IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 April 2023, accepted 12 May 2023, date of publication 16 May 2023, date of current version 23 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3276932

RESEARCH ARTICLE

AdaMoW: Multimodal Sentiment Analysis Based on Adaptive Modality-Specific Weight Fusion Network

JUNLING ZHANG, XUEMEI WU, AND CHANGQIN HUANG, (Member, IEEE)
Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, Zhejiang 321004, China

Corresponding author: Changqin Huang (cqhuang@zju.edu.cn)

This work was supported by the Key Research and Development Program of Zhejiang Province under Grant 2022C03106 and Grant 2021C03141.

ABSTRACT Multimodal sentiment analysis (MSA) is a crucial task in the field of natural language processing (NLP), with a wide range of applications. This paper proposes an adaptive modality-specific weight fusion network (AdaMoW) to address issues in the process of multimodal data fusion. Specifically, we use different weight calculation methods at various stages of the model. In the model training stage, diverse weights are assigned to different modalities by calculating the correlation between the single-modal sentiment prediction value and the real multimodal sentiment labels, and a weight-mapping network is designed to learn this ''data-weight'' mapping relationship. In the testing and verification phase of the model, the trained weight-mapping network is used to obtain the weights of different modalities. In addition, in order to optimize the multimodal fusion data, we designed a generator, which reversely generates the unimodal feature vector through the multimodal fusion vector, and compares it with the original unimodal feature extraction obtained after unimodal feature extraction. The modal feature vectors are compared and optimized, so that the fusion results can maintain the uniqueness of the modality while obtaining the multimodal data interaction information. AdaMoW is verified on two benchmark MSA datasets CMU-MOSI and CMU-MOSEI. The experimental results show that the effectiveness of AdaMoW surpasses the previous baseline model and achieves state-of-the-art results.

INDEX TERMS Deep neural networks, adaptive weight-based feature fusion, modality-specific feature optimization, multimodal sentiment analysis.

I. INTRODUCTION

Sentiment plays a vital role in interpersonal communication, and human cognition and behavior are almost all driven by sentiment, which helps us communicate, learn and understand people's intentions. Early tasks on sentiment analysis mainly focused on using text or images to mine emotional expression tendancies. With the development of the Internet and communication technology, single-modal data such as text or images cannot meet the needs of sentiment analysis. Therefore, sentiment analysis tasks based on multimodal data have become popular [1]. Multimodal sentiment analysis (MSA) primarily uses multiple modal data (e.g. text, vision, and audio data) to design an effective fusion strategy to jointly analyze sentiment tendencies.

The previous approaches in MSA include tensor-based fusion, which directly adds or connects multimodal feature tensors [2], [3]. Despite the gains, the original feature tensor data are very large, which makes tensor fusion methods challenging, and even simple operations require a lot of storage and many computational resources. Moreover, in practical applications, multimodal sentiment analysis needs to process a larger amount of data than single-modal sentiment analysis, so the calculation speed of multimodal sentiment analysis algorithms is required to be higher. Several deep-learning approaches have been developed to probe into the novel, well-designed and efficient fusion methods [4], [5], [6], [7], [8].

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Asif.

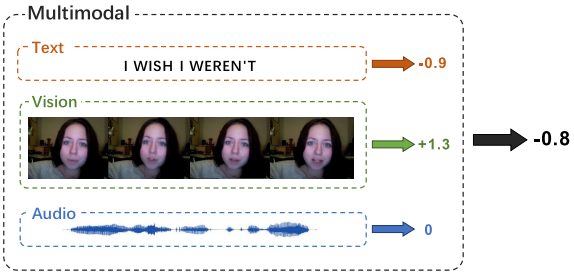


FIGURE 1. An instance of a gap in sentiment between different modalities on the CMU-MOSI dataset. Colored numbers indicate unimodal emotional polarity, and the black number indicates overall multimodal data emotional orientation.

However, none of these methods consider the gap between unimodal and multimodal affective results, i.e., the presence of noisy modalities. Figure 1 shows an instance in the CMU-MOSI dataset. In this example, ''I wish I weren't'' in the text modality indicates a negative polarity of sentiment. In contrast, the human face in the visual modality has the characteristics of ''slightly raised corners of the mouth'', which expresses a positive sentiment. The audio modality does not contain much sentiment information, and has median speech speed and volume, no emotional tendency and belongs to a neutral emotional state. Nevertheless, the overall sentiment is generally still negative, given the combination of multiple modalities. In fact, in real-life data, this kind of problem is common. Different modalities reflect various emotional tendencies, and these multimodal data may have independent or consistent information [9]. Reference [10] also considered this problem and proposed a multimodal routing to dynamically adjust fusion weights by exploring the correlation between modalities.

To fully mine the sentiment-related information contained in the modalities, researchers began to integrate the attention mechanism into multimodal sentiment research [11]. For example, [12] uses the attention mechanism to focus on the critical time-steps of emotional changes in audio modals. Reference [13] proposes a multimodal transformer, which designs a cross-modal attention mechanism to capture the long-range dependencies of elements between different modalities of unaligned data. Although attention mechanisms have been extensively studied for multimodal fusion, they did not consider intermodal dynamics due to modal differences. In general, modality information has complementary characteristics. That is, one kind of modality information can provide supplementary information for another kind of modality. Therefore, different modalities also make various contributions to the final sentiment decision-making.

In addition, as shown in Figure 1, the sentiment polarity of the visual modality is opposed to the final multimodal sentiment polarity, and we can also call this video modality the noise modality. For this problem, [14] utilizes canonical correlational to analyze the noise modality and generate a proxy eigenvector in place of this noise modality. Although this method can resolve the problem to a certain extent, directly removing the noise modality also loses the unique information of the modality itself. Therefore, it is necessary to maintain the modality's unique information and solve the noise modality problem. A good solution is to assign various weights to different modalities. Giving a low weight to a noisy modality allows it to have a lower impact on the final multimodal sentiment analysis decision, and retains the modality-unique information that the modality should have.

Based on the above problems, we designed a novel multimodal fusion framework for MSA, called adaptive modality-specific weight fusion network (AdaMoW). Specifically, we proposed a weight-based feature fusion method by calculating the correlation between unimodal sentiment prediction values and the true multimodal sentiment label and assigning various weights to different modalities. Since the true label only exists during model training, we also designed a weight-mapping network to learn the unimodal weight calculation method. The network can obtain the model's corresponding modal weight according to its unimodal feature vector in testing and verification. We also designed a modality feature generator to generate a unimodal feature vector through the multimodal fusion result, and compared this with the unimodal features before fusion, so that the multimodal fusion result retains more unique information about each modality. The contributions of this paper can be summarized as follows:

1) We introduce a novel adaptive modality-specific weight fusion network (AdaMoW) for multimodal sentiment analysis, which can effectively analyze the dynamic relationship between modalities through data-driven techniques and is robust to noise.
2) We propose a weight-based feature fusion method and a weight-mapping network to effectively utilize the specific sentiment information of each modality and reduce the impact of modality noise on multimodal sentiment prediction.
3) We tailor a unimodal feature generator to optimize the multimodal fusion result to retain more modality-specific information for multimodal sentiment analysis.

The remainder of the paper is organized as follows: Section II reviews the related work on MAS. Section III presents the proposed AdaMoW in detail. Section IV reports the comparative experiment results. Finally, Section V concludes this paper.

II. RELATED WORK

The main task of our model is sentiment analysis, and both unimodal feature extraction and multimodal feature fusion are included in our model. Therefore, this section discusses the related works of unimodal and multimodal sentiment analysis.

A. UNIMODAL SENTIMENT ANALYSIS

Unimodal sentiment analysis mainly includes text-based, vision-based, and audio-based sentiment analysis.

1) TEXT-BASED SENTIMENT ANALYSIS

Is the most common unimodal sentiment recognition method, and its analysis methods mainly include dictionary-based models [15] and machine learning models [16]. Reference [17] first infers the sentiment orientation of the entire sentence through the sentiment polarity of adjectives in the text. However, this method has limitations because, in real scenarios, there will be many double-negatives expressing affirmation or irony with positive emotional words. Later, with the development of natural language processing, [18] used the machine learning method of Naive Bayes, maximum entropy classification, and support vector machines for the first time to explore text-based sentiment analysis. Inspired by [18], various text analysis methods based on machine learning came into being. Reference [19] utilizes different levels of GRU for text embedding and contextual semantic capture.

2) AUDIO-BASED SENTIMENT ANALYSIS

Is mainly based on the acoustic features of speech intonation for emotional modeling. Reference [20] proposes a learning framework to standardize the speaker's characteristics in the feature representation, which can be used for sentiment analysis in different scenarios that are related or irrelevant to the speaker. Reference [21] also combined convolutional neural network (CNN) and long-short-term memory (LSTM) to study the speech context in audio information. Reference [22] proposes using bi-directional long-short-term memory with directional self-attention (BLSTM-DSA) to automatically mark the weight of speech frames, and then select key frames that have a more significant impact on the emotional state in the time series.

3) VISION-BASED SENTIMENT ANALYSIS

Is mainly used to model sentiment expressed by facial expressions or body gestures with visual multimedia, which is an important field of affective computing. Initially, [23] used pixel-level features to model visual emotional information. Later, with the development of neural networks, many neural network frameworks were used for visual sentiment analysis, such as [24], which proposes a suitable CNN for vision sentiment analysis. Reference [25] designed a coupled convolutional network, which obtains overall and local information by coupling the sentiment map. In addition to the traditional neural network used in sentiment analysis, there has been much related sentiment analysis research on the attention mechanism. For example, [26] used the attention mechanism to study the influence of local areas in the video on visual sentiment analysis, which further improves the accuracy of sentiment analysis. Attention is a dynamic feature extraction mechanism that mainly models contextual information. It can be flexibly embedded in deep learning frameworks and achieves an excellent performance in many fields related to time series data. The emotion itself changes with time, and many attention-based visual and audio emotion analysis models have emerged [22], [27], [28].

Intuitively, human emotion is highly subjective and extremely complex, and unimodal sentiment analysis can only analyze sentiment one-sidedly through one modal feature, and cannot comprehensively analyze from multiple perspectives. Therefore, this paper starts from the information characteristics and complementarity of different modalities, selects various feature extraction methods according to the uniqueness of the modalities, and provides high-quality unimodal feature representation for later multimodal feature fusion.

B. MULTIMODAL SENTIMENT ANALYSIS

The task of multimodal sentiment analysis (MSA) is to analyze the emotional tendency of the information by integrating modal information such as text, audio and vision. How to effectively fuse multimodal information to obtain multimodal joint representation and learn intra-modal and cross-modal dynamics is the key challenge of MSA [29].

The multimodal information fusion strategy can be divided into three types: early, late, and hybrid fusion. In the early fusion, the modality features obtained by connecting various modality features are used as the fusion features to input the model [30]. Reference [2] improved the original early fusion method and proposed the tensor fusion network (TFN). TFN expands each mode and calculates the fusion result using the Cartesian product. Inspired by TFN, [3] adopts a low-rank weight for multimodal fusion, which reduces the number of parameters and improves the calculation speed. The late fusion is the weighted fusion of the sentiment analysis results of every unimodality [31]. Despite being straightforward and useful for modeling modality-specific interactions, this technique does not account for cross-modal interactions [32]. Hybrid fusion is the combination of the first two methods, which can model the relationships within and between modes of multiple modalities, better excavate the complementarity between multiple modalities, and obtain a better representation of multimodal features [33].

Reference [34] proposes a hierarchical feature fusion method, which first fuses two of the three modalities, and then fuses all the modalities. Although this approach is helpful for sentiment analysis, it does not consider the dynamic interactions between modalities over time series. Later, inspired by the self-attention mechanism in the NLP field, [13] proposes a multimodal transformer (MulT), which uses a crossmodal attention mechanism to analyze the dynamic relationship between modalities. MulT also adopts a hierarchical fusion method. Based on the three modalities, this is fused with the other two modalities using a crossmodal transformer, and then the fusion results of the three modalities are input into the traditional transformer for feature learning. Finally, the sentiment prediction result is obtained. However, these methods treat different modalities equally, and do not consider the problem of noise in the model. Multiplicative multimodal

emotion recognition (M3ER) [14] was proposed to solve this problem. By designing the modality inspection module to set the relevant threshold to select the noise modality, M3ER generates proxy feature vectors for this noise modality and introduces the multiplicative fusion module to fuse all modality features. Although this method solves the modality noise problem to a certain extent, it removes the noise model and generates the features of the removed model through the proxy module, which reduces the characteristic information of the model itself to a certain extent.

Therefore, this paper considers the problems with the dynamic relationship between different modalities in the time series, the noise in the modalities and the various contributions of different modalities to the final sentiment analysis to design a cascade fusion method based on the attention mechanism and the strategy of computing the weights.

III. METHOD

In this section, we first define the task of multimodal sentiment analysis, then introduce our proposed adaptive modality-specific weight fusion network (AdaMoW) in detail, including the main sub-modules and the method of model optimization.

A. NOTATIONS

Table 1 summarizes the notations used in this work. We denote the set of modalities as $\mathcal{M} = \{t, v, a\}$, in which t, v and a stand for *text*, *vision* and *audio* modalities, respectively. For the rest of the paper, $m \in \mathcal{M}$ represents a specified modality. Let $\mathcal{I}_t \in \mathbb{R}^{\tau_t \times d_t}$, $\mathcal{I}_v \in \mathbb{R}^{\tau_v \times d_v}$ and $\mathcal{I}_a \in \mathbb{R}^{\tau_a \times d_a}$ be the feature vectors for each modality, where $\tau_{\{t,v,a\}}$ represents the sequence length and $d_{\{t,v,a\}}$ is feature dimension of modality m, respectively. Generally, the task of MSA can be considered a classification task of emotion recognition, or a regression task of sentiment polarity analysis. In this paper, our model regards it as a regression task, but in the later part of the experiment, the evaluation index of the related classification task is also calculated, so that the effectiveness of AdaMoW can be repeatedly verified from two aspects. y is the actual sentiment polarity of multimodal data. The final multimodal prediction of our model is $\hat{y}_s$.

B. OVERVIEW

As shown in Figure 2, our model is mainly divided into four parts: the unimodal feature extraction module, weight-based feature fusion module, modality-specific feature optimization module, and multimodal sentiment prediction module. Specifically, we first input the modality data into the unimodal feature extraction module and adopted different feature extraction methods for diverse modalities to fully consider the modality uniqueness. After that, unimodal features were input to the weight-based feature fusion module, the modal labels were used to learn the contribution of different modalities to the final emotional polarity during the training process and the weight-mapping network was trained to also calculate the contribution of various modalities very well during the non-training state. In addition, we also used the crossmodal attention mechanism to align different modalities, which helps to fuse multimodal information better. Then, different unimodal data were assigned diverse weights according to variant contributions and input to the modality-specific feature optimization module. This module used a unimodal feature generator to generate unimodal features and optimize its representation. Finally, the multimodal feature fusion result was input into the multimodal sentiment prediction module to obtain the final sentiment prediction result.

C. UNIMODAL FEATURE EXTRACTION

To better extract unimodal data features, we adopted diverse feature extraction methods for different modalities. First, for vision and audio modalities, we used self-attention networks for feature extraction:

$$\hat{\mathcal{I}}_m = \text{Attention}_m\left(\mathcal{I}_m; \theta_m\right) \in \mathbb{R}^{\tau_m \times d_m}, \quad m \in \{v, a\}, \tag{1}$$

where $Attention_m(\cdot)$ represents the self-attention network of modality m and θ_m is the model's parameter. $\hat{\mathcal{I}}_m$ is the result of feature extraction of modality m, respectively. τ_m mean the sequence length and d_m is the feature dimension of different modalities. For text modalities, feature representation is a basic task in the NLP field. BERT [35] has been recognized as the SOTA method in the NLP domain by virtue of its powerful ability and tremendous success in text-based machine learning applications. Therefore, we used BERT to extract features of text modality:

$$\hat{\mathcal{I}}_t = \text{BERT}\left(\mathcal{I}_t; \theta_t\right) \in \mathbb{R}^{\tau_t \times d_t}, \tag{2}$$

where $\hat{\mathcal{I}}_t$ is the result of the feature extraction of text modality and θ_t is the parameter of BERT. Then, to provide each element of the input sequence with sufficient information of about its neighborhood elements, we fed three modalities into a 1D temporal convolutional layer:

$$f_m = \text{Conv1D}\left(\hat{\mathcal{I}}_m, k_m\right) \in \mathbb{R}^{\tau_m \times d_m}, \quad m \in \{t, v, a\}, \tag{3}$$

where $Conv1D(\cdot)$ represents the operation of a 1D convolutional layer and k_m is the size of the convolutional kernel. f_m is the output of unimodal feature extraction.

D. WEIGHT-BASED FEATURE FUSION

In the feature fusion stage, we designed a weight of modality calculation method based on unimodality correlation to address the problem of different modalities' contributions to the final sentiment analysis results. Meanwhile, we used the crossmodal attention mechanism and each unimodal weight for multimodal feature fusion to obtain fusion results considering modal correlation and timing.

First, we defined the crossmodal attention mechanism:

$$CA_{m_1 \to m_2} = \text{Softmax}\left(\frac{W_Q f_a \cdot W_K^\top f_{m_2}}{\sqrt{d_k}}\right) W_V f_{m_2} \tag{4}$$

TABLE 1. Summary of basic notations.

Notation	Description	Notation	Description
t, v, a	Text, vision and audio modalities	$\mathcal{M}$	$\mathcal{M} = \{t, v, a\}$, the set of modalities
τ_m	The sequence length of modality m	d_m	The feature dimension of modality m
$\mathcal{I}_m$	Model input of modality m	f_m	The output of unimodal feature extraction
y	The actual sentiment polarity of multimodal data	$\hat{y}_s$	Output of AdaMoW
$\hat{y}_m$	The prediction of modality m	$\mathbb{R}^{d_t}$	d_t matrice
k_m	The size of the convolutional kernel of modality m	$CA_{a\to b}$	The operation of crossmodal attention
Q	Query for multi-head attention	K	Key for multi-head attention
V	Value for multi-head attention	f'_m	The feature vector of modality m
ε_m	The relevance of modality m with the ground truth	ω_m	The weight of modality m
$\hat{\omega}_m$	The learned weight of modality m	x_m	The representation of the unimodal feature with weights
x_s	The fusion result of multimodal	$\bar{x}_m$	The generated feature for modality m by generator $G_m(\cdot)$
α	The hyper-parameter	$\mathcal{L}_m$	The loss of NCE of modality m
$\mathcal{L}_{NCE}$	The NCE loss function	$\mathcal{L}_{weight}$	The loss of the unimodal weight
$\mathcal{L}_{task}$	The task loss function	$\mathcal{L}_{main}$	The main loss function

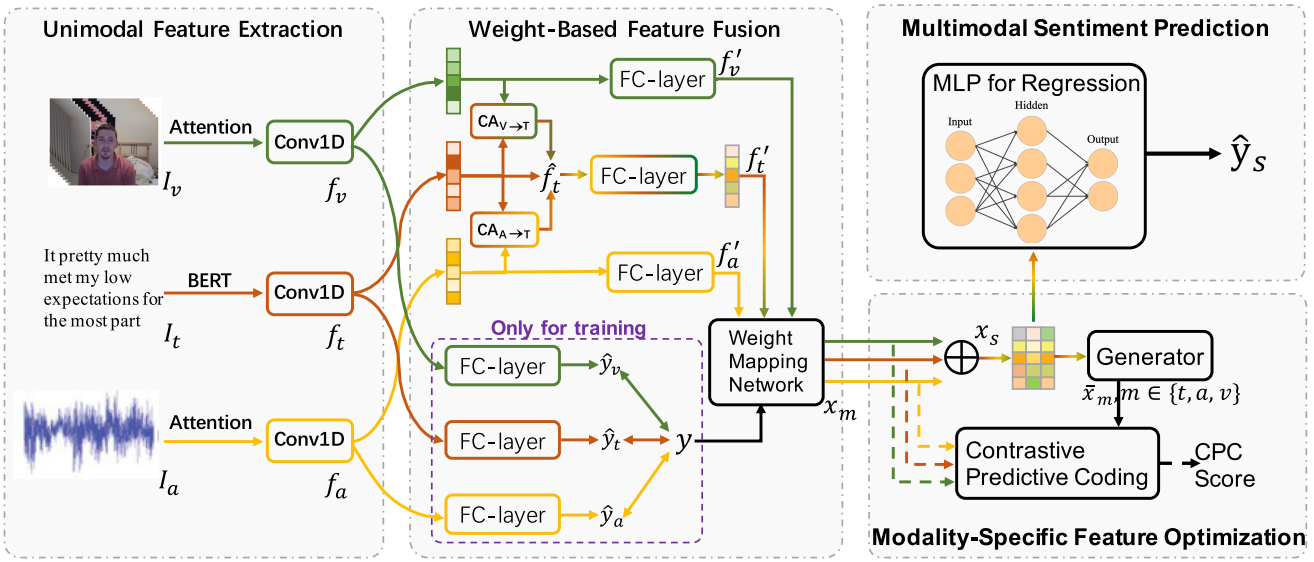


FIGURE 2. The diagram of our proposed AdaMoW.

where $CA_{m_1\to m_2}$ represents the result of the crossmodal attention of modality feature f_{m_1} and f_{m_2}. $Softmax(\cdot)$ is a normalized exponential function whose output vector ranges from 0 to 1 for each element, and the sum of all elements is 1. The attention mechanism is to get a number between 0 and 1 to indicate the importance of different pieces of data, with 1 being very important and 0 being very little important. The inputs f_{m_1} and f_{m_2} are the output of layer normalization operation and we calculated the Query (Q), Key (k) and Value (V) in the attention mechanism using these inputs. Specifically, $Q = W_Q \cdot f_{m_1}, K = W_K \cdot f_{m_2}$ and $V = W_V \cdot f_{m_2}$. W_Q, W_K and W_V are the query, key and value weight, respectively. As text modality contains more sentiment information [36], we used text as the primary modality and applied the crossmodal attention mechanism to obtain the fusion results of text-vision and text-audio modalities. Then, the two fused results were added to the original text features:

$$\hat{f}_t = f_t + CA_{v\to t} + CA_{a\to t} \in \mathbb{R}^{\tau_t \times d_t} \tag{5}$$

where $\hat{f}_t$ represents the feature vector of the text modality after the crossmodal attention. Finally, we mapped the feature representations of the three modalities to a low-dimensional feature representation space using three layers of fully connected and ReLU functions:

$$f'_m = \begin{cases} G_m(\hat{f}_m, \theta_{G_m}) \in \mathbb{R}^{d_m}, & m = t \\ G_m(f_m, \theta_{G_m}) \in \mathbb{R}^{d_m}, & m \in \{v, a\} \end{cases} \tag{6}$$

where $G_m(\cdot)$ represents the network of the feature mapping and θ_{G_m} is the parameter of $G_m(\cdot)$ of modality m, respectively.

Second, we leveraged multimodal sentiment polarity labels to explore unimodal weights during training. Specifically, we first input the feature representation of each model into the prediction network to obtain the predicted value of each modality:

$$\hat{y}_m = P_m(f_m, \theta_{P_m}), \quad m \in \{t, v, a\} \tag{7}$$

where $\hat{y}_m$ is the prediction of modality m, respectively. P_m represents the prediction network with three linear layers

and the activation function of ReLU. The parameters of P_m are θ_{P_m}.

We then calculated its relevance to the ground truth using the L_1 norm:

$$\varepsilon_m = L_1(\hat{y}_m, y) = |\hat{y}_m - y|, \quad m \in \{t, v, a\} \tag{8}$$

where ε_m represents the relevance of modality m with the ground truth; $L_1(\cdot)$ and $|\cdot|$ are the operation of L_1 norm. y is the ground truth. The smaller the ε_m value, the closer the unimodal prediction result is to the final emotional polarity of the multimodal, and the greater the weight that should be given. Therefore, we calculated weights according to their correlation:

$$\omega_m = \frac{e^{-\varepsilon_m * c}}{\sum_i^{t,v,a} e^{-\varepsilon_i * c}}, \quad m \in \{t, v, a\} \tag{9}$$

where ω_m is the weight of different modalities and c is a weight constant. However, the ground truth is only present in the training set, and no true labels are provided in the test and validation data. Therefore, we learned the mapping relationship between unimodal data and its weights by constructing a weight-mapping network:

$$\hat{\omega}_m = W_m(f'_m, \theta_{M_m}), \quad m \in \{t, v, a\} \tag{10}$$

where $W_m(\cdot)$ is a neural network with parameters θ_{W_m} that learn a mapping of $\hat{\omega}_m$ from f'_m. $\hat{\omega}_m$ is the learned weight of modality m and the L_1 norm is used as its objective $\mathcal{L}_{weight}$ function to optimize the network:

$$\mathcal{L}_{weight} = \frac{1}{N} \sum_j^N \left(\sum_i^{\{t,a,v\}} |\omega_i^j - \hat{\omega}_i^j| \right), \tag{11}$$

where N is the batch size.

Finally, the weights of each modality were fused with the above-mentioned unimodal feature vector to obtain the final weight fusion result based on multimodal correlation:

$$\begin{aligned} x_m &= \omega_m * f'_m, m \in \{t, v, a\}, \\ x_s &= Concat(x_t, x_v, x_a) \end{aligned} \tag{12}$$

where x_m is the unimodal feature representation with weights, and we concatenated all modalities' features as the final output x_s of the weight-based feature fusion module.

E. MODALITY-SPECIFIC FEATURE OPTIMIZATION

To ensure that the fusion features better retained the specific information of each modality, and inspired by, but different to, MMIM [37], we constructed a modality-specific feature optimization module with a unimodal feature generator $G_m(\cdot)$ and InfoNCE framework. Specifically, we first fed the multimodal fusion results x_s into the unimodal generator to generate features for three modalities:

$$\bar{x}_m = G_m(x_s, \theta_{G_m}), \quad m \in \{t, v, a\}, \tag{13}$$

where $\bar{x}_m$ is the generated feature for modality m. $G_m(\cdot)$ represents a neural network of the unimodal feature generator with parameters θ_{G_m}. We optimized our generative features $\bar{x}_m$ with the Noise-Contrastive Estimation framework [38], assuming that one of our samples is $\mathcal{X}_m$, and $x_m^i \in \mathcal{X}_m$ is the input of i-th data of modality m. Hence, the negative samples of i-th data are $\tilde{\mathcal{X}}_m^i = \mathcal{X}_m / x_m^i$. The correlation of prediction and truth vectors can be computed as:

$$sim(x_m^i, \bar{x}_m^i) = exp\left(\frac{x_m^i (\bar{x}_m^i)^\top}{||x_m^i||_2 \cdot ||\bar{x}_m^i||_2} \right), \quad m \in \{t, v, a\}, \tag{14}$$

where $sim(\cdot)$ represents the similarity score of modality m, respectively. $||\cdot||_2$ is the Euclidean norm used to normalize the feature vectors. The InfoNCE loss is defined as $\mathcal{L}_{NCE} = \sum_m^{\{t,v,a\}} \mathcal{L}_m$, where $\mathcal{L}_m$ is the loss of modality m:

$$\mathcal{L}_m = -\log \frac{sim(x_m^i, \bar{x}_m^i / \tau)}{\sum_{x_m^j \in \mathcal{X}_m} sim(x_m^j, \bar{x}_m^j / \tau)}, \tag{15}$$

where $\tau > 0$ is a temperature parameter. This can attract the positive sample pairs to each other by InfoNCE loss, that is, the similarity between the generated modal features and the modal feature extraction vectors is improved, and the model fusion features can be promoted to obtain more modality-specific information on each modality.

F. MULTIMODAL SENTIMENT ANALYSIS

Based on the well-designed unimodal feature representation learning and multimodal feature fusion optimization strategy, we used a simple multi-layer perceptron (MLP) to predict sentiment polarity. Specifically, we input the fusion result x_s into the fully connected layer and used ReLU as the activation function:

$$\hat{y}_s = \sigma(x_s) \tag{16}$$

where $\hat{y}_s$ is the output of MLP as the final prediction result of multimodal data and $\sigma(\cdot)$ represents the operation of MLP. In the multi-task of multimodal and unimodal sentiment analysis, we used the L_1 loss as the optimization objective function:

$$\mathcal{L}_{task} = \frac{1}{N} \sum_j^N \left(\sum_i^{\{t,a,v,s\}} |y^j - \hat{y}_i^j| \right), \tag{17}$$

where N is the batch size. Finally, we defined the loss function $\mathcal{L}_{main}$ of the entire model as:

$$\begin{aligned} \mathcal{L}_{main} &= \mathcal{L}_{task} + \mathcal{L}_{weight} + \alpha \mathcal{L}_{NCE} \\ &= \frac{1}{N} \sum_j^N \left(\sum_i^{\{t,a,v,s\}} |y^j - \hat{y}_i^j| + \sum_i^{\{t,a,v\}} |\omega_i^j - \hat{\omega}_i^j| \right) \\ &\quad + \alpha \sum_m^{\{t,v,a\}} -\log \frac{sim(x_m^i, \bar{x}_m^i / \tau)}{\sum_{x_m^j \in \mathcal{X}_m} sim(x_m^j, \bar{x}_m^j / \tau)}, \end{aligned} \tag{18}$$

where α is the hyper-parameter that controls the impact of the NCE loss. We summarize the training process of our entire model in Algorithm 1.

Algorithm 1 AdaMoWAlgorithm

Input: Dataset $\mathcal{D} = \{(\mathcal{I}_t, \mathcal{I}_v, \mathcal{I}_a), Y\}$
Output: Prediction $\hat{y}$ # Predicted sentiment polarity

for each training epoch **do**
  for minibatch $\mathcal{B} = \{(\mathcal{I}_t^i, \mathcal{I}_v^i, \mathcal{I}_a^i)\}_{i=1}^B$ sampled from $\mathcal{D}$ **do**
    # Refer to Subsection (III-C)
    Encode $\mathcal{I}_m^i$ to $\hat{\mathcal{I}}_m^i$ as (1),(2)
    Encode $\hat{\mathcal{I}}_m^i$ to f_m^i using 1D convolution as (3)
    # Refer to Subsection III-D
    Compute $CA_{V\to T}$ and $CA_{A\to T}$ as (4)
    # CA is Crossmodal Attention block
    Produce features of three modalities $f_m'^i$ as (5), (6)
    Produce predictions of each modality $\hat{y}_m^i$ as (7)
    Use L_1 to compute correlations ε_m^i as (8)
    Compute unimodality weights ω_m^i as (9)
    Compute learned weights $\hat{\omega}_m^i$ as (10)
    Produce features with weights x_m^i and x_s^i as (12)
    # Refer to Subsection III-E
    Produce $\bar{x}_m^i$ by generator as (13)
    Use InfoNCE to compute $\mathcal{L}_{NCE}$ as (14),(15)
    # Refer to Subsection (III-F)
    Produce predictions of multimodal data $\hat{y}_s^i$ as (16)
    Use L_1 to calculate $\mathcal{L}_{weight}$ and $\mathcal{L}_{task}$ as (11),(17)
    Add $\mathcal{L}_{task}$, $\mathcal{L}_{weight}$ and $\mathcal{L}_{NCE}$ to compute $\mathcal{L}_{main}$ as (18)
  end for
end for

IV. EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed framework using two public datasets. We introduce detailed comparisons with the baseline of the MSA task and present the results of ablation studies.

A. DATASETS

We used two real-world datasets in the experimental verification, CMU-MOSI [39] and CMU-MOSEI [40].

1) CMU-MOSI

The CMU-MOSI dataset is commonly used in the field of multimodal sentiment analysis. This dataset compiles YouTube video blogs (vlogs) that mostly discuss movie reviews. A total of 93 videos, each between two and five minutes long, were gathered randomly. The 89 speakers included in the films, 41 women and 48 men, were primarily in their 20s and 30s and represented various racial groups. The annotations of these videos were annotated and averaged by five annotators from the Amazon crowdsourcing platform, and annotated as seven types of emotional tendencies, from −3 to +3.

2) CMU-MOSEI

The CMU-MOSEI dataset was selected from review videos on YouTube: 22852 annotated video clips, from 1000 different speakers and 250 topic types. Face detection technology was used during collection to ensure videos were single-person commentary, and videos where the speaker's attention is fully on the camera were selected. The dataset has both sentiment and emotion annotations. The sentiment labeling is a 7-category sentiment labeling for each sentence. Each video was annotated with both emotion and emotion. Sentiment was labeled as seven integers in [−3, 3]. Emotional labeling was an emotional labeling that includes six aspects: angry, disgust, fear, happy, sad and surprise.

TABLE 2. Dataset statistics in CMU-MOSI and CMU-MOSEI.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856

As mentioned in Section III, we used the three modalities of text, video, and audio in the datasets. For the raw data, we directly obtained the preliminary processed features in the dataset publisher's CMU-Multimodal SDK [40], and then input the feature vectors of different modalities into the unimodal feature extraction module of our proposed model. Moreover, we used the same dataset split method of CMU-Multimodal SDK for a fair comparison with the baselines, as shown in Table 2.

B. EVALUATION METRICS

Based on previous works [41], [42] and the characteristics of the datasets, we used the evaluation metrics of both classification and regression tasks to verify the effectiveness of our proposed model. For classification, we used classification accuracy and F1 score as evaluation metrics. Regarding the classification categories, we obtained two classifications for emotional polarity, including binary classification (ACC-2), and also calculated seven types of classification accuracy (ACC-7) for the seven emotional polarities unique to the CMU-MOSI and CMU-MOSEI datasets. In addition, we also considered two different situations to calculate the binary classification, namely positive/negative (P/N) and non-negative/negative (NN/N) classification. We used two evaluation indicators for the regression task: mean absolute error (MAE) and correlation (Corr).

C. BASELINES

To prove the effectiveness of our proposed framework, we selected the latest multimodal sentiment analysis model and classic methods as the baseline to compare with AdaMoW. Here is a brief introduction to these methods:

1) TFN [2]

The tensor fusion network performs fusion, utilizing the outer product of these elements, after breaking up high-order tensors into several low-rank components using a three-fold Cartesian product.

2) LMF [3]
To accomplish an effective fusion, low-rank multimodal fusion decomposes stacked high-order tensors into low-rank weight tensors.

3) MFM [43]
The multimodal factorization model is a generative-discriminative model that uses inference and generative networks to capture information within and between modalities.

4) ICCN [36]
The interaction canonical correlation network accomplishes the fusion process by leveraging mathematical measures to minimize the canonical loss of modality representation pairs.

5) MulT [13]
The multimodal transformer improves the original attention mechanism module that can only input single modal data, constructs an attention mechanism that integrates multimodal information and transcends the time limit, and optimizes the fusion process.

6) MISA [41]
This work projects signals from different modalities into two separate spaces, modality-invariant and specific representations, to complete the fusion process.

7) BIMHA [9]
By adopting the conventional multi-attentional mechanism, bimodal information-augmented multi-head attention improves the interaction between two pair-wise modalities by obtaining the interaction information between the modalities through tensor fusion.

8) MAG-BERT [42]
The multimodal adaptation gate enables BERT and XLNet to use multimodal information for fusion after pre-training by using an alignment gate to convert visual and auditory information.

9) CubeMLP [44]
CubeMLP inputs and mixes relevant modal features across three axes through three independent MLP units, and flattens the mixed multimodal features for task prediction.

D. RESULTS AND DISCUSSION

Table 3 shows the comparative results on the CMU-MOSI and CMU-MOSEI datasets. The experimental accuracy of the baseline model comes from the original text of the paper or the reproduction results of other papers, and the specific sources are marked in the table. We can see from the table that, on the CMU-MOSI, dataset all the evaluation indicators of our model are better than other baselines, and the accuracy of the seven classifications is nearly 3 percent higher than the latest CubeMLP, reaching 48.10%. Corr also improved by more than 0.04 compared to the second-place finisher on this metric, reaching 0.807. On the CMU-MOSEI dataset, we find that MAE, Corr, ACC-2 (NN/N and P/N) and F1 (NN/N and P/N) of AdaMoW achieve the best results, with 0.523, 0.769, 85.1, 85.98, 84.78 and 85.90, respectively. Although the seven classification accuracy could be better, it is only 0.42% behind first place.

Specifically, compared with the TFN and LMF directly fused with tensor, the effect of our proposed model was dramatically improved. The ACC-7 index of the CMU-MOSI dataset increased by more than 13 percent, and the Corr of the CMU-MOSEI dataset also increased by nearly 0.07. This reflects the effectiveness of the fusion strategy we designed. Compared with the MFM, which also has a generative network, our proposed model improves in all evaluation metrics. On the CMU-MOSI dataset, the MAE index was improved from 0.877 to 0.696, and the ACC-7 was improved from 35.4 to 48.1. On the CMU-MOSEI dataset, Corr increased from 0.717 to 0.769, and ACC-7 increased from 84.4 to 54.48. The ICCN and MISA improve the learning of multimodal feature representations. Compared with these two methods, our proposed model also exhibits a higher performance. On the CMU-MOSI dataset, the ACC-2 (P/N) value increased from 83.0 and 82.1 to 86.58, and the F1 (P/N) increased from 83.0 and 82.03 to 86.57. On the CMU-MOSEI dataset, the ACC-2 (P/N) value increased from 84.2 and 84.23 to 85.98, and the F1 (P/N) increased from 84.2 and 83.97 to 85.9. MulT, BIMHA, and MAG-BERT all use attention mechanisms to fuse modal information. The method proposed in this paper also uses the attention mechanism, but we also consider the different weights of fusion modalities. The experimental results on the CMU-MOSI and CMU-MOSEI datasets show that the evaluation indicators of our proposed method are better than these methods, which proves the effectiveness of our proposed weight calculation method. Compared with the latest SOTA method, CubeMLP, all evaluation indicators of our model on the dataset CMU-MOSI were improved, and the ACC-7 indicator increased by nearly 3 percent. On the CMU-MOSEI dataset, while the ACC-7 evaluation index of AdaMoW is not as high as that of CubeMLP, the other metrics are higher. The difference between the ACC-7 of CubeMLP and AdaMoW is only 0.42, but AdaMoW's F1(P/N) is 1.4 higher than that of CubeMLP. This indicates that AdaMoW's accuracy for multi-category sentiment analysis is not as high as that for binary classification sentiment analysis. However, based on the comparison results across all datasets and baseline models, the proposed method shows improvement over previous methods and achieves high accuracy for multimodal sentiment analysis tasks.

E. SENSITIVITY ANALYSIS FOR α

The loss function $\mathcal{L}_{main}$ of our proposed model consists of three parts: the sentiment prediction loss function $\mathcal{L}_{task}$, the weight calculation loss function of each modality of the weight-mapping network $\mathcal{L}_{weight}$, and the loss function of the unimodal feature generator $\mathcal{L}_{NCE}$. The sentiment prediction

TABLE 3. Prediction accuracy of the AdaMoW on the test dataset of CMU-MOSI and CMU-MOSEI. †, ‡, ∗ and ✳ indicate the results from [37], [41], [45] and [9]. We tested two sets of non-negative/negative (left) and positive/negative (right) evaluations for the ACC-2 and F1 indexes. Our outcomes were the best or close to the best. The top performances are highlighted in bold. ↑ means higher is better and ↓ means lower is better.

Model	CMU-MOSI					CMU-MOSEI				
	MAE ↓	Corr ↑	ACC-7 ↑	ACC-2 ↑	F1 ↑	MAE ↓	Corr ↑	ACC-7 ↑	ACC-2 ↑	F1 ↑
TFN†	0.901	0.698	34.9	- /80.8	- /80.7	0.593	0.700	50.2	- /82.5	- /82.1
LMF†	0.917	0.695	33.2	- /82.5	- /82.4	0.623	0.677	48.0	- /82.0	- /82.1
MFM†	0.877	0.706	35.4	- /81.7	- /81.6	0.568	0.717	51.3	- /84.4	- /84.3
ICCN†	0.862	0.714	39.0	- /83.0	- /83.0	0.565	0.713	51.6	- /84.2	- /84.2
MulT‡	0.861	0.711	-	81.5/84.1	80.6/83.9	0.580	0.703	-	- /82.5	- /82.3
MISA‡	0.804	0.764	-	80.79/82.10	80.77/82.03	0.568	0.724	-	82.59/84.23	82.67/83.97
BIMHA✳	0.925	0.671	36.44	78.57/80.3	78.5/80.03	0.559	0.731	52.11	84.07/83.96	83.35/83.5
MAG-BERT*	0.727	0.781	43.62	82.37/84.43	82.50/84.61	0.543	0.755	52.67	82.51/84.82	82.77/84.71
CubeMLP	0.770	0.767	45.50	-/85.60	-/85.50	0.529	0.760	**54.90**	-/85.10	-/84.50
Ours	**0.696**	**0.807**	**48.10**	**85.09/86.58**	**85.13/86.57**	**0.523**	**0.769**	54.48	**85.10/85.98**	**84.78/85.90**
Rank	**1**	**1**	**1**	**1/1**	**1/1**	**1**	**1**	2	**1/1**	**1/1**

TABLE 4. Analysis of the sensitivity of α on the CMU-MOSI and CMU-MOSEI datasets. The parameters selected by this framework are shown in bold.

Description		MAE ↓	Corr ↑	ACC-7 ↑	ACC-2 ↑	F1 ↑
MOSI	α=0.1	0.767	0.775	46.94	81.09/82.61	81.05/82.62
	α=0.5	**0.696**	**0.807**	**48.10**	**85.09/86.58**	**85.13/86.57**
	α=1	0.715	0.789	45.19	82.06/84.08	82.07/84.15
MOSEI	α=0.1	**0.523**	**0.769**	**54.48**	**85.10/85.98**	**84.78/85.90**
	α=0.5	0.560	0.749	52.26	83.22/84.95	83.32/84.73
	α=1	0.571	0.746	50.68	83.85/84.25	83.71/83.65

TABLE 5. Ablation experiments on the fusion of different modal numbers and whether to assign modal weights on the dataset CMU-MOSI. T, V, and A represent the text, video and audio modalities, respectively, while W indicates the computation of weights for unimodal data. The group with the best experimental results is shown in bold.

Num	Model	MAE ↓	Corr ↑	ACC-7 ↑	ACC-2 ↑	F1 ↑
1	T+V	1.442	0.052	17.06	54.96/54.91	54.96/55.03
2	T+A	1.397	0.284	15.60	57.60/56.82	56.27/55.64
3	V+A	1.407	0.267	16.18	57.62/56.13	55.69/54.42
4	T+V+W	0.986	0.697	30.47	80.10/83.20	78.86/82.32
5	T+A+W	0.806	0.758	44.17	82.46/84.63	82.36/84.6
6	V+A+W	1.172	0.472	23.18	71.76/73.11	71.72/73.17
7	T+V+A	0.728	0.780	45.19	82.89/85.06	82.65/84.91
8	T+V+A+W	**0.696**	**0.807**	**48.10**	**85.09/86.58**	**85.13/86.57**

loss includes unimodal and multimodal predictions. Since both $\mathcal{L}_{task}$ and $\mathcal{L}_{weight}$ use L_1 to calculate the loss, and $\mathcal{L}_{NCE}$ uses NCE loss for calculation, we set a hyperparameter α for $\mathcal{L}_{NCE}$ when calculating the loss of the entire model to adjust its impact on the overall loss. Table 4 shows the outcomes of the experiments. Based on our experiments and statistical analysis, we set three values of 0.1, 0.5, and 1 for α to analyze the best hyperparameter. The specific experimental results are shown in Table 4. It can be seen from the table that, for the CMU-MOSI dataset, when α is 0.5, the experimental results were the best. For the CMO-MOSEI dataset, when α is 0.1, the model's performance is better than that using other values. This further demonstrates the effectiveness of using a unimodal feature generator to optimize multimodal fusion results.

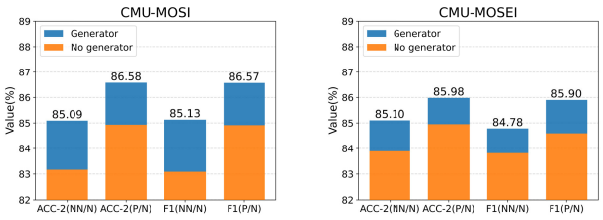
F. ABLATION STUDIES

To verify the effectiveness of our proposed weight-based fusion method, we further explored the influence of the number of modalities and modal weights on the sentiment analysis results of the CMU-MOSI dataset. The specific experimental results are shown in Table 5.

Specifically, we divided the experiments into four groups: the first group represents the fusion results of two modalities, and the second group represents the fusion results of two modalities with weights. The third and fourth groups represent the results of the joint fusion of three models without weight and with weight, respectively. The experimental results with modal weights are generally better than those without weights. For example, the effect of experiment 8 is better than that of experiment 7, and the effects of various mode settings in the second group are better than those without weight in the corresponding settings in the first group.

From the perspective of the number of modalities, through modal correlation analysis, we can eliminate the impact of modal noise on the final multimodal sentiment analysis while maintaining the unique information of the modalities. Therefore, from the comparison between the second and fourth groups of experiments, it can be seen that, under the influence of weights, the effect of three-modal fusion is better than that of all two-modal fusions in the second group.

In the case of the same number of modes, it can be seen that the ACC-7 index of experiment 8 is nearly 3 percent higher than that of experiment 7 without weights for the simultaneous fusion of the three modalities. The ACC-2 (NN/N) and F1 (NN/N) indexes increased to 85.09 and 85.13 from 82.89 and 82.65. For the fusion of the two modalities, the ACC-2 and F1 values of the second group with weighted fusion increased by about 20 percent compared with the unweighted fusion of the first group. In addition, the Corr



(a) CMU-MOSI (b) CMU-MOSEI

FIGURE 3. Ablation studies to evaluate the effect of generator on the CMU-MOSI and CMU-MOSEI datasets.

index of the weighted text and audio modalities fusion in experiment 4 increased from the original 0.284 to 0.752. In summary, all the above experiments demonstrate that the weight-based modality fusion method can effectively utilize the modality-specific information of a unimodality, eliminate the influence of modality noise on the final sentiment, and promote the effect of multimodal sentiment analysis.

In addition, we explore the effectiveness of the generator in the modality-specific feature optimization module. As shown in Figure 3, we experimented with the model with and without the generator on the CMU-MOSI and CMU-MOSEI datasets. The experimental results show that the experiment with the generator is better in all datasets. Specifically, on the CMU-MOSI dataset, ACC-2 (NN/N) and F1 (NN/N) indicators increased by 2 percent. ACC-2 (P/N) and F1 (P/N) indicators also rose by more than one percent. On the CMU-MOSEI dataset, the F1 (P/N) index rose the most, exceeding one percent. Overall, the effect of the model with a generator is better than other cases, which reflects the effectiveness of the modality-specific feature optimization module.

V. CONCLUSION

To better explore unimodal information for multimodal sentiment fusion, this paper proposed a data-driven multimodal sentiment analysis framework based on a weight-based feature fusion method to capture more modality-specific information. Following this method, noisy modalities or those that contributed little to multimodal sentiment analysis decisions were given less fusion weight while preserving modality-specific information. During testing and validation, we obtained the weights for each modality through a weight-mapping network that learns optimal parameters during training. To make the fused features retain as much unimodal-specific information as possible, we tailored a unimodal feature generator to optimize the fusion result. We compared our proposed model with the baseline on the CMU-MOSI and CMU-MOSEI datasets, and the results show that our model is competitive and reaches the state-of-the-art. Ablation experiments could further validate the effectiveness of our proposed framework and modules. In future work, we can explore multi-task learning research for multimodal sentiment analysis, and also add prior knowledge to guide weight analysis in the process of feature fusion.

REFERENCES

[1] 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, Sep. 2017.

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

[3] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[4] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 122–137, Jan. 2020.

[5] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification," *Comput. Intell.*, vol. 36, no. 2, pp. 861–881, May 2020.

[6] Y. Yuan, X. Mu, X. Shao, J. Ren, Y. Zhao, and Z. Wang, "Optimization of an auto drum fashioned brake using the elite opposition-based learning and chaotic k-best gravitational search strategy based grey wolf optimizer algorithm," *Appl. Soft Comput.*, vol. 123, Jul. 2022, Art. no. 108947.

[7] M. Frank, D. Drikakis, and V. Charissis, "Machine-learning methods for computational science and engineering," *Computation*, vol. 8, no. 1, p. 15, Mar. 2020.

[8] Y. Yuan, J. Ren, S. Wang, Z. Wang, X. Mu, and W. Zhao, "Alpine skiing optimization: A new bio-inspired optimization algorithm," *Adv. Eng. Softw.*, vol. 170, Aug. 2022, Art. no. 103158.

[9] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, and Y. Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107676.

[10] Y.-H.-H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1823–1833.

[11] D. Gu, J. Wang, S. Cai, C. Yang, Z. Song, H. Zhao, L. Xiao, and H. Wang, "Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021.

[12] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 320–334, Jan. 2022.

[13] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[14] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1359–1367.

[15] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, and X. Wu, "Chinese text sentiment analysis based on extended sentiment dictionary," *IEEE Access*, vol. 7, pp. 43749–43762, 2019.

[16] C. R. Aydin and T. Gungör, "Combination of recursive and recurrent neural networks for aspect-based sentiment analysis using inter-aspect relations," *IEEE Access*, vol. 8, pp. 77820–77832, 2020.

[17] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, 2011, pp. 142–150.

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

[19] W. Jiao, H. Yang, I. King, and M. R. Lyu, "HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2019, pp. 397–406.

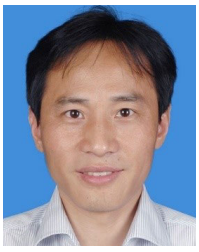
[20] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7342–7346.

[21] M. A. Jalal, R. Milner, and T. Hain, "Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition," in *Proc. Interspeech*, Oct. 2020, pp. 4113–4117.

[22] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Exp. Syst. Appl.*, vol. 173, Jul. 2021, Art. no. 114683.
[23] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 715–718.
[24] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–8.
[25] D. She, J. Yang, M. Cheng, Y. Lai, P. L. Rosin, and L. Wang, "WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1358–1371, May 2020.
[26] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.
[27] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer, "PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 192–201.
[28] F. Huang, X. Li, C. Yuan, S. Zhang, J. Zhang, and S. Qiao, "Attention-emotion-enhanced convolutional LSTM for sentiment analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4332–4345, Sep. 2022.
[29] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song, "What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis," *Inf. Fusion*, vol. 66, pp. 184–197, Feb. 2021.
[30] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal DNN feature fusion," in *Proc. Grand Challenge Workshop Human Multimodal Lang. (Challenge-HML)*, 2018, pp. 11–19.
[31] S. A. Abdu, A. H. Yousef, and A. Salem, "Multimodal video sentiment analysis using deep learning approaches, a survey," *Inf. Fusion*, vol. 76, pp. 204–226, Dec. 2021.
[32] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.
[33] R. Li, J. Zhao, J. Hu, S. Guo, and Q. Jin, "Multi-modal fusion for video sentiment analysis," in *Proc. 1st Int. Multimodal Sentiment Anal. Real-Life Media Challenge Workshop*, Oct. 2020, pp. 19–25.
[34] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.
[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2019, pp. 4171–4186.
[36] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.
[37] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.
[38] M. Gutmann and A. Hyvarinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 297–304.
[39] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.
[40] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[41] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and—Specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131.
[42] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.
[43] Y.-H. H. Tsai, P. Pu Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," 2018, *arXiv:1806.06176*.
[44] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3722–3729.
[45] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.

JUNLING ZHANG received the B.Eng. degree from the College of Mathematics, Physics and Information Science, Zhejiang Ocean University, Zhoushan, China, in 2020. She is currently pursuing the degree with the College of Computer Science and Technology, Zhejiang Normal University, Jinhua, China. Her research interests include machine learning, natural language processing, and affective computing.

XUEMEI WU received the Ph.D. degree in intelligence education from South China Normal University. She is currently a Lecturer with Zhejiang Normal University, Jinhua, Zhejiang, China. Her main research interests include teacher professional development, computer-supported collaborative learning, high-order cognitive thinking, and educational data mining.



CHANGQIN HUANG (Member, IEEE) received the Ph.D. degree in computer science and technology from Zhejiang University, China, in 2005. He completed two Postdoctoral Fellowship positions with the ECNU-TCL Joint Workstation on Educational Technology and Sun Yat-sen University on Computer Software and Theory. He had completed visiting research with the University of California at Irvine, Irvine, CA, USA, in 2011, and La Trobe University, Australia, in 2018. He is currently a Distinguished Professor with Zhejiang Normal University, China, and also the Director of the Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, China. He has published several articles in prestigious journals, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS. His research interests range from big data in education to machine learning and intelligent education. He serves as an Associate Editor for IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES.