Information Fusion 123 (2025) 103230

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

CCIN-SA:Composite cross modal interaction network with attention enhancement for multimodal sentiment analysis

Li Yang *, Junhong Zhong, Teng Wen, Yuan Liao

School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu, 610500, China

ARTICLE INFO

Keywords:
Cross-modal
Attention enhancement
Multimodal fusion
Bi-GRU
Multimodal sentiment analysis

ABSTRACT

The main challenges of multimodal sentiment analysis are unimodal feature extraction techniques and multimodal feature fusion strategies. Considering different inherent sampling rates of temporal information between different modalities, cross-modal context dependency, and weak inter-modal interaction caused by semantic variability, this study proposed a composite cross modal interaction fusion network with attention enhancement for multimodal sentiment analysis (CCIN-SA). First of all, CNN, bidirectional gated recurrent neural network, and multi-modal attention mechanism were used to extract unimodal feature information, unique features within the modality, and sequence timing information. Second, an enhanced cross-modal interaction framework was developed, which persistently fortified the primary modality by integrating lower-level cues from the supplementary modality. Through this process, the primary modality absorbed knowledge from the other modality and identified inter-modal potential adaptations. The enhanced interaction features were inputted into the interaction layer of the gating mechanism network for fusion, and the similarity between two modalities was calculated by using the two cross-modal interaction features as condition vectors. The association degree of important features between the modalities was enhanced, and the association of secondary features was weakened, so as to capture deeper cross modal interaction. Finally, composite hierarchical fusion method was taken to splice and fuse information-enhanced cross-modal joint features and lower-order signals through residual network structure and multi-head attention mechanism. Probability distribution matrices were calculated to enhance the weight of important features within modalities in a weighted manner, so that the final results could be applied to the classification task. Benchmark model comparison experiments were carried out under CMU-MOSI and CMU-MOSEI datasets. The results demonstrate that the accuracy and F1 score were improved and increased by 1% in this study. With better performance in all metrics, the model can explore potential interaction among multimodalities and reflect the advancement and effectiveness of the model.

1. Introduction

Textual sentiment analysis focuses on textual modality. In view of the existing sentiment information sources, sentiment expression presented by a single modality has certain defects and limitations. Therefore, the study of multimodal sentiment analysis (MSA) is attracting a lot of researchers. Video and auditory data can compensate for the absence of textual data. By identifying the underlying relationship between text, images, and speech, they can collaboratively address challenges in sentiment analysis [1,2]. At present, the main problem of MSA is that the model's effectiveness is largely contingent upon the effectiveness of unimodal feature extraction techniques and the proficiency of preprocessing methods for these features. However, how to better extract the implicit sentimental features of different modalities is a major problem. More scholars are devoted to the study of multimodal fusion [3–5].

Based on this, scholars have carried out numerous studies on MSA [6–8]. However, different extracting technologies for multimodal features lead to huge differences in temporal sequences and semantic aspect. The acquired unimodal features are unaligned data, and the visual and speech feature information contains a large number of lower-order signals and redundant information, which affects the effective fusion of cross-modal feature information [6]. Therefore, how to find the potential correlation between unaligned data and solve the difficulty in cross-modal fusion caused by information differences is critically important for MSA.

This study proposed a composite cross modal attention enhancement interaction fusion network for MSA (CCIN-SA). Considering the

* Corresponding author.
E-mail address: xnsy_ly@swpu.edu.cn (L. Yang).

https://doi.org/10.1016/j.inffus.2025.103230
Received 26 December 2024; Received in revised form 30 March 2025; Accepted 16 April 2025
Available online 5 May 2025

unique feature information within each modality and contextual information of target sequence, the temporal information of text, speech and visual feature sequences were obtained by using convolution neural network (CNN), bidirectional gated recurrent unit(BiGRU) and multi-head attention mechanism(MHA), respectively. A composite cross-modal interaction network model was established, and feature enhancement between two modalities was realized with the cross-modal attention mechanism. The inter-modality correlations were obtained with the gating mechanism in composite fusion layer, and the weight of important sentiment features in the multimodal fusion layer was improved by self-attention. The final results were applied in sentiment analysis task, which provided a reference for the study of multimodal sentiment computation. Our main contributions to this paper can be summarized as follows:

(1) A novel model for modality-specific temporal feature extraction was presented that combines CNNs, BiGRU, and MHA to explore unique features within each modality and to extract contextual information.

(2) We employ a gating mechanism-improved cross modal attention network model that integrates information enhanced interaction features into the gated mechanism based interaction network to further capture the similarities between different modalities with condition vectors and to enhance the correlation of significant features and to uncover deeper levels of interaction between modalities.

(3) We design a composite multimodal fusion module that employ multi-head attention mechanisms to concatenate and integrate information-enhanced cross-modal joint features with low-order signals, and retain the affective information inherent in diverse modalities.

(4) Extensive tests demonstrate that this study achieves higher ac curacy and F1 score and superior performance across all metrics.

2. Related work

In text mining, Researchers are increasingly focusing on sentiment analysis. Initially, sentiment analysis concentrated on a single text modality based on sentiment features of lexicon [7]. Machine learning method was used to extract word features as text representation to further realize sentiment analysis tasks. Deep learning has been extensively applied in sentiment analysis since its introduction, and there is a great prospect for development [9]. The current sentiment analysis fused the features of multimodality, including the combination of two or three modalities [10]. Scholars have improved the performance of multimodal sentiment analysis by using single-channel feature extraction techniques and multimodal fusion strategies [11].

For the fusion method of different modal information, deep learning-based feature fusion mechanisms have been put forward and widely applied in recent years, and many results have been achieved. Lin et al. [12] developed separate models for extracting features from text and images, utilizing the attention mechanism to focus on the most sentiment-laden regions within images and textual content. However, in this method, the interaction between text and image modalities relied solely on attention weight-based aggregation, without incorporating dynamic gating mechanisms or cross-modal contrastive learning, resulting in constrained modality complementarity. Liu et al. [13] developed context-enhanced LSTM (long short-term memory) approach to obtained inter-modal information through multimodal feature fusion and feature dimensionality reduction. But this method adopted a strategy of first independently modeling unimodal features and then fusing interactive information, which led to the absence of early-stage real-time cross-modal alignment. The multi-tensor fusion network with cross-modal, proposed by Yan et al. involved the extraction of multi-modal features to establish relationships carrying sentiment information across different modalities [14]. However, this method only modeled pairwise inter-modal relationships (e.g., text-image, text-audio) and failed to explicitly capture interactions among triple or higher-order multimodality. Gkoumas et al. [15] segmented the feature fusion into multiple distinct stages. At each stage, various model architectures were employed to capture the interactions between inter-modal information. However, multi-stage fusion pipelines (e.g., CNN followed by RNN) were manually designed, the method lacked adaptive mechanisms to flexibly handle differences in cross-modal interaction strengths. Hou et al. [16] proposed an early fusion method of fusion long short-term memory (EF-LSTM), which spliced three kinds of modal feature information, text, audio and visual, and then captured the long-distance dependencies in the sequences using LSTM. This method was prone to lead to redundant information input. The early-stage feature concatenation approach forced LSTMs to process high-dimensional redundant information (e.g., text word vectors + image pixel features), which impaired long-range dependency modeling efficiency and introduced redundant input noise. Bian et al. [17] developed the BiMNet to thoroughly extract features across diverse time scales and to fuse these features effectively. However, the aforementioned techniques overlooked the differential impact of various modalities, merely aggregating them in the fusion stage without ignoring their individuality, which leaded to a failure in capturing the most effective fusion of information.

At the later stage, the fusion methods focus on fusion at the decision level, which cannot effectively obtain interaction information among multimodalities. How to improve the interaction between modalities has become the main concern of researchers. The Memory Fusion Network (MFN) was put forward by Zadeh et al. [18], which used LSTM to capture view-specific interaction and the Delta attention network to recognize cross-view interaction. Liang et al. [19] proposed the Recurrent Multilevel Fusion Net work (RMFN), which used multilevel fusion to model cross modal interactions. Others focused on multimodal representation using the expressive power of tensor. In order to capture the intra-modal features and the interactions between modalities, Zadeh et al. [20] proposed a tensor fusion network (TFN) to increase the feature dimension and cause the problems of large parameter size, difficulty in training, and low efficiency. Liu et al. [21] introduced a Low-Rank Multimodal Fusion (LMF) technique, employing low-rank tensor decomposition to enhance the efficiency of multimodal data integration. However, this method still encountered issues with parameter proliferation due to lengthy feature vectors. Additionally, there existed several specialized methods for feature fusion. Mai et al. [22] suggested a hierarchical multimodal fusion approach based on a divide-and-conquer strategy. Unlike direct fusion at the global level, this method considered the interplay between local and global interactions in the fusion process. Tsai et al. [23] proposed a Multimodal Factorization Model (MFM) to learn multi modal representations by decomposing modal information into public discriminative factors and specific generative factors. However, MFM's factorization strategy risked losing critical cross-modal correlations by enforcing strict separation of shared and private factors, while hierarchical fusion approaches (e.g., Mai et al.'s method) might disrupt the synergistic interplay between global and local features. Although the above methods explored feature fusion by diverse approaches, they commonly suffered from parameter explosion and efficiency bottlenecks (e.g., TFN's high-dimensional tensors and LMF's limitations with long features) and failed to adequately optimize pre-fusion feature quality (e.g., inadequate coordination of heterogeneity and noise redundancy). Additionally, weak inter-modal interactivity further restricted the effectiveness of cross-modal information fusion.

In recent years, by introducing the idea of attention mechanism, the model is able to focus more on important information, thus improving the performances. Hazarika Hazarika et al. [24] encompassed dual representational spaces for each modality and minimized redundant data while capturing higher-order feature insights, thereby enhancing the multimodal fusion capability. Rahman et al. [25] utilized the pre-processing method BERT (Bidirectional Encoder Representation from Transformers) to receive both visual and acoustic non-textual data during fine-tuning. Wenmeng et al. [26] employed a self-supervised learning strategy for label generation, enabling the simultaneous training of multiple objectives across different tasks. A complex form of

fusion was used for multimodal representation of decision making, which did not consider the effect of different modal sentiment semantic expression strengths on sentiment analysis. Tsai et al. [27] utilized the cross-modal Transformer module to capture the long distance dependencies between different modal elements, thus realizing multimodal fusion. Gan et al. [28] incorporated varying numbers of hidden neurons in multimodal correlation modules. However, the above cross-modal attention methods did not fully consider the contextual in formation of modalities in unimodal feature extraction, and only the fused modal information was considered in sentiment analysis and prediction, without considering the unique features within modalities.

Based on the above analyses, due to the variability of semantic features among different modalities, this study proposed a composite cross-modal attention enhancement interaction network for MSA (CCIN-SA) to realize better modality interaction, solve the difficulty in modality fusion caused by information differences, and improve the accuracy of sentiment classification.

3. Methodology

As shown in Fig. 1, the main structure of the composite cross modal interaction fusion network with attention enhancement for MSA (CCIN-SA) consists of an input layer, unimodal temporal feature extraction layer, improved cross-modal interaction layer (ICA), gating mechanism network layer (GM), and composite multimodal fusion Layer.

(1) Input Layer: modal primary features, text data is extracted, 300-dimensional Glove word embedding are used as text features; for each sentiment polarity audio, CO VAREP speech analysis framework and 74-dimensional speech higher-order statistical features in the speech signal are extracted; Open face facial expression analysis framework is used to obtain visual features,detect speaker's facial action units in each frame, extract 35 muscle movement units, and obtain the dimensionality of text, audio, and visual features of each video sequenced_T=300, d_A =74 and d_V=35.

(2) Unimodal Temporal Feature Extraction Layer: a unimodal temporal feature representation method is proposed by integrating CNNs, BiGRU [29], and MHA to mine unique features inside modality and ex tract contextual information.

(3) Improved Cross-modal Attention Layer (ICA): a multimodal fusion method is employed based on improved cross modal attention mechanism, which utilize slow-order signals of auxiliary modes to continuously strengthen the target modes, so that the target modes can learn the information of auxiliary modes and capture potential adaptations between different modes to realize inter-modal information enhancement.

(4) Gating Mechanism Layer (GM): a bimodal joint feature representation method similar to gating mechanism is designed. With two semantically related modal features as condition vectors, the similarity between modes is calculated through bimodal interaction attention, the weights of important inter-modal correlation features are reinforced, and the interaction between different modes is explored at a deeper level.

(5) Composite Multimodal Fusion Layer: inter-modal interaction through composite hierarchical fusion is continuously strengthened. Considering inter-modal consistency, specific information inside each modality is captured, and the influence of noise in low-order feature in formation is effectively reduced. The probability distribution matrix is calculated through self-attention mechanism to enhance more important feature information within the modality in a weighted manner. The resulting multimodal sentiment features are used for the final sentiment analysis task.

3.1. Input layer

Word Embedding is a vector representation that uses all the feature dimensions, implying more semantic relevance and potential relationships between each word. This study used Glove Word Embedding method. The text modal (Text, T) used 300-dimensional Glove word embeddings as the text feature d_T. By mapping the text vocabulary t_i to the corresponding embedding vector $V_i \in \mathbb{R}^{d_w}$, the embedding matrix $G \in \mathbb{R}^{d_w * |V|}$ was obtained, where d_w is the dimension of the text word embedding and |V| is the size of the vocabulary.

For speech signals, this study used Collaborative Voice Analysis Repository (COVAREP) [30] to extract speech features, as shown in Fig. 2. The repository extracted more than 30 sentimentally relevant speech features through pitch tracking and polarity detection, spectral envelope analysis, resonance peak tracking, sinusoidal modeling, gating analysis, and phase processing, including pitch, turbulence/clearness segmental features, gating source coefficients, energy, normalized amplitude quotient (NAQ), Mel frequency cepstrum coefficient, maximum dispersion quotient, peak slope, and energy slope. Speech modality (Audio, A) was applied to extract speech higher-order statistical features for tone and mood of each sentimentally polarized audio d_A and 74-dimensional speech features d_A in the speech signal through the COVAREP.

For visual features in video data (Video, V), the picture frames were extracted by segmenting the video clips at a frequency of 30 Hz. Multitask convolution neural network (MTCNN) extracted the alignment and detected face. The face part was cropped and saved. The MultiComp OpenFace2.0 toolkit was applied to extract feature information and 35-dimensional visual features d_V=35 were obtained.

In this study, unimodal features in video sequences were obtained through the Multimodal SDK [31] provided by Carnegie Mellon University (CMU). For text data, 300-dimensional Glove word embeddings were used as text features. For each sentimentally polarized audio, the COVAREP speech analysis framework extracts 74-dimensional speech higher-order statistical features in speech signal. Openface facial expression analysis framework was adopted to obtain visual features, detect the speaker's facial action units in each frame, and extract 35 muscle movement units. The dimensionality of textual, audio, and visual features for each video sequence is d_T=300, d_A =74 and d_V=35, respectively.

3.2. Unimodal temporal feature extraction

In order to explore the unique features within modalities and extract textual, visual and speech modal temporal features with contextual semantic information, the dimensionality of different modalities was unified by convolution neural network. The contextual information was captured by using bi-directional gated recurrent neural network and further enriched based on multi-head attention mechanism to generate the final feature representation containing temporal information.

The textual, audio, and visual feature in each video sequence was obtained through the input layer, and each video sequence contains distinct sentiment polarity. The feature dimensions are denoted as d_T=300, d_A =74 and d_V=35,respectively. Each sample data $\{x_1, x_2, \ldots, x_L\}$ is divided into a sequence of length, and each sample sequence is divided into three sequential modalities of textual (T), visual (V), and audio (A), represented by $X = [X_T, X_V, X_A]$. The feature signals obtained by the above method are input to the temporal single-modal feature extraction layer, as shown in Fig. 3.

CNNs have also been used to extract local information about sequences. Local temporal features are extracted using a set of convolution kernels with fixed feature dimensions $d_k (k \in \{T, V, A\})$. The features of different modalities are mapped to the same dimension d by CNN, which is not only beneficial for dot product operation of

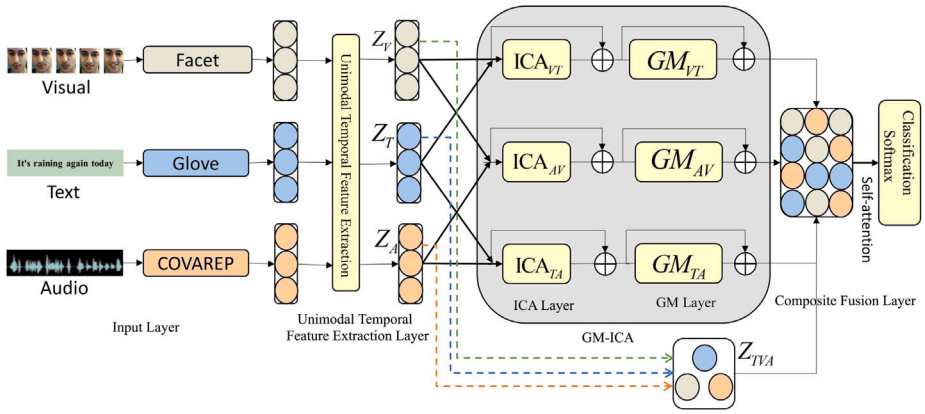


Fig. 1. Composite Cross Modal Interaction Network with Attention enhancement for Multimodal Sentiment Analysis.

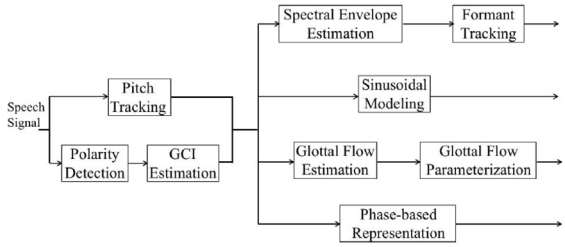


Fig. 2. COVAREP Speech Analysis Repository.

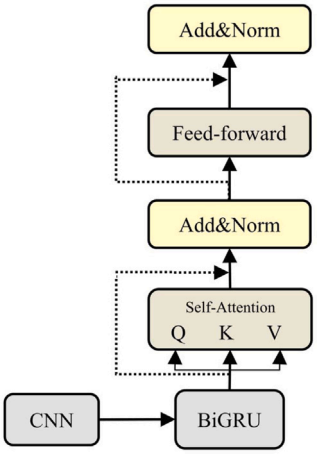


Fig. 3. Unimodal temporal feature extraction layer.

cross-modal attention module, but also ensures that there is sufficient awareness between neighboring elements of input sequence.

$$X_{\{T,V,A\}} = Conv1D(X_{\{T,V,A\}}) \in \mathbb{R}^{d\times L}; \quad k \in \{T,V,A\} \tag{1}$$

Where, L is the sequence length.

After the processing of one-dimensional convolution layer, the obtained features are inputted into BiGRU to obtain bidirectional hidden state of the features at each moment and capture long-term dependency of context. The weight of the contextual features in the modal state is enhanced by MHA, and the similarity of the query indexes in different subspaces is calculated.

$$H = BiGRU(X\{T,V,A\}), k \in \{T,V,A\} \tag{2}$$

$$a_i(H) = softmax\left(\frac{(W^Q H)^T (W^K H)}{\sqrt{d_h/M}}\right)(W^V H)^T \tag{3}$$

Where, M represents the number of attention heads, d_h is the hidden state dimension at a certain moment, W^Q, W^K and W^V are the corresponding mapping matrices of Querys(Q), Keys(K) and Values(V) respectively.

All the attention heads are spliced to obtain the complete output results. The cumulative output results and the query matrix elements are processed through Layer Normalization (LN) to avoid gradient explosion caused by too large value. Residual connection is added to prevent network degradation, and unimodal eigenvector Z is obtained through fully connected layer integration.

$$M_a(H) = concat(a_1, a_2, \ldots, a_M) \tag{4}$$

$$M_{LN} = LN(H + M_a(H)) \tag{5}$$

$$Z = LN(FC(M_{LN}) + M_{LN}) \in \mathbb{R}^{d\times L} \tag{6}$$

Where $Z = [Z_T, Z_V, Z_A]$, $Z_T \in \mathbb{R}^{d\times L}$, $Z_V \in \mathbb{R}^{d\times L}$ and $Z_A \in \mathbb{R}^{d\times L}$ denote the final output textual, visual and audio feature vectors with contextual timing information, respectively.

3.3. Gating mechanism-improved cross-modal attention network(GM-ICA)

The Gating Mechanism-Improved Cross-modal Attention (GM-ICA) network consists of two parts: the Improved Cross-modal Attention (ICA) Layer and the Gating Mechanism (GM) Layer.

3.3.1. Improved cross-modal attention enhancement interaction network layer(ICA)

A multimodal fusion method based on the improved cross-modal attention (ICA) layer is designed in the Improved Cross-modal Attention (ICA) layer to realize the information enhancement among different modalities. The target modality is continuously reinforced through low-order signals of auxiliary modality to learn the corresponding information and capture the potential adaptations among modalities.

Due to the different sampling rates of the sequences in different modalities, the vast majority of the time series are unaligned data. Cross Modal Transformer (CM) considers the interactions between different time series and solves the problem of unaligned data during multimodal fusion [32]. As shown in Fig. 4, for target mode α and auxiliary mode β, each modal (unaligned) sequence feature is denoted as $X_\alpha \in \mathbb{R}^{L_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{L_\beta \times d_\beta}$, where L is the length of the sequence and d denote the sequence length,respectively. As shown in Fig. 4, the potential

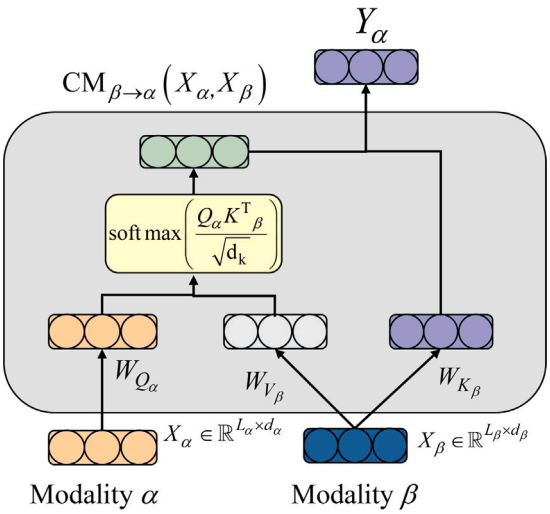


Fig. 4. Improved cross-modal attention network.

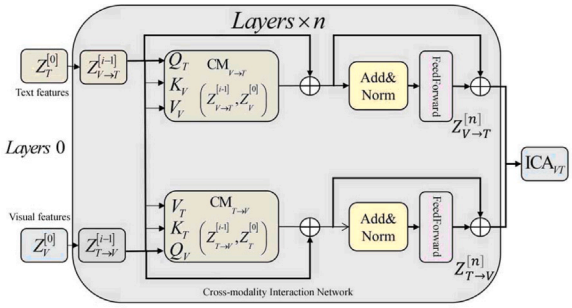


Fig. 5. Improved cross-modality information enhancement model structure (ICA).

adaptation from modality β to modality α is represented as follows:

$$\begin{aligned} Y_a &= CM_{\beta\to\alpha}(X_\alpha, X_\beta) = \\ & softmax(\frac{Q_\alpha K_\beta^T}{\sqrt{d_K}})V_\beta = \\ & softmax(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^T K_\beta^T}{\sqrt{d_K}})X_\beta W_{V_\beta} = \end{aligned} \tag{7}$$

Where, Y_α is the cross-modal attention realization operator, the augmented query $= Q_\alpha X_\alpha W_{Q_\alpha}$ comes from the target modality α, the key $= K_\beta X_\beta W_{K_\beta}$ and the value $= V_\beta X_\beta W_{V_\beta}$ come from the modality β, W_{Q_α}, W_{K_β}, and W_{V_β} are the corresponding weight matrices, respectively. The source modality,such as β reconstructs the feature information of the target modality such as α with low-order signals, and captures important correlation information of different modalities through key/value interaction.

Taking textual modality (Z_T) and visual modality (Z_V) as examples, the information enhancement process is shown in Fig. 5.

The sequence of unimodal features with contextual timing information is input to the cross-modal attention interaction layer. $Z_{V\to T}$ and $Z_{T\to V}$ denote the visual low-order signals reconstruct textual features, and the textual low-order signals reconstruct visual features. In order to avoid gradient explosion caused by the large value, layer normalized LN and residual connectivity are added [33]. The similarity between auxiliary modality and target modality is computed using the two CM modules in both directions, The target modes are continuously strengthened, and feedforward calculation is carried out according to

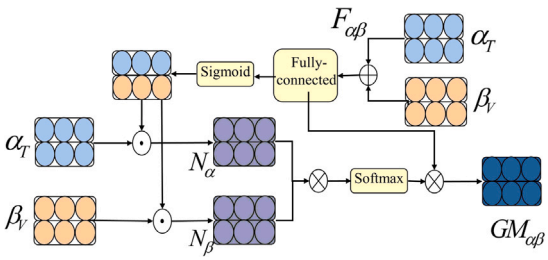


Fig. 6. Bimodal joint feature structure.

i=1,2, . . . ,n layers with $Z_{V\to T}$ as an example, the calculation process is shown in Eqs. (8)–(11).

$$Z_{V\to T}^{[0]} = Z_T^{[0]} \tag{8}$$

$$P_1 = CM_{V\to T}^{[i]}(LN(z_{V\to T}^{[i-1]}), LN(Z_V^{[0]})) \tag{9}$$

$$Z_{V\to T}^{[i]} = P_1 + LN(Z_{V\to T}^{[i-1]}) \tag{10}$$

$$Z_{V\to T}^{[i]} = f_{\theta_{V\to T}^i}(LN(Z_{V\to T}^{[i]})) + LN(Z_{V\to T}^{[i]}) \tag{11}$$

Where, P_1 serves as an intermediate variable to obtain feed-forward values for cross-modal attention, f is a positional feed-forward sublayer parameterized by θ. Sufficient attention is given to the $i-1$ time step of the visual modality via the ith time step of the textual modality.

Textual modality is continuously strengthened using the low-order visual modality information, which can obtain the relevant data from visual modality. After n-layer CM module superposition, the unidirectional cross-modal textual features $Z_{V\to T}^{[n]}$ and the visual features $Z_{T\to V}^{[n]}$ after the textual modality is strengthened by the visual modality are obtained. The enhanced modal features are spliced by the activation function tanh to capture the hidden inter-modal correlations, and the feature vector $Z_{VT}^{[n]}$ is obtained. The attention distribution matrix att_1, can be computed by Softmax, which is multiplied by the multimodal fusion matrix. The final weighted cross-modal interaction features ICA_{VT} of text and vision are computed as shown in Eqs. (12), (13) and (14).

$$Z_{VT}^{[n]} = \tanh(W_{VT}(Z_{V\to T}^{[n]} \oplus Z_{T\to V}^{[n]}) + b_{VT}) \tag{12}$$

$$att_1 = softmax(Z_{VT}^{[n]}) \tag{13}$$

$$ICA_{VT} = att_1 \bullet Z_{VT}^{[n]} \tag{14}$$

Where $\oplus$ represents the vector splicing operation, W_{VT} and b_{VT} represent the mapping matrix and bias term after feature splicing respectively, and $\bullet$ represents the matrix multiplication.

3.3.2. Bimodal joint feature representation

The ICA layer enhances the semantic correlation between two modalities by continuously reinforcing the target modality through low-order signals of the auxiliary modality. Aiming at the acquired bimodal features with semantic correlation, a bimodal joint feature representation method similar to Gating Mechanism (GM) is designed. With two semantically related modal features as the conditional vectors, the similarity between different modalities is explored more deeply through bimodal interaction. The weight of important inter-modal correlation features is strengthened, and the interaction between different modalities is explored at a deeper level. The overall structure is shown in Fig. 6.

Taking modality α_T and modality β_V as examples, which represent different modal features, the feature vectors of α_T and β_V are firstly spliced by Fully Connected Layer (FC) to capture the association information, and the bimodal joint feature vector $F_{\alpha\beta}$ is obtained. The

computational process is shown in Eqs. (15).

$$F_{\alpha\beta} = \tanh(W_{\alpha\beta}(\alpha_T \otimes \beta_V) + b_{\alpha\beta}) \tag{15}$$

Where, $\otimes$ is the splicing between vectors, $W_{\alpha\beta}$ and $b_{\alpha\beta}$ and are randomly initialized weights and bias terms, respectively.

The joint feature vector $F_{\alpha\beta}$ is input into the activation function Sigmoid to generate the joint condition vector $N_{\alpha\beta}$, which expresses the similarity within different modalities and enhances the weights of correlation features. The calculation process is shown in Eqs. (16).

$$N_{\alpha\beta} = Sigmoid(F_{\alpha\beta}) \tag{16}$$

The joint condition vector $N_{\alpha\beta}$ is multiplied with the corresponding elements of the initial modal vectors α_T and β_V to obtain different degrees of attention on different modes. The condition vector matrix N_α of α_T and the condition vector matrix N_β of β_V are obtained. Matrix multiplication operation is performed on N_α and N_β, and the probability distribution a_1 of feature matrix is obtained by Softmax function. Matrix multiplication operation is performed by using joint feature vector $F_{\alpha\beta}$ and probability distribution a_1. The weight of the key information is strengthened by improving the weight of joint feature $F_{\alpha\beta}$, and the final bimodal joint feature $GM_{\alpha\beta}$ is obtained. The calculation process is shown in Eqs. (17)–(21).

$$N_\alpha = \alpha_T \odot N_{\alpha\beta} \tag{17}$$

$$N_\beta = \beta_V \odot N_{\alpha\beta} \tag{18}$$

$$R_{\alpha\beta} = N_\alpha \bullet N_\beta \tag{19}$$

$$a_1 = \frac{e^{R_{\alpha\beta}(i,j)}}{\sum_{k=1}^{n} e^{R_{\alpha\beta}(i,k)}}, i, j = 1, 2, \ldots, n \tag{20}$$

$$GM_{\alpha\beta} = a_1 \bullet F_{\alpha\beta} \tag{21}$$

Where, $\bullet$ represents matrix multiplication, $\odot$ represents multiplication of corresponding elements, $R_{\alpha\beta}$ is the joint feature matrix for obtaining the probability distribution a_1, and $GM_{\alpha\beta}$ is the final bimodal joint eigenvector.

3.3.3. Interactive network based on gating mechanisms (GM)

By improving the cross-modal interaction layer ICA, the information-enhanced text–visual features ($Z^n_{V\to T}$, $Z^n_{T\to V}$), visual–speech features ($Z^n_{A\to V}$, $Z^n_{V\to A}$), and speech–text features ($Z^n_{T\to A}$, $Z^n_{A\to T}$) are obtained respectively. However, due to the variability of semantics between different modalities, there are still some problems of insufficient fusion and weak interactivity in the model. Using bimodal joint feature representation method, text–visual features ($Z^n_{V\to T}$, $Z^n_{T\to V}$), visual–speech features ($Z^n_{A\to V}$, $Z^n_{V\to A}$), and speech–text features ($Z^n_{T\to A}$, $Z^n_{A\to T}$) are inputted into the interaction layer of the gating mechanism network for fusion, and the similarity between the two modalities is computed to enhance the correlation degree of the important features so as to capture the deeper interactions between modalities. The obtained joint features are used for multi-modal fusion to carry out the final sentiment classification task. The structure of the cross-modal sentiment interaction network model based on gating mechanism is shown in Fig. 7.

In the GM-ICA layer, the enhanced modal features are spliced by the activation function tanh to capture the hidden correlations between the modes and obtain the joint features F_{VT}, F_{AV} and F_{TA}. The calculation process is shown as Eqs. (22), (23), and (24).

$$F_{VT} = \tanh(W_{VT}(Z^{[n]}_{V\to T} \oplus Z^{[n]}_{T\to V}) + b_{VT}) \tag{22}$$

$$F_{AV} = \tanh(W_{AV}(Z^{[n]}_{A\to V} \oplus Z^{[n]}_{V\to A}) + b_{AV}) \tag{23}$$

$$F_{TA} = \tanh(W_{TA}(Z^{[n]}_{T\to A} \oplus Z^{[n]}_{A\to T}) + b_{TA}) \tag{24}$$

Where, $\otimes$ denotes the vector splicing operation, W_{VT}, W_{AV}, W_{TA} and b_{VT}, b_{AV} and b_{TA} denote the mapping matrix and bias term after feature splicing respectively.

Sigmoid function is utilized to generate conditional feature vectors N_{VT}, N_{AV} and N_{TV}, which are calculated in Eqs. (25), (26), and (27).

$$N_{VT} = Sigmoid(F_{VT}) \tag{25}$$

$$N_{AV} = Sigmoid(F_{AV}) \tag{26}$$

$$N_{TA} = Sigmoid(F_{TA}) \tag{27}$$

Conditional vectors are multiplied with the corresponding elements of unidirectional cross-modal features respectively. Different degrees of attention are paid to the features of different modalities, and then multiplied by feature mapping matrix to obtain the intermediate eigenvectors R_{VT}, R_{AV} and R_{TA},as shown in Eqs. (28), (29), and (30).

$$R_{VT} = (Z^{[n]}_{V\to T} \odot N_{VT}) \bullet (Z^{[n]}_{T\to V} \odot N_{VT}) \tag{28}$$

$$R_{AV} = (Z^{[n]}_{A\to V} \odot N_{VT}) \bullet (Z^{[n]}_{V\to A} \odot N_{AV}) \tag{29}$$

$$R_{TA} = (Z^{[n]}_{T\to A} \odot N_{VT}) \bullet (Z^{[n]}_{A\to T} \odot N_{TA}) \tag{30}$$

Where, $\odot$ represents multiplication of corresponding elements, and $\bullet$ represents matrix multiplication.

The attention score of the intermediate feature vector $R_{VT,AV,TA}$ is obtained by Softmax, which is calculated as Eq. (31). The weight of the key information is strengthened by improving the weight of joint feature. Finally, the bimodal joint features GM_{VT}, GM_{AV} and GM_{TA} are calculated as Eqs. (32),(33), and (34).

$$c_{\{VT,AV,TA\}} = softmax(R_{\{VT,AV,TA\}}) \tag{31}$$

$$GM_{VT} = c_{VT} \bullet F_{VT} \tag{32}$$

$$GM_{AV} = c_{AV} \bullet F_{AV} \tag{33}$$

$$GM_{TA} = c_{TA} \bullet F_{TA} \tag{34}$$

3.4. Composition fusion layer

First, the low-order feature vectors $Z_{TVA} = [Z_T, Z_V, Z_A]$ of text, speech and vision are extracted by unimodal temporal features. Second, the text–visual, visual–audio and audio–text bimodal joint features GM_{VT}, ,GM_{AV} and GM_{TA} are obtained by the improved multimodal fusion method. Based on the fusion of three modalities, the structure of residual-like network is used for the fusion of composite hierarchy, so as to alleviate the problem of gradient explosion or disappearance in the case of deep model network layers, as shown in Fig. 8. Bimodal joint features are spliced to obtain the trimodal joint features GM_{TVA}, which are computed as Eqs. (35), (36) and (37). The probability distribution matrix is calculated by the Softmax function, which enhances more important feature information inside the modality in a weighted way.

$$c_3 = GM_{VT} \oplus GM_{AV} \oplus GM_{TA} \tag{35}$$

$$F_3 = ReLU(W_{F_3 c_3} + b_{F_3}) \tag{36}$$

$$GM_{TVA} = Softmax(F_3) \tag{37}$$

Where F_3, is the probability distribution of the feature vector c_3, W_{F_3} and b_{F_3} are initialization weights and bias terms of the activation function ReLU.

Composite hierarchical fusion is performed, and the multimodal sentiment feature U is obtained by fusing the obtained GM_{TVA} with the

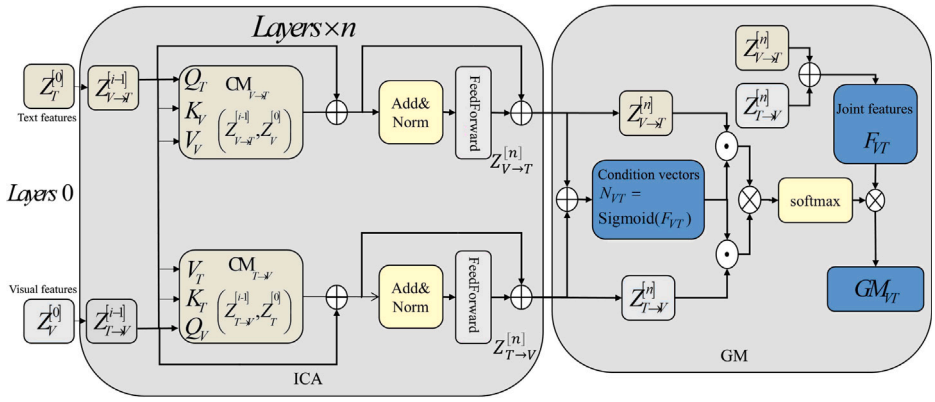


Fig. 7. Gating mechanism-improved cross-modal attention network (GM-ICA).

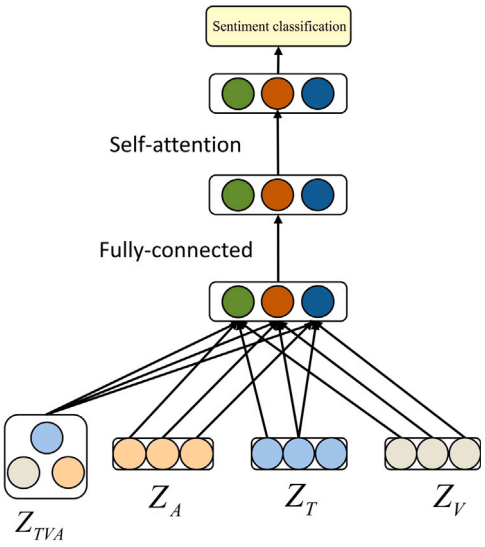


Fig. 8. Structure of composite hierarchical fusion model.

unimodal sentiment features Z_V, Z_T, Z_A through the full connectivity layer. The calculation process is shown in Eq. (38).

$$U = \tanh([GM_{TVA} \oplus Z_V \oplus Z_T \oplus Z_A] \bullet W_U + b_U) \tag{38}$$

Where, W_U and b_U are initialized weights and bias terms of the activation function tanh.

Finally, the multimodal sentiment feature U is input to the self-attention layer, and the final multimodal fusion features are conducted with sentiment classification using the Softmax function,as shown in Eq. (39).

$$CCIN = softmax(ReLU(W_s U + b_3)) \tag{39}$$

Where, W_3 and b_3 are the initialized weights and bias terms of the activation function $ReLU$, and $CCIN$ is the final classification result.

4. Result analysis and discussion

4.1. Data sets

The modeling experiments used two publicly available multimodal sentiment analysis datasets, CMU-MOSI [34] and CMU-MOSEI [35] collected by Carnegie Mellon University from the YouTube website.

(1) The MOSI dataset randomly collected 93 videos with sentiment polarity from the YouTube website. In the 93 videos, 89 speakers, including 41 are female and 48 male aged from 20 to 30 years old, express opinions on different topics. The entire video library is divided into 2199 video clips with sentiment labels, each of which is composed of three modalities of data: text, speech, and video.

The sentiment polarity of each video clip was subjectively annotated from very negative to very positive, with a linear range of [−3, 3], as shown in Table 1. For sentiment dichotomy problem, sentiment values labeled [−3, 0] denote negativity and sentiment values labeled [1, 3] denote positivity.

(2) The MOSEI dataset expands the data volume based on the MOSI dataset, which contains more than 1000 monologue speakers and 250 different types of speech videos. Each video is labeled with the gender of the speaker, and the total length reaches 651 h with the ratio of male to female of 0.57:0.43. This dataset contains 22,676 video segments with sentiment labels, and the average length is 7.28 s, containing 19 words on average. The distribution of sentiment intensity is similar to that of the MOSI dataset, with the number of positive sentiment labels accounting for 71% of the total video segments. The sample cuts of the two datasets are shown in Table 2.

4.2. Experimental setup

In the experiment, RTX3070 graphics card and 16 GB memory GPU were used for model training. PyTorch deep learning framework was used as the main experimental platform for multimodal sentiment analysis model. The detailed configuration of hardware and software environments required for the experiment is shown in Table 3.

Cross-entropy is selected as the loss function, Dropout is used to prevent overfitting, and Adam method is chosen to optimize the learning parameters. The experimental parameters of the model are set in Table 4.

4.3. Metrics

In this study, five common evaluation metrics in multimodal sentiment analysis tasks were used: binary accuracy (Accuracy2, Acc-2), seven-classification accuracy (Accuracy7, Acc-7), F1 (F1-Measure) scores, Mean Absolute Error (MAE), and Pearson correlation coefficient (Pearson correlation, Corr). The details of the evaluation metrics are as follows: The binary confusion matrix was created based on the true value and the predicted value of the sample data as shown in Table 5.

Acc indicates the ratio of the number of correctly predicted samples to the total number of samples. Eq. (40) directly reflects the probability of correction prediction, distributed to [0,1], which can be classified into two, three and seven classifications according to the sentiment label. The number of classifications set up in this experiment is the second and seven classifications.

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{40}$$

Table 1
Meaning of sentiment category labels.

Label	−3	−2	−1	0	1	2	3
Meaning	Very negative	More negative	Negative	Neutral	Positive	More positive	Very positive

Table 2
Data set statistics.

Data set	Training set	Validation set	Test set	Subtotal
MOSI	1151	276	772	2199
MOSEI	16,216	1835	4625	22,676

Table 3
Experimental hardware and software environment.

Software	Configuration details
CPU	I7 5800
GPU	RTX3070
Random access memory (RAM)	16G
Implementation language	Python 3.7
Deep Learning Framework	PyTorch 1.11.0
CUDA	11.31
cuDNN	8.2.1

Table 4
Experimental parameter settings.

Hyperparameter setting	CMU-MOSI	CMU-MOSEI
Learning rate	0.001	0.001
Training cycle	50	30
Batch size	32	32
Text/Speech/Vision Convolutional Kernel	3/3/3	3/3/3
BiGRU hidden unit	300	300
Number of cross-attention	10	10
Cross-modal module stacking n	4	4
Self-attention Dropout	0.1	0.1
Cross-modal Attention Dropout	0.35	0.35
Residual Network Dropout	0.25	0.25
Gradient trimming threshold	0.8	1

Table 5
Confusion matrix.

Predicted value/true value	Positive sample	Negative sample
Positive sample	TP (True Positive)	FP (False Positive)
Negative sample	FN (False Negative)	TN (True Negative)

F1 (F1-Measure) score is used as an evaluation criterion for sentiment classification performance, and the F1 score is the reconciled value of Precision (Pre) and Recall (Recall, Rec), which takes the value interval in [0,1].

The precision rate indicates the proportion of true positive samples inside the predicted positive samples, which is calculated as below:

$$pre = \frac{TP}{TP + FP} \tag{41}$$

Recall represents the proportion of the samples predicted to be positive in all positive samples, and it is calculated as shown in (42):

$$Rce = \frac{TP}{TP + FN} \tag{42}$$

F1 score is calculated from the precision and the recall. In case of sample imbalance in the dataset, F1 score can better measure the performance than the precision, and it is calculated as shown in Eq. (43):

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{43}$$

MAE (Mean Absolute Error) can measure the degree of difference between the estimated results and the measured data more accurately, and the formula is shown in (44). It can better reflect the error between the predicted value and the actual value, and MAE is [0,+). When the predicted value is exactly the same as the true value, it is equal to 0, and the idealized modeling effect is achieved. The larger the value of MAE, the larger the error, and the smaller the value of MAE, indicating that the model has better accuracy and reliability in predictive performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{44}$$

Pearson's correlation coefficient, also known as product-moment correlation, is a commonly used metric for evaluating the performance of an algorithmic model. It is used to measure the degree of linear relationship between two variables, as shown in the formula in (45):

$$Corr = \frac{1}{n}\sum_{i=1}^{n} \frac{E(y_i\hat{y}_i) - E(y_i)E(\hat{y}_i)}{\sigma_{y_i}\sigma_{\hat{y}_i}} \tag{45}$$

Where, n represents the sample size, E is the expected value $\sigma_{\hat{y}_i}$, and represents the standard deviation of y_i.

4.4. Analysis and discussion

4.4.1. Baseline models

Comparative experiments were conducted using various multimodal baseline models for sentiment analysis and the proposed CCIN-SA model.

Early Fusion LSTM (EF-LSTM): This approach concatenates the inputs from three modalities to obtain fused features, which are then processed by an LSTM for sentiment analysis.

Later Fusion LSTM (LF-LSTM): In contrast, LF-LSTM first employs LSTMs to learn individual modality features separately. These features are then concatenated and fed into a classifier to determine the sentiment polarity.

TFN [20]: An outer-product based approach was used to establish a multidimensional tensor for modality-specific and cross-modality features, and capture the intra-modal interaction information.

LMF [21]: On the basis of TFN, feature fusion was realized using low-order factors to reduce the computational complexity of tensor fusion.

MFM [23]: Different modal information was decomposed into shared discriminative features and specific generative factor features to learn multimodal representations.

MulT [27]: Using the idea of cross-modal attention mechanism, the similarity between two modalities was computed to directly deal with unaligned data.

MISA [24]: Each modal feature was mapped to two different representation spaces and learn modal specific and invariant features.

MAG-Bert [25]: Multimodal information by fine-tuning non-textual data can be obtained using Bert and XLNet preprocessing methods.

Self-MM [26]: A self-supervised method for generating unimodal labels was employed in tandem with a multi-task learning strategy to explore the consistencies and variations within multimodal representations.

NHFNET [6]: Computational complexity of cross-modal attention is reduced and the efficiency of multimodal fusion is improved by enhancing information from audio and visual modalities.

4.4.2. Comparative experiment of GM-ICA model

Comparison experiment results on MOSI and MOSEI dataset are shown in Table 6, Table 7. Results demonstrate that the GM-ICA model surpasses the baseline model in most evaluation indexes.

Experiment results show that GM-ICA model has better performance in all indicators compared with other models. Compared to the simple

Table 6
Comparison experiment results of MOSI dataset.

Model	Acc-2	F1-Score	MAE	Corr	Acc-7
EF-LSTM	75.32	75.20	1.023	0.607	33.71
LF-LSTM	76.82	76.73	1.015	0.624	35.33
TFN	77.66	77.56	1.040	0.586	34.50
LMF	79.98	79.89	0.971	0.677	35.01
MFM	79.70	79.67	0.951	0.683	36.25
MulT	82.57	82.34	0.871	0.691	40.04
MISA	82.79	82.75	0.855	0.711	41.71
MAG-BERT	83.93	83.81	0.817	0.729	42.67
Self-MM	83.14	83.09	0.837	0.729	42.69
NHFNET	83.25	83.28	0.815	0.725	43.03
GM-ICA	**83.93**	**83.90**	**0.810**	**0.731**	**43.35**

Table 7
Comparison experiment results of MOSEI dataset.

Model	Acc-2	F1-Score	MAE	Corr	Acc-7
EF-LSTM	78.21	77.96	0.642	0.616	47.42
LF-LSTM	80.06	80.63	0.619	0.659	48.88
TFN	79.41	79.75	0.610	0.671	49.87
LMF	80.09	80.07	0.627	0.673	49.73
MFM	79.91	78.03	0.610	0.640	50.00
MulT	82.62	82.44	0.580	0.693	51.81
MISA	82.93	82.91	0.578	0.701	52.20
MAG-BERT	83.32	83.32	0.553	0.725	52.67
Self-MM	83.58	83.61	0.550	0.719	53.02
NHFNET	84.04	84.29	0.552	0.723	53.59
GM-ICA	**84.70**	**84.91**	**0.545**	**0.736**	**54.12**

feature-level fusion EF-LSTM and decision-level fusion LF-LSTM models, the GM-ICA model demonstrates significant improvements across all metrics, with binary classification accuracy increasing by approximately 5%. Compared with the models based on non-attention methods for multimodal fusion such as TFN, LMF and MFM, the effect of GM-ICA model is significantly improved, and the attention-based multimodal fusion method can better explore the potential correlation between modalities. Obviously, it is difficult to take into account both the intra-modal feature representations and inter-modal information interactions with the feature-level fusion strategy. Compared with complex fusion network models such as MulT and MISA models, the binary classification accuracy is improved by about 2%, indicating that the cross-modal interaction network model based on the gating mechanism focuses more accurately on deeper correlations within the modality while considering effects of modality temporality and lower and higher-order modal information on sentiment classification results. The model can learn more complete sentiment expression and thus make more accurate sentiment judgment. The proposed model is superior to the existing models such as MAG-BRET, Self-MM and NHFNET, and the classification accuracy is improved by about 1%, which fully demonstrates that the interaction information between different modalities can be learned by the improved cross-modal interaction network.

4.4.3. Modality ablation study of GM-ICA

In order to verify whether the multimodal combination makes up for the insufficient ability of unimodality, this study carried out feature extraction for a single modality (text T, visual V, audio A), a two-by-two combination of modalities (T+V,V+A,A+T) and a three-modal combination (T+V+A) and conducted experiments on sentiment analysis, respectively.

Unimodality only uses unimodal temporal feature extraction technique to obtain feature information, which is directly used for sentiment classification task. Bimodality used cross-modal interaction network model to realize feature fusion on the basis of unimodality, and the fused feature information was used for sentiment classification.

Tri-modality carried out feature extraction using the proposed GM-ICA method, and the obtained results were used for the final sentiment analysis task. Tables 8 and 9 depict comparison experiments on MOSI dataset and MOSEI dataset.

Table 8
Modality ablation experiments of MOSI dataset.

Modalities combination	Acc-2	F1-Score	MAE	Corr	Acc-7
T	78.14	78.77	0.954	0.677	39.35
V	67.10	59.08	1.104	0.574	36.11
A	65.36	57.74	1.101	0.570	36.03
T+V	78.72	78.85	0.947	0.680	39.75
V+A	64.83	58.96	1.115	0.567	35.58
A+T	78.32	78.26	0.945	0.680	39.44
T+V+A	**83.93**	**83.90**	**0.810**	**0.731**	**43.35**

Table 9
Modality ablation experiments of MOSEI dataset.

Modalities combination	Acc-2	F1-Score	MAE	Corr	Acc-7
T	80.23	80.07	0.614	0.711	49.95
V	71.86	79.16	0.647	0.640	47.73
A	70.55	78.58	0.673	0.633	45.00
T+V	81.53	81.91	0.535	0.706	51.88
V+A	70.69	79.62	0.679	0.625	47.95
A+T	81.54	81.74	0.579	0.702	51.13
T+V+A	**84.70**	**84.91**	**0.545**	**0.736**	**54.12**

Experimental results show that the three modal features fusion of text (T), visual (V) and audio (A) have the optimal result for sentiment analysis, which fully proves the importance of multimodal fusion. Each modality contains specific sentiment information, which plays an auxiliary effect for multimodal fusion. Comparing the performance of the two datasets, there is an improvement of approximately 1% in binary classification accuracy and about 10% in seven-class classification accuracy, respectively. Sentiment classification can be improved in a certain degree on the larger-scale training data. Particularly, the improvement of sentiment category in finer granularity is obvious.

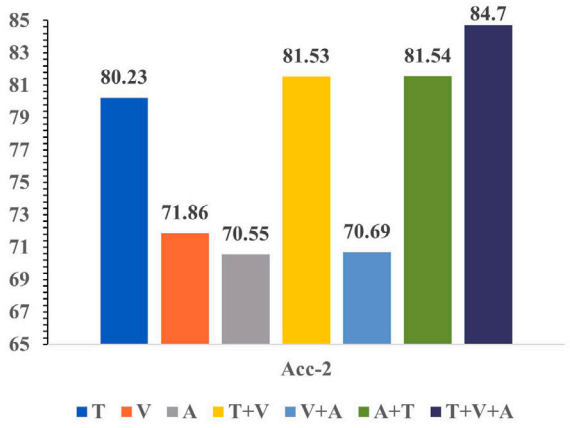
As shown in Fig. 9, in unimodal and bimodal experiments under the same dataset, the results of textual unimodal and bimodal with textual modality are significantly better than the other two, indicating that the sentiment polarity of textual modal features is the most significant in general, and the sentiment information embedded in textual features is more abundant and intuitive. Compared to unimodality, the results of the two metrics of the model in bimodality are significantly better, but the accuracy of V+A modality is slightly lower than that of visual modality, which is caused by the relatively weak emotional polarity of visual and audio modalities, the low interactivity, and the interference of redundant information. Therefore, feature fusion using text, visual, and audio modalities effectively improves the performance of multimodal emotion classification, and the validation also compensates for the problem of incomplete single-modal sentiment expression.

4.4.4. Model ablation study of GM-ICA

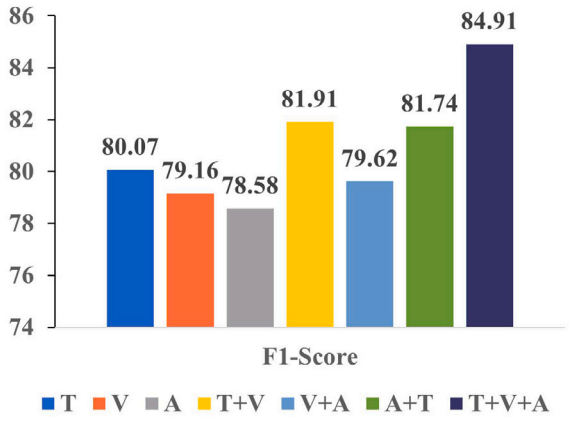
GM-ICA model further fuses bimodal features based on ICA with improved cross-modal information augmentation to capture the correlations between different modalities. Considering whether this approach introduces modal correlations that do not exist through modal augmentation, the final sentiment classification results may be influenced in the gating mechanism cross-modal interaction. To further assess the impact of each module in GM-ICA network model, the following model ablation study was designed.

Comprehensive ablation studies were conducted to evaluate the impact of the four modules, E, B, C, and D, on the overall performance of the model, and the control experiment modules was combined as shown in Table 10.

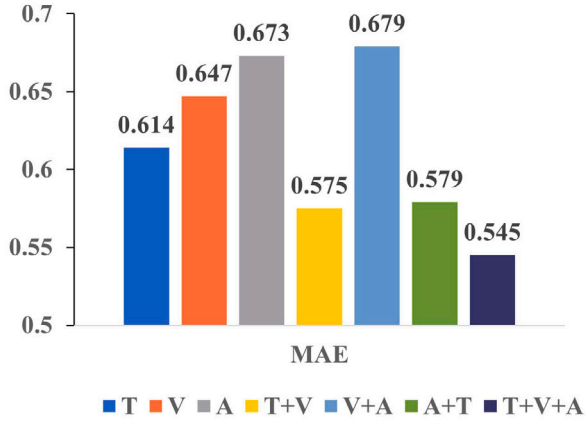
E: GM-ICA (Bi-GRU): Bi-GRU unimodal feature extraction module;
B: GM-ICA (ICA): Improved cross-modal interaction module
C: GM-ICA (Temp): Unimodal timing feature extraction module;
D: GM-ICA (GM): Gating Mechanism Network Module;



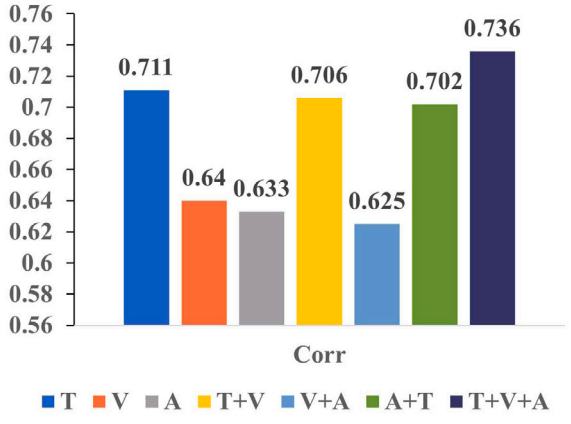
(a) Acc-2



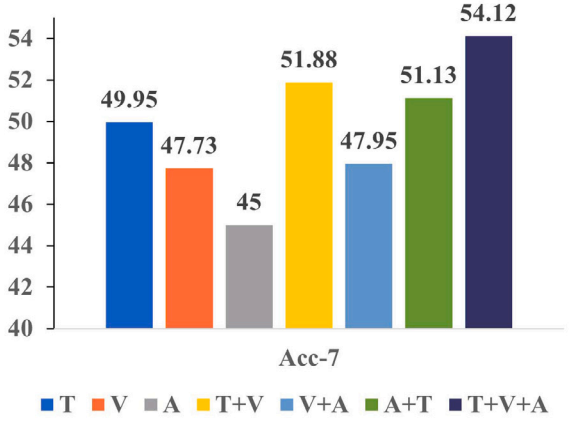
(b) F1-Score



(c) MAE



(d) corr



(e) Acc-7

Fig. 9. Modality ablation experiments of MOSEI.

Table 10
Modules Configuration.

Modules	B:GM-ICA(ICA)	D:GM-ICA(GM)	B+D
E:GM-ICA(Bi-GRU)	**E+B**	E+D	E+B+D
C:GM-ICA(Temp)	C+B	C+D	**C+B+D**

Table 11
Module ablation experiments of MOSI dataset.

Modules combination	Acc-2	F1-Score	MAE	Corr	Acc-7
E+B	83.32	83.12	0.867	0.697	40.31
C+B	83.41	83.37	0.835	0.719	42.47
E+D	83.18	83.07	0.869	0.709	40.27
C+D	83.47	83.45	0.830	0.711	42.69
E+B+D	83.52	83.52	0.827	0.717	43.03
C+B+D	**83.93**	**83.90**	**0.810**	**0.731**	**43.35**

Table 12
Module ablation experiments of MOSEI dataset.

Modules combination	Acc-2	F1-Score	MAE	Corr	Acc-7
E+B	83.37	83.19	0.579	0.699	51.83
C+B	83.69	83.75	0.550	0.725	52.07
E+D	83.61	83.59	0.563	0.719	51.98
C+D	83.97	83.95	0.558	0.713	52.77
E+B+D	84.37	84.19	0.549	0.727	53.83
C+B+D	**84.70**	**84.91**	**0.545**	**0.736**	**54.12**

Tables 11 and 12 illustrate the effects of different modules on the overall model performance, respectively.

Results indicate that GM-ICA outperforms other models in all metrics, and that the absence of any module can influence the overall model. The experimental results of Acc-2 and F1 values are shown in Fig. 10. Under a larger dataset, all indexes are improved, in which the binary classification accuracy and F1 value are improved by about 1.5%, and the model performance is better. For complex neural networks with high computational power, a larger dataset is favorable to capture more implicit sentiment information. For sentiment label classification with finer granularity, the seven classification accuracy rate is significantly improved compared to binary classification accuracy rate, and it performs better on sentiment division with finer granularity, reflecting the advanced and effective model improvement.

The E+B, C+B, E+B+D, and C+B+D experiments were compared on the MOSEI dataset, and the values of Acc-2 and F1 are shown in Figs. 11. By comparing E+B, C+B and E+B+D, C+B+D models on MOSEI, the accuracy and F1 value of binary classification are improved by about 1%, which indicates that the interaction information between different modalities can be captured by the improve gating mechanism interactive network module, and potential correlation between modalities can be more deeply explored. Comparing the experiment results of E+B and C+B, E+B+D and C+B+D, the improved temporal unimodal feature extraction method also contributes about 0.5% compared to the unimodal feature extraction of Bi-GRU. Multi-head attention mechanism method strengthens the weight of the intra-modal contextual features. Comparing the experiment results of E+D, C+D, E+B+D and C+B+D, the cross-modal interaction module contributes about 1% on binary classification accuracy and F1 value. The cross-modal interaction module still has a certain enhancement effect on the model under the strong interaction of the gating mechanism network. In cross-modal interaction, the low-order signals of the auxiliary modalities continuously strengthen the target modality, which retains some unique feature information. Comparing the experiment results of E+B and E+D, C+B and C+D, the gating mechanism network shows stronger interactivity than the cross-modal interaction module in multimodal fusion, and pays better attention to the potential correlations between modality.

The quantity of layers in the cross-modal module stack is a critical hyper parameter influencing the model's ultimate performance, determining the count of cross-modal Transformer modules. Feed-forward computation was performed from i=1 to n layers, with the auxiliary modality constantly updating its sequence information. To explore the impact of varying the number of cross-modal attention layers, denoted as n, on the model's overall performance, experiments were conducted on the CMU-MOSEI dataset using ICA-MSA model, with n ranging from 1 to 9. The metrics assessed in these experiments were accuracy for binary classification (Acc-2) and the F1 score. Fig. 12(a) represents Acc-2 value and Fig. 12(b) depicts F1-Score value. According to the analysis of the experimental results, the two metrics show similar change trends, and both the accuracy and the F1 value rise first and then fall. Optimal model performance is achieved when the cross-modal module has four stacked layers, suggesting that there is not a direct positive correlation between the quantity of layers and performance. When the quantity of layers exceeds four, the model's performance starts to deteriorate markedly, and the training of the model becomes more challenging and less manageable.

Table 13
Comparison experiment of CCIN-SA on MOSI.

Model	Acc-2	F1-Score	MAE	Corr	Acc-7
ICA	83.32	83.12	0.867	0.697	40.31
GM-ICA	83.93	83.90	0.810	0.731	43.35
w/o TMSA	84.11	84.15	0.815	0.711	43.59
w/o GM	84.02	84.00	0.805	0.729	43.48
w/o ICA	83.89	83.58	0.823	0.705	43.37
CCIN-SA	**84.55**	**84.57**	**0.805**	**0.733**	**43.77**

Table 14
Comparison experiment of CCIN-SA on MOSEI.

Model	Acc-2	F1-Score	MAE	Corr	Acc-7
ICA	83.37	83.19	0.579	0.699	51.83
GM-ICA	84.70	84.91	0.545	0.736	54.12
w/o TMSA	84.94	84.97	0.541	0.741	54.17
w/o GM	84.57	84.59	0.565	0.733	53.49
w/o ICA	84.01	84.03	0.568	0.717	53.60
CCIN-SA	**85.26**	**85.28**	**0.534**	**0.745**	**54.36**

4.4.5. Comparison experiments of CCIN-SA model

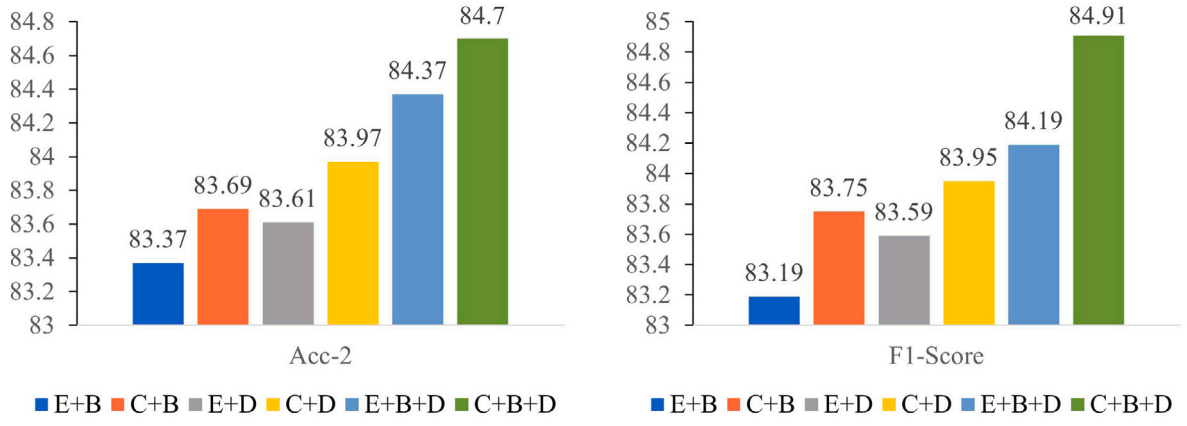
In order to verify the effect of CCIN-SA on the overall multimodal sentiment analysis results, the following control experiments were conducted. The results are shown in Tables 13 and 14.

The improved cross-modal interaction layer and the gating mechanism network layer (CCIN-TMSA) are removed, and a composite hierarchical fusion of the three modalities was performed directly after unimodal feature extraction; Remove the gating mechanism network layer (CCIN-GM) and directly perform composite hierarchical fusion of text–visual, visual–audio and audio–text cross-modal interaction features with and unimodal $Z_{TVA} = [Z_T, Z_V, Z_A]$ after improving the cross-modal interaction.

The cross-modal interaction layer (CCIN-ICA) was removed. After unimodal feature extraction, a modal two-by-two approach was taken to extract joint bimodal features using the gating mechanism network layer, and then composite level fusion was carried out.

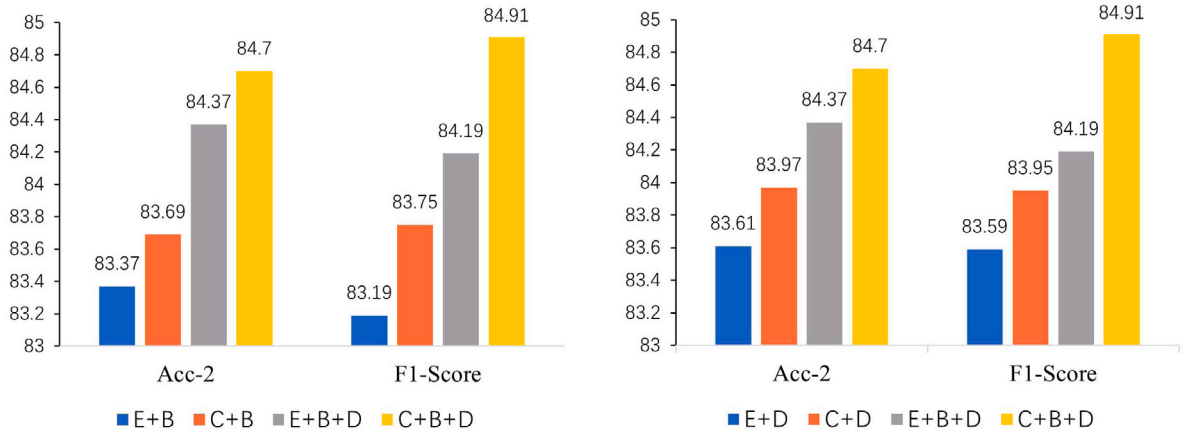
The experimental findings indicate that CCIN-SA outperforms all other models across all metrics. Composite hierarchical fusion method was used to continuously enhance the inter-modal interaction and increase the weights of important features within the modes. While considering inter-modality consistency, the particular details also affects the overall performance within each modality. The distinctive feature details from original modality can effectively improve the impact of noise in the low-order feature details.

Compared with the other models, the accuracy rates decrease to certain extent, indicating that the absence of any one module in the model can influence the overall model. The absence of the gating mechanism interaction network module results in the most notable degradation in the model's performance, indicating that the module is



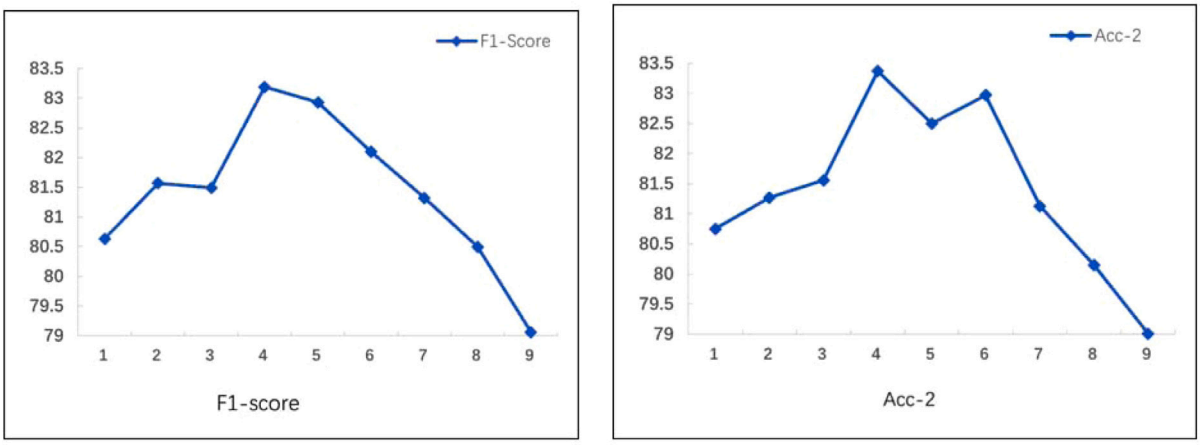
(a) Acc-2 (b) F1-Score

Fig. 10. Module ablation experiments of MOSEI.



(a) Ablation Experiments on Gating Mechanisms (b) Ablation Experiments on Cross-Modal Interaction Modules

Fig. 11. Results of Module Ablation Experiments on MOSEI.



(a) (b)

Fig. 12. Effect of the number of stacking layers of cross-modal module on the evaluation metrics for MOSEI.

Table 15
Modality ablation experiments of CCIN-SA on MOSI dataset.

Modalities combination	Acc-2	F1-Score	MAE	Corr	Acc-7
T	78.31	78.94	0.937	0.684	39.52
V	67.27	59.25	1.087	0.581	36.28
A	65.53	57.91	1.084	0.577	36.2
T+V	78.89	79.02	0.93	0.687	39.92
V+A	65.00	59.13	1.098	0.574	35.75
A+T	78.49	78.43	0.928	0.687	39.61
T+V+A	84.10	84.07	0.809	0.731	43.52
TVA+T+V+A	**84.55**	**84.57**	**0.805**	**0.733**	**43.77**

Table 16
Modality ablation experiments of CCIN-SA on MOSEI dataset.

Modalities combination	Acc-2	F1-Score	MAE	Corr	Acc-7
T	80.40	80.24	0.597	0.718	50.12
V	72.03	79.33	0.630	0.647	47.90
A	70.72	78.75	0.656	0.640	45.17
T+V	81.70	82.08	0.548	0.713	52.05
V+A	70.86	79.79	0.662	0.632	48.12
A+T	81.71	81.91	0.562	0.709	51.30
T+V+A	84.40	84.24	0.547	0.738	54.12
TVA+T+V+A	**85.26**	**85.28**	**0.534**	**0.745**	**54.36**

better at acquiring the interaction details across different modalities, deeper explore the potential correlation between modalities, and reflect the advancement and effectiveness of the GM-ICA model. The absence of cross-modal interaction layer makes the model drop by about 1%, indicating that the modal features enhanced by cross-modal feature information reflect stronger interaction. Unimodal temporal feature extraction layer contributes only about 0.5%, probably because the low-order unimodal information extracted by the multi-attention mechanism during the decision-level fusion process has already implied some of the temporal information, which leads to a low model enhancement.

4.4.6. Ablation study of CCIN-SA

In order to verify whether the multimodal combination makes up for the insufficient sentiment expression ability of single modality, this study carried out feature extraction for single modality, two–two combination modality and three-modal combination, and conducted sentiment analysis experiments. Unimodality only uses the unimodal temporal feature extraction technique to obtain feature information, which is directly used in sentiment classification tasks. Bimodality uses the CCIN method to realize feature fusion on the basis of unimodality. Tri-modality adopts the proposed CCIN-SA method to carry out feature extraction, and the obtained results are used in the final sentiment analysis task. Tables 15 and 16 illustrate modality ablation experiments.

Results indicate that the CCIN-SA model is the most effective for sentiment classification in composite hierarchies, followed by tri-modality, which fully demonstrates that multimodal fusion has a great impact on the performance. Also, composite hierarchical fusion method can further enrich the potential correlation between different modalities, thereby increasing the model's accuracy. In unimodal experiments, the binary classification accuracy and seven-classification accuracy of text sentiment analysis are the highest. In two-by-two modal combination, the modal performance of text-included modality is much higher than that of other modal combinations, which indicates that the sentiment characteristics of textual modal features are the most significant. Combined with the five evaluation indexes of classification results, the best results are achieved when using the three modal fusion features for classification. Therefore, the effective fusion of the three features, text, speech and image, contributes to the enhancement of sentiment classification performance.

5. Conclusions

This study proposed a composite cross-modal attention interaction fusion network for MSA. Firstly, internal temporal information of each modality was obtained by unimodal timing feature extraction technique. Second, through cross-modal attention mechanism, the lower-order information of the source modality was used to enhance the target modality to achieve potential adaptation to the source modality. By fusing the bimodal feature vectors after information enhancement of the improved cross-modal interaction, the deeper inter-modal correlations were further explored, and the fusion features of the two modalities were taken as conditional vectors to capture the similarities between different modalities, strengthen the correlation with important inter-modal interaction features, and weaken the correlation with the secondary inter-modal features. Finally, composite hierarchical fusion method was used to realize feature fusion of the information-enhanced cross-modal joint features with low-order signals by using the multi-attention mechanism, which can mine and retain the sentiment information of different modalities to the largest extent, and continuously strengthen the inter-modal interaction through composite hierarchical fusion. Considering that the distinctive feature details from original modality can influence the result of sentiment classification, increase the weights of important features within the modality, and retain the important features of the initial modality, on the basis of three modality fusion, the composite hierarchical fusion was conducted by using structure similar to residual network. This study proves that the CCIN-SA has certain superiority in MSA's performance.

Based on the preprocessing method, the current unimodal feature extraction technique has been improved, but the improvement after multimodal fusion is not obvious. Exploring implicit correlation information between different modes is still our main task. Therefore, further study will be carried out to explore the modal fusion problem and the noise problem. For the CCIN-SA model, the addition of multiple fusion layers increases model complexity, potentially leading to a significant rise in floating-point operations (FLOPs) and higher experimental configuration/hardware requirements. In the next work, pruning or distillation techniques will be explored for efficiency optimization. While this study primarily utilizes publicly available English datasets, future efforts will extend the research to non-English and medical domain datasets.

CRediT authorship contribution statement

Li Yang: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Junhong Zhong:** Methodology, Investigation. **Teng Wen:** Writing – review & editing, Software. **Yuan Liao:** Writing – review & editing, Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] L. Songning, H. Xifeng, X. Haoxuan, R. Zhaoxia, L. Zhi, Multimodal sentiment analysis: A survey, Displays 80 (2023) 102563.

[2] Y. Sun, Z. Liu, Q.Z. Sheng, D. Chu, J. Yu, H. Sun, Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities, Inf. Fusion 110 (2024) 102454.

[3] J.M. Liu, P.X. Zhang, Y. Liu, et al., Summary of multimodal sentiment analysis technology, J. Front. Comput. Sci. Technol. 15 (7) (2021) 1165–1182.

[4] K. Kim, S. Park, AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis, Inf. Fusion 92 (2023) 37–45.

[5] C. Zhu, M. Chen, S. Zhang, C. Sun, H. Liang, Y. Liu, J. Chen, SKEAFN: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis, Inf. Fusion 100 (2023) 101958.

[6] Z. Fu, F. Liu, Q. Xu, et al., NHFNET: a non-homogeneous fusion network for multimodal sentiment analysis, in: Proc of IEEE International Conference on Multimedia and Expo, IEEE, New York, NY, 2022, pp. 1–6.

[7] B. Liang, H. Su, L. Gui, et al., Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks, Knowl.-Based Syst. 235 (2022) 107643..

[8] S. Liu, P. Gao, Y. Li, W. Fu, W. Ding, Multi-modal fusion network with complementarity and importance for emotion recognition, Inform. Sci. 619 (2023) 679–694.

[9] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 8 (8) (2018) e1253.

[10] Y. An, J. Zhang, L. Xia, A survey of sentiment analysis on social media, Data Anal. Knowl. Discov. 4 (1) (2020) 1–11.

[11] L. Liu, Y. Yang, J. Wang, ABAFN: Aspect-based sentiment analysis model for multimodal, Comput. Eng. Appl. 58 (10) (2022) 193–199.

[12] M. Lin, Z. Meng, Multimodal sentiment analysis based on attention neural network, Comput. Sci. 47 (11) (2020) 508–514.

[13] Q. Liu, D. Zhang, L. Wu, et al., Multi-modal sentiment analysis with context-augmented LSTM, Comput. Sci. 46 (11) (2019) 181–185.

[14] X. Yan, H. Xue, S. Jiang, et al., Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling, Appl. Artif. Intell. 36 (1) (2021) 2000688.

[15] D. Gkoumas, Q. Li, C. Lioma, et al., What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis, Inf. Fusion 66 (2021) 184–197.

[16] M. Hou, J. Tang, J. Zhang, et al., Deep multimodal multilinear fusion with high-order polynomial pooling, Adv. Neural Inf. Process. Syst. 32 (2019) 12113–12122.

[17] G.-B. Bian, J.-Y. Zheng, Z. Li, J. Wang, P. Fu, C. Xin, D.S. da Silva, W.-Q. Wu, V.H.C.D. Albuquerque, BiMNet: A multimodal data fusion network for continuous circular capsulorhexis action segmentation, Expert Syst. Appl. 238 (2024) 121885.

[18] A. Zadeh, P. Liang, N. Mazumder, et al., Memory fusion network for multi- view sequential learning, 2018, ArXiv 1802.00927.

[19] P. Liang, L. Z., A. Zadeh, et al., Multimodal languageanalysis with recurrent multistage fusion, 2018, ArXiv 1808.03920.

[20] A. Zadeh, M. Chen, S. Poria, et al., Tensor fusion network for multimodal sentiment analysis, in: 2017 Conference on Empirical Methods in Natural Language Processing, 2017 Conference on Empirical Methods in Natural Language Processing, ACL, Copenhagen, 2017, pp. 1103–1114.

[21] Z. Liu, Y. Shen, V. Lakshminarasimhan, et al., Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 2018 56th Annual Meeting of the Association for Computational Linguistics, 2018 56th Annual Meeting of the Association for Computational Linguistics, ACL, Stroudsburg, 2018, pp. 2247–2256.

[22] S. Mai, S. Xing, H. Hu, Locally confined modality fusion network with a global perspective for multimodal human af-fective computing, IEEE Trans. Multimed. 22 (1) (2019) 122–137..

[23] Y. Tsai, P. Liang, A. Zadeh, et al., Learning factorized multimodal representations, in: Proceedings of the 7th International Conference on Learning Representations, the 7th International Conference on Learning Representations, OpenReview.net, New Orleans, 2019.

[24] D. Hazarika, R. Dmmermann, S. Poria, et al., MISA: modality-invariant and -specific representations for multimodal sentiment analysis, 2020, ArXiv 2005.03545.

[25] W. Rahman, M. Hasan, S. Lee, et al., Integrating multimodal information in large pretrained transformers, in: Proc of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Stroudsburg, 2020, pp. 2359–2369.

[26] W. Yu, H. Xu, Z. Yuan, et al., Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proc of the 35th AAAI Conference on Artificial Intelligence, AAAI, Palo Alto, 2021, pp. 10790–10797.

[27] Y. Tsai, S. Bai, P. Linag, et al., Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, the 57th Annual Meeting of the Association for Computational Linguistics, ACL, Florence, 2019.

[28] C. Gan, X. Fu, Q. Feng, Q. Zhu, Y. Cao, Y. Zhu, A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis, Expert Syst. Appl. 242 (2024) 122731.

[29] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, the 2015 Conference on Empirical Methods in Natural Language Processing, ACL, Lisbon, 2015, pp. 1422–1432.

[30] G. Degottex, J. Kane, T. Drugman, et al., COVAREP-a collaborative voice analysis repository for speech technologies, in: Proceedings of the 2014 IEEE International Conference on Acoustics, the 2014 IEEE International Conference on Acoustics, IEEE, Florence, Italy, 2014, pp. 960–964.

[31] A. Zadeh, P. Liang, S. Poria, et al., Multi-attention recurrent network for human communication comprehension, in: Proceedings of the 2018 32nd AAAI Conference on Artificial Intelligence, the 2018 32nd AAAI Conference on Artificial Intelligence, AAAI, Palo Alto, 2018, pp. 5642–5649.

[32] Y. Ma, Research on Emotion Recognition Methods Based on Speech and Facial Expressions (mathesis), Huazhong University of Science and Technology, 2023.

[33] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, the 2016 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas,USA, 2016, pp. 770–778.

[34] A. Zadeh, R. Zellers, E. Pincus, et al., Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88.

[35] A. Zadeh, P. Liang, S. Poria, et al., Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, the 56th Annual Meeting of the Association for Computational Linguistics, ACL, Melbourne, 2018, pp. 2236–2246.