M^{3}SA: Multimodal Sentiment Analysis Based on Multi-Scale Feature Extraction and Multi-Task Learning

Changkai Lin, Hongju Cheng, *Senior Member, IEEE*, Qiang Rao, and Yang Yang, *Member, IEEE*

***Abstract*—Sentiment analysis plays an indispensable part in human-computer interaction. Multimodal sentiment analysis can overcome the shortcomings of unimodal sentiment analysis by fusing multimodal data. However, how to extracte improved feature representations and how to execute effective modality fusion are two crucial problems in multimodal sentiment analysis. Traditional work uses simple sub-models for feature extraction, and they ignore features of different scales and fuse different modalities of data equally, making it easier to incorporate extraneous information and affect analysis accuracy. In this paper, we propose a Multimodal Sentiment Analysis model based on Multi-scale feature extraction and Multi-task learning (M^{3}SA). First, we propose a multi-scale feature extraction method that models the outputs of different hidden layers with the method of channel attention. Second, a multimodal fusion strategy based on the key modality is proposed, which utilizes the attention mechanism to raise the proportion of the key modality and mines the relationship between the key modality and other modalities. Finally, we use the multi-task learning approach to train the proposed model, ensuring that the model can learn better feature representations. Experimental results on two publicly available multimodal sentiment analysis datasets demonstrate that the proposed method is effective and that the proposed model outperforms baselines.**

***Index Terms*—Multimodal sentiment analysis, multi-scale feature extraction, multi-task learning, multimodal data fusion.**

I. Introduction

THE sentiment is essentially important in human decision-making, social activities, and creativity. Sentiment analysis is an indispensable part of human-computer interaction [1]. The earliest systematic research in sentiment analysis with machine learning dates back to 2002. Pang et al. [2] utilized three machine learning algorithms to analyze the sentiment of movie reviews and verified the feasibility of employing computers to analyze human sentiment. Subsequently, many machine learning algorithms, such as Naive Bayes [3] and support vector machine [4], had been exploited for sentiment analysis. Human sentiments are generally composed of multiple modalities, and different modalities may express opposite sentiments, which results in less accuracy if we analyze the sentiment with only unimodal data.

It is a crucial way to improve the accuracy of sentiment analysis and overcome the limitations of unimodal by exploiting the complementarity of multimodal data [5]. Zadeh et al. [6] proposed a Tensor Fusion Network (TFN) to calculate a high-dimensional tensor (based on the outer product operation) to fuse bimodal and trimodal. Tsai et al. [7] proposed a Multimodal Transformer (MuLT) that utilizes the cross-modal attention mechanism to fuse two modality data and does not need multimodal data alignment operation. Hazarika et al. [8] projected each modality into two subspaces. The first one learns similar features of the modalities, and the latter captures the individual independent characteristics.

The current multimodal sentiment analysis works can be improved in two aspects. Firstly, the previous multimodal sentiment analysis works only utilized a simple feature extraction sub-network to extract a single scale feature for each modality without considering the influence of the features with different scales on sentiment analysis. Secondly, The previous multimodal sentiment analysis works adopted an equal strategy to fuse multimodal data and ignored the impact of different modalities on the sentiment analysis results.

In the paper, we propose a multi-scale feature extraction method with the channel attention mechanism to address the first problem above. The method is carried out with two steps. *a*) we use the channel attention mechanism to model the different scale features of each hidden layer and assign weights to each feature in the feature extraction process. *b*) we dynamically adjust the weights of features of different scales according to the accuracy of sentiment analysis results so that the model can extract more valuable features in the model training process. We also propose a novel multimodal fusion strategy based on the key modality to solve the second problem. Firstly, we sort the result accuracy of unimodal sentiment analysis and select the one with the highest

Manuscript received 11 May 2023; revised 25 December 2023; accepted 27 January 2024. Date of publication 1 February 2024; date of current version 14 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62372111 and in part by the Natural Science Foundation of Fujian Province of China under Grant 2023J01267. The work of Yang Yang was supported in part by the National Natural Science Foundation of China under Grant 62372110, in part by the Natural Science Foundation of Fujian Province of China under Grant 2023J02008, and in part by the Special Project for Research and Development in Key areas of Guangdong Province under Grant 2020B0101090005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zoraida Callejas. *(Corresponding author: Hongju Cheng.)*

Changkai Lin and Hongju Cheng are with the College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China (e-mail: changkai.lin@foxmail.com; cscheng@fzu.edu.cn).

Qiang Rao is with Fujian Jingwei Digital Technology Corporation Ltd., Fuzhou 350002, China (e-mail: 534368277@qq.com).

Yang Yang is with the School of Computing and Information Systems, Singapore Management University, Singapore 188065 (e-mail: yang.yang.research@gmail.com).

Digital Object Identifier 10.1109/TASLP.2024.3361374

value as the key modality. Secondly, we use the key modality to strengthen the remaining in the process of multimodal fusion, which makes the amount of helpful information more abundant. Finally, we adjust the proportion of the key modality according to the accuracy of sentiment analysis in model training.

The main contributions are summarized as follows:

- We propose a multi-scale feature extraction method with the channel attention to model different scale features. The weights of those different-scale features are dynamically adjusted according to the accuracy of sentiment analysis results during the training process.
- We propose a novel multimodal fusion strategy based on the key modality. During the fusion process, using the key modality enhances the information of the other modalities and enriches the fused helpful information.
- We carry out extensive experiments on the two publicly available datasets, CMU-MOSI and CH-SIMS. The experimental results demonstrate that our proposed method surpasses the baselines.

II. Related Work

A. Multi-Scale Features Extraction

Multi-scale feature extraction is a widespread technique in deep learning [9-11], which can extract different scales of features by performing multiple convolution and pooling operations for the input data, enabling a more helpful feature representation. Lin et al. [12] have developed a feature pyramid network with a laterally connected top-down architecture that aggregates different scale image features extracted from different convolutional kernels for target detection. Li et al. [13] proposed a novel multi-scale residual network that utilizes convolutional kernels of different sizes to detect image features for the most meaningful image information adaptively. Sun et al. [14] proposed a multi-scale feature extraction network for decoding electromyographic signals, which uses channel attention and spatial attention to capture more critical features for gesture recognition.

Besides, many researchers also introduced multi-scale feature extraction into the sentiment analysis problem. Lei et al. [15] proposed a multi-scale emotional speech synthesis framework that can generate speech with emotion in three approaches. Cao et al. [16] proposed a multi-label sentiment analysis model with multi-scale CNNs by fusing different scales of text features to predict text sentiment. Guo et al. [17] proposed a multi-scale transformer by introducing prior knowledge and multi-scale structure into the self-attentive module. They designed a strategy to control the scale distribution of each layer. Song et al. [18] proposed a sentiment network with visual attention, which joins several scaled features extracted from different network layers to predict image sentiment.

B. Multi-Task Learning

Multi-task learning refers to fusing training data from different tasks and jointly optimizing model parameters to enhance the model's learning ability and generalization performance. Jin et al. [19] proposed a multi-task and multi-scale text sentiment analysis model. The model enhances the encoder's performance and improves sentiment classification accuracy by introducing multi-task learning. Yang et al. [20] proposed a two-phase multi-task sentiment analysis framework, which applies a two-phase training strategy to fully utilize the pre-trained model and multi-task learning strategy to improve the sentiment analysis accuracy. Yang et al. [21] proposed a multi-task learning framework called cross-modal multi-task transformer, which unites two unimodal assistance tasks to learn different intrinsic representations of modality. Majumder et al. [22] proposed a framework based on multi-task learning. The framework uses deep neural networks to model the correlation between sentiment classification tasks and sarcasm detection tasks to improve performance. Li et al. [23] propose a multi-task learning-based approach to multimodal sentiment analysis and emotion recognition. This approach employs a shared and private model to separately and jointly process sentiment and emotion features and utilizes the attention mechanism and BI-LSTM network for feature fusion. Experiments on the CMU-MOSEI dataset demonstrate the model's effectiveness and the potential of multi-task learning in enhancing multimodal sentiment analysis performancey.

C. Multimodal Sentiment Analysis

Cambria et al. [24] proposed a multimodal sentiment analysis framework based on deep learning and proved that multimodal data could result in more accurate sentiment analysis than unimodal data. Xu et al. [25] designed a novel multi-interactive memory network with two interactive memory networks to capture the interaction information between text and visual. Chu et al. [26] designed a novel deeply-fused audio-text bi-modal transformer with a cross-modal fusion and a staged cross-modal pre-training scheme to judge sentiment accurately. Han et al. [27] proposed a bi-bimodal fusion network by exploring the importance of text, audio, and visual modality. They introduced a gating mechanism to control the optimal weight of each modality in the fusion.

In addition to text, vision, and audio, physiological signals are also frequently used in sentiment analysis. The physiological signals that are often used for sentiment analysis. [28]. Zheng et al. [29] studied how to utilize eye movement and EEG for sentiment analysis. The results showed that the accuracy of sentiment analysis with EEG was 2.51% higher than that with eye movement, and the accuracy after fusion was 5.55% and 8.06% higher compared with EEG and eye movement separately. Zhu et al. [30] explored the use of EEG, peripheral physiological signals, and facial expressions for sentiment analysis. The results showed that facial expression achieved the best arousal (about 71%) in unimodal sentiment analysis. After fusing the facial expression with peripheral physiological signals, the arousal is increased by 0.52%. After further fusing EEG signals, the arousal is increased by 0.68% again. Zhong et al. [31] studied the fusion between facial expressions and physiological signals. They reported a classification accuracy of 50.57% for valence and 53.64% for arousal when only facial expressions were used. The classification accuracy for valence and valence was

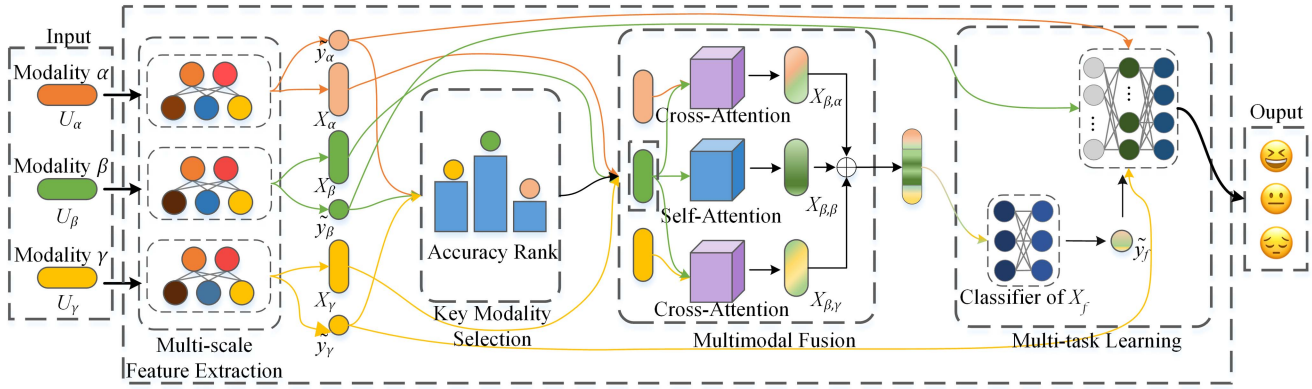

Fig. 1. Framework of the M^{3}SA in which Modality β is selected as the key modality.

improved by about 19.96% and 19.89% after fusing the two modalities. In addition to physiological signals, body movement also plays an important role in sentiment analysis problems.

In addition to physiological signals, body movement also plays an important role in sentiment analysis problems. Gunes et al. [32] presented a multimodal sentiment analysis approach to analyzing sentiment by fusing expressive face and upper-body gestures. Their experimental results show that the emotion classification using the two modalities combined achieves better recognition accuracy in general, outperforming the classification using the face or body alone. Castellano et al. [33] studied the effect of information from facial expressions, body movements, and speech. The results showed that the body movement achieved the highest accuracy (about 67.1%) in unimodal sentiment analysis. These studies show that physiological signals and body data help improve the accuracy of sentiment analysis.

Inspired by the prior works, In this paper, we propose a multimodal sentiment analysis model based on multi-scale feature extraction and multi-task learning. The model uses channel attention to model multi-scale unimodal features and dynamically adjusts the weights of different scale features according to sentiment analysis results in the training phase. A multimodal fusion strategy based on key modality is also proposed. The key modality is used to reinforce the information contained in the other modalities during multimodal fusion, making the amount of helpful information richer.

III. The Proposed Method

In this section, we present the overall framework of the M^{3}SA model and describe multi-scale feature extraction, key modality selection, multimodal fusion, and multi-task learning in the following sections. The overall structure of M^{3}SA is shown in Fig. 1.

A. Multi-Scale Feature Extraction

Multi-scale feature extraction can enhance a model's ability to understand multimedia data, such as images or text, and thus improve its performance on the task. We introduce a multi-scale feature extraction method to extract features at different scales for any modality data, as shown in Fig. 2.

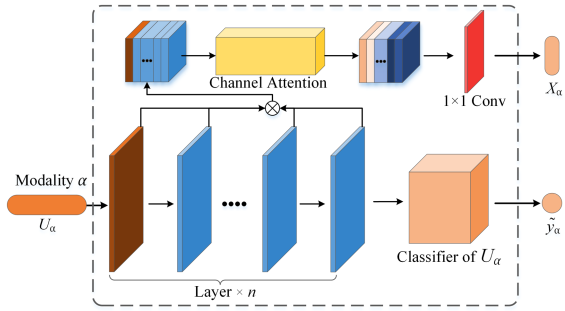

Fig. 2. Multi-scale feature extraction for modality α.

Suppose that a multimodal dataset contains N data samples $U = \{U_1, \ldots, U_N\}$, each sample U_i contains three different modality data $U_i = (U_\alpha, U_\beta, U_\gamma)$, where $U_\alpha = \{u_\alpha^1, u_\alpha^2, \ldots, u_\alpha^{T_\alpha}\}, U_\alpha \in \mathbb{R}^{T_\alpha \times D_\alpha}$; $U_\beta = \{u_\beta^1, u_\beta^2, \ldots, u_\beta^{T_\beta}\}, U_\beta \in \mathbb{R}^{T_\beta \times D_\beta}$; $U_\gamma = \{u_\gamma^1, u_\gamma^2, \ldots, u_\gamma^{T_\gamma}\}, U_\gamma \in \mathbb{R}^{T_\gamma \times D_\gamma}$, in which T_m denotes sequence length for modality m; D_m is the low-level feature dimensions for modality m; u_m^t is the low-level feature at timestep t; $m \in \{\alpha, \beta, \gamma\}$.

For low-level feature U_α, we use the LSTM network to extract the temporal features of modality α. LSTM is a temporal recurrent neural network that can solve the long-term dependence problem of traditional recurrent neural networks. It is suitable for processing and predicting events with long intervals and delays. The temporal feature of the modality α is extracted as shown:

$$F_\alpha^1 = \mathrm{LSTM}(U_\alpha). \tag{1}$$

After the temporal processing with the LSTM network, the first feature $F_\alpha^1 \in \mathbb{R}^{T_\alpha \times d_\alpha}$ of modality α is obtained, where d_α denotes the dimensional size of the temporal feature.

We use $(n-1)$ linear layers to extract $(n-1)$ different scales of features for F_α^1, as described in (2):

$$F_\alpha^{k+1} = \mathrm{LinearLayer}\left(\mathrm{NormLayer}\left(\mathrm{Relu}\left(\mathrm{F}_\alpha^{\mathrm{k}}\right)\right)\right), \tag{2}$$

where $k = 1, 2, \ldots, n-1$.

We stack all the scale features of modality α and obtain the multi-scale feature F_α:

$$F_\alpha = \mathrm{stack}\left(\mathrm{F}_\alpha^1, \mathrm{F}_\alpha^2, \ldots, \mathrm{F}_\alpha^{\mathrm{n}}\right), \tag{3}$$

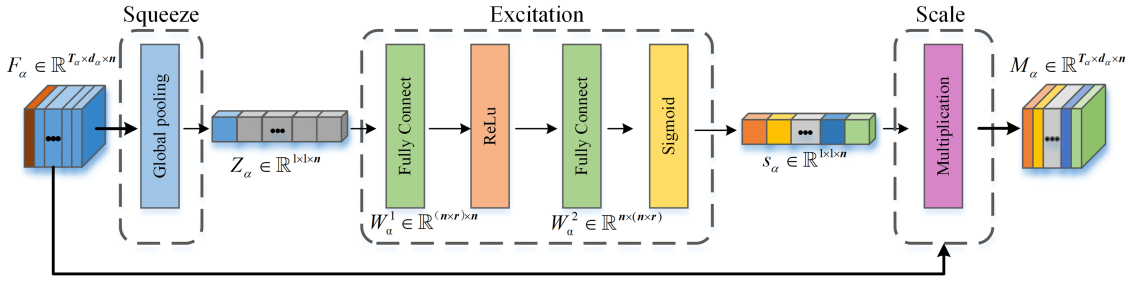

Fig. 3. Architecture of channel attention module.

where $F_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha \times n}$.

The attention mechanism can automatically learn and calculate the contribution of the input data to the final result and assign different weights to different input data to concern task-relevant data. We use the channel attention proposed in [34] to model the degree of influence of features at different scales on the results of sentiment analysis and adjust the weight coefficients of each feature. The channel attention module consists of three operations, Squeeze, Excitation, and Scale, as shown in Fig. 3.

The Squeeze operation calculates the global spatial information by compressing and aggregating the data of the channels, in which the spatial information of the *k*-th channel is calculated according to the following equation:

$$z_\alpha^k = \text{Squeeze}\left(F_\alpha^k\right) = \frac{1}{T_\alpha \times d_\alpha} \sum_{i=1}^{T_\alpha} \sum_{j=1}^{d_\alpha} F_\alpha^k(i,j), \quad (4)$$

where $k = 1, 2, ..., n$. By executing the Squeeze operation on all channels, we obtain the global spatial information $\mathbf{Z}_\alpha = \{z_\alpha^1, z_\alpha^2, ..., z_\alpha^n\}$, $Z_\alpha \in \mathbb{R}^{1\times 1\times n}$.

Similar to the gate mechanism in LSTM, the Excitation operation uses two fully connected layers to learn the weight parameter s_α:

$$\begin{aligned} s_\alpha &= \text{Excitation}\left(Z_\alpha, W_\alpha\right) = \text{Relu}\left(g\left(Z_\alpha, W_\alpha\right)\right) \\ &= \text{Relu}\left(W_\alpha^2 \delta\left(W_\alpha^1 Z_\alpha\right)\right), \end{aligned} \quad (5)$$

where $s_\alpha \in \mathbb{R}^{1\times 1\times n}$, $W_\alpha^1 \in \mathbb{R}^{(n\times r)\times n}$, $W_\alpha^2 \in \mathbb{R}^{n\times(n\times r)}$ and r is the channel transformation ratio.

The Scale operation controls the output of the different channels with the learned weight parameter s_α:

$$\tilde{m}_\alpha^k = \text{Scale}\left(F_\alpha, s_\alpha\right) = s_\alpha^k F_\alpha^k. \quad (6)$$

By executing the Scale operation on all channels, we obtain $\tilde{M}_\alpha = \{\tilde{m}_\alpha^1, \tilde{m}_\alpha^2, \ldots, \tilde{m}_\alpha^n\}$. Then we use a 1×1 size convolution kernel to integrate the information of all channels to obtain the modality α high-level features X_α:

$$X_\alpha = \text{Conv}_{1\times 1}\left(\tilde{M}_\alpha\right). \quad (7)$$

B. Key Modality Selection

It shows that the useful information related to sentiment analysis is not evenly distributed among different modalities. An equal fusion strategy may bring in unnecessary information, which in turn impacts the accuracy of sentiment analysis. We address the above problem by selecting one modality as the key modality and adopting an unbalanced fusion strategy.

First, we perform sentiment analysis with unimodal data. In Fig. 2, we feed the features $F_m^n, m \in \{\alpha, \beta, \gamma\}$ extracted from (2) into the respective sentiment analysis classifier to obtain the sentiment analysis results for each modality:

$$\tilde{y}_\alpha = \text{classifier}_\alpha\left(F_\alpha^n\right), \quad (8)$$

$$\tilde{y}_\beta = \text{classifier}_\beta\left(F_\beta^n\right), \quad (9)$$

$$\tilde{y}_\gamma = \text{classifier}_\gamma\left(F_\gamma^n\right). \quad (10)$$

Subsequently, we rank the accuracy of unimodal sentiment analysis and select the modality with the highest accuracy as the key modality.

C. Multimodal Fusion Based on Key Modality

A good multimodal fusion strategy should be able to extract and integrate meaningful information from multiple modalities to improve the accuracy of sentiment analysis. If we adopt an equal fusion strategy for multimodal data fusion without considering the relative importance of different modalities, this may introduce non-essential information and thus reduce the accuracy of sentiment analysis. In this paper, we propose a fusion strategy based on key modality to balance the contributions of different modalities to sentiment analysis. The strategy consists mainly of key modality multi-head self-attention and cross-modal multi-head attention.

In [35], the authors proposed a method to calculate the attention value of the input data by dot product:

$$\begin{aligned} \text{head} &= \text{Attention}(Q, K, V) \\ &= \text{softmax}\left(\frac{QK^\tau}{\sqrt{d_h}}\right) V, \end{aligned} \quad (11)$$

where Q, K, and V mean the query, key, and value vectors.

Multi-head attention transforms query, key, and value by utilizing different projection matrices, and these metrics will be sent to the attention aggregation layer in a parallel manner:

$$\text{MutiHead}(Q, K, V) = \text{Concat}\left(\text{head}^1, \ldots, \text{head}^h\right). \quad (12)$$

1) Key Modality Multi-Head Self-Attention: Suppose the key modality is β, we transform the high-level features X_β by matrix projection and obtaining the Q_β^k, K_β^k, and V_β^k vectors,

respectively. As shown in the following:

$$Q_\beta^k = X_\beta W_{Q,\beta}^k, \tag{13}$$

$$K_\beta^k = X_\beta W_{K,\beta}^k, \tag{14}$$

$$V_\beta^k = X_\beta W_{V,\beta}^k. \tag{15}$$

where $W_{Q,\beta}^k, W_{K,\beta}^k$, and $W_{V,\beta}^k$ are projection matrices; k means for the k-th head.

Subsequently, the self-attentive value of the k-th head is calculated with (16):

$$\begin{aligned} \text{head}^k &= \text{Attention}\left(Q_\beta^k, K_\beta^k, V_\beta^k\right) \\ &= \text{softmax}\left(\frac{Q_\beta^k K_\beta^{k\tau}}{\sqrt{d_h}}\right) V_\beta^k, \end{aligned} \tag{16}$$

Based on the multi-head attention described above, the high-level feature X_β is mapped with h different projection matrices to obtain h different sets of query, key, and value vectors $\{Q_\beta^k, Q_\beta^k, Q_\beta^k\}$, $k = 1, 2, ..., h$. Then, we perform the self-attentive operations, according to (16), in parallel for each set of query, key, and value vectors to obtain h different heads. Finally, all these heads are concatenated together to obtain the enhanced feature $X_{\beta,\beta}$:

$$\begin{aligned} X_{\beta,\beta} &= \text{MutiHead}(Q_\beta, K_\beta, V_\beta) \\ &= \text{Concat}\left(\text{head}^1, \ldots, \text{head}^h\right). \end{aligned} \tag{17}$$

2) Cross-Modal Multi-Head Attention: Cross-modal multi-head attention is an extension of the traditional attention mechanism, which can capture the interaction information between different modality data. In this paper, we need to attach the meaningful information from key modality β to other modalities, such as modality α, thus fusing the features of key modality β and modality α. Therefore, the cross-modal multi-head attention in this paper is focused on the key modality and another modality. The following is a detailed description with modality α as the example.

First, we project the high-level feature X_β of the key modality β to obtain the Q_β^k vector and the high-level feature X_α of the modality α to obtain K_α^k and V_α^k vectors.

$$Q_\beta^k = X_\beta W_{Q,\beta}^k, \tag{18}$$

$$K_\alpha^k = X_\alpha W_{K,\alpha}^k, \tag{19}$$

$$V_\alpha^k = X_\alpha W_{V,\alpha}^k. \tag{20}$$

Secondly, we calculate the cross-modal attention value between modality β and α, which attaches the meaningful information from key modality β to the modality α:

$$\text{Cross\_head}_{\beta,\alpha}^k = \text{softmax}\left(\frac{Q_\beta^k K_\alpha^{k\tau}}{\sqrt{d_h}}\right) V_\alpha^k. \tag{21}$$

We also project the features X_β and X_α by utilizing h different projection matrices to obtain h different sets of query, key and value vectors $\{Q_\beta^k, Q_\alpha^k, Q_\alpha^k\}$, $k = 1, 2, ..., h$. Then cross-modal attention operations are performed in parallel for each set to get h different heads. Finally, all these heads are collocated together to obtain the fusion feature $X_{\beta,\alpha}$ between key modality β and modality α:

$$\begin{aligned} X_{\beta,\alpha} &= \text{MutiHead}\left(Q_\beta, K_\alpha, V_\alpha\right) \\ &= \text{Concat}\left(\text{Cross\_head}_{\beta,\alpha}^1, \ldots, \text{Cross\_head}_{\beta,\alpha}^h\right). \end{aligned} \tag{22}$$

Given key modality β and modality γ, we can use the same operation to obtain fusion feature $X_{\beta,\gamma}$.

Finally, we concatenate the fusion feature $X_{\beta,\alpha}$, $X_{\beta,\gamma}$ and the key modality reinforcement feature $X_{\beta,\beta}$ to obtain the multimodal fusion features X_f:

$$X_f = \text{Concat}\left(X_{\beta,\alpha}, X_{\beta,\gamma}, X_{\beta,\beta}\right). \tag{23}$$

D. Multi-Task Learning

Multi-task learning can improve model performance by mining the relationships between different tasks to jointly optimize the model parameters, which can preserves the independent features of each task data. In the process of experiments, we found that the distributions of the features of each modality in the potential high-dimensional feature space are very similar, which leads to the model not being able to fully utilize the difference information between the multimodal data, thus reducing the sentiment analysis precision. We set multiple different subtasks for M^{3}SA to solve the problem and obtain the loss function of M^{3}SA by weighted summation of the loss function of each subtask.

We feed the multimodal fusion features X_f extracted from (23) into the multimodal sentiment analysis classifier to obtain the multimodal sentiment analysis score $\tilde{y}_f$:

$$\tilde{y}_f = \text{classifier}_f(X_f). \tag{24}$$

As shown in (8), (9), (10), and (24), M^{3}SA has four sentiment analysis subtasks, i.e., sentiment analysis with modality α, β, γ, and multimodal fusion feature X_f, respectively, and each subtask predicts an independent sentiment analysis result. We calculate the loss values between the predicted sentiment results and the true sentiment results for each subtask and then weigh the sum of them to gain the loss values of the M^{3}SA model. The loss function of M^{3}SA is shown in (25):

$$L = \frac{1}{N}\sum_{1}^{N}\sum_{i}^{\{\alpha,\beta,\gamma,f\}} \partial_i L_i\left(\tilde{y}_i, y_i\right), \tag{25}$$

where $L_i(\tilde{y}_i, y_i)$ is the loss function for each subtask; N is the number of training samples; ∂_i is the weight used to balance the loss function for each subtask; y_i is the true sentiment scores. The weights of each subtask are gradually reduced with dynamic weight decay in the training process so that each subtask can be fully trained.

Comparative learning can extract discriminative features with incomplete scores by identifying the differences between the data. There are no unimodal sentiment scores in some cases. For example, we only have true sentiment score y_f, and don't have labels for each modality such as y_α, y_β, and y_γ. In this case, we can choose comparative learning to carry out the multitask training process.

The first step is to construct positive sample pairs and negative sample pairs. Let U_1 and U_2 represent the two multimodal samples, $U_1 = \{U_\alpha^1, U_\beta^1, U_\gamma^1\}$, $U_2 = \{U_\alpha^2, U_\beta^2, U_\gamma^2\}$. A positive sample pair is pair of two different modalities of data from the same sample. For example, there are three positive sample pairs with U_1, i.e., $\{U_\alpha^1, U_\beta^1\}$, $\{U_\alpha^1, U_\gamma^1\}$ and $\{U_\beta^1, U_\gamma^1\}$. We can also construct three positive sample pairs with U_2. A negative sample pair is pair of distinct modalities of data from two samples. Given two samples U_1 and U_2, there are six negative sample pairs, namely, $\{U_\alpha^1, U_\beta^2\}$, $\{U_\alpha^1, U_\gamma^2\}$, $\{U_\beta^1, U_\alpha^2\}$, $\{U_\beta^1, U_\gamma^2\}$, $\{U_\gamma^1, U_\alpha^2\}$ and $\{U_\gamma^1, U_\beta^2\}$.

The second step is to define the comparative loss function. Here, we show an example to illustrate how it is defined with sample pairs of modality α and modality β. Assuming the batch size is N, there are totally N positive sample pairs and $(N^2 - N)$ negative sample pairs. Given a positive or negative sample pair, we use a dot product to calculate the similarity between the different modalities. For a positive sample, following the idea of the InfoNCE loss function [36], we choose the similarity of the positive sample as the numerator and choose the sum of the similarity of all negative sample pairs and the similarity of this positive sample as the denominator. The comparative loss of one positive sample is the value after a logarithmic operation. The comparative loss of all positive samples is the sum of pairs:

$$L_{\alpha,\beta} = -\sum_{\{U_\alpha^i, U_\beta^i\}\in\mathcal{N}} \times \log\left(\frac{\exp(U_\alpha^{i\mathrm{T}} U_\beta^i/\tau)}{\exp(U_\alpha^{i\mathrm{T}} U_\beta^i/\tau) + \sum_{\{U_\alpha^j, U_\beta^k\}\in\mathcal{N}'} \exp\left(U_\alpha^{j\mathrm{T}} U_\beta^k/\tau\right)}\right), \tag{26}$$

where $\mathcal{N}$ and $\mathcal{N}'$ denote the set of positive and negative sample pairs in the batch, and τ is a temperature parameter used to control the smoothness of probability distribution.

We also can calculate the comparative loss for modality α and γ, modality β and γ in a similar way, i.e., $L_{\alpha,\gamma}$, $L_{\beta,\gamma}$.

Finally, we define the multi-task learning loss function. It is described as follows.

$$L = \frac{1}{N}\left(\partial_{\alpha,\beta} L_{\alpha,\beta} + \partial_{\alpha,\gamma} L_{\alpha,\gamma} + \partial_{\beta,\gamma} L_{\beta,\gamma} + \sum_1^N \partial_f L_f\right), \tag{27}$$

where $\partial_{\alpha,\beta}, \partial_{\alpha,\gamma}, \partial_{\beta,\gamma}$, and ∂_f are hyper-parameters used to balance the losses of different sub-tasks.

IV. Experiment and Analysis

A. Datasets

In this section, we test and analyze the model's performance with the CMU-MOSI [37] and CH-SIMS [38]. Both datasets provide three different modality data for sentiment analysis: text (T), audio (A), and vision (V).

CMU-MOSI: The CMU-MOSI dataset contains 2199 video clips. These clips were mainly taken by segmenting 93 videos of 89 speakers by sentences that describe the speaker's comments on a movie. The CMU-MOSI manually labeled each video clip with a sentiment score ranging from -3 to 3, indicating positive sentiment (sentiment score > 0) or negative sentiment (sentiment score < 0). The larger the absolute value of the sentiment score, the stronger the sentiment.

CH-SIMS: The CH-SIMS dataset contains 2281 video clips. These clips are mainly taken by segmenting 61 videos from different movies, TV series, and variety shows, which is a Chinese multimodal video sentiment analysis dataset. Compared with CMU-MOSI, CH-SIMS not only provides the sentiment scores of each video clip but also annotates the sentiment scores for text, audio, and vision modalities, respectively. The sentiment score ranges from -1 (strongly negative) to 1 (strongly positive).

1) Text feature: BERT [39] is a pre-trained deep learning model developed by Google that has already learned rich linguistic features on large-scale text datasets, including lexical semantics, sentence structure, and contextual relationships. It means that BERT has powerful language understanding ability. In our experiments, we utilize BERT to extract the text modality high-level features X_t of the CMU-MOSI and CH-SIMS datasets, and the dimension of X_t is 768.

2) Audio feature: We use different tools to extract low-level audio features U_a for each dataset. On the CMU-MOSI dataset, we use the COVAREP [40] to extract 74-dimensional low-level audio features with frame rates of 1000 Hz. The Features include pitch, energy, NAQ, MFCCs, peak, and energy slopes. On the CH-SIMS dataset, the LibROSA [41] is used to extract 33-dimensional low-level audio features such as log F0, MFCCs, and Constant-Q chromatograms, with a frequency of 22050 Hz.

3) Vision feature: We extract low-level vision features U_v with different tools. On the CMU-MOSI dataset, we utilize the Facet [42] to extract 47-dimensional low-level vision features, including facial action units, facial pose, head pose, and orientation with a 30 Hz sampling frequency. On the CH-SIMS dataset, we align faces with MTCNN [43] models and use the OpenFace2.0 [44] tool to extract facial action unit, head pose, head orientation, eye movements, and so on 709-dimensional low-level vision features, which are also sampled at a frequency of 30 Hz.

B. Baselines

- *Early Fusion LSTM (EF-LSTM)* [45]. LSTM concatenates the initial inputs of the three modalities and utilizes the LSTM network to capture temporal information in each sequence.
- *Later Fusion Deep Neural Network (LF-DNN)* [45]. LF-DNN first extracts three modality features with three separate DNNs, then concatenates them for sentiment prediction.
- *Tensor Fusion Network (TFN)* [6]. TFN captures bimodal and trimodal information through the outer product operation and uses the output of the tensor fusion layer for sentiment analysis.
- *The Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA)* [8]. MISA uses an encoder to decompose each modality into modality-invariant and modality-specific representations. Also, the

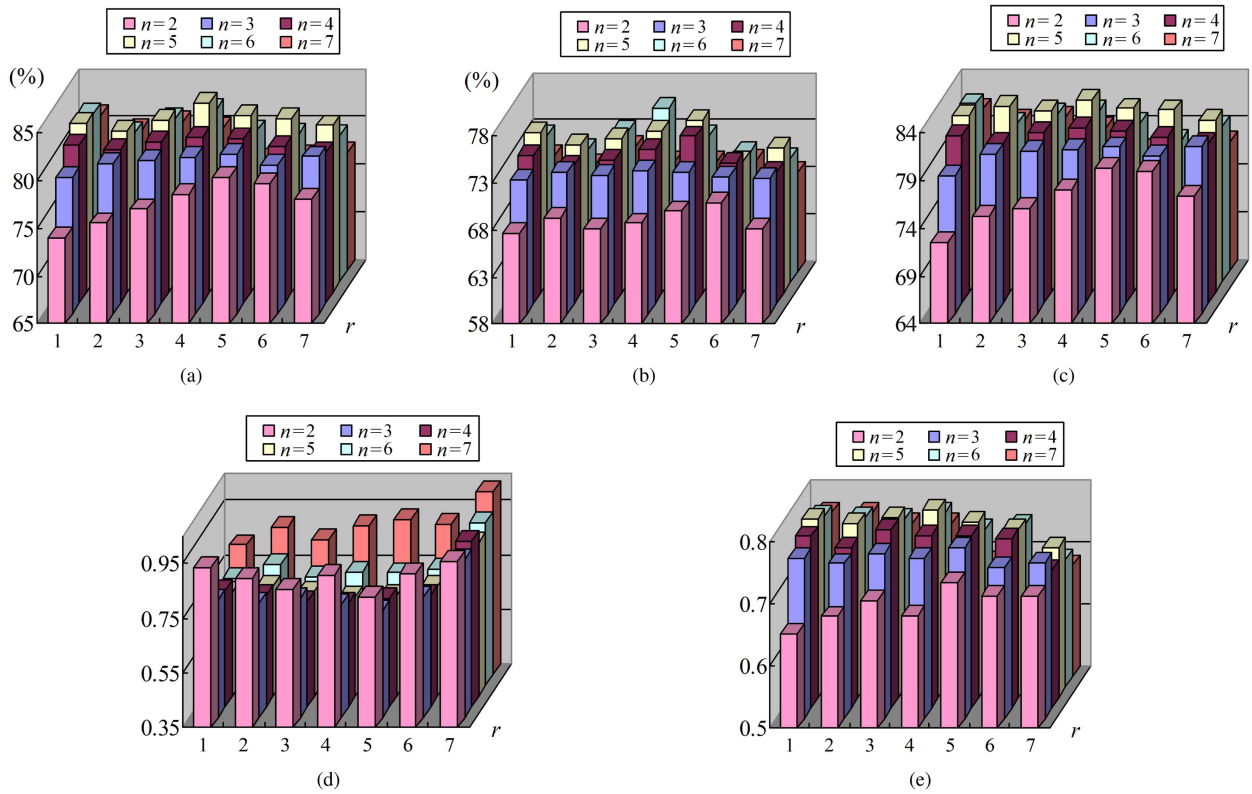

Fig. 4. Impact of different hidden layers n and channel transformation ratio r. (a) Acc 2. (b) Acc 3. (c) F1. (d) MAE. (e) Corr.

loss function of MISA includes similarity loss, orthogonal loss, reconstruction loss, and prediction loss.

- *The Low-rank Multimodal Fusion (LMF)* [8]. LMF proposes a low-rank fusion method, which decreases the computational complexity and improves the efficiency of model training by decomposing a high-rank tensor into a low-rank tensor.
- *Multi-task Later Fusion Deep Neural Network (MLF-DNN)* [38]. MLF-DNN is an extension of LF-DNN. MLF-DNN first extracts features of each modality with three separate DNNs and concatenates the extracted features for sentiment prediction through multi-task learning.
- *Multi-task Tensor Fusion Network (MTFN)* [38]. MTFN is an extension of TFN. MTFN first models bi-modality and tri-modality information through the outer product operation and uses the output of the tensor fusion layer for sentiment analysis through multi-task learning.

In this paper, we use five evaluation metrics to evaluate the performances of each sentiment analysis model, which include binary classification accuracy (Acc_2), triple classification accuracy (Acc_3), F1 score (F1), Mean Absolute Error (MAE) and Pearson Correlation (Corr). For all metrics, a higher value means better performance, except MAE. In the binary classification, we classify sentiment types into negative and positive, in which sentiment score < 0 is negative, and sentiment score $\geq$ is positive. In triple classification, we classify sentiment types into negative, neutral, and positive, in which sentiment score < 0 is negative, sentiment score $= 0$ is neutral, and sentiment score > 0 is positive.

TABLE I
HYPERPARAMETER SETTINGS OF THE DATASETS

Hyper-parameter	CMU-MOSI	CH-SIMS
Batch size	64	64
Learning rate	0.00005	0.000001
Dropout	0.4	0.3
Epoch	50	100
Head number h	5	5
Weight decay	5e-3	5e-3

C. Experimental Setup

We reproduce all baselines in the same experimental platform for a fair comparison. The experimental platform is built with Python 3.6.13 and Pytorch 1.2.0 + Cuda9.2 framework, and all models are trained on Intel Xeon E5-2620V4 CPU + NVIDIA Tesla P100 16 GB GPU. Considering the differences between the CMU-MOSI and CH-SIMS datasets, we set different hyper-parameters for each dataset, respectively, as shown in Table I.

D. Impact of Different Hidden Layers and Channel Transformation Ratio

To explore the impact of the different number of hidden layers n and channel transformation ratio r on the results of sentiment analysis, we have carried out a set of experiments on the CMU-MOSI dataset, in which n is changed from 2 to 7 and r is changed from 1 to 7.

1) Impact of the Number of Hidden Layers: In Fig. 4, it can be seen that with the number of hidden layers increasing,

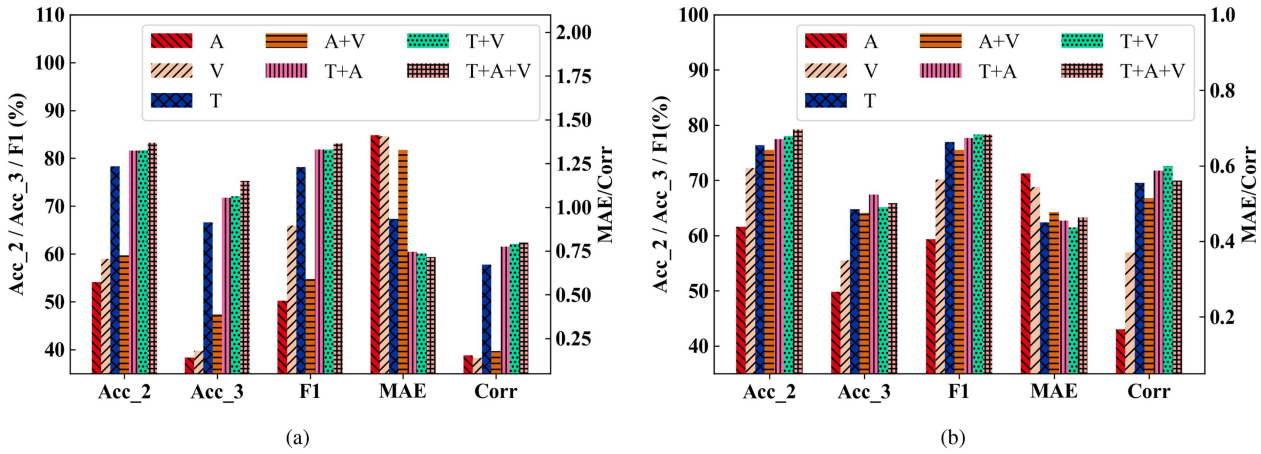

Fig. 5. Impact of modality number. (a) CMU-MOSI. (b) CH-SIMS.

the performance measured with these five metrics shows a tendency of first increasing and then decreasing. In case $n = 2$, Acc_2, Acc_3, F1 and Corr have low values and MAE has a high value, which indicates that the neural network has a weak learning ability with only two hidden layers. In case $n = 3$, the performance shown with these different metrics is improved. In case $n = 4$, the performance is further improved and tends to stabilize. The metrics, Acc_2, F1, Corr, and MAE, have reached the best value with $n = 5$, while Acc_3 reached the best with $n = 6$. After that, the model performance decreases instead with the number of layers increasing. Especially in case $n = 7$, the model performance suffers a severe decline. Compared with the best ones, the values of Acc_2, Acc_3, F1, and Corr have decreased by about 3.5%, 5.1%, 4.690, and 0.0375, respectively, and MAE has increased by 0.1099.

The experimental results show that adding a certain number of hidden layers appropriately can help the model to efficiently learn the features from training data, and finally improve the performance. However, the model complexity will increase with too many hidden layers, and it will lead to issues such as over-learning of the noise and details of the training data, which finally results in a drop in the sentiment analysis accuracy. Therefore, in the subsequent experiments, we choose the number of hidden layers n to be 5.

2) Impact of the Channel Transformation Ratio: The channel dimension of the channel attention mechanism can be changed by adjustment of channel transformation ratio r. In this way, the learning ability of the model can be influenced by adding or reducing the number of parameters in the model. In Fig. 4, it can be seen that the best r varies with the number of hidden layers. For example, in case $n = 2$, the value of F1 increases firstly and then decreases as r increases. This conclusion may be explained by noticing that the learning ability of the model is poor, with fewer hidden layers. With r increasing from 2 to 5, the metric F1 increases accordingly because additional parameters can increase the learning ability of the model. However, in case $n = 5$, the tendency of F1 change is not so obvious. The reason may be that the model has enough parameters to learn the sentimental information in the data, while more parameters won't help to improve the model's learning ability.

In Fig. 4, it can also be seen that the values of F1 and Acc_2 are optimal in case $r = 4$, and Acc_3 and MAE are optimal in case $r = 5$. The sub-optimal value is achieved at $r = 4$ for MAE, about 0.0038 behind the optimal one. The optimal value is attained at $n = 5$ and $r = 4$ for Corr, indicating that the model can adapt well to the data. According to the experimental results, we choose the number of hidden layers n to be 5 and the channel change ratio r to 4.

E. Impact of Modality Number

To evaluate the impact of modality number on sentiment analysis results, we perform comparative experiments on CMU-MOSI and CH-SIMS datasets for unimodal sentiment analysis, bimodal sentiment analysis, and trimodal sentiment analysis, respectively. The experimental results are shown in Fig. 5, in which A, V, and T indicate the sentiment analysis with audio, vision, and text modality; "+" means the fusion operation. For example, T+V means fusing the text and vision.

It can be seen that the performance obtained with different modalities varies greatly, with the lowest performance for the audio and the highest for the text modality according to the experimental result of unimodal. That is because text data usually contains more sentiment-related information, such as sentiment words, emojis, etc. Compared with audio and vision modality, the improvement in the performance of text modality is significant, about 20% for Acc_2 and 30% for Acc_3, and other evaluation metrics also improve considerably on CMU-MOSI and CH-SIMS datasets. The experimental results prove that the helpful information relevant to sentiment analysis is not equally distributed among the different modalities, in which the text modality makes the most significant contribution to sentiment analysis. Based on the experimental results of unimodal sentiment analysis on both datasets, we select the text modality with the highest accuracy of unimodal sentiment analysis as the key modality.

The bimodal sentiment analysis has a higher accuracy rate than that of unimodal sentiment analysis. The sentiment analysis accuracies are higher for T+A and T+V than that for T, A, and V. The experimental results on the CMU-MOSI dataset show

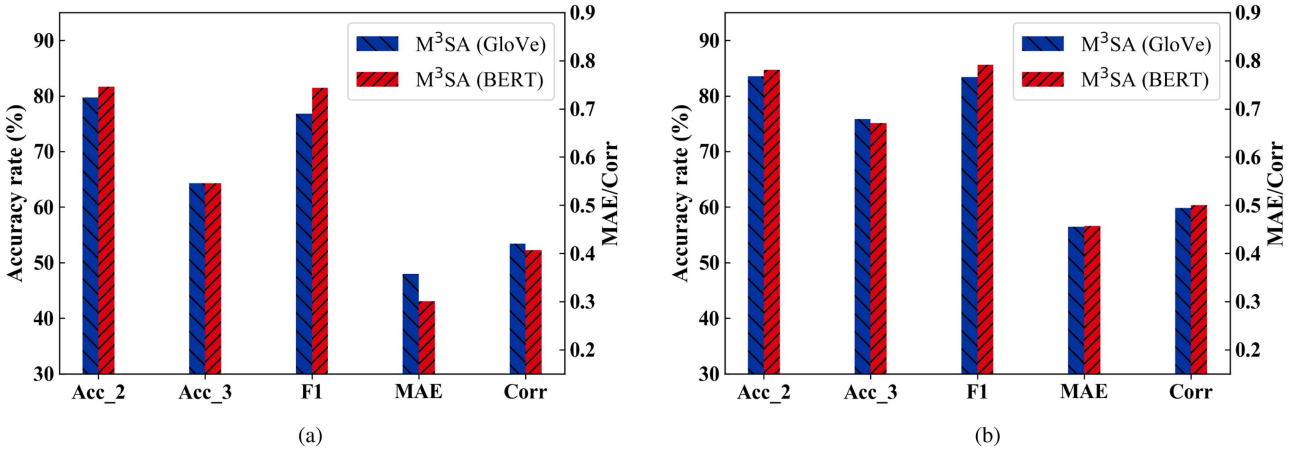

Fig. 6. Impact of transfer learning. (a) CMU-MOSI. (b) CH-SIMS.

that Acc_2 of T+A and T+V significantly improved by about 20% compared to modality A and V. However, there is only a slight performance improvement for modality A+V compared to modality A and V. From the above analysis. We can conclude that fused text modality is the most effective for sentiment analysis.

The trimodal sentiment analysis is not always better than the bimodal. Comparing the results of T+A+V, T+A, T+V, and A+V on the CMU-MOSI dataset, the highest accuracy has been obtained when all three modalities are involved, which is because when multiple modalities are fused, the model can utilize the complementary information from between different modalities to get more accurate sentiment analysis. However, The experimental results on the CH-SIMS dataset show that trimodals are not always better than bimodals. For example, modality T+A has higher Acc_3 than T+A+V, and modality T+V has higher Corr and F1 than T+A+V, which indicates that the performance of sentiment analysis does not increase singularly with the number of modalities.

There are two possible reasons for this issue. *a*) The feature extraction method is not appropriate. Data from different modalities need to extract features before they can be input into the model for training and prediction. If the feature extraction method is unsuitable, it may lead to insufficient feature extraction, which affects the performance of sentiment analysis. *b*) The fusion strategy is not appropriate. Multimodal sentiment analysis needs to fuse information from different modalities, but various fusion methods may lead to diverse sentiment analysis results, and if an inappropriate fusion strategy is chosen, it may also hurt the results of sentiment analysis. Therefore, finding a suitable feature extraction method and fusion strategy is key to achieving a more accurate multimodal sentiment analysis.

F. Impact of Transfer Learning

To verify whether transfer learning helps improve the accuracy of sentiment analysis, we have added a set of experiments on CMU-MOSI and CH-SIMS datasets, respectively. In the experiments, we deleted the BERT model and directly converted the raw text into word vectors with the GloVe model, and then extracted the feature with the LSTM network. The results of the experiments are shown as follows, in which M^{3}SA (GloVe) and M^{3}SA (BERT) represent the results with the GloVe model and BERT, respectively.

Fig. 6(a) shows the results on CMU-MOSI, and a similar conclusion can be found except that there is a little decrease with Acc_3 (about -2.21%). Fig. 6(b) shows the results of CH-SIMS. The performance of BERT is better than GloVe with all metrics, in which F1 is the most significant. It demonstrates that transfer learning helps improve the performance of sentiment analysis. Especially, BERT has been pre-trained on large-scale datasets and has learned rich representations of language. These representations can capture the deep semantic information of words and provide richer and more complex contextual features when compared with the traditional word embedding technique GloVe. Note that sentiment often depends heavily on the context, and this advantage is particularly important for the model to understand the text sentiment. BERT can transfer knowledge from large datasets to specific multimodal sentiment analysis tasks. Even in relatively small datasets, it is possible to significantly improve the performance of the model with this pre-training and fine-tuning approach.

G. Impact of Multi-Scale Feature Extraction

The impact on the model's performance after introducing multi-scale feature extraction on the CMU-MOSI is shown in Fig. 7(a), where MS indicates that the model uses multi-scale feature extraction. Comparing the experimental results of T+A+V with T+A+V (MS), we can see that Acc_2, Acc_3, F1, and MAE are improved after introducing multi-scale feature extraction, and only Corr has a slight decrease. The model's performance is significantly improved after introducing the multi-scale feature extraction method on CH-SIMS. As shown in Fig. 7(b), comparing the experimental results of T+A+V and T+A+V (MS), five evaluation metrics are improved after the introduction of multi-scale feature extraction. Comparing the experimental results of T+A and T+A+V (MS), Acc_3, MAE, and Corr are improved after introducing the multi-scale feature extraction. It indicates that multi-scale feature extraction can relieve the problem of decreasing the accuracy of sentiment

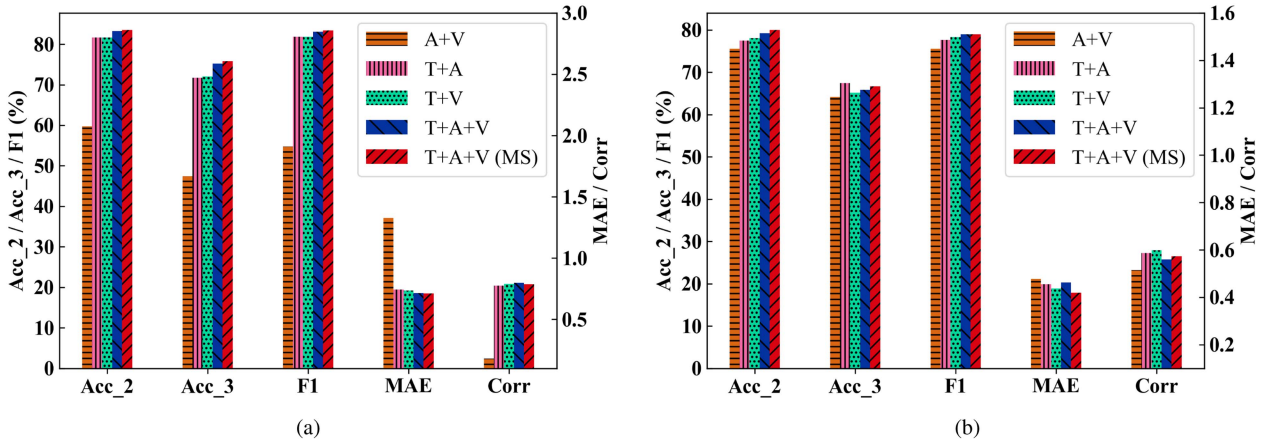

Fig. 7. Impact of multi-scale feature extraction. (a) CMU-MOSI. (b) CH-SIMS.

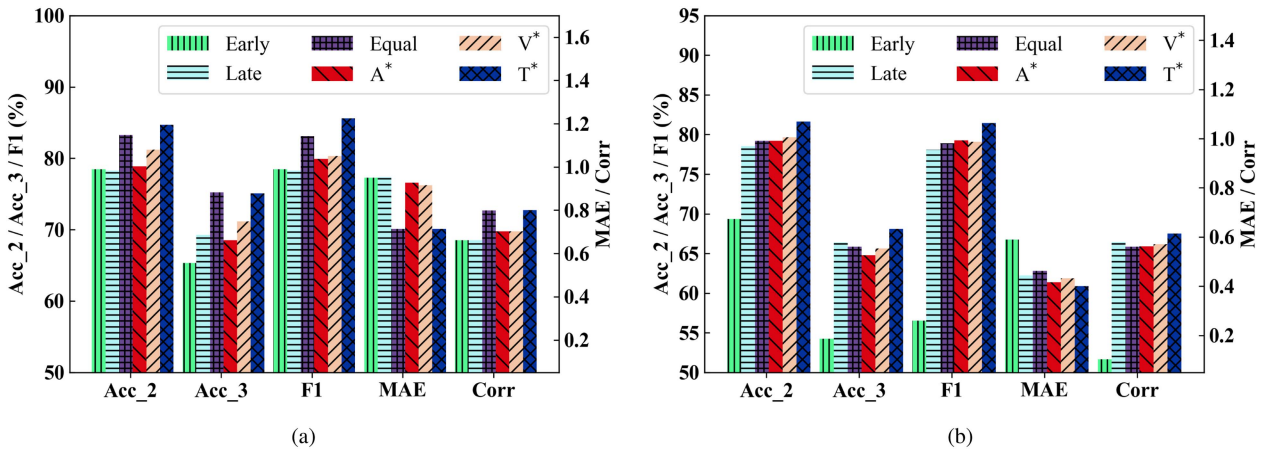

Fig. 8. Impact of different fusion strategies. (a) CMU-MOSI. (b) CH-SIMS.

analysis due to the increase in the number of modalities. Based on the experimental results of the two datasets, we can conclude that our proposed multi-scale feature extraction method can improve the accuracy of sentiment analysis.

H. Impact of Fusion Strategy

To explore the impact of different fusion strategies for sentiment analysis, we carried out four sets of experiments on CMU-MOSI and CH-SIMS, respectively. The experimental results are shown in Fig. 8, in which Equal represents the attention-based fusion strategy; Early and Late represent the early fusion strategy and late fusion strategy; T^*, A^*, and V^* represent selecting audio, vision, and text modality as the key modality. In the attention-based fusion strategy, multi-head self-attentive operations are performed on high-level features of text, vision, and audio modality to get $(X_{t,t}, X_{a,a}, X_{v,v})$ according to (18)–(22), and then collocate them to obtain the fused feature $X_f = \text{Concat}(X_{t,t}, X_{a,a}, X_{v,v})$.

First, from Fig. 8, we can see that the result of late fusion is similar to that of early fusion with four metrics (such as Acc_2, F1, MAE, and Corr) on CMU-MOSI, while the former performs significantly better than the latter with some metrics (especially Acc_2, Acc_3, and F1) on CH-SIMS. In general, the sentiment analysis results with the late fusion strategy are more accurate than those with early fusion. The reason is probably that late fusion requires extraction of the independent features of different modalities before decision fusion, which helps to retain the unique information of each modality, while early fusion directly combines the features of different modalities at the extraction stage and information of some modalities might be lost during the fusion process. The difference in fusion stage results in better performance when late fusion is compared with early fusion.

Then, we analyze the effect of the attention mechanism. On CMU-MOSI, the performance of the attention-based fusion strategy has the best results with all metrics compared with the early fusion and late fusion. On CH-SIMS, the performance of the attention-based fusion strategy was similar to that of the late fusion. The fusion based on attention usually automatically learns the importance of different modalities, which may explain why it outperforms the late fusion with some evaluation metrics. However, the attention mechanism makes it difficult to capture the long-time dependencies among the training dataset because the attention weights will decay with increasing distance, and thus the model is hard to learn the relationship between sequences of elements that are far away from each other.

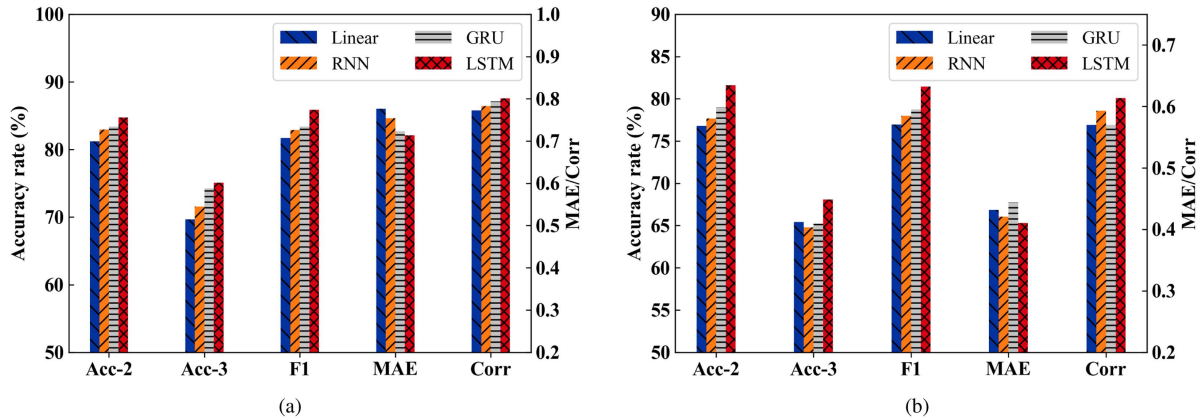

Fig. 9. Impact of different feature extraction methods. (a) CMU-MOSI. (b) CH-SIMS.

Finally, we analyze the impact of different key modalities on the final result of sentiment analysis. In Fig. 8, we can see that the fusion strategy with T as the key modality is better than the Equal strategy in all metrics except for Corr on the CMU-MOSI dataset. On CH-SIMS, the attention-based fusion strategy is worse than those with the strategies of A^*, V^*, and T^* in all metrics except for Acc_3. It demonstrates that the fusion strategy based on key modality is effective and can improve the performance of sentiment analysis. Moreover, comparing the experimental results of Equal strategy with A^* and V^* on CMU-MOSI, we find that the choice of key modality is very important. By comparing the experimental results of A^*, V^*, and T^* on two datasets, we find that the performance of sentiment analysis with A as the key modality is the lowest, and the one with T as the key modality is the highest. Therefore, it is reasonable to select the modality with the highest unimodal sentiment analysis accuracy as the key modality. The experimental results on both datasets demonstrate the effectiveness of our proposed fusion strategy based on key modality.

I. Impact of Different Feature Extraction Methods

In the paper, we use the pre-trained BERT from Google to extract the text feature. For audio and vision modalities feature extract, we tested four different neural network layers, namely, linear layer, LSTM, RNN, and GRN, to explore the impact of different feature extraction methods on the sentiment analysis results. The experimental results are shown in Fig. 9.

As shown in the figure, the feature extraction with LSTM and GRU performs better than the linear layer and the RNN with the metrics Acc_2 and Acc_3. It shows that a complex recurrent network structure is more effective in capturing sentiment-relevant features. The metric F1, as a combination of precision and recall, is usually a better indicator of the overall performance of a model. LSTM achieves the best performance in both datasets with metric F1. In addition, the MAE with LSTM is also lower than Linear, RNN, and GRU. It can be concluded that feature extraction using recurrent neural networks (RNN) and its variants (LSTM and GRU) is usually more effective than Linear. In particular, LSTM can obtain more accurate results for feature extraction.

TABLE II
RESULT OF MULTIMODAL SENTIMENT ANALYSIS ON CH-SIMS DATASET

Model	CH-SIMS				
	Acc_2	Acc_3	F1	MAE	Corr
EF-LSTM	69.37	54.27	56.55	0.5901	0.1033
LF-DNN	78.55	66.41	78.11	0.4440	0.5794
TFN	77.24	65.97	76.80	0.4352	0.5814
LMF	78.55	66.30	78.08	0.4401	0.5881
MISA	76.48	61.05	74.61	0.4781	0.5224
MLF-DNN	78.99	67.27	79.19	0.4289	0.5849
MTFN	79.43	**68.27**	79.92	0.4057	**0.6298**
M^{3}SA	**81.62**	68.08	**81.46**	**0.4010**	0.6140

TABLE III
RESULT OF MULTIMODAL SENTIMENT ANALYSIS ON CMU-MOSI DATASET

Model	CMU-MOSI				
	Acc_2	Acc_3	F1	MAE	Corr
EF-LSTM	78.50	65.31	78.46	0.9503	0.6605
LF-DNN	78.28	69.24	78.28	0.9529	0.6607
TFN	79.59	70.21	79.25	0.8925	0.6856
LMF	80.03	70.26	81.18	0.9081	0.6879
MISA	82.51	73.47	82.39	0.7460	0.7905
MLF-DNN	81.10	70.93	83.81	0.7719	0.7822
MTFN	82.97	72.62	83.09	0.7573	0.7986
M^{3}SA	**84.72**	**75.11**	**85.61**	**0.7133**	**0.8010**

J. Multimodal Sentiment Analysis

We perform experiments comparing the performance of the M^{3}SA model with other popular multimodal sentiment analysis models on the CMU-MOSI and CH-SIMS datasets. Since unimodal sentiment scores are not provided on CMU-MOSI, we adopt comparative learning to perform multi-task learning on this dataset. The sentiment analysis results are shown in Table III, in which TLF-DNN, MTFN, and M^{3}SA have adopted comparative learning. The loss function is shown in (27). The experimental results with baselines on two datasets are listed in Tables II and III, in which the bolded data means the best results and the added underlined data means the second best.

From Table II, we can find that M^{3}SA significantly outperforms other baselines in all metrics except the MTFN model from the comparative experimental results. Compared with the MLF-DNN modal, Acc_2, Acc_3, F1, and Corr improve by

2.63%, 2.54%, 2.27%, and 0.0291, respectively, and MAE decreases by 0.0279. This result indicates that the performance of the M^{3}SA model reaches the top level. Compared with the MTFN model, M^{3}SA have progressed in Acc_2, F1, and MAE, while Acc_3 and Corr do not do as well as the MTFN model. However, the performance of M^{3}SA is still better. First, M^{3}SA outperforms all baseline models in Acc_2, F1, and MAE and achieved the second-best rankings in Acc_2 and Corr. Second, M^{3}SA only decreases by 0.19% and 0.0102 in Acc_3 and Corr, respectively, which is not unacceptable compared to the improvement in model performance in Acc_2, F1, and MAE. The above data and analysis can prove that the performance of M^{3}SA is better on the CH-SIMS dataset.

From Table III, compared with the traditional models, i.e., TFN and MTFN, the performance of the model is improved after introducing multi-task learning on CMU-MOSI. All the evaluation metrics are better than those without multi-task learning. In particular, the binary classification accuracy (Acc_2) is improved by 3.38%, and the triple classification accuracy (Acc_3), improves by 2.41%, which is a larger improvement. It shows the efficiency of multi-task learning in sentiment analysis. The proposed M^{3}SA achieves the best performance. Compared with MTFN, the Acc_2, and Corr are increased by 1.75% and 0.24, and MAE is reduced by 0.044. Compared with MISA, Acc_3 is improved by 1.64%. While compared with MLF-DNN, F1 is increased by 1.8. It shows that M^{3}SA outperforms all baseline models. The experimental results demonstrate that our proposed multi-scale feature extraction and multi-task learning strategy helps understand different modality data and improving sentiment analysis accuracy.

V. Discussion

In this section, we have discussed the adaptability of the sentiment model, the fusion of potential modalities, and the fusion strategy for real-time or streaming data.

A. Adaptability Model for Multimodal Sentiment Analysis

It is a challenging issue to develop a sentiment analysis model that is general and robust across different domains and languages. The final accuracy of one model trained on one dataset may fall outside expectation when used in another dataset due to some observations.

1) Complicated Dataset Background: For example, the humor elements on UR-FUNNY [46] are highly culturally relevant and contain a much more comprehensive range of topics, while the data on CMU-MOSI focuses only on movie reviews and is just concerned with individuals' opinions on the movie plots, the actor's performance, and the director's work, etc.

2) Directness of Emotional Expression: The CMU-MOSI trends to describe direct sentiments, such as happiness, sadness, anger, and so on, and this information can be easily reflected in speech, voice, and facial expressions. In contrast, the UR-FUNNY involves subtler / complex affective and cognitive processes, including puns, sarcasm, hyperbole, or metaphors.

3) Differences in Feature Engineering: Feature extraction methods depend heavily on the modality in the dataset, and thus we cannot directly adopt the method from one dataset to another. The CMU-MOSI and CH-SIMS provide the raw textual modality data, and thus we can directly use BERT to extract the appropriate features. However, the UR-FUNNY provides only the word vectors extracted from the GloVe model but not the raw textual data.

B. Fusion of Additional Modalities

In addition to text, vision, and audio, physiological signals are also frequently used in sentiment analysis. The physiological signals that are often used for sentiment analysis include electromyogram, electroencephalogram, electrocardiogram, electrodermal activity, blood volume pulse, respiratory volume, skin temperature, heart rate, etc. [28] It might be helpful to improve the accuracy of sentiment analysis by utilizing more modalities in the model.

There are still a lot of constraints that limit the applications of physiological signals in sentiment analysis, such as privacy and ethical issues, complexity and cost of data collection, challenges of data processing and analysis, and difficulty of data annotation. If we consider the cost, complexity, and universality of models simultaneously, the text, vision and audio, can already satisfy the current sentiment analysis requirement. However, it is still interesting to investigate the impact of additional modalities on sentiment analysis.

C. Multimodal Fusion in Real-Time or Streaming Data

Our manuscript focuses on designing an accurate multimodal sentiment analysis model with stable modality data. We aim to mine the correlation and independence features between modalities and propose a multi-scale feature extraction method based on the channel attention mechanism. It is an important and perspective issue on how to perform sentiment analysis on real-time or streaming data. This idea shall face entirely new challenges, such as variable input modalities and dynamic variations in the time series. There are some feasible ideas for this issue.

1) Adaptive Learning: The data features and distributions may change with time in real-time situations. Online learning or incremental learning can be used to solve this problem.

2) Real-time Data Processing: The key is to quickly react and adapt to variable data in real time and adjust the fusion strategy. Feasible strategies to solve this problem include data window technology and rapid feature extraction.

3) Multimodal Dynamic Fusion: It means exploring fusion methods for multiple data sources and different types of inputs and real-time fusion strategies to provide model adaptability in the dynamic environment.

VI. Conclusion

In this paper, we present a novel multimodal sentiment analysis model called Multimodal Sentiment Analysis based on Multi-scale feature extraction and Multi-task learning (M^{3}SA). The model extracts multi-scale features with a channel attention mechanism to model the output of different hidden layers, which

allows the model to extract more helpful features. We have proposed a novel multimodal fusion strategy based on the key modality. During the fusion process, the key modality enhances the other modality's information and enriches the fused helpful information. Furthermore, multi-task learning is introduced to ensure that the model learns different modality's correlation features while preserving independence features. Experimental results on two datasets demonstrate that M^{3}SA outperforms baseline models.

In current multimodal sentiment analysis research, it is a common assumption that data from different modalities, such as text, audio, and visual, are complete. However, this assumption does not hold in real-world applications. Data from various sources may be lost or corrupted for various reasons. For instance, audio data may be degraded due to noise interference or interruptions during transmission. These issues significantly limit the practicality of multimodal sentiment analysis models. We are currently engaged in researching and addressing this challenge. Our next objective is to develop a model capable of handling incomplete modality data more effectively, and enhance its performance in real-world applications. It involves exploring methods for efficiently integrating data from various modalities and adjusting the model to maintain its performance when specific modality data is missing.

REFERENCES

[1] C. Lin and M. S. Obaidat, "Behavioral biometrics based on human-computer interaction devices," in *Biometric-Based Physical and Cybersecurity Systems*, The Netherlands: Springer, 2019, pp. 189–209.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, 2002, pp. 79–86.

[3] S. Tan, X. Cheng, Y. Wang, and H. Xu, "Adapting naive bayes to domain adaptation for sentiment analysis," in *Proc. 31th Eur. Conf. IR Res. Adv. Inf. Retr.*, 2009, vol. 5478, pp. 337–349.

[4] N. Zainuddin and A. Selamat, "Sentiment analysis using support vector machine," in *Proc. Int. Conf. Comput. Commun. Control Technol.*, 2009, pp. 333–337.

[5] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, 2017.

[6] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[7] Y. H. H. Tsai, S. Bai, P. P. Liang, J. P. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2019, vol. 2019 pp. 6558–6569.

[8] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[9] P. Wang, R. Yu, N. Gao, C. Lin, and Y. Liu, "Task-driven data offloading for fog-enabled urban IoT services," *IEEE Internet Things J.*, vol, 8, no. 9, pp. 7562–7574, May 2021.

[10] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2014, pp. 474–490.

[11] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with multiscale similarity learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1648–1659, Oct. 2013.

[12] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[13] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 517–532.

[14] B. Sun, B. Song, J. Lv, P. Chen, C. Ma, and Z. Gao, "A multi-scale feature extraction network based on channel-spatial attention for electromyographic signal classification," *IEEE Trans. Cogn. Dev. Syst.*, vol. 15, no. 2, pp. 591–601, Jun. 2023.

[15] Y. Lei, S. Yang, X. Wang, and L. Xie, "MsemoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE-ACM Trans. Audio Speech Lang.*, vol. 30, pp. 853–864, 2022.

[16] X. Cao, H. Liangwen, H. Wang, and L. Liu, "Microblog-oriented multi-scale CNN multi-label sentiment classification model," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, 2020, pp. 626–631.

[17] Q. Guo, X. Qiu, P. Liu, X. Xue, and Z. Zhang, "Multi-scale self-attention for text classification," in *Proc. AAAI Conf. Artif. Intell*, 2020, pp. 7847–7854, doi: 10.1609/aaai.v34i05.6290.

[18] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, 2018.

[19] N. Jin, J. Wu, X. Ma, K. Yan, and Y. Mo, "Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification," *IEEE Access*, vol. 8, pp. 77060–77072, 2020.

[20] B. Yang, L. Wu, J. Zhu, B. Shao, X. Lin, and T. Y. Liu, "Multimodal sentiment analysis with two-phase multi-task learning," *IEEE-ACM Trans. Audio Speech Lang.*, vol. 30, pp. 2015–2024, 2022.

[21] L. Yang, J. Na, and J. Yu, "Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis," *Inf. Process. Manag.*, vol. 59, no. 5, 2022, Art. no. 103038.

[22] N. Majumder, S. Poria, H. Pang, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and sarcasm classification with multitask learning," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 38–43, May/Jun. 2019.

[23] J. LI, H. Zhao, G. Shi, D. Zhang, C. Zhang, and H. Ma, "Multimodal sentiment analysis method based on multi-task learning," in *Proc. 6th Int. Confer. Signal Process. Mach. Learn.*, 2023, pp. 308–314.

[24] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. B. V. Subramanyam, "Benchmarking multimodal sentiment analysis," in *Proc. Int. Conf. Computat. Linguistics Intell. Text Process.*, 2017, pp. 166–179.

[25] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 371–378.

[26] L. H. Chu, Z. Chen, X. Yu, M. Xiao, and P. Chang, "Self-supervised cross-modal pretraining for speech emotion recognition and sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics*, 2022, pp. 5105–5114.

[27] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L. P. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interact.*, 2021, pp. 6–15.

[28] L. Shu et al., "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, 2018, Art. no. 2074.

[29] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[30] Q. Zhu, G. Lu, and J. Yan, "Valence-arousal model based emotion recognition using EEG, peripheral physiological signals and facial expression," in *Proc. 4th Int. Conf. Mach. Learn. Soft Comput.*, 2020, pp. 81-85.

[31] B. Zhong et al., "Emotion recognition with facial expressions and physiological signals," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2017, pp. 1–8.

[32] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *J. Netw. Comput. Appl.*, vol. 30, no. 4, pp. 1334–1345, 2007.

[33] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Proc. Affect Emotion Hum.- Comput. Interaction*, 2008, pp. 92–103.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[35] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[36] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[37] A. Zadeh, R. Zellers, E. Pincus, and L. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Jun. 2016.

[38] W. Yu et al., "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Ann. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.

[39] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2019, pp. 4171–4186.
[40] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP–A collaborative voice analysis repository for speech technologies," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 960–964.
[41] B. McFee et al., "Librosa: Audio and music signal analysis in python," in *Proc. Python Sci. Conf.*, 2015, vol. 8, pp. 18–25.
[42] P. Ekman and L. E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*. Oxford, U.K.: Oxford Univ. Press, 2005.
[43] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[44] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.
[45] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.
[46] M. K. Hasan et al., "UR-FUNNY: A multimodal language dataset for understanding humor," in *Proc. Conf. Empir. Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 2046–2056.

Changkai Lin received the B.S. degree in software engineering in 2021 from Fuzhou University, Fuzhou, China, where he is currently working toward the M.S. degree. His researchfocuses on multimodal fusion.

Hongju Cheng (Senior Member, IEEE) received the B.E. and M.E. degrees from the Wuhan University of Hydraulic and Electric Engineering, Wuhan, China, in 1997 and 2000, respectively, and the Ph.D. degree in computer science from Wuhan University, Wuhan, in 2007. He is currently a Professor with the College of Computer and Data Science, Fuzhou University, Fuzhou, China. His research interests include the Internet of Things, mobile ad hoc networks, and wireless sensor networks.

Qiang Rao received the B.S. degree in land resources management and the M.E. degree in engineering of surveying and mapping from Wuhan University, Wuhan, China, in 2013 and 2015, respectively. He is currently a Department Manager with the Fujian Jingwei Digital Technology Corporation Ltd. He is currently engaged in the planning, development, and management of smart cities.

Yang Yang (Member, IEEE) received the B.Sc. and Ph.D. degrees from Xidian University, Xi'an, China, in 2006 and 2011, respectively. She is also a Research Fellow (postdoctor) with the School of Information System, Singapore Management University, Singapore. She is currently a Full Professor with the College of Computer and Data Science, Fuzhou University, Fuzhou, China. Her research interests include information security and privacy protection.