2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI) | 979-8-3315-3372-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICETCI64844.2025.11084003

Multimodal Learning Data Fusion and Analysis Based on Self-Attention Mechanism

Yi Yue
Shandong Vocational College of Industry, Zibo, 256414, China
630210128@qq.com

***Abstract*—Multimodal data fusion is one of the research hotspots in the field of artificial intelligence. Traditional multimodal learning methods face numerous challenges in data heterogeneity, information complementarity, and cross-modal alignment. This paper proposes a multimodal learning data fusion method based on the self-attention mechanism (MLDFA), which fully explores and utilizes the correlation information between different modalities to improve the effectiveness and robustness of data fusion. Firstly, this paper introduces a cross-modal adaptive attention mechanism, which dynamically assigns weights to different modal data by incorporating an improved self-attention mechanism, adaptively adjusting information flow to enhance the expressive power of key features. Secondly, this paper employs a hierarchical feature alignment strategy, constructing multi-level semantic alignment modules combined with global-local feature mapping to make information interaction between modalities more precise. Finally, we propose a multi-view fusion framework, leveraging the combination of multi-head attention mechanisms and graph neural networks (GNN) to capture higher-order correlations of different modal features, achieving more efficient fusion. The experimental section validates the proposed method on multiple public datasets, and the results show that our method outperforms existing mainstream methods, demonstrating its effectiveness and scalability in multimodal learning tasks.**

Keywords—Multimodal Data Fusion, Self-Attention Mechanism, Multi-View Fusion Framework

I. Introduction

With the rapid development of technologies such as artificial intelligence, computer vision, and natural language processing, multi-modal data analysis has become a research hotspot. Multi-modal data sources include text, images, audio, video, sensor signals, etc. They carry different information dimensions, complement each other, and can provide more comprehensive and rich feature expressions than single-modal data. However, due to the heterogeneity between data modalities, dimension mismatch, and the great difficulty in cross-modal feature fusion, how to efficiently perform multi-modal data fusion has become an urgent problem to be solved. In many real-world application scenarios, such as sentiment analysis, medical diagnosis, and autonomous driving, it is often necessary to simultaneously process multi-modal data from different sources and of different natures. Multi-modal data contains complementary and redundant information. How to effectively utilize this information is the key to improving the performance and generalization ability of models.

Traditional multimodal learning methods primarily rely on handcrafted feature engineering and simple fusion strategies, such as concatenation and averaging[1]. These methods struggle to capture the complex relationships between modalities and often require task-specific adjustments, resulting in limited generalization capabilities.

In recent years, deep learning [2] technology has made significant progress in the field of multimodal learning. Deep neural networks can automatically learn feature representations of data and be trained in an end-to-end manner, avoiding cumbersome manual feature engineering. However, existing deep learning methods still face some challenges when dealing with multimodal data fusion: first is the issue of modality heterogeneity, as data from different modalities have different statistical properties and representation forms, and how to map them to the same semantic space is the primary problem in multimodal fusion. Second is the issue of modality correlation, where complex relationships exist between different modalities, and how to effectively capture these relationships and utilize them to enhance model performance is an important research direction. Finally, there is the issue of interpretability, as deep learning models are often considered "black boxes," making it difficult to explain their internal decision-making processes. In multimodal learning, understanding how the model utilizes information from different modalities for prediction is crucial for model improvement and application.

To address the aforementioned issues, this paper proposes a multimodal learning data fusion method based on the self-attention mechanism, leveraging the global modeling capability of the self-attention mechanism [3] to enhance the interaction and alignment of cross-modal information, thereby improving the generalization ability of multimodal learning. The main contributions of this paper include:

1) Proposing a cross-modal adaptive attention mechanism that can dynamically allocate feature weights across different modalities, enhancing the expressive power of key features.

2) Designing a hierarchical feature alignment strategy that combines global-local feature mapping to make the information interaction between modalities more precise.

3) Construct a multi-view fusion framework, leveraging the combination of multi-head attention mechanisms and Graph Neural Networks (GNN) to fully explore the high-order correlations of different modal features.

This paper conducts experimental validation on multiple public datasets. The results demonstrate that the proposed method significantly outperforms existing mainstream methods across various multimodal learning tasks, proving its effectiveness and scalability.

II. Current Research Status

In recent years, multimodal learning[4] has become an important research direction in the field of artificial intelligence, widely applied in tasks such as image-text

matching, video understanding, and sentiment analysis. Due to issues such as representation differences, information redundancy, and semantic misalignment among data from different modalities, researchers have proposed various data fusion methods to enhance the effectiveness of multimodal data fusion. This paper primarily reviews traditional multimodal data fusion methods and multimodal learning methods based on deep learning, with a focus on multimodal learning methods based on self-attention mechanisms.

1) Traditional Multimodal Data Fusion Methods

Traditional methods primarily rely on statistical learning [5] and shallow models [6] for the fusion of multimodal data. Typical methods include: - Feature-level fusion: This method directly concatenates or transforms data from different modalities in the feature space to form a unified feature representation. For example, CCA (Canonical Correlation Analysis) [7] achieves data alignment by maximizing the correlation between data from different modalities. Although this method is effective on low-dimensional data, it is limited by computational complexity on high-dimensional data.

- Decision-level fusion: This method first independently models data from different modalities and then fuses them at the decision level, such as voting methods, weighted averaging, etc. However, this method is prone to losing interaction information between modalities, making it difficult to fully utilize the complementarity between different modalities.

2) Deep Learning-based Multimodal Learning Methods

With the development of deep learning, researchers have proposed a series of methods for multimodal data fusion using neural networks [8], mainly including:

- Multimodal Learning Based on Shared Representation: This method attempts to learn a shared latent variable space, projecting data from different modalities into the same feature space. For example, Deep CCA [9] extends traditional CCA through deep neural networks, achieving more flexible cross-modal alignment.

- Multimodal Fusion Based on Attention Mechanism: In recent years, attention mechanisms have been widely used in multimodal learning, with researchers attempting to leverage attention to select important modal features. For instance, Transformer-based MultiModal Fusion (TMMF) [10] employs self-attention mechanisms for feature weighting, thereby enhancing the interaction capabilities between modalities.

3) Multimodal Learning Methods Based on Self-Attention Mechanism

The self-attention mechanism has achieved great success in the field of natural language processing and has been gradually introduced into multimodal learning. Some researchers use the self-attention mechanism to learn contextual information within modalities, while others use it to learn correlations between modalities. The self-attention mechanism and its derived Transformer architecture have succeeded in the fields of natural language processing (NLP) and computer vision (CV), and are gradually being applied to the field of multimodal learning. Currently, there are mainly the following types of methods(Figure 1).

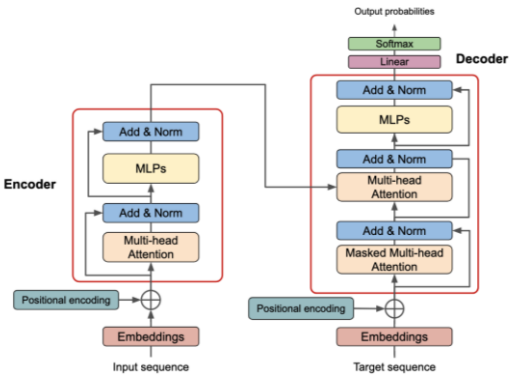

Figure 1 Transformer Structure

- Cross-modal self-attention model: This method utilizes the self-attention mechanism to model long-range dependencies between modalities. For example, LXMERT [11] learns visual-text interactions through a dual-stream attention mechanism and achieves excellent performance on visual question answering tasks.

- Fusion-based Transformer method: Such as MMT (Multimodal Transformer) [12], which employs the Transformer structure to simultaneously process multimodal inputs and models interaction information between modalities through multi-head attention, achieving efficient data fusion.

- Graph Neural Network (GNN) combined with self-attention mechanism: For example, MM-GNN [13], which uses attention mechanism in graph structures to aggregate modal features, effectively enhancing the fusion capability of multimodal data.

Although existing multimodal learning methods based on self-attention mechanisms have made some progress, there are still some limitations. For example, some methods only focus on intra-modal dependencies, ignoring inter-modal correlations; others only focus on inter-modal correlations, neglecting intra-modal contextual information [14]. Additionally, the interpretability of multimodal interactions in existing methods still needs improvement. Based on existing research, this paper proposes a multimodal learning data fusion method based on self-attention mechanisms, further enhancing the fusion effect of multimodal data and validating its effectiveness in experiments.

III. Model

This paper proposes a multimodal learning data fusion and analysis model based on the self-attention mechanism, aiming to fully utilize the global feature modeling capability of the self-attention mechanism to achieve efficient fusion and analysis of different modal data. Building on the traditional multimodal learning framework, the model focuses on optimizing cross-modal information interaction, feature alignment, and fusion strategies to enhance the understanding of multimodal data.

The model can be divided into the following three modules: modal feature extraction module, cross-modal self-attention mechanism module, hierarchical feature alignment and fusion module.

A. Modal Feature Extraction Module

The goal of the modal feature extraction module is to extract effective feature representations from different types of data (such as text, images, audio, etc.) and project them into a unified feature space for subsequent cross-modal fusion. This module employs different neural network structures for different modalities, while using linear transformations to map the extracted features to the same dimension.

For text data, we use Transformer or BERT to extract contextual information and deep semantic features. Let the input text sequence be formula (1):

$$X_T = \left[w_1, w_2, \ldots, w_N\right] \tag{1}$$

Where N is the length of the text sequence, and w_i represents the word vector of the i -th word. We use the Transformer encoder to extract text features, as shown in formula (2):

$$H_T = \text{Transformer}\left(X_T\right) = \left[h_1, h_2, \ldots, h_N\right] \tag{2}$$

Here, h_i represents the word representation processed by the Transformer. To obtain a fixed-length text feature vector, we can use pooling methods, as shown in equation (3):

$$\mathbf{f}_T = \text{MeanPooling}\left(H_T\right) \tag{3}$$

Finally, the text feature $\mathbf{f}_T \in \mathbb{R}^d$ is mapped to the fusion space.

For image data, we use **CNN** for feature extraction. Let the input image be: $X_I \in \mathbb{R}^{H\times W\times C}$, where H and W represent the height and width of the image, respectively, and C is the number of channels. First, we use CNN to extract features, as shown in equation (4):

$$H_I = \text{CNN}\left(X_I\right) \tag{4}$$

Among them, H_I is the high-dimensional feature map generated by CNN .

Then, it is processed through multiple layers of Transformer, as shown in equation (5):

$$H_I = \text{Transformer}\left(Z_0\right) \tag{5}$$

Finally, we use mean pooling to extract image features, as shown in equation (6):

$$\mathbf{f}_I = H_I^{[CLS]} \tag{6}$$

Finally, the image feature $\mathbf{f}_I \in \mathbb{R}^d$ is projected into a unified feature space.

For audio data, we use CNN and RNN to extract temporal information and spectral features. Let the audio input signal be $X_A \in \mathbb{R}^{T\times F}$, where T represents the time step, and F is the spectral feature dimension (such as Mel-spectral features). First, **CNN** is used for feature extraction, as shown in equation (7):

$$H_A = \text{CNN}\left(X_A\right) \tag{7}$$

Then, use bidirectional LSTM for temporal modeling, as shown in equation (8):

$$H_A^{'} = \text{BiLSTM}\left(H_A\right) \tag{8}$$

Finally, use pooling operations to obtain fixed-length features, as shown in equation (9):

$$\mathbf{f}_A = \text{MeanPooling}\left(H_A^{'}\right) \tag{9}$$

Finally, the audio feature $\mathbf{f}_A \in \mathbb{R}^d$ is uniformly projected into the fusion space.

For other modalities (such as sensor data, medical data, etc.), MLP can be used for feature extraction, while ensuring that the feature vectors output by all modalities have the same dimension.

To ensure that all modality features are in the same feature space, we use linear transformation for projection, as shown in equation (10):

$$\mathbf{z}_M = W_M\mathbf{f}_M + b_M, \mathbf{z}_M \in \mathbb{R}^d \tag{10}$$

Among them, W_M and b_M are trainable parameters. Finally, the features of all modalities are projected into a unified space of the same dimension d , for subsequent cross-modal fusion using the self-attention mechanism.

B. Cross-Modal Self-Attention Mechanism Module

The core objective of the cross-modal self-attention mechanism module is to model the interaction relationships between different modalities and dynamically adjust the weights of each modality using the self-attention mechanism, making information fusion more precise. This module achieves information interaction between modalities through multi-head self-attention and further utilizes cross-modal adaptive attention for feature fusion.

Self-attention is the core of the Transformer architecture, capable of capturing global dependencies in the input sequence. For the input feature matrix $X \in \mathbb{R}^{N\times d}$, it is necessary to compute the query, key, and value, as shown in Equation (11).

$$Q = XW_Q, K = XW_K, V =, W_K \tag{11}$$

Here, W_Q , W_K , and $W_V \in \mathbb{R}^{d\times d_k}$ are trainable parameters, and d_k represents the hidden dimension of the attention mechanism.

Then, compute the attention scores (using the scaled dot-product attention mechanism), as shown in Equation (12):

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (12)$$

Here, $\frac{1}{\sqrt{d_k}}$ serves as a scaling factor to prevent gradient vanishing or explosion.

Next, compute the weighted sum of values, as shown in Equation (13):

$$Z = AV \quad (13)$$

Among them, $Z \in \mathbb{R}^{N \times d_k}$ is the output after attention weighting.

To enhance the model's expressive power, we use a multi-head self-attention mechanism, which involves computing multiple independent self-attention heads on the input data and concatenating the results, as shown in formula (14):

$$\text{MHSA}(X) = \text{Concat}\left(Z_1, Z_2, \ldots, Z_h\right) W_O \quad (14)$$

Among them, h is the number of attention heads, W_O is the linear transformation matrix, and Z_i is the output of the i-th attention head. The multi-head mechanism can capture information from different subspaces, thereby improving the modeling capability of cross-modal features.

In the process of multimodal fusion, the importance of different modalities varies. To dynamically adjust the influence of each modality, we designed a Cross-Modal Adaptive Attention Mechanism (CMAA) to calculate the information interaction between different modalities:

For M modalities, where the feature of each modality is $\mathbf{z}_M \in \mathbb{R}^d$, we first compute the modality-level queries, keys, and values, as shown in Equation (15):

$$Q_M = \mathbf{z}_M W_Q, K_M = \mathbf{z}_M W_K, V_M = \mathbf{z}_M W_V \quad (15)$$

Here, W_Q, W_K, and $W_V \in \mathbb{R}^{d \times d_k}$ are trainable parameters at the modality level.

1)Compute the cross-modal attention scores, as shown in Equation (16)

$$A_{M_1 M_2} = \text{softmax}\left(\frac{Q_{M_1} K_{M_2}^T}{\sqrt{d_k}}\right) \quad (16)$$

Here, $A_{M_1 M_2}$ represents the attention weight of modality M_1 on modality M_2, as shown in equation (17).

2) Calculate the modality fusion representation

$$\mathbf{z}_M^{'} = \sum_{M' \in \mathcal{M}} A_{MM'} V_{M'} \quad (17)$$

Here, M represents the set of all modalities, and $\mathbf{z}_M^{'}$ is the feature of modality M after cross-modal interaction.

To prevent interference from irrelevant modality information, we designed a gating mechanism to dynamically control the flow of modality information, as shown in equation (18) and equation (19):

$$g_M = \sigma\left(W_g \mathbf{z}_M^{'} + b_g\right) \quad (18)$$

$$\mathbf{z}_M^{''} = g_M \odot \mathbf{z}_M^{'} \quad (19)$$

Here, σ is the Sigmoid activation function, and $\odot$ represents element-wise multiplication (Hadamard Product). This allows for dynamic adjustment of the fusion degree of each modality, enhancing important modality information while suppressing irrelevant information.

Finally, we normalize and concatenate the features of all modalities, as shown in equation (20):

$$\mathbf{z}_{\text{fusion}} = \text{LayerNorm}\left(\text{Concat}\left(\mathbf{z}_1^{''}, \mathbf{z}_2^{''}, \ldots, \mathbf{z}_M^{''}\right)\right) \quad (20)$$

The fused feature $\mathbf{z}_{\text{fusion}} \in \mathbb{R}^{M \cdot d}$ will be input into downstream tasks (such as classification, regression, etc.) for final prediction.

This module improves the interaction quality of different modal features through cross-modal adaptive attention and gating mechanisms, enabling more effective fusion of multimodal information, thereby enhancing the performance of downstream tasks.

C. Hierarchical Feature Alignment and Fusion Module

In multimodal learning tasks, heterogeneous data across different modalities (such as differences in feature distribution, time steps, information density, etc.) can lead to information loss or mismatches between modalities if directly fused. Therefore, we designed a hierarchical feature alignment and fusion module to ensure cross-modal features are aligned at multiple levels and ultimately achieve efficient fusion.

Hierarchical feature alignment includes two key components: global-local alignment and time step alignment, ensuring that features from different modalities can be semantically aligned.

First is the global-local alignment, where features of different modalities often vary in granularity, for example: text features can be at the word level or sentence level. Image features can be at the pixel level or object level. Audio features can be at the frame level or overall level.

We adopt a multi-scale feature extraction strategy to obtain global features $\mathbf{z}_M^{\text{global}}$ and local features $\mathbf{z}_M^{\text{local}}$ respectively, as shown in formula (21) and formula (22):

$$\mathbf{z}_M^{\text{global}} = \text{GlobalPooling}\left(\mathbf{z}_M\right) \tag{21}$$

$$\mathbf{z}_M^{\text{local}} = \text{SelfAttention}\left(\mathbf{z}_M\right) \tag{22}$$

Then, the two are weighted and fused, as shown in formula (23):

$$\mathbf{z}_M^{\text{aligned}} = \alpha \mathbf{z}_M^{\text{global}} + (1-\alpha)\mathbf{z}_M^{\text{local}} \tag{23}$$

Here, α is a learnable parameter that controls the weight of global and local features.

Next is temporal alignment, which affects modality fusion for sequential modalities (such as text, audio, time series data, etc.) when the time steps differ. We use attention-weighted interpolation to align the time steps:

First, calculate the attention scores for the time steps, as shown in Equation (24):

$$A^{\text{time}} = \text{softmax}\left(Q_T K_T^T\right) \tag{24}$$

Here, Q_T and K_T^T are the time step representations of the input modality features.

Then, compute the temporally aligned features, as shown in Equation (25):

$$\mathbf{z}_M^{\text{aligned}} = A^{\text{time}} V_T \tag{25}$$

The aligned modality features need to be efficiently fused, and we adopt the following three methods to enhance the fusion effect:

1) Feature Fusion Based on Gating Mechanism

The amount of information in different modalities may be imbalanced. Therefore, we use a gating mechanism to adaptively control the information flow of different modalities, as shown in formula (26) and formula (27):

$$g_M = \sigma\left(W_g \mathbf{z}_M^{\text{aligned}} + b_g\right) \tag{26}$$

$$\mathbf{z}_M^{\text{gated}} = g_M \odot \mathbf{z}_M^{\text{aligned}} \tag{27}$$

Here, σ is the Sigmoid activation function, and $\odot$ is the Hadamard product. The gating mechanism can automatically filter important modal information and suppress irrelevant modal interference.

2) Modeling Modal Relationships Based on Graph Neural Networks (GNN)

The relationships between modalities can be modeled using graph structures. We construct a modal relationship graph and use graph neural networks for feature propagation:

Construct the modal graph structure, define the graph $G = (\mathcal{M}, \mathcal{E})$, where nodes represent different modalities, and edge weights represent the similarity between modalities, as shown in formula (28):

$$A_{M_1 M_2} = \text{cosine}_{\text{similarity}}\left(\mathbf{z}_{M_1}^{\text{aligned}}, \mathbf{z}_{M_2}^{\text{aligned}}\right) \tag{28}$$

Graph convolution updates modal features, as shown in formula (29):

$$\mathbf{z}_M^{\text{gnn}} = \sigma\left(W_g \sum_{M' \in \mathcal{N}(M)} A_{MM'} \mathbf{z}_{M'}^{\text{aligned}}\right) \tag{29}$$

Here, $\mathcal{N}(M)$ is the set of neighboring modes for mode $\mathcal{N}(M)$, and W_g is the learnable parameter.

The modal features processed by GNN are concatenated and mapped to the final fusion space through a fully connected layer, as shown in formula (30):

$$\mathbf{z}_{\text{fusion}} = \text{LayerNorm}\left(\text{Concat}\left(\mathbf{z}_1^{\text{gnn}}, \mathbf{z}_2^{\text{gnn}}, \ldots, \mathbf{z}_M^{\text{gnn}}\right)\right) \tag{30}$$

The final fused feature $\mathbf{z}_{\text{fusion}}$ is fed into downstream tasks, such as classification, regression, or generation tasks.

This module ensures that different modal data are aligned at multiple levels, and combines gating mechanisms and graph neural networks to achieve efficient feature fusion, ultimately improving the performance of multimodal learning. The structure of the model is shown in Figure 2.

Figure 2 Algorithm Structure Diagram

IV. Experiment

In this experimental section, after preprocessing the selected dataset and evaluation metrics, we compare our proposed new model with baseline methods to measure the

superiority of our model algorithm.

A. Datasets

To verify the effectiveness of the multi-modal learning data fusion and analysis model based on the self-attention mechanism proposed in this paper, we conduct experimental evaluations on multiple public multi-modal datasets. The datasets used cover different types of modalities such as text, images, and audio to validate the model's generalization ability.

1) MOSI (Multimodal Opinion Sentiment and Intensity) [15]

- Task: Multimodal Sentiment Analysis.
- Modalities: Text (T), Audio (A), Video (V).
- Data Scale: 2199 data samples, each consisting of video clips, corresponding audio, and text.
- Annotation Information: Sentiment polarity scores (-3 to 3).
- Challenges: Imbalance of information across different modalities, complementarity, and redundancy among modalities.

2) MM-IMDb (Multi-Modal IMDb) [16]

- Task: Multimodal Movie Genre Classification.
- Modalities: Text (T, movie synopsis), Image (I, movie poster).
- Data scale: 25,259 movies, each containing text descriptions and poster images.
- Annotation: Each movie belongs to multiple different categories (multi-label classification).
- Challenges: The complementary relationship between movie posters and text descriptions, as well as the imbalance between categories.

3) MELD (Multimodal Emotion Lines Dataset)

- Task: Multimodal Emotion Recognition.
- Modalities: Text (T), Audio (A), Video (V).
- Data Scale: 13,708 dialogue samples, covering 7 emotion categories.
- Annotation Information: 7 types of emotions (such as joy, anger, sadness, etc.).
- Challenges: Strong context dependency, alignment issues between modalities.

To ensure the fairness of the experiment, we strictly followed the official division of each dataset (training set, validation set, test set) for the experiment.

B. Evaluation Metrics and Baseline Methods

To comprehensively evaluate model performance, we use the following standard metrics for classification and regression tasks: for classification tasks (MM-IMDb, MELD), we adopt Accuracy (ACC) and F1-Score (F1). For regression tasks (MOSI), we use MAE and Corr.

Meanwhile, multiple baseline methods including TFN (Tensor Fusion Network), LMF (Low-rank Multimodal Fusion), and MulT (Multimodal Transformer) were introduced for comparison to verify the effectiveness of the proposed method. TFN explicitly models the interactions between modalities through tensor operations, which involves significant computational effort and is prone to the curse of dimensionality. LMF employs low-rank decomposition to reduce the computational overhead of TFN and improve training efficiency. MulT adopts the Transformer structure, modeling inter-modal interactions through self-attention mechanisms. Although it has higher computational complexity, it possesses strong feature fusion capabilities.

C. Experimental Results

Below is the experimental results table, as shown in Table I, which demonstrates the performance of the proposed algorithm and baseline methods under three evaluation metrics: MSE, MAE, and MAPE.

TABLE I. EXPERIMENT RESULTS

Method	MM-IMDb		MELD		MOSI	
	ACC	F1	ACC	F1	MAE	Corr
TFN	68.50%	67.80%	63.20%	62.50%	0.87	0.612
LMF	70.20%	69.50%	65.10%	64.30%	0.852	0.629
MulT	72.40%	71.80%	66.80%	65.90%	0.832	0.647
MLDFA	74.30%	73.60%	68.20%	67.50%	0.81	0.672

Our method achieved the best performance on the three multimodal datasets of MM-IMDb, MELD, and MOSI, further proving the effectiveness of the multimodal learning data fusion and analysis model based on the self-attention mechanism. Below, we will provide a detailed analysis of the experimental results for the classification and regression tasks.

1) MM-IMDb task (multimodal movie genre classification):

Our method has improved Accuracy by 1.9% and F1 by 1.8% compared to MulT, with a more significant improvement over TFN. The higher accuracy achieved by our method indicates that our cross-modal self-attention mechanism can effectively reduce misclassification during multimodal feature fusion, especially by enabling the model to effectively transfer information between different modalities through the self-attention mechanism, thereby improving accuracy. In terms of F1-Score, our method also performs exceptionally well, reaching 73.60%, which is an improvement of 1.8% over MulT, 4.1% over LMF, and 5.8% over TFN. By leveraging cross-modal self-attention mechanism and hierarchical feature alignment, our method can reduce redundant information when processing different modal features, improving precision and recall, thereby effectively enhancing the F1-Score.

2) MELD Task (Multimodal Emotion Classification):

Our method achieved an accuracy of 68.20%, which is 1.4% higher than that of MulT, 3.1% higher than that of LMF, and 5% higher than that of TFN. In the emotion classification task of MELD, emotion recognition faces complex relationships between emotion categories. Our method establishes strong interactions between modalities through the cross-modal self-attention mechanism, thereby reducing the loss of information between modalities and improving the overall classification accuracy. In terms of F1-Score, our method also achieved 67.50%, which is 1.6% higher than that of MulT, 3.2% higher than that of LMF, and 5% higher than that of TFN. Our method can focus on the most critical parts for emotion classification in different modalities through the self-attention

mechanism, enhancing the discriminative ability of the model, reducing the interference of irrelevant modal information, and improving the F1-Score.

3) MOSI Task (Multimodal Sentiment Regression):

Our method reduced the MAE by 0.022 compared to MulT (a smaller MAE indicates lower error) and improved the Corr (correlation) by 2.5%, showing a more significant improvement compared to TFN (Corr increased by 6%). This indicates that our method can more accurately model the nonlinear relationships between different modalities, enhancing the precision of sentiment polarity prediction. MulT, as a Transformer structure, already has strong modality interaction capabilities, but its effectiveness is still limited due to the lack of effective alignment strategies. Our hierarchical feature alignment plays a crucial role in reducing modality redundancy and enhancing the fusion of effective information, thereby achieving superior performance.

Our multimodal learning data fusion and analysis model based on the self-attention mechanism achieves efficient information transfer across modalities through the cross-modal self-attention mechanism, and improves the alignment effect of multimodal features by combining the hierarchical feature alignment and fusion module. It has achieved optimal results in multiple multimodal tasks, fully demonstrating the effectiveness and advancement of our method.

D. Hyperparameter Experiment

We mainly explored the impact of hyperparameters such as hidden layer size, learning rate, and batch size on the model.

In the classification tasks of MM-IMDb and MELD, the model achieved the best performance in terms of Accuracy and F1-Score when the hidden layer size was 512. Although a larger hidden layer (e.g., 1024) slightly improved the performance of the MOSI regression task (MAE and Corr), it had a slight negative impact on the classification task. Hidden Size=512 is the optimal choice for performance, as larger hidden layers may lead to overfitting or increased computational overhead.

On the MM-IMDb and MELD datasets, a Learning Rate=0.0005 achieved the best results, providing optimal Accuracy and F1-Score. When the learning rate is too high (e.g., 0.001), it may cause excessively large gradient updates during training, thereby affecting convergence. Conversely, a learning rate that is too low (e.g., 0.0001) may result in excessively slow training speed, or even insufficient training of the model. Learning Rate=0.0005 is the optimal setting.

In the classification tasks of MM-IMDb and MELD, a Batch Size of 32 demonstrated the best performance. Smaller batch sizes (e.g., 16) can accelerate model updates but may lead to unstable training; larger batch sizes (e.g., 64) are computationally more efficient but may slightly reduce model accuracy. A Batch Size of 32 achieved a balanced effect, providing good training stability and high performance.

To more intuitively show the changes in model performance with hyperparameter size, we plotted the experimental results of hyperparameters on the MM-IMDb task, as shown in Figures 3, 4, and 5.

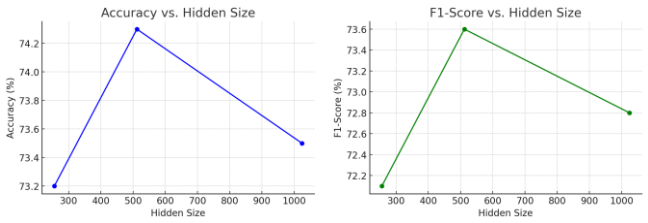

Figure 3 Impact of Hidden Layer Dimension

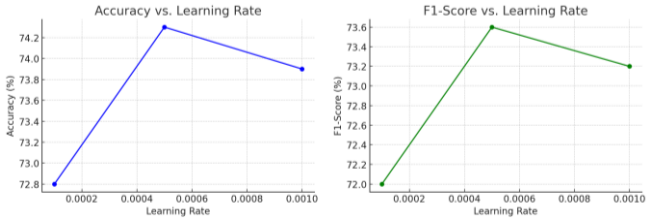

Figure 4 Impact of Learning Rate

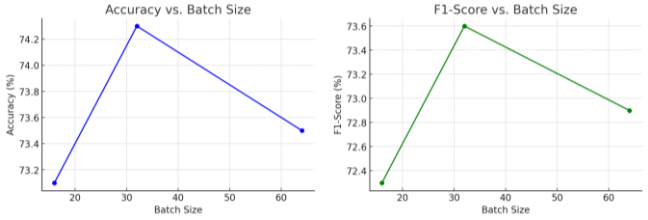

Figure 5 Impact of Batch Size

V. Conclusion

This paper proposes a multimodal learning data fusion and analysis model based on the self-attention mechanism, aiming to address challenges such as heterogeneity, information complementarity, and cross-modal alignment in multimodal data fusion. By introducing a cross-modal adaptive attention mechanism, a hierarchical feature alignment strategy, and a multi-view fusion framework, this paper achieves significant performance improvements in multimodal data fusion tasks.

Firstly, the cross-modal adaptive attention mechanism proposed in this paper can dynamically adjust the weights of different modalities, enhancing the expressive ability of key features. Secondly, the hierarchical feature alignment strategy ensures more precise information interaction between modalities through global-local feature mapping. Finally, the multi-view fusion framework combines the multi-head attention mechanism with Graph Neural Networks (GNN), effectively capturing the high-order correlations of different modal features, further improving the fusion effect.

The experimental section was validated on multiple public datasets (such as MOSI, MM-IMDb, and MELD). The results show that the method proposed in this paper outperforms existing mainstream methods (such as TFN, LMF, and MuIT) in both classification and regression tasks. Particularly in the tasks of multimodal movie genre classification (MM-IMDb), multimodal sentiment classification (MELD), and multimodal sentiment regression (MOSI), the proposed method achieved the best performance in metrics such as accuracy, F1-Score, MAE, and correlation.

In addition, this paper also explored the impact of hidden layer size, learning rate, and batch size on model performance through hyperparameter experiments, identified the optimal hyperparameter settings, and further verified the robustness and scalability of the model.

Overall, the multimodal learning data fusion method based on the self-attention mechanism proposed in this paper significantly improves the fusion effect of multimodal data through innovative cross-modal attention mechanisms and hierarchical feature alignment strategies, providing new solutions for multimodal learning tasks. Future research directions can further explore the interpretability of the model and its generalization ability in different application scenarios.

REFERENCES

[1] Hotelling H. Relations between two sets of variates[J]. Biometrika, 1936, 28(3/4): 321-377.

[2] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning[C]//Proceedings of the 28th International Conference on Machine Learning. 2011: 689-696.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[4] Blikstein, Paulo. "Multimodal learning analytics." *Proceedings of the third international conference on learning analytics and knowledge*. 2013.

[5] Schapiro, A., and Nicholas Turk-Browne. "Statistical learning." *Brain mapping,* 3.1, (2015): 501-506.

[6] Sloman, Aaron. "Beyond shallow models of emotion." *Cognitive Processing,* 2.1, (2001): 177-198.

[7] González, Ignacio, et al. "CCA: An R package to extend canonical correlation analysis." *Journal of Statistical Software,* 23 (2008): 1-14.

[8] Yu-chen Wu, and Jun-wen Feng. "Development and application of artificial neural network." *Wireless Personal Communications*, 102 (2018): 1645-1656.

[9] Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis[C]//International Conference on Machine Learning. 2013: 1247-1255.

[10] Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 6558-6569.

[11] Tan H, Bansal M. LXMERT: Learning cross-modality encoder representations from transformers[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019: 5100-5111.

[12] Sun C, Myers A, Vondrick C, et al. VideoBERT: A joint model for video and language representation learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7464-7473.

[13] Zhang D, Li S, Zhang X, et al. Multi-modal graph neural network for joint reasoning on vision and scene text[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12746-12756.

[14] Ramachandram, Dhanesh, and Graham W. Taylor. "Deep multimodal learning: A survey on recent advances and trends." *IEEE signal processing magazine,* 34.6, (2017): 96-108.

[15] Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv:1707.07250, 2017.

[16] Kiela D, Bhooshan S, Firooz H, et al. Supervised multimodal bitransformers for classifying images and text[J]. arXiv preprint arXiv:1909.02950, 201.