IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 9, 2022, accepted February 26, 2022, date of publication March 11, 2022, date of current version March 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3157712

MEDT: Using Multimodal Encoding-Decoding Network as in Transformer for Multimodal Sentiment Analysis

QINGFU QI[1,2], LIYUAN LIN[1], RUI ZHANG[1,2], AND CHENGRONG XUE[1,2]

[1] College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin 300222, China
[2] School of Software and Communications, Tianjin Sino-German University of Applied Sciences, Tianjin 300222, China

Corresponding author: Rui Zhang (zhangrui@tsguas.edu.cn)

This work was supported in part by the Tianjin Sino-German University of Applied Sciences Technology Project under Grant ZDKT2018-006, and in part by the Tianjin Sino-German University of Applied Sciences Video Image Intelligent Analysis and Processing Technology Innovation Team and Tianjin ''131'' Innovative Talent Team.

ABSTRACT Multimodal sentiment analysis is a challenging task in the field of natural language processing (NLP). It uses multimodal signals (natural language, facial gestures, and acoustic behavior) in videos to generate emotional understanding. However, the importance of single modality data in the video to emotional outcomes is not static. With the extension of the time dimension, the emotional attributes of a specific natural language will be affected by non-natural language data, resulting in a vector shift in the feature space. At the same time, long-term dependencies within a specific modality and long-term dependencies between multiple modalities that are ''unaligned'' need to be considered. In response to the above problems, this paper proposes Multimodal Encoding-Decoding Network with Transformer. The network model encodes multimodal data through a Bidirectional Encoder Representations from Transformers (BERT) network and Transformer encoder to resolve long-term dependencies within modalities. And the network reconstructs the Transformer decoder to solve the weight problem of multimodal data in an iterative way. The network fully considers the long-term dependencies between modalities and the offset effect of non-natural language data on natural language data. Under the same experimental conditions, we validated our model on general multimodal sentiment analysis datasets. Compared with state-of-the-art models, the network achieves good progress and strong stability.

INDEX TERMS Auxiliary information, long-term dependence, multimodal sentiment analysis, transformer, vector offsets.

I. INTRODUCTION

Sentiment analysis has always been a popular research direction in the field of NLP. In the early days, most of the work was focused on unimodal research–mainly plain text sentiment analysis [1], [2] –in which the investigations were limited to determining the usage of words in positive and negative scenarios [3] and obtaining emotional results by analyzing the meaning of specific word combinations. Further analysis of human behavior shows that humans transmit information not only through natural language but also through non-natural language (visual and acoustic) [4]. This rich behavioral information can better help us understand human emotional intentions [5]. This behavioral information is considered to be the multimodal language of human beings. With the rapid development of online media, more and more people tend to use video to record their comments and opinions on products or movies. This requires a multi-dimensional analysis of people's opinions and emotions in the video to better understand the information it conveys [6]. Moreover, with the maturity of audio and video feature extraction methods [7], the research progress of multimodal sentiment analysis has also been accelerated. Currently, modeling multimodal language for emotional understanding has become the central research direction of NLP and multimodal machine learning [8]–[10].

With further research, we found that although multimodal language information is processed at the same time, it is still

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa Rahimi Azghadi.

the natural language that plays a decisive role in the final emotional understanding. It is difficult for us to analyze the intentions of an actor by relying only on visual or acoustic behaviors because the non-natural language behaviors of people expressing the same emotion are usually different. Suppose a person shows an emotional state for a certain thing, but it is almost difficult for us to determine whether it is a positive emotion or a negative emotion through facial expressions. When we combine facial behavior with different natural language descriptions, we can clearly understand an actor's emotional intentions, but this will enhance or weaken the original emotion expressed in the current natural language. This leads to another problem. Since multimodal language communication occurs through natural language and non-natural language channels, the meanings of words and sentences transmitted by humans through natural language change dynamically in different non-natural language contexts [11], [12]. In other words, for a sentence that expresses positive emotions in the field of purely natural language, the meanings of words within the language are fixed in the vector space. When nonnatural language behavior is introduced, it will cause the words to shift in the original vector space. The specific change is reflected in the strength and direction and even causes its meaning to be biased to the opposite side.

Furthermore, the heterogeneity of cross-modality typically increases the difficulty of the analysis of human language because the variable sampling rate for each modal sequence can lead to misaligned inherent data [13], expressed as an ‘‘unaligned’’ multimodal language sequence. Therefore, the final result of multimodal emotional discrimination is affected not only by the long-term dependence relationship within a specific mode but also by the long-term dependence relationship between ‘‘unaligned’’ multiple modalities. Therefore, how to coordinate the long-term dependencies within the modalities and the long-term dependencies between ‘‘unaligned’’ multiple modalities is a very important research topic.

In response to the above problems, we proposed the multimodal encoding-decoding network with transformer, which is a model for handling human ‘‘unaligned’’ multimodal languages. The main contributions of this paper are:

- A new model for processing multi-modal data, which is used to solve the problem of the dynamic change of the weight of multi-modal data in the time dimension, and update the cross-modal weight value in an iterative manner.
- Solve long-term dependencies within a single modality and long-term dependencies across modalities, focusing on solving the problem of the offset of the meaning of words in natural language data caused by non-natural language data.

In order to verify the performance of our model in multimodal sentiment analysis, we conducted experiments on the benchmark CMU-MOSI and CMU-MOSEI datasets. We retrained our model and the latest model previously proposed in the same experimental environment and evaluated and compared the final results. In all benchmark tests, our model can outperform the benchmark and is more stable than other models.

The remainder of this article is organized as follows. In Sect. II, we introduce some related work on multimodal emotion recognition. In Sect. III, we elaborate on the overall architecture of our model. In Sect. IV, we describe the data set and baseline model used in the experiment in detail. In Sect. V, we present the results of the experiment and report the necessary analysis. We summarized our model and elaborated on future work in Sect. VI.

II. RELATED WORKS

In this section, we mainly discuss the related work of multimodal sentiment analysis and briefly introduce the two basic models we will use.

A. MULTIMODAL SENTIMENT ANALYSIS

Multimodal sentiment analysis is now a popular research direction. It models natural language and non-natural language to gain emotional understanding. With the emergence of a large number of multimodal datasets (such as CMU-MOSI [14] and CMU-MOSEI [15]), scholars have successively proposed many models for multimodal sentiment analysis. In early work, fusion methods directly connected multiple modal data [16]–[19], and the primary and secondary relationships between the modes were not studied. For example, in literature [16], the author regards the problem of multimodal sentiment analysis as dynamic modeling within and between modalities. The single-mode, dual-mode, and three-mode dynamics are explicitly modeled by calculating the vector field of the triple Cartesian product, and the multimodal emotion fusion tensor is calculated. In [17], the author applies LSTM to each modal view to learn the interaction of a specific view and reconstructs the LSTM memory network to learn multimodal cross-attempt interaction information. In [18], the author decomposes the fusion problem into multiple stages, and each stage focuses on a subset of multimodal signals for specialized and effective fusion. Then, the fusion method is combined with the recurrent neural network system to model the interactions in time and modalities. In literature [19], the author proposed a multimodal attention framework based on recurrent neural networks to learn the joint relationships between multiple modalities and discourse and used contextual information for discourse-level emotion prediction. The multimodal fusion methods mentioned above all put multiple modal information in the same position without emphasizing the primary and secondary relationships between each modal information. Our research is closer to the work reported in [12], [20], [21] confirming that natural language information occupies an important position in multimodal sentiment analysis. In literature [12], for the first time, the author proposed that a speaker's intentions usually change dynamically according to different nonlanguage environments. When modeling human language, not only the

TABLE 1. A summary and comparison of two multimodal fusion methods.

	References	Features
Early Fusion	"Tensor fusion network for multimodal sentiment analysis" [16]	1. Feature vectors are cascaded in the time dimension. 2. The default weight of multimodal data is the same. 3. Flexible placement in different stages of model building.
	"Memory Fusion Network for Multi-view Sequential Learning" [17]	
	"Multimodal language analysis with recurrent multistage fusion" [18]	
	"Contextual Inter-modal Attention for Multi-modal Sentiment Analysis" [19]	
Weighted Fusion	"Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors" [12]	1. Assign different weight properties to each modality data.
	"MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis" [20]	2. Modeling the way of weight acquisition.
	"Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis" [21]	3. Non-natural language data often changes the strength of a particular natural language meaning.

literal meaning of words but also the nonverbal context in which these words appear must be considered. To this end, the author proposes a gated modal hybrid network, which dynamically moves word representations based on nonverbal cues. In literature [21], the author combines the gated modal hybrid network mentioned above with the BERT model without changing the basic structure of the BERT model, which can actually accept nonverbal information. These studies all use textual information as an important carrier and then introduce nonverbal behavior as auxiliary information to form a multimodal emotional understanding. The core of our work is to use natural language information as the dominant modality and non-natural language information as the auxiliary modality to obtain the fusion vector representation in the natural language vector space. Table. 1 summarizes and compares the two multimodal fusion methods.

B. TRANSFORMER AND BERT

In our experiment, two basic encoding networks, the transformer [22] and BERT (Bidirectional Encoder Representations from Transformers) [23], are mainly involved. A transformer is an acyclic neural architecture designed for sequence data modeling. It discards the RNN and CNN as the basic models of sequence learning and completely adopts the attention mechanism; therefore, the architecture does not have the ability to capture sequential sequences. For this reason, the author uses position embedding in the architecture to represent time-series information and finally achieves better performance than the loop structure in terms of results, speed, and depth. BERT is a successful application of a transformer and a successful language model. The input embedding of this model is generated by adding token embedding, segment embedding, and position embedding. Then, multiple encoder layers are applied on top of these input embeddings. Each encoder has a multi-head attention layer and a feedforward layer, and each layer has a residual connection with layer normalization. BERT adopts the automatic coding method to learn the vector representation of the masked mark in the process.

III. PROPOSED APPROACH

In this section, we introduce in detail the **Multimodal Encoding-Decoding Network with Transformer** (MEDT). The purpose of MEDT is to solve the problem of the "unalignment" characteristic of multimodal language sequences, introduce word transfer representation, and finally obtain multimodal emotional fusion vector representation. Different from the previous strategy, we adopted a joint encoding and decoding method with text as the main information and sound and image as auxiliary information to obtain the emotion fusion vector representation. Our model can be divided into two parts: 1) The unimodal encoder, which is used to handle the long-term dependencies within the modal and to encode unimodal information; and 2) The multimodal joint-decoder, which is used to solve the long-term dependencies between "unaligned" multiple modalities, dynamically update the weight attribute values between different modalities at different times, and finally obtain multimodal fusion feature representation. Fig. 1 shows the overall architecture of the model.

The input of the MEDT is multimodal sequence data. This article mainly handles the following three types of multimodal sequence data: natural language $\{Language(l)\}$ and non-natural language $\{Visual(v), Acoustic(a)\}$, where Language is the original text data I_l input into the BERT model (see Sect. III-A1). The initial feature vector of Visual and Acoustic is expressed as $I_m \in \mathbb{R}^{T_m \times d_m}$ $(m = \{a, v\})$, where d_m and T_m represent their respective time dimension and feature dimension.

A. UNIMODAL ENCODER

In this part, we explained the unimodal encoder, and we used different encoding methods for natural language and non-natural language.

1) NATURAL LANGUAGE ENCODER

We used a pre-trained BERT [23] model that performed well in the text domain to encode plain text to extract sentence representations with long-term dependencies. We consider that text information plays a leading role in the final results of sentiment analysis. In order to ensure that the BERT model can extract sentence representations containing sentiment information, we roughly fine-tuned the BERT network on the pure text sentiment classification dataset. The fine-tuned BERT model does not need to achieve the best accuracy and is only used to ensure that a general sentence representation with emotional attributes is obtained. When we apply the model to text in multimodal data, we will also perform synchronous training. We apply the 12-layer BERT to the IMDB dataset [24], which contains 50,000 positive and negative reviews from the movie database, including 36,000 training set samples, 4,000 validation set samples, and 10,000 tests set samples. The fine-tuned BERT model can achieve 89.42% accuracy on the IMDB dataset sentiment dual classification task.

Given the original text data $I_l = [I_1, I_2, \cdots, I_N]$, N is the number of samples. Each sample I_n $(n \in N)$ is a language sequence $I_n = [i_1, i_2, \cdots, i_T] \in \mathbb{R}^{T \times d}$ that carries T word-piece tokens. Two special tokens $[CLS]$ and $[SEP]$ are added to I_n, and we will use the former to predict emotions later. Then, we input I_n into the input embedder, and its output is the input encoding vector $E_n = [e_{CLS}, e_1, e_2, \cdots, e_T, e_{SEP}]$ of BERT after adding markers, segments, and position embedding.

$$E_n = InputEmbedder\left(I_n\right) \in \mathbb{R}^{T_l \times d} \tag{1}$$

T_l is equal to T plus two special symbols. d is the initial encoding vector dimension. Finally, we input E_n into the fine-tuned BERT model and obtain the lexical embedding X_n of the last layer as the text embedding X_l with long-term dependencies.

$$X_l = X_n = Finetuned - BERT\left(I_n\right) \in \mathbb{R}^{T_l \times d_l} \tag{2}$$

where d_l represents the feature dimension of the language modality after passing through the BERT network, which is 768 dimensions.

2) NON-NATURAL LANGUAGE ENCODER

For visual and acoustic data $I_m \in \mathbb{R}^{T_m \times d_m}$ $(m = \{a, v\})$, we emulate the way the transformer encodes text data, and we apply the encoder of the transformer to non-natural language data. In this paper, for convenience, we call it a Non-natural Language Transformer Encoder (NNLE). For any natural language, the position and order of words in a sentence are very important. They are not only part of the grammatical structure of a sentence but also important concepts that express semantics. If the position or sequence of a word in a sentence is different, the meaning of the entire sentence may deviate. Similar to natural language, for non-natural language data with a time dimension, such as continuous changes of facial expressions or voice intonations, if the arrangement order is different, the meaning expressed will also be affected.

Since the transformer model discards the RNN (Recurrent Neural Network) and CNN (Convolutional Neural Network) as the basic models of sequence learning, it completely adopts the attention mechanism, which means that the transformer model does not have the ability to capture time series. In order to enable the sequence to carry time information, following [22], we add position information embedding (PE) to I_m and then apply a Position-wise Feedforward Network (PFN) to obtain non-natural language embedding data $P_m \in \mathbb{R}^{T_m \times d_m}$ $(m = \{a, v\})$ with relative position information:

$$\begin{aligned} PFN &= xW + b \\ P_m &= PFN\left(I_m + PE\left(I_m\right)\right) \end{aligned} \tag{3}$$

The reason why the network is position-wise is that the transformation parameters of each position t are the same when passing the linear layer. $PE\left(I_m\right) \in \mathbb{R}^{T_m \times d_m}$ calculates the fixed position embedding of each position index of the non-natural language data in the time dimension. We leave more details of the positional embedding to I.

NNLE is the same as the traditional transformer encoder (see Fig. 1, right). It consists of N identical coding layers, each layer consists of two sublayers, and each sublayer introduces residual connections and layer normalization. The overall structure is summarized as:

$$LayerNorm\left(x + SubLayer\left(x\right)\right) \tag{4}$$

The two sublayers are the multi-head attention mechanism (MHA) and position-wise fully connected feed-forward network (FFN). The first sublayer MHA utilizes a self-attention block defined as a scaled dot product function:

$$Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5}$$

where Q, K, and V are input vectors with the same shape. Expression $\sqrt{d_k}$ is a scaling factor, where d_k is the feature dimension of the input vector. Multi-head means projecting Q, K, and V through h different linear transformations and finally stitching together different attention results:

$$\begin{aligned} &MultiHeadAttention\left(Q, K, V\right) = Concat\left(head_1, \ldots, head_h\right) W^o \\ &\quad wherehead_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{6}$$

The second sublayer FFN consists of two linear transformations, and the first linear transformation is followed by a ReLU activation function. Similar to the PFN, the FFN is position-wise because the transformation parameters of each position t are the same when passing through the linear layer:

$$FeedForward\left(x\right) = max\left(0, xW_1 + b_1\right) W_2 + b_2 \tag{7}$$

Assume that the input of the 0th layer $Z_m^{[0]} = P_m$. In the ith coding layer, the output $Z_m^{[i-1]}$ of the previous layer first passes through the multihead attention block to obtain the intermediate output $\overline{Z}_m^{[i]}$, which can be expressed as:

$$\begin{aligned} \overline{Z}_m^{[i]} &= LayerNorm \\ &\times \left(Z_m^{[i-1]} + MultiHeadAttention\left(Z_m^{[i-1]}, Z_m^{[i-1]}, Z_m^{[i-1]}\right)\right) \end{aligned} \tag{8}$$

Then, through the second sublayer feedforward network, the final output $Z_m^{[i]}$ of the ith coding layer is obtained:

$$Z_m^{[i]} = LayerNorm\left(\overline{Z}_m^{[i]} + FeedForward\left(\overline{Z}_m^{[i]}\right)\right) \tag{9}$$

Finally, after N coding layers, we obtain the non-natural language embedding $X_m = Z_m^{[N]} \in \mathbb{R}^{T_m \times d_m}$ $(m \in \{a, v\})$ with timing information.

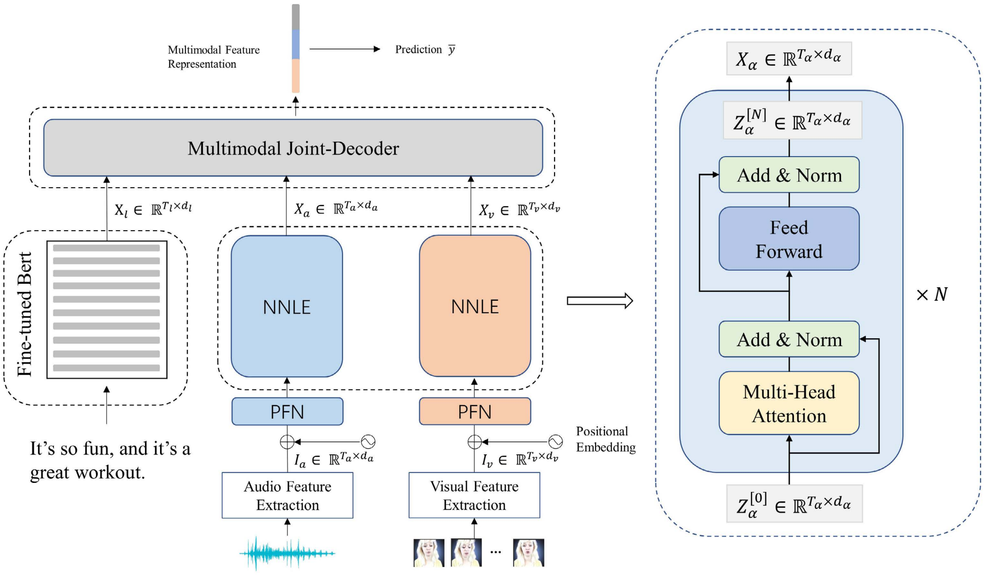

Multimodal Feature Representation
Prediction $\bar{y}$
Multimodal Joint-Decoder
$X_l \in \mathbb{R}^{T_l \times d_l}$
$X_a \in \mathbb{R}^{T_a \times d_a}$
$X_v \in \mathbb{R}^{T_v \times d_v}$
Fine-tuned Bert
NNLE
NNLE
PFN
PFN
It's so fun, and it's a great workout.
$I_a \in \mathbb{R}^{T_a \times d_a}$
$I_v \in \mathbb{R}^{T_v \times d_v}$
Positional Embedding
Audio Feature Extraction
Visual Feature Extraction
...
$X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$
$Z_\alpha^{[N]} \in \mathbb{R}^{T_\alpha \times d_\alpha}$
Add & Norm
Feed Forward
× N
Add & Norm
Multi-Head Attention
$Z_\alpha^{[0]} \in \mathbb{R}^{T_\alpha \times d_\alpha}$

FIGURE 1. The overall architecture of the MEDT. It consists of two parts: 1) the unimodal encoder and 2) the multimodal joint-decoder. The unimodal encoder is divided into the natural language encoder of the fine-tuned BERT and the non-natural language encoder (NNLE), and the PFN is the Positionwise Feedforward Network. The multimodal joint-decoder is a reconstruction of the transformer's decoder. The right side of the figure is the encoder architecture of the transformer.

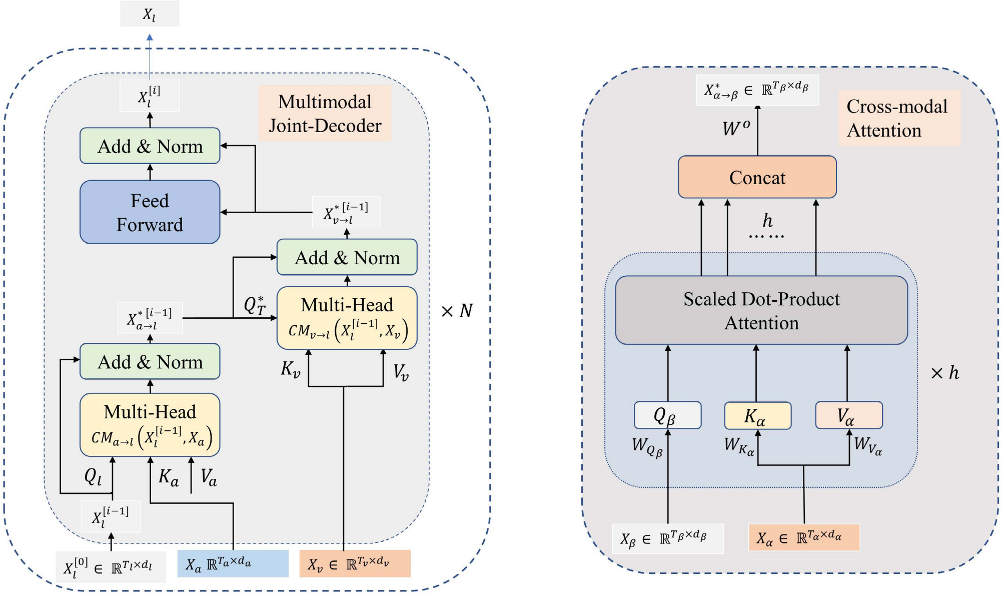

X_l
$X_l^{[i]}$
Multimodal Joint-Decoder
Add & Norm
Feed Forward
$X_{v\to l}^{*\,[i-1]}$
Add & Norm
Q_T^*
Multi-Head
$CM_{v\to l}\left(X_l^{[i-1]}, X_v\right)$
$X_{a\to l}^{*\,[i-1]}$
× N
Add & Norm
K_v
V_v
Multi-Head
$CM_{a\to l}\left(X_l^{[i-1]}, X_a\right)$
Q_l
K_a
V_a
$X_l^{[i-1]}$
$X_l^{[0]} \in \mathbb{R}^{T_l \times d_l}$
$X_a\ \mathbb{R}^{T_a \times d_a}$
$X_v \in \mathbb{R}^{T_v \times d_v}$
$X_{\alpha\to\beta}^{*} \in \mathbb{R}^{T_\beta \times d_\beta}$
Cross-modal Attention
W^o
Concat
h
……
Scaled Dot-Product Attention
× h
Q_β
K_α
V_α
W_{Q_β}
W_{K_α}
W_{V_α}
$X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$
$X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$

FIGURE 2. The left side shows the overall architecture of the multimodal joint decoder, and the right side shows the cross-modal attention mechanism we use.

B. MULTIMODAL JOINT-DECODER

In this part, we reconstruct the decoder of the transformer to obtain the multimodal fusion embedding representation, which is called the multimodal joint-decoder. First, this network considers the characteristics of word vectors in the feature space that are affected by non-natural language data; and second, the network solves the problem of long-term dependence between cross-modalities. (Fig. 2 shows the overall structure). The multimodal joint-decoding layer is composed of two sublayers. The second sublayer adopts a position-wise fully connected feed-forward network such as NNLE. The difference between the two networks is that in the multi-head cross-modal attention mechanism, we use a cross-modal attention block (CM) containing a scaled dot product function [13].

In the cross-modal attention block (CM), two different modal vectors $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$ and $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ are given. We define the queries as $Q_\beta = X_\beta W_{Q_\beta}$, the keys as $K_\alpha = X_\alpha W_{k_\alpha}$, and the values as $V_\alpha = X_\alpha W_{V_\alpha}$, where $W_{Q_\beta} \in \mathbb{R}^{d_\beta \times d_k}$, $W_{K_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$, and $W_{V_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$ are the weights of the linear transformation. Therefore, the cross-modal embedding $Y_\beta \in \mathbb{R}^{T_\beta \times d_k}$ from β to α can be obtained:

$$\begin{aligned} Y_\beta &= CM_{\alpha\to\beta}\left(Q_\beta, K_\alpha, V_\alpha\right) \\ &= softmax\left(\frac{Q_\beta K_\alpha^T}{\sqrt{d_k}}\right) V_\alpha \\ &= softmax\left(\frac{X_\beta W_{Q_\beta}\left(X_\alpha W_{k_\alpha}\right)^T}{\sqrt{d_k}}\right) X_\alpha W_{V_\alpha} \end{aligned} \tag{10}$$

Among the variables, Y_β and Q_β have the same length (that is, T_β), but the data information comes from the feature space of V_α. Expression $\sqrt{d_k}$ is a scaling factor. In particular, the function softmax calculates the attention score matrix $S \in \mathbb{R}^{T_\beta \times T_\alpha}$ from modality β to modality α, where the (i, j)th score indicates the degree of correlation between the information at the ith time step of modality β and the information at the jth time step of modality α. Hence, the ith time step of Y_β is a weighted summary of V_α, with the weight determined by the ith row in the attention score matrix S. However, $CM_{\alpha\to\beta}$ is just single-head cross-modal attention. The multi-head uses h different linear transformations to project X_β and X_α and finally splices the different attention results to obtain the cross-modal embedding representation $\overline{X}_{\alpha\to\beta} = X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$ of modality β to modality α:

$$\begin{aligned} \overline{X}_{\alpha\to\beta} &= MultiheadCM_{\alpha\to\beta}\left(X_\beta, X_\alpha\right) \\ &= Concat\left(head_1, \cdots, head_h\right) W^O \\ wherehead_i &= CM_{\alpha\to\beta}\left(X_\beta W^i_{Q_\beta}, X_\alpha W^i_{K_\alpha}, X_\alpha W^i_{V_\alpha}\right) \end{aligned} \tag{11}$$

For the convenience of the following description, we summarize the multi-head cross-modal attention mechanism as:

$$LayerNorm\left(X_\beta + MultiHeadCM_{\alpha\to\beta}\left(X_\beta, Z_\alpha\right)\right) \tag{12}$$

The input of the multimodal joint-decoder is three types of coded multimodal embeddings $X_l \in \mathbb{R}^{T_l \times d_l}$, $X_a \in \mathbb{R}^{T_a \times d_a}$, and $X_v \in \mathbb{R}^{T_v \times d_v}$, where X_l is the text embedding through the BERT encoder, and X_a and X_v are the acoustic and visual embeddings with time information through the NNLE, respectively. In this paper, we always use the natural language data X_l as the query vector and use X_a and X_v as the key and value vectors, respectively, to obtain the cross-modal fusion embedding representation after the text embedding is offset under the influence of non-natural language data.

The multimodal decoder is composed of N multimodal decoding layers. Assuming that the text embedding of the 0th layer is represented as $Z_l^{[0]} = X_l \in \mathbb{R}^{T_l \times d_l}$, for the ith decoding layer, we first use audio X_a as the initial keys and values to obtain the cross-modal text embedding representation $Z^{*[i]}_{a\to l} \in \mathbb{R}^{T_l \times d_l}$:

$$\begin{aligned} Z^{*[i]}_{a\to l} &= LayerNorm \\ &\times \left(Z_l^{[i-1]} + MultiHeadCM_{a\to l}\left(Z_l^{[i-1]}, Z_a\right)\right) \end{aligned} \tag{13}$$

At this time, the data in $Z^{*[i]}_{a\to l}$ belong to the text feature space, but compared to before, the data shift in direction under the influence of audio information. Then, X_v is passed into the cross-modal attention mechanism as the new keys and values, and $Z^{*[i]}_{a\to l}$ is used as the query vector to further obtain the new cross-modal text embedding representation $Z^{*[i]}_{v\to l} \in \mathbb{R}^{T_l \times d_l}$:

$$\begin{aligned} Z^{*[i]}_{v\to l} &= LayerNorm \\ &\times \left(Z^{*[i]}_{a\to l} + MultiHeadCM_{v\to l}\left(Z^{*[i]}_{a\to l}, Z_v\right)\right) \end{aligned} \tag{14}$$

Here, $Z^{*[i]}_{v\to l}$ is a text embedding representation containing two types of non-natural language information. Finally, after a position-wise feed-forward network, the cross-modal text embedding representation $Z_l^{[i]}$ of the ith layer is obtained.

$$Z_l^{[i]} = LayerNorm\left(Z^{*[i]}_{v\to l} + FeadForward\left(Z^{*[i]}_{v\to l}\right)\right) \tag{15}$$

Finally, after n decoding layers, we obtain the final cross-modal text embedding representation $X_{(a,v)\to l} = Z_l^{[N]} \in \mathbb{R}^{T_l \times d_l}$. Then, we choose the feature vector of the special symbol token [CLS] in the text embedding as the embedding representation $X_f \in \mathbb{R}^{d_l}$ of multimodal fusion information and further use it for sentiment analysis.

IV. EXPERIMENTAL SETTINGS

In this section, we introduce our experimental settings, including the experimental datasets, baselines, and evaluations.

A. DATASETS

In this work, we use two public multimodal sentiment analysis datasets: MOSI and MOSEI. Here, we give a brief introduction to the above datasets.

1) MOSI

The CMU-MOSI [14] dataset is one of the most popular benchmark datasets for multimodal sentiment analysis. It comprises 2,199 short monologue video clips taken from 93 YouTube movie review videos. Human annotators label each sample with a sentiment score from -3 (strongly negative) to 3 (strongly positive). We further processed the dataset and divided it into a training set containing 1,284 samples, a validation set containing 229 samples, and a test set containing 686 samples.

2) MOSEI

The CMU-MOSEI [15] dataset expands its data with a higher number of utterances and greater variety in samples, speakers, and topics than CMU-MOSI. The dataset contains 22,856 annotated video segments (utterances) from 5,000 videos, 1,000 distinct speakers and 250 different topics. We also processed the dataset further and divided it into a training set containing 16,326 samples, a validation set containing 1,871 samples, and a test set containing 4,659 samples.

B. BASELINES

In order to verify the performance of the MEDT, we conducted a fair comparison with the following various state-of-the-art models for multimodal language analysis. These

TABLE 2. Results for multimodal sentiment analysis on CMU-MOSI and CMU-MOSEI with aligned and unaligned multimodal sequences. For the performance indicators, ^{h} means higher is better and $^\ell$ means lower is better. (SD) is the standard deviation of the results of five experiments. For the model, (B) means that the language features are based on BERT. In Acc-2 and F1-Score, the left of the "/" is calculated as "neg./non-neg." and the right is calculated as "neg./pos.".

Metric	MAE$^\ell$(SD)	Corr^h(SD)	Acc-5^h(SD)	Acc-7^h(SD)	Acc-2^h(SD)	F1_score^h(SD)
CMU-MOSI Sentiment (unaligned)						
TFN(B)	95.92(4.68)	66.07(2.54)	40.29(2.49)	35.39(1.92)	77.00(1.79)/77.90(1.84)	77.02(1.76)/77.99(1.80)
LMF(B)	93.60(3.27)	66.76(1.20)	39.63(2.86)	34.84(2.42)	78.31(0.83)/79.36(0.96)	78.28(0.79)/79.39(0.90)
Mult(B)	88.65(1.92)	70.43(0.71)	41.92(3.96)	36.73(2.50)	79.30(0.57)/80.64(0.61)	79.22(0.56)/80.62(0.58)
MISA(B)	77.23(2.35)	78.25(0.66)	47.37(1.28)	42.39(1.01)	81.55(0.33)/83.05(0.64)	81.52(0.34)/83.08(0.60)
Self-MM(B)	71.52(0.44)	79.65(0.70)	53.41(0.20)	46.74(0.53)	82.92(0.38)/84.70(0.50)	**85.84**(0.42)/84.68(0.50)
MEDT(our)	**69.99**(0.89)	**80.92**(0.38)	**55.67**(0.36)	**47.92**(0.56)	**83.32**(0.41)/**84.91**(0.27)	83.25(0.45)/**84.89**(0.31)
CMU-MOSI Sentiment (aligned)						
MISA(B)	76.37(2.15)	78.20(0.85)	48.37(2.06)	42.83(1.59)	82.27(0.73)/**83.90**(0.54)	82.26(0.74)/83.94(0.54)
Self-MM(B)	72.11(1.73)	78.94(0.82)	53.15(1.27)	46.68(1.50)	82.22(1.21)/83.72(1.20)	82.19(1.20)/83.75(1.18)
MEDT(our)	**70.21**(1.35)	**81.09**(0.35)	**54.02**(0.23)	**47.04**(0.34)	**82.30**(0.43)/**83.78**(0.53)	**82.27**(0.45)/**84.81**(0.50)
CMU-MOSEI Sentiment (unaligned)						
TFN(B)	56.83(0.45)	72.17(0.68)	53.36(0.27)	51.86(0.31)	78.15(4.65)/82.25(1.53)	78.74(4.01)/82.18(1.26)
LMF(B)	57.67(0.49)	71.62(0.52)	52.97(0.59)	51.60(0.55)	80.38(0.62)/83.37(0.26)	80.82(0.53)/83.28(0.26)
MISA(B)	55.77(1.13)	74.96(1.25)	53.12(1.35)	51.67(1.22)	81.05(1.18)/84.58(0.19)	81.54(0.96)/84.55(0.25)
Self-MM(B)	**53.04**(0.36)	76.75(0.33)	55.63(0.35)	53.95(0.48)	79.69(4.44)/83.96(1.45)	80.28(3.97)/83.97(1.29)
MEDT(our)	53.10(0.30)	**77.49**(0.34)	**55.97**(0.44)	**54.21**(0.40)	**81.94**(0.61)/**85.57**(0.25)	**82.39**(0.50)/**85.53**(0.30)
CMU-MOSEI Sentiment (aligned)						
MISA(B)	55.63(0.51)	75.16(0.84)	53.72(0.29)	52.09(0.27)	79.25(2.19)/84.17(0.72)	79.95(1.93)/84.23(0.61)
Self-MM(B)	**53.75**(0.27)	**76.50**(0.34)	**55.42**(0.47)	**53.75**(0.27)	81.94(3.38)/84.63(1.33)	82.31(2.91)/84.56(1.18)
MEDT(our)	53.96(0.67)	76.31(0.67)	55.08(0.57)	53.40(0.53)	**82.10**(0.47)/**85.00**(0.46)	**82.36**(0.42)/**84.82**(0.59)

models are trained using extracted BERT word embeddings as their language input:

TFN: The Tensor Fusion Network [16] creates a multidimensional tensor to capture unimodal, bimodal, and trimodal interactions and explicitly model intramodal and intermodal dynamics.

LMF: Low-rank Multimodal Fusion [25] is an improvement of the TFN that uses a low-rank tensor to perform multimodal fusion to improve efficiency.

MulT: The Multimodal Transformer [13] adopts directional pairwise cross-modal attention, which attends to interactions between multimodal sequences across distinct time steps and latently adapts streams from one modality to another.

MISA: Modality-Invariant and -Specific Representations [20] project each modality to two subspaces of modal invariant and specific modalities to capture cross-modal commonality and unimodal private features for task prediction fusion.

SelF-MM: The Self-Supervised Multitask Multimodal sentiment analysis network [21] obtains an informative unimodal representation by jointly learning a multimodal task and three unimodal subtasks. Among the subtasks, the label of the unimodal subtask is obtained through a label generation module based on a self-supervised learning strategy. Then, the multimodal and single-modal tasks are jointly trained to learn consistency and differences, respectively.

C. EXPERIMENTAL DESIGN

1) EXPERIMENTAL DETAILS

We use Adam as the optimizer and the initial learning rate of 5e-5 for the BERT natural language encoder. The learning rate of the two non-natural language encoders is 0.001, and the learning rate of the multimodal decoder and other networks is 0.0001. For a fair comparison, we conducted experiments on our model and the model mentioned above under the same experimental conditions. We ran each model five times and reported the average performance.

2) EVALUATION METRICS

Following previous work [15], [20], the emotional intensity predictions using the MOSI and MOSEI datasets are regression tasks, and the mean absolute error (MAE) and Pearson correlation (Corr) are used as the performance indicators. Additionally, the benchmark also involves classification scores that include seven-class accuracy (Acc-7) and five-class accuracy (Acc-5) ranging from -3 to 3, binary accuracy (Acc-2), and the F-Score. For the binary accuracy score, we chose two different evaluation methods. The first is negative/non-negative classification, where non-negative labels are based on scores ≥ 0 [26]. The second is the more accurate negative/positive classification, where negative and positive classes are assigned to sentiment scores of < 0 and > 0, respectively [13]. We use the segment mark -/- to report the results of these two indicators, where the score on the left represents neg./non-neg. and the score on the right is neg./pos. Furthermore, we calculate the standard deviation (SD) of the five experimental results of the abovementioned evaluation indexes and use it as the stability index of the model.

V. RESULTS AND DISCUSSION

In this section, we have conducted a detailed analysis and discussion of the experimental results on the CMU-MOSI and CMU-MOSEI datasets.

A. RESULTS

Table. 2 shows the comparison results on the MOSI and MOSEI datasets. For a fair comparison, we experimented with our model and the benchmark models under the same experimental conditions. Following previous work [20], [21],

TABLE 3. Examples from the CMU-MOSI data set. The true emotional label lies between strongly negative (−3) and strongly positive (+3). According to the different "data settings", we performed fitting experiments on the "aligned" and "unaligned" data.

	Language + Visual + Audio	Truth	MEDT (aligned)	MEDT (unaligned)
1	"I really think that I really like the mirandas character." + smile + excited voice	1.75	1.66	1.69
2	"And that some are very relatable." + calm + peaceful voice	0.80	1.06	0.98
3	"I did not find it all that good." + frown + irritable voice	-1.80	-1.72	-1.86
4	"Or big collector of the action figures." + calm + gentle voice	0	0.25	0.19

we tested our model (MEDT) on "aligned" data and "unaligned" data according to the different "data settings" and compared its results with those of the benchmark models. First, we apply our model and the benchmark models to "unaligned" data. Compared with the benchmark models, our model achieved significant improvements in all evaluation indicators. As mentioned earlier, when TFN and LMF networks perform multimodal sentiment analysis, each modality data has an equal effect on the final sentiment result. Our model MEDT takes natural language data as the dominant information iteratively updates the weight ratio of non-natural language data to natural language data, and dynamically obtains multimodal fusion feature representations to obtain emotional results. As can be seen from Table. 2, our method has a substantial improvement in the evaluation indicators of classification or regression compared with the previous two networks. For example, on the "unaligned" MOSI data, especially in the regression task, the mean squared error (MAE) dropped directly by 23.61 points, and in the binary classification accuracy metric and F1 score, it also steadily increased by 5 points. Moreover, compared with the state-of-the-art model Self-MM, the accuracy has also been improved. Then, for the "aligned" data, we applied the MISA and Self-MM models that performed well on the "unaligned" data to the "aligned" data, and our model still had the best results. In addition to the basic evaluation indicators, we also recorded the results of five experiments and calculated the standard deviation to show the stability of the model. The results show that the standard deviation of our model for all evaluation indicators is low, and the fluctuation is generally maintained between 0.2-0.9. Compared with other models, our model has strong stability and resistance to randomness, which means that the model can obtain relatively stable output results under different conditions. In Fig. 3, we show the results of five experiments on the Pearson correlation (Corr) indicator of the two datasets in the regression task. Our model guarantees a high Corr index while also ensuring the stability of the model.

B. DISCUSSION

Weighting each modality data and enhancing the associated modality data for a particular task can better achieve the desired results. Our model further validates this idea. However, our model is less flexible than the method that automatically obtains the dominant mode data by building a weight distribution model. Taking the three modalities of natural language (language) and non-natural language (audio and visual) in this paper as examples, our model assumes in advance that natural language is dominant for emotion acquisition. Taking non-natural language as the influencing factor of natural language offset dynamically adjusts the weight value between natural language and non-natural language. This idea is slightly different from the above, the acquisition of the dominant modal data is changed from the model's self-learning to artificial assumption, but the final presentation result is ideal. Therefore, this provides a good idea for our next step, that is, how to let the model learn and determine the dominant modal data autonomously. At the same time, it can be seen from Table. 2 that the evaluation results of our model on aligned data and unaligned data are also slightly different, which is reflected in the fact that the evaluation indicators of unaligned data are generally better than aligned data. This is because the multi-modal feature fusion used in the early stage is based on the cascade of feature vectors, which requires multiple modal sequence data to be consistent in the time dimension, which is convenient for model building. But this leads to the unavoidable loss of some information. Our model does not need to take this factor into account, and access to more information ensures the superiority of our model.

C. QUALITATIVE ANALYSIS

In order to verify the more intuitive performance of our model in regression tasks, in Table. 3, we selected multimodal samples from the MOSI dataset to display the results. We applied the model to "aligned" data and "unaligned" data and fitted the true values separately. Our model has a better fitting effect on samples showing strong emotions. Regardless of whether it is for strong positive emotions or negative emotions, the fitting error of our two experimental results to the true value is maintained at ±0.1; furthermore, for the neutral sample, the fitting error of the emotional result is approximately ±0.25. Overall, our model shows an excellent emotional fitting effect.

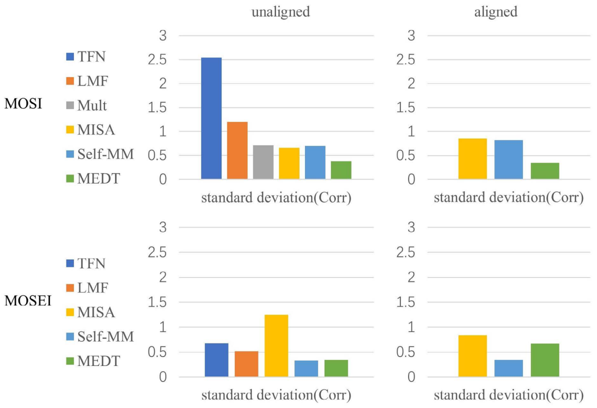


FIGURE 3. On the MOSI and MOSEI datasets, we show the standard deviations of the five experimental results on the Corr indicator on the aligned data and the unaligned data, respectively.

VI. CONCLUSION

In this article, we introduce a Multimodal Encoding-Decoding Network with Transformer (MEDT) for multimodal sentiment analysis. We use different encoding methods for the three multimodal information: for language data, we use a pre-trained BERT model to obtain lexical embedding; and for visual and acoustic language data, we use the transformer's encoder to encode non-natural language data to obtain embedding representation. Finally, we also reconstructed the decoder to obtain cross-modal multimodal embedding representation. Our model finally solves the long-term dependence relationship between specific modalities and multimodalities and considers the offset characteristics of text embedding under the influence of non-natural language information. Our experiments have proven the superior performance of the MDET. However, our model was designed around the idea of taking natural language as the dominant and non-natural language as auxiliary information. This greatly limits the flexibility to switch dominance between multiple modalities in the model. Next, we will focus on the idea of flexibly obtaining dominant information among multiple modal information, and propose a better model for multimodal sentiment analysis. In addition to this, we will also envision better multimodal data alignment methods to better fit our models.

DISCLOSURES

The authors declare no conflict of interest.

APPENDIX I. POSITIONAL EMBEDDING

Because the Transformer model abandons RNN and CNN as the basic model of sequence learning and completely adopts the attention mechanism, the Transformer model does not have the ability to capture sequential sequences. At the same time, the order of arranging the input sequence does not change the behavior of the Transformer or change its output. To solve this problem, following the work of [13], we use sin and cos functions to encode the position information of the sequence of length T, and the frequency is determined by the feature dimension index. In particular, we define the positional embedding (PE) of the sequence $X \in \mathbb{R}^{T\times d}$ (where T is the length) as a matrix, where:

$$PE[pos, 2i] = sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$
$$PE[pos, 2i+1] = cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{16}$$

where pos is the position index in the time dimension, i is the dimension, and the value is 0, $\left[\frac{d}{2}\right]$. Therefore, each feature dimension of PE is a position value showing a sinusoidal pattern. After calculation, the position embedding is directly added to the sequence, so that $X + PE$ encodes the position information of the element at each time step.

REFERENCES

[1] P. Bo and L. Lillian, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. Assoc. Comput. Linguistics*, 2004, p. 271, doi: 10.3115/1218955.1218990.

[2] G. Vinodhini and R. M. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," *Int. J.*, vol. 2, no. 6, pp. 282–292, 2012, doi: 10.1007/978-1-4899-7502-7.

[3] H. Pham, T. Manzini, P. Liang, and B. Poczos, "Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis," in *Proc. Challenge-HML*, Jul. 2018, pp. 53–63, doi: 10.18653/v1/w18-3308.

[4] R. G. Milo, "Tools, language and cognition in human evolution," *Int. J. Primatol.*, vol. 16, no. 6, pp. 1029–1031, Dec. 1995, doi: 10.1007/BF02696116.

[5] C. Manning, M. Surdeanu, and J. Bauer, "The Stanford CoreNLP natural language processing toolkit," in *Proc. Assoc. Comput. Linguistics*, Jun. 2014, pp. 55–60, doi: 10.3115/v1/P14-5010.

[6] S. Poria, E. Cambria, and D. Hazarika, "Context-dependent sentiment analysis in user-generated videos," in *Proc. Assoc. Comput. Linguistics*, vol. 1, Jul. 2017, pp. 873–883, doi: 10.18653/v1/P17-1081.

[7] A. Abate, P. Barra, S. Barra, C. Molinari, and M. Nappi, "Clustering facial attributes: Narrowing the path from soft to hard biometrics," *IEEE Access*, vol. 8, pp. 9037–9045, 2020, doi: 10.1109/ACCESS.2019.2962010.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[9] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018, doi: 10.1016/j.knosys.2018.07.041.

[10] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proc. Assoc. Comput. Mach.*, New York, NY, USA, vol. 9, Jul. 2018, pp. 350–358, doi: 10.1145/3219819.3219853.

[11] R. Bamler and S. Mandt, "Dynamic word embeddings," in *Proc. JMLR*, Sydney, NSW, Australia, vol. 70, 2017, pp. 380–389, doi: 10.5555/3305381.3305421.

[12] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, and A. Zadeh, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 7216–7223. [Online]. Available: https://arxiv.org/abs/1811.09362

[13] Y. Tsai, S. Bai, P. Liang, and J. Kolter, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Assoc. Comput. Linguistics*, Jul. 2019, pp. 6558–6569, doi: 10.18653/v1/P19-1656.

[14] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *CoRR*, vol. abs/1606.06259, pp. 1–10, Jul. 2016.

[15] A. Zadeh, P. Liang, and S. Poria, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. Assoc. Comput. Linguistics*, vol. 1, Jul. 2018, pp. 2236–2246, doi: 10.18653/v1/P18-1208.

[16] A. Zadeh, M. Chen, and S. Poria, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Assoc. Comput. Linguistics*, Sep. 2017, pp. 1103–1114, doi: 10.18653/v1/d17-1115.

[17] A. Zadeh, P. Liang, N. Mazumder, and S. Poria, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 5634–5641.

[18] P. Liang, Z. Liu, A. Zadeh, and L. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Assoc. Comput. Linguistics*, Nov. 2018, pp. 150–161, doi: 10.18653/v1/D18-1014.

[19] D. Ghosal, M. S. Akhtar, D. Chauhan, and S. Poria, "Contextual intermodal attention for multi-modal sentiment analysis," in *Proc. Assoc. Comput. Linguistics*, Nov. 2018, pp. 3454–3466, doi: 10.18653/v1/D18-1382.

[20] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and specific representations for multimodal sentiment analysis," in *Proc. Assoc. Comput. Mach.*, New York, NY, USA, vol. 10, 2020, pp. 1122–1131, doi: 10.1145/3394171.3413678.

[21] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, Nov. 2021, vol. 35, no. 12, pp. 10790–10797, doi: 10.18653/v1/D18-1382.

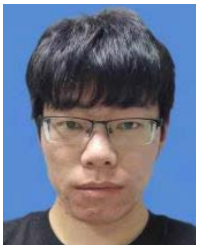
[22] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Nov. 2017, pp. 5998–6008.

[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Nov. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[24] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2011, pp. 142–150, doi: 10.5555/2002472.2002491.

[25] Z. Liu, Y. Shen, and V. Lakshminarasimhan, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. Assoc. Comput. Linguistics*, Jul. 2018, pp. 2247–2256, doi: 10.18653/v1/P18-1209.

[26] A. Zadeh, P. Liang, S. Poria, P. Vij, and E. Cambria, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.



QINGFU QI was born in Liaocheng, Shandong, in 1997. He received the bachelor's degree from the Qilu University of Technology, in 2019. In 2019, he studied for a master's degree in control engineering at Tianjin University of Science and Technology. He is currently pursuing the joint master's degree with the Tianjin University of Science and Technology and the Tianjin Sino-German University of Applied Sciences. Currently, he has published an academic paper as the first author. His research interest includes multimodal sentiment analysis.

LIYUAN LIN received the M.S. degree in communication engineering and the Ph.D. degree in information communication engineering from Tianjin University, China, in 2009 and 2016, respectively. She is currently working with the Tianjin University of Science and Technology. Her research interests include computer vision, virtual reality, and deep learning theory.

RUI ZHANG received the Doctorate degree in information and communication engineering. He is currently the Deputy Dean of the School of Software and Communication, Tianjin Sino-German University of Applied Sciences, a member of the Communist Party of China, an Associate Professor, a Master Tutor, and a person in charge of communication engineering. He is a member of Tianjin "131" Innovative Talent Team. In the past few years, he has published more than ten high-level papers, of which the first author or corresponding author has published two SCI searches, six EI searches, and four national invention patents. The main research projects consist of four horizontal topics with funding of more than 1.5 million and three provincial and ministerial-level topics with funding of more than ten million. His research interests include machine learning, multi-modal analysis, water, air communication, and heterogeneous detection.

CHENGRONG XUE was born in Tongcheng, Anhui, in 1996. He received the bachelor's degree from Huaibei Normal University, in 2019. He is currently pursuing the master's degree in electronic information with the Tianjin University of Science and Technology, with a focus on natural language processing and machine vision.