Computers, Materials & Continua
DOI: 10.32604/cmc.2023.028291
Article

Tech Science Press
check for updates

A Multi-Level Circulant Cross-Modal Transformer for Multimodal Speech Emotion Recognition

Peizhu Gong[1], Jin Liu[1], Zhongdai Wu[2], Bing Han[2], Y. Ken Wang[3] and Huihua He[4,*]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China
[2]Shanghai Ship and Shipping Research Institute, Shanghai, 200135, China
[3]Division of Management and Education, University of Pittsburgh, Bradford, USA
[4]College of Early Childhood Education, Shanghai Normal University, Shanghai, 200234, China
*Corresponding Author: Huihua He. Email: hehuihua@shnu.edu.cn
Received: 07 February 2022; Accepted: 14 March 2022

Abstract: Speech emotion recognition, as an important component of human-computer interaction technology, has received increasing attention. Recent studies have treated emotion recognition of speech signals as a multimodal task, due to its inclusion of the semantic features of two different modalities, i.e., audio and text. However, existing methods often fail in effectively represent features and capture correlations. This paper presents a multi-level circulant cross-modal Transformer (MLCCT) for multimodal speech emotion recognition. The proposed model can be divided into three steps, feature extraction, interaction and fusion. Self-supervised embedding models are introduced for feature extraction, which give a more powerful representation of the original data than those using spectrograms or audio features such as Mel-frequency cepstral coefficients (MFCCs) and low-level descriptors (LLDs). In particular, MLCCT contains two types of feature interaction processes, where a bidirectional Long Short-term Memory (Bi-LSTM) with circulant interaction mechanism is proposed for low-level features, while a two-stream residual cross-modal Transformer block is applied when high-level features are involved. Finally, we choose self-attention blocks for fusion and a fully connected layer to make predictions. To evaluate the performance of our proposed model, comprehensive experiments are conducted on three widely used benchmark datasets including IEMOCAP, MELD and CMU-MOSEI. The competitive results verify the effectiveness of our approach.

Keywords: Speech emotion recognition; self-supervised embedding model; cross-modal transformer; self-attention

1 Introduction

Speech emotion recognition (SER) [1–3] is of significant importance in subjective cognitive research, which aims to determine a human's emotional states towards a certain topic by understanding the characteristics of speech in media. Traditional emotion recognition methods tend to be based on unimodality like image or text. However, the limited amount and single distribution of information

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

makes the results unsatisfactory. Unlike the traditional paradigm, recent studies consider emotion recognition of speech signals as a multimodal task because it contains semantic features of two different modalities, i.e., audio and text. Yoon et al. [4] combined the information from audio and text sequences using dual recurrent neural networks (RNNs) and then predicted the emotion class. Yoon et al. [5] put forward multi-hop attention mechanism to compute the relevant segments of the textual data and audio signal. In recent years, Transformer [6,7] has been widely used in various research areas including Computer Vision (CV) and Natural Language Processing (NLP) and achieved state-of-the-art results. Naturally, the cross-modal Transformer [8] has been proposed for interaction problems in multimodal task. However, studies [9] have shown that using the Transformer purely at an early stage is not good for results or requires an extremely large dataset. Therefore, we propose a multi-level framework to extract features and interactions in a stepwise manner. In addition, when it comes to representing raw data, traditional approaches usually employ Mel-frequency cepstral coefficients (MFCCs) [10] for audio signals and glove embeddings [11] for textual sequences. Rather than using those low-level feature extractor, self-supervised embedding models (SSE) [12] would be a better choice due to its powerful representation capabilities. SSE models are usually pre-trained on pretext tasks with a large number of unlabeled data in advance, and then generate more valuable feature representations for downstream tasks [13–15]. Bidirectional Encoder Representations from Transformer (BERT) [16] is known as one of the most outstanding SSE for text representation, while Masked Autoencoder (MAE) [17] gives unlimited possibilities for vision learning.

In this paper, we propose a multi-level circulant cross-modal Transformer (MLCCT) for multimodal speech emotion recognition. To the best of our knowledge, it is the first time that a Transformer-based progressive framework is used in multimodal speech emotion recognition, which may better capture correlations between different modalities. MLCCT is composed of three parts, feature extraction, interaction and fusion. SSE models are introduced for feature extraction, which give a powerful representation of the original data. Specially, Transformer Encoder Representations from Alteration (TERA) [18] and a robustly optimized BERT pretraining approach (RoBERTa) [19] are used to represent audio signal and textual sequence, respectively. In particular, MLCCT contains two types of feature interaction processes. The first one employs a bidirectional Long Short-term Memory (Bi-LSTM) with circulant interaction mechanism for low-level features, while in the second stage a two-stream residual cross-modal Transformer block (RCTB) is applied. Finally, we choose self-attention blocks for fusion and a fully connected layer to make predictions. Comprehensive experimental results on three benchmark datasets including IEMOCAP, MELD and CMU-MOSEI [20–22] show that the proposed MLCCT has achieved competitive results. In summary, the major contributions of our research can be summarized as follows:

- A multi-level framework is proposed for a progressive cross-modal interaction, which combines local and global features for accurate predictions, which will serve as an inspiration for future research.
- A circulant interaction mechanism is presented to take full advantage of capturing correlations between different modalities in an early stage.
- Most of the work done focuses on handcrafted feature extracting by using machine learning methods, while SSE models are introduced in this paper to generate more valuable feature representations for downstream classification.
- A two-stream residual cross-modal Transformer block is put forward to establish deep interactions between modalities.

The remaining of this paper is structured as follows: In Section 2, we introduce related work on multimodal speech emotion recognition and self-supervised learning. In Section 3, the details

of proposed MLCCT are presented. Finally, we evaluate the experimental results three benchmark datasets and draw a conclusion in Sections 4 and 5.

2 Related Work

2.1 Multimodal Speech Emotion Recognition

Emotion recognition models can be divided into two categories, discrete models and dimensional models. Discrete models use a few adjectives such as anger, disgust, fear, happiness, sadness, and surprise to describe emotional states. Thus, it is widely accepted for its simplicity despite of the limitations in presenting dynamic processes. In contrast, several basic attributes are usually selected as coordinate measures of the emotion space in the dimensional emotion model. Then all emotional states can be found on this space with its own coordinate points. The Pleasure-Arousal-Dominance model (PAD) [23] is one of the most well-known dimensional models, which can theoretically represent an infinite number of emotions. However, the dimensional model is difficult to understand and complex to manipulate, resulting in it not being mainstream.

Traditional methods recognize speech emotion by acoustic feature extraction or semantic information analysis. Acoustic features are classified into three categories: rhythmic features, spectral-based correlation features, and sound quality features, which describe information about the pitch, amplitude and timbre of speech, respectively. Bhargava et al. [24] improved automatic emotion recognition from speech by incorporating rhythm and temporal features. Different metrics of speech rhythm are investigated with the aim to determine general emotional tendencies at the overall level by studying the regularity of the appearance of certain linguistic elements in speech. Palo et al. [25] put forward a Wavelet-based MFCCs model to perform affective computing with respect to spectral features, and the improved model is more resistant to interference from noise. Recently, deep learning techniques have demonstrated breakthrough performance and have been considered as an alternative to tradition approaches in SER. The two most popular neural network architectures are convolutional neural networks (CNNs) [26–29] and recurrent neural networks (RNNs) [30–32]. Among them, CNN is beneficial for spatial information and RNN helps to capture temporal information in SER. Yenigalla et al. [33] presented a CNN-based model to classify emotion using phoneme and spectrogram. Zhao et al. [34] proposed a hybrid model, in which a 1D CNN-LSTM network and a 2D CNN-LSTM network were constructed to learn local and global emotion-related features from speech and log-mel spectrogram, respectively. In addition, the attention mechanism is particularly favored in multimodal interaction problems, especially Transformer. Tsai et al. [35] proposed Multimodal Transformer (MulT), which leverages inter-modal connections to individually reinforce target modal characteristics. A Transformer-based joint-encoding for emotion recognition [36] is put forward, which relies on a modular co-attention and a glimpse layer to jointly encode one or more modalities. However, using transformers purely at an early stage is detrimental to the results, so we propose a progressive framework to address this issue.

2.2 Self-supervised Embedding

In recent years, supervised learning has hit a bottleneck, which relies heavily on expensive manual labeling and suffers from poor generalization. As an alternative, self-supervised embedding models (SSE) has showed its soaring performance and gained increasing attention. SSE aims to learn a generic representation for downstream tasks, where training data is labeled by using a semi-automatic process. Specifically, the process may predict a portion of the data from the others. The general procedure is that the SSE models are firstly pre-trained on a set of pretext tasks, and then the pre-trained SSE models extract features for downstream tasks.

SSE models can be categorized into generative, contrastive and adversarial. Generative models encode the input x into a high-dimensional vector and then reconstructs x from the vector as the output. GPT and GPT-2 [37,38] reconstruct the sentence by predicting the next word. BERT incorporates masked language model on the basis of next word prediction and achieves better results. PixelCNN [39] fixes the picture by next pixel predicting, while VQ-VAE [40] acts on the whole picture. Generative models recover the original data distribution without making assumptions for downstream tasks, leading to its wide applications. However, generative models are extremely sensitive to rare samples and there is a semantic gap between pretext tasks and downstream tasks. Contrastive models map pairs of data to a common space and measure similarity. Deep InfoMax [41] focus on modeling the mutual information between local feature and global context, while MoCo [42] and SimCLR [43] tend towards instance-level representations. In addition, works including CLEAR [44] and BERT-CT [45] have presented overwhelming performances on various benchmark datasets. In comparison, without decoder, contrastive models are usually lightweight and take classification tasks as downstream tasks. Adversarial models train an encoder to generate fake samples and a decoder to distinguish them from real samples. Adversarial models, especially Generative Adversarial Networks (GAN) [46], have shown significant results in image generation and style transformation. However, there are still challenges for its future development due to its easy collapse and limited application in NLP.

3 Method

As shown in Fig. 1, the structure of proposed MLCCT is composed of three parts, feature extraction, interaction and fusion. SSE models are introduced for feature extraction, which give a more powerful representation of the original data than those using spectrograms or MFCCs. In particular, MLCCT contains two types of feature interaction processes, where a Bi-LSTM with circulant interaction mechanism is proposed for low-level features, while a two-stream RCTB is applied when high-level features are involved. Finally, we choose self-attention blocks for fusion and a fully connected layer to make predictions. It is believed that the progressive framework may better capture correlations between different modalities.

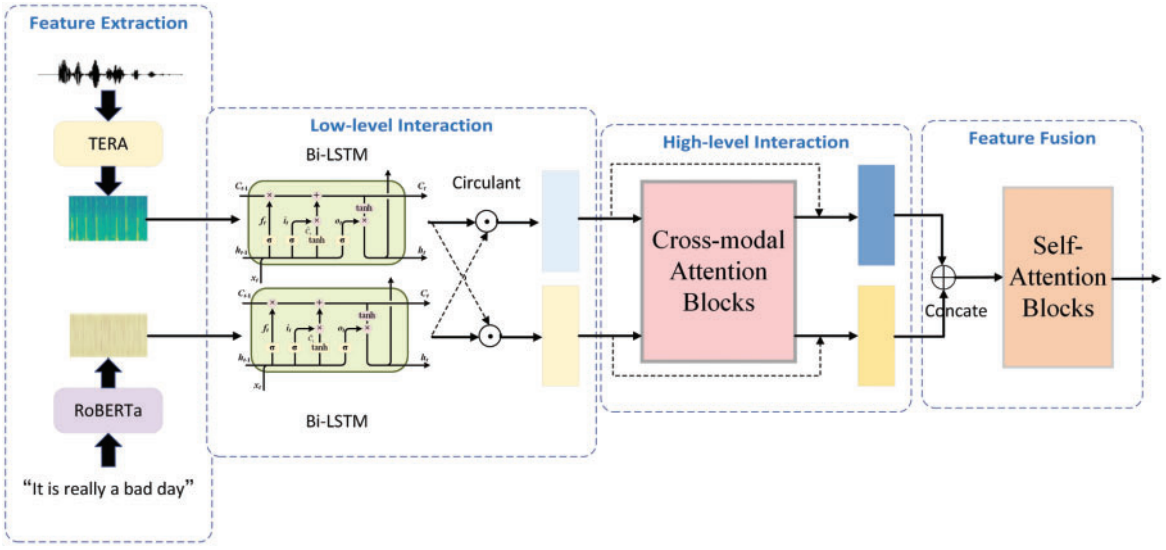

Figure 1: Overall architecture of proposed MLCCT

3.1 Self-supervised Embedding Layer

Rather than using low-level feature extractor, SSE models would be a better choice due to its powerful representation capabilities. SSE models are usually pre-trained on pretext tasks with a large number of unlabeled data in advance, and then generate more valuable feature representations for downstream tasks. Specially, TERA and RoBERTa are used to represent audio signal and textual sequence, respectively.

TERA introduces a total of three types of alteration to pre-train Transformer encoders on a large amount of unlabeled speech, namely time alteration, frequency alteration and magnitude alteration. Among them, the time alteration enables to learn richer phonetic content through contextual understanding of previous and future content., the frequency alteration effectively encodes speaker identity and the magnitude alteration improves performance by increasing data diversity for pre-training. We download the checkpoint of TERA from the open-sourced fairseq toolkit, which is pre-trained on LibriSpeech dataset for 960 h. The models' input is an 80-dimensional log ME.

RoBERTa is an extension study on BERT pre-training, carefully measuring the effects of many key hyperparameters and training data sizes. The improvements specifically include the following points. Firstly, RoBERTa has a larger training dataset with the addition of CC-NEWS, OPEN WEB TEXT and STORIES, while expanding batch size from 256 to 8 K. Secondly, dynamic masks are proposed to replace the original static masks to perform data augmentation for larger datasets. Finally, the next sentence prediction mechanism is eliminated, which is proved to be of little use. In this work, we train RoBERTa following the $\text{BERT}_{\text{LARGE}}$ architecture, which contains a 24-layer Transformer with 16 self-attention heads and 1024 hidden dimensions. The model is pre-trained on the BOOKCORPUS and WIKIPEDIA datasets with 100 K steps.

3.2 Multi-level Interaction Module

The proposed multi-level interaction module contains low-level interaction sub-module and high-level interaction sub-module, which capture the correlations between modalities in a progressive way. With the low-level feature interaction sub-module, MLCCT can extract fine-grained local features from speech signals and text sequences. Considering the superiority of RNN in handling sequence tasks, we employ a Bi-LSTM with circulant interaction mechanism, as shown in Fig. 2. LSTM is an extension of RNN, which can effectively solve the problem of gradient disappearance or explosion. LSTM makes some improvements and optimization based on the structure of RNN, adding memory cells and updating memory by input gates and forget gates. When LSTM is dealing with the information at current timestamp t, it receives a total of three input vectors x_t, h_{t-1} and C_{t-1}, where x_t is the input of current timestamp, h_{t-1} and C_{t-1} are the output and memory cell state of previous timestamp, respectively. The specific process can be described by the following equation.

$$f_t = \sigma\left(\delta_f \cdot x_t + \varphi_f \cdot h_{t-1} + \varepsilon_f\right) \tag{1}$$

$$i_t = \sigma\left(\delta_i \cdot x_t + \varphi_i \cdot h_{t-1} + \varepsilon_i\right) \tag{2}$$

$$o_t = \sigma\left(\delta_o \cdot x_t + \varphi_o \cdot h_{t-1} + \varepsilon_o\right) \tag{3}$$

$$\tilde{C}_t = \tanh\left(\delta_o \cdot [x_t,\ h_{t-1}] + \varepsilon_C\right) \tag{4}$$

where f_t, i_t and o_t stand for forget gate, input gate and output gate, respectively. δ is weight matrix from the input layer to the hidden layer, while φ denotes weight matrix from the hidden layer to the hidden layer and ε is the offset matrix. In addition, σ represents sigmoid activation function. The LSTM first determines current state C_t by controlling the memory cell how much prior information C_{t-1} is

forgotten and how much temporary state $\tilde{C}_t$ is retained through forgetting gates f_t and input gates i_t. And then, the output gate o_t is used to decide how to output the hidden state layer h_t by computing C_t in memory cell.

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

$$h_t = o_t * \tanh(C_t) \tag{6}$$

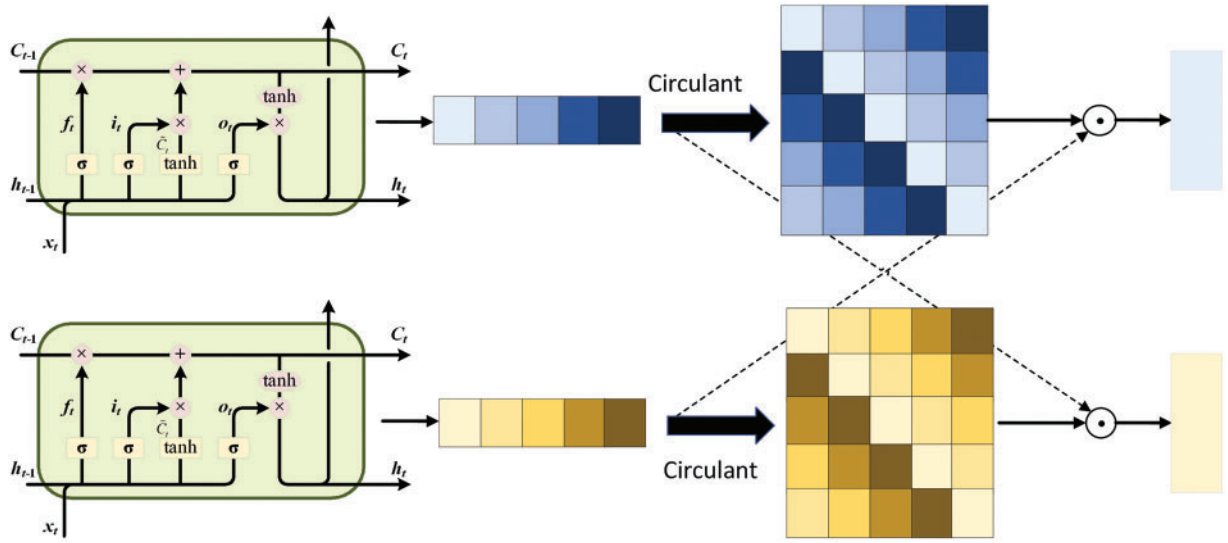

Figure 2: Detailed procedures of Bi-LSTM with circulant interaction mechanism

After Bi-LSTM, the circulant interaction mechanism is put forward to capture inter-modal correlation in an early stage. Specially, the intermediate results of audio $a \in \mathbb{R}^a$ and text $s \in \mathbb{R}^s$ will constructed as circulant matrices and multiplied by each other, where elements in each row perform a shift operation without changing the values as a whole. To further reduce computational cost, two weight matrices $W_a \in \mathbb{R}^{d\times a}$ and $W_s \in \mathbb{R}^{d\times s}$ are involved to project both feature vectors to a lower dimensional space. The detailed procedures are illustrated in Eqs. (7)–(9).

$$A = \tilde{A} \odot (W_s \cdot s) = \frac{1}{d}\sum_{i=1}^{d} a_i \odot (W_s \cdot s) \tag{7}$$

$$S = \tilde{S} \odot (W_a \cdot a) = \frac{1}{d}\sum_{i=1}^{d} s_i \odot (W_a \cdot a) \tag{8}$$

where

$$\tilde{A} = circ\_matrix(W_a \cdot a)\ ,\ \tilde{S} = circ\_matrix(W_s \cdot s) \tag{9}$$

where $a_i \in \mathbb{R}^d$ and $s_i \in \mathbb{R}^d$ are row vectors of circulant matrices, $\odot$ is denoted as Hadmard product. Through the element-wise product, we complete the preliminary interaction and get the integrated vectors $A \in \mathbb{R}^d$ and $S \in \mathbb{R}^d$.

As shown in Fig. 3, a two-stream RCTB is proposed for high-level interaction between audio and text. The RCTB takes the low-level integrated vectors as input. Since A and S share same dimension, there are no additional adjustment operations. RCTB takes advantage of the cross-modal attention

mechanism to explore inter-modal relationships and strengthen the target modal representation. In addition, shortcuts are employed to learn the residuals, which makes it easier for the model to converge. Take audio as the query vector $Q_A = AW_{Q_A} \in \mathbb{R}^{d\times 1}$, text as the key vector $K_S = SW_{K_S} \in \mathbb{R}^{d\times 1}$ and value vector $V_S = SW_{V_S} \in \mathbb{R}^{d\times 1}$, where W_{Q_A}, W_{K_S} and W_{V_S} are weights. Then a scaled dot-product attention is computed, and the specific process can be seen in the following formula.

$$Y_A = A + softmax\left(\frac{\sum_{i=1}^{d} Q_{A_i} \odot K_S}{\sqrt{d}}\right) V_S \tag{10}$$

where $Y_A \in \mathbb{R}^{d\times 1}$ is the final integrated vector of audio. In the same way, we can also obtain the final interaction vector with text as the target modality Y_s.

$$Y_S = S + softmax\left(\frac{\sum_{i=1}^{d} Q_{S_i} \odot K_A}{\sqrt{d}}\right) V_A \tag{11}$$

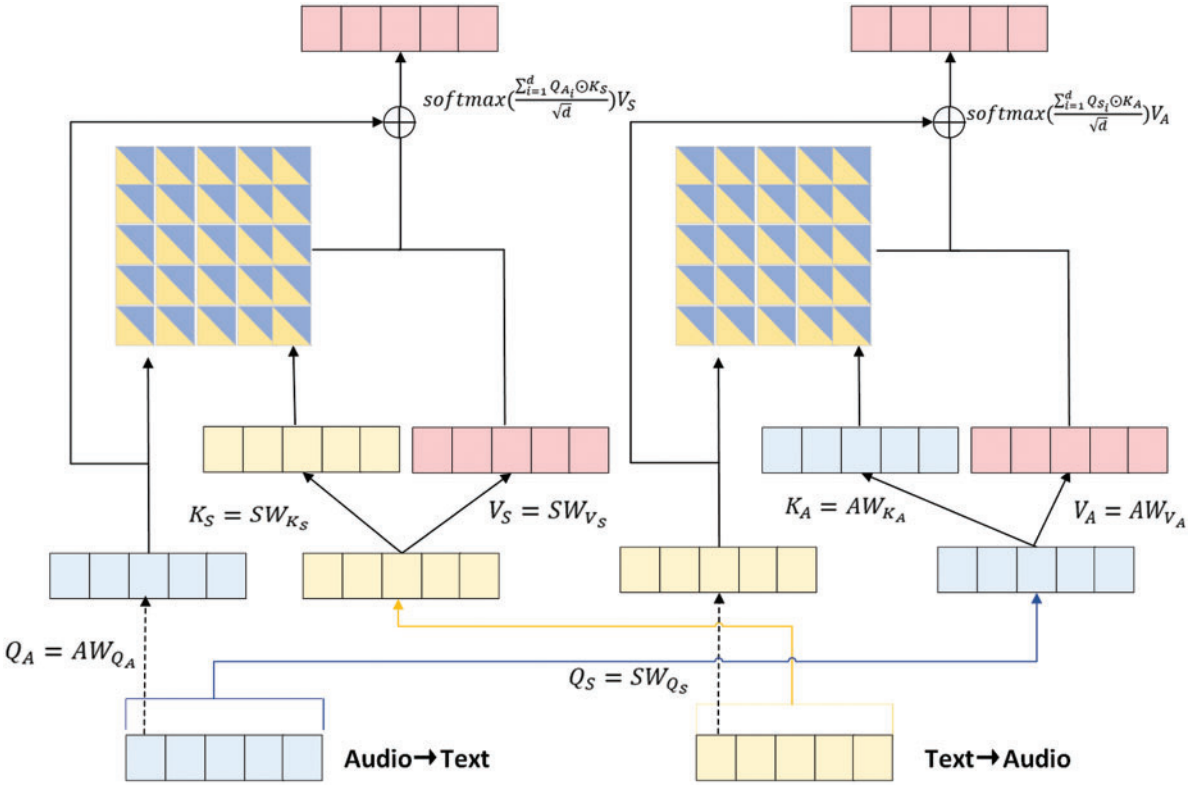

Figure 3: Structure of two-stream residual cross-modal Transformer block

For fusion, we first concatenate two final integrated vectors Y_A and Y_S, and then feed them into multi-head self-attention blocks. The output from self-attention layer will pass through a fully connected layer to make prediction. Multi-head attention allows the model to jointly attend to information from different representation subspaces.

$$Y = FC\left(W^O\left(\sum_{i=1}^{h} \Theta\left(Q_\tau W_i^Q, K_\tau W_i^K, V_\tau W_i^V\right)\right)\right) \tag{12}$$

where

$$Q_\tau = K_\tau = V_\tau = concat\,(Y_A, Y_S) \tag{13}$$

where FC denotes fully connected layer, Θ is self-attention computing and h is the number of head. Compared with somewhat straightforward fusion, self-attention blocks may lead to a further performance gain.

4 Experiment and Discussion

We carried out our experiments on three widely used benchmark datasets including IEMOCAP, MELD and CMU-MOSEI to verify the effectiveness of our proposed MLCCT. The testing environment was conducted on one single Nvidia RTX 3090 graphic card with 16GB memory, and an Intel(R) Core(TM) i7–7700 3.60 GHz.

4.1 Dataset

IEMOCAP collects data from five sessions of ten male and female participants, each session consisting of two unique participants. IEMOCAP is segmented by dialogue and each dialogue is annotated with categorical labels, such as angry, happy, sad, neutral, surprised, etc. We divided the whole dataset into five subsets, using the first four dialogues as training and validation and the last one as testing, which excludes speaker-related interference and is useful for real-life scenario applications.

The MELD dataset has over 12,000 discourses from the Old Friends TV series. Unlike other datasets, MELD is a conversational dataset, with several speakers participating in a single dialogue. Each dialogue is annotated with any of seven emotions: anger, disgust, sadness, joy, surprise, fear, and neutrality.

The CMU-MOSEI dataset is the largest sentence-level sentiment analysis and emotion recognition dataset available in online video. CMU-MOSEI contains over 65 h of annotated video from over 1,000 speakers and 250 topics. Similar to the MELD dataset, CMU-MOSEI is divided into three groups: training set, validation set, and testing set. We performed data statistics for the three benchmark datasets mentioned above, which are recorded in Tabs. 1 and 2.

Table 1: Amount of data for each type of emotion in IEMOCAP

Emotion	IEMOCAP		
	Training	Validation	Testing
Happy	198	28	58
Sad	426	60	122
Angry	202	29	58
Neutral	769	110	220

Table 2: Amount of data for each type of emotion in MELD and CMU-MOSEI

Emotion	MELD			CMU-MOSEI		
	Training	Validation	Testing	Training	Validation	Testing
Neutral	4513	450	1204	-	-	-
Happy	1661	157	380	7136	636	1438
Fear	264	36	50	257	28	55
Sad	674	107	204	2669	279	502
Disgust	266	21	68	767	61	150
Angry	1065	148	328	1607	136	384
Surprise	1150	142	207	348	29	78

4.2 Comparative Study

4.2.1 Experiments on IEMOCAP

Due to the uneven distribution of samples of each category in the IEMOCAP dataset, we selected the four most used emotion categories (neutral, angry, sad and happy) for our classification experiments. The comparison of the proposed MLCCT with prior studies on the IEMOCAP dataset can be clearly seen in Tab. 3, and mean accuracy is used as the evaluation metric. It confirms that multimodal speech emotion recognition tends to outperform those based on unimodal, audio or text. Compared with Lex-eVector, which is also based on multimodal data, MLCCT improves nearly 7% in accuracy, further verifying that our proposed progressive framework has a greater effect on the interaction and fusion in multimodal speech emotion recognition.

Table 3: Comparison of MLCCT with prior models on IEMOCAP

Model	Modality	Acc
LSTM [47]	Audio	54
Extreme Learning Machine (ELM) [48]	Audio	54.3
Hierarchical Decision Tree [49]	Audio	56.83
RNN-ELM [50]	Audio	62.85
ACNN [51]	Audio	62.11
RNN + Attention [52]	Audio	63.5
CNN + LSTM [53]	Audio	64.5
Bi-LSTM-ELM + LLDs [54]	Audio	64.2
Greedy-Attention [55]	Audio	59.4
FCN + Attention [56]	Audio	70.2
Bi-LSTM + Context-aware Attention [57]	Audio	68.8
Lex-eVector [58]	Audio	57.4
	Text	53.5
	Multimodal	69.2

(Continued)

Table 3: Continued

Model	Modality	Acc
Ensemble model of CNN and LSTM [59]	Audio	62.72
	Text	64.78
MLCCT	Audio	66.19
	Text	56.5
	Multimodal	**75.92**

To further analyze the recognition results, Tab. 4 shows the accuracy and F1 scores for each category, and Fig. 4 plots the confusion matrix. From the experimental results, we can find that our model has better performance in terms of accuracy compared to RAVEN [60] and MCTN [61] models. In terms of the MulT model, MLCCT performs better overall, except for a 1.1% lower classification accuracy in Anger. We believe that the improvement stems from placing the Transformer further back in the network and thus more adapted to high-level features. In brief, our model achieved significant recognition results on the IEMOCAP dataset.

Table 4: Performance for each category on IEMOCAP

	Happy		Sad		Angry		Neutral	
Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RAVEN [60]	77	76.8	67.6	65.6	65	64.1	62	59.5
MCTN [61]	80.5	77.5	72	71.4	64.9	65.6	49.4	49.3
MulT [35]	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
MLCCT	**84.7**	**82.8**	**78.1**	**75.2**	**72.8**	**72.1**	**65.1**	**62.5**

Fig. 4 shows in more detail the specific effects of the model on each category. It can be seen that MLCCT achieves the best recognition in "Happy", followed by "Sad" and "Angry", and the last one is "Neutral". We speculate that it may be "Neutral" that is the most vaguely defined and is more difficult to represent in a multimodal form.

4.2.2 Experiments on MELD

The evaluation of our proposed model compared to the conventional studies on MELD is presented in Tab. 5. It can be seen that MLCCT outperforms any other model in terms of average accuracy and F1 score. Although the improvement is limited, for example, its accuracy is only 1.5% higher than that of [102]. It is believed that the reason leading to this result may be the more complex speakers and scenes in MELD.

A more detailed analysis is conducted based on each category on the MELD. As can be seen in Fig. 5, the recognition accuracy of our proposed model inevitably decreases as the number of emotion categories increases. Specifically, MLCCT performs well on Neutral, Joy and Surprise, reaching an accuracy of 78.68 on Neural. However, its recognition results on Sadness and Disgust are not satisfactory, especially sadness is often misclassified as Neutral. It is speculated that the reason leading to this result may lie in the uneven distribution of the various types of data in the MELD dataset, with relatively little disgust and sadness data.

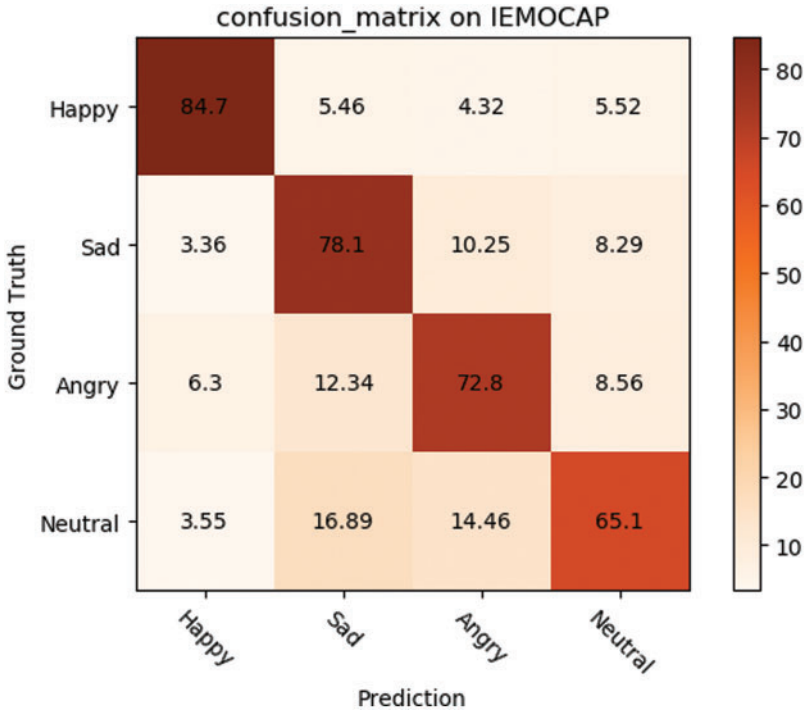

Figure 4: The recognition rate confusion matrix on IEMOCAP

Table 5: Comparison of MLCCT with prior models on MELD

Model	Modality	Acc	F1 score
AGHMN [62]	Text	60.3	58.1
KET [63]	Text	60.6	58.2
Confidence-estimator Ensemble Model [64]	Multimodal	61.2	59.5
Con-GCN [65]	Audio	49.32	47.4
	Text	45.61	42.2
	Multimodal	61.7	59.4
MLCCT	Audio	48.8	45.34
	Text	61.7	58.9
	Multimodal	**63.2**	**62.4**

4.2.3 Experiments on CMU-MOSEI

Tab. 6 presents the comparison of the proposed model and prior works on CMU-MOSEI dataset in terms of mean accuracy and F1 score. The experimental results again demonstrate the fact that speech emotion recognition based on multimodal data tend to have better accuracy than those relying on unimodal data alone. Moreover, our proposed multi-level model also proves to be a better choice, achieving an accuracy of 51.2% and an F1 score of 82.0.

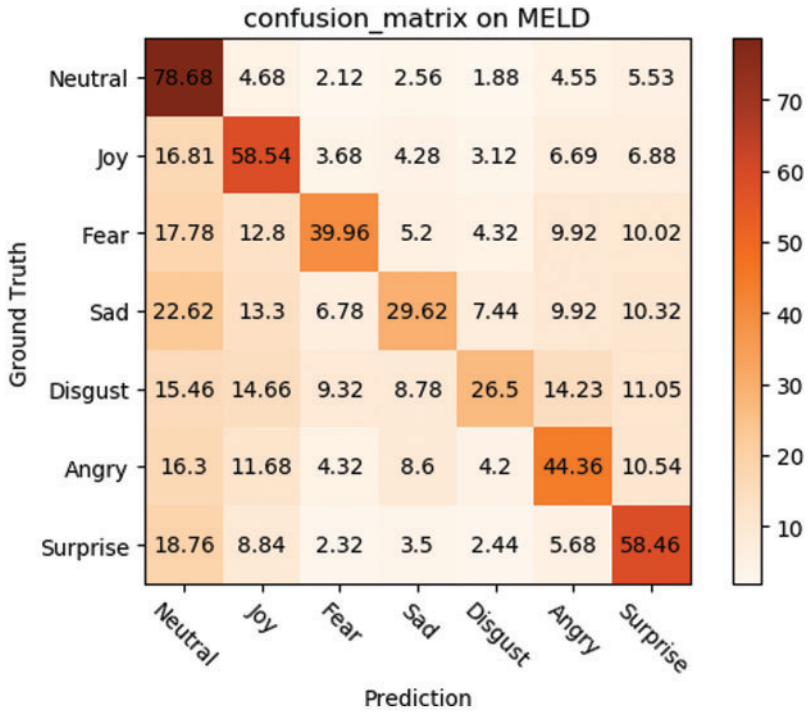

Figure 5: The recognition rate confusion matrix on MELD

Table 6: Comparison of MLCCT with prior models on CMU-MOSEI

Model	Modality	Acc	F1 score
RTN [66]	Multimodal	48.12	62.12
CIM [67]	Audio	30.25	38.72
	Text	42.6	59.22
	Multimodal	49.12	66.6
GMU + Attention [68]	Audio	31.75	43.65
	Text	43.58	60.82
	Multimodal	50.31	72.21
MLCCT	Audio	35.6	46.56
	Text	48.2	62.78
	Multimodal	**51.2**	**82.0**

As in the previous two subsections, we also provide a more specific analysis of the results for each category on CMU-MOSEI shown in Fig. 6. It can be further seen that our model has certain generalization ability and can still cope well with the overall accuracy when facing the problem of unbalanced data across emotion labels. In particular, MLCCT achieved the best accuracy of 79.82% in Happy, and the worst performance in Fear.

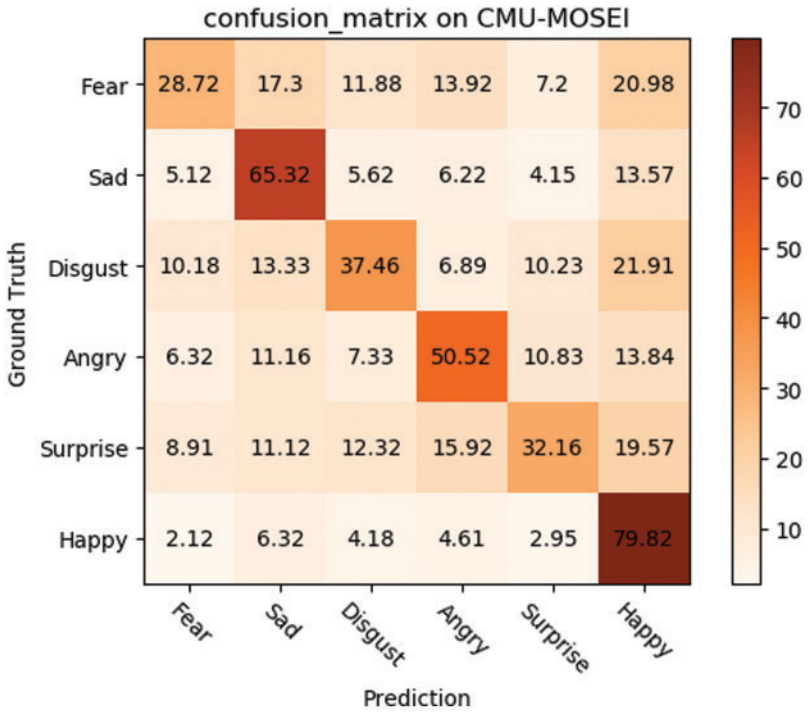

Figure 6: The recognition rate confusion matrix on CMU-MOSEI

4.3 Ablation Study

4.3.1 Ablation Study on Multi-Level Interaction

To demonstrate the effectiveness of the progressive framework, we conduct an ablation study on multi-level interaction. Specifically, we design the following strategies for comparison, as shown in Tab. 7.

Table 7: Ablation study on multi-level interaction on CMU-MOSEI

	SSE	Low-level interaction	High-level interaction	Acc	F1 score
A	√	√		28.12	30.18
B	√		√	42.71	72.51
C		√	√	45.25	75.62
D	√	√	√	51.20	82.03

The experimental results in Tab. 7 show that the model containing only single-level interactions is limited in terms of recognition results. In particular, the accuracy of Strategy B, which only employs a Bi-LSTM with circulant interaction mechanism for low-level features, dropped significantly to 28.12%. Without the low-level interactions, the model in Strategy C also encountered a bottleneck, achieving an accuracy of 42.71% and an F1 score of 72.51. In addition, SSE also proved to be effective, which helped to improve the model's accuracy rate by about 6%.

4.3.2 Ablation Study on Attention Blocks

A total of two types of attention mechanisms are included in the structure of this network, which are cross-modal attention mechanism and self-attention mechanism. To further explore the influence of these two attention mechanisms on the final recognition accuracy, we select the CMU-MOSEI dataset with the largest amount of data for the ablation experiments, as shown in Tab. 8.

Table 8: Ablation study on attention blocks on CMU-MOSEI

Cross-modal attention block	Self-attention block	Cross-model head	Self-attention head	ACC-6	ACC-2	F1 score
1	1	1	1	49.7%	79.6%	80.2
1	1	2	4	50.1%	80.2%	80.6
2	4	1	1	50.9%	82.4%	81.5
2	4	2	4	51.2%	82.9%	82.0
2	4	4	8	51.3%	83.6%	82.7
4	8	2	4	51.4%	83.5%	82.5
4	8	4	8	51.2%	82.1%	81.7

It is known that as the number of attention blocks increases, the accuracy of the model improves more slowly and the training difficulty and recognition time also increase. Therefore, combining various factors, we finally chose two cross-modal attention blocks with four heads and two self-attention blocks with four heads as hyperparametric settings.

5 Conclusion

In this paper, we propose a multi-level circulant cross-modal Transformer (MLCCT) for multimodal speech emotion recognition. Different from prior works, MLCCT adopts a progressive framework, which combines local and global features for accurate predictions. To the best of our knowledge, this is the first time that a Transformer-based progressive framework is used in multimodal speech emotion recognition, which will serve as an inspiration for future research. Specifically, MLCCT better captures inter-modal relationships through two modal interaction processes. The first one employs a bidirectional Long Short-term Memory (Bi-LSTM) with circulant interaction mechanism for low-level features, while a two-stream residual cross-modal Transformer block is applied when high-level features are involved. Finally, self-attention blocks are used for fusion. Comprehensive experimental results on three benchmark datasets including IEMOCAP, MELD and CMU-MOSEI show that the proposed MLCCT has achieved competitive results.

Acknowledgement: Peizhu Gong: Conceptualization, Methodology, Writing-original draft, Software. Jin Liu: Supervision, Project administration, Formal analysis, Data curation. Huihua He: Supervision, Project administration, Data curation. Bing Han: Visualization, Software, Resources. Zhongdai Wu: Resources, Validation. Y. Ken Wang: Data curation.

Funding Statement: This work was supported by the National Natural Science Foundation of China (No. 61872231), the National Key Research and Development Program of China (No. 2021YFC2801000), and the Major Research plan of the National Social Science Foundation of China (No. 2000&ZD130).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. M. S. A. Abdullah, S. Y. A. Ameen and S. Zeebaree, "Multimodal emotion recognition using deep learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 2, pp. 52–58, 2021.

[2] M. Ayadi, M. S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, pp. 1–9, 2016.

[4] S. Yoon, S. Byun and K. Jung, "Multimodal speech emotion recognition using audio and text," in *IEEE Spoken Language Technology Workshop (SLT)*, Greece, pp. 112–118, 2018.

[5] S. Yoon, S. Byun, S. Dey and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, pp. 2822–2826, 2019.

[6] P. Gong, J. Liu, Y. Yang and H. He, "Towards knowledge enhanced language model for machine reading comprehension," *IEEE Access*, vol. 8, pp. 224837–224851, 2020.

[7] X. Jiang, F. Yu, T. Song and V. C. Leung, "Resource allocation of video streaming over vehicular networks: A survey, some research issues and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 37, pp. 1–30, 2021.

[8] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self-supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.,* "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[10] N. Sato and Y. Obuchi, "Emotion recognition using Mel-frequency cepstral coefficients," *Information and Media Technologies*, vol. 2, no. 3, pp. 835–848, 2007.

[11] N. Alsaaran and M. Alrabiah, "Classical arabic named entity recognition using variant deep neural network architectures and BERT," *IEEE Access*, vol. 9, pp. 91537–91547, 2021.

[12] X. Zhai, A. Oliver, A. Kolesnikov and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea, pp. 1476–1485, 2019.

[13] J. Xia, Y. Lu, L. Tan and P. Jiang, "Intelligent fusion of infrared and visible image data based on convolutional sparse representation and improved pulse-coupled neural network," *Computers, Materials & Continua*, vol. 67, no. 1, pp. 613–624, 2021.

[14] M. H. Changrampadi, A. Shahina, M. B. Narayanan and A. N. Khan, "End-to-end speech recognition of tamil language," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1309–1323, 2022.

[15] R. Chen, L. Pan, C. Li, Y. Zhou, A. Chen *et al.,* "An improved deep fusion CNN for image recognition," *Computers, Materials & Continua*, vol. 65, no. 2, pp. 1691–1706, 2020.

[16] J. Devlin, M. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv: 1810.04805, 2018.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollar *et al.,* "Masked autoencoders are scalable vision learners," arXiv preprint arXiv: 2111.06377, 2021.

[18] A. Liu, S. Li, and H. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi *et al.,* "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv: 1907.11692, 2019.

[20] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower *et al.,* "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] S. Poria, N. Majumder, R. Mihalcea and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
[22] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, L. Morency *et al.,* "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, vol. 1, pp. 2236–2246, 2018.
[23] S. Arifin and P. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1325–1341, 2008.
[24] M. Bhargava and T. Polzehl, "Improving automatic emotion recognition from speech using rhythm and temporal feature," arXiv preprint arXiv: 1303.1761, 2013.
[25] H. K. Palo, M. Chandra and M. N. Mohanty, "Recognition of human speech emotion using variants of Mel-frequency cepstral coefficients," *Advances in Systems, Control and Automation*, vol. 442, pp. 491–498, 2018.
[26] S. Chang and J. Liu, "Multi-lane capsule network for classifying images with complex background," *IEEE Access*, vol. 8, pp. 79876–79886, 2020.
[27] D. Zeng, K. Liu, S. Lai, G. Zhou and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. of COLING 2014, the 25th Int. Conf. on Computational Linguistics: Technical Papers*, Dublin, Ireland, pp. 2335–2344, 2014.
[28] W. Sun, G. Z. Dai, X. R. Zhang, X. Z. He, X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, pp. 1–13, 2021.
[29] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 8, pp. 1–16, 2021.
[30] W. Liu, J. Wang, L. Chen and B. Chen, "Prediction of protein essentiality by the improved particle swarm optimization," *Soft Computing*, vol. 22, no. 20, pp. 6657–6669, 2018.
[31] J. Liu, Y. Yang and H. He, "Multi-level semantic representation enhancement network for relationship extraction," *Neurocomputing*, vol. 403, no. 5, pp. 282–293, 2020.
[32] J. Liu, Y. Yang, S. Lv, J. Wang and H. Chen, "Attention-based BiGRU-CNN for Chinese question classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 1, pp. 1–12, 2019.
[33] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar *et al.,* "Speech emotion recognition using spectrogram and phoneme embedding," in *Proc. of Interspeech 2018*, Hyderabad, India, pp. 3688–3692, 2018.
[34] J. Zhao, M. Xia and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.
[35] Y. H. Tsai, S. Bai, P. Liang, J. Z. Kolter, L. Morency *et al.,* "Multimodal transformer for unaligned multimodal language sequences," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Firenze, pp. 6558–6569, 2019.
[36] J. Delbrouck, N. Tits, M. Brousmiche and S. Dupont, "A Transformer-based joint-encoding for emotion recognition and sentiment analysis," arXiv preprint arXiv: 2006.15955, 2020.
[37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei *et al.,* "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, pp. 9, 2019.
[38] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings," arXiv preprint arXiv: 1909.00512, 2019.
[39] A. Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals and A. Graves, "Conditional image generation with pixel-CNN decoders," in *Advances in Neural Information Processing Systems*, Spain, vol. 29, pp. 4790–4798, 2016.
[40] A. Oord and O. Vinyals, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, USA, vol. 16, pp. 6306–6315, 2017.
[41] R. Hjelm, A. Fedorov, S. Marchildon, K. Grewal, P. Bachman *et al.,* "Learning deep representations by mutual information estimation and maximization," arXiv preprint arXiv:1808.06670, 2018.

[42] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 9729–9738, 2020.
[43] T. Chen, S. Kornblith, M. Norouzi and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Int. Conf. on Machine Learning (PMLR)*, London, UK, pp. 1597–1607, 2020.
[44] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun *et al.,* "Clear: Contrastive learning for sentence representation," arXiv preprint arXiv: 2012.15466, 2020.
[45] F. Carlsson, A. C. Gyllensten, E. Gogoulou, E. Y. Hellqvist and M. Sahlgren, "Semantic re-tuning with contrastive tension," in *Int. Conf. on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, pp. 1–21, 2020.
[46] I. Goodfellow, J. Abadie, M. Mirza, B. Xu, D. Farley *et al.,* "Generative adversarial nets," *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, pp. 1–10, 2014.
[47] V. Chernykh and P. Prikhodko, "Emotion recognition from speech with recurrent neural networks," arXiv preprint arXiv: 1701.08071, 2017.
[48] K. Han, D. Yu and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech*, Singapore, pp. 223–227, 2014.
[49] C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach, " *Speech Commun*, vol. 53, no. 9, pp. 1162–1171, 2011.
[50] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech*, pp. 1–4, 2015.
[51] M. Neumann and N. T. Vu, "Attentive convolutional neural network-based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," arXiv preprint arXiv: 1706.00612, 2017.
[52] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New OLEANS, USA, pp. 2227–2231, 2017.
[53] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers and B. Schmauch, "CNN + lSTM architecture for speech emotion recognition with data augmentation," arXiv preprint arXiv: 1802.05630, 2018.
[54] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Seventh Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, San Antonio, Texas, pp. 190–195, 2017.
[55] C. Huang and S. Narayanan. "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Interspeech*, Singapore, pp. 1387–1391, 2016.
[56] Y. Zhang, J. Du, Z. Wang, J. Zhang and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, Honolulu, Hawaii, USA, pp. 1771–1775, 2018.
[57] G. Ramet, P. N. Garner, M. Baeriswyl and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, Greece, pp. 126–131, 2018.
[58] Q. Jin, C. Li, S. Chen and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Queensland, Australia, pp. 4749–4753, 2015.
[59] S. Tripathi, S. Tripathi and H. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," arXiv preprint arXiv: 1804.05788, 2018.
[60] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh *et al.,* "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Honolulu, Hawaii, USA, vol. 33, no. 1, pp. 7216–7223, 2019.
[61] H. Pham, P. P. Liang, T. Manzini, L. P. Morency and B. Poczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Honolulu, Hawaii, USA, vol. 33, no. 1, pp. 6892–6899, 2019.

[62] W. Jiao, M. Lyu and I. King, "Real-time emotion recognition via attention gated hierarchical memory network," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, USA, vol. 34, no. 5, pp. 8002–8009, 2020.
[63] P. Zhong, D. Wang and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," arXiv preprint arXiv: 1909.10681, 2019.
[64] U. Nadeem, M. Bennamoun, F. Sohel and R. Togneri, "Learning-based confidence estimation for multi-modal classifier fusion," in *Int. Conf. on Neural Information Processing*, Canada, pp. 299–312, 2019.
[65] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu *et al.,* "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proc. of the Twenty-Eighth Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Macao, China, pp. 5415–5421, 2019.
[66] S. Sahay, S. H. Kumar, R. Xia, J. Huang and L. Nachman, "Multimodal relational tensor network for sentiment and emotion classification," arXiv preprint arXiv: 1806.02923, 2018.
[67] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal *et al.,* "Multi-task learning for multi-modal emotion recognition and sentiment analysis," arXiv preprint arXiv:1905.05812, 2019.
[68] S. Sangwan, D. S. Chauhan, M. Akhtar, A. Ekbal, P. Bhattacharyya *et al.,* "Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis," in *Int. Conf. on Neural Information Processing*, Canada, pp. 662–669, 2019.