Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6243347, 12 pages
https://doi.org/10.1155/2022/6243347

Research Article

Multimodal Sentiment Analysis Based on Interactive Transformer and Soft Mapping

Zuhe Li,[1,2] **Qingbing Guo**,[1] **Chengyao Feng**,[3] **Lujuan Deng**,[1] **Qiuwen Zhang**,[1] **Jianwei Zhang**,[2] **Fengqin Wang**,[1] **and Qian Sun**[1]

[1]*School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China*
[2]*Henan Key Laboratory of Food Safety Data Intelligence, Zhengzhou University of Light Industry, Zhengzhou 450002, China*
[3]*Brandeis High School, San Antonio, TX 78249, USA*

Correspondence should be addressed to Qian Sun; 331907020384@zzuli.edu.cn

Received 27 November 2021; Revised 23 December 2021; Accepted 15 January 2022; Published 3 February 2022

Academic Editor: Mohamed Elhoseny

Multimodal sentiment analysis aims to harvest people's opinions or attitudes from multimedia data through fusion techniques. However, existing fusion methods cannot take advantage of the correlation between multimodal data but introduce interference factors. In this paper, we propose an Interactive Transformer and Soft Mapping based method for multimodal sentiment analysis. In the Interactive Transformer layer, an Interactive Multihead Guided-Attention structure composed of a pair of Multihead Attention modules is first utilized to find the mapping relationship between multimodalities. Then, the obtained results are fed into a Feedforward Neural Network. The Soft Mapping layer consisting of stacking Soft Attention module is finally used to map the results to a higher dimension to realize the fusion of multimodal information. The proposed model can fully consider the relationship between multiple modal pieces of information and provides a new solution to the problem of data interaction in multimodal sentiment analysis. Our model was evaluated on benchmark datasets CMU-MOSEI and MELD, and the accuracy is improved by 5.57% compared with the baseline standard.

1. Introduction

Sentiment analysis aims to detect affective states or subjective information from data. It is often used to understand or judge people's attitudes, opinions, and sentiment. Traditional sentiment analysis mainly focuses on text data, using statistical knowledge combined with natural language processing and machine learning techniques to study and analyze the sentiment polarity of sentences or documents [1]. In reality, human sentiment is expressed not only through language, but also through acoustic information (e.g., speakers' tone of voice) and visual information (e.g., speakers' facial expressions and body movements). For example, social media users are no longer satisfied with sharing feelings and emotions in the form of text but tend to use multimedia forms such as pictures or videos when sending blog posts. In this way, they can express their attitudes more abundantly. The multimodal information at different granularity levels is spread out by people, and traditional sentiment analysis methods cannot handle this problem well. On the basis of text information, multimodal sentiment analysis can use multimodal representation learning, multimodal alignment, and multimodal fusion technologies to combine acoustic and visual information to eliminate ambiguity caused by a single modality.

At present, the research of multimodal sentiment analysis can be divided into two categories according to the number of talkers: multimodal narrative sentiment analysis and multimodal conversational sentiment analysis [2]. Multimodal narrative sentiment analysis usually transmits the author's personal attitude with narrative information. The expression of information is relatively independent and does not involve the interaction between multiple speakers. For example, in the analysis of public opinion, multimodal narrative sentiment analysis is used to analyze the information in social media platforms such as microblog and

twitter to obtain users' attitude towards a hot event. In contrast, there is more than one speaker in multimodal conversational sentiment analysis. The sentiment or attitude of each talker is transmitted in the form of dialogue or communication. In the process of interaction, the sentiment state of speakers is mutually influencing, jumping, and unstable. For example, in customer sentiment analysis, multimodal conversational sentiment analysis is used to obtain interaction clues among multiple customers and to predict sentiment evolution trend in the interaction process.

The research of multimodal sentiment analysis is not mature, and there are still some unsolved problems. Among them, the establishment of multimodal information fusion mechanism has become the main bottleneck restricting the development of this field. Multimodal information fusion aims to fuse the representations of different modalities and retain the key information in each modality during the fusion process. To solve this problem, there are several ideas: feature-level fusion, decision-level fusion, and hybrid fusion [3, 4], as shown in Figure 1. The yellow part in Figure 1 shows the feature-level fusion mechanism, which extracts feature vectors from multimodal data, respectively, and fuses them into a multimodal feature vector. The fused vector is used to judge sentiment [5, 6]. This fusion method can obtain the association among various modalities, but the features need to be mapped to a shared space for fusion because the semantic spaces of different modalities are different from each other. The blue section in Figure 1 shows the decision-level fusion mechanism, which first conducts single modality sentiment analysis independently and then fuses the results to obtain the final decision [7, 8]. This fusion method can design a feature extraction method for the semantic space of each modality to obtain the optimal single modality decision, but independent learning will cause the overall time cost to be too large. The green area in Figure 1 shows the hybrid fusion mechanism, which comprehensively uses the feature-level fusion and decision-level fusion methods to reduce the time cost on the basis of fully learning the associated information of each modality.

In addition, multimodal interaction is also a hot topic. Multimodal interaction aims to supplement the information of different modalities. When the information of one modality information is missing, the data of another modality is used to make up the missing part. This kind of interaction is concealed, complex, and dynamic. It makes multiple modalities related to each other and affects the final sentiment judgment. How to accurately and comprehensively model the complex interaction in multimodal data is still troublesome in this field. Multimodal interaction mainly includes two situations: the interaction between features and the interaction between decisions. For the interaction between features, multimodal features in a shared space need to be aligned through semantic space barriers across different domains to achieve semantic fusion. Several methods such as feature concatenation [9, 10] and attention mechanisms [11] have been developed to solve this problem. For the interaction between decisions, it is necessary to understand the correlation between multimodal data, and this is a comprehensive cognitive decision-making process. The voting method and linear weighting method are proposed to solve this problem.

In view of the above two problems, it can be seen that, in the process of multimodal information fusion, making full use of the correlation among different modalities to make each modality learn from each other is the key to multimodal sentiment analysis. Based on the Transformer-Encoder framework [12], we propose a model to learn the information of different modalities. In this process, we utilize the Guided-Attention mechanism [13] to introduce the information of other modalities. Then, the unimodal results are mapped to higher dimensions for fusion, and the final decision is made according to the fusion results. The main contributions of this work can be summarized as follows:

(i) We propose a multimodal sentiment analysis model based on Interactive Transformer and Soft Mapping. This model can achieve the optimal decision of each modality and fully consider the correlation information between different modalities.

(ii) We propose the Interactive Transformer (IT) structure, which can mine the interactive information between modalities.

(iii) We propose the Soft Mapping (SM) structure, which projects each modality representations to a new space for fusion.

(iv) The experimental results on two benchmark datasets show that the proposed method can achieve better accuracy than those existing methods only using linguistic and acoustic information.

The remaining sections of this paper are arranged as follows. In Section 2, we will review the related work on CMU-MOSEI dataset and MELD dataset. In Section 3, we will describe our method and the core content of the model structure in detail. In Section 4, we will describe the datasets and the method of data preprocessing. In Section 5, we will provide experimental results and analysis. In Section 6, we will summarize the paper and discuss the potential future work.

2. Related Work

In the early studies, researchers focused on the problem of multimodal fusion, and they tried to solve the problem of multimodal sentiment analysis by improving the fusion method. For example, some models fuse each modality representations according to different granularities.

Amir Zadeh et al. proposed the Memory Fusion Network (MFN) [14] to solve the problem of multimodal sequence modeling. The interaction problem is divided into view-specific interactions and cross-view interactions, which is implemented in three steps. Firstly, they use LSTM to learn view-specific interactions individually. Then, they learn the cross-view interactions by Delta-memory Attention Network (DMAN). Finally, they summarized it through time with a Multiview Gated Memory. MFN takes the interaction between modalities as a breakthrough point, which ensures the performance without sacrificing the time complexity and

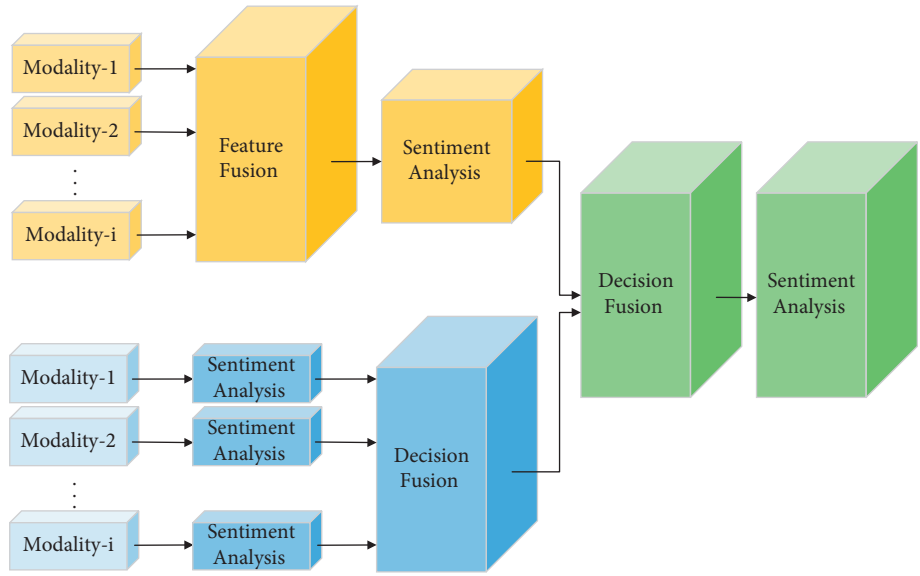


Figure 1: Multimodal fusion mechanisms.

space complexity of the model. Amir Zadeh et al. proposed Graph Memory Fusion Network (Graph-MFN) [15]to improve the Memory Fusion Network (MFN). In this way, they use a new fusion model called the Dynamic Fusion Graph (DFG) to build the n-modal interactions and replace the original fusion component in MFN. This improvement enables the method to dynamically select the appropriate fusion graph according to its importance in the process of multimodal information fusion.

The above methods provide ideas for the research in the field of multimodal sentiment analysis, but they all ignore the interaction between modality representations, such as supplementing by one modality when the information of another modality is scarce. This defect makes these methods unstable. With the development of deep learning, researchers begin to pay attention to the interaction between modalities. Some researchers try to use LSTM model to obtain the contextual information around each discourse from the perspective of time. Some researchers use gating mechanism or attention mechanism to couple features. These ideas promote the development of multimodal sentiment analysis research.

Kumar et al. proposed a gating mechanism, named learnable gates [16]. This mechanism considers that each modality is not of equal importance and needs to be dynamically adjusted according to the linguistic information, tone of the speaker, and facial expressions of utterance. This method explains how multiple modalities contribute to sentient, selectively learning cross fusion vectors to solve the noise problem in the process of multimodal information fusion. Sahay et al. proposed Relational Tensor Network architecture [17]. Tensor fusion is applied to the modality features of each video clip, and LSTM network is used to model the sequence between clips. In this method, the interaction between modalities is refined to a single time segment from the time level, and the operation can be performed frame by frame. Shenoy and Sardana proposed an end-to-end RNN architecture [18], named Multilogue-Net. They assume that the sentiment or emotion governing a particular utterance predominantly depends on 4 factors: interlocutor state, interlocutor intent, the preceding and future emotions, and the context of the conversation. Firstly, the model represents the speaker's state by learning multiple state vectors of a given utterance and then uses pairwise attention mechanism to obtain the relationship among all modalities. This method starts with the emotional states of both sides of the conversation and fully obtains the relevance and context information between modalities to assist multimodal emotion recognition.

Transformer draws the global dependency between input and output completely based on the attention mechanism without recurrence and convolutions [12]. It utilizes the Multihead Attention to replace the recurrent layers and achieves excellent results in translation tasks. At present, researchers have tried to apply it to multimodal sentiment analysis, providing new research directions for solving problems in this field. Delbrouck et al. describe a Transformer-based joint-encoding (TBJE) architecture for the task of Emotion Recognition and Sentiment Analysis [19]. The model uses Transformer-Encoder framework for emotion recognition, and the proposed joint-coding framework can fuse any kind of modality information.

In summary, we believe that it is necessary to grasp the implicit correlation between modalities. For example, in the issue of sentiment ambiguity, because of the difference in intonation and context, the true meaning of the language can be very different. Implicit correlation is used to analyze the language environment. At the same time, acoustic information and visual information are used to analyze intonation and body language. After comprehensive analysis, we can get the real sentiment and emotion contained in the information. Finally, we choose to use the Transformer-Encoder framework to draw the internal dependence of a single modality, which is doped with residual transformation

and layer normalization to enhance the adaptability of the model. The Multihead Attention module in the coding framework is improved by the idea of Guided-Attention mechanism, so that the information of a certain position can notice the representation information in other modalities.

3. The Proposed Method

This section mainly introduces the framework structure of the model. Interactive Transformer layer is based on the Transformer model, which uses the coding framework to learn the representation information of different modalities. It only needs to rely on the attention mechanism combined with feedforward neural network to achieve the effects of other network models. In fact, the model can obtain the global dependency between input and output without involving the recursive structure of sequence coding. In this process, the Interactive Multihead Guided-Attention (IMHGA) structure proposed by us can introduce the information of other modalities to complete interaction. IMHGA structure is a combination of two improved Multihead Attention (MHA) modules; we will elaborate its principle in Section 3.2. Finally, Soft Mapping is used to map the local results of each modality to higher dimensions for fusion, and the final decision is based on the fusion results.

3.1. Overall Architecture. As shown in Figure 2, in addition to the data preprocessing part, the model consists of two parts: Interactive Transformer (IT) layer and Soft Mapping (SM) layer. The Interactive Transformer layer is composed of N blocks stacked side by side; each block is composed of IMHGA structure and Feedforward Neural Networks (FNN). The Soft Mapping layer consists of stacking Soft Attention (SA) module and the output of SA module. After preprocessing, the data is transferred to Interactive Transformer layer, in which the representation information of each modality is learned. Because the information of other modalities is introduced, it can superimpose the information from different representation subspaces in the learning of a single modality representation information. Then, the result is passed into the FNN layer which is composed of full connection layer and nonlinear activation function. In each block, the two sublayers are finally subjected to a residual transformation and layer normalization (A & L), as shown in the following formula:

$$\text{LayerNorm}\,(x + \text{Sublayer}\,(x)). \tag{1}$$

Through our experiments, it is found that the best result can be obtained when $N = 4$ and all blocks' parameters are independent. The output of the previous block is used as the input of the next block. Finally, the output of coding module is input into Soft Mapping layer. They will be mapped to a higher dimensional space for fusion in order to obtain the final result.

3.2. Interactive Multihead-Guided Attention. The Interactive Multihead Guided-Attention (IMHGA) structure is composed of a pair of improved Multihead Attention (MHA) modules, which we call Guided-Attention (GA). The core idea of it is to use attention mechanism to determine the corresponding relationship between two languages, and there is no dependency during forward propagation. Therefore, attention mechanism can execute operations in parallel and speed up the training of the model to reduce the time cost. We apply this idea to multimodal problems, hoping to find the mapping relationship between two-modality information. On this basis, with the help of Multihead Attention mechanism, we use the other modality information as a guide when learning one modality information.

The query (Q), key (K), and value (V) in IMHGA structure come from multiple modality data. It is not like traditional MHA module, which comes from the same modality data. As shown in Figure 3, GA-x and GA-y learn modality-x and modality-y, respectively, where the vectors K and V come from the currently learning modality and the vector Q comes from another modality. Taking GA-x's learning of modality-x as an example, all vectors are subjected to a linear transformation. Then, the query (Q_y) from the modality-y and the key (K_x) from the modality-x are used to calculate the similarity weight by the dot product function, and results are normalized by Softmax function. Finally, the weight is used to perform a weighted summation of the value (V_x) from modality-x. The calculation method in this step is the result of Guided-Attention module. The specific process is shown in the following formula.

$$\text{Guided} - \text{Attention}\left(Q_y, K_x, V_x\right) = \text{soft}\max\left(\frac{\left(Q_y K_x^T\right)}{\sqrt{d}}\right)V_x. \tag{2}$$

The above operations are performed for a total of h times, and each time is regarded as a head module. The result of IMHGA structure can be obtained by splicing and linear changing the results of h times. Note that, in order to make the dot product not too large, the calculated similarity weight is usually divided by the dimension of K and the parameter W of linear transformation in each head is different. The specific process is shown in the following formulas:

$$\text{head}_i = \text{Guided} - \text{Attention}\left(Q_y W_i^Q, K_x W_i^k, V_x W_i^V\right), \tag{3}$$

$$\text{IMHGA}\left(Q_y, K_x, V_x\right) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)W^O. \tag{4}$$

Through a large number of experiments, we found that the model can play the best effect when using the language and acoustic data. In this paper, we mainly use the fusion of two-modality data. Generally speaking, the language modality is used as main information, and the acoustic modality is used as auxiliary information. However, some important information may be ignored, such as mood and intonation in acoustic modality. They can help us identify such special situations as polysemy and irony in the language modality. Therefore, we regard the status of the two as equal and make them modulate each other. This is the meaning of Interactive in IMHGA structure. After IMHGA structure, the two generated matrices are output to the next Soft Mapping layer in parallel through FNN for fusion.

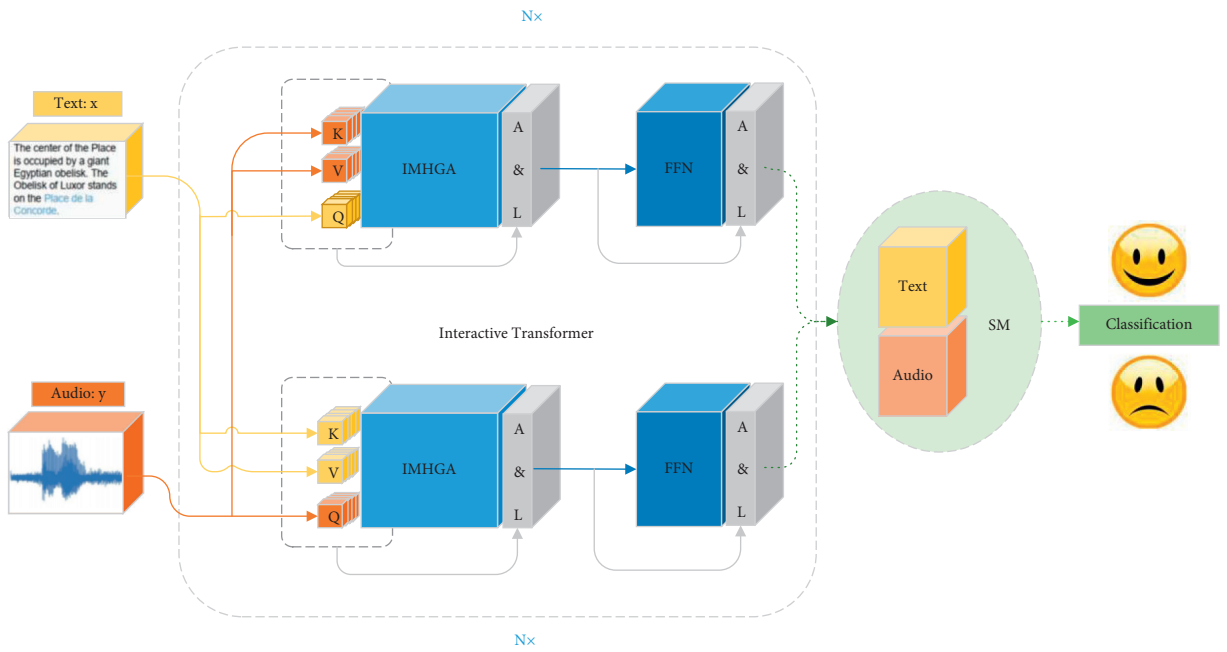


Figure 2: Overview of the proposed framework.

3.3. Soft Mapping. So far, the model has learned the interaction information between the modalities. Before sentiment classification, the learning results of each modality need to be projected into a new performance space in SM for fusion, as shown in Figure 4. Specifically, we map vectors $f_{\text{text}}, f_{\text{audio}}$ from each FNN to a higher dimension, as shown in the following formula:

$$E_i = w_i f_i^T, \quad \{i \in \text{text}, \text{audio}\}, \tag{5}$$

where w_i is a $2k \times 1$ transformation matrix and the vector f_i is embedded into a higher dimension $2k \times k$. Then, we use the set $\{v_j\}$ of vector size $1 \times 2k$ to do Soft Attention for matrix in the high dimensional space. After the results are weighted and summed, they are integrated into the vector m_i of size k. The calculation process is shown in the following formulas:

$$a_{ij} = \text{Soft max}(v_j E_i), \quad j \in [1, N], \tag{6}$$

$$\text{SoftAttention}(E_i) = m_i = \sum_{j=1}^{N} (a_{ij} E_j), \tag{7}$$

where m_i is the calculation result on a single node on the sequence. Therefore, we need to stack the results of all nodes in the whole sequence to get the Soft Mapping feature. As is shown in the following formula, we have

$$s = \text{Stacking}\left(\sum_{j=0}^{N} (m_j)\right). \tag{8}$$

Note that a residual transformation and layer normalization are carried out to ensure that the input of the next round contains the result of the previous round, at the end of this process. It is shown in the following formula:

$$E = \text{LayerNorm}(E + s). \tag{9}$$

The vector s is the result of each modality. On this basis, the vectors obtained from the two modalities are summed according to the element order, and the sum results are classified and predicted according to the following formula:

$$y \sim p = W_a(\text{LayerNorm}(s_{\text{text}} + s_{\text{audio}})). \tag{10}$$

4. Data Preparation

4.1. CMU-MOSEI Dataset and MELD Dataset. We use CMU-MOSEI dataset [15] to verify our model. It is mainly composed of personal monologue video, including 250 different themes and thousands of different speakers. The dataset decomposes the video into video, audio, and text forms and contains a variety of tags. The first is sentiment, which is divided into seven levels [−3, 3], including the following: [−3: highly negative, −2 negative, 1 weakly negative, 0 neutral, +1 weakly positive, +2 positive, and +3 highly positive]. The other is based on Ekman emotions; Ekman emotions of the following: {happiness, sadness, anger, fear, disgust, and surprise} are annotated on a [0, 3] Likert scale for presence of emotion x: [0: no evidence of x, 1: weakly x, 2: x, 3: highly x]. The label distribution of CMU-MOSEI dataset is shown in Figure 5.

In addition, we use MELD dataset [20] to evaluate the model. It selects the multiperson dialogue scene in the TV series Friends as its material, which also contains linguistic, acoustic, and visual information. Its label is also multi-category, which divides emotions into anger, trouble, fear, joy, neutral, sadness, and surprise. In order to adapt to different needs, the above classification is divided into rougher negative, positive, and neutral sentiments. The label distribution of MELD dataset is shown in Figure 6.

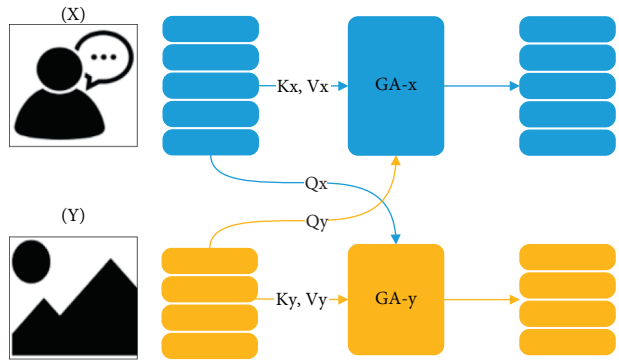


Figure 3: The structure of interactive Multihead Guided-Attention.

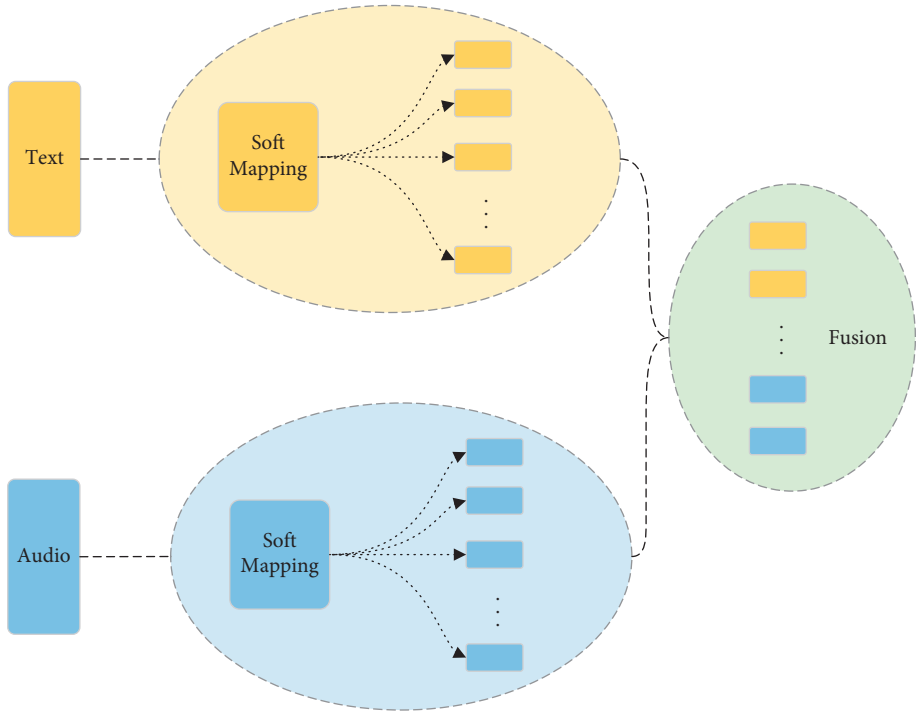


Figure 4: The structure of Soft Mapping.

4.2. Feature Extraction. Next, the preprocessing process of modality data is introduced, and different methods are selected to extract features according to their respective characteristics.

4.2.1. Linguistic Feature. Sentiment analysis has a long history of development in linguistics. It has its own characteristics from semantic-based sentiment dictionary methods to machine learning-based sentiment classification methods. In this study, in order to enable the fusion of multimodal data, the above methods will no longer be applicable and should be processed from a more abstract point of view. For linguistic data, we need to transform it into a vector containing semantic and grammatical information to represent it. Firstly, the original linguistic data is analyzed to construct a cooccurrence matrix for words. Then, based on the distributed representation of the matrix, the cooccurrence matrix is decomposed by using the association between words to obtain the representation vector of words. Specifically, we process the text data to obtain valid words and then count the frequency of word occurrences and record them in the cooccurrence matrix X. The element of X is $x_{(i,j)}$, which indicates the number of times word-i and word-j appear in the same window. Because there are 14176 independent words, we create a cooccurrence matrix X with dimension 14176×14176. We use GloVe [21] to embed the matrix X, and each word is embedded into a vector of 300 dimensions.

4.2.2. Acoustic Feature. Audio is a way for human beings to express their emotions. Intuitively speaking, acoustic information is another form of linguistic information, and the emotion at this time is consistent with the expression of linguistic information. However, acoustic data is more

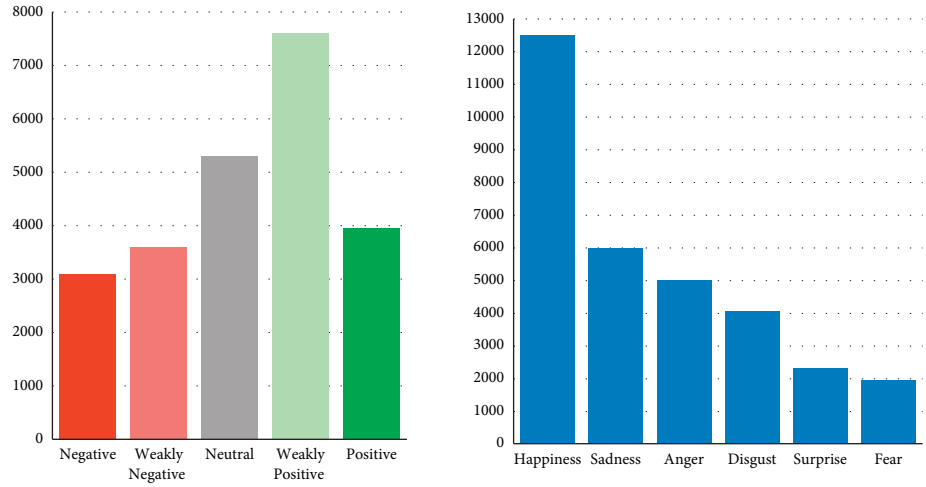


Figure 5: Motion histogram of CMU-MOSEI dataset [15].

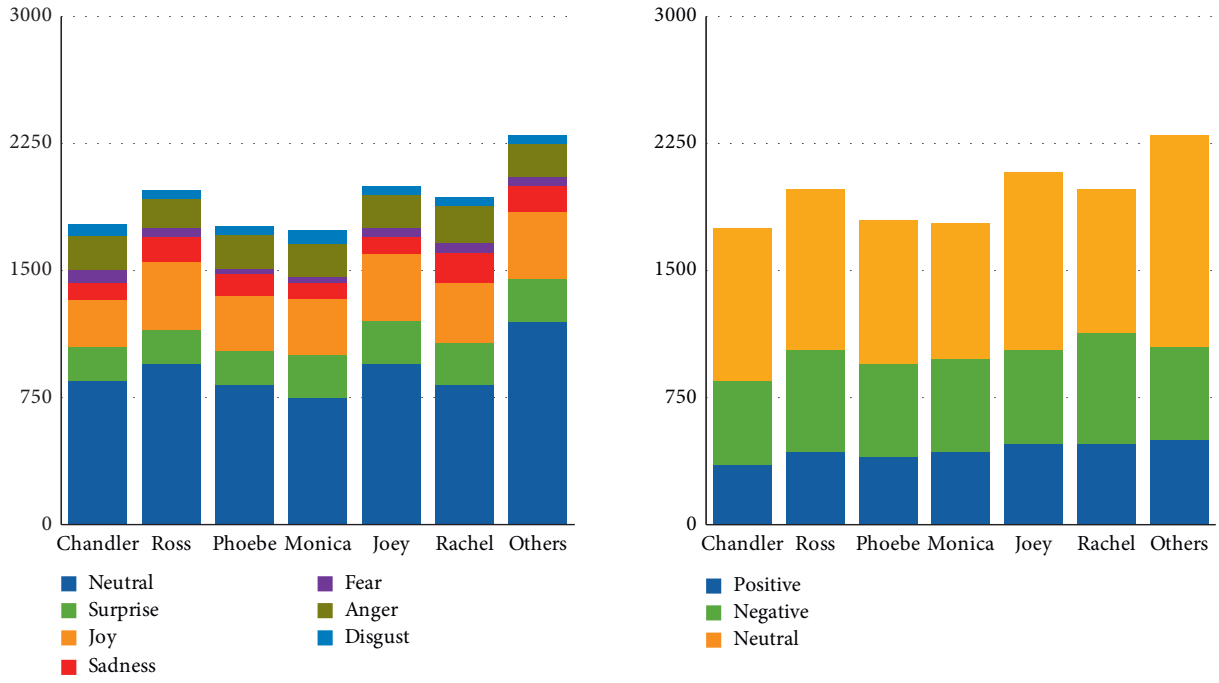


Figure 6: Character distribution across MELD [20].

complex than linguistic data and also contains a large part of unique acoustic information, such as laughter, sighs, high intonation, and low intonation. These are the key tasks of acoustic data in multimodal emotion recognition. The most abundant emotions in acoustic data are contained in human voice. When extracting speech features, it is necessary to remove irrelevant noise and focus on the human voice. Mel scale can be used to divide the sensitivity of human ear to frequency, so in this study Mel-Frequency Cepstral Coefficients (MFCC) [22] are used to extract acoustic features. Specifically, the 40 ms time scale is used to synthesize multiple sampling points of continuous audio signal within the time scale, which is called "frame." Then, the signal is preenhanced through a high-pass filter to compensate for the high frequency part of the speech signal, and Fourier transform is used to transform it from the time domain to the frequency domain to observe the state of energy part. Next, the frequency spectrum obtained by each frame is filtered by Mel filter to remove the frequency information that cannot be distinguished by human ear. After extracting the logarithmic energy on each Mel scale, the inverse discrete Fourier transform is performed. Finally, we can get acoustic features; the vector for it contains 80 dimensions.

4.2.3. Visual Feature. Video contains a lot of extremely abstract visual information, such as human facial expressions, human body movements, and even the color of the

entire video picture. They can play a certain role in emotion recognition. In this study, we use pretrained CNN to extract visual features [23], which uses a two-dimensional convolution kernel to extract spatial information and a one-dimensional convolution kernel to extract temporal information. Because the introduction of visual information brings more noise, we will not show the content of visual information in the follow-up experiments.

4.3. Baselines

4.3.1. Graph-MFN. This method uses a new fusion model called the Dynamic Fusion Graph (DFG) to build the n-modal interactions and replace the original fusion component in MFN [15].

4.3.2. B2 + B4 W/Multimodal Fusion. It utilizes self-attention to capture long term context and gating mechanism to selectively learn cross attended features [16].

4.3.3. Multilogue-Net. The model focuses on effectively capturing the context of a conversation and treats each modality independently, taking into account the information a particular modality is capable of holding [18].

4.3.4. TBJE. The approach relies on a modular coattention and a glimpse layer to jointly encode one or more modalities [19].

4.3.5. Text-CNN. The method achieves excellent results by a simple CNN with little hyperparameter tuning and static vectors [24].

4.3.6. BcLSTM. The model designed based on LSTM can capture the contextual information in the conversation [8].

4.3.7. DialogueRNN. The method is based on recurrent neural networks that keeps track of the individual party states throughout the conversation and uses this information for emotion classification [25].

5. Experiments

In this section, we will report the result on CMU-MOSEI dataset and MELD dataset. It is worth mentioning that the model can achieve good results in the case of only using linguistic feature and acoustic feature.

5.1. Implementation Details. In order to ensure the training speed and training results at the same time, the SWATS optimization method proposed by Keskar and Socher [26] was used in the experiment. Using Adam optimizer in the early stage can bring the advantage of fast convergence. Using SGD optimizer in the later stage can help the model find the optimal solution in a small range. Specifically, when using Adam optimizer, calculate the learning rate of SGD optimizer after each iteration. If it is found that the learning rate basically remains unchanged, it means that the bottleneck has been reached and you can switch at this time. At this time, the orthogonal projection of SGD optimizer on the descending direction of Adam optimizer should be exactly equal to the descending direction of Adam optimizer. It is shown in the following formula:

$$\mathrm{proj}_{\eta t}^{\mathrm{SGD}} = \eta_t^{\mathrm{Adam}}. \tag{11}$$

Therefore, the initial learning rate of SGD optimizer is shown in the following formula:

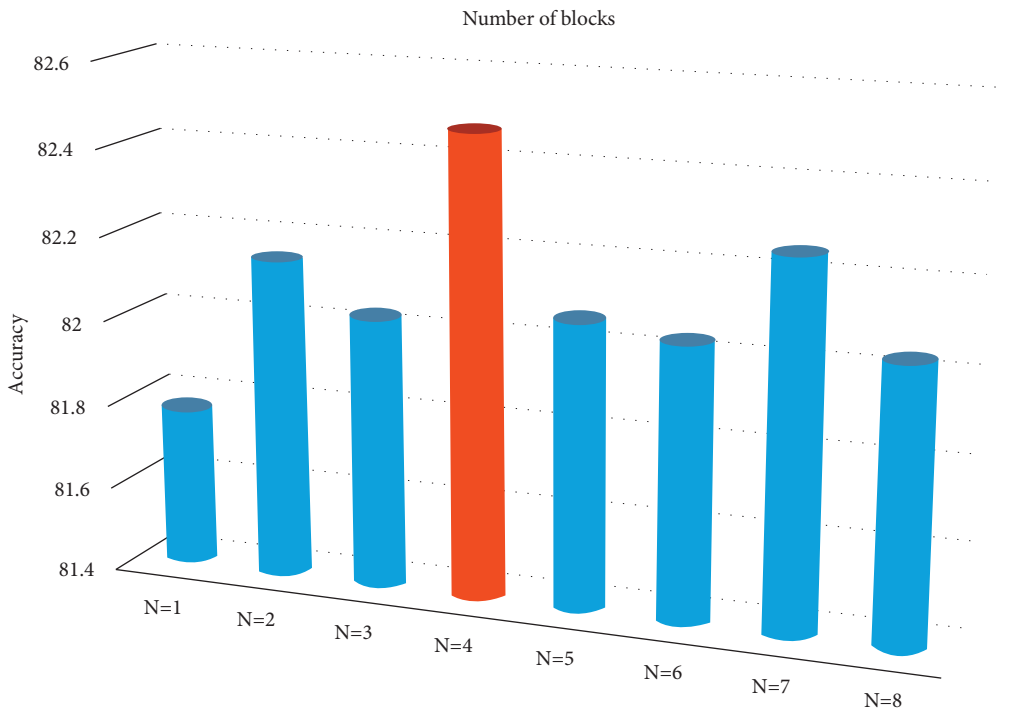
$$\alpha_t^{\mathrm{SGD}} = \frac{\left(\left(\eta_t^{\mathrm{Adam}}\right)^T \eta_t^{\mathrm{Adam}}\right)}{\left(\left(\eta_t^{\mathrm{Adam}}\right)^T\right) g_t}. \tag{12}$$

Then, we found that the number of blocks in interactive transformer directly affects the final result. Considering that the change of structure will bring some influence, we find the best setting of value by comparing the change of 2-class sentient accuracy under different N values. The specific results are shown in Figure 7. The results use five times average value and finally use $n = 4$ to ensure the best performance of the model, and the hidden layer size of each coding block is 512. In addition, we set the Interactive Multihead Guided-Attention structure to 4-head modules and the hidden layer size of Feedforward Neural Networks to 1024. In order to prevent the overfitting phenomenon, we set dropout of 0.1 on the output of each FNN and of 0.5 on the input of classification.

5.2. The Result of CMU-MOSEI Dataset. We compare the evaluation results of the model on CMU-MOSEI dataset with Graph-MFN [14], B2 + B4 w/multimodal fusion [16], Multilogue-Net [18], and TBJE [19]. The results of 2-class-sentiment are shown in Table 1. It should be noted that our model does not use visual information and integrating it into the noise makes the results unsatisfactory, which is also the direction of our next efforts. In general, only using linguistic feature and acoustic feature has been able to make our model achieve good results, which has been improved compared to other methods.

The results of 7-class-sentiment and 6-class-emotion are shown in Table 2. We show the average value of the classification results and only compare with the model using the same calculation method. It can be seen that the effect of the model on emotion multiclassification task is only slightly improved compared with other methods, which will be the direction of our future work. In addition, the performance of 7-class-sentiment classifications is not as good as that of 6-class-emotion classifications, which is in line with the expectation. It is also a common fault of all methods, because 7-class-sentiment deals with more fine-grained classification, with only very subtle differences between each category.

We randomly selected 4662 samples from the test set to verify 2-class-sentiment results and calculated True Positive Rate (TPR) and False Positive Rate (FPR) by using the



	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N=8
Accuracy	81.774	82.148	82.033	82.472	82.067	82.039	82.255	82.042

Figure 7: 2-class-sentiment accuracy according to the number of blocks per Interactive Transformer.

Table 1: 2-class-sentiment of CMU-MOSEI dataset.

Model	2-class-sentiment	
	Accuracy (%)	F1-score (%)
Graph-MFN (T + A + V) [14]	76.90	77.00
B2 + B4 w/multimodal fusion (T + A + V) [16]	81.14	78.53
Multilogue-Net (T + A) [18]	80.18	79.88
Multilogue-Net (T + A + V) [18]	82.10	80.01
TBJE (T + A) [19]	82.30	/
TBJE (T + A + V) [19]	81.50	/
The proposed approach (T + A)	82.47	81.23

Table 2: 7-class-sentiment and 6-class-emotion of CMU-MOSEI dataset.

Model	7-class-sentiment		6-class-emotion	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Graph-MFN (T + A + V) [14]	45.00	/	/	/
TBJE (T + A) [19]	45.36	/	81.48	/
TBJE (T + A + V) [19]	44.40	/	80.68	/
The proposed approach(T + A)	45.58	42.93	81.57	81.16

prediction results of the model and the real labels of the samples, so as to approximate the continuous Receiver Operating Characteristic (ROC) curve. As shown in Figure 8, we can see that our model has excellent performance.

5.3. The Result of MELD Dataset. We compare the evaluation results of the model on MELD dataset with text-CNN [24], bcLSTM [8], DialogueRNN [25], and the weighted F-score shown in Table 3. Neither text-CNN model nor our model uses context information, so the results are not very different. But bcLSTM model and DialogueRNN model using context information have better performance. Horizontal comparison shows that the performance of the model on the two datasets is different. The reason may be that the content of CMU-MOSEI dataset is different from MELD dataset. The CMU-MOSEI dataset is mainly composed of individual monologue scenes. The data of each modality is only related

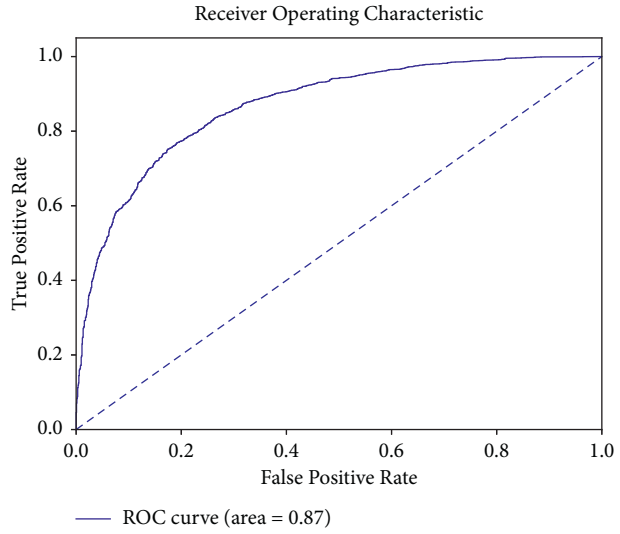
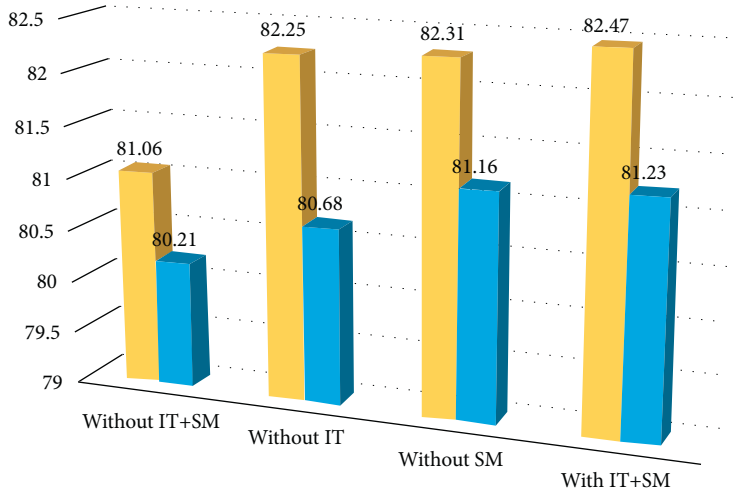


Figure 8: Receiver Operating Characteristic curve of CMU-MOSEI dataset.

Table 3: The result of MELD dataset.

Model	F1-score (%)	
	3 sentiments	7 emotions
Text-CNN (T) [24]	64.25	55.02
BcLSTM (T + A) [8]	66.68	59.25
DialogueRNN (T + A) [25]	67.56	60.25
The proposed approach(T + A)	64.67	55.23



	Without IT+SM	Without IT	Without SM	With IT+SM
Accuracy	81.06	82.25	82.31	82.47
F1-score	80.21	80.68	81.16	81.23

Figure 9: Ablation experiment on CMU-MOSEI dataset.

to the emotion of a single person, and there are no other interference factors. The MELD dataset is mainly composed of multiperson dialogue scenes. Taking acoustic modality data as an example, there may be multiple people talking at the same time, which will cause interference to sentiment analysis. And the emotional state of multiple people will affect each other. When analyzing a person's emotional state, it is necessary to consider the influence of other people. Our model has not been improved for the multiperson dialogue scene, which is the direction of our next work.

5.4. Ablation Study. In this study, we propose two structures: Interactive Transformer (IT) and Soft Mapping (SM). The Interactive Transformer layer interacts with each other when learning a modality to improve the learning effect, while the Soft Mapping layer fuses the learning results of each modality to better classify emotions. In order to verify the effectiveness of these two structures, we have carried out the ablation experiment. The experiment was divided into four situations: the first is completely not using Interactive Transformer layer and Soft Mapping layer, only Interactive Transformer layer is removed, only Soft Mapping layer is removed, and Interactive Transformer layer and Soft Mapping layer are used at the same time.

When we do not use Interactive Transformer, we choose traditional Transformer to replace it, which means that we lose the ability of interaction between modalities. When SM is not used, we directly perform weighted average calculation on the learning results of each modality and then perform sentiment classification. The results are shown in Figure 9. It can be seen that the improvement is obvious when using IT or SM alone, but the improvement is very limited when using both at the same time. The reason may be that the roles of IT and SM are duplicated, and they are both designed for modal interaction. So, when they are used alone, the information between modalities can be complementary, and the results are similar. When using them at the same time, the supplementary information that they mined is repeated, so the improvement is not obvious. In the follow-up, we will try to dig out related information in different directions in a targeted manner to make the division of labor between IT and SM clearer.

6. Conclusions and Future Work

In this paper, we propose an Interactive Transformer and Soft Mapping based method for multimodal sentiment analysis. The proposed model can fully consider the relationship between multiple modality pieces of information, which is helpful for sentiment analysis after data fusion. Although our model has achieved competitive results on the CMU-MOSEI dataset, there are still some shortcomings. Our model does not make full use of the visual modality information, and it only uses data from linguistic modalities and acoustic modalities. Trying to add data from visual modalities makes the results unsatisfactory. In the next step, we will continue to look for the method of integrating visual data, because expression and body movements of characters in the visual data also contain rich and delicate emotions, which can be of great help to emotion recognition. In addition, the evaluation results of MELD dataset also reflect some problems. Our model ignores that people's emotions affect each other in multiperson dialogue scenarios. For example, when a person expresses negative emotions externally, the emotional state of other people will also shift negatively. In the future, we will focus on the emotional analysis of multiperson dialogue from four aspects: role of context, interspeaker influence, emotion shifts, and contextual distance.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethical Approval

This paper does not contain any studies with animals performed by any of the authors.

Conflicts of Interest

The author(s) declare that there are no conflicts of interest with respect to the research, authorship, and/or publication of this paper.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China under Grants 61702462, 62072416, 61873246, and 61771432, the Scientific and Technological Project of Henan Province under Grants 222102210010, 192102210108, 202102210137, 202102210517, and 192102210109, and the Research and Practice Project of Higher Education Teaching Reform in Henan Province under Grants 2019SJGLX320 and 2019SJGLX020.

References

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: from unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[2] Y. Zhang, L. Rong, D. Song, and Z. Peng, "A survey on multimodal sentiment analysis," *Pattern Recognition and Artificial Intelligence*, vol. 33, no. 5, pp. 426–438, 2020.

[3] S. Verma, C. Wang, L. Zhu, and W. Liu, "DeepCU: integrating both common and unique latent information for multimodal sentiment analysis," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3627–3634, Macao, China, August 2019.

[4] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[5] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Copenhagen, Denmark, January 2017.

[6] Z. Sun, P. K. Sarma, W. Sethares, and E. Bucy, "Multi-modal sentiment analysis using deep canonical correlation analysis," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, pp. 1323–1327, Graz, Austria, July 2019.

[7] M. Wöllmer, F. Weninger, T. Knaup, and B. Schuller, "YouTube movie reviews: in, cross, and open-domain sentiment analysis in an audiovisual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[8] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-Dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 873–883, Vancouver, BC, Canada, June 2017.

[9] V. Pérez-Rosas, R. Mihalcea, and L. P. Morency, "Utterance-level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 973–982, Sofia, Bulgaria, August 2013.

[10] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2539–2544, Lisbon, Portugal, July 2015.

[11] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 206–218, 2011.

[12] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pp. 5999–6009, Long Beach, CA, USA, December 2017.

[13] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6281–6290, Long Beach, CA, USA, June 2019.

[14] P. Amir Zadeh, P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 5634–5641, New Orleans, Lousiana, USA, February 2018.

[15] A. A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246, Melbourne, Australia, January 2018.

[16] A. Kumar and J. Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4477–4481, Barcelona, Spain, May 2020.

[17] S. Sahay, S. H. Kumar, R. Xia, and J. Huang, "Multimodal relational tensor network for sentiment and emotion classification," in *Proceedings of the 1st Grand Challenge and Workshop on Human Multimodal Language*, pp. 20–27, Melbourne, Australia, January 2018.

[18] A. Shenoy and A. Sardana, "Multilogue-net: a context aware RNN for multi-modal emotion detection and sentiment analysis in conversation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 19–28, Stroudsburg, PA, USA, February 2020.

[19] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Proceedings of the 2nd Grand Challenge and Workshop on Multimodal Language*, pp. 1–7, Seattle, WA, USA, June 2020.

[20] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: a multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, January 2018.

[21] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, Doha, Qatar, January 2014.

[22] B. Mcfee, C. Raffel, D. Liang et al., "Librosa: audio and music signal analysis in Python," in *Proceedings of the 14th Python in Science Conference*, pp. 18–24, San Jose, CA, USA, August 2015.

[23] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, Salt Lake City, UT, USA, June 2018.

[24] K. Yoon, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1746–1751, Doha, Qatar, August 2014.

[25] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: an attentive RNN for emotion detection in conversations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6818–6825, 2019.

[26] N. S. Keskar and R. Socher, "Improving generalization performance by switching from Adam to SGD," in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, December 2018.