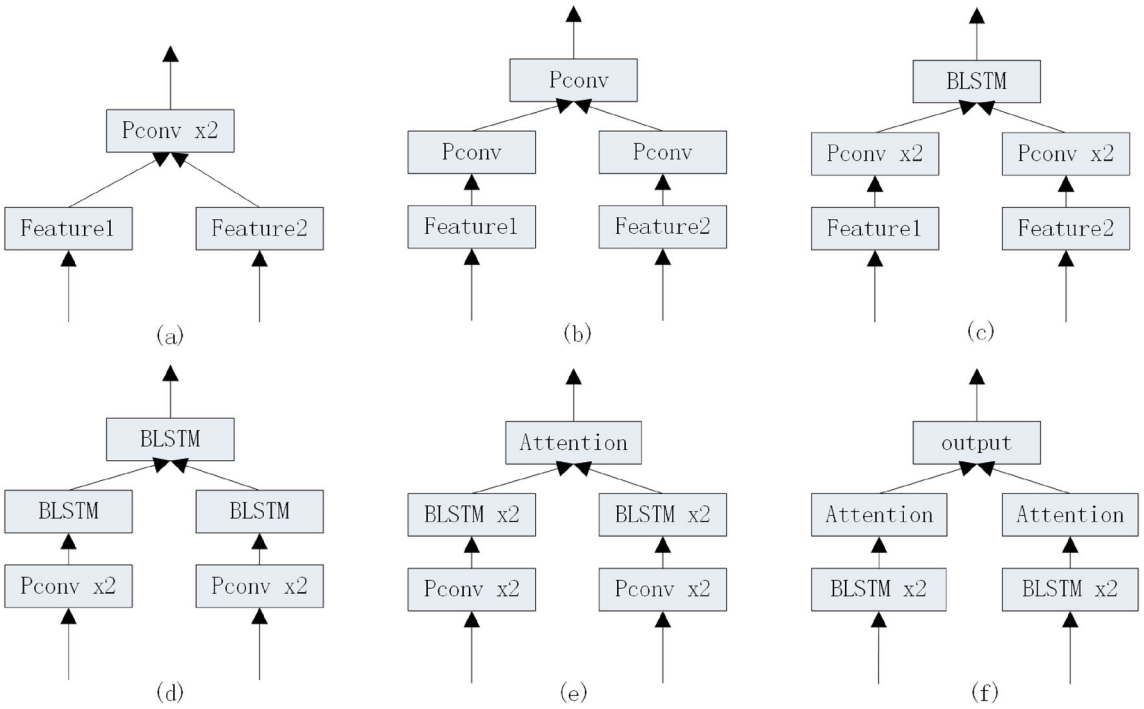

Fig. 6 The partial structure of different fusion methods

Table 5 Performance results of different fusion methods

Fusion layer	Accuracy	Prediction	*F*1-score		
			0	1	2
2-layer Pconv	0.8662	0.865	0.91	0.85	0.91
1-layer Pconv	0.9002	0.904	0.93	0.71	0.94
2-layer BLSTM	0.8248	0.773	0.88	0.78	0.88
1-layer BLSTM	0.8370	0.821	0.87	0.80	0.90
Attention	0.8589	0.869	0.87	0.61	0.90
FC	0.8491	0.878	0.92	0.82	0.93

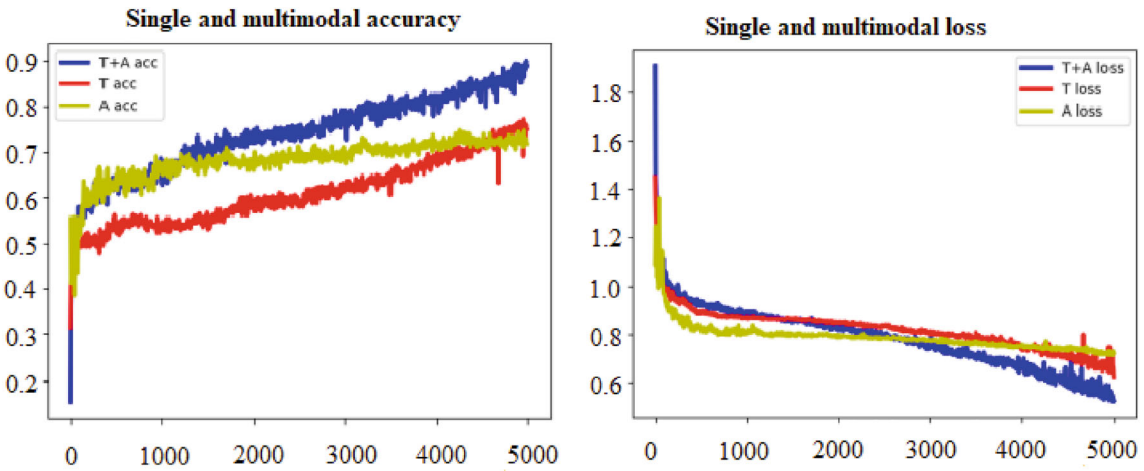

Fig. 7 Comparison of accuracy and loss between the fusion model and single model

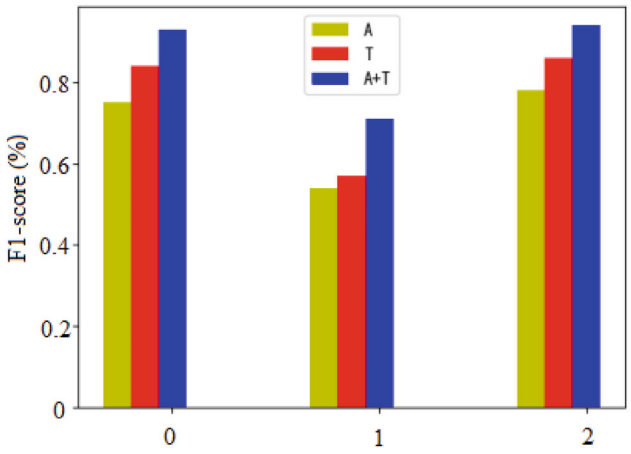

Fig. 8 Classification accuracy comparison between the fusion model and single model

and the training is relatively stable. It can also be observed in Fig. 7 that the accuracy of the audio model rose slowly in the later stage, while the accuracy of the text model fluctuated greatly in the later stage. These phenomena were solved well in the PBA model, which proves that the effectiveness of the proposed fusion methodology.

Figure 8 shows the comparison results of these three models in $F1$-score. It can be seen in the figure that the classification accuracy of the fused model on various emotions was significantly higher than that of any single-modal emotion model. Different emotional signals reflect the emotional expressions to different degrees. It is worth noting that the accuracy of the positive emotion was lower than that of the neutral and negative emotions. However, under the same emotional state, the accuracy of the positive emotion after fusion was significantly higher than that of the single-modal model in this emotion. The integration of audio and text for emotion recognition is beneficial to learn from each other and improve the accuracy of final emotion recognition. This proves the necessarily to perform multimodal fusion.

As is well-know, to better evaluate the model, the dataset is processed and divided into training data, validation data and testing data. The training data is used to fit model towards the best weight and bias. The validation data is used to determine the model hyperparameters and select the optimal model. Testing data are used to test the model's ability to learn from new data. In this experiment, we compared the dataset with emotion classification, and the comparison results are shown in the Fig. 9, the division (7:2:1) is adopted due to the best train accuracy of emotion classification.

Meanwhile, we compared the proposed framework with other existing work in the last 2 years, the result is shown in Table 6. The Table 6 shows that the proposed PBA models have the highest accuracy level in three different tests (Audio, Text and Fusion). At the same time, the table also proves that the recognition effect of fusion modality is better than that of single-modality model.

Conclusions

In this paper, an dual-modal emotion recognition model is proposed. This model is composed of a parallel convolution (Pconv) module and attention-based bidirectional long short-term memory (BLSTM) module. The Pconv module employ parallel structure to extract social features in several dimensions, this is conducive to enhance the capacity of representation. Additionally, attention-based BLSTM module is utilized to strengthen the correlation between information and improve the accuracy of recognition. Experiments conducted on the CH-SIMS dataset indicate that the recognition accuracy reaches 74.70% on audio data and 77.13% on text, while the accuracy of the dual-modal fusion model reaches 90.02%. Through experiments it proves the feasibility to

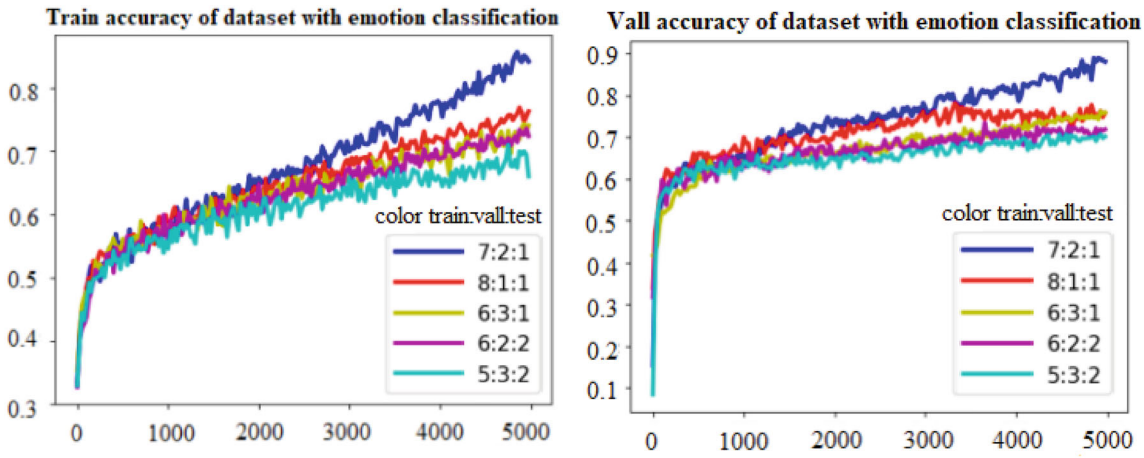

Fig. 9 Comparing of dataset with emotion classification accuracy result

Table 6 Accuracy comparison with other papers

S. no.	Method	Training parameter	Accuracy (%)		
			Audio	Text	Fusion
1	DNN (Singh et al. [36])	8:2	68.3	68.0	74.5
2	SVM (Vashishtha and Susan [37])	7:3	58.0	71.2	82.5
3	BERT (Pepino et al. [38])	N/A	72.26	68.11	81.31
4	DCNN (Priyasad et al. [39])	8:1:1	69.89	70.81	79.22
5	Transformer (Makiuchi et al. [40])	N/A	70.1	66.1	73.0
6	CNN (Krishna and Patil [41])	N/A	55.60	65.90	72.82
7	CTNet (Lian [42])	8:2	67.5	80.8	83.6
8	Resnet and BERT (Padi [43])	N/A	70.33	65.97	76.07
9	Proposed (PBA)	7:2:1	74.01	77.13	90.02

process heterogeneous information within homogeneous network component, and demonstrates great flexibility for the modality expansion and architecture design. In future work, efforts will be taken to balance all kinds of emotional states, and broaden the proposed framework to other modalities.

Acknowledgements This paper is funded by the National Key Research and Development Program of China (no. 2020YFB1313600).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Nayak S, Nagesh B, Routray A et al (2021) A human–computer interaction framework for emotion recognition through time-series thermal video sequences. Comput Electr Eng 93:107–118
2. Bouhlal M, Aarika K, Ait Abdelouahid R et al (2020) Emotions recognition as innovative tool for improving students' performance and learning approaches. Procedia Comput Sci 175:597–620
3. Krause FC, Linardatos Ef, Fresco DM et al (2021) Facial emotion recognition in major depressive disorder: a meta-analytic review. J Affect Disord 293:320–328
4. Cui Y, Ma Y, Li W et al (2020) Multi-EmoNet: a novel multi-task neural network for driver emotion recognition. IFAC PapersOnLine 53:650–655
5. Mumenthaler C, Sander D, Manstead ASR (2020) Emotion recognition in simulated social interactions. IEEE Trans Affect Comput 11(2):308–312
6. Volpert-Esmond HI, Bartholow BD (2021) A functional coupling of brain and behavior during social categorization of faces. Personal Soc Psychol Bull 47:1580–1595
7. Liu L, Xu H, Wang J, Li J, Xu H (2020) Cell type-differential modulation of prefrontal cortical gabaergic interneurons on low gamma rhythm and social interaction. Sci Adv 6(30):eaay4073
8. Baltruaitis T, Ahuja C, Morency LP (2019) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41:423–443
9. Poria S, Hazarika D, Majumder N et al (2020) Beneath the tip of the iceberg: current challenges and new directions in sentiment analysis research. IEE Trans Affect Comput 14:1–29
10. Sharma R, Pachori RB, Sircar P (2020) Automated emotions recognition based on higher order statistics and deep learning algorithm. Biomed Signal Process Control 58:101867
11. Singh K, Malhotra J (2022) Two-layer LSTM network based prediction of epileptic seizures using EEG spectral features. Complex Intell Syst 8:2405–2418
12. Sharma R, Sircar P, Pachori RB (2020) Seizures classification based on higher order statistics and deep neural network. Biomed Signal Process Control 59:101921
13. Qi X, Wang W, Guo L et al (2019) Building a Plutchik's wheel inspired affective model for social robots. J Bionic Eng 16(002):209–221
14. Hossain MS, Muhammad G (2018) Emotion recognition using deep learning approach from audio-visual emotional big data. Inf Fusion 49
15. Srivastava N, Salakhutdinov R (2014) Multimodal learning with deep Boltzmann machines. J Mach Learn Res 15:2949–2980
16. Xu G, Li W, Liu J (2020) A social emotion classification approach using multi-model fusion. Future Gener Comput Syst 102:347–356
17. Cai H, Qu Z, Li Z et al (2020) Feature-level fusion approaches based on multimodal EEG data for depression recognition. Inf Fusion 59:127–138
18. Nguyen D, Nguyen K, Sridharan S et al (2018) Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. Comput Vis Image Underst 174:33–42
19. Liu Y, Fu G (2021) Emotion recognition by deeply learned multichannel textual and EEG features. Future Gener Comput Syst 119:1–13

20. Li J, Selvaraju RR, Gotmare AD et al (2021) Align before fuse: vision and language representation learning with momentum distillation. In: Paper Presented at the Proceedings of the 35th Conference on Neural Information Processing System, Sydney, pp 104–121
21. Li W, Gao C, Niu G et al (2020) Unimo: towards unified-modal understanding and generation via cross-modal contrastive learning. In: Paper Presented at the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Conference on Natural Language Processing, Thailand, pp 2592–2607
22. Wang X, Peng M, Pan L, Hu M, Jin C, Ren F (2018) Two-level attention with two-stage multi-task learning for facial emotion recognition. J Vis Commun Image Represent 62(JUL.):217–225
23. Ancilin J, Milton A (2021) Improved speech emotion recognition with mel frequency magnitude coefficient. Appl Acoust 179(3):108046
24. Farhoudi Z, Setayeshi S (2020) Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition. Speech Commun 127:92–103
25. Lu J, Batra D, Parikh D et al (2019) Vilbert: pretraining task-agnostic visiolinguistic representations for vision and language tasks. In: Paper Presented at the Proceedings of 33rd Conference on Neural Information Processing Systems, Vancouver, Canada, pp 13–23
26. Liunian LH, Yatskar M, Yin D et al (2019) Visualbert: a simple and performant baseline for vision and language. arXiv arXiv:1908.03557
27. Chen YC, Li L, Yu L et al (2020) Uniter: universal image-text representation learning. In: European conference on computer vision. Paper Presented at the Proceedings of European Conference on Computer Vision, Glasgow, pp 1303–1313
28. Wang Z, Zhou X, Wang W et al (2020) Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video. Int J Mach Learn Cybern 11:923–934
29. Xu H, Zhang H, Han K et al (2019) Learning alignment for multimodal emotion recognition from speech. In: Proceedings of InterSpeech 2019, September 15-19, Graz, Austria, pp 3569–3573
30. Narotam S, Nittin S, Abhinav D (2017) Continuous multimodal emotion recognition approach for AVEC 2017. arXiv arXiv:1709.05861
31. Meng Z (2021) Research on timbre classification based on BP neural network and MFCC. J Phys Conf Ser 1856(1):012006
32. Kolesnikova O, Gelbukh A (2020) A study of lexical function detection with word2vec and supervised machine learning. J Intell Fuzzy Syst 39(2):1–8
33. Shobana J, Murali M (2021) An efficient sentiment analysis methodology based on long short-term memory networks. Complex Intell Syst 7:2485–2501
34. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. Comput Sci 23:1399–1409
35. Yu W, Xu H, Meng F et al (2020) CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proceedings of the 58th annual meeting of the association for computational linguistics, Seattle, pp 3718–3727
36. Singh P, Srivastava R, Rana K et al (2021) A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowl Based Syst 229:107–119
37. Vashishtha S, Susan S (2020) Inferring sentiments from supervised classification of text and speech cues using fuzzy rules. Procedia Comput Sci 167:1370–1379
38. Pepino L, Riera P, Ferrer L et al (2020) Fusion approaches for emotion recognition from speech using acoustic and text-based features. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, pp 6484–6488
39. Priyasad D, Fernando T, Denman S et al (2020) Attention driven fusion for multi-modal emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, pp 3227–3231
40. Makiuchi MR, Uto K, Shinoda K (2021) Multimodal emotion recognition with high-level speech and text features. In: Proceedings of the 2021 IEEE automatic speech recognition and understanding workshop, Cartagena, pp 350–357
41. Krishna D, Patil A (2020) Multimodal emotion recognition using cross-modal attention and ID convolutional neural network. In: Interspeech, Shanghai, China: ISCA, 2020, pp 4243–4247
42. Lian Z, Liu B, Tao J (2021) CTNet: conversational transformer network for emotion recognition. IEE/ACM Trans Audio Speech Lang Process 29:985–1000
43. Padi S, Sadjadi SO, Manocha D et al (2022) Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models. arXiv:2202.08974, pp 407–414