Multimodal Sentiment Analysis Based on Attentional Temporal Convolutional Network and Multi-Layer Feature Fusion

Hongju Cheng, *Member, IEEE*, Zizhen Yang, Xiaoqi Zhang, and Yang Yang

***Abstract*—Multimodal sentiment analysis aims to extract and integrate information from different modalities to accurately identify the sentiment expressed in multimodal data. How to effectively capture the relevant information within a specific modality and how to fully exploit the complementary information among multiple modalities are two major challenges in multimodal sentiment analysis. Traditional approaches fail to obtain the global contextual information of long time-series data when extracting unimodal temporal features, and they usually fuse the features from multiple modalities with the same method and ignore the correlation between different modalities when modeling inter-modal interactions. In this paper, we first propose an Attentional Temporal Convolutional Network (ATCN) to extract unimodal temporal features for enhancing the feature representation ability, then introduce a Multi-layer Feature Fusion (MFF) model to improve the effectiveness of multimodal fusion, which fuses the different-level features by different methods according to the correlation coefficient between the features, and cross-modal multi-head attention is used to fully explore the potential relationship between the low-level features. The experimental results on SIMS and CMU-MOSI datasets show that the proposed model achieves superior performance on sentiment analysis tasks compared to state-of-the-art baselines.**

***Index Terms*—Attentional temporal convolutional network, cross-modal multi-head attention, multimodal sentiment analysis, multi-layer feature fusion.**

I. Introduction

SENTIMENT plays an important role in human cognition, especially in decision-making, perception, and interpersonal communication. The sentiment can be judged by information from different sources such as voice, facial expressions, text, body movements, and physiological signals, each of which is regarded as one modality [1]. Multimodal sentiment analysis refers to sentiment identification in combination with multimodal information, which can effectively improve the accuracy and robustness of the sentiment analysis by exploiting complementary information among different modalities. A survey published by D'mello et al. [2] shows that multimodal sentiment analysis systems are consistently more accurate than their unimodal counterparts. However, the heterogeneous and temporal nature of multimodal data [3] still makes it challenging to explore the intrinsic relationship among modalities and fuse information from multiple modalities in an effective way.

The core challenge of multimodal sentiment analysis is to design a fusion strategy that can better capture the interactions among multimodal signals. Multimodal interactions are split into two types: intra-modal interactions and inter-modal interactions. The former focuses on interactions among features within a specific modality when data at different timesteps are correlated. The modeling of intra-modal interactions will affect specific-modal representation, thus it has a significant influence on the final decisions. While the latter centers on the interactions among different modalities, which aims at exploiting the complementary information from different modalities to provide more accurate predictions. Most existing approaches tend to consider inter-modal interactions, turning a blind eye to intra-modal interactions. Note that modalities are temporal and heterogeneous, and it is necessary to design a multimodal fusion model that can simultaneously learn modality-specific features within modality and capture the interactions among different modalities [4].

The most common multimodal fusion technique in the literature generally integrates features from different modalities directly by concatenation or dot product operation [5]. This method can capture the interactions among modalities, but it loses the temporal dependencies within each modality and potentially suppresses intra-modal interactions. Inspired by the success of attention models, many attention-based fusion strategies have been developed to capture multimodal interactions. Tsai et al. [6] proposed Multimodal Transformer (MulT) that utilized multi-head attention to compute the correlations among modalities from different subspaces at different positions and capture the inter-modal representations directly from unaligned multimodal sequential data. However, it did not explicitly capture the inherently sequential nature of multimodal data, and only used the fused features for prediction, which ignores specific characteristics within each modality.

Some multimodal fusion models with the recurrent neural networks, such as LSTMs and GRUs [7], [8], [9], can effectively capture the temporal structure within a single modality, but they are short of specific components to capture intra-modal and inter-modal interactions explicitly and fuse multimodal

Manuscript received 30 June 2022; revised 5 February 2023; accepted 4 April 2023. Date of publication 7 April 2023; date of current version 29 November 2023. This work was supported in part by the National Science Foundation of China under Grant 61872091. Recommended for acceptance by E. Mower Provost. *(Corresponding author: Hongju Cheng.)*

The authors are with the College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China (e-mail: cscheng@fzu.edu.cn; 200327108@fzu.edu.cn; 1404908269@qq.com; yang.yang.research@gmail.com).

Digital Object Identifier 10.1109/TAFFC.2023.3265653

features at each timestamp in the recurrent structure. Zadeh et al. [10] proposed Tensor Fusion Network that calculated a multi-dimensional tensor representation by Cartesian product to describe both intra-modality and inter-modality interactions, though with high computational complexity. Liu et al. [11] proposed the Low-rank Multimodal Fusion method that leveraged low-rank weight tensors to make multimodal fusion. Murphy et al. [12] revealed that late fusion is better than early fusion when the correlation between modalities is high, and early fusion performs better when the correlation is low. The existing multimodal fusion models tend to design more complex networks to fuse the features from different modalities by the same method and ignore the correlations between different modalities. It is essential to look further into the implicit relationships among different modalities. However, it is difficult to quantify the correlation between the high-dimensional features of different modalities. How to select an appropriate fusion strategy for fusing different-level features is another critical issue.

In this paper, we propose an Attentional Temporal Convolutional Network (ATCN) to capture the global contextual information within a single modality. The basic idea is to insert an attention block on the basis of traditional temporal convolutional layers, which distributes higher weights for vital information to suppress irrelevant information. Low-level features contain more detailed information and less semantic information, while high-level features contain the opposite. If the features from different modalities are fused in the same way, the potential correlation information among modalities cannot be fully exploited, and it may even lead to redundancy; for example, although multi-head attention can help learn more abundant implicit information, it will also lead to the low-rank bottleneck problem and adversely affect representation information when attention heads are projected into a low-dimensional space [13]. To fully use complementary information among modalities and reduce redundant information, a novel Multilayer Feature Fusion (MFF) model is proposed in this paper. MFF first explores the correlations between different modalities, then fuses the low-level features of low-correlation modalities by using cross-modal multi-head attention, and concatenates the high-level features of high-correlation modalities. It can effectively improve the sentiment analysis performance with lower computational cost.

The main contributions of this paper can be summarized as follows:

1) We develop ATCN to effectively extract unimodal temporal features. The network can better capture the long-range dependencies of the sequential input data and adaptively focus on mocriticalant information in combination with the temporal convolutional network and the attention mechanism;
2) We propose a novel MFF model to fuse the features from different modalities by exploring the feature correlations between modalities. The model can fully utilize the complementary information among multiple modalities and effectively reduce the interactions of irrelevant information to improve fusion efficiency;
3) To evaluate the effectiveness of the proposed model, we conduct extensive experiments on SIMS and CMU-MOSI multimodal datasets. The experiment results show our method outperforms current state-of-the-art baselines.

The rest of this paper is organized as follows: Section II presents the related work in recent years. We define the problem formulation in Section III. Section IV introduces the details of the proposed model in this paper. Section V describes the experimental datasets, baseline models, evaluation metrics, and the analysis of experimental results. Section VI shows the conclusions and future research directions.

II. Relate Work

In this section, we briefly review current related work on unimodal temporal feature extraction, multimodal data fusion, and sentiment analysis.

A. *Unimodal Temporal Feature Extraction*

The temporal models commonly use convolutional neural networks or recurrent neural networks to capture the temporal structure and long-term dependencies within data, thus extracting the temporal features. TextCNN model [14] can effectively capture their adjacent relations by utilizing a set of convolutional kernels with different heights and identical widths and feature dimensions to extract local features. Recurrent neural networks extract the hidden features from input data sequentially which simulates the memory, forgetting, and updating of the brain to obtain new feature representations. Zhao et al. [15] proposed a time-weighted LSTM model, which attempted to capture the temporal dependencies of data and assign weights to data via the time weight function to improve prediction accuracy. Zhang et al. [16] proposed a new architecture, termed Deep and Wide Neural Networks (DWNN) that combined convolutional neural network and recurrent neural network, thus outperforming the general RNN models. However, traditional RNNs are usually challenging to store long-history information due to the forgetgating mechanism.

Recently, Temporal Convolutional Networks (TCNs) have achieved some breakthrough success in sequence modeling, particularly for long-range sequential data. Compared to traditional recurrent architectures such as LSTMs and GRUs, TCNs can exhibit longer effective memory, which is composed of causal convolution, dilated convolution, and residual connection. Lei et al. [17] introduced a Temporal Deformable Residual Network (TDRN) that computed the residual stream and the pooled/unpooled stream by temporal residual modules. Pandey et al. [18] inserted an additional Temporal Convolution Module (TCM) into the fully convolutional temporal network, called Temporal Convolutional Neural Network (TCNN), resulting in a real-time audio enhancement in the temporal domain. Du et al. [19] proposed a temporal hourglass convolutional neural network, which can effectively capture the long-term dynamic dependencies of videos and outperform canonical recurrent networks without using any attention.

B. Multimodal Data Fusion

Multimodal fusion combines information from two or more modalities to make predictions, which can be categorized into two major types: early fusion and late fusion.

Late fusion first acquires the features of each modality and makes independent predictions, after which the decisions obtained will be integrated to generate the final result via different mechanisms, such as weighted sum [20], averaging [21], majority voting [22], or a learnable model [23]. Late fusion builds separate models for each input modality, and its improved flexibility ensures a seamless extension to more modalities. However, the low-level interactions among different modalities are usually not modeled effectively, posing a potential threat to adverse performance.

Early fusion usually extracts separate feature representations from each modality, then fuses them at the feature level, and makes final predictions. The most common early fusion technique [24] is the concatenation of the features from each modality to construct a multimodal joint representation. This method is easy to handily feasible while stopping short of looking into the interactions among modalities. In recent years, deep learning-based techniques have been employed to fuse information from multiple modalities to exploit the complementarity of multimodal heterogeneous data and provide more accurate predictions. Considering the time-series nature of multimodal data from different sensors, some researchers have exploited recurrent neural structures like LSTMs and GRUs for multimodal fusion. He et al. [25] designed a unimodal reinforced Transformer with temporal squeeze fusion, which automatically explored the time-dependent interactions within unaligned multimodal sequences by modeling unimodal and multimodal sequences from the perspective of compressing the temporal dimension.

Attention selectively focuses on important information and simultaneously filters out irrelevant information, and more researchers employ attention mechanisms to explore intra-modal and inter-modal interactions [26]. Wang et al. [27] proposed a hierarchical attention-LSTM model based on the cognitive limbic system (HALCB) that contained two modules-the low-path module for the binary classification and the high-path module for the multi-classification. Xiao et al. [28] proposed a Hierarchical Self-Attention Fusion (H-SATF) model to acquire fusion features within modality and among different modalities, which can effectively improve the performance by retaining important information and filtering irrelevant information. Xue et al. [29] proposed a Multi-level Attention Map Network (MAMN) that captured the consistent and heterogeneous correlation between multi-granular features and explored the interactions between multi-level attention maps.

Tensor-based approaches have been commonly used for multimodal data fusion, which computes the outer product between vectors and exploits the interactions among all elements of the vectors to obtain a joint feature representation. Zadeh et al. [10] proposed a Tensor Fusion Network (TFN) that calculated a multi-dimensional tensor representation by Cartesian product to model both intra-modality and inter-modality dynamic interactions, though with high computational complexity. Liu et al. [11] proposed the Low-rank Multimodal Fusion method that leveraged low-rank weight tensors to make multimodal fusion much more efficient. Mai et al. [30] proposed a new local confined modality fusion network, which modeled local interactions by using the tensor fusion method and then modeled global interactions by capturing dependencies between local tensors via a bidirectional multi-connected LSTM, thus obtaining an integral comprehension of multimodal information.

C. Sentiment Analysis

With the main focus on a single modality, such as text and image, early sentiment analysis work aims to obtain the most effective sentiment features. The in-depth study of unimodal sentiment analysis laid the foundation for multimodal sentiment analysis. Liu et al. [31] proposed a new hierarchical neural network model based on dynamic word embedding for text sentiment analysis, in which dynamic word embedding points to the semantic information of polysemous words, while the hierarchical neural network can capture global and local features from the sentence representation. Wu et al. [32] proposed a Cascade EF-GAN network to extract richer expression features by focusing on local expression focal information to reduce artifacts and blurring of facial images. Truong et al. [33] found that the combination of image and text leads to the comprehension of the sentiment information expressed in documents and achieve more accurate sentiment analysis, which inspired more researchers for the introduction of more modal information for sentiment analysis. Dashtipour et al. [34] proposed a novel context-aware multimodal sentiment analysis framework, which simultaneously exploited audio, visual, and text to accurately judge the expressed sentiment.

Besides, sentiment analysis research based on bio-sensing signals such as galvanic skin Reply (GSR), electroencephalogram (EEG), heart rate (HR), etc., has attracted researchers' great interest in recent years. Soleymani et al. [35] studied how to instantaneously detect speakers' sentiment by using EEG signals and facial expressions, and analyzed the relationship between the EEG features and the facial expression features. Zheng et al. [36] developed a novel multimodal emotion recognition framework based on multimodal deep neural networks, to explore the complementary information between EEG and eye movements. They demonstrated that the fusion of EEG and eye movements could significantly improve the emotion recognition accuracy. Song et al. [5] proposed a multichannel EEG sentiment recognition model based on a novel dynamical graph convolutional neural network (DGCNN) to learn the intrinsic relationship among different EEG channels and better extract the non-linear discriminative features.

In summary, prior approaches tend to fuse the features of different modalities by using the same strategy and seldom consider the correlations among different modalities. The proposed multimodal sentiment analysis model based on ATCN and MFF can effectively improve performance by capturing both specific characteristics within each modality and interactions among multiple

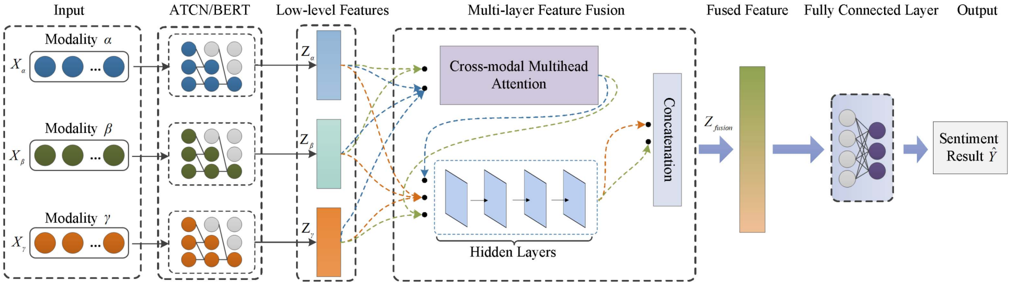


Fig. 1. Model framework.

modalities, which can fully exploit the complementary information while reducing the interference of invalid information.

III. Problem Definitions

In this paper, we aim to judge the expressed sentiment by combining multimodal data. These multimodal data include but are not limited to text, visual, acoustic, etc. Our work focuses on how to effectively integrate three different modalities for sentiment analysis. Suppose that a multimodal sentiment dataset contains N multimodal samples, $X = \{X_i\}_{i=1}^N$, each of which corresponds to a sentiment label Y_i. Each multimodal sample X_i consists of data from three modalities with different sequence lengths, namely X_α, X_β, and $X_\gamma . X_i = (X_\alpha, X_\beta, X_\gamma)$, where $X_\alpha = \{x_{\alpha,t} | t = 1, 2, \ldots, T_\alpha\} \in \mathbb{R}^{T_\alpha \times d_\alpha}, X_\beta = \{x_{\beta,t} | t = 1, 2, \ldots, T_\beta\} \in \mathbb{R}^{T_\beta \times d_\beta}$, and $X_\gamma = \{x_{\gamma,t} | t = 1, 2, \ldots, T_\gamma\} \in \mathbb{R}^{T_\gamma \times d_\gamma}$, in which T_m denotes sequence length for modality m and d_m is the corresponding feature dimensions, $x_{m,t}$ is the feature representation at timestep t for modality m, $m \in \{\alpha, \beta, \gamma\}$.

Given a multimodal dataset $\{(X_i, Y_i)\}_{i=1}^N$, our task is to establish a multimodal sentiment analysis model $\mathcal{M}$ to predict the sentiment for a given multimodal input, where N is the number of training samples. The optimization objective can be defined as follows:

$$\min_\theta \frac{1}{N} \sum_{i=1}^N l(\mathcal{M}(X_i; \theta), Y_i), \tag{1}$$

where l is a loss function, and θ represents trainable parameters.

IV. Multimodal Sentiment Analysis Model

A. Model Framework

Multimodal sentiment data includes external signals such as facial expressions, text, and audio, and internal signals such as respiration rate and heart rate. These multimodal data are often presented in the form of sequences. Multimodal fusion is more effective, and promising, and has better generalizability over unimodal learning by exploring the complementarity among multiple modalities. However, various modalities differ from each other in their internal structure. For example, a video is a continuous sequence of images while text consists of discrete words. It is challenging to make full use of the correlation and complementarity of multimodal data. To better capture interactions within a single modality and complementarity among multiple modalities, we propose a novel multimodal sentiment analysis model based on ATCN and MFF in this paper. The framework is presented (details can be found in the following sections) in Fig. 1.

First, we utilize ATCN to extract temporal features of unimodal sequential data, which can simultaneously explore complex temporal dependencies within a single modality and suppress interference of irrelevant information by distributing higher attention weights to important information. Multimodal fusion needs to take both complementarity and redundancy in multimodal data into account to fully explore the hidden connections among them. So in the second step, we introduce a novel MFF model, which integrates the low-level features of low-correlation modalities by using cross-modal multi-head attention and concatenates the high-level features of high-correlation modalities. The cross-modal multi-head attention focuses on information from different representation subspaces at different positions and repeatedly reinforces the target modality with low-level information from another auxiliary modality. In this way, it can better capture interactions among the low-level features of different modalities. Finally, the fused multimodal representation is set as the input of the fully connected layer for the final sentiment predictions.

B. Unimodal Feature Extraction

TCN is a special type of convolutional neural network, used for sequence modeling tasks, which can effectively enlarge the receptive field without increasing the number of model parameters. Dilated convolution is a technique that expands the convolution kernel by inserting holes between its consecutive elements. It helps to expand the covered area of the input. Although TCN can efficiently explore the global contextual information from long time-series data, consecutive input sequences may often contain irrelevant information, which introduces additional noise. In this paper, ATCN is proposed to better capture the specific characteristics within each modality.

Fig. 2 shows the structure of ATCN. K is the number of hidden layers of the ATCN. The filter size $k = 2$, and the dilation factor $d = 2^{l-1}$, where $l = \{1, 2, \ldots, K\}$. When $d = 1$, a dilated convolution reduces to a normal convolution; and $d = 2$ means skipping one element per input; $d = 4$ means skipping three elements, and so on. The dilation factor d is increased exponentially with the network depth, thus it can cover more information of the input.

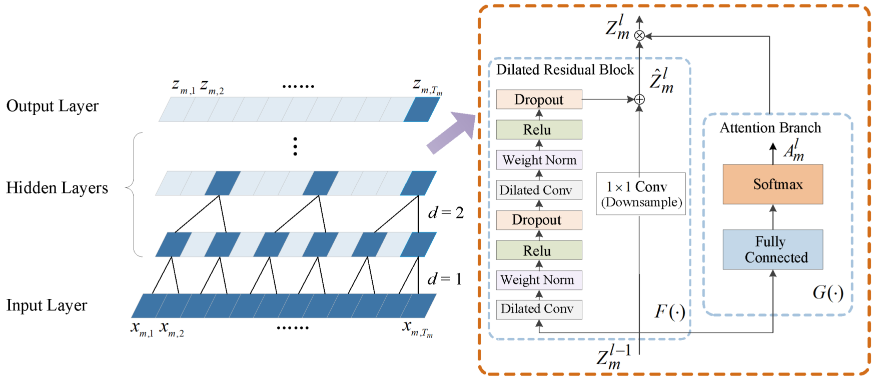


Fig. 2. The structure of ATCN.

Considering the sequence $X_m = \{x_{m,t} | t = 1, 2, \ldots, T_m\} \in \mathbb{R}^{T_m \times d_m}$ of modality *m* as the input of ATCN, the output is $Z_m^{l-1} = \{z_{m,t}^{l-1} | t = 1, 2, \ldots, T_m\} \in \mathbb{R}^{T_m \times N_{l-1}}$, where N_{l-1} is the number of convolutional filters in the $(l-1)$th hidden layer. We can compute the output $\hat{Z}_m^l \in \mathbb{R}^{T_m \times N_l}$ from the *l*th layer with the following formula:

$$\hat{Z}_m^l = Z_m^{l-1} + F(Z_m^{l-1}, W_{m,l}), \tag{2}$$

where $W_{m,l} \in \mathbb{R}^{k \times N_{l-1} \times N_l}$ are the weights of each dilated convolution filter, and $F(\cdot)$ denotes a series of operations of dilated residual block.

Different from traditional TCNs, our ATCN endeavors to better extract temporal features by adding an attention block between two hidden layers. We utilize the attention block of ATCN to generate feature matrix $A_m^l \in \mathbb{R}^{T_m \times N_l}$ with the same size of $\hat{Z}_m^l \in \mathbb{R}^{T_m \times N_l}$. The operations at the *l*th layer $Z_m^l \in \mathbb{R}^{T_m \times N_l}$ is formally described as

$$A_m^l = G(Z_m^{l-1}, W_{m,Att}), \tag{3}$$

$$Z_m^l = A_m^l \cdot \hat{Z}_m^l, \tag{4}$$

where $G(\cdot)$ denotes a series of operation of attention block, For each timestep $t \in \{1,2,\ldots, T_m\}$, $A_m^l \in \{A_m^l(1), A_m^l(2), \ldots, A_m^l(T_m)\}$ indicates the importance of each timestep and its weights are added to $\hat{Z}_m^l$, which can ensure that the network focuses on those important information.

ATCN has an extraordinarily long memory, which means that there is no information leakage from the future into the past. The output of ATCN is dependent on the whole input sequence and achieves full input coverage. We take the last element of the output of the last hidden layer as the final output of ATCN, $Z_m = z_{m,T_m} \in \mathbb{R}^{d_m}$. For modality α, β, γ, we use ATCN to respectively transform their input $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$, $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$, $X_\gamma \in \mathbb{R}^{T_\gamma \times d_\gamma}$, into the corresponding low-level features, namely $Z_\alpha \in \mathbb{R}^{d_\alpha}$, $Z_\beta \in \mathbb{R}^{d_\beta}$, $Z_\gamma \in \mathbb{R}^{d_\gamma}$, for the following fusion process.

C. Multi-Layer Feature Fusion (MFF)

The deep neural networks with different numbers of layers extract different-level features. Low-level features contain more detailed information but less semantic information, while high-level features are just the opposite. If different modalities are fused at the same layer, it is difficult to strike a balance between low-level detailed information and high-level semantic information. The information of different modalities is usually correlated and it is argued that the closer semantics of two modalities means a higher correlation between them. In such a case, concatenating the high-level features at the upper layer can minimize redundancy and lead to performance improvements with lower computational costs. To make multimodal fusion more effective, we design an MFF model to fuse different modal features by exploring the correlations between the modalities in this paper. MFF mainly consists of two components: 1) Deep Canonical Correlation Analysis (DCCA) for measuring correlations between different modalities (Section IV-C1); 2) Cross-modal multi-head attention for fusing the features of different modalities (Section IV-C2). Section IV-C3 explains how MFF gets the multimodal fusion feature by combining DCCA and cross-modal multi-head attention in detail.

1) Deep Canonical Correlation Analysis (DCCA): Considering the high-dimensional features of each modality, traditional methods are difficult to measure the correlations between them. In this paper, DCCA [37] is employed to compute the correlation coefficient between two high-dimensional features by inputting them through multiple stacked layers of nonlinear transformation. Given the inputs $Z_p \in \mathbb{R}^{d_p}$, $Z_q \in \mathbb{R}^{d_q}$, we first use different deep neural networks f_p and f_q to perform nonlinear transformation on them respectively, which is described as

$$H_p = f_p(Z_p; \theta_p), \tag{5}$$

$$H_q = f_q(Z_q; \theta_q), \tag{6}$$

where θ_p and θ_q represent the parameter of the network f_p and f_q respectively.

The optimization objective of DCCA is to maximize *corr*(H_p, H_q) to obtain the best optimal network parameters (θ_p^*, θ_q^*), where *corr*(H_p, H_q) denotes the correlation between H_p and H_q. The objective can be defined as follows.

$$\begin{aligned}(\theta_p^*, \theta_q^*) &= \underset{(\theta_p, \theta_q)}{\arg\max} corr(H_p, H_q) \\ &= \underset{(\theta_p, \theta_q)}{\arg\max} \frac{cov(H_p, H_q)}{\sqrt{\sigma(H_p)} \cdot \sqrt{\sigma(H_q)}}, \end{aligned} \tag{7}$$

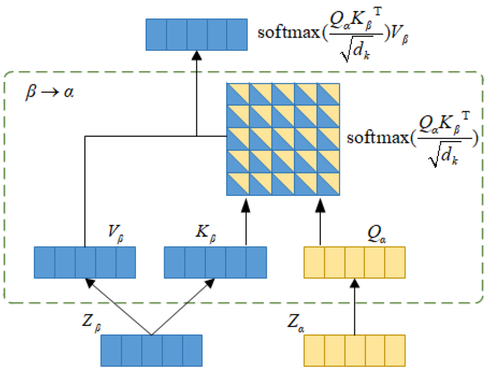

softmax($\frac{Q_\alpha K_\beta^{\mathrm{T}}}{\sqrt{d_k}}$)$V_\beta$
$\beta \rightarrow \alpha$
softmax($\frac{Q_\alpha K_\beta^{\mathrm{T}}}{\sqrt{d_k}}$)
V_β
K_β
Q_α
Z_β
Z_α

Fig. 3. Cross-modal attention between two different modalities α and β.

where *cov*(H_p, H_q) is the covariance between H_p and H_q, and $\sigma(H_p)$, $\sigma(H_q)$ is the variance of H_p and H_q, respectively. Note that covariance and variance are statistical characteristics that describe the relationship between two feature sets.

The maximum value calculated is the correlation coefficient between the high-dimensional features Z_p and Z_q, i.e.,

$$\rho(Z_p, Z_q) = \frac{cov(f_p(Z_p;\theta_p^*), f_q(Z_q;\theta_q^*))}{\sqrt{\sigma(f_p(Z_p;\theta_p^*))} \cdot \sqrt{\sigma(f_q(Z_q;\theta_q^*))}}, \tag{8}$$

where $\rho(Z_p, Z_q) \in$ [-1,1]. $\rho(Z_p, Z_q)$ is closer to 0 when the correlation between Z_p and Z_q is lower. Following the properties of the covariance, we have $\rho(Z_q, Z_p) = \rho(Z_p, Z_q)$.

2) Cross-Modal Multi-Head Attention: Cross-modal multi-head attention is utilized to deeply explore the latent interactions between different low-level features, where cross-modal attention focuses on the cross-modal interactions, and multi-head attention focuses on information from different representation subspaces at different positions.

A) Cross-modal multi-head attention between two different modalities

Consider two modalities α and β. The unidirectional cross-modal attentional process from modality β to modality α, namely $\beta \rightarrow \alpha$, is shown in Fig. 3. During the process, we take the feature of target modality α as the query vector and compute the similarity between α and β. In this way, the information of modality β is attached to modality α.

We first apply linear projections for features $Z_\alpha \in \mathbb{R}^{d_\alpha}$ and $Z_\beta \in \mathbb{R}^{d_\beta}$, respectively. They are projected to the same d_u-dimensional vector space, as shown in the following:

$$Z_\alpha = W_\alpha Z_\alpha + b_\alpha, \tag{9}$$

$$Z_\beta = W_\beta Z_\beta + b_\beta, \tag{10}$$

where $W_\alpha \in \mathbb{R}^{d_u \times d_\alpha}$, $W_\beta \in \mathbb{R}^{d_u \times d_\beta}$ are projection matrices for α and β modalities, respectively.

The projected feature $Z_\alpha' \in \mathbb{R}^{d_\alpha}$ is transformed as follows.

$$Q_\alpha = Z_\alpha' W_Q, \tag{11}$$

$$K_\alpha = Z_\alpha' W_K, \tag{12}$$

$$V_\alpha = Z_\alpha' W_V, \tag{13}$$

where the queries Q_α, keys K_α, and values V_α are vectors, and W_Q, W_K, W_V are weights, $W_Q \in \mathbb{R}^{d_u \times d_u}$, $W_K \in \mathbb{R}^{d_u \times d_u}$, $W_V \in \mathbb{R}^{d_u \times d_u}$.

Given modality β, we can use the same operation to obtain Q_β, K_β, and V_β. Then we carry out the unidirectional cross-modal fusion from modality β to modality α by utilizing Q_α, K_β, and V_β. The process is defined as follows. Modality α keeps receiving correlated meaningful information from modality β.

$$CA(Q_\alpha, K_\beta, V_\beta) = \text{softmax}\left(\frac{Q_\alpha \mathbf{K}_\beta^{\mathrm{T}}}{\sqrt{d_k}}\right) V_\beta, \tag{14}$$

where softmax is used to normalize the attention scores.

Instead of the single cross-modal attention with d_u-dimensional keys, values, and queries, the proposed cross-modal multi-head attention linearly projects these queries, keys, and values h times with different parameters. On each of these projected versions, we perform the cross-attention in parallel and obtain d_v-dimensional outputs. These outputs are concatenated and projected again, and we have the final result with the dimension d_u. Cross-modal multi-head attention allows the model to focus on information from different representation subspaces at different positions, thus leading to better fusion performance. Considering the cross-modal multi-head attention between modalities α and β, the specific process is shown as follows:

$$\begin{aligned} C_{\beta \rightarrow \alpha} &= MultiHead(Q_\alpha, K_\beta, V_\beta) \\ &= Concat(head_1, head_2, head_h)W, \end{aligned} \tag{15}$$

where $head_i$ is calculated as follows:

$$head_i = CA(Q_\alpha W_{Q,i}, K_\beta W_{K,i}, V_\beta W_{V,i}), \tag{16}$$

where $W_{Q,i}, W_{K,i}, W_{V,i}, W$ are the learnable projection matrixes, $W_{Q,i} \in \mathbb{R}^{d_u \times d_k}$, $W_{K,i} \in \mathbb{R}^{d_u \times d_k}$, $W_{V,i} \in \mathbb{R}^{d_u \times d_v}$, $W \in \mathbb{R}^{d_u \times d_u}$, and $d_k = d_v = d_u/h$, $head_i \in \mathbb{R}^{d_u/h}$. We can better learn potential interactions without additional time complexity with multi-head attention.

The information learned from the auxiliary modality β is attached to the target modality α. Then, the cross-modal feature of modality α is obtained after batch normalization (BN), which is represented as follows:

$$F_\alpha = \text{ReLU}(\text{BN}(Concat(C_{\beta \rightarrow \alpha}, Z_\alpha)), \tag{17}$$

Since the interactions between different modalities are mutual, we obtain the cross-modal feature F_β in the similar way. The fused multimodal feature is the concatenation of the feature of modality α and modality β.

$$Z_{\alpha,\beta} = Concat(F_\alpha, F_\beta). \tag{18}$$

B) Cross-modal multi-head attention between three different modalities

It should be noted that the fusion method based on cross-modal multi-head attention can seamlessly extend to more than two modalities. Considering two auxiliary modalities, α and β, and the target modality γ, we first compute the cross-modal features of $\alpha \rightarrow \gamma$ and $\alpha \rightarrow \gamma$ separately, with the (15), and then attach the meaningful correlated information from modalities

Algorithm 1: Multi-Layer Feature Fusion (MFF).

```
Input  : The low-level features of modality $\alpha$, $\beta$, $\gamma$: $Z_\alpha$, $Z_\beta$, $Z_\gamma$
Output: The fusion feature $Z_{fusion}$
// $Z_{\alpha+\beta}$=[$Z_\alpha$, $Z_\beta$], $Z_{\alpha+\gamma}$=[$Z_\alpha$, $Z_\gamma$], $Z_{\beta+\gamma}$=[$Z_\beta$, $Z_\gamma$]
Calculate $\rho_{\alpha,\beta}$, $\rho_{\alpha,\gamma}$, $\rho_{\beta,\gamma}$ , $\rho_{\alpha+\beta,\gamma}$, $\rho_{\alpha+\gamma,\beta}$, and $\rho_{\beta+\gamma,\alpha}$ with Eq. (8);
$\delta$ is the average of $\rho_{\alpha,\beta}$, $\rho_{\alpha,\gamma}$, $\rho_{\beta,\gamma}$ , $\rho_{\alpha+\beta,\gamma}$, $\rho_{\alpha+\gamma,\beta}$, and $\rho_{\beta+\gamma,\alpha}$;
// $i$, $j \in \{\alpha, \beta, \gamma\}$, $i \neq j$, and $u = \{\alpha, \beta, \gamma\} - \{i, j\}$
Assume that $\rho_{i+j,u}$ is the maximum one in $\{\rho_{\alpha+\beta,\gamma}, \rho_{\alpha+\gamma,\beta}, \rho_{\beta+\gamma,\alpha}\}$ ;
if $\rho_{i+j,u} \leq \delta$ then
    Calculate $Z_{\alpha,\beta,\gamma}$ with Eq. (20);
    $Z_{fusion} \leftarrow \mathcal{A}_{\alpha,\beta,\gamma}(Z_{\alpha,\beta,\gamma}; \theta_{\alpha,\beta,\gamma})$;
else
    if $\rho_{i,j} \leq \delta$ then
        Calculate $Z_{i,j}$ with Eq. (18);
        $\tilde{Z}_{i,j} \leftarrow \mathcal{A}_{i,j}(Z_{i,j}; \theta_{i,j})$, $\tilde{Z}_u \leftarrow \mathcal{A}_u(Z_u; \theta_u)$;
        $Z_{fusion} \leftarrow [\tilde{Z}_{i,j}, \tilde{Z}_u]$ ;
    else
        // extract high-level features
        $\tilde{Z}_\alpha \leftarrow \mathcal{A}_\alpha(Z_\alpha; \theta_\alpha)$, $\tilde{Z}_\beta \leftarrow \mathcal{A}_\beta(Z_\beta; \theta_\beta)$, $\tilde{Z}_\gamma \leftarrow \mathcal{A}_\gamma(Z_\gamma; \theta_\gamma)$;
        $Z_{fusion} \leftarrow [\tilde{Z}_\alpha, \tilde{Z}_\beta, \tilde{Z}_\gamma]$ ;
    end
end
return $Z_{fusion}$
```

α and β to modality γ. The operation is formally described as follows:

$$F_\gamma = \text{ReLU}(\text{BN}(Concat(C_{\alpha\rightarrow\gamma}, C_{\beta\rightarrow\gamma}, Z_\gamma)). \tag{19}$$

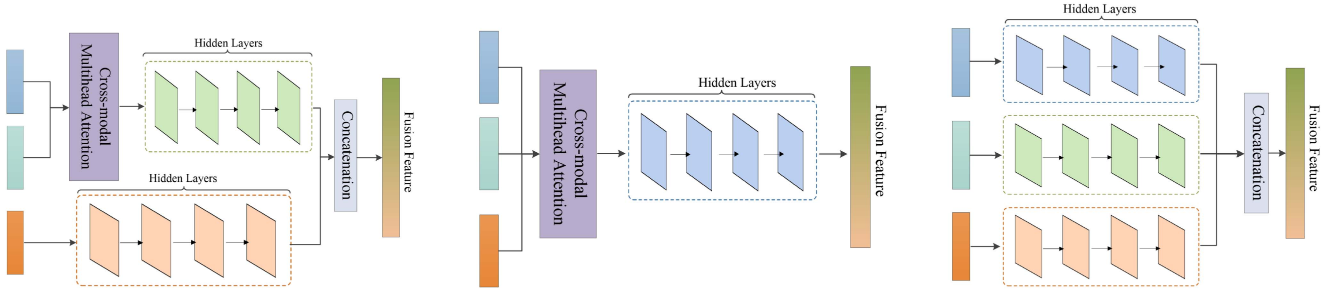
Similarly, we can obtain the fused feature F_β, and the cross-modal feature F_α. The fused feature of each modality contains the low-level information received from the other two modalities. The final multimodal fused feature representation $Z_{\alpha,\beta,\gamma}$ is the concatenation of all the fused features, which is shown as follows:

$$Z_{\alpha,\beta,\gamma} = Concat(F_\alpha, F_\beta, F_\gamma). \tag{20}$$

3) Multimodal Fusion: We should note that the features of different modalities are often not independent, but correlated with each other. In normal cases, different modalities are expected to express similar emotional information. The correlation between their features shall be high. The expressed sentiment can be evaluated accurately by fusing these high-level semantic features of different modalities. Meanwhile, the fusion operation of low-level features would bring superabundant information and the final accuracy might be decreased. However, there are also some cases where the sentiments expressed by different modalities are distinct, but the correlation between them is low. For example, facial expressions, voices, and words might convey opposite information by trying to hide the criminal behaviors when one suspect is talking to the investigators. In this case, the model should pay more attention to the detailed information concealed in different modalities, while this information is contained in the low-level features. With the low-level fusion, we can more accurately evaluate the expressed sentiment.

Considering both different-level features and correlations between modalities, MFF integrates the low-level features of low-correlation modalities by utilizing cross-modal multi-head attention, and fuses the high-level features of high-correlation modalities by concatenation, as shown in Fig. 4(a). The two extreme types of MFF are low-layer feature fusion (the low-level features of all modalities are fused at the low layer, as shown in Fig. 4(b)) and upper-layer feature fusion (the high-level features of all modalities are fused at the upper layer, as shown in Fig. 4(c)).

The whole process of MFF is presented in Algorithm 1. We first use DCCA method to calculate the correlation coefficient between different modalities, i.e., $\rho_{\alpha,\beta}$, $\rho_{\alpha,\gamma}$, $\rho_{\beta,\gamma}$, $\rho_{\alpha+\beta,\gamma}$, $\rho_{\alpha+\gamma,\beta}$, and $\rho_{\beta+\gamma,\alpha}$. Let δ be the correlation threshold, used as the selection criterion for the optimal fusion strategy. The value of δ depends on the characteristics of different modalities and is not fixed. We calculate the threshold δ with the average of all the correlation coefficients. Assuming that $\rho_{i+j,u}$ is the maximum one in $\{\rho_{\alpha+\beta,\gamma}, \rho_{\alpha+\gamma,\beta}, \rho_{\beta+\gamma,\alpha}\}$, where $i, j \in \{\alpha, \beta, \gamma\}$, $i \neq j$, and $u = \{\alpha, \beta, \gamma\}$. Then we choose an appropriate fusion strategy by comparing the values of $\rho_{i+j,u}$, $\rho_{i,j}$ with δ. Case 1) If $\rho_{i+j,u} \leq \delta$, we fuse the low-level features of the three modalities by using cross-modal multi-head attention to obtain the fusion feature Z_{fusion}; Case 2) If $\rho_{i+j,u} > \delta$ and $\rho_{i,j} \leq \delta$, we adopt fusion strategy (a) to fuse the modalities; Otherwise, we directly concatenate the high-level features of three modalities to get the fusion feature Z_{fusion}. MFF model fuses the features from different modalities at different layers according to the correlation coefficient between different features. The model can better balance the relationship between modalities and fully explore the important-information interactions, which will be proved in Section V-E2.



(a) Different-level feature fusion (b) Low-level feature fusion (c) High-level feature fusion

Fig. 4. Multi-layer feature fusion.

TABLE I
THE DETAILED STATISTICS OF SIMS AND CMU-MOSI DATASETS

Datasets	#Train	#Valid	#Test	Feature dimensions	Sequence lengths	Range of labels
SIMS	1368	456	457	768/709/33	39/55/400	[-1, 1]
CMU-MOSI	1284	229	686	768/47/74	50/375/500	[-3, 3]

D. Sentiment Analysis

In this paper, we aim to address multimodal sentiment analysis tasks, which is to combine the information from multiple modalities to determine the expressed sentiment. Through the abovementioned processes, we have obtained the multimodal fused feature Z_{fusion} generated by MFF. Last, The fused feature Z_{fusion} passes through the fully connected layer for the final sentiment prediction.

$$\hat{Y} = WZ_{fusion} + b, \tag{21}$$

where W and b represent the projection parameters and the bias, respectively. We use L2 regularization to constrain these parameters to improve the generalization ability.

V. EXPERIMENTS AND ANALYSIS

In this section, extensive experiments are conducted to evaluate the performance of the proposed model in this paper, with all experiments involving three modalities of completely different natures: text, video, and audio.

A. Datasets

In this work, we use the two most widely used datasets in multimodal sentiment analysis, SIMS [38] and CMU-MOSI [39]. Here, we give a brief introduction to the above datasets.

SIMS dataset collects 60 raw videos from different movies, TV series, and variety shows, and a total of 2,281 video clips are selected from these videos after frame-level cropping. The average length of video clips is 3.67 seconds. Each video clip contains text, video, and audio modalities, and human annotators label each clip with a sentiment score from -1 (strongly negative) to 1 (strongly positive).

CMU-MOSI dataset is a collection of opinion videos on the online website YouTube, which contains 2,199 video clips sliced from 93 videos. The average length of video clips is 4.2 seconds. Each video clip inherently contains acoustic, visual, and textual modalities, while manually annotated with a continuous sentiment score from -3 (strongly negative) to 3 (strongly positive).

The detailed statistics of SIMS and CMU-MOSI datasets are shown in Table I. The contents in the third and fourth columns (from left to right) represent the feature dimensions and sequence lengths corresponding to text, video, and audio modalities. We use the same split dataset for all the models in our experiments.

B. Preprocessing

For fair comparisons with other baselines, we extract the features from textual, visual, and acoustic raw data as introduced below. We use these pre-processed features as the input for all models in our experiments.

Text Features. We utilize the pre-trained BERT [40] as the feature extractor for text. The raw sentence is fed into the encoder to generate contextualized word embeddings as the input of text modality. Because BERT has achieved state-of-the-art performance in natural language processing, and thus we replace the operations of ATCN in (2)–(4) with $Z_t{=}BERT(X_t; \theta_t^{bert})$ when processing the text modality.

Visual Features. For SIMS dataset, we extract frames from the video segments at 30 Hz, and OpenFace2.0 toolkit [41] is used to extract 709-dimensional visual features, including the 68 facial landmarks, 17 facial action units, head pose, and orientation, and eye gaze. For CMU-MOSI dataset, Facet [42], an analytical tool based on the Facial Action Coding Systems (FACS), is used to extract 47-dimensional facial expression features, which include facial action units, facial landmarks, head pose, gaze tracking, and HOG features.

Audio Features. For SIMS dataset, LibROSA speech toolkit [43] is used to extract acoustic features at 22050 Hz. We totally extract 33-dimensional frame-level acoustic features, which include log F0, MFCCs, and CQT. These features are related to emotions and tone of speech. For CMU-MOSI dataset,

TABLE II
HYPER-PARAMETER SETTINGS OF THE MODEL

Hyper-parameter	CMU-MOSI	SIMS
Learning_rate	0.00001	0.0001
Batch size	64	64
Att_dropout	0.2	0.2
Head_number	10	10
d_u	50	50
Dropout(Output)	0.2	0.4
Dimension of Z_{fusion}	100	128
Epochs	150	100

COVAREP [44], a professional acoustic analysis framework, is used to extract 74-dimensional acoustic features, which include multiple important features for determining emotion in speech, such as pitch, VUV, NAQ, MFCCs, and MDQ.

Modality Alignment. We use the forced alignment tool P2FA [45] to align the text with visual and acoustic signals at the word level. The tool automatically aligns the visual and acoustic segments with the corresponding word of text according to the standard pronunciation. And we then perform averaging on the visual and acoustic features within segments. Note that our model can also be directly applied to unaligned multimodal signals.

C. Experimental Setup

In this paper, we build and train all the models based on the NVIDIA TeslaP100 GPU platform and the python3.7 + pytorch1.4.0 + cuda11.2 deep learning framework. In our experiments, the Adam optimizer is used to optimize model parameters. With differences between CMU-MOSI and SIMS datasets in mind, we set two sets of hyper-parameters respectively, as shown in Table II, in which d_u represents the dimension of cross-modal multi-head attention output and Z_{fusion} represents the final fused feature.

D. Baselines and Metrics

We compare our results with several advanced multimodal fusion algorithms as detailed below.

(1) *Early Fusion LSTM (EF-LSTM)* [7]. This model concatenates the input textual, acoustic, and visual features at each timestamp, and builds an LSTM network to construct a multimodal representation. The last hidden state of the LSTM is taken to predict the sentiment.

(2) *Later Fusion DNN (LF-DNN)* [46]. This model builds a DNN for textual, acoustic, and visual inputs separately, and concatenates the last hidden state of the three DNNs as the multimodal feature which is taken to predict the sentiment.

(3) *The Tensor Fusion Network (TFN)* [10]. This approach concurrently models intra-modality and inter-modality dynamics, which calculates a multi-dimensional representation tensor by Cartesian product.

(4) *Low-rank Multimodal Fusion (LMF)* [11]. It is an improvement over TFN, which decomposes high-order tensors into a set of low-rank factors, where a low-rank multimodal tensor fusion technique is used to improve efficiency.

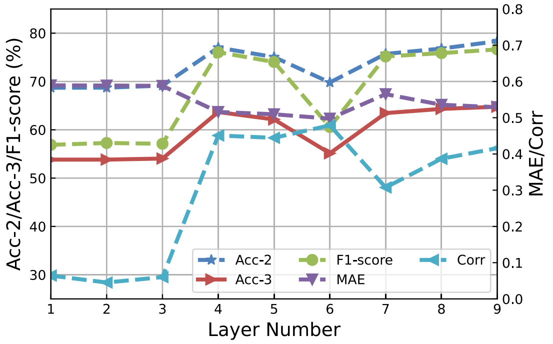


Fig. 5. Impact of ATCN with different number of layers *K* for modeling video sequences on SIMS dataset.

(5) *Memory Fusion Network (MFN)* [9]. It utilizes LSTMs and Delta-memory Attention Network to model the view-specific and cross-view interactions, then summarizes them over time with a Multi-view Gated Memory.

(6) *Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA)* [47]. It projects each modality into two different subspaces and incorporates a combination of losses, including distributional similarity loss, orthogonal loss, reconstruction loss, and task prediction loss.

(7) *Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM)* [48] It designs a label generation module based on the self-supervised learning strategy to acquire unimodal representations, and designs a weight-adjustment strategy to balance the learning progress among different subtasks. In this paper, two types of evaluation metrics, i.e., regression and classification, are employed to evaluate the model's performance. Regression metrics include mean absolute error (MAE) and Pearson correlation (Corr). Classification metrics include seven-class accuracy (Acc-7), three-class accuracy (Acc-3), binary-class accuracy (Acc-2), and F1-score. For Acc-2 and F1-score, which is regarded as positive with the score ≥ 0, and negative with score < 0. For all metrics, the higher value indicates better performance except MAE.

E. Analysis and Results

1) Impact of ATCN With Different Hidden Layers: To explore the impact of ATCN with different numbers of hidden layers *K* on unimodal data sequence, we have carried out a set of experiments on SIMS and CMU-MOSI datasets, in which *K* is set from 1 to 9, respectively.

Intuitively, the wider receptive field allows ATCN to capture long-term dependencies and obtain more meaningful information from the network input. However, excessive unnecessary parameters can easily lead to overfitting and degraded performance, especially when too many layers are in the network. Figs. 5 and 6 show the experimental results on SIMS dataset. As we can see from Fig. 5, the accuracy with four layers is approximate to that with nine layers when measured with the same metrics, such as Acc-2, Acc-3, F1-score, MAE, and Corr.

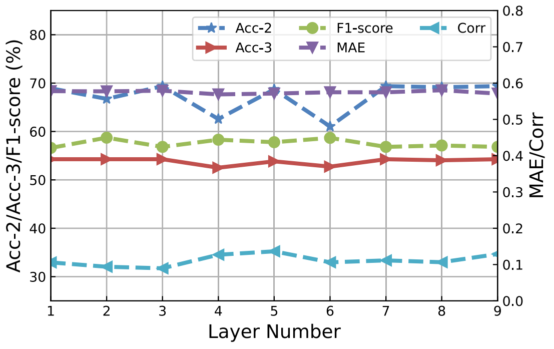


Fig. 6. Impact of ATCN with different number of layers *K* for modeling audio sequences on SIMS dataset.

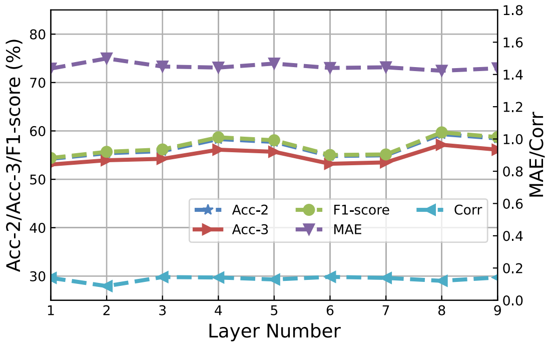


Fig. 7. Impact of ATCN with different number of layers *K* for modeling video sequences on CMU-MOSI dataset.

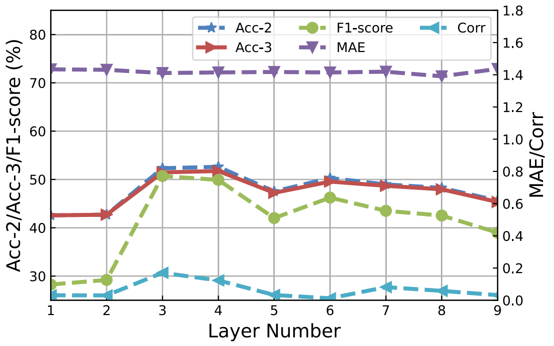


Fig. 8. Impact of ATCN with different number of layers *K* for modeling audio sequences on CMU-MOSI dataset.

The trend in Fig. 6 is not so obvious. MAE, Acc-3, and Corr have the best performance when $K = 3$, while F1 and Acc-2 show slightly better performance on $K = 4$. For a balance between the accuracy and computational cost, we choose $K = 4$ for video sequences and $K = 3$ for audio sequences as the number of hidden layers in ATCN for the following experiments.

Note that the sequence lengths and feature dimensions of SIMS and CMU-MOSI datasets are distinct from each other, we conduct similar experiments on CMU-MOSI dataset. The experimental results are shown in Figs. 7 and 8. We can find that the accuracy with eight layers performs best in Fig. 7. In Fig. 8, the accuracy rates on $K = 3$ and $K = 4$ are very close. Considering the balance between computational cost and accuracy, we choose $K = 8$ for video sequences and $K = 3$ for audio sequences.

To further evaluate the effectiveness of ATCN in sequence modeling, we compare it with traditional models, such as LSTM, BiLSTM, GRU, and BiGRU. Fig. 9 shows the results on SIMS dataset. It can be found that ATCN outperforms all the others when modeling the video sequences. ATCN is superior to others with metrics Acc-2, Acc-3, and MAE, and its performance with F1-score and Corr is second satisfactory when modeling the audio sequences. Similar cases can be found on CMU-MOSI dataset, as shown in Fig. 10. The superiority may be that ATCN can capture the global information of different sequential data, and extract effective unimodal temporal features by capturing the long-term dependencies.

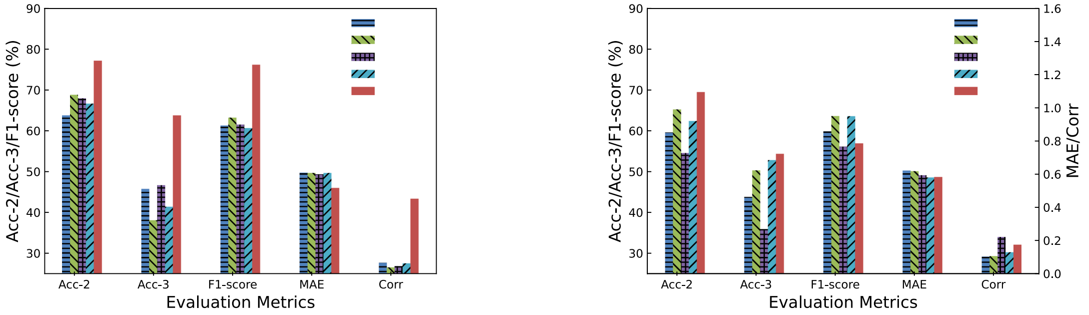
2) Impact of Modality Number: Table III shows the experimental results on two different multimodal sentiment analysis datasets, SIMS and CMU-MOSI, from which we can have the following observations.

First, in the case of unimodal, the text provides more accurate emotion recognition results compared with the audio and video. We attribute the reason to the fact that the text should contain more information. In contrast, audio and video are generally unfiltered/original signals, which are likely to lose some information due to the potential interference from the susceptible environment. In this case, the sentiment analysis results with text are intended to be more accurate than the others.

Secondarily, the fusion results with bimodal are not certainly better than that with unimodal. As mentioned, the accuracy of text is higher than that of audio and video. Combining the text with audio or video separately does not bring improved accuracy, and the combination might even reduce the final accuracy. For example, on SIMS, the accuracy with modality T and A+T is 79.65% and 79.43%, respectively, which means the accuracy is slightly improved after fusion; on CMU-MOSI, the accuracy is 80.32% and 81.34%, respectively, which means the accuracy is reduced after fusion.

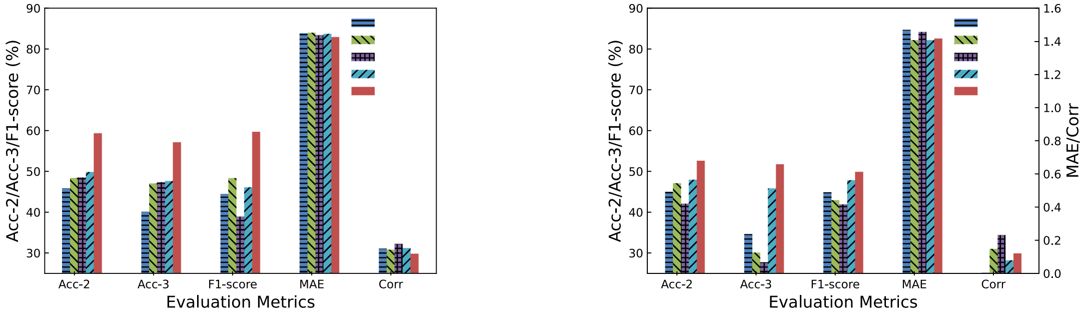
Finally, the trimodal helps to improve the accuracy of the sentiment analysis significantly. In general, with the increase in the number of fused modalities, the performance should be improved since there is complementary information among different modalities. Multimodal fusion is conducive to improving the accuracy with full utilization of this information. On the other hand, it is not guaranteed that trimodal fusion can achieve better results, or is inferior to unimodal in some cases. It is because trimodal brings more related but redundant information which is possible to confuse the model and affect the accuracy in turn. It is critical to select a suitable fusion strategy when we design an effective fusion algorithm.

3) Impact of Different Fusion Strategies: We have conducted a group of experiments to compare the performance of different fusion strategies with trimodal on sentiment analysis. *a)* The low-level detailed features of three different modalities are fused, namely, A+V+T; *b)* The high-level semantic features of three modalities are fused, namely, A $\oplus$ V$\oplus$T; *c)* fuse the low-level detailed features of two modalities, and then fuse the high-level semantic features with the last modality. The last case can be further divided into three types, i.e., (A+V)$\oplus$T, (V+T)$\oplus$A, (A+T)$\oplus$V. The operation + means low-level feature fusion, and $\oplus$ means high-level feature fusion.



(a) Video modality (b) Audio modality

Fig. 9. Performance of different sequence models on SIMS datasets.



(a) Video modality (b) Audio modality

Fig. 10. Performance of different sequence models on CMU-MOSI datasets.

TABLE III
RESULTS ON SIMS AND CMU-MOSI DATASETS

Task type		SIMS					CMU-MOSI				
		Acc-2	Acc-3	F1-score	MAE	Corr	Acc-2	Acc-3	F1-score	MAE	Corr
Unimodal	A	69.37	54.27	56.82	0.5791	0.1290	52.62	51.75	49.91	1.4145	0.1233
	V	77.02	63.68	76.06	0.5159	0.4510	59.33	57.14	59.70	1.4231	0.1267
	T	79.43	66.74	79.06	0.4077	0.6020	81.34	78.57	81.43	0.7550	0.7758
Bimodal	A+V	77.41	64.69	76.88	0.4979	0.4881	59.74	58.88	56.91	1.0566	0.6115
	A+T	79.65	65.21	79.21	0.4369	0.5782	80.32	77.99	80.50	0.9546	0.7555
	V+T	80.53	66.30	80.20	0.4343	0.5790	81.22	78.17	81.04	0.9385	0.7674
Trimodal	A+V+T	80.09	65.65	79.55	0.4445	0.5785	82.10	79.04	81.88	0.9324	0.7841
	(A+V)⊕T	**81.40**	**68.71**	**80.56**	**0.4318**	**0.6038** 👍	79.04	77.87	78.91	0.9719	0.6789
	(V+T)⊕A	80.12	67.40	80.28	0.4470	0.5900	81.66	78.60	81.69	0.8331	0.7865
	(A+T)⊕V	78.99	65.21	76.75	0.4401	0.5700	**83.97**	**79.30**	**83.00**	**0.7524**	**0.7789** 👍
	A⊕V⊕T	79.24	69.37	79.67	0.4362	0.5977	73.18	70.70	73.39	1.1038	0.5656

The experimental results show that (A+V) ⊕ T performs best on SIMS dataset, and (A+T)⊕V performs best on CMU-MOSI dataset. These results indicate that the multi-layer feature fusion model can achieve better performance on different datasets by comparing the trimodal with unimodal and bimodal. But it should be pointed out that not all multi-layer feature fusion strategies will produce rewarding results. For example, the performance of (A+T)⊕V on SIMS dataset and (A+V)⊕T on CMU-MOSI dataset is less impressive than that of T or V+T. We speculate that the effect of multi-layer feature fusion strategies might depend on the feature correlation between different modalities. Therefore, we have proposed a novel MFF model, which can adaptively select an appropriate fusion strategy regarding the relationship between the correlation and the given threshold.

4) Impact of Threshold Value: The proposed MFF model would choose a proper fusion strategy by comparing the correlation coefficient with the threshold. We illustrate this idea with

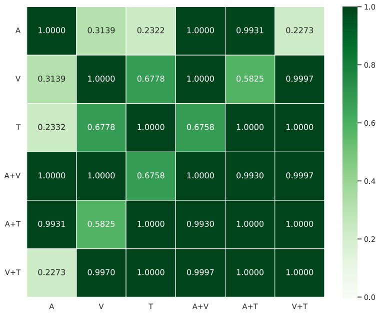


Fig. 11. Correlation matrix between the various modalities of SIMS dataset.

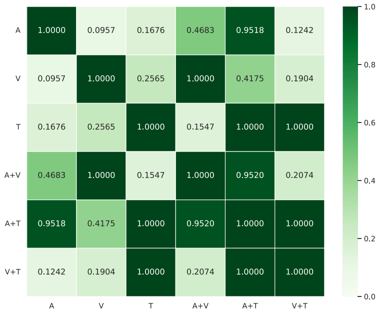


Fig. 12. Correlation matrix between the various modalities of CMU-MOSI dataset.

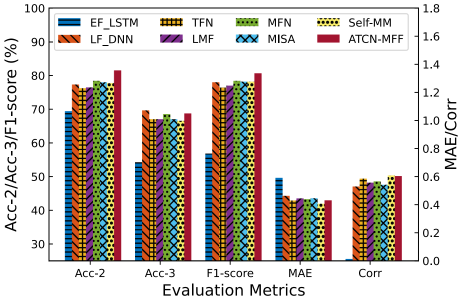


Fig. 13. Performance of different multimodal models on SIMS dataset.

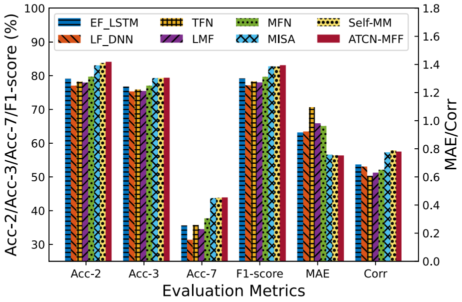


Fig. 14. Performance of different multimodal models on CMU-MOSI dataset.

TABLE IV
IMPACT OF THRESHOLD ON SIMS DATASET

Range of δ	Selected strategy
(0, 0.2273]	A⊕V⊕T
(0.2273, 0.2322]	A⊕V⊕T
(0.2322, 0.3139]	A⊕V⊕T
(0.3139, 0.5825]	(A+V)⊕T 👍
(0.5825, 0.6758]	(A+V)⊕T 👍
(0.6758, 0.6778]	A+V+T
(0.6778,1)	A+V+T

TABLE V
IMPACT OF THRESHOLD ON CMU-MOSI DATASET

Range of δ	Selected strategy
(0, 0.0957]	A⊕V⊕T
(0.0957, 0.1242]	A⊕V⊕T
(0.1242, 0.1547]	A⊕V⊕T
(0.1547, 0.1676]	A⊕V⊕T
(0.1676, 0.2565]	(A+T)⊕V 👍
(0.2565, 0.4175]	(A+T)⊕V 👍
(0.4175,1)	A+V+T

SIMS and CMU-MOSI datasets. The correlation coefficients between different modalities can be found in Figs. 11 and 12. As is shown, $\rho_{A,V}$, $\rho_{A,T}$, $\rho_{V,T}$, $\rho_{A+V,T}$, $\rho_{A+T,V}$, and $\rho_{V+T,A}$, are significantly different on the two datasets. In this way, a fixed threshold is not suitable for different datasets, and its value should be selected by explicitly displaying the difference in the correlation coefficients.

The threshold plays an important role in the proposed MFF model. If the value is too small, the measured correlations are all intended to be larger than the threshold, and thus the chosen fusion strategy is the high-level feature fusion, i.e., A $\oplus$ V$\oplus$T. In this way, the MFF model degenerates into the traditional concatenation among all different modalities with a big loss in terms of accuracy. On the other hand, if the chosen value is too large, the measured correlations are smaller than the threshold, and the A+V+T strategy is chosen, which fuses the low-level features of all modalities with cross-modal multi-head attention. This strategy will increase the model's computational cost, and it may also bring lots of redundant information into the model, affecting the final performance adversely.

In order to verify the impact of the threshold on the MFF model, we sort the correlation coefficients of all possible modal combinations from small to large, and obtain possible ranges of the threshold which might influence the strategy selection. As we can see from Tables IV and V, the selected strategies are different with distinct thresholds. It shows that the threshold has a crucial impact on the final performance of the model, because its value can decide which kind of fusion strategy (low-level feature fusion, high-level feature fusion, or different-level feature fusion) is chosen.

In this paper, we choose the average of all these measured correlations, i.e., $\rho_{A,V}$, $\rho_{A,T}$, $\rho_{V,T}$, $\rho_{A+V,T}$, $\rho_{A+T,V}$, and $\rho_{V+T,A}$, as the threshold. The value equals 0.4516 and 0.2027 separately on SIMS and CMU-MOSI datasets. It can be found that the threshold is matched to the optimal fusion strategy on both datasets. Specifically, we can find that the correlation between A+V and T, $\rho_{A+V,T}$, is the maximum one in $\{\rho_{A+V,T}$, $\rho_{A+T,V}$ $\rho_{V+T,A}\}$, and $\rho_{A+T,V}$ is the higher than the threshold and $\rho_{A,V}$ is lower than the threshold on SIMS dataset. In this case, the final chosen strategy, (A+V)⊕T, has significantly better performance over the others (as shown in Table III). This result verifies the rationality of the MFF model designed in this paper. In general, the proposed MFF will choose the average of all possible combinations as the threshold. It means that the optimal fusion strategies selected in different datasets may be completely different, which indicates that the MFF model has certain adaptability.

5) Results Analysis of Multimodal Sentiment: In this paper, we first compare the results of the proposed ATCN-MFF model with other important multimodal sentiment analysis models on SIMS dataset. As indicated in Fig. 13, compared with other baseline algorithms, ATCN-MFF has achieved improvements with Acc-2, F1-score, and Corr. Especially, Acc-2 is improved by about 4% compared with MFN, and F1-score is improved by more than 2% compared with Self-MM. We assume that the reason behind it may be two-fold. *a*) ATCN can deal with the long-term dependencies between data at different moments within a unimodal sequence, resulting in meaningful and modality-specific feature representations. *b*) The proposed MFF can learn the interactions among modalities effectively. By adaptively selecting the most appropriate multi-layer fusion strategy according to the correlations between modalities, MFF can make full use of the complementary information among modalities while reducing the redundancy as well as the information interference, and finally effectively improve the accuracy in sentiment analysis.

We have also carried out the performance analysis on the ATCN-MFF model on CMU-MOSI dataset. As we can see from Fig. 14, the proposed ATCN-MFF model achieves the best results on almost all metrics compared with other baseline algorithms. For example, Acc-2 is improved by about 1% compared with MISA, and we also attain a slight improvement compared with the state-of-art Self-MM. The experiments on both SIMS and CMU-MOSI datasets show that the proposed model achieves better performance, which demonstrates the effectiveness of our model.

VI. Discussion

Sentiment is a complex psychological state that includes subjective experiences, physiological responses, and behavioral or expressive responses. A person's sentiment state can be recognized through external signals such as facial expressions, text, audio, and body movements, as well as internal signals such as EMG, EEG, respiratory rate, and heart rate. Our work only involves three different modalities, it can be easily extended to *N* different modalities. But its computational complexity will exponentially increase with the number of modalities, and lead to high computational costs. However, it is worth noting that the fusion of more modalities cannot always produce better results in many cases. Although multimodal data contains more information than a single modality, extensive modalities bring lots of redundant information which easily confuses the model and adversely affects performance. Most multimodal sentiment analysis works focus on how to effectively integrate image, text, and audio modalities, and some works combine audio, image, and physiological signals for sentiment analysis. These works usually turned out fruitful. The study on the more effective fusion of three different modalities in this paper has certain values in practical applications.

VII. Conclusion

Multimodal sentiment analysis requires models that capture both specific characteristics within each modality and interactions among multiple modalities. To better capture the long-term dependencies within sequences, this paper introduces ATCN to extract unimodal temporal features. We also propose an MFF model that fuses the different modal features by different methods at different layers according to the correlation coefficients between modalities. This method can make multimodal fusion more effective and enhance the performance of sentiment analysis. In the model, cross-modal multi-head attention is used to deeply explore the interactions between low-level detailed information from different modalities, and concatenating the high-level features can reduce the interference induced by redundant information. Experimental results on both SIMS and CMU-MOSI multimodal datasets show that our model outperforms the current state-of-the-art approaches.

This work just focuses on how to improve the accuracy of sentiment analysis effectively by combining multimodal data, instead of the computational efficiency. In future work, we plan to study how to effectively improve the computational efficiency of the model while increasing the accuracy for the sentiment analysis problem, and try to find a balance between the computational efficiency and accuracy. Additionally, the interactions and correlations between different modalities are very dynamic and complex, we intend to find an effective multimodal model that can dynamically adjust the fusion strategy from time to time and from one video clip to another, and verify the feasibility on diverse datasets.

Acknowledgments

We would like to express sincere thanks to the anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript.

References

[1] Z. Ren, Z. Wang, Z. Ke, Z. Li, and S. Wushour, "Survey of multimodal data fusion," *Comput. Eng. Appl.*, vol. 57, no. 18, pp. 49–64, 2021.

[2] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surveys*, vol. 47, no. 3, pp. 1–36, 2015.

[3] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov./Dec. 2018.

[4] Y. Peng and J. Qi, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1, pp. 1–24, 2019.

[5] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Third Quarter 2020.

[6] Y. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[7] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal DNN feature fusion," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 11–19.

[8] 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, First Quarter 2022.

[9] A. Zadeh et al., "Memory fusion network for multi-view sequential learning," in *Proc. 32th AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[10] A. Zadeh et al., "Tensor fusion network for multimodal sentiment analysis," in *Proc. 22th Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[11] L. Zhun et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[12] R. R. Murphy, "Computer vision and machine learning in science fiction," *Sci. Robot.*, vol. 4, no. 30, pp. 7221–7235, 2019.

[13] S. Bhojanapalli et al., "Low-rank bottleneck in multi-head attention models," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 864–873.

[14] Z. Chen, A. Feng, and H. E. Jia, "Text sentiment classification based on 1D convolutional hybrid neural network," *J. Comput. Appl.*, vol. 39, no. 7, pp. 1936–1941, 2019.

[15] Z. Zhao, R. Rao, S. Tu, and J. Shi, "Time-weighted LSTM model with redefined labeling for stock trend prediction," in *Proc. IEEE 29th Int. Conf. Tools Artif. Intell.*, 2017, pp. 1210–1217.

[16] R. Zhang, Z. Yuan, and X. Shao, "A new combined CNN-RNN model for sector stock price analysis," in *Proc. IEEE 42th Annu. Comput. Softw. Appl. Conf.*, 2018, pp. 546–551.

[17] P. Lei and S. Todorovic, "Temporal deformable residual networks for action segmentation in videos," in *Proc. IEEE 36th Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6742–6751.

[18] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE 45th Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6875–6879.

[19] Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang, "Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 565–578, Third Quarter 2021.

[20] G. Evangelopoulos et al., "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[21] E. Shutova, D. Kiela, and J. Maillard, "Black holes and white rabbits: Metaphor identification with visual features," in *Proc. 15th Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, 2016, pp. 160–170.

[22] E. Morvant, A. Habrard, and A. Stéphane, "Majority vote of diverse classifiers for late fusion," in *Proc. Joint IAPR Int. Workshops Statist. Techn. Pattern Recognit. Struct. Syntactic Pattern Recognit.*, Berlin, Heidelberg: Springer, 2014, pp. 153–162.

[23] S. Peng, C. Lin, J. Tan, and L. Hsu, "The g-good-neighbor conditional diagnosability of hypercube under PMC model," *Appl. Math. Comput.*, vol. 218, no. 21, pp. 10406–10412, 2012.

[24] F. Hen et al., "Complementary fusion of multi-features and multi-modalities in sentiment analysis," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 82–99.

[25] J. He, S. Mai, and H. Hu, "A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis," *IEEE Signal Process Lett.*, vol. 28, no. 5, pp. 992–996, May 2021.

[26] D. Gkoumas et al., "What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis," *Inf. Fusion*, vol. 66, pp. 184–197, 2021.

[27] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "A cognitive brain model for multimodal sentiment analysis based on attention neural networks," *Neurocomputing*, vol. 430, no. 10, pp. 159–173, 2021.

[28] G. Xiao et al., "Multimodality sentiment analysis in social Internet of Things based on hierarchical attentions and CSAT-TCN with MBM network," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12748–12757, Aug. 2021.

[29] X. Xue, C. Zhang, Z. Niu, and X. Wu, "Multi-level attention map network for multimodal sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 5105–5118, May 2023, doi: 10.1109/TKDE.2022.3155290.

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

[31] F. Liu, L. Zheng, and J. Zheng, "HieNN-DWE: A hierarchical neural network with dynamic word embeddings for document level sentiment classification," *Neurocomputing*, vol. 403, pp. 21–32, 2020.

[32] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade EF-GAN: Progressive facial expression editing with local focuses," in *Proc. IEEE 38th Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5021–5030.

[33] Q. T. Truong and H. W. Lauw, "VistaNet: Visual aspect attention network for multimodal sentiment analysis," in *Proc. 33th AAAI Conf. Artif. Intell.*, 2019, pp. 1–9.

[34] K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "A novel context-aware multimodal framework for persian sentiment analysis," *Neurocomputing*, vol. 457, no. 2, pp. 377–388, 2021.

[35] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, First Quarter 2016, doi: 10.1109/TAFFC.2015.2436926.

[36] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019, doi: 10.1109/TCYB.2018.2797176.

[37] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[38] W. Yu et al., "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.

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

[40] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 18th North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[41] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.

[42] P. Ekman and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. London, U.K., Oxford Univ. Press, 2005.

[43] B. McFee et al., "Librosa: Audio and music signal analysis in python," in *Proc. 14th Scipy. Conf.*, 2015, pp. 18–25.

[44] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREPA collaborative voice analysis repository for speech technologies," in *Proc. IEEE 40th Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 960–964, doi: 10.1109/ICASSP.2014.6853739.

[45] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *J. Acoustical Soc. Amer.*, vol. 5, pp. 3878–3878, 2008.

[46] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguistics*, 2018, vol. 2, pp. 606–611, doi: 10.18653/v1/P18-2096.

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

[48] W. Yu et al., "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.

Hongju Cheng (Member, IEEE) received the BE and ME degrees in EE from the Wuhan University of Hydraulic and Electric Engineering, in 1997 and 2000, respectively, and the PhD degree in computer science from Wuhan University, in 2007. He is currently a professor with the College of Computer and Data Science, Fuzhou University. His interests include Internet of Things, mobile ad hoc networks, wireless sensor networks, and wireless mesh networks. Since 2007, he has been with the College of Computer and Data Science, Fuzhou University, Fuzhou, China. He is serving as editor/guest editor for more than 10 international journals. He has published almost 80 papers in international journals and conferences.

Zizhen Yang received the BS degree in computer science and technology from East China Jiaotong University, China, in 2018. She is currently working toward the MS degree with Fuzhou University, China. Her research includes multimodal fusion.

Xiaoqi Zhang received the BE degree in network engineering from Fuzhou University, China, in 2016, and the ME degree in software engineering from Heilongjiang University, China, in 2019. She is currently working toward the PhD degree with Fuzhou University, China. Her research includes Internet of Things, intelligent edge computing, cognitive radio sensor networks and mobile ad hoc networks.

Yang Yang received the BSc degree from Xidian University, Xi'an, China, in 2006, and the PhD degrees from Xidian University, China, in 2011. She is currently a full professor with the College of Computer and Data Science, Fuzhou University. Her research interests are in the area of information security and privacy protection. She is also a research fellow (postdoctor) with the School of Information System, Singapore Management University. She has published papers on *IEEE Transactions on Information Forensics and Security*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE Transactions on Services Computing*, *IEEE Transactions on Cloud Computing*, *IEEE Transactions on Industrial Informatics* etc.