Mathematical Biosciences and Engineering

MBE, 21(2): 2488–2514.
DOI: 10.3934/mbe.2024110
Received: 16 November 2023
Revised: 29 December 2023
Accepted: 08 January 2024
Published: 17 January 2024

http://www.aimspress.com/journal/MBE

Research article

Research on cross-modal emotion recognition based on multi-layer semantic fusion

Zhijing Xu and Yang Gao*

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

* **Correspondence:** Email: gy_kaixia0728@163.com.

Abstract: Multimodal emotion analysis involves the integration of information from various modalities to better understand human emotions. In this paper, we propose the Cross-modal Emotion Recognition based on multi-layer semantic fusion (CM-MSF) model, which aims to leverage the complementarity of important information between modalities and extract advanced features in an adaptive manner. To achieve comprehensive and rich feature extraction from multimodal sources, considering different dimensions and depth levels, we design a parallel deep learning algorithm module that focuses on extracting features from individual modalities, ensuring cost-effective alignment of extracted features. Furthermore, a cascaded cross-modal encoder module based on Bidirectional Long Short-Term Memory (BILSTM) layer and Convolutional 1D (ConV1d) is introduced to facilitate inter-modal information complementation. This module enables the seamless integration of information across modalities, effectively addressing the challenges associated with signal heterogeneity. To facilitate flexible and adaptive information selection and delivery, we design the Mask-gated Fusion Networks (MGF-module), which combines masking technology with gating structures. This approach allows for precise control over the information flow of each modality through gating vectors, mitigating issues related to low recognition accuracy and emotional misjudgment caused by complex features and noisy redundant information. The CM-MSF model underwent evaluation using the widely recognized multimodal emotion recognition datasets CMU-MOSI and CMU-MOSEI. The experimental findings illustrate the exceptional performance of the model, with binary classification accuracies of 89.1% and 88.6%, as well as F1 scores of 87.9% and 88.1% on the CMU-MOSI and CMU-MOSEI datasets, respectively. These results unequivocally validate the effectiveness of our approach in accurately recognizing and classifying emotions.

Keywords: multimodal emotion recognition; multimodal fusion; cascade encoder; inter-modal information complementation; Mask-gated Fusion Networks (MGF-module)

1. Introduction

Social media has become a vital medium for sharing and receiving information, containing implicit emotional content in text, voice, and video formats. The recorded data holds significant potential to support decision-making processes [1]. Over the past few years, sentiment analysis research has made considerable progress [2]. However, previous research solely relied on unimodal sentiment analysis for judgments, resulting in unsatisfactory accuracy due to the limited and easily influenced sentiment features extracted from a single modality. Given that people operate in a multimodal environment, there is an increasing demand in artificial intelligence for multimodal capabilities, aiming to replicate human-like meticulous observation and rich emotional understanding to judge human emotions based on specific features. Consequently, comprehensive analysis of multiple modalities for emotion judgment has garnered attention among researchers in the field of emotion analysis, as it can significantly enhance recognition accuracy. Emotion recognition serves as the foundation for emotional interaction and has witnessed numerous advancements in human-computer interaction [3]. The powerful analytical and decision-making capabilities of sentiment recognition technology have broad applications across various aspects of life.

Sentiment analysis based solely on unimodal modalities can yield unsatisfactory results due to the homogeneity of emotions within each modality, leading to suboptimal test outcomes in the presence of interference issues. Consequently, in subsequent studies, researchers have turned to multimodal combined analysis for sentiment determination, recognizing that different modalities can complement each other to enhance sentiment recognition accuracy. For instance, Yoon et al. [4] proposed the Multimodal Dual Recurrent Encoder (MDRE) architecture, which leverages both textual and audio data to understand speech using a recurrent neural network (RNN) to predict sentiment categories by integrating information from both audio and text sequences. Similarly, Jeong et al. [5] introduced a method to enhance the performance of speech sentiment recognition through multimodal cue learning with a pre-trained text-based model, employing on-the-fly learning using textual and audio information via a language model pre-trained on natural language text. Furthermore, Batbaatar et al. [6] presented the Semantic Emotional Neural Network (SENN), a neural network architecture utilizing bi-directional long and short-term memory (BiLSTM) to capture contextual information and emphasize semantic relationships, as well as a Convolutional Neural Network (CNN) to extract affective features focusing on emotional word relationships in the text. This approach effectively incorporates semantic, syntactic, and emotional information by employing pre-trained word representations.

In the realm of multimodal emotion recognition, numerous researchers have focused on data fusion and feature fusion. Zadeh et al. [7] introduced the tensor fusion network (TFN) pair, which computes the correlation between two or three modalities while preserving unimodal correlation, but this approach significantly increases feature dimension. The resulting increase in feature dimension impacts computational efficiency and memory consumption, leading to exponentially escalating time and space complexity as the number of input modalities grows, thereby raising the risk of overfitting with a high number of parameters. In response, Liu et al. [8] employed low-rank weight for multimodal fusion, reducing the number of parameters and enhancing computational speed, representing an

improvement over the challenges associated with TFN. Mai et al. [9] proposed a graph fusion network to simulate interactions between different modalities. Kratzwald et al. [10] developed sent2affect, a transfer learning model for affective computing. This specific task necessitates the customization of recurrent neural networks for bi-directional processing, the utilization of loss layers for regularization, and the application of weighted loss functions. Zheng et al. [11] presented a new task termed EMER, which differs from traditional emotion recognition as it not only predicts emotional states but also provides explanations for these predictions. Moreover, EMER aims to address the enduring challenge of label ambiguity and enhance the reliability of emotion recognition systems.

The advancement of cascade modeling in natural language processing has significantly enhanced the performance of data encoders. These encoders, comprising a combination of multiple networks, have proven effective in focusing on different dimensions and depth levels of input data, leading to improved results in sentiment recognition. For instance, Sun et al. [12] developed a speech encoder by cascading Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to extract deep semantic features. They then integrated these features with textual information at a feature layer, effectively capturing emotional information embedded in speech and achieving a 4% to 5% improvement in final accuracy compared to using a single network as a data encoder. In a similar way, Liu et al. [13] utilized three cascaded channels based on deep learning techniques to extract features for the fusion of three modalities in successive phases. Additionally, Lee et al. [14] employed a heterogeneous feature fusion approach based on BERT to merge multiple low-level and high-level features of text and speech for emotion recognition. Furthermore, Kumar et al. [15] applied cascaded ResNet and Deep CNN models for recognizing facial features in a given facial image. These studies collectively demonstrate the efficacy of cascade modeling in leveraging multiple networks to enhance feature extraction and fusion across various modalities, thereby advancing the field of sentiment recognition.

The above approach is an aggregation-based fusion paradigm, but the gap between modalities severely impairs multimodal fusion. In order to bridge the modal gap, there is some work applying the attention mechanism to multimodality can satisfy the features extracted by different models to realize the alignment operation [16], and make one modality potentially adapt to the other modality to achieve modal complementarity. For different modalities, they are fused into the joint embedding space to extract common features, and the diversity of each modality as well as private features are also considered. In order to reduce the distribution gap and redundant information between different modalities and obtain more accurate modal representations. Hazarika et al. [17] used similarity loss and difference loss to explore the consistency and complementarity between multiple modalities. Yang et al. [18] proposed a feature de-entanglement multimodal emotion recognition (FDMER) method to learn the public and private feature representations of each modality. Han et al. [19] used the cascaded modular multimodal cross-attention network to capture deep textual semantic information. Zaidi et al. [20] design a multimodal dual-attention transformer to refine feature representations while enhancing cross-modal and cross-linguistic interactions.

Previous multimodal approaches help to utilize complementary information across modalities [21]. Previous research in cross-modal studies has primarily focused on leveraging the complementary information between different bimodalities, such as images and text, audio and text, etc. Various common approaches have been utilized, including fusion models based on Convolutional Neural Networks (CNNs) and recurrent neural networks (RNNs), attention-based fusion models, and deep learning models. Specifically, in the context of bimodal information complementarity, cross-attention

mechanisms are commonly employed to guide one modality to attend to the other modality for updating features accordingly. For instance, Sun et al. [12] introduced the Multi-Modal Cross and Self-Attention Network (MCSAN), which addresses the effective fusion of two modalities by utilizing self-attention mechanisms to propagate information within each modality. Yang et al. [22] proposed Cross-Modal BERT (CM-BERT), which fine-tunes pre-trained BERT models through bimodal interactions, and incorporates a masked multi-modal attention mechanism for dynamically adjusting word weights by combining text and audio modalities. In recent years, with advancements in technologies like virtual reality and augmented reality, research on cross-modal information complementarity has extended towards the combination of multiple modalities, such as audiovisual, haptic, etc. Yang et al. [23] presented the Self-Adaptive Context and Modality Interaction Modeling (SCMM) framework, which comprises three interaction sub-modules designed to handle different levels of complexity in cross-modal interactions. By effectively exploiting the potential complementarity between these modalities, multi-modal features can exhibit greater discriminative power, resulting in improved performance compared to single-modal models. Paraskevopoulos et al. [24] proposed a neural network architecture that incorporates feedback mechanisms during forward propagation to capture top-down cross-modal interactions and extract high-level representations of each modality.

While the extraction of semantic features from each modality using deep learning algorithms has shown improvement, there is considerable potential for enhancing inter-modal interaction, information complementation, and adaptive feature extraction. To address these challenges, a cross-modal emotion recognition model based on multilayer semantic fusion is proposed. The model can perform semantic supplementation of modal features, fully utilize the complementary information of other modalities for reinforcement operations in case of partial missing sentiment data, adaptively focus on the importance of different modal features in sentiment categorization, and capture intra- and inter-modal contained sentiment information. By utilizing multimodal information and learning contextual relationships, the model can adaptively prioritize the importance of different modal features in emotion classification, thereby addressing challenges such as low recognition accuracy and emotion misclassification resulting from feature complexity, noise, and redundant information.

The contributions of this paper can be summarized as follows:

1) We present the CM-MSF deep learning model to tackle cross-modal semantic complementarity and multi-level fusion. In order to achieve a more thorough and enriched feature extraction across various dimensions and depths, this research utilizes a dual encoder structure for extracting modality-specific features. Furthermore, a cost-effective modality alignment technique is employed to align the extracted feature vectors. This improvement significantly enhances the overall performance of the CM-MSF model in capturing and integrating cross-modal information.

2) By employing a cascaded encoder with embedded Bidirectional Long Short-Term Memory (BiLSTM) layers and Convolutional Neural Network (Conv1d), we achieve adaptive modality interaction within the modality interaction module. This approach enables the complementary integration of multiple modalities by reducing differences between modalities and eliminating redundant information caused by signal heterogeneity. The BiLSTM layers enhance the modeling of sequential structural information, while the Conv1d layers perform convolutional operations on the input, capturing local details and dependencies between adjacent tokens. This integration of Conv1d layers significantly improves the model's ability to model local structures and extract language features from the input sequence.

3) In the final module, we introduce the Mask-gated Fusion Module (MGF-module). This

innovative module combines mask techniques with gate structures, allowing for precise control over the weights and influences of different modalities during the feature fusion process. As a result, it enables flexible and adaptive selection and propagation of information. The gate mechanism intelligently filters and adjusts the features of each modality, effectively governing the flow of information from each modality.

2. Related work

Multimodal sentiment analysis can capture the diversity and complexity of human emotional expressions more comprehensively by combining features from different modalities. These data sources are fused and combined to improve the performance and effectiveness of the model [25]. Some commonly used fusion methods are described below, each of which has its advantages and disadvantages：

1) Early Fusion: Features from different data sources are spliced together to form a larger feature vector, which is then fed into the model for training. The advantage of this method is that it is simple and direct, but it may lead to too much noise and redundant information.

2) Late Fusion: Features from different data sources are input into different models for training, and then the outputs of the models are fused. The advantage of this method is that it can make full use of the information of each modality, but it requires multiple training and fusion, and the computation is large.

3) Cross Fusion: A model is used to process features from different data sources simultaneously, and the relationship between different data sources is fused together by sharing a part of the parameters. The advantage of this method is that it can deal with the interaction effects between different data sources, but it requires complex parameter design and adjustment.

4) Attention-based Fusion: Uses the attention mechanism to weight the features of different data sources for fusion, and reinforces the importance of each data source to the model output. The advantage of this method is that it can dynamically learn the weights between data sources, but it requires complex model design and training.

5) Multi-level Fusion: Hierarchical combination of different fusion methods to form a multi-level fusion structure. The advantage of this method is that it can fully utilize the advantages of different fusion methods and reduce redundant information and noise interference, but it requires complex structure design and adjustment.

In the previous modal fusion, one fusion method is used for the model. The limitations are that the generated high-dimensional features are oriented in such a way that complex relationships cannot be modeled and the interconnections between different modalities cannot be captured. Zhang et al. [26] fused speech emotion features and face emotion features using DBN network and then trained them to learn new emotion features after fusion of the two modalities. Hazarika et al. [27] used self-attention mechanism features to fuse audio features and text features to get new sentiment features. Hossain et al. [28] used Extreme Learning Machine (ELM) to fuse speech and video features at feature level. The features of the full connectivity layer will be extracted and feature fusion will be performed by two consecutive ELMs. The first ELM contains 100 hidden units and the second ELM contains 250 hidden units. Then, the output of the second ELM is sent to SoftMax and SVM for emotion recognition to get the final result. Cheng et al. [29] introduced the multilayer feature fusion (MFF) model to improve the effectiveness of multimodal fusion by fusing different levels of features through different methods. Wang et al. [30] used feature-level and decision-level fusion to classify the features extracted from

EEG signals and facial expressions.

Model-level fusion aims at obtaining a joint feature representation of the three modalities, and its implementation depends mainly on the fusion model used. Model-level fusion is a deeper fusion approach that produces more optimized joint discriminative feature representations for classification and regression tasks, taking full account of the relationships between the models. In the modal interaction module of this study, a cascade model is designed to fuse two of the three modalities separately to realize the flow of information. Feature-level fusion of unimodal features and features after modal interaction is performed to enhance feature distinctiveness. The model presented in this paper employs a hybrid approach, combining both model-level and feature fusion, resulting in significantly enhanced effectiveness. The utilization of multimodal fusion contributes to improved emotion recognition performance by leveraging the complementary characteristics of different modalities. Model-level fusion, in particular, excels in harnessing the strengths of deep neural networks more effectively compared to decision-level and feature-level fusion techniques.

3. Methodological algorithm

3.1. System model

To effectively leverage the complementary nature of essential information across modalities and extract and filter advanced features in an adaptive manner, we present a multi-layer semantic fusion-based Cross-Modal Sentiment Recognition (CM-MSF) model. The framework is presented (details can be found in the following sections) in Figure 1. The model in this paper is divided into four major modules: Feature extraction, unimodal feature processing, cascading cross-modal information interaction, fusion and classification.

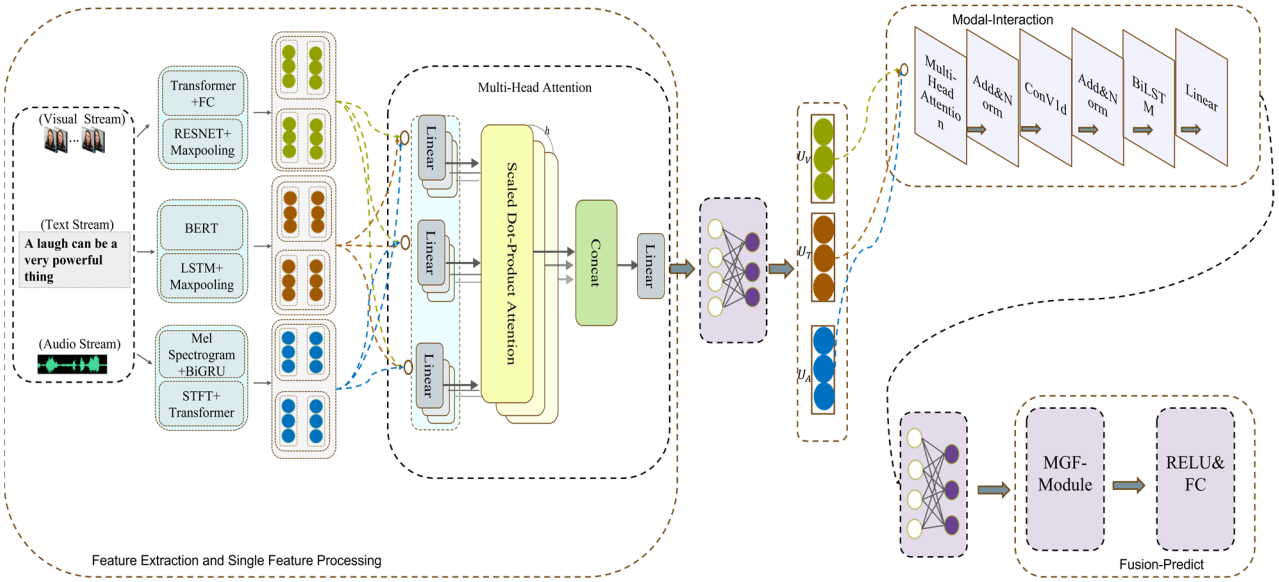


Figure 1. Framework diagram of the CM-MSF network proposed in this paper.

First, we employ dual encoders to extract temporal features from unimodal sequence data at various dimensions and depths. This allows us to enhance the model's ability to represent and learn

from individual modalities. We then perform feature fusion and implicit alignment using a multi-head attention mechanism, which significantly improves the integration of unimodal data. To reduce dimensionality, we apply a fully-connected layer. Next, we utilize cascaded cross-modal encoders to facilitate the flow of information between modalities, giving higher attention weights to important information. This ensures that crucial details are properly emphasized. Finally, we pass the complemented features through a mask gating network, enabling flexible adaptive filtering and masking of information. The fused multimodal representations are then fed into a fully connected layer for the final sentiment prediction.

3.2. Multi-model integration

In recent studies, researchers have explored methods to enhance the performance of emotion recognition by constructing data encoders using multiple network models cascaded together. This approach allows for the specialization of each model in capturing specific dimensional features of the input data, resulting in more comprehensive emotional representations. However, increasing the number of cascaded layers can lead to a potential risk of gradient explosion. To address this issue, we propose a parallel approach in this study where different data encoders process each modality independently. By doing so, we effectively mitigate the risk of gradient explosion caused by stacking models, thereby optimizing the overall performance of emotion recognition.

3.2.1. Text data feature extraction

In the text modality, BERT and LSTM are chosen to extract text features. BERT, a pre-trained language model based on the Transformer architecture, has demonstrated exceptional success in natural language processing tasks. Through unsupervised learning on extensive textual data, BERT can acquire a rich representation of contextually relevant word vectors, capturing semantic relationships between words and dependencies at the sentence level. This makes BERT highly effective for feature extraction in the text modality. On the other hand, LSTM, a variant of recurrent neural networks (RNN), has shown outstanding performance in modeling sequence data. With its memory units and gating mechanisms, LSTM adeptly handles long-term dependencies in textual sequences, effectively capturing temporal information and contextual relevance within the text. For tasks like sentiment analysis and text generation, LSTM excels in the text modality. The combination of BERT and LSTM enables more comprehensive extraction of text modality features, integrating word-level and sentence-level representations to express semantic and contextual information more comprehensively.

3.2.2. Speech data feature extraction

The Transformer model excels in semantic modeling as it can learn contextual dependencies in speech signals and capture higher-level semantic information. It achieves this through the self-attention mechanism, which allows it to acquire relevant information globally and assign importance to each signal during feature extraction, thereby enhancing the interpretability of the model. On the other hand, the bidirectional BiGRU is capable of capturing temporal sequence information. By processing both forward and backward sequence information, it effectively utilizes contextual information, making it highly effective for modeling long temporal dependencies in speech tasks. By utilizing both models

side by side, features can be extracted at different time scales, resulting in a more comprehensive representation of the speech signal. Parameter sharing between the Transformer and bidirectional BiGRU models improves the training process efficiency and enhances the generalization performance over the dataset. Additionally, both models are naturally computationally parallel and can be computed quickly on GPUs or other accelerated devices. Figure 2 showcases the BiGRU structure.

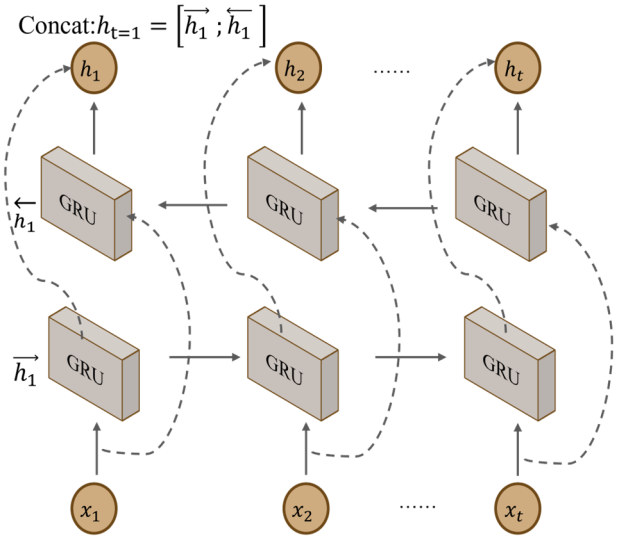


Figure 2. Structural diagram of BiGRU.

3.2.3. Video data feature extraction

As a deep Convolutional Neural Network, ResNet has gained widespread recognition and achieved remarkable results in the field of image processing. It excels in visual feature representation learning and possesses strong capabilities in this regard. Furthermore, ResNet exhibits excellent temporal modeling abilities, enabling it to effectively model and learn the time series relationships within videos. When comparing ResNet-101 to ResNet-50, ResNet-101 is more suitable as it can better capture fine-grained information and richer features. Additionally, the Transformer model is capable of capturing temporal sequence information in various types of sequential data, including video. Through its self-attention mechanism, the Transformer can adaptively allocate attention to different regions within the video during the learning process. This enables the model to better focus on important spatial regions and time segments within the video. By combining these two models side by side in video processing tasks, it becomes possible to fully capture the spatial, temporal, and semantic information embedded in the video. This integration enhances the feature representation of video content, leading to improved performance and understanding of the video data.

3.2.4. Data preprocessing

In this paper, the video stream is considered as a temporal sequence, where each time step corresponds to a frame of the video. The frame sequence of the video is fed into the model to obtain the hidden state sequence for each time step. Subsequently, the temporal sequence information is

transformed into a fixed-dimension video feature vector using the maximum pooling operation. Similarly, text data undergoes a comparable process. The audio information stream is preprocessed to prepare it for model input through short-time Fourier transform. Specifically, the audio stream is preprocessed into a suitable format for further processing, known as the Mel spectrogram. This transformed representation is then utilized to extract feature vectors from the audio stream.

3.3. Unimodal attention mechanism alignment and convergence

In the model presented in this paper, the features acquired by different deep learning models for each modality undergo feature alignment and fusion using the multi-head attention mechanism. This approach enables the computation of multiple attention heads in parallel, allowing each head to learn distinct correlation patterns and conduct feature alignment and fusion. As each attention head can focus on specific aspects, it becomes more adept at capturing feature dependencies and local patterns, thereby enhancing its modeling capabilities. For instance, Figure 3 illustrates the process of aligning and fusing text features using the multi-head attention mechanism.

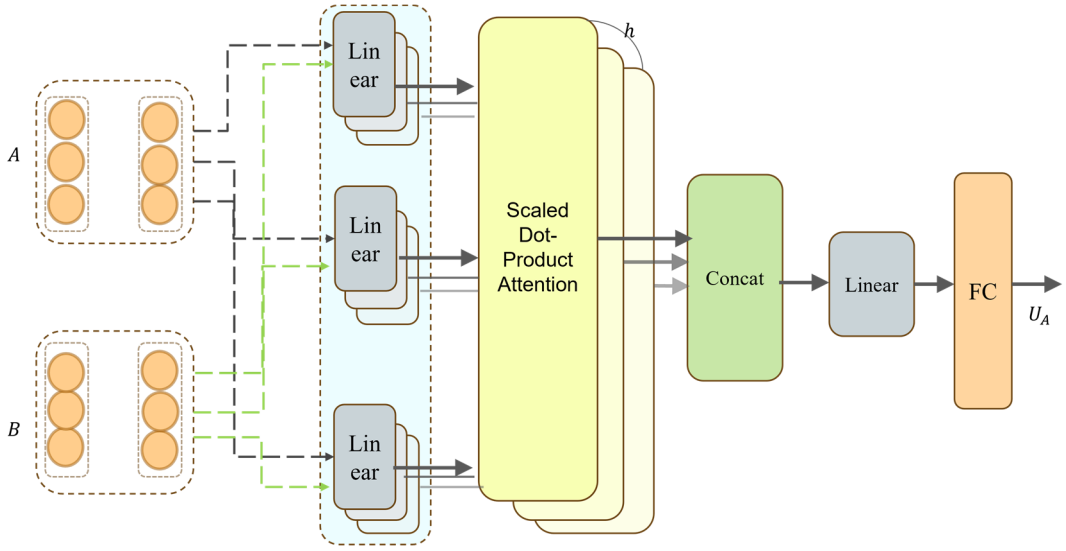


Figure 3. Fusion alignment model diagram for text modality.

Multi-model feature alignment and fusion steps for text based on multiple attention mechanisms:

Step1: The text is represented by two features extracted by BERT with LSTM $A = \{a_1^t, a_2^t, \ldots \ldots a_{nA}^t\}$ and $B = \{b_1^t, b_2^t, \ldots \ldots b_{nB}^t\}$, their dimensions are nA, nB respectively.

Step 2: Linear transformation and attention score calculation. In order to introduce the attention mechanism and map the feature vectors into the attention space, a linear transformation is first applied to A and B. Using different weight matrices $W_i^Q \in R^{nT \times d_q}, W_i^K \in R^{nT \times d_k}, W_i^V \in R^{nT \times d_v}$ transform them into query (query), key (key) and value (value) representations.

For the feature vector A, compute multiple attention heads as follows:

$$AttentionScores(ai) = softmax\left(Q_a \cdot \frac{(K_b W_{ai}^K)}{\sqrt{d_a^K}}\right) \tag{1}$$

The above equation takes the shape of (nA, nB), where W_{ai}^{K} is the key transformation matrix of the i-th attention head, and $\cdot$ denotes the matrix multiplication.

For the feature vector B:

$$AttentionScores(bi) = softmax\left(Q_b \cdot \frac{(K_a W_{bi}^{K})}{\sqrt{d_b^K}}\right) \tag{2}$$

The above equation takes the shape of (nB, nA), where W_{bi}^{K} is the key transformation matrix of the i-th attention head. SoftMax function is used to normalize the attention scores to a probability distribution that guarantees that the sum of the attention scores in each row is 1. Here the attention scores of multiple attention heads can be computed, capturing the diversity of information among the different heads.

where $Q_T = AW_i^Q$, $K_T = AW_i^K$, $V_T = AW_i^V$ are three vectors generated by linear variation of the text sequence, which correspond to query vector, key vector and value vector, respectively. Where $T \in (A, B), Q_T, K_T, V_T$ have the shapes of $(nT, d_{qT}), (nT, d_{kT}), (nT, d_{vT}), d_{qT}, d_{kT}, d_{vT}$ are the dimensions of query vector, key vector, and value vector respectively.

Step 3: Feature alignment and fusion.

$$C_{ai} = AttentionScores_{ai} \cdot (V_b \cdot W_{ai}^{V}) \tag{3}$$

Based on the attention scores of each attention head, the fused feature vectors are computed using the above equation in the shape of (nA, d_{vb}), where W_{ai}^{V} is the value transformation matrix of the i-th attention head.

$$C_{bi} = AttentionScores_{bi} \cdot (V_a \cdot W_{bi}^{V}) \tag{4}$$

Use this equation to compute the fused feature vectors of feature vector B.

Step 4: Multi-head mechanism. The multi-head attention mechanism introduces multiple attention heads, each with different weight matrices W_q, W_k, W_v. Each of the h heads produces a feature fusion result, and finally these results are stitched together to obtain the final feature vector.

$$C_T = concatC_{Ti} \tag{5}$$

Take all the fused feature vectors C_{ai}, C_{bi} Splice them by columns with shapes of $(nA, d_{vb} \times h), (nB, d_{va} \times h)$.

The final aligned and fused feature vectors are obtained by performing a linear transformation on the spliced feature vectors C_a, C_b The shapes are respectively $(nA, d), d = d_{vb} \times h$ and $(nB, d), d = d_{va} \times h$.

Step 5: Dimensionality transformation. Finally, the two fused feature vectors obtained are linearly varied to obtain the final deep feature representation of the text.

$$H_T = W \cdot [C_a, C_b] \tag{6}$$

Define a weight matrix W, with dimension $m \times (d_{vb} \times h + d_{va} \times h)$. In the above equation, the $[C_a, C_b]$ denotes the input matrix formed by concatenating two feature vectors by columns of a $(d_{vb} \times h + d_{va} \times h) \times 1$ of the column vectors. Matrix multiplication $W \cdot [C_a, C_b]$ Multiply this

input vector with the weight matrix W to obtain an output vector of $m \times 1$. That is, the output feature vector. Similarly based on the above five-step algorithm, the video and speech processed features are obtained H_V, H_A.

Step 6: Tri-modal alignment. For better modal interaction, in this paper, the obtained modal representations are fed into the fully connected layer for inter-modal dimension conversion. The number of output nodes of the fully connected layer is set to d, the desired feature dimension.

$$U_i = FC[ReLU(H_i \cdot W_o + b_1)] \tag{7}$$

where W_o is the weight matrix of the fully-connected layer with size (d, m), b_1 is the set bias vector $(d, 1)$, and $H_i = \{H_{i1}, H_{i2}, \ldots \ldots H_{id}\}, i = \{A, T, V\}$, the input feature vector is computed by forward propagation through the fully connected layer. That is, the input vector is multiplied with the weight matrix and the bias vector is added and then nonlinearly transformed by the ReLU activation function. Where W_o and b_1 are trainable parameters. According to the loss function, the gap between the output feature vector and the target value is calculated and the weights and bias of the fully connected layer are updated to reduce the loss through the back propagation algorithm.

3.4. Cascade model cross-modal information interaction

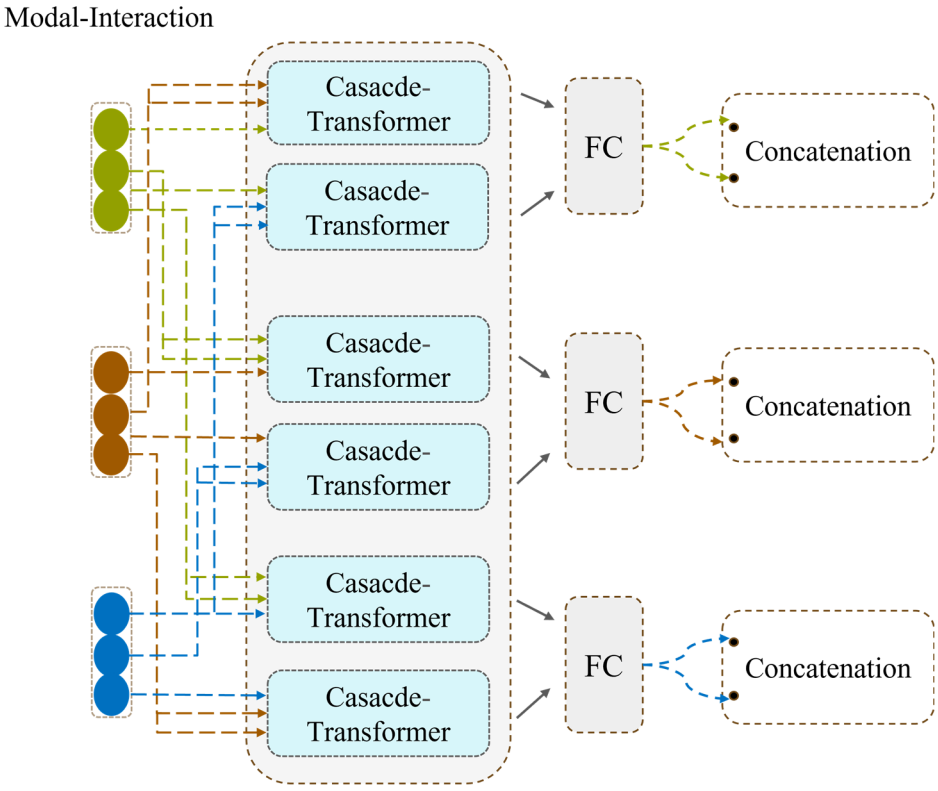


Figure 4. Flowchart of the intermodal interaction.

Following the aforementioned sequence of operations, given the multimodal features U_A, U_V, U_T, the multimodal interaction module takes these three feature vectors as input. By effectively leveraging the potential complementarity of information across these modalities, the multimodal features can be

rendered more discriminative, thereby enabling the model to outperform a unimodal model. In this module, we consider the incorporation of a modified cascaded cross-modal model for cross-modal information interaction fusion [31] (as shown in Figure 4), which integrates a Bidirectional Long Short-Term Memory (BiLSTM) layer to enhance sequential structural information. Pairwise complementary information interactions are conducted, potentially allowing the flow of information from one modality to the other through attention mechanisms. The introduction of the Conved1 layer within the Transformer encoder involves subjecting the input to a convolutional operation aimed at capturing local information and dependencies among adjacent tokens. Unlike traditional Transformer encoders that solely employ a self-attentive mechanism to capture global information in the input, without explicitly considering the local features and contextual information of each token in the sequence, the inclusion of the Conved1 layer enhances the model's capacity to model local structures and effectively captures linguistic features within the input sequence [32]. Moreover, this convolutional layer operation can to some extent reduce the number of model parameters, thereby lowering computational complexity and training costs.

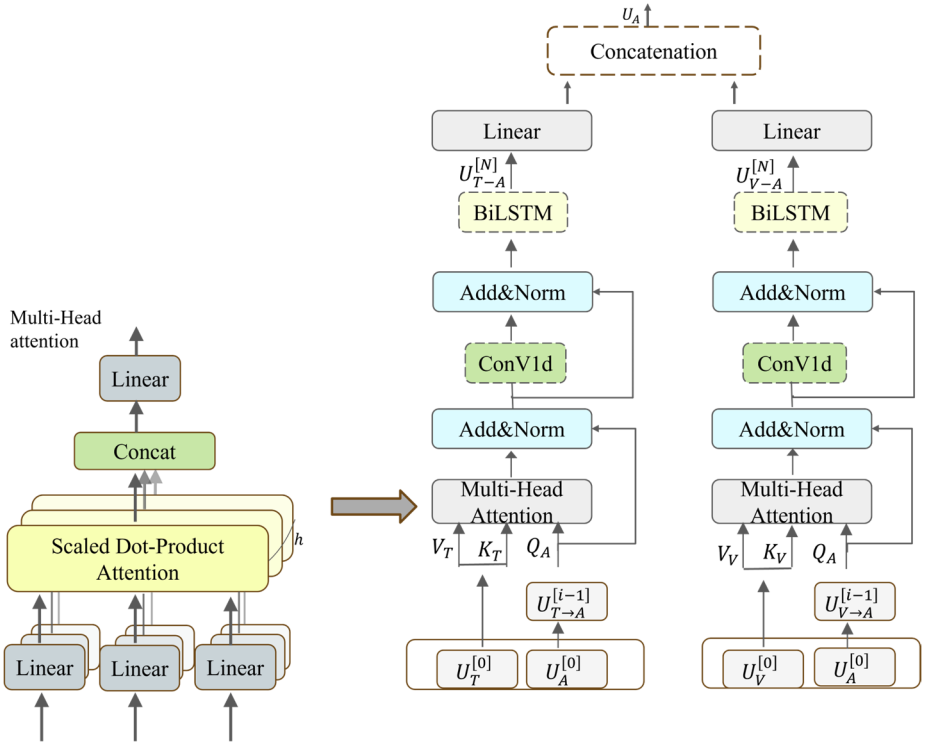


Figure 5. Cascade cross-modal information interaction map.

This module takes an encoder as an example, as shown in Figure 5, in cross-modal information interaction based on the cascade cross-modal model, to understand is the association between the video and the text, it is first necessary to serially convert each feature into a query, a key, and a value using a linear projection: $Q_A = U_A W_V^Q, K_T = U_T W_K^K, V_T = U_T W_T^V$, where $W_m^Q, W_m^K, W_m^V \in R^{d \times d}, m \in \{V, T, A\}$ is the corresponding weight matrix, the cross-modal computation process is shown in Figure 6. Here, a bi-directional LSTM with a hidden size of 100 is used and the forward and backward passes are summed.

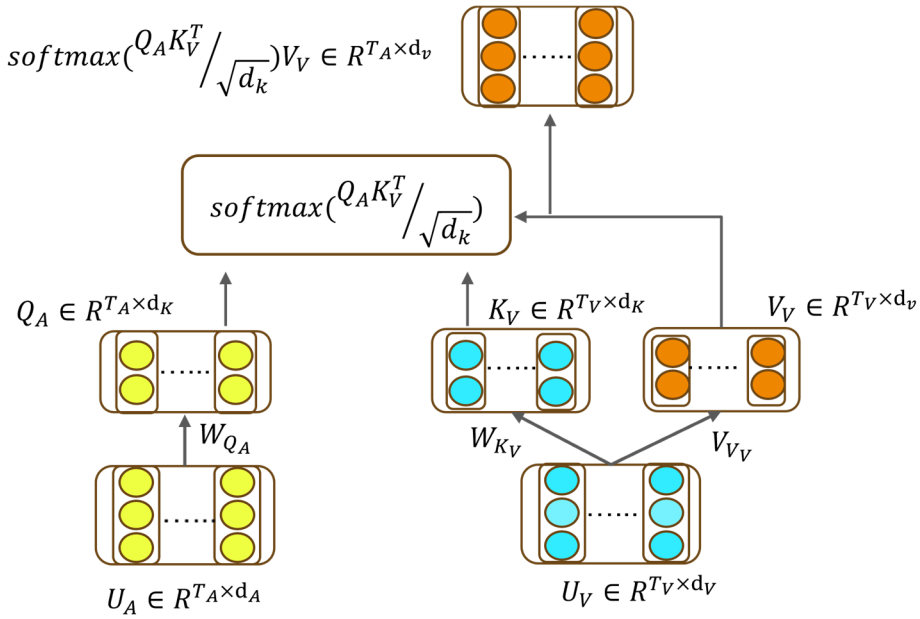


Figure 6. Diagram of the cross-modal calculation process.

The input text and video features are passed through N cascaded encoder structures to obtain a video feature representation that focuses on the text features, which is computed as follows:

$$U_{TA} = softmax\left(\frac{Q_A K_T}{\sqrt{d}}\right) V_T \tag{8}$$

$$U_{TA}^{[0]} = U_A^{[0]}, U_{T\to A} = LN(U_A + U_{T\to A}) \tag{9}$$

$$\widehat{U_{TA}^{[i]}} = LN\left\{CM_{TA}^{[i],mul}\left(U_{TA}^{[i-1]}, U_T^{[0]}\right)\right\} + LN\left(U_{TA}^{[i-1]}\right) \tag{10}$$

$$U_{TA}^{[N]} = BiLSTM\left\{LN\left\{Conv1d\left(\widehat{U_{TA}^{[i]}}\right)\right\} + LN\left(\widehat{U_{TA}^{[i]}}\right)\right\} \tag{11}$$

where LN denotes layer normalization, and $CM_{TA}^{[i],mul}$ denotes the i-th layer's multiple attention, and $\widehat{U_{TA}^{[i]}}$ denotes the output of the i-th layer, $i = \{1, 2, \ldots N\}$.

The aforementioned equation is computed using the attention mechanism to determine the interaction information between the two modalities. The SoftMax function is utilized to calculate the attention weights, enabling the extraction of propagation information from the computed value terms of each feature sequence. Each of the multiple attention modules in this module is followed by a conv1d layer that focuses on short-term contexts. Models that combine Transformer and BiLSTM (as depicted in Figure 7) can more effectively learn the relationship between affective temporal dependencies, and their combination further enhances performance while tightly integrating the two modalities at the model level for fusion. Ultimately, the module facilitates the updating of information from one modality to the features of the other.

According to the above algorithm, we finally get $U_{TV}, U_{VT}, U_{TA}, U_{AT}, U_{AV}, U_{VA}$ six two-by-two

for information complementation. After that, $(U_{TV}, U_{AV}), (U_{TA}, U_{VA}), (U_{VT}, U_{AT})$ are connected in pairs after dimensionality reduction in the linear layer to get the updated feature vectors U_A, U_T, U_V respectively. The updated feature vector is transferred to the next layer.

$$U_A = Concat\big((U_{TA}W_{TA} + b_{TA}), (U_{VA}W_{VA} + b_{VA})\big) \tag{12}$$

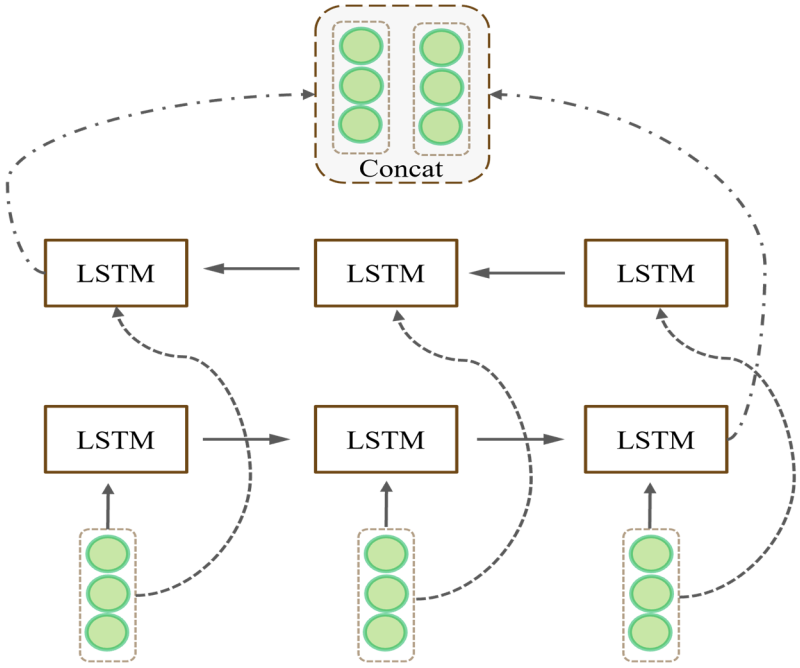


Figure 7. Diagram of BiLSTM structural model.

3.5. Multimodal mask gating fusion and classification module

In the multimodal mask gating module (depicted in Figure 8), the gating network integrates the multimodal masking mechanism with the GRU gating structure to fully leverage multimodal information, learn contextual relationships, and extract high-level features for subsequent emotion recognition classification. Leveraging the relationships between modal feature representations at each time step allows for enhanced focus on significant features and the capture of richer contextual information, which is crucial for emotion classification as emotions are often context-dependent. Through the design of the masking mechanism and gating network combination module, the GRU gating structure can learn abstract high-level feature representations. Additionally, leveraging the gating mechanism of the GRU (as shown in Figure 9) to model contextual information of sequential data aids in identifying features with stronger discriminative power for sentiment classification. Introducing the masked multimodal mechanism into the GRU gating structure enables adaptive focus on the importance of different modal features in sentiment classification, thereby better capturing the associations across modalities and further enhancing sentiment categorization performance. The masking multimodal mechanisms offer the capability to interpret model predictions. By analyzing the weights and selectively masking or ignoring certain inputs, it becomes possible to understand the extent to which different modalities contribute to the results in sentiment classification.

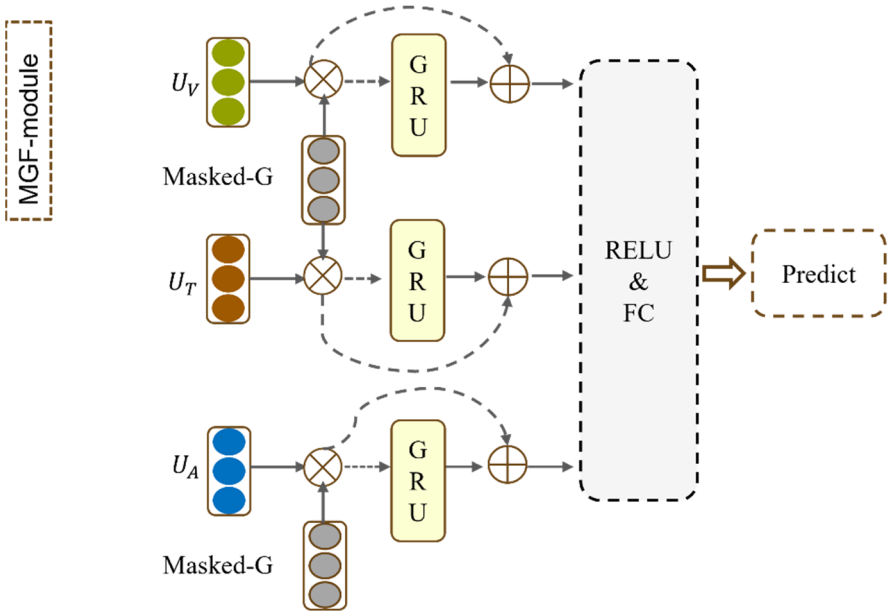


Figure 8. Model diagram of the mask gating network proposed in this paper.

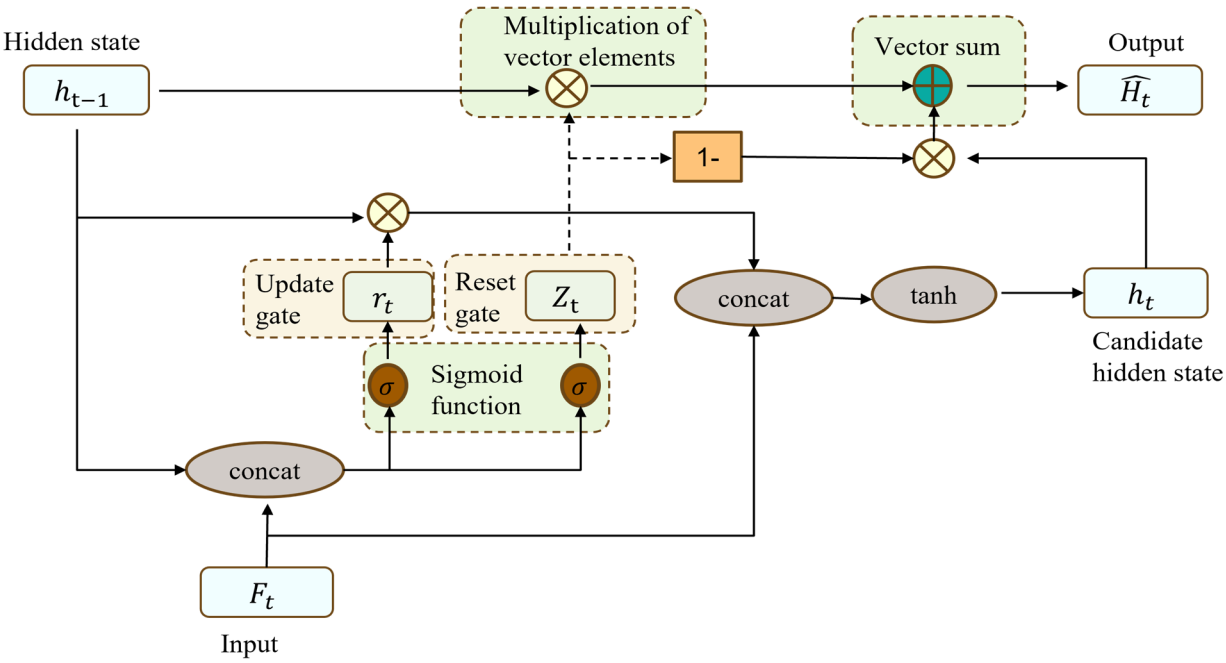


Figure 9. Network model structural diagram of GRU.

Step 1: Obtain the corresponding position encoding vectors for each modality from the previous module U_A, U_T, U_V, generate binary mask vectors using One-HOT encoding in the text position encoding feature vector, and generate binary mask vectors for the video and audio position encoding feature vectors using adaptive thresholding method to indicate whether each frame has a significant feature or not. Thus, the mask vectors of corresponding dimensions are represented M_A, M_T, M_V, The

elements in the mask vector are either 0 or 1, which are used to indicate the parts that need to be masked or ignored. Afterwards, the output modal feature representation is multiplied with the mask gate vector G at the element level to obtain the mask-gated adjusted input features.

$$G = Sigmoid\left(W_g \cdot U_i + b_g\right) \tag{13}$$

$$F_i = U_i \cdot G \tag{14}$$

where W_g, b_g is the weight parameter to be obtained by learning, $i \in \{A, T, V\}$, F_i is the mask feature vector.

Step 2: After that, the introduction of gated network is considered to dynamically adjust the behavior and output structure of the network according to the inputs, and the fusion feature vector is obtained by summing the mask feature vector with the GRU output.

$$r_t = sigmoid(W_r * [F_i, h_{t-1} + b_r]) \tag{15}$$

$$z_t = sigmoid(W_z * [F_i, h_{t-1} + b_z]) \tag{16}$$

$$h_t = tanh(W_h * [F_i, r_i * h_{t-1} + b_h]) \tag{17}$$

$$\widehat{H_t} = (1 - z_i) * (h_{t-1}) + z_i * h_i \tag{18}$$

$$H_i = \widehat{H_t} + F_i \tag{19}$$

One of them is r_t, z_t, h_t are the updated, reset and updated hidden states, respectively, and w_r, h_{t-1}, b_r are the weight matrix of the updaters, the hidden state of the previous time step, and the bias vector. w_z, b_z, W_h, b_h are the weight matrices of the reset door, the candidate hidden states and the bias vector respectively. All the above parameters are trainable parameters for learning.

Step 3: Finally, the obtained features are sent to the classification output layer for sentiment classification. In this paper, the nonlinear layer ReLU is used to fuse them to capture their interactions, the linear layer is used to predict the final emotion classification labels, the SoftMax function is used for multicategory classification and the cross-entropy function is used to optimize the model in this paper. The formulas are as follows:

$$F = Relu(W_F([H_A, H_T, H_V]) + b_F) \tag{20}$$

$$y = W_y \cdot F + b_y \tag{21}$$

$$\begin{cases} \hat{y} = softmax(y) \\ softmax(x_i) = \frac{e^{x_i}}{\sum_{i=1}^{n} e^{x_i}} \\ \mathcal{L} = -\sum_i y_i \, log \, \hat{y}_i \end{cases} \tag{22}$$

where $W_F \in R^{d_F \times (d_H^A + d_H^T + d_H^V)}$, $b_F \in R^{d_F}$, $W_y \in R^{1 \times d_F}$, $b_y \in R^1$ are the trainable parameters of the network. e^{x_i} represents the output value of the i-th node in the classification module, and n represents the value that is n categorized. $\hat{y}$ represents the predicted category probability value of the model. $\hat{y}_i$ denotes the prediction result of the model.

4. Results

4.1. Benchmarks and assessment indicators

In this experiment, the study performed experimental evaluations on the widely-used multimodal emotion recognition datasets CMU-MOSI and CMU-MOSEI. The datasets were preprocessed to align the word-level multimodal signals for each sample. This ensured that the input data was properly synchronized across modalities, allowing for accurate analysis and classification of emotions.

CMU-MOSI. The CMU-MOSI dataset is a multimodal dataset utilized for sentiment analysis and opinion research. It encompasses multimodal data derived from 228 videos encompassing diverse domains such as movie reviews, speeches, interviews, and more. The CMU-MOSI dataset offers multimodal features including video, speech, and text, along with comprehensive labeling information that includes sentiment category labels, sentiment intensity labels, and opinion polarity labels. Comprising 2119 video clips of brief monologues, the dataset follows a standard partitioning approach, utilizing 1798 samples for training and the remaining 399 samples for testing.

CMU-MOSEI. The CMU-MOSEI dataset contains a series of multimodal clips from movie review videos covering a wide range of information forms including audio, video and text. The main feature of this dataset is that each clip is annotated with the corresponding sentiment label and sentiment intensity. It contains 23,453 video clips extracted from 239 movies. Emotions consisted of happiness, sadness, anger, fear, disgust, and surprise, and the annotation scores ranged from (0, 3), where 0 indicates no emotion and 3 indicates strong emotion. The dataset is divided into training and test sets corresponding to 16,216 and 6525 discourse counts, respectively, and 4659 test sets are selected for the experiment.

Assessment metrics. For the MOSI and MOSEI datasets, this paper uses mean absolute error (MAE), F1 score, correlation coefficient (Corr) between the model and human prediction, binary classification accuracy (ACC-2) and seven-category classification accuracy (ACC-7) as assessment metrics.

$$MAE(i) = \left(\frac{1}{M}\right) * \sum\left|y(i) - \widehat{y(i)}\right|, MAE = \left(\frac{1}{N}\right) * \sum MAE(i) \tag{23}$$

$$Corr = \frac{\left(\Sigma\left((y_\alpha - \overline{y_\beta}) * (y_\beta - \overline{y_\beta})\right)\right)}{\sqrt{\Sigma(y_\alpha - \overline{y_\beta})^2} * \sqrt{\Sigma(y_\beta - \overline{y_\beta})^2}} \tag{24}$$

where the actual sentiment value of the i-th sample $y(i)$, the predicted sentiment value $\widehat{y(i)}$, the overall mean absolute error is obtained by averaging so the sample mean absolute error, $\alpha, \beta \in \{A, T, V\}$.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{25}$$

$$F1 - score = \frac{2*TP}{2*TP+FN+FP} \tag{26}$$

where TP is the number of correct samples where the model predicted correctly, TN is the number of incorrect samples where the model predicted correctly, FP is the number of correct samples where

the model predicted incorrectly, and FN is the number of incorrect samples where the model predicted incorrectly.

4.2. Experimental setup

This experiment is built based on PyTorch toolkit. The experiment was implemented on Windows 10 operating system with a GPU version of RTX3080Ti, the development platform Pytorch, and the development tool Pycharm.

In the unimodal fusion of feature vectors of different dimensions and levels extracted by different models, in order to prevent information redundancy due to too many attention layers and the number of multi-head attention, and too few to cause insufficient information about the long-distance interactions between the extracted modalities, which makes the information omitted, the number of video and audio Transformer encoder layers is set to 5, the number of multi-head attention heads is 8, and the attention head To prevent overfitting, the dropout rate of the model is set to 0.3. The batch size of the model is set to 32, and the model is trained 100 times. The initialized learning rate is 0.00001, and the learning rate automatically decreases when the training effect of the model no longer rises. The Adam algorithm optimizer was used to optimize the parameters in the model. Visual and audio features are sampled at 15 HZ and 12.5 HZ respectively and the text serial length is unified. The detailed distribution of the CMU-MOSI data in the dataset is shown in Table 1 below.

Table 1. CMU-MOSI dataset information.

Dataset		Positive	Negative	Neutral	Total
CMU-MOSI	Train	833	866	81	1780
	Test	190	194	15	399

4.3. Experimental analysis

4.3.1. Experimental analysis I

The following state-of-the-art methods were chosen as baselines for comparison with the models in this paper.

The CM-MSF model is compared to all baselines as presented in Table 2. In the CMU-MOSI dataset, CM-MSF demonstrates superior performance and surpasses the baseline on numerous metrics, encompassing both regression and classification tasks. Moreover, the model in this study outperforms complex fusion mechanisms like TFN [7] and LMF [8], underscoring the significance of effective and sufficient semantic complementation prior to fusion. Among the baseline models, MULT, MFN, and MISA employ a single deep learning model for feature extraction in unimodal mode. However, this approach may lead to incomplete capture of feature information and fail to consider the extraction of both local and global information. On the other hand, the CM-BERT baseline model utilizes text and audio bimodalities, overlooking the mutual information present in the text and audio modalities. While SSE-FT and CMC-HF models incorporate layered fusion and attention mechanism-based fusion, the performance of the layered fusion concept stands out prominently among the results of other baseline systems.

Table 2. Analysis of experimental comparison models.

Model	Model analysis
HFusion [33]	A novel fusion strategy for feature fusion in a hierarchical manner, where two modalities are first fused and only then all three modalities are fused for subsequent analysis.
LMF [8]	An efficient method for multimodal fusion using low-rank weight tensor modeling is an improvement of TFN tensor fusion in. Not only does it greatly reduce the computational complexity, but also significantly improves the performance.
MULT [34]	Multimodal transformer models use directed pairwise cross-modal attention to focus on interactions between multimodal serials in different time steps and potentially adapt the flow from one modality to another.
MISA [35]	Learning effective multimodal representations to aid the fusion process by means of feature de-entanglement, the model reduces the impact of modal gaps due to inter-modal heterogeneity and avoids complex fusion techniques.
MFN [36]	The Memory Fusion Network is a multi-view gated memory network that enables internal cross-view interactions. It explicitly considers interactions in neural architectures and continuously models them over time.
SSE-FT [37]	The model is studied using the method of pre-trained networks and the mechanism of hierarchical fusion in the feature extraction phase, which effectively fuses multiple modalities.
CMC-HF [13]	A novel multimodal emotion recognition framework is proposed to realize cascaded multichannel and hierarchical fusion, and in this paper, an improved hierarchical fusion module is introduced to facilitate cross-modal interaction among three modalities.
CM-BERT [22]	Designing Cross-Modal BERT using textual and audio modalities to interact to fine-tune pre-trained BERT models, introducing masked multimodal attention combining textual and audio modal information to adjust word weights.
AOBERT [38]	All-modal unity BERT can reduce intra- and inter-modal losses caused by other fusion methods using one single-flow transformer for pre-training.

As indicated in Table 3, the CM-MSF model attains an ACC2-h of 89.1% in binary classification, surpassing the CMC-HF model, which previously exhibited the best performance among the baselines, by 0.9%. The CMC-HF method employs a cascade model to extract local and long-distance information and conducts hierarchical fusion of features based on modal interactions, outperforming other models among the aforementioned baselines but falling slightly short compared to the model proposed in this paper. The CM-MSF model achieves an effectiveness score of 59.8% on ACC7-h, which is an improvement of 4.1% compared to the SSE-FT model and 1.6% compared to the CMC-HF model. Notably, the improvement of the CM-MSF model is particularly evident in the emotion classification task, in which the MAE and F1 scores outperform the other compared models. Additionally, Figure 10 illustrates the confusion matrix of this model on the dataset, providing further insight into its performance. Table 4 shows the experimental results of CM-MSF on the CMU-MOSEI dataset, where Acc (7/2class) and F1 scores outperform the other comparison models.

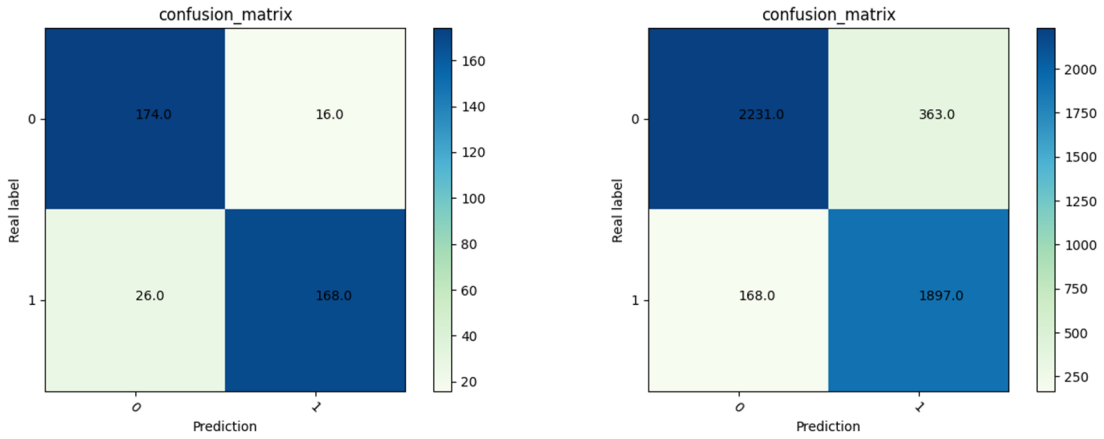
Table 3. Comparison experiments between CM-MSF and other baseline models based on the CMU-MOSI dataset.

Model	MAE ↓	F1 Score ↑	Corr ↑	ACC (2 class -h) ↑	ACC (7 class -h) ↑
HFusion [33]	0.890	79.8	0.703	77.9	35.3
LMF [8]	0.912	75.7	0.668	76.4	32.8
MULT [34]	0.889	81.0	0.686	81.1	39.1
MISA [35]	0.783	81.7/83.6	0.761	81.8/83.4	42.3
MFN [36]	0.965	77.3	0.632	77.4	34.1
SSE-FT [37]	0.592	87.0	0.792	87.3	55.7
CMC-HF [13]	0.581	87.3	**0.798**	88.2	58.2
CM-BERT [22]	0.729	84.5	0.791	84.5	44.9
AOBERT [38]	0.856	85.4/86.4	0.700	85.2/85.6	40.2
CM-MSF	**0.576**	**87.9**	0.795	**89.1**	**59.8**

Notes: ↑ indicates that larger values of the following are better, and ↓ indicates that smaller values are better.

Table 4. Comparison experiments between CM-MSF and other baseline models based on the CMU-MOSEI dataset.

Model	MAE ↓	F1 Score ↑	Corr ↑	ACC (2 class -h) ↑	ACC (7 class -h) ↑
MISA [35]	0.555	83.8/85.3	0.756	83.6/85.5	52.2
MMIM [39]	0.526	82.7/85.9	0.772	82.2/86.0	54.2
MMLATCH [24]	0.582	82.9	0.704	82.8	52.1
Hycon [40]	0.601	--/85.6	**0.778**	--/85.4	52.8
AOBERT [38]	**0.515**	85.0/85.9	0.763	84.9/86.2	54.5
CM-MSF	0.545	**88.1**	0.769	**88.6**	**56.3**



(a) (b)

Figure 10. Confusion matrix of CM-MSF model on MOSI (a) and MOSEI (b) datasets.

4.3.2. Experimental analysis II

To assess the advantages of multimodality over unimodality and understand the impact of different modalities in the task, several experiments were conducted using single modal inputs of text,

video, and audio, as well as bimodal fusion of various combinations. The results of these experiments are presented in Table 5. The table clearly demonstrates that the performance suffers significantly when using a single modality. Specifically, the performance is worse when using only the video modality or only the audio modality compared to when using only the text modality. This observation suggests that the text modality may play a dominant role in overall emotion recognition effectiveness. On the other hand, combining two modalities for sentiment analysis leads to improved performance compared to using independent modalities. This finding indicates that modalities can learn from each other, leveraging information from one modality to enhance features and thereby improve overall performance.

Table 5. Comparison results of unimodal, bimodal, and trimodal emotion recognition experiments based on this model on CMU-MOSI dataset.

	Modal	$\boldsymbol{MAE}\downarrow$	$\boldsymbol{Corr}\uparrow$
unimodal	V	0.910	0.531
	A	0.989	0.314
	T	0.796	0.726
bimodal	T+A	0.878	0.772
	T+V	0.710	0.805
	A+V	1.03	0.328
Tri-modal	V+A+T	0.576	0.795

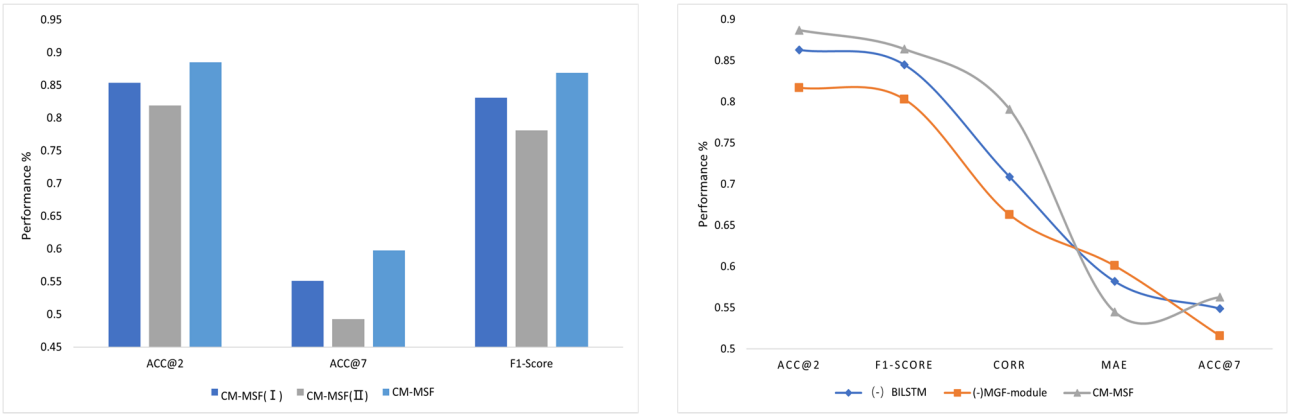
The model presented in this paper incorporates a mask gating module, which applies gating mask restrictions subsequent to the cascade attention mechanism for extracting mutual information. This enables the selection of discriminative emotion features that facilitate accurate categorization.

Furthermore, with reference to Table 6, where (-) denotes that the module is excluded, the cascade model in the Modal Interaction Module and the Mask Gating Module were subjected to ablation study. When only the Transformer is utilized as the fusion encoder for the Modal Interaction Module, the MAE performance experiences a decrease of approximately 0.069%, and the Corr performance decreases by about 0.098%. The Transformer model can generate highly expressive high-level emotion representations, while the BILSTM model excels at inferring these representations. Models that combine the Transformer and BILSTM models exhibit enhanced capabilities in learning temporal dependencies of emotions and exploring advanced representations, leading to promising performance improvements. Upon removing the MGF-module, the MAE performance decreases by about 0.144% and the Corr performance decreases by about 0.093%. This experiment aims to validate the efficacy of the proposed mask gating network in enhancing classification accuracy, and the analysis underscores the effectiveness of the two-part module enhancement.

Table 6. Ablation experiment.

	$MAE\downarrow$	$Corr\uparrow$
(-) BILSTM	0.648	0.697
Transformer+BiLSTM	0.576	0.795
(-) MGF-module	0.720	0.649

To assess the effectiveness of the parallel model, this study conducted comparative experiments on single modality using both a single encoder and a parallel encoder. The CM-MSF (I) model consists of video modal Transformer, text modal BERT, and audio modal Transformer, while the CM-MSF (II) model consists of video modal ResNet, text modal LSTM, and audio modal BiGRU. The experimental results are shown in Figure 11(a). From the data shown in the figure, it can be seen that the unimodal parallel bimodal model shows excellent performance in sentiment classification. Figure 11(b) shows the analysis of the ablation experiments done by the model of this study under the CMU-MOSEI dataset, from which it can be seen the importance of each module of the model in improving the recognition accuracy. Moreover, this paper also separately analyzed the text modality to verify the effectiveness of the strategy, as demonstrated in Table 7.



(a) (b)

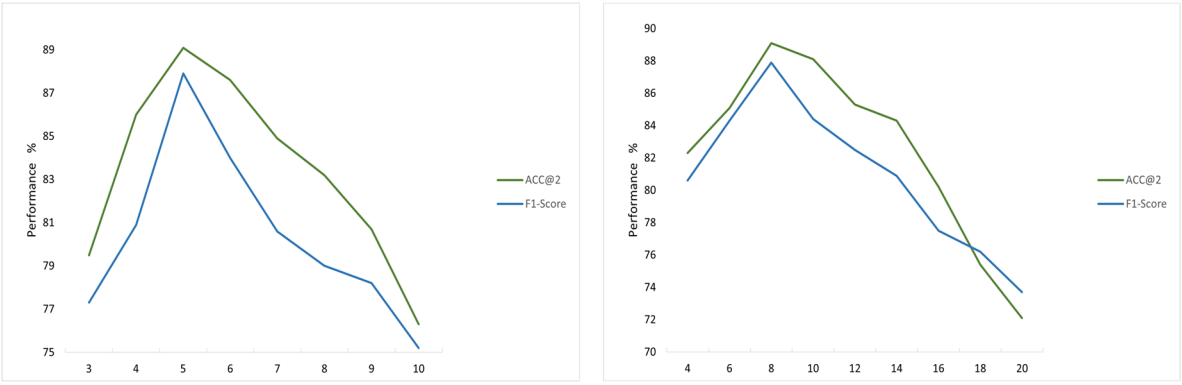
Figure 11. Performance analysis of single-modal parallel dual model.

Table 7. Text modal through single and dual model performance analysis.

	ACC@2	ACC@7	F1-Score
BERT	87.6	56.7	86.0
LSTM	84.9	51.3	83.6
BERT+LSTM	89.1	59.8	87.9

4.3.3. Experimental analysis III

In the bimodal modeling approach described in this paper, the feature vectors obtained from each unimodal source are fused and aligned using a multi-head attention network. In the cascade cross-modal information interaction module, stacked N cascade encoders are utilized to enhance the cross-modal information integration. Since the number of multiple attention heads and the cascade model have a significant impact on the performance of the model, we conducted separate experiments on the CMU-MOSI dataset to determine the optimal parameters for these two factors. The performance metrics, including dichotomous, tetrachotomies and F1 values, at different numbers of attention heads and cascade encoders are illustrated in Figure 12 (a),(b) respectively. These figures provide insights into the impact of varying these parameters on the model's performance.



(a) (b)

Figure 12. Performance analysis of different number of attention heads with cascade encoder.

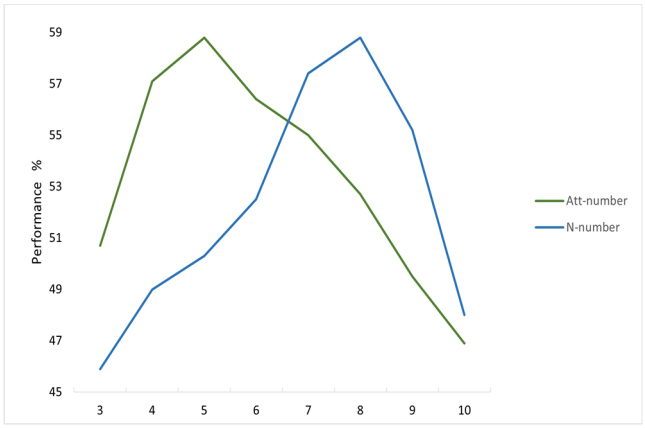


Figure 13. Performance analysis of seven classifications under different number of attention heads with cascade encoder.

In Figure 13, "Att-number" represents the number of attention heads, and "N-number" represents the number of cascade encoders. Based on the experimental results, it can be inferred that CM-MSF achieves optimal recognition performance when the number of attention heads is set to 5 and the number of cascade encoders is set to 8. Deviating from these optimal values, either by selecting smaller or larger values, may result in unsatisfactory recognition performance due to the omission of valid information or information redundancy.

5. Conclusions

In this paper, we introduce CM-MSF, a novel multimodal sentiment analysis framework that leverages a single modality and a dual encoder to extract features. These features are then fed into a modal interaction module and a mask gating network for adaptive selection, which ultimately predicts the sentiment state. The proposed approach utilizes a multi-head attention mechanism for model fusion and alignment in the fusion module, while feature enhancement is achieved through cascaded BILSTM and Conv1d layers in the modal interaction phase. To evaluate the performance of our model, we conducted experiments on the CMU-MOSI and MOSEI datasets. Our results demonstrate that CM-

MSF outperforms previously proposed models, highlighting the effectiveness of our fusion techniques and the importance of utilizing multimodal data. We also performed ablation experiments to demonstrate the necessity of each component in our proposed framework. Overall, our study underscores the significance of multimodal sentiment analysis and demonstrates the potential of our proposed CM-MSF framework.

Although the method has made promising progress in terms of performance improvement, there are some shortcomings that need to be addressed in the course of future research. On the one hand regarding the dataset, most of the video materials in the currently available datasets are from the Internet, and the emotions inside these videos are not spontaneous emotions that flow out of the human door, so they lack authenticity to a certain extent. In addition, the type of modality and the amount of data in the dataset is also an important issue that restricts the research of multimodal emotion recognition. On the other hand, further research is needed in multimodal fusion, and how to efficiently fuse and interact between different modalities and improve the recognition rate of the seven emotions is also a difficult issue that needs to be continued in the future.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant No. 62271303) and Shanghai Pujiang Talents Program (Grant No. 22PJD029).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. R. K. Patra, B. Patil, T. S. Kumar, G. Shivakanth, B. M. Manjula, Machine learning based sentiment analysis and swarm intelligence, in *2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)*, IEEE, (2023), 1–8. https://doi.org/10.1109/ICICACS57338.2023.10100262
2. R. Das, T. D. Singh, Multimodal sentiment analysis: A survey of methods, trends, and challenges, *ACM Comput. Surv.*, **55** (2023), 1–38. https://doi.org/10.1145/3586075
3. S. Peng, K. Chen, T. Tian, J. Chen, An autoencoder-based feature level fusion for speech emotion recognition, *Digital Commun. Networks*, 2022. https://doi.org/10.1016/j.dcan.2022.10.018
4. S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, (2018), 112–118. https://doi.org/10.1109/SLT.2018.8639583
5. E. Jeong, G. Kim, S. Kang, Multimodal prompt learning in emotion recognition using context and audio information, *Mathematics*, **11** (2023), 2908. https://doi.org/10.3390/math11132908

6. E. Batbaatar, M. Li, K. H. Ryu, Semantic-emotion neural network for emotion recognition from text, *IEEE Access*, **7** (2019), 111866–111878. https://doi.org/10.1109/ACCESS.2019.2934529
7. A. Zadeh, M. Chen, S. Poria, E. Cambria, L. P. Morency, Tensor fusion network for multimodal sentiment analysis, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, (2017), 1103–1114. https://doi.org/10.18653/v1/D17-1115
8. Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, L. P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, (2018), 2247–2256. https://doi.org/10.18653/v1/P18-1209
9. S. Mai, H. Hu, S. Xing, Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, in *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press, (2020), 164–172. https://doi.org/10.1609/aaai.v34i01.5347
10. B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, H. Prendinger, Deep learning for affective computing: Text-based emotion recognition in decision support, *Decis. Support Syst.*, **115** (2018), 24–35. https://doi.org/10.1016/j.dss.2018.09.002
11. L. Zheng, L. Sun, M. Xu, H. Sun, K. Xu, Z. Wen, et al., Explainable multimodal emotion reasoning, preprint, ArXiv:2306.15401.
12. L. Sun, B. Liu, J. Tao, Z. Lian, Multimodal cross- and self-attention network for speech emotion recognition, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, (2021), 4275–4279. https://doi.org/10.1109/ICASSP39728.2021.9414654
13. X. Liu, Z. Xu, K. Huang, Multimodal emotion recognition based on cascaded multichannel and hierarchical fusion, *Comput. Intell. Neurosci.*, **5** (2023), 9645611. https://doi.org/10.1155/2023/9645611
14. S. Lee, D. K. Han, H. Ko, Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification, *IEEE Access*, **9** (2021), 94557–94572. https://doi.org/10.1109/ACCESS.2021.3092735
15. P. Kumar, X. Li, Interpretable multimodal emotion recognition using facial features and physiological signals, preprint, arXiv:2306.02845.
16. F. Lv, X. Chen, Y. Huang, L. Duan, G. Lin, Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2021), 2554–2562. https://doi.org/10.1109/CVPR46437.2021.00258
17. D. Hazarika, R. Zimmermann, S. Poria, MISA: Modality-invariant and -specific representations for multimodal sentiment analysis, in *Proceedings of the 28th ACM International Conference on Multimedia*, ACM, (2020), 1122–1131. https://doi.org/10.1145/3394171.3413678
18. D. Yang, S. Huang, H. Kuang, Y. Du, L. Zhang, Disentangled representation learning for multimodal emotion recognition, in *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*, ACM, (2022), 1642–1651. https://doi.org/10.1145/3503161.3547754
19. H. Han, J. Yang, W. Slamu, Cascading modular multimodal cross-attention network for rumor detection, in *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, IEEE, (2023), 974–980. https://doi.org/10.1109/ICCECT57938.2023.10140211

20. S. A. M. Zaidi, S. Latif, J. Qadir, Cross-language speech emotion recognition using multimodal dual attention transformers, preprint, arXiv:2306.13804.
21. Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press, (2020), 8992–8999. https://doi.org/10.1609/aaai.v34i05.6431
22. K. Yang, H. Xu, K. Gao, CM-BERT: Cross-Modal BERT for text-audio sentiment analysis, in *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*, ACM, (2020), 521–528. https://doi.org/10.1145/3394171.3413690
23. H. Yang, X. Gao, J. Wu, T. Gan, N. Ding, F. Jiang, et al., Self-adaptive context and modal-interaction modeling for multimodal emotion recognition, in *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, (2023), 6267–6281. https://doi.org/10.18653/v1/2023.findings-acl.390
24. G. Paraskevopoulos, E. Georgiou, A. Potamianos, Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, (2022), 4573–4577. https://doi.org/10.1109/ICASSP43922.2022.9746418
25. L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, *Inf. Fusion*, **95** (2023), 306–325. https://doi.org/10.1016/j.inffus.2023.02.028
26. S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio–visual emotion recognition, *IEEE Trans. Circuits Syst. Video Technol.*, **28** (2018), 3030–3043. https://doi.org/10.1109/TCSVT.2017.2719043
27. D. Hazarika, S. Gorantla, S. Poria, R. Zimmermann, Self-Attentive feature-level fusion for multimodal emotion detection, in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, (2018), 196–201. https://doi.org/10.1109/MIPR.2018.00043
28. M. S. Hossain, G. Muhammad, Emotikon recognition using deep learning approach from audio–visual emotional big data, *Inf. Fusion*, **49** (2019), 69–78. https://doi.org/10.1016/j.inffus.2018.09.008
29. H. Cheng, Z. Yang, X. Zhang, Y. Yang, Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion, *IEEE Trans. Affective Comput.*, **14** (2023), 3149–3163. https://doi.org/10.1109/TAFFC.2023.3265653
30. S. Wang, J. Qu, Y. Zhang, Y. Zhang, Multimodal emotion recognition from EEG signals and facial expressions, *IEEE Access*, **11** (2023), 33061–33068. https://doi.org/10.1109/ACCESS.2023.3263670
31. C. Xu, K. Shen, H. Sun, Supplementary features of BiLSTM for enhanced sequence labeling, preprint, arXiv:2305.19928.
32. L. Zhu, M. Xu, Y. Bao, Y. Xu, X. Kong, Deep learning for aspect-based sentiment analysis: A review, *PeerJ Comput. Sci.*, **8** (2022), e1044. https://doi.org/10.7717/peerj-cs.1044
33. Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, S. Ruslan, Multimodal transformer for unaligned multimodal language sequences, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, NIH Public Access, (2019), 6558–6569. https://doi.org/10.18653/v1/p19-1656

34. Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, (2019), 6558–6569. https://doi.org/10.18653/v1/p19-1656
35. D. Hazarika, R. Zimmermann, S. Poria, MISA: Modality-invariant and -specific representations for multimodal sentiment analysis, in *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*, ACM, (2020), 1122–1131. https://doi.org/10.1145/3394171.3413678
36. A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, L. P. Morency, Memory fusion network for multi-view sequential learning, AAAI Press, (2018), 5634–5641. https://doi.org/10.1609/aaai.v32i1.12021
37. S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, S. Nanayakkara, Multimodal emotion recognition with transformer-based self supervised feature fusion, *IEEE Access*, **8** (2020), 176274–176285. https://doi.org/10.1109/ACCESS.2020.3026823
38. K. Kim, S. Park, AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis, *Inf. Fusion*, **92** (2023), 37–45. https://doi.org/10.1016/j.inffus.2022.11.022
39. W. Han, H. Chen, S Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, (2021), 9180–9192. https://doi.org/10.18653/v1/2021.emnlp-main.723
40. S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, *IEEE Trans. Affective Comput.*, **14** (2023), 2276–2289. https://doi.org/10.1109/TAFFC.2022.3172360

AIMS Press