electronics

MDPI

Article

A Multimodal Sentiment Analysis Approach Based on a Joint Chained Interactive Attention Mechanism

Keyuan Qiu [1], Yingjie Zhang [1], Jiaxu Zhao [1], Shun Zhang [2], Qian Wang [3] and Feng Chen [1,*]

[1] College of Information Science and Technology, Shihezi University, Shihezi 832003, China; nightwish2024@stu.shzu.edu.cn (K.Q.); zhangyingjie@xjshzu.com (Y.Z.)
[2] College of Sciences, Shihezi University, Shihezi 832003, China
[3] Department of Computer Science, Durham University, Durham DH1 3LE, UK; qian.wang173@hotmail.com
* Correspondence: cf_inf@shzu.edu.cn

Abstract: The objective of multimodal sentiment analysis is to extract and integrate feature information from text, image, and audio data accurately, in order to identify the emotional state of the speaker. While multimodal fusion schemes have made some progress in this research field, previous studies still lack adequate approaches for handling inter-modal information consistency and the fusion of different categorical features within a single modality. This study aims to effectively extract sentiment coherence information among video, audio, and text and consequently proposes a multimodal sentiment analysis method named joint chain interactive attention (VAE-JCIA, Video Audio Essay–Joint Chain Interactive Attention). In this approach, a 3D CNN is employed for extracting facial features from video, a Conformer is employed for extracting audio features, and a Funnel-Transformer is employed for extracting text features. Furthermore, the joint attention mechanism is utilized to identify key regions where sentiment information remains consistent across video, audio, and text. This process acquires reinforcing features that encapsulate information regarding consistency among the other two modalities. Inter-modal feature interactions are addressed through chained interactive attention, and multimodal feature fusion is employed to efficiently perform emotion classification. The method is experimentally validated on the CMU-MOSEI dataset and the IEMOCAP dataset. The experimental results demonstrate that the proposed method significantly enhances the performance of the multimodal sentiment analysis model.

Keywords: sentiment analysis; attention mechanisms; decision fusion; deep learning; model optimization

Citation: Qiu, K.; Zhang, Y.; Zhao, J.; Zhang, S.; Wang, Q.; Chen, F. A Multimodal Sentiment Analysis Approach Based on a Joint Chained Interactive Attention Mechanism. *Electronics* **2024**, *13*, 1922. https://doi.org/10.3390/electronics13101922

Academic Editor: Ioannis Hatzilygeroudis

Received: 5 April 2024
Revised: 5 May 2024
Accepted: 10 May 2024
Published: 14 May 2024

1. Introduction

As a crucial task in the field of Natural Language Processing (NLP), sentiment analysis aims to identify the affective tendencies of each aspect entity in textual utterances [1] or other modal data [2]. Multimodal sentiment analysis, on the other hand, is a processing task that integrates multiple modalities to analyze and make decisions about human emotions [3]. In recent years, with the rapid development of social media, the volume of texts, pictures, and videos shared on online platforms has been increasing steadily. Visual and speech information can compensate for the expressive limitations of textual information. Exploring the potential correlations among text, visual, and speech modalities can enhance sentiment analysis solutions. User-generated content is increasingly adopting a multimodal approach. Besides comments or tweets in pure text form, users are also publishing data containing additional modalities such as images and videos to enrich their viewpoints [4]. This trend has resulted in a significant surge in video content on the Web, thereby making the effective classification and management of these vast video datasets a pressing research concern. Consequently, the academic community has increasingly focused on this issue, leading to numerous research endeavors on the multimodal sentiment analysis of video content [5,6].

While current research predominantly focuses on training deep learning models for extracting multimodal features from data and has yielded satisfactory results, two main challenges persist. First, many existing models utilize a global self-attention mechanism. However, the temporal sequences in speech and video data may not be perfectly aligned. Even if alignment is achieved during data preprocessing, the models may struggle to direct attention to emotionally relevant regions due to fixed time scales and receptive fields, thereby limiting their effectiveness [6]. On the other hand, the order of modal interactions between different modalities can also have an impact on the final judgment of the model, and the sequential fusion order of modalities may also cause information loss.

This paper introduces a novel multimodal emotion recognition approach based on joint attention and chained interactive attention to tackle the aforementioned challenges. Joint attention facilitates modal feature alignment by identifying regions of emotional consistency across different modalities, addressing the issue of imperfect temporal alignment between audio and video data. Meanwhile, chained interactive attention aims to enhance the effective fusion of modalities and mitigate information loss during feature fusion. The proposed model encompasses video, audio, and text modalities, offering a more comprehensive approach to capturing emotional features compared to single-modal emotion recognition models. It enables the more accurate recognition of emotional information and demonstrates superior effectiveness compared to existing multimodal models.

In pursuit of this goal, this paper presents a multimodal sentiment analysis method grounded in cross-modal cross-attention networks. The objective is to enhance the reliability and robustness of intra-modal self-interaction and inter-modal cross-interaction feature representation.

2. Related Works

The differing sampling methods of text, visual, and speech modal features result in significant disparities in time series and semantic information, leading to unaligned unimodal features. Moreover, visual and speech feature information comprises numerous low-order signals and redundant data, hindering effective inter-modal feature fusion [7]. Therefore, multimodal sentiment analysis poses greater challenges compared to unimodal sentiment analysis. Earlier sentiment analysis generally used traditional machine learning methods, such as Cao et al. [8,9]. Jointly with text captions, they proposed a concept called the Visual Sentiment Theme Model in order to make fuller use of text features, text, and image features, which are used for SVM training. After obtaining local classification results, decision-making interactions between modalities are accomplished by fusing the results at the decision level. In recent years, Zhang et al. [10] proposed a quantum-like multimodal network framework for multimodal emotion recognition based on interaction dynamics modeling. Zadeh et al. [11] proposed a new model TFN (tensor fusion network) which solves the problem of intra-channel and inter-channel dynamics modelling. Liu et al. [12] proposed a low-rank tensor fusion method which solves the problem of the complexity of multimodal modal tensor fusion computational complexity, but these methods are unable to selectively filter the required contextual information by virtue of cognitive evidence. Bai et al. [13] introduced the Multimodal Transformer (MulT) to solve the data explicit alignment problem in an end-to-end manner. Hazarika et al. [14] proposed to project a single modality into modality-invariant and modality-specific subspaces using the multi-headed attention mechanism to fully learn modal context information. However, these methods simply input multimodal data into a designed network and do not effectively deal with the random noise of the raw data, making the data fusion less effective and making it difficult for the model to learn advanced features. In addition, when the distribution of unimodal representations is inconsistent with the rest of the representations, it is difficult to fuse the different modal features, which ultimately affects the emotion recognition results to a certain extent. Xu et al. [15] designed an interactive memory network with a multi-hop attention mechanism to achieve the interaction of textual and visual information and fuse them in the final splice. Devlin et al. [16] explored the relationship between the target

aspect words and the text, image, and data using the attention mechanism data. They used the attention mechanism to mine local features related to aspect words in text and visual representations and, finally, fused the features to improve the prediction. Bao et al. [17] used interactive attention to try to explore the correlations within and across modalities to obtain common features across modalities, but they only spliced the fused features without enhancing the modal feature correlations. Hu et al. [18] proposed a joint interaction attention based on graphic and textual sentiment analysis, which was applied to picture and text modalities but still had limitations, and the model lacked generalization ability.

Table 1 provides a summary of the relevant work mentioned above. While existing methods for multimodal sentiment analysis have shown progress, they still exhibit significant shortcomings. First, these methods often rely on a global self-attention mechanism to model long-term dependencies in a time series, making them vulnerable to noise and resulting in suboptimal data fusion. Second, difficulties arise in feature fusion from different modalities when the distributions of unimodal representations are inconsistent, negatively impacting sentiment recognition accuracy. Lastly, there remains considerable scope for enhancing the generalization capability of existing models, particularly in addressing complex and dynamic real-world application scenarios.

Table 1. Summary of related works in multimodal sentiment analysis.

Authors	Contribution	Methodology
Cao et al. [8,9]	Visual Sentiment Theme Model using SVMs.	Decision-level fusion of text and image features.
Zhang et al. [10]	Multimodal emotion recognition network.	Quantum-like network and interaction dynamics modeling.
Zadeh et al. [11]	Tensor Fusion Network for dynamics.	Handles intra and inter-channel dynamics.
Liu et al. [12]	Low rank tensor fusion method.	Reduces complexity in tensor fusion.
Bai et al. [13]	Multimodal Transformer for data alignment.	End-to-end explicit alignment approach.
Hazarika et al. [14]	Subspace projection for modality contexts.	Uses multi-headed attention for context learning.
Xu et al. [15]	Interactive memory network with attention.	Fuses textual and visual information interactively.
Devlin et al. [16]	Attention mechanism for aspectual words.	Local feature mining and fusion in modalities.
Bao et al. [17]	Interactive attention for modality correlations.	Fuses features without enhancing correlations.
Hu et al. [18]	Interaction attention for sentiment analysis.	Applied to graphic-textual modalities, with limits.

3. Model Establishment

The method proposed in this paper can be divided into a feature extraction module, joint attention module, and interactive attention fusion module, and the method framework is shown in Figure 1.

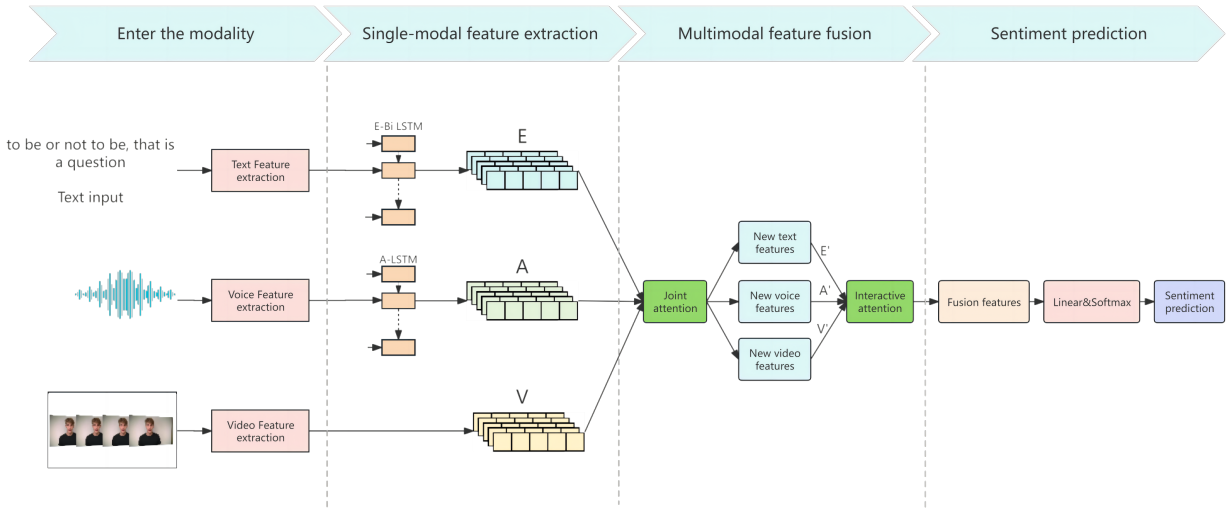

Figure 1. General model architecture.

The feature extraction module is designed based on 3D CNN, Conformer, and Funnel-Transformer modules to extract features expressing emotions in video, audio, and text. This paper is different from most of the models that directly use the Transformer model for fusion or directly use cross-attention to force the fusion. The Joint Attention module derives the similarity between different modal data by calculating the crossover matrix and mines and refines the region of emotional consistency between different modalities; the Interactive Attention is based on the Transformer to focus on the complementary information between different modalities and fuses the outputs, which are Linear, Linear, and Linear. Subsequently, the sentiment prediction is categorized using Linear and softmax.

For example, there was a social Tweet that contained the following three modal messages. Text: It's an amazing day! Video: showing people screaming on a roller coaster. Audio: recorded screams and background cacophony. When using the joint attention mechanism, the model considers these modalities simultaneously to synthesize the sentiment conveyed by the post, which is different from interactive attention, which considers more complementary information between different modal data. For example, post text may contain ambiguous expressions of emotion, such as Today was amazing! The emotional color of this statement depends on the context and may be positive or negative. The corresponding video may show a surprised face or an accident. By interacting with the attention mechanism, the model is able to recognize key emotional cues in the video, such as subtle changes in expression, which complement the emotional vocabulary in the text to more accurately define the overall emotional tendency.

3.1. Overview of the Methodology

Convolutional neural networks (CNNs) are commonly utilized to automatically extract key features from input data in multimodal sentiment analysis tasks. However, traditional 2D CNNs fail to consider the temporal correlation in video frame sequences [19], leading to decreased accuracy in video sentiment classification. Furthermore, as the network depth increases, traditional CNN-based methods are prone to getting trapped in local optimal solutions, resulting in decreased classification accuracy and significantly slower training speeds. To address these challenges, this study employs a 3D CNN that incorporates the time dimension for feature fusion. The specific structure of the network is illustrated in Figure 2 below.

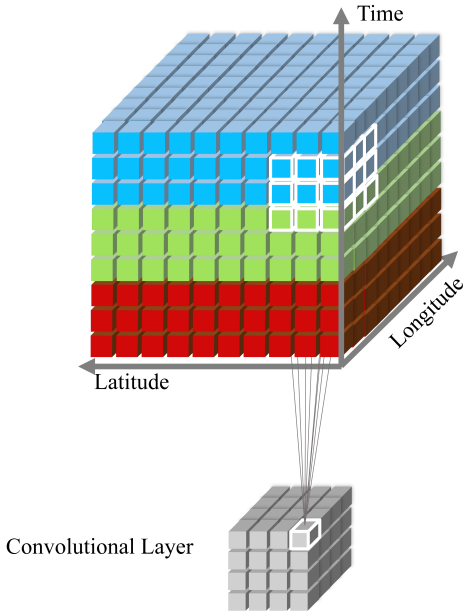

Figure 2. 3D CNN structure.

3D CNN is implemented by stacking multiple consecutive frames into a cube and applying a 3D convolutional kernel. In this architecture, the feature maps of each convolutional layer are connected to adjacent consecutive frames from the previous layer to capture temporal information.

Moreover, in sentiment analysis tasks, sentiment judgment depends not only on changes in expression but also on the entire sentence. The BERT model [16] adopts the encoder structure of the Transformer model and is trained with large-scale unlabeled data to acquire a comprehensive representation of semantic information. However, its pre-training primarily relies on unimodal data. While the model can be adaptably fine-tuned for multiple data types, its computational demands increase significantly when processing longer sequences, thereby impacting training and fine-tuning efficiency. To enhance multimodal data processing capability and improve the efficiency of processing long sequences, this study utilizes the Funnel-Transformer model [20] to handle text features. The model structure is depicted in Figure 3 below.

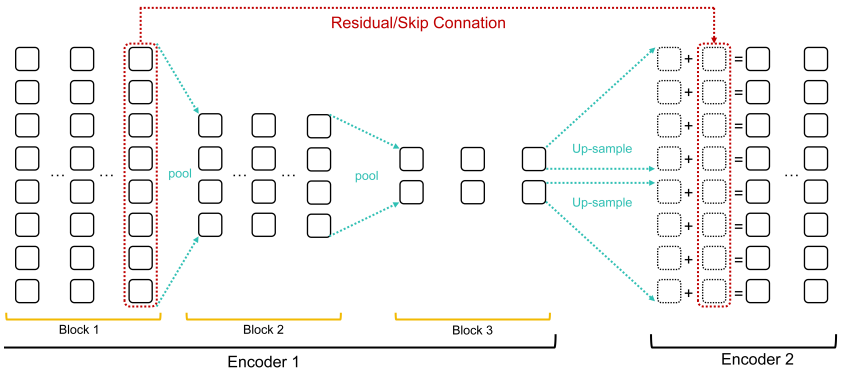

Figure 3. Funnel-Transformer structure.

As illustrated in Figure 3 above, as the model depth increases, the number of model parameters decreases by compressing the sequence length through pooling operations. In contrast to the BERT model, which solely employs the Transformer's decoder structure,

the Funnel-Transformer model preserves the decoder and restores the compressed vectors to the original sequence length through up-sampling. This approach maintains the traditional Transformer model's capability to handle multimodal data.

Regarding speech feature extraction, while the Transformer excels in managing global dependencies, it has limited capability to capture local features such as subtle changes in phonemes or syllables. Conversely, Conformer [21] integrates CNN technology, which is more effective at capturing local features such as subtle changes in phonemes. Given that speech signals contain rich local temporal information, the Conformer exhibits outstanding generalization capability in speech recognition tasks across various languages and accents.

3.2. Video Feature Extraction

The input video is initially segmented into a sequence of consecutive video frames, each representing a two-dimensional image. These frames collectively compose a three-dimensional video tensor. The input video tensor undergoes a 7×7 3D convolution operation. Each convolutional block consists of convolutional layers, residual connections, and layer normalization. This hierarchical structure facilitates the gradual extraction of higher-level features, transitioning from low-level spatial information to high-level abstract features. The outputs from each convolutional block are interconnected in a sequential manner through residual connections [22]. The specific process is as shown in Figure 4.

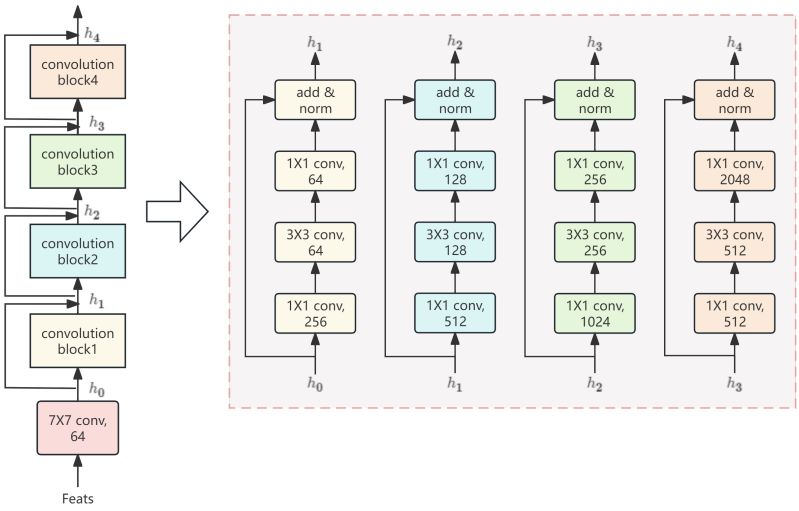

Figure 4. Video feature extraction steps.

3.3. Audio Feature Extraction

The audio feature extraction method uses MFCC features and rhythmic features as input. First, pre-emphasis is applied to the speech signal to enhance the high-frequency part and to smooth the signal. Subsequently, Discrete Fourier Transform (DFT) is applied to each frame of the speech signal to achieve the conversion from time domain to frequency domain. In order to simulate the hearing mechanism of the human ear, a Mel filter bank is used to divide the spectrum into a number of Mel frequency bands to obtain a frequency distribution that is closer to the perception of the human ear. Finally, a discrete cosine transform is performed on the Mel spectrum of each speech signal frame to obtain the final MFCC features. Meanwhile, the metrical features are extracted using the OpenSMILE toolkit, version 2.1.2. In the feature extraction process, the MFCC features (denoted as m) and the rhythmic features (denoted as p) are connected together to generate feature information with rich information $A \in \mathbb{R}^{T \times D}$. Here, T is the number of feature tokens and D is the feature dimension.

$$A = \text{concatenate}(m, p)$$

As seen in Figure 5, the speech feature $A = \{a_1, a_2, \ldots, a_T\}$. Initially, it is input into the Conformer encoders. Each encoder conducts residual concatenation and layer normalization on the extracted feature vector a_0, followed by convolutional and linear layers to produce a new feature vector a1. Following processing by the i-layer encoders, the resulting feature sequence is forwarded to a three-layer LSTM [23,24] (Long Short-Term Memory) network.

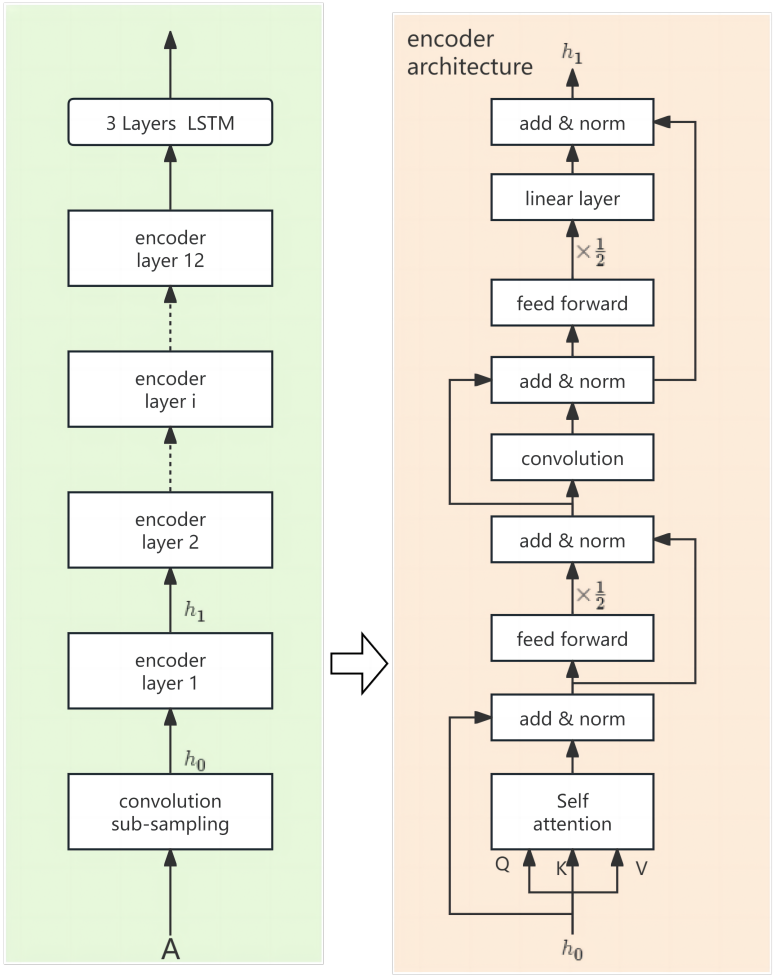

Figure 5. Audio feature extraction steps.

3.4. Text Feature Extraction

This study employs the Funnel-Transformer to extract features from all tokens in the text. The structure of the module is depicted in Figure 6 below.

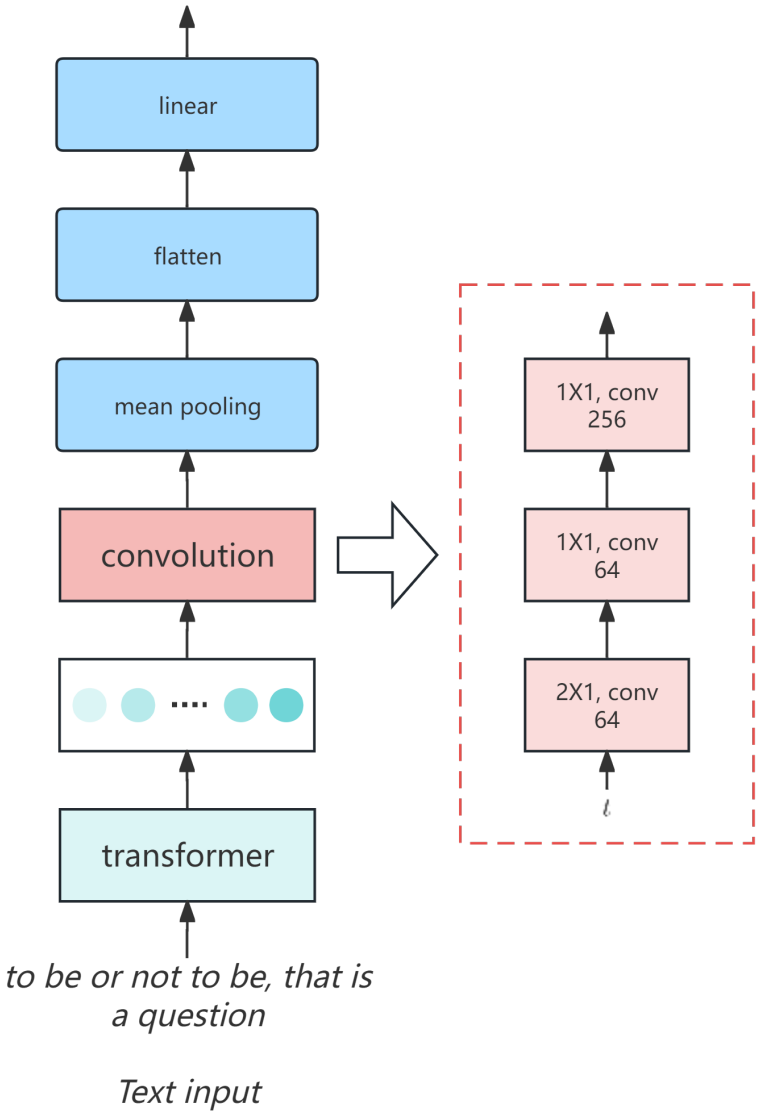

Figure 6. Text feature extraction steps.

Because of the parallelized nature of the model, directly capturing the positional information of the markers is not feasible. Thus, this paper employs the position-encoding technique. In this approach, each position is encoded using a sine–cosine function with varying frequencies to produce two-dimensional values. These values constitute unique position-encoding vectors, representing specific positions in the text sequence. The calculation process for odd and even positions is illustrated below.

$$\begin{cases} P_{i,2k} = \sin\left(\dfrac{i}{10000^{2k/d}}\right) \\ P_{i,2k+1} = \cos\left(\dfrac{i}{10000^{2k/d}}\right) \end{cases} \tag{1}$$

i is the position index, k is the dimension index, d is the dimension of the feature vector, and $P_{i,j}$ are the elements in the position-encoding matrix.

When text features $E = \{e_1, e_2, \ldots, e_T\}$ pass through the LSTM layer, Bi-LSTM [25,26] processes the input feature eT at each time step t, thereby learning and remembering the information in the sequence. Bi-LSTM processes the sequence E step-by-step in this way,

with each time step combining the current input feature and the previous state information to efficiently capture long-term dependencies.

3.5. Joint Attention Module

The model utilizes a joint attention mechanism to capture consistency information across different modalities. This mechanism is designed to capture consistency information among images, videos, and texts. It enhances the weights of features exhibiting consistent emotional expression while reducing the weights of features with inconsistent emotional expression. The structure of the model is depicted in Figure 7 below. The arrows in the figure indicate the flow of information from different features, and the portion across the line segments is represented in this paper using dashed lines. Different modalities explore and strengthen regions of emotional coherence by seeking feature information from the other two modalities.

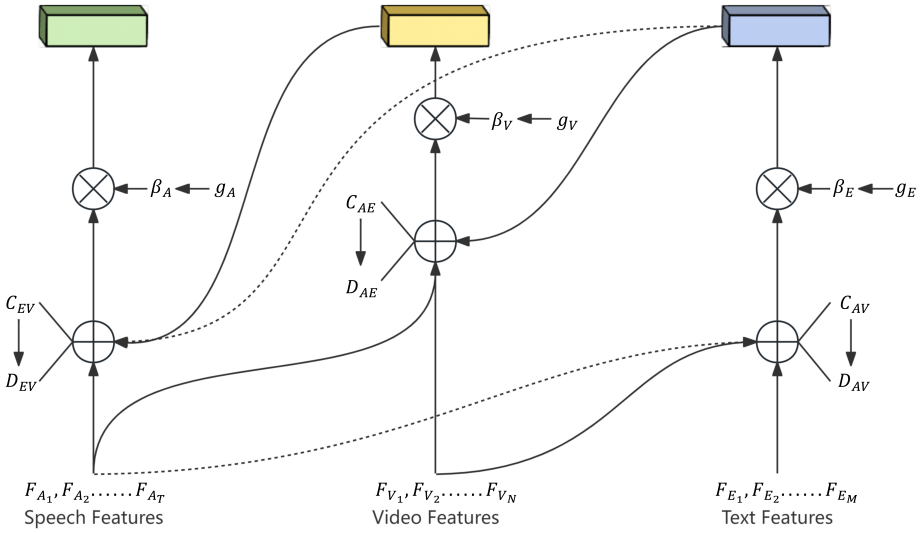

Figure 7. Joint attention mechanism.

The relationship between each pair of modal combinations (e.g., speech–text, text–video, video–speech) is captured by computing the crossover matrix. The video features V, speech features A, and text features E, as defined in the previous section, are used to construct the cross-matrix formula, where W represents the trainable weight matrix.

Audio–text crossover matrix (audio-to-text attention):

$$C_{AE} = \tanh(AW_{AE}E^{\top})$$

Audio–video cross-matrix (audio-to-video attention):

$$C_{AV} = \tanh(AW_{AV}V^{\top})$$

Text–video cross-matrix (text-to-video attention):

$$C_{EV} = \tanh(EW_{EV}V^{\top})$$

Text–audio cross-matrix (text-to-speech attention):

$$C_{EA} = \tanh(EW_{EA}A^{\top})$$

Video–audio cross-matrix (video to audio attention):

$$C_{VA} = \tanh(VW_{VA}A^{\top})$$

Video–text cross-matrix (video-to-text attention):

$$C_{VE} = \tanh(VW_{VE}E^{\top})$$

C_{MN} is the attention matrix from mode M to mode N and W_{MN} is the corresponding trainable weight matrix. D_{MN} is the similarity calculated from C_{MN}. For each modality, the similarity matrix guided by the other two modalities is then calculated.

Similarity matrix of audio A guided by video V and text E:

$$S_A^{(V,E)} = \tanh[(C_{AV} \otimes W_{AV}) \cdot (C_{AE} \otimes W_{AE})]$$

Similarity matrix of video V guided by audio A and text E:

$$S_V^{(A,E)} = \tanh[(C_{VA} \otimes W_{VA}) \cdot (C_{VE} \otimes W_{VE})]$$

Similarity matrix of text E guided by audio A and video V:

$$S_E^{(A,V)} = \tanh[(C_{EA} \otimes W_{EA}) \cdot (C_{EV} \otimes W_{EV})]$$

$S_X^{(Y,Z)}$ denotes the similarity matrix of the modes X induced by the modes Y and Z. Additionally, the model employs a gating mechanism to generate feature weights for each modality.

The audio feature weight is as follows:

$$g_A = \sigma\left(S_A^{(V,E)}\right) W_{g_A}$$

$$\beta_A = W_{p_A}^{\top} \tanh(g_A)$$

The text feature weight is as follows:

$$g_E = \sigma\left(S_E^{(A,V)}\right) W_{g_E}$$

$$\beta_E = W_{p_E}^{\top} \tanh(g_E)$$

The video feature weight is as follows:

$$g_V = \sigma\left(S_V^{(A,E)}\right) W_{g_V}$$

$$\beta_V = W_{p_V}^{\top} \tanh(g_V)$$

Here, g_x represents the gating signal, and W denotes the trainable weight matrix used to generate the eigenweights for the modality X. σ is the sigmoid activation function, which constrains the output between 0 and 1. Using the obtained eigenweights, the softmax function computes the attention weights for each modality, which are then utilized to obtain the weighted modal features.

Audio attention weights and weighted features:

$$a_A = \text{softmax}(\beta_A), \quad F = \sum_{i=1}^{K} a_{A_i} A_i$$

Text attention weights and weighted features:

$$a_E = \text{softmax}(\beta_E), \quad F = \sum_{i=1}^{K} a_{E_i} E_i$$

Video attention weights and weighted features:

$$a_V = \text{softmax}(\beta_V), \quad F = \sum_{i=1}^{K} a_{V_i} V_i$$

3.6. Chained Interactive Attention Module

The aim of this study is to offer insights into the affective complementarities among video, language, and text and to enable effective interactive integration across these modalities. To achieve this objective, this paper introduces an innovative chained interaction attention mechanism specifically designed for integrating multimodal information. The core of the mechanism is constructed based on a scaled dot product attention model with input parameters including query (Q), key (K), and value (V). The specific structural details of the model are depicted in Figure 8 below.

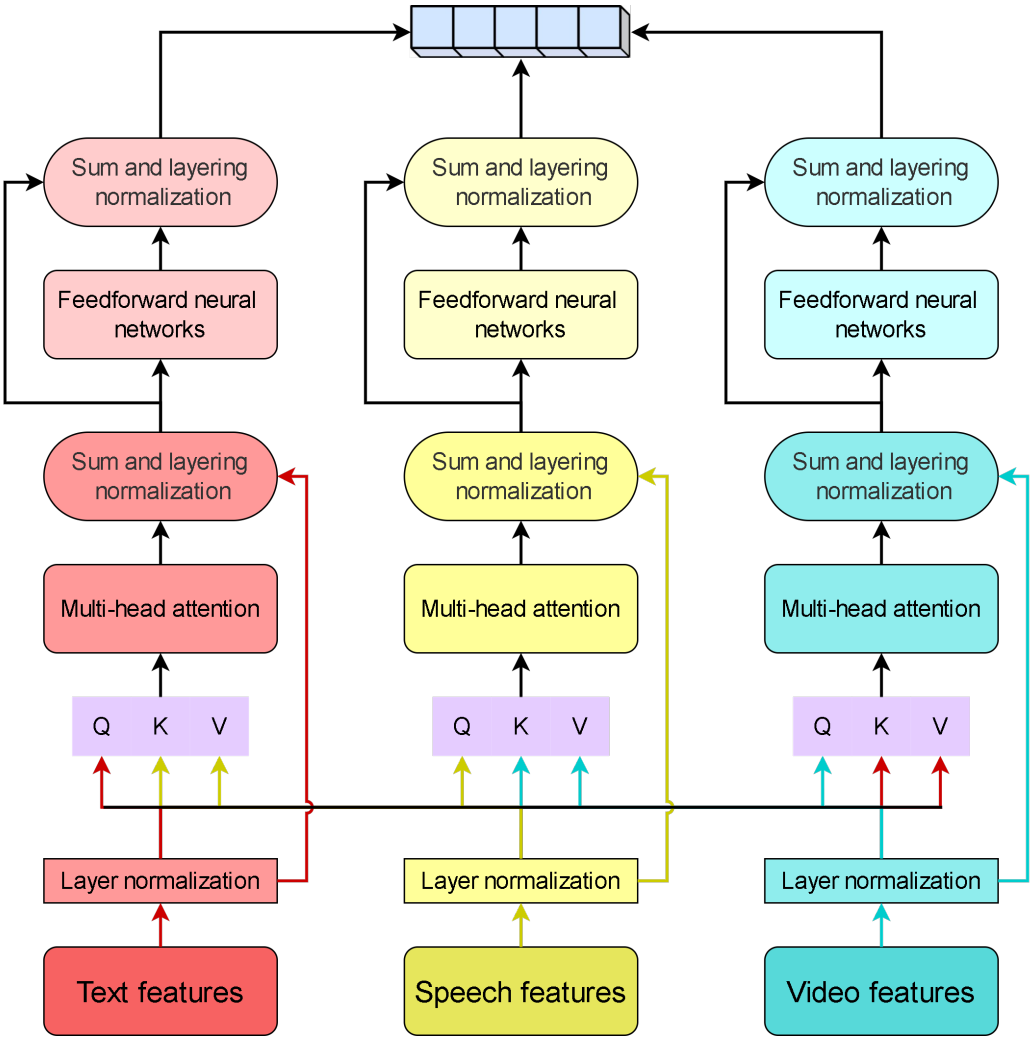

Figure 8. Chained interactive attention mechanism.

For the interactive fusion of text towards video, the video features are used as Q, and the text features are used as K and V, where $Q_v = W_{Q_v}$, $K_E = W_{K_E}$, and $V_E = W_{V_E}$.

The interactive attention of the text toward the video features is represented as follows:

$$f_V = V(Q, K, V) = \text{softmax}\left(\frac{Q_V(K_E^\top)}{\sqrt{d_k}}\right) V_E$$

Similarly, for interactive fusion from video toward audio, with audio features as Q and video features as K and V, the attention for adaptive fusion from video toward audio is represented as follows:

$$f_A = A(Q, K, V) = \text{softmax}\left(\frac{Q_A(K_V^\top)}{\sqrt{d_k}}\right)V_V$$

For interactive fusion from audio toward text, with text features as Q and audio features as K and V, the attention for adaptive fusion from audio toward text is represented as follows:

$$f_E = E(Q, K, V) = \text{softmax}\left(\frac{Q_E(K_A^\top)}{\sqrt{d_k}}\right)V_A$$

The video interaction feature P_V is obtained from a Feedforward Neural Network (FNN) [27] and Layer Normalisation (LN) [28]:

$$P_V = \text{LN}(\text{FFN}(f_V(F_V, F_E, F_E)))$$

Similarly, the speech interaction characterized by the interaction attention module is the following:

$$P_A = \text{LN}(\text{FFN}(f_A(F_A, F_V, F_V)))$$

The textual interactions characterized by the interactive attention module is the following:

$$P_E = \text{LN}(\text{FFN}(f_E(F_E, F_V, F_V)))$$

Finally, all the interaction features are spliced to obtain the desired emotion fusion feature P:

$$P = P_A \oplus P_E \oplus P_V$$

3.7. Loss Function

To fine-tune the overall model for the multimodal sentiment analysis task, this paper employs the Sentiment Consistency Cross Entropy Loss (MSCCE Loss). For the sentiment classification task, the cross-entropy loss is utilized to measure the discrepancy between the model output and the actual labels, as depicted in the following formula:

$$\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{2}$$

where C is the sample, y is the true label vector, $\hat{y}$ is the probability vector predicted by the model, y_c and $\hat{y}_c$ are the values of the cth element of these vectors. Additionally, to promote consistency across modalities, this paper introduces a loss term aimed at minimizing the distance between different modal representations of the same affective state and maximizing the distance across different affective states. Cosine similarity is employed as the consistency metric.

$$\mathcal{L}_{\text{consistency}} = 1 - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{F_{V_i} \cdot F_{A_i}}{\|F_{V_i}\|\|F_{A_i}\|} + \frac{F_{V_i} \cdot F_{E_i}}{\|F_{V_i}\|\|F_{E_i}\|} + \frac{F_{A_i} \cdot F_{E_i}}{\|F_{A_i}\|\|F_{E_i}\|}\right) \tag{3}$$

where F_{V_i}, F_{A_i}, and F_{E_i} denote the video, audio, and text feature vectors of the ith sample, respectively, and N is the total number of samples. Therefore, combining the cross-entropy loss and modal consistency loss, the total loss function is the following:

$$\mathcal{L}_{\text{MSCCE Loss}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{consistency}} \tag{4}$$

λ is a balancing factor to control the weight of the modal coherence loss in the total loss.

4. Experiments

4.1. Experimental Environment Settings

The setup of this experiment is shown in Table 2 below.

Table 2. Experimental environment configuration.

Experimental Environment	Environmental Configuration
System	Linux
GPU	RTX 4090
CPU	16 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10 GHz
Pytorch	1.11.0
Python	3.8
Cuda	11.3

The public datasets CMU-MOSEI [29] and IEMOCAP [30] were utilized in this experiment. The IEMOCAP dataset consists of recorded content featuring 10 male and female actors paired in pairs. It spans a total duration of 12 h, divided into 5 sessions. Dialogues are transcribed as text and labeled with 10 emotion categories, including anger, pleasure, and sadness. The utilization of this dataset aligns with previous studies and facilitates performance comparisons. CMU-MOSEI comprises over 20,000 film video clips sourced from YouTube, encompassing a broad spectrum of emotions such as happy, sad, and neutral. Each discourse represents a distinct multimodal instance conveying the speaker's perception of the film's theme. Emotion intensity was annotated by five assessors, with ratings ranging from −3 to +3. Here, −3 indicates strong negative emotions, while +3 indicates strong positive emotions. CMU-MOSEI also employed the same [−3, 3] emotion intensity scoring system. The datasets were divided into two groups corresponding to two tasks: polarity analysis and emotion recognition. Sentiment recognition is a classification task aimed at identifying specific sentiment categories within multimodal discourse. Sentiment analysis, on the other hand, is a regression task that represents the sentiment polarity of discourse through continuous values. The classification and division of the dataset into training, validation, and test sets are illustrated in Table 3 below; the ratio of training sets, validation sets, and test sets follow 7:1:2.

Table 3. Divided datasets.

Tasks	Data Set	Emotions	Divide	
			Train + Val	Test
Emotional polarity judgment	CMU-MOSEI	−3, −2, −1, 0, 1, 2, 3	17,830	4759
Multimodal sentiment analysis	IEMOCAP	Happy, Frustrated, Angry, Neutral, Sad	5568	1623

4.2. Evaluation Indicators

To comprehensively assess the performance improvement of the VAE-JCIA method in sentiment analysis tasks, this study introduces several evaluation metrics: mean absolute error (MAE) and Pearson's correlation coefficient (Corr).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

$$\mathrm{Pearson} = \frac{1}{n}\sum_{i=1}^{n}\frac{E(y_i\hat{y}_i) - E(y_i)E(\hat{y}_i)}{\sigma_{y_i}\sigma_{\hat{y}_i}} \tag{6}$$

Additionally, the predictions of sentiment analysis for each model can be rounded to 7 different sentiment polarities (−3 to 3) and used to calculate the 7 category accuracy (Accuracy7) as well as the 2 category accuracy (Accuracy2) and the 2 category F1 value ($F1_2$) of positivity and negativity, where positivity corresponds to sentiment polarities greater than zero and negativity corresponds to sentiment polarities less than zero.

For the emotion recognition task, this study employed weighted accuracy (WA) and unweighted accuracy (UA) as the primary metrics to evaluate the model's performance. Both of these accuracy metrics provide a comprehensive assessment of the model's performance, particularly when dealing with an unbalanced dataset. Weighted accuracy considers the imbalance in the sample size of each category, while unweighted accuracy reflects the model's average classification accuracy across all categories.

$$\mathrm{WA} = \frac{\sum_{i=1}^{n}(TP_i)}{\sum_{i=1}^{n}(TP_i + FP_i + FN_i)} \tag{7}$$

$$\mathrm{acc}_i = \frac{TP_i}{TP_i + FP_i} \times 100\% \tag{8}$$

$$\mathrm{UA} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{acc}_i \times 100\% \tag{9}$$

4.3. Parameter Settings

The hyperparameter settings of this paper are shown in Table 4 below. Additionally, a learning rate adjustment strategy with warm-up is employed, where the learning rate gradually increases in the first few epochs to reach 0.0001 and then progressively decays at a rate of 0.8, with a lower limit of 0.00001.

Table 4. Hyperparameters configuration.

Hyperparameter	Value
Optimization	Adam
Audio Shape	74
Video Shape	35
Text Shape	768
Output Dimension	1
Hidden Audio Size	64
Hidden Video Size	64
Hidden Text Size	64
Hidden Size	32
Learning Rate	0.00005
Batch Size	32
Number of Epochs	200
Early Stopping	10
Dropout Input	0.2
Attention Heads	5
Gradient Clipping	4.0

4.4. Comparison Models

To validate the performance of the VAE-JCIA method on the sentiment analysis task, several competitive tri-modal sentiment analysis benchmark models were selected for the experiments.To test the effectiveness of the models on sentiment polarity detection, the experiments were selected to include the following:

Recurrent neural model using memory fusion Network: MFN. Models focusing on temporal multimodal analysis and feature splicing fusion: MARN, RMFN. Models that use attention mechanisms and Transformer modules for cross-modal information learning: MulT. A model for multimodal fusion based on graph neural networks: graph-MFN. Model for multimodal fusion using tensor: TFN. Commonly used subspace learning model: the MCTN with cyclic displacement mechanism. Multimodal sentiment analysis based on long-short-time features and decision-level fusion: the LSTF-DFusion. Multimodal fusion model for extracting contrast information from multimodal time series: Deep-HOSeq. A multilayer perceptron-based model for sentiment analysis: CubeMLP. A model for representing common properties of modalities: MISA. Self-supervised multi-task learning model: Self-MM.

To test the effectiveness of the models on sentiment classification tasks, the experiments were selected to include the following:

Multimodal dual recurrent encoder model: MDRE. Fine-grained cross-modal excitation speech recognition model: FG-CME. Bidirectionally aligned network multimodal emotion recognition model: GBAN. Multimodal dynamic fusion network-based emotion recognition model: MM-DFN. Multi-view layer attention network-based sentiment recognition model: DIMMN.

4.5. Results

In the experimental part, we conducted multiple rounds of a randomized validation of the model by splitting the dataset and randomly selecting the validation according to the above ratio; the results of each validation are shown in Tables 5 and 6 below.

Table 5. Randomized validation results for dichotomous and heptachotomous polarity.

Round	$Accuracy_2$	$F1_2$	$Accuracy_7$	$F1_7$
1	85.92	0.8894	53.09	0.5436
2	85.81	0.8887	53.17	0.5460
3	85.89	0.8891	52.92	0.5427
4	85.78	0.8882	52.92	0.5421
5	85.87	0.8890	53.09	0.5450
6	85.78	0.8876	52.85	0.5422
7	85.81	0.8886	52.94	0.5436
8	85.91	0.8822	53.07	0.5433
9	85.99	0.8897	53.17	0.5462
10	85.99	0.8897	53.11	0.5433
Average	85.88	0.8889	53.03	0.5438

As can be seen through Table 5, in the dichotomous polarity task weight, $Accuracy_2$ has a mean value of 85.88%, with values fluctuating between 85.78% and 85.95% over ten rounds of randomized validation.The metrics for $F1_2$ are more stable, with a mean value of 0.8889, with a low of 0.8876 (round 6) and a high of 0.8897 (rounds 9 and 10). The seven-category polarity task weighs heavily, with a mean of 53.03% for $Accuracy_7$, and its value fluctuates slightly between 53.03% and 53.17%. $F1_7$ has a mean of 0.5438, and this metric also varies modestly over the 10 rounds, ranging from 0.5421 (round 4) to 0.5462 (round 9).

Table 6. Randomized validation results for a five-category sentiment analysis task.

Round	WA%	UA%
1	71.94	70.31
2	70.11	69.77
3	71.55	70.22
4	71.89	70.28
5	71.01	70.12
6	71.54	70.28
7	70.55	69.93
8	71.89	70.38
9	71.74	70.15
10	71.34	70.73
Average	71.36	70.22

Table 6 shows that the average WA score of the model is 71.36%, the maximum score is 71.94%, and the minimum score is 70.11%. The average UA score of the model is 70.22%, the maximum score is 70.73%, and the minimum score is 69.77%.

4.5.1. Polarity Analysis Results

The results of the comparative experiments for multimodal sentiment analysis are shown in Table 7 below, where the data are derived from the literature [11,13,14,31–41] and from the available open source code replications.

Table 7 with present in detail the performance of various models when performing polarity analysis on the CMU-MOSEI dataset. The results show that all the models proposed in this paper achieve excellent performance on the CMU-MOSEI dataset.

Table 7. Sentiment analysis results of different models on CMU-MOSEI dataset.

Model	MAE/%	Corr/%	$Accuracy_2$/%	$Accuracy_7$/%	$F1_2$
Graph-MFN	0.623	0.677	81.58	45.00	0.827
TFN	0.593	0.700	80.80	-	0.825
MARN	0.587	0.627	79.80	34.70	0.836
RMFN	0.565	0.679	82.10	38.30	0.814
MCTN	0.551	0.667	83.66	-	0.823
Deep-HOSeq	0.551	0.688	84.22	44.17	0.846
LSTF-Dfusion	0.546	0.691	85.73	46.35	0.851
CubeMLP	**0.529**	0.760	85.10	-	0.845
MISA	0.568	0.724	84.20	-	0.840
Self-MM	0.724	0.762	85.20	52.90	0.851
MuIT	0.580	0.703	82.50	-	0.823
Ours	0.542	**0.774**	**85.88**	**53.03**	**0.889**

The analysis of the confusion matrix in Figure 9 indicates that the model demonstrates overall accuracy in classifying the polarity of emotions. The diagonal line exhibits the highest number of correct classifications, but there are still instances of misclassifications on both sides. It is important to consider that emotions' polarities may be adjacent in intensity, such as -2 and -3 or 2 and 3, and may exhibit similar expression, posing challenges for the model in differentiation. Consequently, this level of error remains within acceptable tolerance levels.

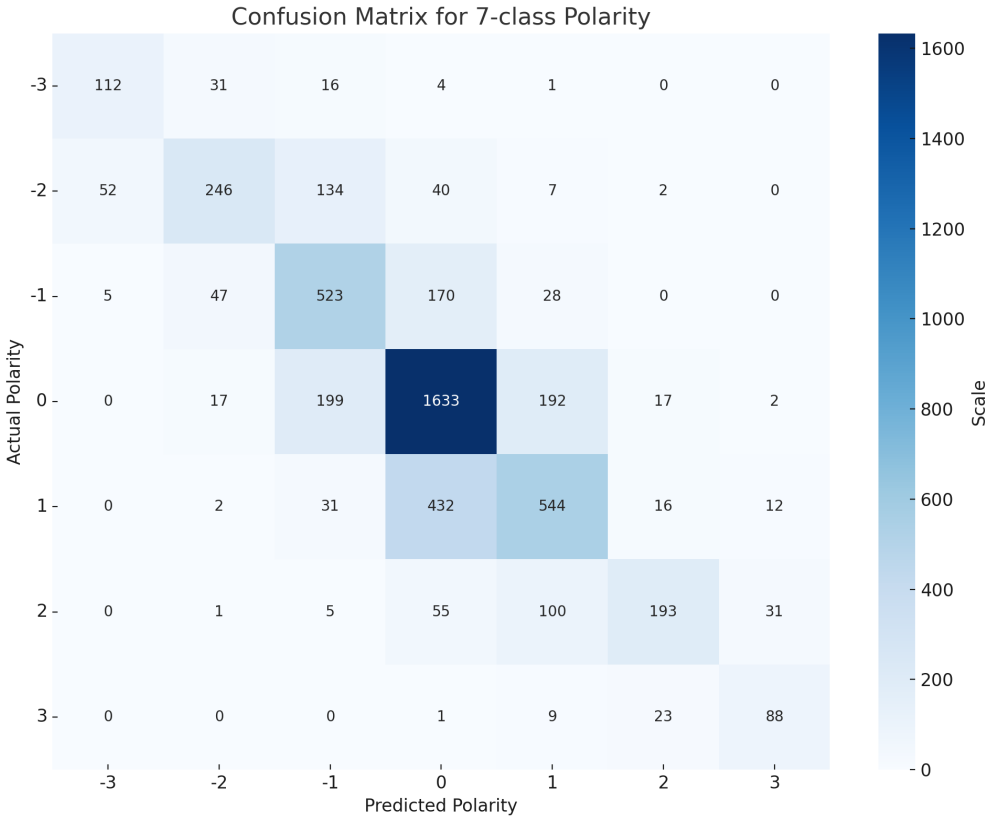

Figure 9. Polarity analysis confusion matrix.

4.5.2. Sentiment Recognition Results

The results of emotion recognition are shown in Table 8, where the data are taken from the literature [42–44].

Table 8. Sentiment recognition results of different models on the IEMOCAP dataset.

Model	WA/%	UA/%
MM-DFN	68.21	-
GBAN	71.39	70.08
MDRE	**71.80**	71.40
DIMMN	64.70	-
FG-CME	71.01	**71.66**
Ours	71.26	70.15

The results of the comparison experiments among different models for the emotion recognition task are presented in Table 8. The evaluation using the IEMOCAP dataset indicates that the experimental models proposed in this paper yield favorable results compared to other bimodal and trimodal sentiment analysis models. Although the values of WA and UA cannot surpass the MDRE model, considering that the MDRE model is only a bimodal model (audio and text), the fact that the model in this paper can achieve such values can still be called an excellent result.

The confusion matrix for IEMOCAP is depicted in Figure 10. The analysis of the confusion matrix reveals a notable proportion of transitions between similar emotions in the dataset, such as happy and excited. Additionally, a considerable number of instances labeled as happy are incorrectly predicted as excited by the model. The author speculates that the unbalanced distribution of training samples may contribute to this issue, particu-

larly as samples labeled as happy are the least represented in the dataset. Consequently, the model may assign lower training priority to a few classes, leading to confusion.

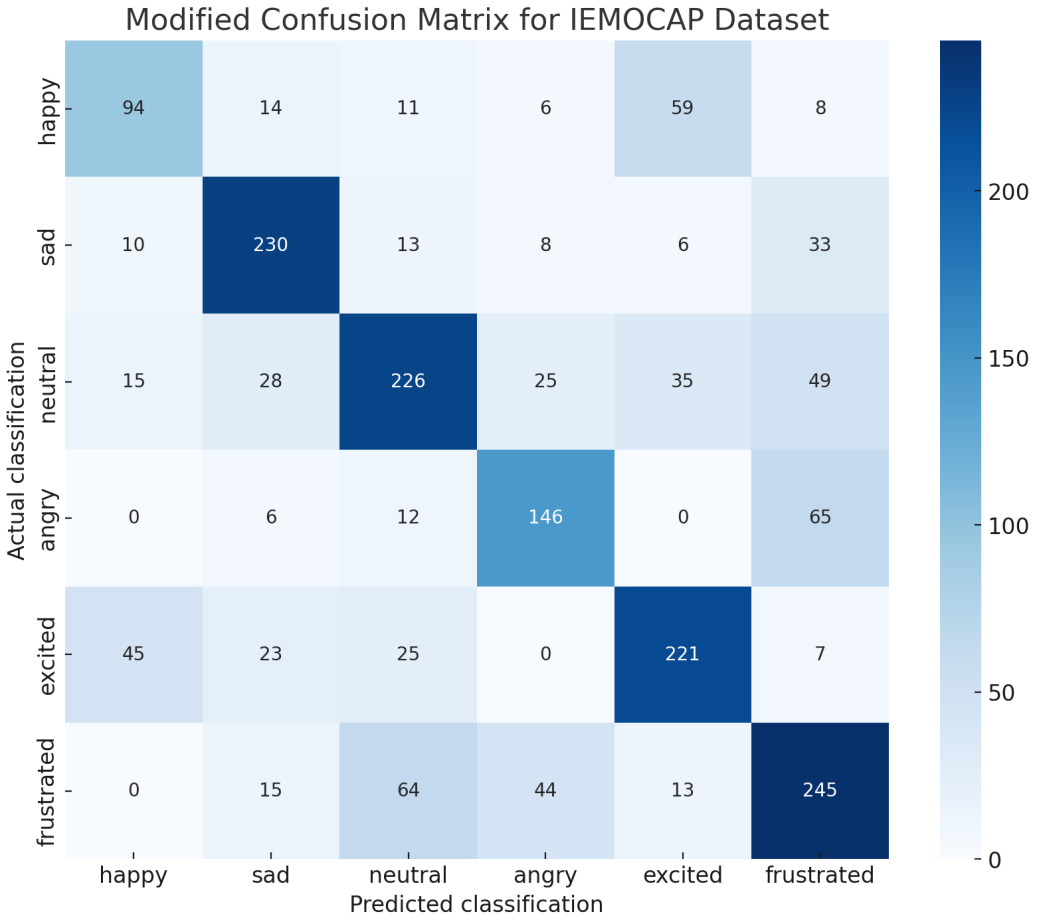

Figure 10. Sentiment analysis confusion matrix.

4.6. Analysis of Ablation Experiments

4.6.1. VAE-JCIA Methodological Validity

To evaluate the impact of the VAE-JCIA method on sentiment analysis performance across different modules and structures, this study conducted ablation experiments on the IEMOCAP and CMU-MOSEI datasets. The ablation experimental methods included the following: VAE-JCIA, a multimodal fusion sentiment analysis method utilizing joint chained interactive attention; VAE-JA, a multimodal fusion sentiment analysis method utilizing joint attention (without chained interactive attention); and VAE-CIA, a multimodal fusion sentiment analysis method based on chained interactive attention (without joint attention).

The specific results are shown in Table 9. In order to uniformly compare the two datasets, we separately calculated the accuracy and F1 scores for the five-category classification of the IEMOCAP dataset and placed them in the same table with the results from the CMU-MOSEI dataset.

Table 9. Results of ablation experiments on IEMOCAP and CMU-MOSEI datasets.

Model	IEMOCAP			CMU-MOSEI		
	$Accuracy_5$	$Recall_5$	$F1$	$Accuracy_2$	$Recall_2$	$F1$
VAE-CIA	0.670	0.724	0.6968	0.8180	0.8891	0.8517
VAE-JA	0.649	0.614	0.6302	0.7920	0.8837	0.8351
VAE-JCIA	0.722	0.778	0.7492	0.8510	0.9096	0.8846

4.6.2. Modal Fusion Sequential Validity

To further investigate the impact of the fusion order of different modes on multimodal sentiment analysis, this study conducted cross-modal fusion with text, audio, and visuals as the primary modes, and the remaining modes were labeled as the auxiliary modes. The results of the ablation experiments are presented in Table 10.

Table 10. Results of ablation experiments adjusting different modal fusion orders.

Modal Fusion Order	MAE	Corr	Acc_2/%	$F1$	Acc_7/%
A→E→V→A	0.711	0.603	83.71	0.8491	51.09
A→V→E→A	0.728	0.609	83.64	0.8352	49.79
V→A→E→V	0.886	0.646	83.18	0.8291	51.34
V→E→A→V	0.643	0.678	83.04	0.8281	52.45
E→V→A→E	0.545	**0.775**	84.82	0.8757	**53.59**
E→A→V→E	**0.542**	0.774	**85.88**	**0.8889**	53.03

The experimental results demonstrate that the cross-modal fusion strategy with text (E) as the primary modality outperforms the strategy with audio (A) or visual (V) as the primary modality, underscoring the crucial role of text in multimodal sentiment analysis. The sequence of fusing text modalities with audio modalities first and then with visual modalities may mirror the way the human brain processes multimodal emotions, potentially resulting in superior performance compared to other cascading fusion orders. Furthermore, if the video or speech modality serves as the primary information carrier and textual information is limited, utilizing text as a secondary modality may be more effective.

4.6.3. Balancing Factor Adjustments

In the context of the VAE-JCIA model, λ serves as an equilibrium coefficient that influences not only the model's emphasis on single-modal sentiment categorization but also the emphasis on inter-modal consistency. Thus, this experiment focuses on exploring the optimal value of λ within a reasonable range while keeping other variables constant.

Upon analyzing the fold trend, according to Figure 11, it can be seen that the corresponding values of the balance coefficient λ when the model achieves the best results under the two datasets are approximately 0.45 and 0.55. This phenomenon is attributed to the significant differences in data distribution between the different corpora, which necessitates adjusting the distance between similar samples with different balancing coefficients to align with the loss function consistency during model operation. Additionally, the model's effectiveness decreases when the balancing coefficient λ takes on larger or smaller values. This occurs because smaller values of λ cause the model to focus more on the data in the respective modality and disregard the correlation between different modalities. Moreover, larger values of λ diminish the model's ability to mine sample-specific features.

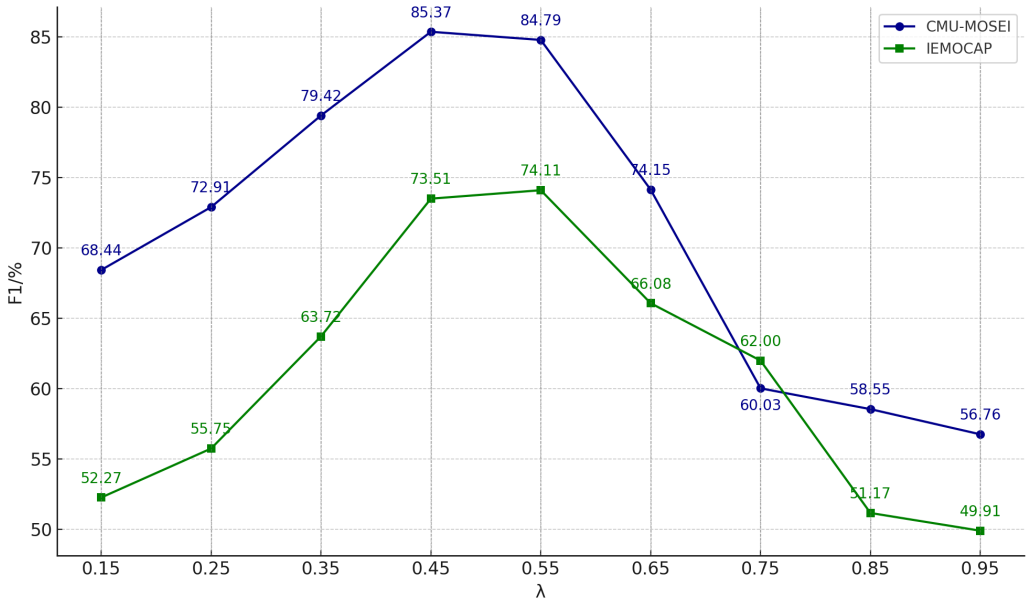

Figure 11. Variation in F1 value with equilibrium coefficient.

4.6.4. Noise Robustness Testing

To evaluate the model's robustness against noise in both video and audio, Gaussian noise was incorporated into the video data to replicate scenarios like device quality issues and transmission errors, while white noise was introduced to the audio data to mimic environmental noise. Specifically, a noise matrix mirroring the dimensions of the video or audio input is created. Each value within this noise matrix is independently sampled from a Gaussian distribution with a mean of zero. The noise values generated are then scaled by a predefined noise level (e.g., 0.05), aimed at regulating the extent of noise influence on the original data. Following scaling by the noise level, the standard deviation of the noise becomes 0.05 times that of the original data, and subsequently, the resulting noise value is added to the original input data. Thus, each data point undergoes perturbation to a certain extent, with the magnitude of perturbation governed by the noise value. This method of introducing noise allows us to simulate random errors that could arise during data acquisition and transmission, thereby assessing the model's resilience to such noise. The specific experimental results are shown with reference to Figure 12 below.

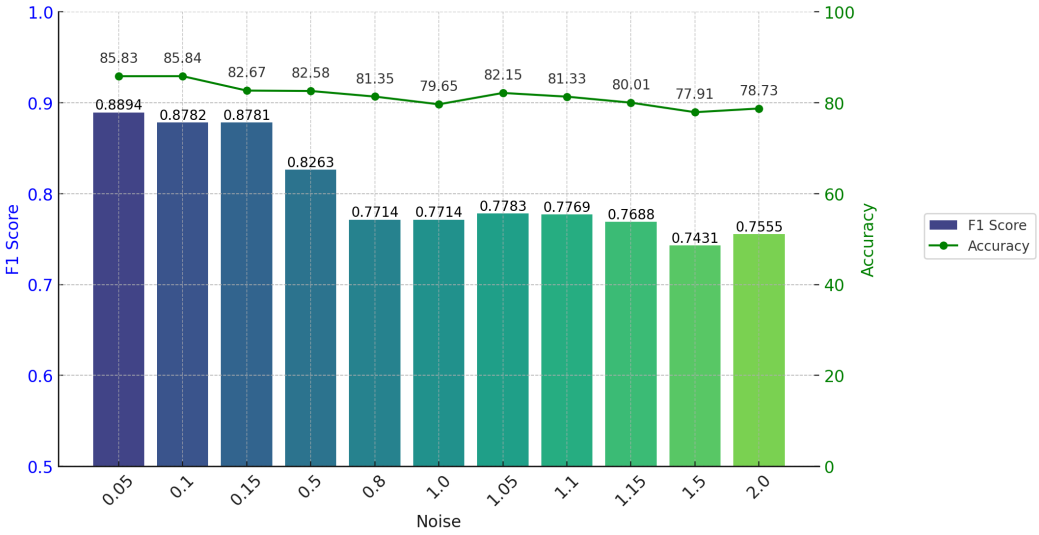

Figure 12. Changes in F1 score and accuracy with noise.

From the data in the figure above, it is clear that as the noise level increases, the F1 value and accuracy are affected to a certain extent. When the noise level is at a low level, it has a slight effect on the model, but as the noise level increases, especially when it exceeds 0.5, the F1 value appears to fall off a cliff and then increases the noise level again until the noise level is raised to 2. The F1 values are leveled off. The authors speculate that the constant influence of noise makes it difficult for the model's attentional mechanism to extract key features from the other two modalities, so the model gradually degenerates into a sentiment analysis judgment based on the information of the textual modality, which affects the correctness and stability of the model's judgment.

5. Conclusions

In this paper, we propose a VAE-JCIA method based on joint chained interactive attention, aimed at addressing the challenge of integrating coherent information across video, audio, and text modalities in multimodal sentiment analysis, a concern often overlooked in mainstream approaches. This method utilizes a joint attention network to concentrate on the user's emotional expression across three modalities and improves the emotional coherence among them based on their correlations. Subsequently, these features are progressively fused using chained interactive attention to enhance sentiment analysis performance. Experimental results demonstrate that the approach proposed in this study offers a novel perspective for multimodal sentiment analysis. In the future, it will help to be applied within areas such as AI dialogue, intelligent interviewing, and character profiling. However, the current approach has some limitations.

Although the approach in this paper demonstrates significant innovation and effectiveness in handling multimodal data fusion, the authors still identified several noteworthy limitations and potential room for improvement during the experimental process:

(1) Although the information fusion between different modalities is enhanced by the joint chained interactive attention mechanism, the fusion depth and efficiency of the mechanism may be limited by the information contained in the initial representation of the modal features. It is now difficult to achieve a 100% complete parsing of the absolute emotions of characters in video and audio, and it is a challenge for the model to achieve deeper information fusion while maintaining computational efficiency.
(2) Although the paper was validated on the CMU-MOSEI and IEMOCAP datasets, these datasets may not be able to encompass all the multimodal variations in emotional expressions in real-world scenarios. Therefore, the generalization ability of the model and its applicability in different application contexts are issues that need to be considered in future research.
(3) In real scenarios, there may be information inconsistency or direct conflict between different modalities. Whether current models can effectively solve such problems and how to optimally deal with such modal inconsistencies remain issues worth exploring in the future.

In the future, we will try to implement other methods to overcome the problems mentioned above.

Author Contributions: K.Q. was responsible for the design of the experiments and the writing of the manuscript; Y.Z. was responsible for the discussion and design of the experiments; S.Z. and J.Z. were responsible for the editing of the manuscript format and the processing of the pictures; and F.C. was responsible for the supervision of the manuscript and the suggestion of revisions. Q.W. was responsible for providing technical support and financial assistance. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that were used are confidential.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl. Based Syst.* **2022**, *235*, 107643. [CrossRef]
2. Zhu, Y.; Dong, J.; Xie, L.; Wang, Z.; Qin, S.; Xu, P.; Yin, M. Recurrent Multi-View Collaborative Registration Network for 3D Reconstruction and Optical Measurement of Blade Profiles. *Knowl. Based Syst.* **2024**, *295*, 111857. [CrossRef]
3. Chen, L.; Guan, Z.Y.; He, J.; Peng, J. A survey on sentiment classification. *J. Comput. Res. Dev.* **2017**, *54*, 1150–1170.
4. Zhou, J.; Zhao, J.; Huang, J.X.; Hu, Q.V.; He, L. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing* **2021**, *455*, 47–58. [CrossRef]
5. Zhu, L.; Zhu, Z.; Zhang, C.; Xu, Y.; Kong, X. Multimodal sentiment analysis based on fusion methods: A survey. *Inf. Fusion* **2023**, *95*, 306–325. [CrossRef]
6. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2023**, *91*, 424–444. [CrossRef]
7. Fu, Z.; Liu, F.; Xu, Q.; Qi, J.; Fu, X.; Zhou, A.; Li, Z. NHFNET: A non-homogeneous fusion network for multimodal sentiment analysis. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18-22 July 2022; IEEE: New York, NY, USA, 2022; pp. 1–6.
8. Cao, D.; Ji, R.; Lin, D.; Li, S. A cross-media public sentiment analysis system for microblog. *Multimed. Syst.* **2016**, *22*, 479–486. [CrossRef]
9. Cao, D.; Ji, R.; Lin, D.; Li, S. Visual sentiment topic model based microblog image sentiment analysis. *Multimed. Tools Appl.* **2016**, *75*, 8955–8968 [CrossRef]
10. Zhang, Y.; Song, D.; Li, X.; Zhang, P.; Wang, P.; Rong, L.; Yu, G.; Wang, B. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Inf. Fusion* **2020**, *62*, pp. 14–31. [CrossRef]
11. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
12. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.-P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
13. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. *Proc. Conf. Assoc. Comput. Linguist. Meet.* **2019**, *2019*, 6558.
14. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
15. Xu, N.; Mao, W.; Chen, G. Multi-interactive memory network for aspect based multimodal sentiment analysis. *Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 371–378. [CrossRef]
16. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* , **2018**, arXiv:1810.04805.
17. Guangbin, B.; Li, G.; Wang, G. Bimodal Interactive Attention for Multimodal Sentiment Analysis. *J. Front. Comput. Sci. Technol.* **2022**, *16*, 909.
18. Hu, H.; Ding, Z.; Zhang, Y.; Liu, M. Images-Text Sentiment Analysis in Social Media Based on Joint and Interactive Attention. *J. Beijing Univ. Aeronaut. Astronaut.* **2023**. (In Chinese)
19. Fan, T.; Wu, P.; Wang, H.; Ling, C. Sentiment Analysis of Online Users Based on Multimodal Co-attention. *J. China Soc. Sci. Tech. Inf.* **2021**, *40*, 656–665. (In Chinese)
20. Dai, Z.; Lai, G.; Yang, Y.; Le, Q. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4271–4282.
21. Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Shen, Y.; Mariconti, E.; Vervier, P.-A.; Stringhini, G.T. Predicting security events through deep learning. *arXiv* **2019**, arXiv:1905.10328
24. Shahid, F.; Zameer, A.; Muneeb, M. A novel genetic LSTM model for wind power forecast. *Energy* **2021**, *223*, 120069. [CrossRef]
25. Fang, X.; Xu, M.; Xu, S.; Zhao, P. A deep learning framework for predicting cyber attacks rates. *EURASIP J. Inf. Secur.* **2019**, *2019*, 5. [CrossRef]
26. Yao, Z.; Zhang, T.; Wang, Q.; Zhao, Y. Short-term power load forecasting of integrated energy system based on attention-CNN-DBILSTM. *Math. Probl. Eng.* **2022**, *2022*, 1075698. [CrossRef]
27. Bengio, Y.; Ducharme, R.; Vincent, P. A neural probabilistic language model. *Adv. Neural Inf. Process. Syst.* **2000**, *33*, 4271–4282.
28. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.

29. Zadeh, A.A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.-P. 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 (Volume 1: Long Papers); Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, VIC, Australia, 2018; pp. 2236–2246.
30. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359 [CrossRef]
31. Akhtar, M.S.; Ekbal, A.; Cambria, E. How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Comput. Intell. Mag.* **2020**, *15*, 64–75. [CrossRef]
32. Krommyda, M.; Rigos, A.; Bouklas, K.; Amditis, A. An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media. *Informatics* **2021**, *8*, 19. [CrossRef]
33. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.-P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
34. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.-P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
35. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.-P.; Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. *Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 6892–6899. [CrossRef]
36. Liang, P.P.; Liu, Z.; Zadeh, A.; Morency, L.-P. Multimodal language analysis with recurrent multistage fusion. *arXiv* **2018**, arXiv:1808.03920.
37. Wang, H. Sentiment Analysis Based on Multimodal Feature Fusion. Master's Thesis, Nanjing University of Posts and Telecommunications, Nanjing, China, 2023.
38. Verma, S.; Wang, J.; Ge, Z.; Shen, R.; Jin, F.; Wang, Y.; Chen, F.; Liu, W. Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis. In Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, 17–20 November 2020; IEEE: New York, NY, USA, 2020; pp. 561–570.
39. Sun, H.; Wang, H.; Liu, J.; Chen, Y.-W.; Lin, L. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3722–3729.
40. Shi, P.; Hu, M.; Shi, X.; Ren, F. Deep Modular Co-Attention Shifting Network for Multimodal Sentiment Analysis. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 109. [CrossRef]
41. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proc. Aaai Conf. Artif. Intell. **2021**, *35*, 10790–10797. [CrossRef]
42. Yoon, S.; Byun, S.; Jung, K. Multimodal speech emotion recognition using audio and text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; IEEE: New York, NY, USA, 2018; pp. 112–118.
43. Hu, D.; Hou, X.; Wei, L.; Jiang, L.; Mo, Y. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 7037–7041.
44. Wen, J.; Jiang, D.; Tu, G.; Liu, C.; Cambria, E. Dynamic interactive multiview memory network for emotion recognition in conversation. *Inf. Fusion* **2023**, *91*, 123–133. [CrossRef]