

Multimodal Mutual Attention-Based Sentiment Analysis Framework Adapted to Complicated Contexts

Lijun He^{1b}, Ziqing Wang, Liejun Wang^{1b}, and Fan Li^{1b}, *Senior Member, IEEE*

Abstract—Sentiment analysis has broad application prospects in the field of social opinion mining. The openness and invisibility of the internet makes users' expression styles more diverse and thus results in the blooming of complicated contexts in which different unimodal data have inconsistent sentiment tendencies. However, most sentiment analysis algorithms only focus on designing multimodal fusion methods without preserving the individual semantics of each unimodal data. To avoid misunderstandings caused by ambiguity and sarcasm in complicated contexts, we propose a multimodal mutual attention-based sentiment analysis (MMSA) framework adapted to complicated contexts, which consists of three levels of subtasks to preserve the unimodal unique semantics and enhance the common semantics, to mine the association between unique semantics and common semantics and to balance decisions from unique and common semantics. In the framework, a multiperspective and hierarchical fusion (MHF) module is developed to fully fuse multimodal data, in which different modalities are mutually constrained and the fusion order is adjusted in the next step to enhance cross-modal complementarity. To balance the data, we calculate the loss by applying different weights to positive and negative samples. The experimental results on the CH-SIMS multimodal dataset show that our method outperforms existing multimodal sentiment analysis algorithms. The code of this work is available at <https://gitee.com/viviziqing/mmsacode>.

Index Terms—Sentiment analysis, multitask framework, unimodal unique semantics, multimodal fusion.

I. INTRODUCTION

SENTIMENT analysis technology is important in the fields of social public opinion mining, enterprise information

Manuscript received 2 April 2023; revised 27 April 2023; accepted 9 May 2023. Date of publication 15 May 2023; date of current version 7 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U1903213, in part by the Project funded by the China Postdoctoral Science Foundation under Grant 2021M692587, and in part by the Natural Science Foundation of Sichuan Province under Grant 2022NSFSC0966. This article was recommended by Associate Editor C. Chen. (Corresponding author: Lijun He.)

Lijun He is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Sichuan Digital Economy Industry Development Research Institute, Chengdu, Sichuan 610036, China (e-mail: lijunhe@mail.xjtu.edu.cn).

Ziqing Wang and Fan Li are with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: wangziqing@stu.xjtu.edu.cn; lifan@mail.xjtu.edu.cn).

Liejun Wang is with the College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China (e-mail: wljxju@xju.edu.cn).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TCSVT.2023.3276075>.

Digital Object Identifier 10.1109/TCSVT.2023.3276075



Fig. 1. Examples of sentiment analysis in complicated contexts. The ground-truth sentiments of the whole video clip, vision modality, text modality and audio modality are also given. The ground-truth sentiments of different modalities in the same video can vary.

analysis, and financial transactions. The aim of sentiment analysis is to predict the polarity of a person's opinion or behavior [1]. This polarity can be stated as positive, negative or neutral, and the intensity of the polarity is also considered. Previous sentiment models consisted of mostly text. With the development of social media, an increasing number of people share their viewpoints in multimodal forms, such as video, image and text, over the internet. The large amount of these multimodal data has laid a solid data foundation for sentiment analysis [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. On the other hand, increasingly more research focuses on multimodal sentiment analysis by reinforcement learning [12], [13] or deep learning. However, semantic inconsistencies among the data in different modalities may exist in some complicated contexts, such as implied [14], metaphorical and ironic contexts, as shown in Fig. 1. Distinguishing the real sentiment in these contexts needs to simultaneously keep the uniqueness of each modality and fully fuse the multimodal data. Most multimodal sentiment analysis methods focused on exploring the commonness of different modalities to obtain a joint multimodal representation [10], which disregarded the uniqueness of each modality. Thus, they were not suitable for application in complicated contexts. Therefore, an investigation of the uniqueness of each modality and commonality among multimodal data to adapt to sentiment analysis in complicated contexts is urgent and worthwhile.

Fusing the features of different modalities is a key problem in any multimodal task [15]. Current multimodal sentiment

analysis research is divided into three categories according to the selected multimodal fusion method: feature-level fusion [16], decision-level fusion [17] and hybrid fusion [18], [19]. Feature-level fusion methods usually perform simple concatenation [20], addition and weighted operations. These methods can achieve some improvement compared with unimodal sentiment analysis [21]. However, these methods only integrated multimodal data; thus, they were redundant because of repeated information among multimodal data. With the development of deep learning, researchers began to employ neural networks [23], ICON memory networks [24], MHMAN [25] memory attentive networks, Bi-Bimodal Fusion Network (BBFN) [26] and Multimodal Transformer (MulT) [51] based on Transformer, CubeMLP [27] based entirely on MLP, MMLGAN [28] local-global attention networks, visual aspect attention networks [29] and tensor-fusion methods [30] for multimodal data. The tensor-fusion method is highly complex due to the outer product operation. Attention-based fusion methods usually perform cross-modal fusion first, followed by multimodal fusion. These phased fusions not only increase the computational complexity but also make the multimodal data unable to pay attention to each other, leading to inadequate fusion. For the decision-level fusion methods, each unimodal data point was independently analyzed, and then the analyzed results were fused to determine the final sentiment result. Feature-level fusion methods and decision-level fusion methods focus on different stages of multimodal fusion, and each has its own advantages. The hybrid fusion methods inherited the advantages of feature-level fusion and decision-level fusion but increased the model complexity and implementation difficulty. On the basis of these studies, The authors in [31] and [32] have focused more on the specificity of multimodal sentiment analysis tasks or emotion recognition in conversation tasks. The authors in [32] mentions modality-level uncertainty and invariance and equivariance among modalities. MISA identifies modality-invariant and modality-specific representations of multimodal data and uses them to aid fusion for prediction. CH-SIMS [10] and AV-MC [11] focus on capturing the difference between two modalities. They propose a multitask learning framework to analyze the interaction between two modalities on datasets that have both multimodal annotations and independent unimodal annotations.

In general, the following problems need to be addressed: (1) Disregard for the uniqueness of each unimodal data. Unimodal data uniqueness complements multimodal common features and is a key clue for complex contextual sentiment analysis. (2) The existing multimodal fusion methods are usually unable to simultaneously interconstrain multimodal data. Thus, the fusion models are redundant, and the sentiment semantics may deviate. (3) Lack of interaction between unimodal tasks and multimodal tasks. The sentiment tendency of different unimodal data in complicated contexts may be inconsistent or even contradictory. Without an interaction study, it is difficult to coordinate sentiment tendencies among different tasks.

To address the above problems, we propose a multimodal mutual attention-based progressive sentiment analysis

algorithm adapted to complicated contexts. The contributions are presented as follows:

- **Progressive multitask sentiment analysis framework.** To essentially explore the uniqueness of each modality and commonality among multimodal data, we construct a three-level progressive multitask framework. At the first level, we formulate three parallel single modality subtasks and one multimodal fusion subtask, which are responsible for the independent sentiment representation of each unimodality and multimodality. At the second level, we design the feature-level decision fusion subtask to discover the common features among the four first-level subtasks. To jointly consider unique and common decisions, the decision fusion subtask is performed to realize multilevel comprehensive sentiment analysis. The loss value of each sample contains the prediction results of three unimodal subtasks in the first level and the decision fusion subtask in the third level. The loss value of a batch of samples is weighted by the positive and negative samples.
- **Multiperspective and hierarchical fusion module.** To promote multimodal information interaction, a multiperspective and hierarchical fusion module is carried out based on our established closed-loop mutual attention structure. The fusion module has two layers. At the first layer, the mutual attention is dominated by text information and constrained by audio, vision and text information. In the following layer, we adjust the fusion order to obtain the multiperspective and multimodal fusion features. In this module, fusion of any two modalities will be constrained by the third modality, which can avoid the fusion feature deviation caused by semantic polarization of some modality. The fusion feature generated by the module is the integration and symbiosis of three modalities.
- **Dual decision fusion strategy adapted to complex contexts.** In view of the semantic inconsistency among the different modalities in complicated contexts, we design a dual decision fusion strategy to explore the interaction between unimodal tasks and multimodal tasks from two aspects. First, we fuse the higher semantic features of each unimodal and multimodal task. Then, we design a weighted learning network to balance the influence of different decisions of multiple subtasks.

The remainder of this article is organized as follows: Section II reviews previous related work. Section III presents the details of our methods. Section IV gives the experimental results. Section V concludes the paper.

II. RELATED WORK

Primary unimodal sentiment analysis methods such as [21] employed textual information to analyze sentiment tendency. However, it is difficult to achieve accurate sentiment analysis with only limited textual semantic information. Fortunately, with the development of social media and mobile terminals, a large amount of multimodal data emerged over the internet, which laid a solid data foundation for multimodal sentiment

analysis [2], [3], [4], [5], [6], [7], [8], [9], [10]. Integrating speech and facial expression for audio-visual emotion recognition [22] has attracted many researchers. Multimodal fusion methods are the core of multimodal sentiment analysis, which is usually divided into three categories according to the selected multimodal fusion method: feature-level fusion, decision-level fusion and hybrid fusion. We classify related work according to the specific operations employed during feature-level fusion, such as tensor computation, neural networks, and attention mechanisms.

A. Tensor Computation-Based Fusion Method

To explore cross-modality interactions, an increasing number of researchers have focused on designing new tensor-based fusion methods to learn the information interaction among different modalities. A regularization method based on tensor rank minimization was proposed to learn intermodality interactions [33]. The authors in [30] proposed the Tensor Fusion Network, which can provide a new tensor representation by calculating the outer product among multimodal features. To overcome its high complexity, the authors in [34] improved the fusion efficiency by decomposing the weights of high-dimensional tensors. Moreover, the work in [35] employed the low-rank multimodal fusion method to decompose the weight tensor to more effectively reduce the computational complexity. These tensor-based fusion methods enhance multimodal interaction to render fusion features more expressive. However, it is difficult to remove the introduced large amount of redundancy information. Inspired by the abovementioned tensor-based fusion works, some studies began to fuse the local tensor and global tensor [36]. Similar to [36], Hou et al. [37] recursively integrated the local relations into global relations via multilinear fusion, and Mai et al. [38] carried out the time correlation processing of the LSTM network for the results of tensor fusion. Although essential cross-modality interactions can be achieved, these methods cannot be directly implemented in practice due to the large amount of redundant information.

B. Neural Network-Based Fusion Method

Neural networks can automatically learn multimodal interactions for efficient multimodal sentiment analysis. Hazarika et al. [39] designed a two-layer gate recurrent unit (GRU) to store the speaker information and context information of the current multimodal content. Then, they added a global gating GRU to store the global information of the current multimodal content in [24]. Poria et al. [40] proposed a convolution multicore learning model for feature fusion to analyze sentiment polarity. Based on the idea of translation, some robust and learning multimodal joint representation methods were proposed in [41]. In addition, the multimodal decomposition model (MFM) [42] and multimodal baseline (MMB) [43] used the factor decomposition method to learn multimodal embedding. Bi-MHG [44] models the relevance among three modalities while allowing missing modalities. WSCNet [45] is a weakly supervised framework that combines localized information with global representation. The aim of

these methods was mostly to map the unimodal features of different subspaces to a unified semantic space. CubeMLP [27] is an MLP-based multimodal sentiment analysis framework for multimodal feature processing that consists of three independent MLP units, each of which has two affine transformations to mix all modality features across three axes. These methods can simultaneously concentrate on global and local information to improve the efficiency of multimodal data fusion without an attention mechanism.

C. Attention Mechanism-Based Fusion Method

Recently, with the advantages of the attention mechanism, many multimodal sentiment analyses have tried to employ it to highlight key sentiment features. In [29], image data was taken by the attention mechanism to help lock important text information. Zadeh et al. designed a memory fusion network with a special attention mechanism in [46]. To make the attention mechanism more effective, Poria et al. [47] proposed an LSTM-based multilayer attention mechanism to distribute higher weights to more informative modalities. Similarly, Chen et al. [48] employed a multimodal attention mechanism and Bi-LSTM selective learning to fuse complementary information from audio and text. Ghosal et al. [49] proposed an attention model based on an RNN to fuse the attention vector among multimodal pairs. To reduce the mutual interference of multimodal information and to enhance the complementarity of information, the authors in [50] adopted Bi-GRU in the intramodal representation and developed a cross-modal interactive gating mechanism. Tsai et al. [51] proposed a transformer-based multimodal fusion algorithm to map one modal to the other modal feature space using the attention between the two modal features. The algorithm has 6 cross-modal attention branches, so the model complexity is relatively high. The Bi-Bimodal Fusion Network (BBFN) [26] is aimed at leveraging dynamics of independence and correlation between two modalities. The fusion scheme of the BBFN consists of two transformer-based bimodal learning modules. The BBFN performs fusion and separation on pairwise modality representations to achieve excellent sentiment analysis performance. In summary, feature fusion methods based on attention mechanisms can selectively learn key features and improve the efficiency of feature fusion. However, current research on multimodal sentiment analysis has failed to use multimodal information as a whole for multimodal mutual attention constraints and to highlight the important sentiment clues, which limited the sentiment analysis accuracy, especially in complicated context scenarios.

III. PROPOSED METHOD

As shown in Fig. 2, MMSA is divided into two stages. At the first stage, we propose a data preprocessing module to remove the redundancy of original textual raw features from CH-SIMS [10] and CMU-MOSEI [7]. At the second stage, we design a Progressive Multitask Sentiment Analysis Framework to promote the collaboration and interaction among unimodal data. Details of the algorithm are presented as follows:

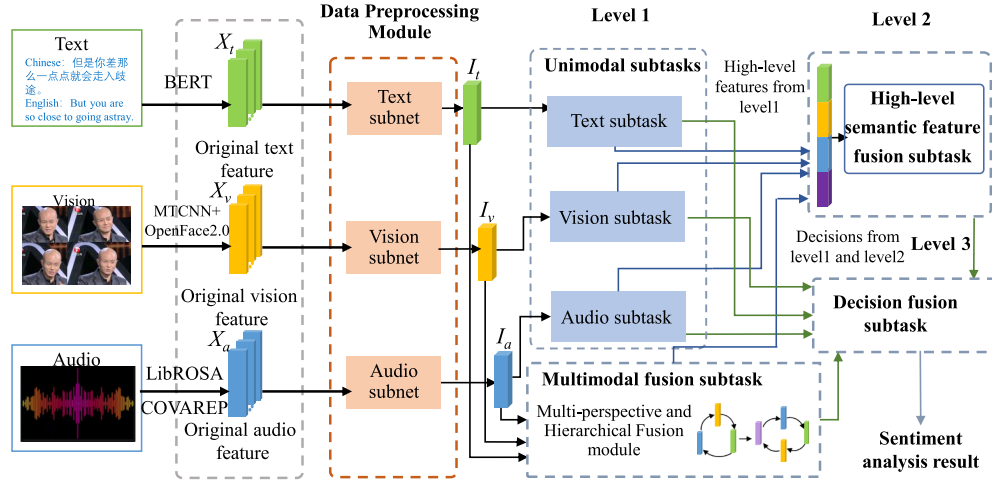


Fig. 2. Multimodal mutual attention-based progressive multitask sentiment analysis framework. The proposed framework. The figure shows the sentiment prediction process of one video clip. First, we separately process text, visual, and audio raw features through the Data Processing Module. Second, we separately input them into three unimodal subtasks and the multimodal fusion subtask at Level 1 for sentiment analysis. At Level 2, we fuse high-level sentiment features from subtasks at Level 1. Last, we balance the decisions of the above two levels to obtain the final sentiment analysis results.

A. Progressive Multitask Sentiment Analysis Framework

As shown in Fig. 2, we remove the redundancy of original textual raw features \mathbf{X}_t , visual raw features \mathbf{X}_v and audio raw features \mathbf{X}_a in the data preprocessing stage. These raw features were extracted by BERT [52], LibROSA [53] or COVAREP [54], MTCNN [55] and OpenFace2.0 [56]. Due to the redundancy and distribution of different modal features, we apply LSTM to learn the long-term dependency of \mathbf{X}_t and obtain textual compact semantic features $\mathbf{I}_t \in \mathbb{R}^{d_t}$. We separately input \mathbf{X}_a and \mathbf{X}_v into three FC layers activated by \tanh and obtain audio compact semantic features $\mathbf{I}_a \in \mathbb{R}^{d_a}$ and visual compact semantic features $\mathbf{I}_v \in \mathbb{R}^{d_v}$. In the second stage, we design a Progressive Multitask Sentiment Analysis Framework to promote the collaboration and interaction among \mathbf{I}_t , \mathbf{I}_v and \mathbf{I}_a . Details of the framework are presented as follows:

1) *Level 1 Preservation of Complete Unimodal Unique Semantics and Mining of Common Multimodal Semantics:* In complex contexts, there are differences in the sentiment semantics expressed by each individual unimodal data even in one video clip. Therefore, it is important to preserve unimodal unique semantics and further mine multimodal common semantics for analysis. Therefore, we apply four parallel subtasks: three unimodal subtasks and one multimodal subtask. Because unimodal compact features \mathbf{I}_t , \mathbf{I}_v and \mathbf{I}_a have similar data distributions and redundancies, unimodal subtasks are similar. We use text subtasks as an example to present the network structures of unimodal subtasks. In the text subtask, we input \mathbf{X}_t into two FC layers to obtain the text high-level feature $\mathbf{h}_t \in \mathbb{R}^{d_{ht}}$ as follows:

$$\mathbf{h}_t = \tanh(\mathbf{W}_{t2} \cdot \tanh(\mathbf{W}_{t1} \cdot \mathbf{I}_t + \mathbf{b}_{t1}) + \mathbf{b}_{t2}) \quad (1)$$

where d_{ht} represents the dimension of text high-level semantic feature \mathbf{h}_t , and \mathbf{h}_t are input to the third FC layer to predict the text sentiment analysis results \hat{P}_t as follows:

$$\hat{P}_t = \mathbf{W}_{t3} \cdot \mathbf{h}_t + \mathbf{b}_{t3} \quad (2)$$

where \mathbf{W}_{tu} and \mathbf{b}_{tu} denote the weights and biases, respectively, of the u^{th} FC layer. We use the same method to obtain visual

high-level features $\mathbf{h}_v \in \mathbb{R}^{d_{hv}}$, audio high-level features $\mathbf{h}_a \in \mathbb{R}^{d_{ha}}$, visual sentiment analysis results \hat{P}_v and audio sentiment analysis results \hat{P}_a .

We input \mathbf{I}_t , \mathbf{I}_a and \mathbf{I}_v into the multimodal fusion subtask. In the multimodal fusion subtask, we design multiperspective and hierarchical fusion modules to fuse multimodal data, the details of which will be introduced in Section B.

2) *Level 2 Exploration of the Association Between Unique Semantics and Common Semantics:* Subtasks of the first level can preserve the unimodal unique semantics. However, the multimodal fusion subtask pursues a common sentiment expression, which leads to the bluntness of unimodal unique semantics in the fusion process. In addition, the whole framework lacks information interaction between the unique semantic features at the first level and the common semantic features at the first level. To enrich the sentiment semantics of multimodal data and take into account both unique semantics and common semantics, we design the second-level subtask to predict the sentiment results based on the high-level semantic features extracted from each subtask at the first level. We present the second level in Section C.

3) *Level 3 Balancing of the Diversity Among Decisions:* The sentiment feature and decisions of each subtask at the first two levels of the multitask sentiment analysis framework, which are critical to the final sentiment prediction, are different. To fully balance the decisions of subtasks at the first two levels, as shown in Section D, we design a decision fusion subtask at the third level to balance each sentiment decision and fully utilize the mutual constraints and facilitation relationships between two different subtasks. The third-level subtask obtains the final sentiment prediction, which is more consistent with the real sentiment tendency.

B. Multimodal Fusion Subtask

The fusion module is the core of the multimodal fusion subtask. We design multiperspective and hierarchical fusion modules to perform mutual attention fusion.

As shown in Fig. 3, MHF consists of two steps: (1) a multiperspective and hierarchical fusion step and (2) a fused

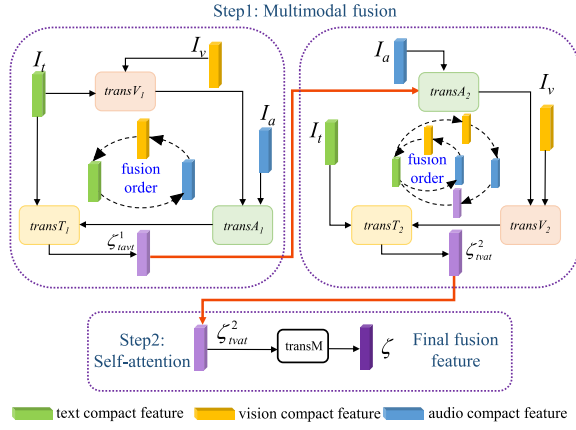


Fig. 3. Multiperspective and hierarchical fusion module. The figure shows two steps of the fusion module. In the first step, we design two layers. There are three cross-modal attention mechanism fusion networks guided by different modalities connected to fuse three modalities in a closed loop in each layer. In the second step, we design self-attention on the fusion feature.

feature self-attention step. In the multimodal fusion step, there are two layers of closed-loop mutual attention structure to fuse unimodal compact features I_t , I_v and I_a . Within each layer, there are cross-modal attention mechanism fusion networks $transV(\cdot)$, $transA(\cdot)$ and $transT(\cdot)$ guided by sequentially connected vision, audio, and text. In the first layer, we use visual features to fuse text features by an attention mechanism and then fuse audio features and visual-text fused features. We fuse text features and audio-visual-text fused features to obtain text-audio-visual-text fused features. Three modalities pay attention to each other in each layer, so it is a closed-loop fusion structure. Between layers, we adjust the fusion order to compensate for the semantic correlation differences between two modalities to obtain multiperspective multimodal fusion features. A detailed description is presented as follows:

1) *Cross-Modal Attention Mechanism Fusion Network*: Each layer of the closed-loop mutual attention structure consists of three different modality-guided, cross-modal attention mechanism fusion networks.

We let $transV(\cdot)$, $transA(\cdot)$ and $transT(\cdot)$ denote the vision-, audio-, and text-guided cross-modal attention mechanism fusion network, respectively. The number of neurons and weights among these networks are different, but their network structures are similar.

Using the vision-guided cross-modal attention mechanism fusion network in the first layer $transV_1(\cdot)$ as an example, as shown in Fig. 4, its input consists of vision compact features I_v and text compact features I_t . To ensure the diversity of compact semantic features, we set three FC layers to map I_v to the visual query vector $\mathbf{Q} \in \mathbb{R}^{d_{vq}}$ and I_t to the text key vector $\mathbf{K} \in \mathbb{R}^{d_{tk}}$ and the text value vector $\mathbf{V} \in \mathbb{R}^{d_{tv}}$ ($d_{tk} = d_{tv}$), respectively, as follows:

$$\mathbf{Q} = \text{BN}(\mathbf{W}_{vq} \cdot \mathbf{I}_v + \mathbf{b}_{vq}) \quad (3)$$

$$\mathbf{K} = \text{BN}(\mathbf{W}_{tk} \cdot \mathbf{I}_t + \mathbf{b}_{tk}) \quad (4)$$

$$\mathbf{V} = \text{BN}(\mathbf{W}_{tv} \cdot \mathbf{I}_t + \mathbf{b}_{tv}) \quad (5)$$

where d_{vq} , d_{tk} and d_{tv} represent the dimensions of the corresponding features; \mathbf{W}_{vq} and \mathbf{b}_{vq} denote the weights and biases, respectively, of the FC layer that maps vision compact

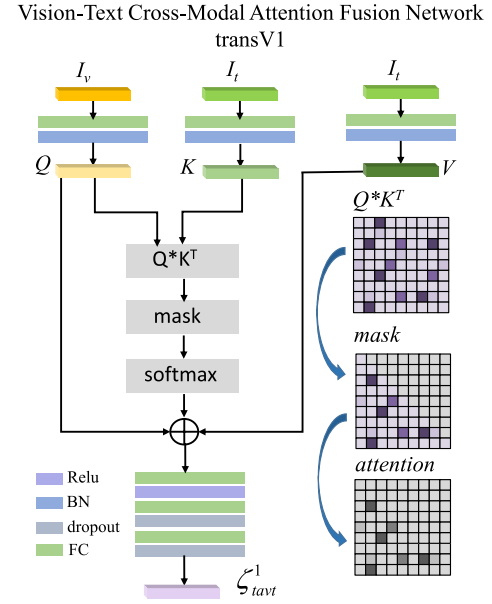


Fig. 4. Cross-Modal Attention Mechanism Fusion Network $transV_1$. The figure shows the process of fusing text and vision modalities. First, we map them to the corresponding vector. Second, we construct a vision-text attention parameter matrix. Third, we weight the text value vector by the attention parameter matrix. Last, we obtain the vision-text fusion feature from the FC layers.

feature I_v to \mathbf{Q} ; and \mathbf{W}_{tk} and \mathbf{b}_{tk} , and \mathbf{W}_{tv} and \mathbf{b}_{tv} denote another two FC layers' weights and biases, respectively.

Next, the correlation matrix $\pi \in \mathbb{R}^{d_{vq} \times d_{tk}}$ between vision and text is calculated by \mathbf{Q} and \mathbf{K} as follows:

$$\pi = \mathbf{Q} \cdot \mathbf{K}^T \quad (6)$$

There are differences in the importance of features in different dimensions of vision and text. For example, features that directly describe crying or laughing have a great impact on the basic sentiment tendency. Focusing on important sentiment features and using relatively unimportant features for fine-tuning can help the network constantly approach the correct sentiment tendency. To make the key features more involved in the fusion feature calculation, we use the mask matrix to mask the attention parameters above by the diagonal.

As shown in Fig. 5, we design the $\mathbf{mask} \in \mathbb{R}^{d_{vq} \times d_{tk}}$ matrix to mask the parameters above the diagonal of the correlation matrix. Then, we obtain the attention matrix $\pi' \in \mathbb{R}^{d_{vq} \times d_{tk}}$ by using $\text{softmax}(\cdot)$ to map the parameters into the range of (0, 1), as shown by Equations (11) and (12):

$$\mathbf{mask} = \begin{bmatrix} 0 & -\infty & \cdots & -\infty \\ 0 & 0 & \cdots & -\infty \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad (7)$$

$$\pi' = \text{softmax}(\pi + \mathbf{mask}) \quad (8)$$

The attention matrix π' is dotted with the text value vector \mathbf{V} to obtain the vision-text, cross-modal fusion feature $\mathbf{z}^1_{vt} \in \mathbb{R}^{d_{vq}}$

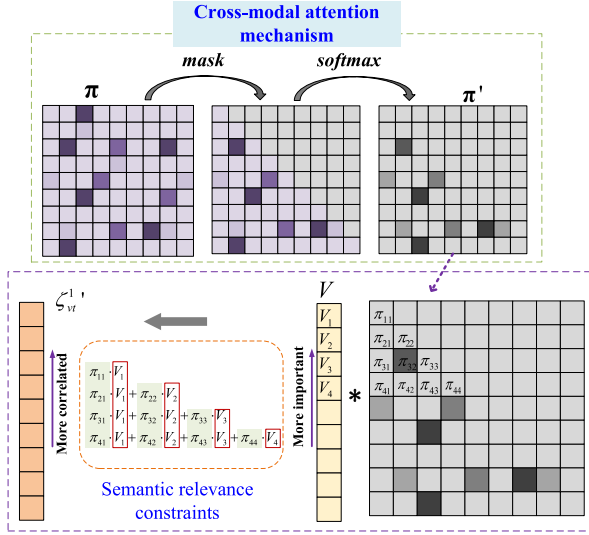


Fig. 5. Cross-Modal Attention Mechanism. In this figure, we show the process of masking the attention parameters matrix in the upper part. Then, in the lower part, the figure shows the process of weighting the value vector by the vision-text attention parameter matrix.

by embedding two FC layers activated by $ReLU(\cdot)$:

$$\zeta_{vt}^{1'} = \pi' \cdot \mathbf{V} \quad (9)$$

$$\zeta_{vt}^1 = ReLU(\mathbf{W}_{vt2} \cdot ReLU(\mathbf{W}_{vt1} \cdot \text{BN}(\zeta_{vt}^{1'}) + \mathbf{b}_{vt1}) + \mathbf{b}_{vt2}) \quad (10)$$

where \mathbf{W}_{vtu} and \mathbf{b}_{vtu} denote the weights and biases, respectively, of the u_{th} FC layer.

As shown in Fig. 5, for the text value vector $\mathbf{V} = [V_1, V_2, \dots, V_{d_{tv}}]^T$, the first dimensional feature V_1 participates in calculating all departments' values of $\zeta_{vt}^{1'}$, which plays the most important role in the fusion process. V_2 is involved in the calculation of ζ_{vt}^1 from its second dimension to the last dimension. Thus, after training, the distribution of \mathbf{V} is sorted by decreasing sentiment semantic importance from its first dimension to the last dimension. The visual query vector \mathbf{Q} computes fusion features such as the text value vector in the form of attention parameters, which also has a similar distribution to the text vector. Therefore, our proposed cross-modal attention mechanism fusion network can not only strengthen the important features of vision and text in two ways weighting and the number of participations in the fusion feature calculation, but also ensure that the importance distribution of the semantic features of vision is similar to that of text features. The semantic relevance between two modalities can be strengthened, and the semantic distribution within modality is more continuous. Thus, the weight values in the obtained attention matrix are more representative of the correlation between the features of different dimensions across the two modalities.

2) *Closed-Loop Mutual Attention Structure*: Since the current text processing methods are relatively mature and text features usually contain the richest information, the closed-loop mutual attention structure at each layer is dominated by text and jointly constrained by three modalities for multimodal fusion. The fusion orders at the two layers are different. The fusion sequence of the first layer is designed as $transV_1(\cdot) \sim$

$transA_1(\cdot) \sim transT_1(\cdot)$. $transV_1(\cdot)$ uses visual features to focus important text features and obtain the vision-text, cross-modal fusion feature ζ_{vt}^1 :

$$\zeta_{vt}^1 = transV_1(\mathbf{I}_v, \mathbf{I}_t) \quad (11)$$

Then, $transA_1(\cdot)$ uses the audio information to constrain ζ_{vt}^1 to obtain the audio-vision-text fusion feature $\zeta_{avt}^1 \in \mathbb{R}^{d_{aq}}$.

$$\zeta_{avt}^1 = transV_1(\mathbf{I}_a, \zeta_{vt}^1) \quad (12)$$

$transT_1(\cdot)$ constrains ζ_{avt}^1 by using text information. In this process, three modalities are mutually constrained to obtain fusion feature $\zeta_{tavt}^1 \in \mathbb{R}^{d_{tq}}$ in text-to-text closed-loop order:

$$\zeta_{tavt}^1 = transT_1(\mathbf{I}_t, \zeta_{avt}^1) \quad (13)$$

To generate fusion features with different complementary relationships of multimodal features, we adjust the fusion order at the second layer and design it as $transA_2(\cdot) \sim transV_2(\cdot) \sim transT_2(\cdot)$. Note that the inputs of $transA_2(\cdot)$ are \mathbf{I}_a and ζ_{tavt}^1 , while the inputs of the other networks are similar to those of the first layer. We obtain $\zeta_{tvat}^2 \in \mathbb{R}^{d_{tq}}$. ζ_{tvat}^2 is fed into the multimodal self-attention mechanism fusion network denoted by $transM(\cdot)$, which performs the self-attention on ζ_{tvat}^2 at the second step as follows:

$$\zeta = transM(\zeta_{tvat}^2, \zeta_{tvat}^2) \quad (14)$$

We employ two FC layers to obtain the multimodal fusion high-level feature $\mathbf{h}_m \in \mathbb{R}^{d_{hm}}$ and then to predict the multimodal sentiment category \hat{P}_m based on \mathbf{h}_m :

$$\mathbf{h}_m = \tanh(\mathbf{W}_{m2} \cdot \tanh(\mathbf{W}_{m1} \cdot \text{BN}(\zeta) + \mathbf{b}_{m1}) + \mathbf{b}_{m2}) \quad (15)$$

$$\hat{P}_m = \mathbf{W}_{m3} \cdot \mathbf{h}_m + \mathbf{b}_{m3} \quad (16)$$

where \mathbf{W}_{mu} and \mathbf{b}_{mu} denote the weights and biases, respectively, of the u_{th} FC layer.

3) *Innovation Summary of Fusion Module*: The proposed mutual attention fusion has two fusion layers. In different fusion layers, we also adjust the fusion order to obtain the multi-perspective multimodal fusion features. In our module, cross-modal fusion of any two modalities will be constrained by the third modality, which can avoid the fusion feature deviation caused by semantic polarization of some modality. In our proposed cross-modal fusion methods, we use the mask matrix to mask the attention parameters above by the diagonal. In this way, the semantic distribution of unimodal features is more relevant. In addition, more important features are more involved in the cross-fusion features computation. Thus, the weight values in the obtained attention matrix are better for cross-modal fusion.

The multiperspective and hierarchical fusion module is not limited to text, audio and vision modalities and has a strong scalability to fuse more modalities.

C. High-Level Semantic Feature Fusion Subtask

Designing independent subtasks at the first level can help preserve unique unimodal semantics and obtain common multimodal semantics. However, these independent subtasks lack essential interaction, which results in insufficient complementarity between unimodal unique semantics and common

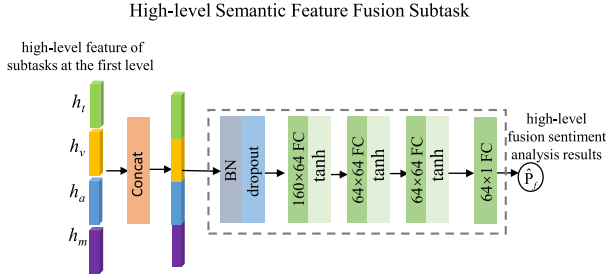


Fig. 6. Network of high-level semantic feature fusion subtask. In the left part of the figure, we concatenate high-level semantic features extracted from subtasks at the first level. Then, we input the fusion feature to FC layers, as shown in the right part of the figure.

semantics. To utilize the uniqueness of the unimodal semantic features and enrich the expressiveness of multimodal data, as shown in Fig. 6, we develop a high-level semantic feature fusion subtask at the second level, whose inputs are the high-level semantic features of each subtask at the first level denoted by $\mathbf{h}_t, \mathbf{h}_v \in \mathbb{R}^{d_{hv}}, \mathbf{h}_a \in \mathbb{R}^{d_{ha}}$ and \mathbf{h}_m .

The high-level semantic features extracted from the unimodal subtask can represent the uniqueness of each unimodal data. On the other hand, high-level semantic features of multimodal fusion subtasks can reflect the commonness of different modalities. As shown in Fig. 6, high-level semantic features are concatenated to enhance the interaction and association between unimodal uniqueness and multimodal commonness.

$$\mathbf{h}_f = \text{Concat}(\mathbf{h}_t, \mathbf{h}_v, \mathbf{h}_a, \mathbf{h}_m) \quad (17)$$

Then, we add four FC layers to predict the sentiment category of the high-level semantic feature fusion subtask \hat{P}_f based on the concatenated feature \mathbf{h}_f , as follows:

$$\begin{aligned} \hat{P}_f = & \mathbf{W}_{f4} \cdot \tanh(\mathbf{W}_{f3} \cdot \tanh(\mathbf{W}_{f2} \cdot \tanh(\mathbf{W}_{f1} \cdot \text{BN}(\mathbf{h}_f) \\ & + \mathbf{b}_{f1}) + \mathbf{b}_{f2}) + \mathbf{b}_{f3}) + \mathbf{b}_{f4}) \end{aligned} \quad (18)$$

where \mathbf{W}_{fu} and \mathbf{b}_{fu} denote the weights and biases, respectively, of the u_{th} FC layer.

D. Decision Fusion Subtask

In complicated context, the decisions $\hat{P}_t, \hat{P}_v, \hat{P}_a, \hat{P}_m$, and \hat{P}_f of the above subtasks may be inconsistent. We propose two strategies for decision fusion subtasks to balance the diversity among different decisions and acquire the final sentiment prediction result \hat{P}_d from a global perspective at the third level. One of the strategies named MMSA-W is to set automatically learnable weights to balance the influence of decisions across subtasks. Let w_t, w_a, w_v, w_m and w_f denote the weights of text, audio, vision, multimodal fusion, and high-level feature fusion subtask decisions, respectively. During the training process, the model automatically learns the optimal weights to reasonably balance the decisions for each subtask and obtain the final predicted result \hat{P}_d as follows:

$$\hat{P}_d = \mathbf{W} \cdot \hat{\mathbf{P}} \quad (19)$$

where weight vectors $\mathbf{W} = [w_t, w_a, w_v, w_m, w_f]$ and decision vectors $\hat{\mathbf{P}} = [\hat{P}_t, \hat{P}_v, \hat{P}_a, \hat{P}_m, \hat{P}_f]^T$. Another strategy, MMSA-L, uses an FC layer with weights \mathbf{W}_d and bias \mathbf{b}_d to map

the decision vectors to the final sentiment analysis result as follows:

$$\hat{P}_d = \mathbf{W}_d \cdot \hat{\mathbf{P}} + \mathbf{b}_d \quad (20)$$

E. Loss Function

To constrain the preservation of the unique semantics of the unimodal subtasks and the accuracy of the final sentiment prediction results, we set the objective function as follows:

$$\text{Loss} = \min \frac{1}{N} \sum_{n=1}^N \sum_i \alpha_i L(P_i^n, \hat{P}_i^n) + \sum_j \beta_r \|W_r\|^2 \quad (21)$$

where N is the number of training samples, $i \in \{d, t, a, v\}$, and $r \in \{t, a, v\}$. $L(P_i^n, \hat{P}_i^n)$ is the training loss between the prediction result \hat{P}_i^n and label P_i^n of the n_{th} sample in subtask i . α_i is the hyperparameter to balance these four results. With the exception of the training losses in different subtasks, we sparse the sharing parameters via the $L2$ norm in the last item, in which W_r is the sharing parameter and β_r represents the step of weight decay of subnet r in the Data Preprocessing Module.

The number of negative samples in the CH-SIMS dataset is approximately 1.8 times the number of positive samples. Imbalanced positive samples and negative samples have different impacts on computing loss and lead to poor training results. To balance the impacts of positive and negative samples, we give different weights to the positive and negative loss values. We name the model trained by this weighted loss Loss_B MMSA-B. The Decision Fusion Subtask of MMSA-B is the same as MMSA-W. The difference between them is the loss function. The weighted loss of MMSA-B is:

$$\text{Loss}_p = W_p \cdot \sum_{n=1}^{N_p} \sum_i \alpha_i L(P_i^n, \hat{P}_i^n) \quad (22)$$

$$\text{Loss}_n = W_n \cdot \sum_{n=1}^{N_n} \sum_i \alpha_i L(P_i^n, \hat{P}_i^n) \quad (23)$$

$$\text{Loss}_B = \min \frac{1}{N} (\text{Loss}_p + \text{Loss}_n) + \sum_j \beta_r \|W_r\|^2 \quad (24)$$

where the loss of positive samples Loss_p and the loss of negative samples Loss_n are calculated by Equation (20) and Equation (21).

IV. EXPERIMENT

A. Dataset

CH-SIMS [10] is a Chinese multimodal sentiment analysis dataset that consists of 2281 video clips from different movies, TV series and variety shows. The time length of each video clip is within $[1\text{ s}, 10\text{ s}]$. Each video clip sample contains three modalities: text, audio and vision. Different from other existing datasets with only the whole sentiment label, each modality in it owns its individual label, which is fine-grained into several levels, including strongly negative (-1), negative (-0.8), weakly negative ($-0.2, -0.4, -0.6$), neutral (0), weakly positive ($0.2, 0.4, 0.6$), positive (0.8), and

strongly positive (1). CMU-MOSEI [7] is a popular dataset for multimodal sentiment analysis, which is enlarged from CMU-MOSI [4]. The dataset contains 23453 video segments extracted from 5000 videos. The dataset is annotated with sentiment tendencies in the interval $[-3, 3]$. Each sample only has a multimodal annotation.

B. Baselines

We compare the performance of our proposed MMSA with the following multimodal sentiment analysis models: (1) EF-LSTM [57]: The early fusion LSTM method splices the initial features of three modalities and then captures the sequence's long-distance dependencies among the spliced features by using LSTM. (2) LF-DNN [58]: The later fusion separately DNN learns each unimodal representation and then concatenates the multimodal features to learn multimodal fusion before classification. (3) MFN [46]: The memory fusion network takes into account view-specific interactions and cross-view interactions in a neural architecture and continuously models them over time. (4) TFN [30]: The Tensor Fusion Network explicitly models view-specific and cross-view dynamics by creating a multidimensional tensor based on the outer product, which can capture unimodal, bimodal and three modal interactions across three modalities. (5) LMF [35]: The low-rank multimodal fusion model improves the TFN model. It learns modality-specific and cross-modal interactions by performing a low-rank multimodal tensor fusion technique to improve efficiency. (6) MULT [51]: The heart of the multimodal transformer model is based on directional pairwise cross-modal attention. This model uses pairwise cross-modal attention to promote interactions between two multimodal sequences across distinct time steps and to latently adapt streams from one modality to another modality. (7) SELF-MM [59]: The Self-Supervised Multi-Task Multimodal sentiment analysis network is based on the self-supervised learning strategy to acquire unimodal representations by jointly learning one multimodal task and three unimodal subtasks. Different from the multimodal task, the labels of unimodal subtasks are autogenerated in the self-supervised method. (8) BIMHA [60]: The bimodal information-augmented multihead attention network is inspired by the observation that the interactions between any two modalities are different. This network obtains the interaction between two modalities by tensor fusion and enhances the weighted bimodal interaction by its proposed multihead attention mechanism. (9)BBFN [26]: The bimodal fusion network performs fusion and separation on pairwise modality representations with two transformer-based bimodal learning modules to analyze the dynamics of independence and correlation between two modalities. (10) AV-MC [11]: The Acoustic Visual Mixup Consistent framework uses unimodal annotations and unsupervised data in CH-SIMS v2.0 to learn different nonverbal contexts for sentiment analysis.

C. Experimental Details

The training is conducted by NVIDIA GeForce GTX 2080 TI with learning rates of 0.001. The Adam optimizer is employed with a batch size of 64. The number of

output channels of subnets in the Data Preprocessing Module are set to 64, 16 and 64 for the vision, audio and text subnets, respectively. The input dimensions of the high-level semantic feature fusion subtask are set to 64, 16, 64 and 64 for vision, audio, text and fusion high-level features, respectively. When we train MMSA on the CH-SIMS dataset, we record the experimental results in terms of multiclass classification and regression. For multiclass classification, we employ 2-class accuracy (ACC2), 3-class accuracy (ACC3), 5-class accuracy (ACC5) and weighted F1 score (F1). For regression, we report the mean absolute error (MAE) and Pearson correlation (CORR). Higher values indicate better performance for all metrics, with the exception of MAE. When we train MMSA on CMU-MOSEI, we record 2-class accuracy (ACC2) and F1-score (F1) for the classification tasks and record MAE and Pearson correlation (CORR) as measurements.

D. Experimental Results on CH-SIMS Dataset

Table I shows the performance on sentiment analysis accuracy and network parameters on the CH-SIMS dataset among our proposed MMSA (MMSA-L, MMSA-W and MMSA-B) and other existing methods, including single-task-based and multitask-based sentiment analysis algorithms. The difference between MMSA-L and MMSA-W is the different methods of decision fusion subtask. The loss function employed for the training process of MMSA-L and MMSA-W is *Loss*, while the loss function used for training MMSA-B is *Loss_B*.

As expected, compared with single-task sentiment analysis methods, most multitask sentiment analysis methods can improve the results by sharing the knowledge contained in different tasks. Poor performance of SELF-MM is attributed to the notion that SELF-MM is based on the self-supervised learning strategy, in which the unimodal labels were generated by its proposed model. Thus, it is more suitable for practical implementation and has very low dataset requirements. We also observe that our proposed method MMSA can outperform other multitask methods by considering the coordination of both uniqueness and commonness from multimodal data and multimodal mutual interactions. ACC2, ACC5, F1 and MAE of MMSA-B can be improved by 1.53%, 2.05%, 1.33% and 0.34%, respectively, over other methods. The results of MMSA-B are better than those of MMSA-L and MMSA-W because it weights the losses of samples to balance the different effects of positive and negative samples on training. This finding shows that our data balancing approach has improved the results. Particularly, compared with BIMHA, the number of our network parameters is 6-7 times smaller, which is even lower than in most of the single-task algorithms. Notably, the network with the lowest complexity is EF-LSTM because it directly splices the three modalities from the lack of intermodal interaction information. From the above analysis, we conclude that our developed fusion module fully fuses multimodal data by providing more possibilities of fusion depth and fusion perspective with fewer parameters. Moreover, the designed multitask framework based on it can promote the interaction of unimodal unique semantics and multimodal common semantics to obtain richer sentiment semantics and

TABLE I

PERFORMANCE COMPARISON OF SENTIMENT ANALYSIS ON THE CH-SMIS DATASET. THE BEST RESULTS ARE SHOWN IN **RED**. THE SECOND BEST RESULTS ARE SHOWN IN **BLUE**. THE THIRD BEST RESULTS ARE SHOWN IN **GREEN**

Model		ACC2	ACC3	ACC5	F1	MAE	CORR	PARAS
Single-task	EF-LSTM	69.37	54.27	21.23	56.82	58.98	2.12	215K
	GRAPH-MFN	79.65	67.4	39.17	80.4	44.7	58.14	1.1M
	TFN	80.31	64.33	36.76	80.66	45.14	58.06	35M
	LMF	78.99	66.74	37.86	78.99	44.19	57.42	1M
	MFN	79.21	66.08	39.39	79.15	43.44	58.12	601K
	LF-DNN	78.99	64.99	41.36	79.72	41.9	58.94	635K
	MULT	79.65	65.86	38.95	79.94	43.87	58.21	1.5M
Multi-task	MLMF	80.09	66.96	39.61	80.63	42.4	59.68	1.4M
	MTFN	81.62	67.4	38.73	81.39	39.82	67.27	140M
	MLF-DNN	81.36	69.08	40.13	81.68	41.02	63.55	283K
	SELF-MM	79.21	65.86	41.14	78.53	45.11	58.21	102M
	BIMHA	82.71	69.23	45.21	82.72	38.5	66	2.5M
	BBFN	83.15	71.99	42.67	83.66	45.14	65.63	\
	AV-MC	82.06	69.58	43.33	81.90	38.34	69.94	\
	MMSA-L	82.74	70.46	44.2	82.31	38.13	68.23	365K
	MMSA-W	83.15	71.12	42.01	83.17	39.13	67.34	365K
	MMSA-B	84.68	71.33	47.26	84.99	38.0	68.64	365K

TABLE II

PERFORMANCE COMPARISON OF SENTIMENT ANALYSIS ON THE CMU-MOSEI DATASET

Model	ACC2	F1	MAE	CORR
TFN	-/82.5	-/82.1	59.3	70.0
LMF	-/82.0	-/82.1	62.3	67.7
MFN	76.0/-	76.0/-	-	-
MULT	-/82.5	-/82.3	58	70.3
SELF-MM	82.81/ 85.17	82.53/ 85.30	53.0	76.5
MMSA-B	82.89 /84.11	83.06 /84.08	57.1	72.5

analyze these differences among different subtasks. In summary, the experimental results suggest that exploring the uniqueness of each modality and commonness among the multimodal data on the basis of multiperspective and hierarchical multimodal fusion modules can yield better performance on sentiment analysis.

E. Experimental Results on CMU-MOSEI Dataset

Table II shows the experimental results on the CMU-MOSEI dataset. As previously mentioned, MMSA-B preserves the individual semantics of each unimodal data. Therefore, MMSA-B has five subtasks, and each subtask needs the supervision of each unimodal label. Unlike CH-SIMS, which has unimodal labels and multimodal labels, CMU-MOSEI only has multimodal labels. Therefore, to train MMSA-B on the CMU-MOSEI dataset, we use this multimodal label as the label for each subtask. The results of other methods on MOSEI are derived from SELF-MM [59]. For ACC2 and F1, the left side of “/” is calculated as “negative or nonnegative”, and the right side is calculated as “negative or positive”.

As shown in Table II, we observe that MMSA-B achieves the best performance in terms of ACC2 calculated as “negative or nonnegative” and the other indicators is similar to SELF-MM, which shows the effectiveness of MMSA-B. In the future, we will explore the application of MMSA-B with mul-

TABLE III

RESULTS FOR MULTIMODAL FUSION SEQUENCE EXPERIMENTS WITH DIFFERENT DOMINANT MODALITIES

Dominant modality	Fusion module	ACC2	F1	MAE	CORR
audio	A-V-T-A-T-V-A	82.71	83.44	45.21	61.85
	A-T-V-A-V-T-A	82.06	82.52	43.7	63.91
vision	V-T-A-V-A-T-V	82.28	82.41	41.18	65.24
	V-A-T-V-T-A-V	82.49	82.77	41.45	65.5
text	T-A-V-T-V-A-T	82.93	83.62	43.52	66.24
	T-V-A-T-A-V-T	83.15	83.17	39.13	67.34

titasking on datasets with only multimodal sentiment labels, thereby improving the results on CMU-MOSEI.

F. Ablation Experiments and Analysis

In this section, we conducted the following ablation experiments on MMSA-W to further explore the functionality of each module in our algorithm.

1) *Efficiency of Multimodal Fusion Module*: To examine the MHFM, we conduct experiments on the performance on ACC2, F1, MAE and CORR from two aspects: different fusion sequences among modalities and fusion hierarchy. For instance, A-V-T-A-T-V-A means that the order of fusion is $transV_1(\cdot) \sim transT_1(\cdot) \sim transA_1(\cdot) \sim transT_2(\cdot) \sim transV_2(\cdot) \sim transA_2(\cdot)$, and others are similar to this definition. In Table III, it is obvious that the final two fusion modules with text as the dominant modality perform relatively better than the other fusion modules. Specifically, T-V-A-T-A-V-T achieves the best performance in terms of the three metrics ACC2, MAE and CORR. Text original features that are extracted based on BERT contain richer sentiment information compared with visual and audio features; thus, the model based on the text-dominated multimodal fusion module can achieve better results.

In Table IV, we provide the results of the existing multimodal fusion algorithms LF-DNN, TFN, our proposed

TABLE IV
RESULTS FOR FUSION MODULE COMPARISON EXPERIMENTS

	Fusion module	ACC2	F1	MAE	CORR
Other methods	LF-DNN	81.4	81.12	40.3	65.43
	TFN	82.28	82.26	41.12	64.76
One layer	T-A-V-T	82.49	82.53	42.14	64.57
	T-V-A-T	82.71	83.05	42.97	62.21
Two layers	T-V-A-T-A-V-T	83.15	83.17	39.13	67.34
Four layers	T-V-A-T-A-V-T -V-A-T-A-V-T	83.1	83.14	39.31	66.72

TABLE V
RESULTS FOR PROGRESSIVE HIERARCHY
SUBTASKS ABLATION EXPERIMENTS

Decision subtask	ACC2	F1	MAE	CORR
$FL(\times)OL(\times)$	79.87	80.28	55.12	55.44
$FL(\checkmark)OL(\times)$	82.06	81.89	50.07	65.31
$FL(\times)OL(\checkmark)$	82.71	82.41	39.53	66.99
$FL(\checkmark)OL(\checkmark)$	83.15	83.17	39.13	67.34

one-layer fusion and two-layer fusion. It is easy to determine that the one-layer fusion mode T-V-A-T fuses three modalities dominated by text, driven by audio, vision and text in a closed-loop mutual attention structure. Both T-A-V-T and T-V-A-T comprise one layer of the closed-loop mutual attention structure. Considering that the complementary relationship among different modalities may be different, we adjust the fusion sequence of V and A and add one layer to drive multimodal fusion from different perspectives as T-V-A-T-A-V-T. For the LF-DNN and TFN, we replace our multiperspective and hierarchical fusion module with the fusion modules of the LF-DNN and TFN.

As shown in Table IV, we observe that the performance of using the two-layer fusion achieves 0.44% improvement over one-layer fusion and achieves 0.05% improvement over four-layer fusion on ACC2. The four-layer fusion is very similar to the best performance, but its calculation is more than that of two-layer fusion. Different fusion orders of the two-layer fusion can compensate for the differences in complementary information among different modalities. Therefore, the two-layer structure achieves the best fusion effect. In addition, our fusion methods based on the one-layer, two-layer and four-layer structure perform better than LF-DNN and TFN in terms of ACC2. Note that the closed-loop mutual attention structure can fully learn the interaction information among modalities compared with other fusion modules.

2) *Efficiency of Progressive Hierarchy Subtasks*: To verify the efficiency of the last two levels' subtasks, we conduct the following ablation experiments. As shown in Table V, FL indicates a high-level feature fusion subtask at the second level, and OL implies a decision fusion subtask at the third level.

As shown in Table V, we find that employing both FL and OL simultaneously acquire the best performance, which is 3.28% better than the result of removing both subtasks in terms of ACC2. Unimodal subtasks and multimodal fusion subtasks cannot fully utilize the mutual promotion and restrict

TABLE VI
RESULTS FOR UNIMODAL SUBTASKS ABLATION EXPERIMENTS

	Subtasks	ACC2	F1	MAE	CORR
Case1	m	79.43	80.46	48.61	55.48
Case2	m, t	80.53	80.95	46.6	55.1
Case3	m, a	79.87	80.56	47.18	57.21
Case4	m, v	80.31	80.42	46.79	58.75
Case5	m, t, a	79.65	80.85	46.91	56.87
Case6	m, t, v	82.28	81.96	40.48	66.33
Case7	m, a, v	81.84	81.72	41.25	64
Case8	t, a, v	80.74	80.86	51.63	65.58
Case9	m, t, a, v	83.15	83.17	39.13	67.34

relationship between each other, and the unique and common semantics among different modalities cannot be taken into account without complete progressive hierarchy subtasks. We also find that OL has a higher positive effect than FL; different unimodal data in one comment exhibit strongly inconsistent sentiment tendencies, especially in a complicated context. Fortunately, OL can balance different sentiment decision suggestions of different subtasks to obtain analysis results that are more similar to the true sentiment tendency.

3) *Efficiency of Unimodal and Multimodal Fusion Subtasks*: Table VI shows the results of the ablation experiments of each unimodal subtask and multimodal fusion subtask. m, t, a, and v indicate whether the framework has a multimodal fusion subtask, text subtask, audio subtask or vision subtask, respectively. For example, in case 5, "m, t, a" indicates that there is no vision subtask in the framework, and the other cases in the table are similar. In case 9, three unimodal subtasks are complete. In Case 2, we only use multimodal labels and text labels, and the whole framework contains text, multimodal fusion, high-level semantic feature fusion and decision fusion subtasks without audio and visual subtasks. We conduct multiple multitask combination experiments to analyze the impact of different unimodal subtasks.

By comparing Case 2, Case 3 and Case 4, we observe that text information has the greatest effect on sentiment analysis, while audio information achieves a poor performance. We can infer that text has the richest information, such as sentiment words or viewpoint descriptions, which is why we dominate the fusion process with textual modality. Compared to case 9, case 8 only lacks one multimodal fusion subtask, but the results are much worse, indicating that the modal fusion function of the multimodal fusion subtask is indispensable. According to the observations of Case 2 and Case 5, it is not certain that the more unimodal subtasks there are, the better the results because asynchrony among subtasks will have a negative impact on sentiment analysis. As more unimodal subtasks are introduced, the negative effects gradually decrease due to the increase in the complementarity of multimodal information. The performance of Case 8 can yield the best results, which indicates that each subtask extracts different semantics, and the best results can only be achieved by synergizing the semantics of all subtasks.

4) *Ablation Experiments on Loss Function and the Effectiveness of Mask*: In this section, we do not need to compare

TABLE VII
RESULTS FOR THE ABLATION EXPERIMENTS ON LOSS FUNCTION

Mask	ACC2	F1	MAE	CORR
Loss1	84.68	84.99	38.0	68.64
Loss2	80.53	79.99	42.44	64.65

TABLE VIII
RESULTS FOR THE ABLATION EXPERIMENTS ON MASK

Mask	ACC2	F1	MAE	CORR
Mask(\times)	84.68	84.99	38.0	68.64
Mask(\checkmark)	82.06	81.83	39.34	67.06

with other methods, so we do not need to consider the fairness issue that other methods do not use balanced loss. These ablation experiments use $Loss_B$.

The original loss function $Loss_B$ includes four parts: $L(P_t^n, \hat{P}_t^n)$, $L(P_a^n, \hat{P}_a^n)$, $L(P_v^n, \hat{P}_v^n)$ and $L(P_d^n, \hat{P}_d^n)$. We performed an ablation experiment on the loss function composition. On the basis of the original loss function, we add two more items: $L(P_d^n, \hat{P}_m^n)$ and $L(P_d^n, \hat{P}_f^n)$.

As shown in Table VII, in $Loss1$, the loss function consists of the loss of four subtasks, while the loss function consists of the loss of six subtasks in $Loss2$. According to the results in Table VII, the result of $Loss1$ is better. The outputs of multimodal fusion and high-level semantic feature fusion subtasks are not the final multimodal prediction results. Compared with directly calculating $L(P_d^n, \hat{P}_m^n)$ and $L(P_d^n, \hat{P}_f^n)$ in $Loss2$, it is better to have \hat{P}_m^n and \hat{P}_f^n assist in predicting \hat{P}_d^n in $Loss1$.

We performed ablation experiments on the cross-modal attention mechanism fusion network with or without mask. As shown in Table VIII, the ACC2 of the network with a mask is 2.92% higher than that without a mask. This finding shows that mask can better fuse multimodal data, thus improving the sentiment prediction results.

G. Subjective Results

To verify that our algorithm can adapt to complicated contexts, we selected this kind of example. In these examples, the sentiment labels of different unimodal data and the label of the whole sample are different, which is consistent with the sentiment expression scene in our real life.

As shown in Fig. 7, the sentiment labels P_d , P_t , P_a , and P_v corresponding to multimodal, text, audio and vision, respectively, in these examples are different. The aim of multimodal sentiment analysis is to move closer to the P_d marked in red in Fig. 7. The multimodal fusion module in MTFN also constrains three modalities to each other, unlike other algorithms that lack interaction or perform fusion in steps for multimodal fusion, which is similar to MMSA. Therefore, we compare the prediction results of MMSA with MTFN in these examples to demonstrate the adaptability of our overall framework to complicated contexts. In example a, the people in the video are smiling, and thus, the sentiment label of vision is positive. However, the content of his speech is serious, and the sentiment labels of the audio and text are negative. This is a common method of sentiment expression



 <p>Audio, Vision information</p> <p>Text information Chinese: 但是你差那么一点点就会走入歧途。 English: But you are so close to going astray.</p>	<p>Label $P_d: 0$ $P_t: -0.8$ $P_v: 1$ $P_a: -0.4$</p>
	<p>Result MMSA: -0.03 MTFN: -0.38</p>
(a)	
 <p>Audio, Vision information</p> <p>Text information Chinese: 三个孩子我都特别喜欢。 English: I especially like all three children.</p>	<p>Label $P_d: 0.6$ $P_t: 1$ $P_v: -0.8$ $P_a: 1$</p>
	<p>Result MMSA: 0.52 MTFN: 0.77</p>
(b)	
 <p>Audio, Vision information</p> <p>Text information Chinese: 给我也查查吧。 English: Check it for me to.</p>	<p>Label $P_d: -1$ $P_t: 0$ $P_v: -1$ $P_a: -0.8$</p>
	<p>Result MMSA: -0.73 MTFN: -0.23</p>
(c)	
 <p>Audio, Vision information</p> <p>Text information Chinese: 有点担心我们三个人自己的安全。 English: A little worried about the safety of the three of us.</p>	<p>Label $P_d: 0$ $P_t: -0.2$ $P_v: 1$ $P_a: 0.4$</p>
	<p>Result MMSA: -0.04 MTFN: -0.28</p>
(d)	

Fig. 7. Results of MTFN versus MMSA in a complicated context. In this figure, we present the comparative results of MTFN and MMSA for these examples, in which the corresponding labels of multimodal, text, audio and vision are different. The closer the predicted result is to the label in red, the better.

in real life. Notably, its sentiment tendency is neutral. Our prediction is very similar to label P_d . In example c, his expression is painful, and its corresponding vision label P_v is negative. However, the textual content of his speech does not have sentiment tendency, and the value of text label P_t is 0. Compared with MTFN, the prediction result of MMSA is more accurate. The results indicate that our predictions are closer to the true label. These results show that our algorithm is more adapted to sentiment analysis in complicated contexts. We fully consider the differences in semantics among different modalities and separately retain them at the first level. Then, we set progressive hierarchy subtasks to coordinate the semantics and balance the decisions of different subtasks to obtain the final prediction result that is more consistent with the real context.

V. CONCLUSION

A multimodal mutual attention-based progressive multitask sentiment analysis framework is proposed to analyze the sentiment tendencies of multimodal data. To fully fuse multimodal data, we develop a multiperspective and hierarchical fusion module, which performs closed-loop mutual attention based on a cross-modal attention mechanism fusion network on different modal data at one layer and adjusts the fusion order to supplement the differences in modal complementarity between two layers. Then, we develop a high-level feature

fusion subtask to coordinate the semantics among the first level of subtasks and a decision fusion subtask to balance the decisions of the subtasks of the first two levels to further enhance the interaction among all the above subtasks. The experimental results show that the proposed multimodal mutual attention-based progressive multitask sentiment analysis framework has superior sentiment analysis performance as well as low complexity, which makes it suitable for adaptation to complicated contexts in real life.

REFERENCES

- [1] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *J. Netw. Comput. Appl.*, vol. 149, Jan. 2020, Art. no. 102447.
- [2] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Trans. Multimedia*, vol. 23, pp. 4014–4026, 2021.
- [3] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 997–1007, Apr. 2018.
- [4] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.
- [5] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2506–2515.
- [6] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an obviously perfect paper)," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4619–4629.
- [7] A. Bagher 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.
- [8] M. K. Hasan et al., "UR-FUNNY: A multimodal language dataset for understanding humor," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 2046–2056.
- [9] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.
- [10] W. Yu, H. Xu, F. Meng, Y. Zhu, and K. Yang, "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.
- [11] Y. Liu, Z. Yuan, H. Mao, Z. Liang, and W. Yang, "Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-mixup," in *Proc. Int. Conf. Multimodal Interact.*, 2022, pp. 247–258.
- [12] X. Cheng, J. Lu, B. Yuan, and J. Zhou, "Identity-preserving face hallucination via deep reinforcement learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4796–4809, Dec. 2020.
- [13] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1034–1047, Mar. 2022.
- [14] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1709–1717, Jun. 2020.
- [15] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.
- [16] V. Pérez-Rosas, R. Mihalcea, and L. P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 973–982.
- [17] T. Liu, J. Wan, X. Dai, F. Liu, Q. You, and J. Luo, "Sentiment recognition for short annotated GIFs using visual-textual fusion," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 1098–1110, Apr. 2020.
- [18] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2539–2544.
- [19] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.
- [20] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.
- [21] J. Chen, S. Yan, and K.-C. Wong, "Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 10809–10818, Aug. 2020.
- [22] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, "Audio-visual emotion-aware cloud gaming framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2105–2118, Dec. 2015.
- [23] Z. Sun, P. K. Sarma, W. Sethares, and E. P. Bucy, "Multi-modal sentiment analysis using deep canonical correlation analysis," in *Proc. Annu. Conf. Int. Speech. Commun. Assoc.*, Sep. 2019, pp. 1323–1327.
- [24] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.
- [25] T. Yu, J. Yu, Z. Yu, Q. Huang, and Q. Tian, "Long-term video question answering via multimodal hierarchical memory attentive networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 931–944, Mar. 2021.
- [26] 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.*, Oct. 2021, pp. 6–15.
- [27] 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.
- [28] Y. Ou, Z. Chen, and F. Wu, "Multimodal local-global attention network for affective video content analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1901–1914, May 2021.
- [29] Q. T. Truong and H. W. Lauw, "VistaNet: Visual aspect attention network for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 305–312.
- [30] 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.
- [31] 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.
- [32] F. Chen, J. Shao, A. Zhu, D. Ouyang, X. Liu, and H. T. Shen, "Modeling hierarchical uncertainty for multimodal emotion recognition in conversation," *IEEE Trans. Cybern.*, early access, Jul. 12, 2022, doi: 10.1109/TCYB.2022.3185119.
- [33] P. P. Liang, Z. Liu, Y.-H.-H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1569–1576.
- [34] E. J. Barezi, P. Momeni, I. Wood, and P. Fung, "Modality-based factorization for multimodal fusion," in *Proc. 4th Workshop Represent. Learn. NLP (RepLANLP-)*, 2019, pp. 260–269.
- [35] 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.
- [36] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 481–492.
- [37] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–10.
- [38] 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.

- [39] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 2122–2132.
- [40] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.
- [41] P. Hai, P. P. Liang, T. Manzini, L. P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6892–6899.
- [42] Y. Tsai, P. P. Liang, A. Zadeh, L. P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–11.
- [43] P. P. Liang, Y. C. Lim, Y.-H.-H. Tsai, R. Salakhutdinov, and L.-P. Morency, "Strong and simple baselines for multimodal utterance embeddings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 2599–2609.
- [44] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, Apr. 2019.
- [45] 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.
- [46] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, and L. P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.
- [47] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1033–1038.
- [48] F. Chen, Z. Luo, Y. Xu, and D. Ke, "Complementary fusion of multi-features and multi-modalities in sentiment analysis," in *Proc. CEUR Workshop*, 2020, pp. 82–89.
- [49] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3454–3466.
- [50] A. Kumar and J. Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4477–4481.
- [51] 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.
- [52] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 4171–4186.
- [53] B. Mcfee, C. Raffel, D. Liang, D. Ellis, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.
- [54] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP: A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014.
- [55] 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.
- [56] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.
- [57] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal DNN feature fusion," in *Proc. 1st Grand Challenge Workshop Human Multimodal Lang. (Challenge-HML)*, 2018, pp. 11–19.
- [58] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. Subramanyam, "Benchmarking multimodal sentiment analysis," in *Int. Conf. Comput. Linguistics Intell. Text Process.* Springer, 2017, pp. 166–179.
- [59] 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.*, vol. 35, no. 12, pp. 10790–10797, May 2021.
- [60] T. Wu, J. Peng, W. Zhang, H. Zhang, and Y. Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107676.



Lijun He received the B.S. and Ph.D. degrees from the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008 and 2016, respectively. She is currently an Associate Professor with the School of Information and Communications Engineering, Xi'an Jiaotong University. Her research interests include video communication and transmission, video analysis, processing, and compression techniques.



Ziqing Wang received the B.E. degree from Jilin University in 2020. She is currently pursuing the master's degree with the School of Information and Communications Engineering, Xi'an Jiaotong University. Her research interests include multimodal sentiment analysis.



Liejun Wang received the Ph.D. degree from the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China, in 2012. He is currently a Professor with the School of Information Science and Engineering, Xinjiang University. His research interests include wireless sensor networks, encryption algorithm, and image intelligent processing.



Fan Li (Senior Member, IEEE) received the B.S. and Ph.D. degrees in information engineering from Xi'an Jiaotong University, Xi'an, China, in 2003 and 2010, respectively. From 2017 to 2018, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of California at San Diego, San Diego, CA, USA. He is currently a Professor with the School of Information and Communications Engineering, Xi'an Jiaotong University. He has published more than 30 technical articles. His research interests include multimedia communication and video quality assessment. He was a member of the Organizing Committee for IET VIE 2008. He served as the Local Chair for ICST Wicon 2011.