DOI: 10.53469/wjimt.2025.08(09).12

Multimodal Fusion Technology for Analyzing Children's Emotions Based on the Attention Mechanism

Chuyao Ma, Caixia Sun, Wei Shen

Huzhou University, Huzhou 313000, Zhejiang, China

Abstract: ***To enhance the emotion recognition ability of preschool education dialogue robots, this paper proposes a multimodal fusion model based on the cross-modal Transformer architecture. The model consists of feature extraction, fusion, and output layers. It extracts multi-source data through BERT, audio via AFEU units, and OpenFace toolkit. The multi-head self-attention mechanism is introduced to obtain high-level features, with text as an auxiliary and audio-video as the main modalities. The improved cross-modal Transformer and AVFSM module are used to fuse features and achieve emotion recognition. Experiments show that in the CH-SIMS and self-built Tea datasets, the model outperforms the baseline model in classification and regression metrics, verifying the effectiveness of each component. It has good robustness and generalization ability, and has a good application prospect in preschool education and other fields.***

Keywords: Emotion analysis of young children; Cross-modal Transformer architecture; Multi-head self-attention mechanism; Multi-modal fusion.

1. INTRODUCTION

With the continuous evolution of intelligent technologies, educational informatization has become a frontier hotspot in research and practice. As a typical representative of intelligent technology, conversational robots have been widely applied in many fields such as medical services and home services. Human-machine emotional interaction, as a core function of intelligent service robots, is crucial to enhancing the interaction experience with its recognition accuracy. Current research on robot emotion recognition is mainly based on convolutional neural networks; Specific research focuses on single-modal emotion classification and prediction, such as emotion recognition based on text information, image emotion recognition, speech emotion recognition, and emotion recognition based on ECG physiological signals, etc. Although a single recognition method helps robots understand human emotions, it has limitations such as a high rate of misjudgment. For this purpose, this study takes the preschool education dialogue robot as the carrier, introduces multimodal fusion based on the cross-modal Transformer architecture on the basis of existing research, and introduces the multi-head self-attention mechanism, combined with the results of expression emotion recognition and continuous speech emotion recognition, aiming to improve the preschool education dialogue robot's ability to recognize children's emotions.

2. ARCHITECTURE OF THE MODEL

The Architecture of the model is shown in Figure 1. From bottom to top, the model consists of the feature extraction layer, the fusion layer, and the output layer. First, text, audio, and visual features are extracted respectively through different feature extraction methods. After obtaining the low-level features of these three, input them into the multi-head self-attention mechanism to extract the high-level features. The attention mechanism is composed of three weight matrices Q, K, and V. In the multimodal feature fusion stage, the K and V matrices are provided by text modality as the auxiliary modality, and the query vector Q is provided by audio and visual modality as the main modality. Through the cross-modal attention mechanism, the output is residually connected to the high-level features obtained by the multi-head self-attention mechanism. The T-A and T-V fusion features for high-level text feature fusion are obtained, while for Audio and video feature fusion, the Audio and video feature sense module (AVFSM) proposed in this paper is used for fusion to obtain A-V fusion features. The three fused features are concatenated by dimension and then input into the soft attention mechanism layer for feature selection. Finally, the results are fed into the fully connected layer for classification prediction.

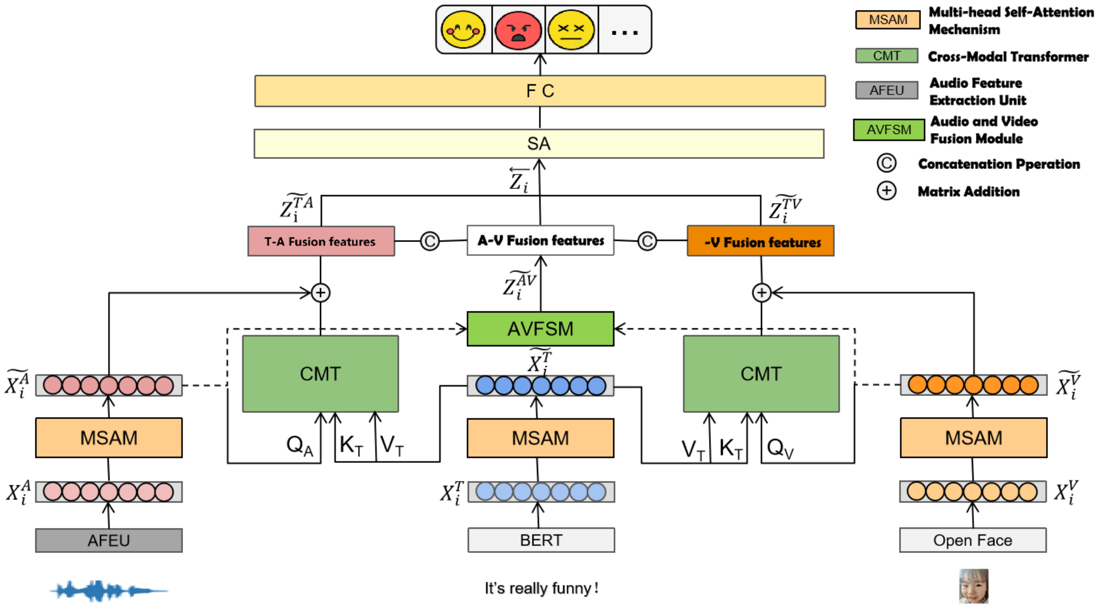


Figure 1: Overall architecture diagram of the model based on cross-modal Transformer fusion

3. THREE-MODAL FEATURE EXTRACTION

3.1 Text Sentiment Modal Feature Extraction Based on BERT Model

For text sentiment features, we use the pre-trained model BERT, which consists of a bidirectional Transformer capable of capturing context information of the text and is widely used in tasks in the NLP field, as shown in Figure 2.

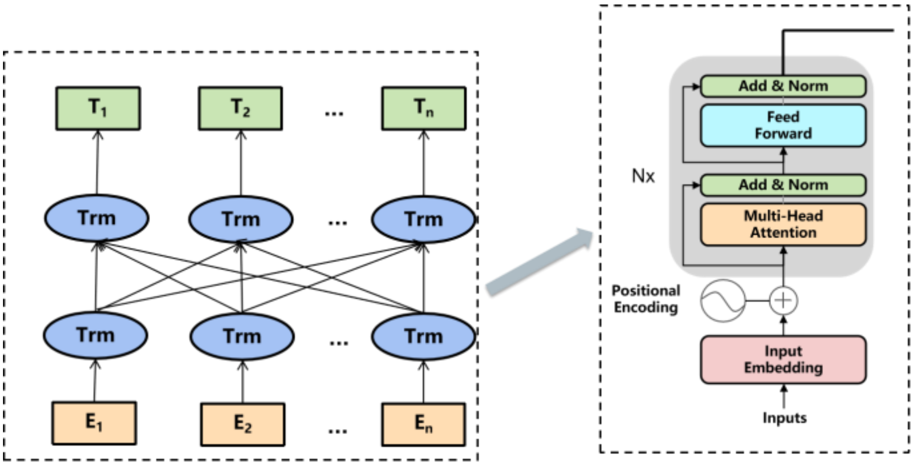


Figure 2: Basic Structure diagram of the BERT model

The process of extracting text features by the BERT model is shown in Figure 3. For a text of length L, after passing through the BERT model, the output is an LX768-dimensional feature matrix.

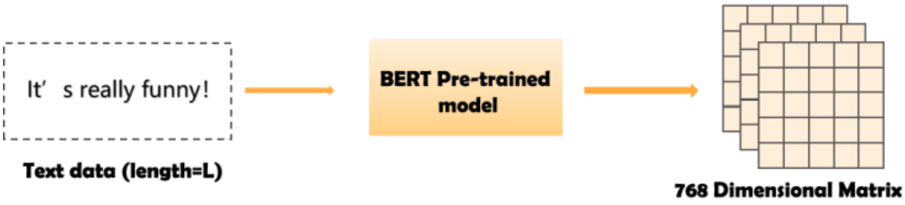


Figure 3: Text Feature Extraction

When using BERT to embed word vectors in sentence-level text, it maps each word fragment (token) to a 768-

dimensional vector representation. For each word segment, it generates word embeddings, segment embeddings, and position embeddings. The three embedding vectors are added together to form the final input representation, which is then fed into the multi-layer bidirectional Transformer encoder to obtain the context information of each layer of words using the self-attention mechanism. Finally, vectors labeled with [CLS] are extracted to represent the overall features of the sentence, or average pooling is performed on each word fragment vector to generate sentenced global features that can be used for downstream tasks such as sentiment classification and similarity calculation, as shown in Figure 4.

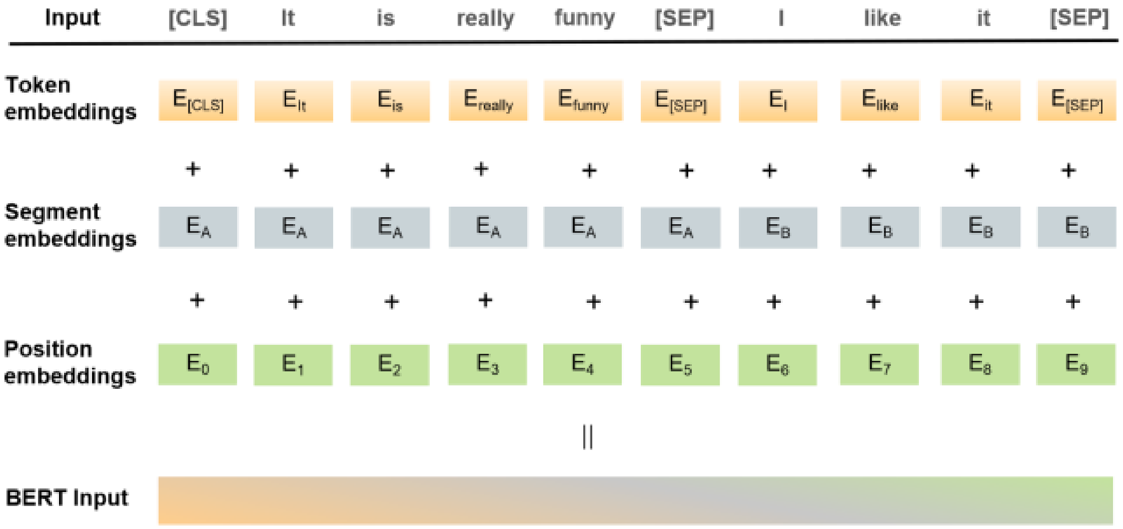


Figure 4: Word Vector Embeddings

3.2 Audio Modal Feature Extraction Based on the Audio Feature Extraction Unit (AFEU)

Accurate extraction of audio features is particularly important for multimodal affective analysis of young children. However, most of the existing audio feature extraction methods rely on a single feature (such as MEL frequency cepstral coefficients MFCC), making it difficult to fully characterize the audio emotional features. Therefore, in order to extract richer and finer acoustic Feature representations, this paper proposes an Audio Feature Extraction Unit (AFEU) whose structure is shown in Figure 5. The unit takes into account spectrogram features, prosodic features, sound quality features, and the widely used MFCC features, aiming to extract rich audio affective features.

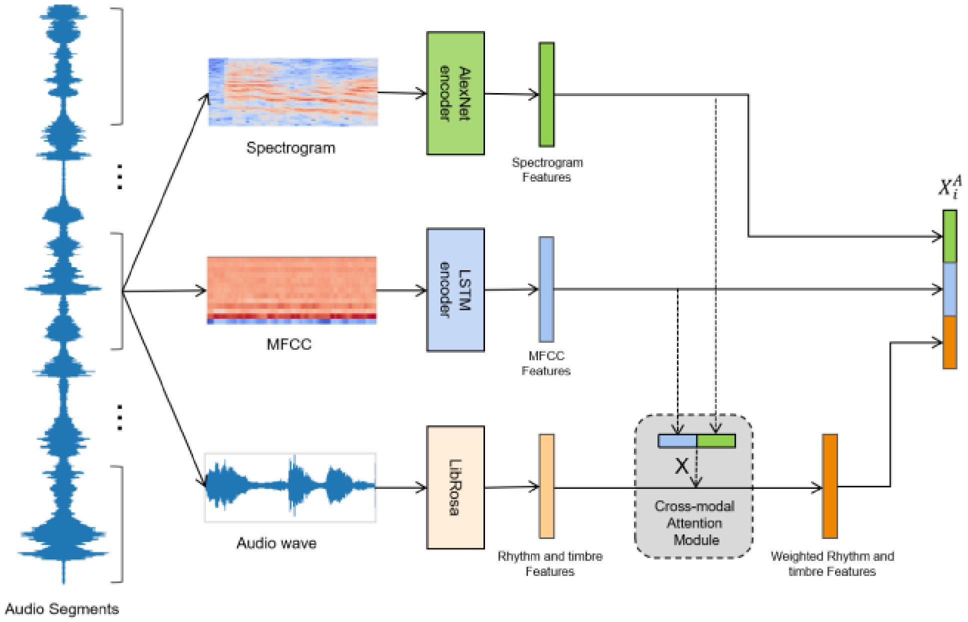
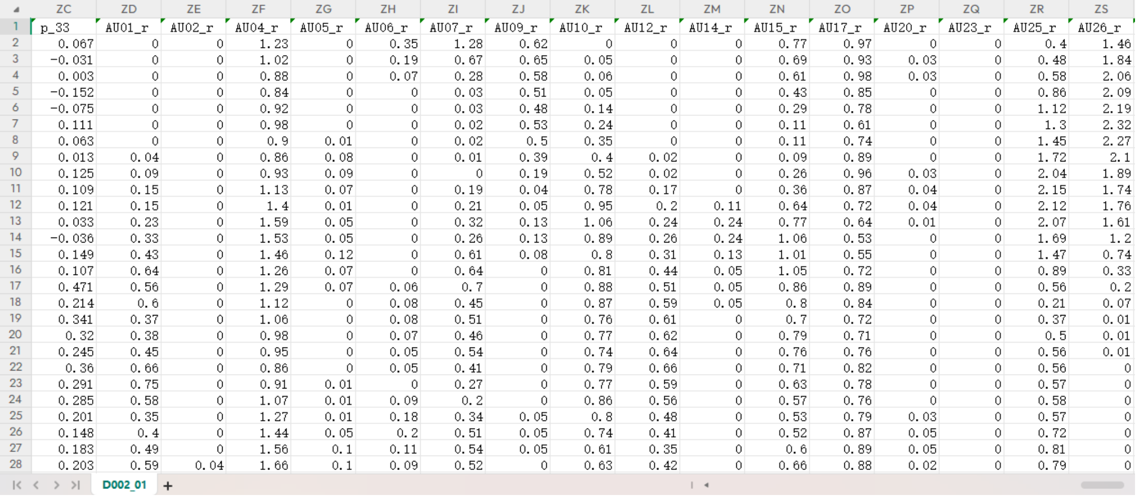


Figure 5: Audio Feature Extraction Unit (AFEU)

Feature extraction of audio spectrograms using the AlexNet network to mine deep information in the frequency domain; The LSTM network was used to model the MFCC features and capture their temporal information. Extract prosodic features using the Librosa tool; Finally, the above features are fused across modal Transformers, and the fused features are feature-level fused with the first two original features to obtain the final audio sentiment feature representation.

3.3 Visual Modal Feature Extraction Based on OpenFace

For the visual modality, this paper uses the OpenFace2.2.0 toolkit to extract information such as facial action units, heads, gaze directions, etc., to obtain the facial features of the video. The core idea is to perform facial detection on each frame of the video, mark the facial area, and identify 68 facial marker points. The 2D and 3D coordinates of each marker are 136 dimensions (68 points * 2D) and 204 dimensions (68 points * 3D), respectively. Extract 17 AUs intensities (from 0 to 5) and 18 AUs existences (0 for absence, 1 for presence) for a total of 35 dimensions from the facial action unit. In the head orientation estimation, the three rotation angles of the head (Pitch, Yaw, Roll) and the head translation vector (x,y,z) were estimated for a total of 6 dimensions. Finally, the gaze direction includes a total of 2 dimensions of horizontal and vertical angles, and the gaze Angle data for the left and right eyes is usually 6 dimensions. Ultimately, multiple facial features are obtained, with a total of 389 dimensions. Figure 6 shows the facial features extracted by Open Face.



	ZC	ZD	ZE	ZF	ZG	ZH	ZI	ZJ	ZK	ZL	ZM	ZN	ZO	ZP	ZQ	ZR	ZS
1	p_33	AU01_r	AU02_r	AU04_r	AU05_r	AU06_r	AU07_r	AU09_r	AU10_r	AU12_r	AU14_r	AU15_r	AU17_r	AU20_r	AU23_r	AU25_r	AU26_r
2	0.067	0	0	1.23	0	0.35	1.28	0.62	0	0	0	0.77	0.97	0	0	0.4	1.46
3	-0.031	0	0	1.02	0	0.19	0.67	0.65	0.05	0	0	0.69	0.93	0.03	0	0.48	1.84
4	0.003	0	0	0.88	0	0.07	0.28	0.58	0.06	0	0	0.61	0.98	0.03	0	0.58	2.06
5	-0.152	0	0	0.84	0	0	0.03	0.51	0.05	0	0	0.43	0.85	0	0	0.86	2.09
6	-0.075	0	0	0.92	0	0	0.03	0.48	0.14	0	0	0.29	0.78	0	0	1.12	2.19
7	0.111	0	0	0.98	0	0	0.02	0.53	0.24	0	0	0.11	0.61	0	0	1.3	2.32
8	0.063	0	0	0.9	0.01	0	0.02	0.5	0.35	0	0	0.11	0.74	0	0	1.45	2.27
9	0.013	0.04	0	0.86	0.08	0	0.01	0.39	0.4	0.02	0	0.09	0.89	0	0	1.72	2.1
10	0.125	0.09	0	0.93	0.09	0	0	0.19	0.52	0.02	0	0.26	0.96	0.03	0	2.04	1.89
11	0.109	0.15	0	1.13	0.07	0	0.19	0.04	0.78	0.17	0	0.36	0.87	0.04	0	2.15	1.74
12	0.121	0.15	0	1.4	0.01	0	0.21	0.05	0.95	0.2	0.11	0.64	0.72	0.04	0	2.12	1.76
13	0.033	0.23	0	1.59	0.05	0	0.32	0.13	1.06	0.24	0.24	0.77	0.64	0.01	0	2.07	1.61
14	-0.036	0.33	0	1.53	0.05	0	0.26	0.13	0.89	0.26	0.24	1.06	0.53	0	0	1.69	1.2
15	0.149	0.43	0	1.46	0.12	0	0.61	0.08	0.8	0.31	0.13	1.01	0.55	0	0	1.47	0.74
16	0.107	0.64	0	1.26	0.07	0	0.64	0	0.81	0.44	0.05	1.05	0.72	0	0	0.89	0.33
17	0.471	0.56	0	1.29	0.07	0.06	0.7	0	0.88	0.51	0.05	0.86	0.89	0	0	0.56	0.2
18	0.214	0.6	0	1.12	0	0.08	0.45	0	0.87	0.59	0.05	0.8	0.84	0	0	0.21	0.07
19	0.341	0.37	0	1.06	0	0.08	0.51	0	0.76	0.61	0	0.7	0.72	0	0	0.37	0.01
20	0.32	0.38	0	0.98	0	0.07	0.46	0	0.77	0.62	0	0.79	0.71	0	0	0.5	0.01
21	0.245	0.45	0	0.95	0	0.05	0.54	0	0.74	0.64	0	0.76	0.76	0	0	0.56	0.01
22	0.36	0.66	0	0.86	0	0.05	0.41	0	0.79	0.66	0	0.71	0.82	0	0	0.56	0
23	0.291	0.75	0	0.91	0.01	0	0.27	0	0.77	0.59	0	0.63	0.78	0	0	0.57	0
24	0.285	0.58	0	1.07	0.01	0.09	0.2	0	0.86	0.56	0	0.57	0.76	0	0	0.58	0
25	0.201	0.35	0	1.27	0.01	0.18	0.34	0.05	0.8	0.48	0	0.53	0.79	0.03	0	0.57	0
26	0.148	0.4	0	1.44	0.05	0.2	0.51	0.05	0.74	0.41	0	0.52	0.87	0.05	0	0.72	0
27	0.183	0.49	0	1.56	0.1	0.11	0.54	0.05	0.61	0.35	0	0.6	0.89	0.05	0	0.81	0
28	0.203	0.59	0.04	1.66	0.1	0.09	0.52	0	0.63	0.42	0	0.66	0.88	0.02	0	0.79	0

Figure 6: Teacher facial feature data

Suppose there are N videos in total, and each video contains n segments, then the i-th video can be represented as. $V_i=\{V_{i1},\ V_{i2},\ \cdots V_{in}\}$ The text, audio, and video of the JTH segment in the i-th video are passed into their respective feature extraction modules to obtain the corresponding text feature representation, speech feature representation, and video feature representation. $X_{ij}^{T}X_{ij}^{A}X_{ij}^{V}$ As shown in Formula (1).

$$X_i^m=\left[X_{i1}^m,X_{i2}^m,\cdots,X_{in}^m\right]\in R^{L_i\times d_i^m} \tag{1}$$

Here, T represents the text mode, A represents the audio mode, and V represents the video mode; m∈ $\{T,A,V\}L_i$ Represents the number of segments in the i-th video, d_i^m and represents the feature dimensions of each mode in the i-th video.

4. MULTI-HEAD SELF-ATTENTION MECHANISM

After extracting the low-level features of the three modalities by their respective methods, the low-level features of each modality are modeled using the multi-head Self-attention mechanism (MSAM) to capture context-related information and obtain rich high-level feature information. In the case of the text modality, the text feature representation of the i-th video is input into the multi-head X_i^T Transformer to learn the internal representation of the modality, as shown in equations (2) and (3).

$$Q_T=X_i^TW_Q,\ K_T=X_i^TW_K,\ V_T=X_i^TW_V \tag{2}$$

$$\mathrm{Attention}\left(Q_T,K_T,V_T\right)=\mathrm{Softmax}\left(\frac{Q_TK_T^T}{\sqrt{d_k}}\right)V_T \tag{3}$$

Where is the $\sqrt{d_k}d_k$ dimension scaling factor, which is used to prevent the gradient from vanishing or exploding due to an overly large dot product value, and the corresponding linear transformation weight matrix is the dimension size. $W_Q \in R^{d_i^T \times d_k}$, $W_K \in R^{d_i^T \times d_k}$, $W_V \in R^{d_i^T \times d_v} X_i^T d_i^T = d_k = d_v$

$$\mathrm{head}_h = \mathrm{Attention}\left(Q_T W_h^Q, K_T W_h^K, V_T W_h^V\right) \tag{4}$$

$$\mathrm{MultiHead}\left(Q_T, K_T, V_T\right) = \mathrm{Concat}\left(\mathrm{head}_1, \mathrm{head}_2, \cdots, \mathrm{head}_h\right) W_h^O \tag{5}$$

Where h is the number of attention heads, is the linear transformation weight matrix of the text features of the multi-head self-attention, is the dimension of each head. $W_h^O \in R^{hd_v \times d_i^T} d_k^h = d_v^h = {d_i^T}/{h}$

After the low-level features enter the multi-head Transformer, the internal relation vector representation of the text mode is obtained through residual connection and layer normalization operations, and then the extracted relation vector is subjected to nonlinear transformation through the feedforward Neural Network (FNN) composed of linear layers. To enhance the feature representation ability. Finally, the final high-level text feature representation is obtained again through residual concatenation and layer normalization $\widetilde{X_i^T}$. $\widetilde{X_i^A}, \widetilde{X_i^V}$ In the same way, high-level audio and video features can be obtained.

5. CROSS-TEXT MODAL FUSION BASED ON THE ATTENTION MECHANISM

How to effectively fuse multimodal features has always been a challenge in the field of multimodal sentiment analysis. With the emergence of attention mechanisms, some researchers have significantly improved the performance of their models by introducing them. The Transformer is implemented based on the attention mechanism, which can capture global context relationships and strengthen the internal structure of the modality, enhancing the ability to aggregate information. Studies have shown that the sentiment classification of text modalities often outperforms that of audio and video modalities [13,14]. To reduce the complexity of the model, in the cross-modal interaction part, this paper uses text modality to assist audio and video modality for modeling, and enhances the fusion effect through improved cross-modal attention.

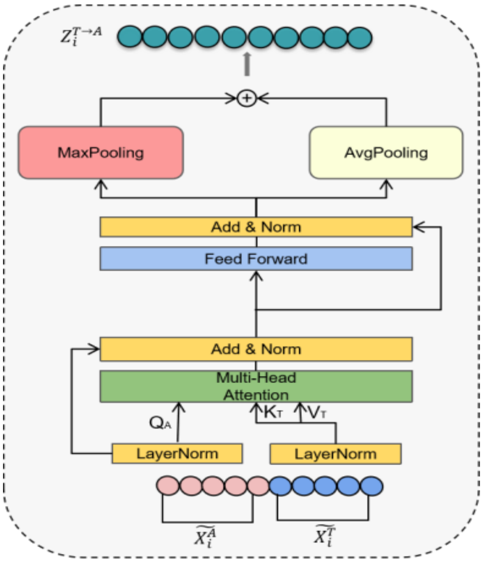


Figure 7: An improved structure diagram based on the cross-modal Transformer

Figure 7 shows the improved Cross-modal Transformer (CMT) structure. After extracting rich high-level features from low-level features through the multi-head self-attention mechanism, the high-level audio features and high-level text features are layer-normalized to provide matrices with high-level text features as auxiliary modalities and Q query vectors with high-level audio modalities as primary modalities for feature sharing through the multi-head self-attention mechanism. $X_i^T, X_i^A, X_i^V \widetilde{X_i^A} \widetilde{X_i^T} K_T, V_T$ The calculation process of the cross-modal Transformer is as shown in formula (6):

$$\mathrm{CMT}_{T\to A} = \mathrm{Softmax}\left(\frac{Q_A K_T^T}{\sqrt{d_k}}\right) V_T = \mathrm{Softmax}\left(\frac{\widetilde{X_i^A} W_{QA} W_{KT}^T \widetilde{X_i^{TT}}}{\sqrt{d_k}}\right) \widetilde{X_i^T} W_{VT} \tag{6}$$

Among them $W_{QA}\in R^{d_i^A\times d_k}$, they are respectively the linear transformation weight matrices. $W_{KT}\in R^{d_i^T\times d_k}W_{VT}\in R^{d_i^T\times d_v}$

After passing through the cross-modal Transformer, consistent with the multi-head self-attention mechanism, text-audio fusion features are obtained using residual connections and layer normalization operations, and then through the feedforward neural network, the model learns the interaction information of the two modalities. $X_i^{T\to A}$ Finally, the output results of high-level text-audio features of cross-text modal fusion are obtained.

Because pooling operations have the advantages of suppressing noise, reducing redundant information, and preventing overfitting, combinatorial pooling is used at the cross-modal output end. $X_i^{T\to A}$ The high-level text-audio features obtained through text-assisted modal fusion are input into the maximum and average pooling layers. The maximum pooling captures local features, and the average pooling captures global features. Finally, the two pooling results are concatenated by dimension to obtain the final high-level text-audio feature output result of cross-text modal fusion, $Z_i^{T\to A}$ as follows:

$$X_{i\_max}^{T\to A}=\mathrm{MaxPooling}\left(X_i^{T\to A}\right) \tag{7}$$

$$X_{i\_avg}^{T\to A}=\mathrm{AvgPooling}\left(X_i^{T\to A}\right) \tag{8}$$

$$Z_i^{T\to A}=\mathrm{Concat}\left(X_{i\_max}^{T\to A},\ X_{i\_avg}^{T\to A}\right) \tag{9}$$

In the same way, a representation of the high-level text-video feature vectors after fusing the text feature information can be obtained. $Z_i^{T\to V}$

As shown in Equations (10) and (11), the input high-level audio and video features are concatenated with their corresponding cross-text modal fusion features to obtain the final output $\widetilde{Z_i^{TA}}$,$\widetilde{Z_i^{TV}}$. Enhance the retention and transmission of single-modal internal information and achieve full fusion of single-modal and cross-modal information.

$$\widetilde{Z_i^{TA}}=\mathrm{Concat}\left(\widetilde{X_i^{A}},\ Z_i^{T\to A}\right) \tag{10}$$

$$\widetilde{Z_i^{TV}}=\mathrm{Concat}\left(\widetilde{X_i^{V}},\ Z_i^{T\to V}\right) \tag{11}$$

In sentiment analysis, audio and video features play an important role, and the fusion process is often affected by data quality, environmental noise, and the imbalance of information between modalities, making it difficult for the model to extract useful information. In order to fuse Audio and video features more effectively, this paper introduces the Audio and video feature sense module (AVFSM) aimed at alleviating the imbalance of modal information and enhancing the expressive power and fusion effect of audio and video features.

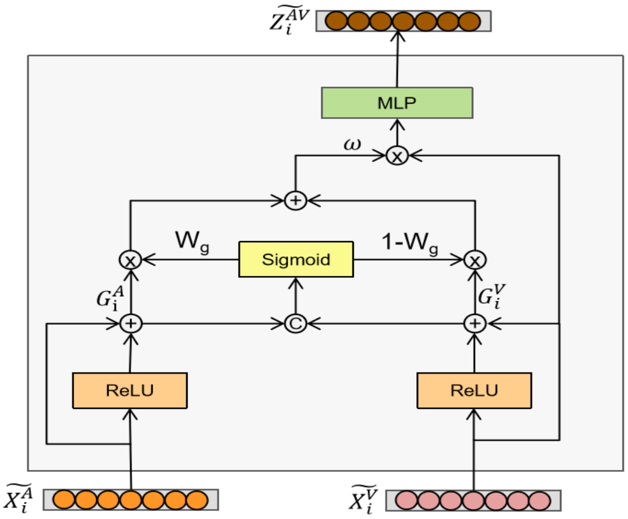


Figure 8: Audio and Video Feature Enhancement Module (AVFSM)

As shown in Figure 8, the working process is described as follows: High-level audio and video features extracted by multi-head self-attention are encoded into modal internal features respectively through ReLU activation functions to optimize single-modal features and improve the representational ability of audio and video features. To extract local and global information and enhance low-quality modal information, residual concatenation is performed to obtain intermediate features and $G_i^A G_i^V$. Information compensation is then achieved Sigmoidby adaptively adjusting the proportion of audio and video features. Finally $\widetilde{Z_i^{AV}}$, A multi-layer perceptron (MLP) is used to obtain fused audio and video A-V features to enhance audio and video synergy. The formula for the audio-video feature fusion mechanism is as follows:

$$G_i^A=\mathrm{ReLU}\left(\widetilde{X_i^A}\right)+\widetilde{X_i^A} \tag{12}$$

$$G_i^V=\mathrm{ReLU}\left(\widetilde{X_i^V}\right)+\widetilde{X_i^V} \tag{13}$$

$$W_g=\mathrm{Sigmoid}\left(\mathrm{Concat}\left(G_i^A,G_i^V\right)\right) \tag{14}$$

$$\omega=G_i^A * W_g+G_i^V*(1-W_g) \tag{15}$$

$$\widetilde{Z_i^{AV}}=\mathrm{MLP}\left(\omega * \widetilde{X_i^V}\right) \tag{16}$$

In the formula: is the high-level audio feature, is the high-level video feature, is the output feature after fusion. $\widetilde{X_i^A} \epsilon R^{L_i \times d_i^A}$ $\widetilde{X_i^V} \epsilon R^{L_i \times d_i^V}$ $\widetilde{Z_i^{AV}} \epsilon R^{L_i \times d_i^{AV}}$

Finally, the three fused modal features are concatenated together as the final multimodal feature representation, as shown in formula (17). $\overleftarrow{Z_i}$

$$\overleftarrow{Z_i}=\mathrm{Concat}\left(\widetilde{Z_i^{TA}},\widetilde{Z_i^{TV}},\widetilde{Z_i^{AV}}\right) \tag{17}$$

6. EXPERIMENTS AND RESULTS ANALYSIS

This section verifies the effectiveness of the model proposed in this paper by conducting comparative experiments on the CH-SIMS dataset. The experimental environment configuration, hyperparameter Settings, evaluation metrics, baseline model, etc. are included, and the results of the comparison experiments and ablation experiments are analyzed in detail.

6.1 Evaluation Indicators

For performance evaluation of multimodal sentiment analysis models, the evaluation metrics are set based on different types of output data of the model. It mainly consists of two tasks: classification and regression. For the classification task, the model’s output is the sentiment category. Therefore, 2-classification (negative/non-negative) accuracy (Acc-2), 3-classification (positive, negative, neutral) accuracy (Acc-3), and 2-classification weighted F1 score are used as metrics to evaluate the model’ s classification performance. The formula for accuracy is as follows:

$$\mathrm{Acc}=\frac{TP+TN}{TP+TN+FP+FN} \tag{21}$$

F1 Score is also a type of classification metric, with values ranging from [0,1]. The larger the F1 value, the stronger the classification ability of the model. The specific calculation formula is as follows:

$$F1=\frac{2\cdot TP}{2\cdot TP+FN+FP} \tag{22}$$

For regression tasks, the model ‘s output is a numerical representation of sentiment polarity, so mean absolute error (MAE) and Pearson correlation coefficient (Corr) are used as evaluation metrics. The smaller the MAE value, the more accurate the model’ s prediction. The specific formula is as follows:

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{23}$$

The Pearson Correlation Coefficient (Corr) [70] is typically used as a statistical indicator to evaluate the correlation between predicted values and true values. Its value range is [-1,1]. A value close to +1 indicates a strong positive correlation, and a value close to -1 indicates a strong negative correlation. A close to 0 indicates no significant linear relationship. The specific formula is as follows:

$$\mathrm{Corr}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{24}$$

Here, and are the values of two variables in the data sample, which are respectively the mean values of the variables x and y: $x_iy_i\bar{x},\bar{y}$

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\,,\ \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{25}$$

6.2 Experimental Setup and Dataset

The experiment was run on an 11-core PC server with Ubuntu 20.04 as the system version, NVIDIA-GeForce-RTX-4090 as the GPU, and the software configuration included deep learning frameworks PyTorch, Python, CUDA, etc. And appropriate hyperparameters were set for different modules.

To verify the validity of the model, experiments were conducted using the CH-SIMS public dataset and the self-made Tea dataset, the former being a Chinese multimodal sentiment dataset created by Tsinghua University, for details see Table 2.3; The latter is a self-made multimodal emotion dataset for young children in real-world scenarios. The partitioning for the CH-SIMS and homemade Tea datasets is shown in Table 1.

Table 1: Partitioning of Experimental Datasets

Categories	CH-SIMS			Self-built -Tea dataset		
Total	Training Set	Validation set	Test Set	Training Set	Validation set	Test Set
	1368	456	457	909	113	115
Positive emotion	419	139	140	228	29	30
Neutral emotions	207	69	69	564	58	59
Negative emotions	742	248	248	117	26	26

6.3 Comparing methods and Results

The experiment selected classic multimodal sentiment analysis models such as TFN, LMFMFM, MulT, MTFN, MLMF, and Self-MM, as well as cross-modal Transformer fusion models as benchmarks to verify the performance of the proposed models.

6.3.1 Comparative experiments on public datasets

As shown in Table 2, the experimental results of the model on the CH-SIMS dataset show that, compared with other model methods, the model proposed in this paper has significant improvements in F1 scores, Acc-2, and Acc-3. Compared with the tensor fusion network (TFN) model, the model in this paper improved the Acc-2 and F1 scores by 3.98 and 3.52 percentage points respectively. TFN has to perform tensor outer product operations when fusing different modalities, with many model parameters, a large amount of computation and a long training time. Therefore, in this paper, combinatorial pooling is used for the final cross-modal output to obtain high-quality fusion features. Compared with the Transformer-based fusion model MulT, the model in this paper also achieved better results. The audio feature extraction unit (AFEU) proposed in this paper, compared with the single MFCC feature, can mine richer and more detailed feature representations of audio features. In addition, By proposing the Audio and Video Feature Enhancement module (AVFSM), important audio and video fusion features can be selected based on weights during the fusion stage to improve the quality of audio and video features.

Table 2: Comparison results on the CH-SIMS dataset

Model	Acc-2 ↑	Acc-3 ↑	F1 ↑	MAE ↓	Corr ↑
TFN)	78.38	65.12	78.62	0.432	0.591
LMF)	77.77	64.68	77.88	0.441	0.576
MFN	77.90	65.73	77.88	0.435	0.582
MulT	78.56	64.77	79.66	0.453	0.564
MTFN	81.09	68.80	81.01	0.395	0.666
MLMF	79.34	68.36	79.07	0.409	0.639
Self-MM	80.04	65.47	80.44	0.425	0.595
Ours	**82.36**	**69.77**	**82.14**	**0.401**	**0.692**

6.3.2 Ablation experiment

To verify the effectiveness of the components presented in this paper, ablation experiments were conducted on the model on the CH-SIMS dataset, and the results are shown in Table 3. The impact of each part on the overall performance of the model is mainly explored in two aspects: one is the selection of different modes; The second is to analyze the extent to which each module contributes. The details are as follows:

1) V: Retain the complete model architecture, use text and audio as the primary modalities when fusing across modalities, and provide the Q matrix; Video is an auxiliary mode, providing K and V matrices.

2) A: Retain the complete model architecture and use text and video as the primary modality and audio as the secondary modality when fusing across modalities.

3) r/a AFEU: Remove the audio feature extraction module from the full model architecture and use the traditional single MEL frequency cepstral coefficients (MFCC) feature as the audio feature input.

4) r/a MSAM: Remove the multi-head self-attention mechanism from the full model architecture and directly use the low-level features extracted by each module as cross-modal input.

5) r/a CMT: Remove the improved cross-modal Transformer from the complete model architecture and directly concatenate the modal features after passing through the multi-head self-attention machine and send them to the soft attention mechanism layer.

6) r/a MP_AP: Remove the combinatorial pooling layer from the cross-modal Transformer module.

7) r/a AVFSM: Remove the audio and video feature enhancement module on the full model architecture.

8) r/a Soft-Attention: Remove Soft attention from the full model architecture and feed vectors concatenated through multi-head self-attention and cross-modal Transformer directly into the fully connected layer.

Table 3: Ablation Experiment Results on CH-SIMS

Item	Method	Acc-2 ↑	Acc-3 ↑	F1 ↑	MAE ↓	Corr ↑
1	V	80.46	67.22	79.84	0.437	0.574
2	A	80.90	67.47	79.91	0.435	0.588
3	r/a AFEU	80.71	66.98	80.12	0.429	0.655
4	r/a MSAM	81.45	68.60	80.79	0.415	0.670
5	r/a CMT	79.18	65.96	79.96	0.446	0.565
6	r/a MP_AP	81.99	68.76	81.07	0.424	0.611
7	r/a AVFSM	80.57	67.08	79.88	0.445	0.590
8	r/a Soft-attention	81.35	68.03	80.56	0.411	0.667
9	**Ours**	**82.36**	**69.77**	**82.14**	**0.401**	**0.692**

As can be seen from the experimental results in Table 3, all indicators decreased significantly after the components were removed. Analysis of Experiments 1, 2 and 9 shows that the model performs best when the text mode is used as the auxiliary mode and audio and video as the main mode.

Analysis of Experiments 3 and 9 revealed that after the removal of the audio feature extraction unit, the Acc-2 and Acc-3 of the model decreased to 1.65% and 2.79% respectively. The introduction of the audio feature extraction unit can enhance the model's ability to extract acoustic features.

Analysis of Experiments 5, 6, and 9 shows that after removing the cross-modal Transformer, the performance degradation of the model is the greatest, which is 3.18%, 3.81%, 2.18%, -0.045%, and 0.127% respectively in table order. This indicates that the cross-modal Transformer is crucial to the model.

Analysis of Experiments 7, 8, and 9 shows that the performance of the model declined after removing the Audio and Video Feature Enhancement module (AVFSM) and the soft attention layer, indicating that these two modules also contributed to the improvement of the model's performance.

The complete model achieved optimal performance on all metrics, verifying the effectiveness of the components in the model proposed in this paper. The modules complement each other and jointly enhance the multimodal sentiment analysis ability of the model.

6.3.3 Comparative experiments on self-built datasets

To verify the effectiveness of the self-built dataset and to evaluate the robustness and generalization ability of the model proposed in this paper in the classroom scenario, comparative experiments were conducted using the self-built dataset on five classic baseline models. The results are shown in Table 4.

Table 4: Comparative experiments on the self-built Tea dataset with the benchmark model

Model	Acc-2 ↑	Acc-3 ↑	F1 ↑	MAE ↓	Corr ↑
TFN	72.45	56.14	71.58	0.593	0.604
LMF	73.38	57.27	72.04	0.589	0.599
MFM	73.32	57.65	72.55	0.577	0.584
MulT	72.88	57.43	71.90	0.592	0.581
Self-MM	74.06	58.21	73.12	0.564	0.622
Ours	**76.40**	**60.66**	**76.51**	**0.488**	**0.676**

As shown in the table above, comparative experiments were conducted on the self-built dataset in this paper, and the results were consistent with those on the CH-SIMS model. Compared with several other major baseline models, the model in this paper achieved good performance in all metrics on the self-built dataset.

7. CONCLUSION

To sum up, the cross-modal Transformer architecture with multi-head self-attention mechanism has good performance in multimodal fusion, resulting in more accurate recognition results. At the same time, the recognition results are more comprehensive and reliable, showing relatively excellent robustness and generalization ability, and can be applied to the recognition of children's emotions in the actual preschool education process. Further enhance the teaching effect. But there are certain limitations, such as the multiple physiological and psychological manifestations of emotional expression, so the next step of research will expand the modal types of multimodal emotion recognition to further enhance the comprehensiveness of recognition results.

ACKNOWLEDGEMENT

This paper is support by the Zhejiang Province College Students' Science and Technology Innovation Activity Plan (New Talent Plan) (Number: 2024R430A005).

REFERENCES

[1] Peng Kaibei, Sun Xiaoming, Chen Haowei, et al. Voice Emotion recognition method for railway stations based on convolutional neural networks [J]. Computer Simulation, 2023, 40(2): 177-180+189.

[2] Gao Lijun, Xue Lei. Speech Emotion Recognition Based on Transformer architecture [J]. Industrial Control Computers, 2023, 36(1): 82-83+86.

[3] Wang Xi, Wang Junbao, Bianba Wangdui. Emotion Recognition of Tibetan Speech Based on Convolutional neural Network [J]. Information Technology and Informationization, 2022, (11): 202-206.

[4] Cui Chenlu, Cui Lin. Lightweight Speech Emotion Recognition for Data Augmentation [J]. Computer and Modernization, 2023, (4): 83-89+100.

[5] Zhu Yonghua, Feng Tianyu, Zhang Meixian, et al. Convolutional speech emotion recognition network based on incremental method [J]. Journal of Shanghai University (Natural Science Edition),2023,29(1):24-40.

[6] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. 2018:4171-4186

[7] Baltrusaitis T, Zadeh A, Lim Y C, et al. OpenFace 2.0: Facial Behavior Analysis Toolkit [C]. Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, 2018: 59-66

[8] Lu Xueqiang, Tian Chi, Zhang Le, et al. Multimodal sentiment analysis model fusing multi-feature and attention mechanism [J]. Data Analysis and Knowledge Discovery, 24,8(05):91-101.

[9] Kim K, Park S. AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis [J]. Information Fusion, 2023, 92: 37 to 45.

[10] Junxi, Y., Wang, Z., & Chen, C. (2024). GCN-MF: A graph convolutional network based on matrix factorization for recommendation. Innovation & Technology Advances, 2024, 2(1), 14–26. https://doi.org/10.61187/ita.v2i1.30

[11] Wang L, Peng J, Zheng C, et al. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning [J]. Information Processing & Management, 2024, 61(3): 103675.

[12] Fu Y, Zhang Z, Yang R, et al. Hybrid cross-modal interaction learning for multimodal sentiment analysis [J]. Neurocomputing, 2024, 571: 127201.

[13] Zeng Y, Mai S, Hu H F. Which is Making the Contribution: Modulating Unimodal and Cross-Modal Dynamics for Multimodal Sentiment Analysis[C]. Findings of the Association for Computational Linguistics: Emnlp, 2021: 1262-1274.

[14] Wu Y, Lin Z, Zhao Y, et al. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis[C]. Findings of the association for computational linguistics: Acl-Ijcnlp ,2021: 4730-4738.