Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8843186, 8 pages
https://doi.org/10.1155/2020/8843186

WILEY | Hindawi

Research Article

Multimodal Fusion Method Based on Self-Attention Mechanism

Hu Zhu,[1] Ze Wang,[2] Yu Shi,[3] Yingying Hua,[1] Guoxia Xu,[4] and Lizhen Deng[5]

[1]*Jiangsu Province Key Lab on Image Processing and Image Communication, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*
[2]*R&D Center, China Academy of Launch Vehicle Technology, Beijing 100176, China*
[3]*Bell Honors School, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*
[4]*Department of Computer Science, Norwegian University of Science and Technology, Gjovik 2815, Norway*
[5]*National Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*

Correspondence should be addressed to Lizhen Deng; alicedenglzh@gmail.com

Received 25 June 2020; Revised 10 August 2020; Accepted 2 September 2020; Published 23 September 2020

Academic Editor: Yin Zhang

Multimodal fusion is one of the popular research directions of multimodal research, and it is also an emerging research field of artificial intelligence. Multimodal fusion is aimed at taking advantage of the complementarity of heterogeneous data and providing reliable classification for the model. Multimodal data fusion is to transform data from multiple single-mode representations to a compact multimodal representation. In previous multimodal data fusion studies, most of the research in this field used multimodal representations of tensors. As the input is converted into a tensor, the dimensions and computational complexity increase exponentially. In this paper, we propose a low-rank tensor multimodal fusion method with an attention mechanism, which improves efficiency and reduces computational complexity. We evaluate our model through three multimodal fusion tasks, which are based on a public data set: CMU-MOSI, IEMOCAP, and POM. Our model achieves a good performance while flexibly capturing the global and local connections. Compared with other multimodal fusions represented by tensors, experiments show that our model can achieve better results steadily under a series of attention mechanisms.

1. Introduction

Multimodal integration has become a popular research direction in the field of artificial intelligence by virtue of its outstanding performance in various applications. Multimodal research has performed well in speech recognition [1], emotion recognition [2, 3], emotion analysis [4], speaker feature analysis [5], and media description [6].

Multimodal fusion is an extremely important research direction and core technology in multimodal field research. Multimodal fusion is aimed at utilizing the complementary information present in multimodal data by combining multiple modalities. It is one of the challenges of multimodal fusion to extend fusion to multimodal while keeping the model and calculation complexity reasonable.

Previous research methods used feature concatenation to fuse different data. These methods [7, 8] take the feature of the input concatenated as input, and some methods [9] even remove the temporal correlation in the modalities. Although these methods have been integrated at the beginning, it is precisely because of this that the interaction within the modal is suppressed at the beginning, causing the modalities to lose its overall correlation or even temporal dependencies.

Some fusion methods [10, 11] use methods such as weighted average or majority voting to fuse modalities together, and these modalities have their own models in later stages. Each of these methods has an inevitable shortcoming. Since each model is modeled separately, the interaction of the modes is lost.

At present, the latest methods [12, 13] try to use tensor representation to model the interactions between modes to solve those shortcomings. The extremely high-dimensional tensor representation caused by various forms of outer products puts a lot of pressure on the calculation

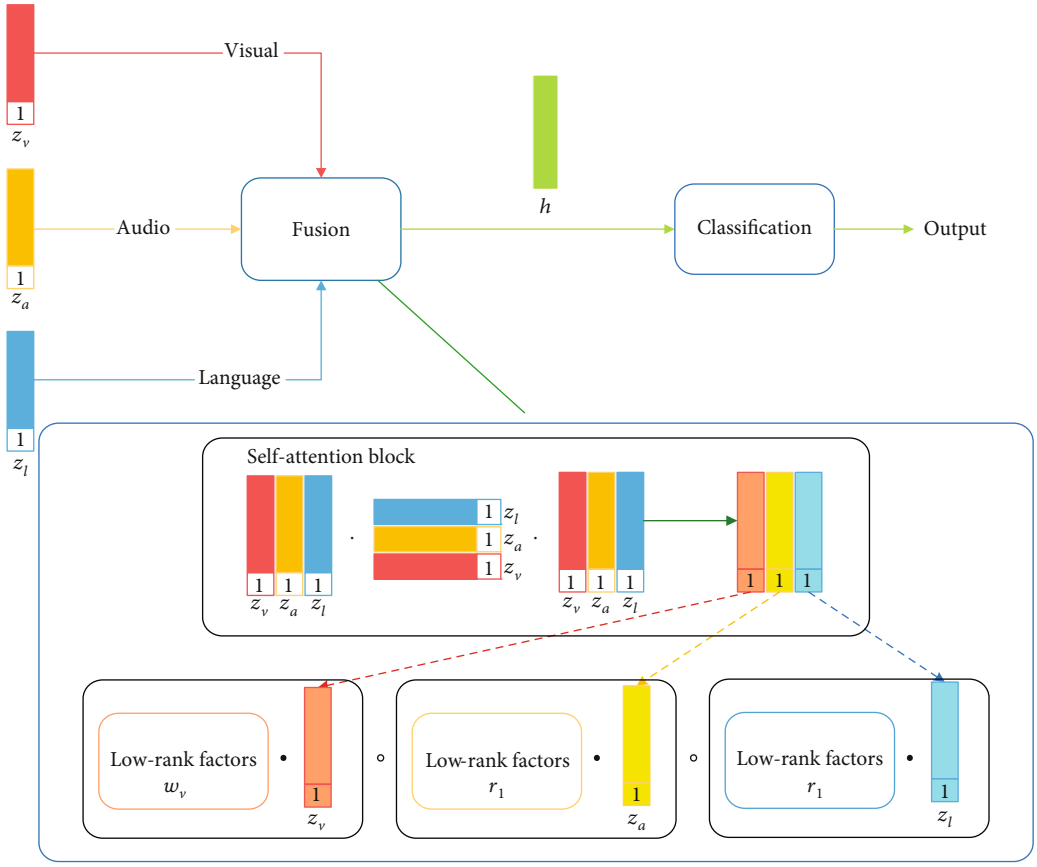

FIGURE 1: Overview of our multimodal fusion model based on self-attention mechanism: the unimodal representations z_v, z_a, and z_l as input to MF (multimodal fusion), which were obtained by passing the unimodal inputs x_v, x_a, and x_l into three subnetworks f_v, f_a, and f_l, respectively. In MF, z_v, z_a, and z_l generate new unimodal representations z_v', z_a', and z_l' through self-attention; then, z_v', z_a', and z_l' produce an output representation by performing low-rank multimodal fusion with modality-specific factors. The output will be multimodal representation, which can be used for applying classification task.

speed and memory occupation. In [14], Liu et al. proposed to use the low-rank multipeak fusion method, which partially solves the problem of large calculation and complicated parameters due to tensor representation but lacks the consideration of the correlation between multiple unimodal inputs.

An attention mechanism has been applied to various fields and has achieved satisfactory results. In [15], Wang et al. proposed "Residual Attention Network," a convolutional neural network using an attention mechanism which can incorporate with the state-of-art feed forward network architecture in an end-to-end training fashion. Lin et al. proposed a novel structure-attention-based LSTM as a hierarchical structure model, which has an advantage in capturing the potential semantic structure. As for applications, Choi et al. [16] proposed a fine-grained attention mechanism for neural machine translation while Ge et al. [17] proposed a leveraged attention mechanism in video action recognition. Hsiao and Chen [18] proposed to integrate the attention mechanism into deep recurrent neural network models for speech emotion recognition. However, none of these previous works aimed at applying an attention mechanism in multimodal fusion.

In this paper, we propose a novel low-rank multipeak fusion model based on a self-attention mechanism, which uses the low-rank weight tensor with an attention mechanism to make multipeak fusion more efficient and more globally relevant. The overall framework of our model is shown in Figure 1. We evaluate the performance of our method through experiments on three multimodal fusion tasks on public data sets and also compare our experiments with the latest models. While reducing the complexity and parameters of the model, we are studying how to improve the applicability and stability of our model. To our knowledge, this is the first time that the self-attention mechanism has been applied to the low-rank factor of multimodal fusion. Compared with other tensor-based models, our model performs very well both in terms of efficiency and performance.

The main contributions of our paper are as follows:

(i) We propose low-rank multimodal fusion based on a self-attention mechanism, which can effectively improve the global correlation

(ii) While maintaining low parameter complexity and high calculation speed, our model has high adaptability and can be applied to various tasks

(iii) We provide the performance of our model on three multimodal tasks evaluated on public data sets compared to other latest models

2. Related Work

2.1. Tensor Representation Method. The tensor representation method is one of the most successful methods for multimodal fusion. The core of tensor representation is to convert the input representation into a high-dimensional tensor, and then map it to a lower-dimensional output vector space. Tensors are usually formed by multiplying the outer product by the input modality. The input tensor Z is calculated from the unimodal representation:

$$Z = \otimes_{n=1}^{N} z_n, z_n \in \mathbb{R}^{d_n}, \tag{1}$$

where $\otimes_{n=1}^{N}$ denotes the tensor outer product over a set of vectors indexed by n, and z_n is the input representation.

The input tensor $Z \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ uses a linear layer $f(\bullet)$ to generate a vector representation:

$$h = f(Z; W; b) = W \cdot Z + b, h, b \in \mathbb{R}^{y}, \tag{2}$$

where W denotes the weight of this layer and b represents the bias. Because Z is an N-order tensor, where N is the number of input modes, the weight W should be an $N+1$-order tensor. The dimension of the weight W is $W \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N \times d_h}$, where the $N+1$th dimension is equal to the size of the output representation d_h. Since $W \cdot Z$ is a dot product, the weight W can be regarded as $d_h N$th order tensor.

Due to the high dimension of tensor Z, the computational difficulty and model complexity of tensor fusion method are greatly improved. The dimension of tensor Z increases exponentially with the number of modes. This makes the tensor fusion method fail to perform more tasks at the same time, which reduces the adaptability of the model.

2.2. Low-Rank Tensor Representation Method. The low-rank multimodal fusion method is aimed at solving the shortcomings of the multimodal fusion model represented by a tensor with the method of decomposing the weight W into a set of low-rank factors.

The method of degrading the weights in the multimodal fusion method represented by the low-rank tensor is to decompose the weight W into N fixed modalities. Because $W \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N \times d_h}$ can be regarded as $d_h N$th order tensor, so our weight can be expressed as follows:

$$\widetilde{W_m} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}, m = 1, \cdots, d_h. \tag{3}$$

A deep understanding of formula (3), so $\widetilde{W_m}$ has the following exact decomposition of the vector in Equation (5). Each $\widetilde{W_m}$ contributes to one dimension in the vector h, so we can simplify Equation (2):

$$h_m = \widetilde{W_m} \cdot Z, \tag{4}$$

$$\widetilde{W_m} = \sum_{i=1}^{R} \otimes_{n=1}^{N} w_{n,m}^{(i)}, w_{n,m}^{(i)} \in \mathbb{R}_n^d, \tag{5}$$

where R is the rank of the tensor, which makes the decomposition most efficient. $\{\{w_{n,m}^{(i)}\}_{n=1}^{N}\}_{i=1}^{R}$ is the decomposition factor of the original weight tensor based on rank R.

In the above formula $\{\{w_{n,m}^{(i)}\}_{n=1}^{N}\}_{i=1}^{R}$, we give the rank R a fixed value r, and the formula $\{\{w_{n,m}^{(i)}\}_{n=1}^{N}\}_{i=1}^{r}$ can be decomposed by the fixed rank, and the model is parameterized at the same time. We expand the vector $w_{n,m}^{(i)}$ by m (where $= 1, \cdots, d_h$) into a set of low-rank factors $w_n^{(i)} = [w_{n,1}^{(i)}, w_{n,2}^{(i)}, \cdots w_{n,dh}^{(i)}]$, so the $\{w_n^{(i)}\}_{i=1}^{r}$ is its corresponding low-rank factors. Therefore, the weights of the multimodal fusion method represented by tensor can be transformed into low-rank weight tensor:

$$W = \sum_{i=1}^{r} \otimes_{n=1}^{N} w_n^{(i)}. \tag{6}$$

Bring Equation (6) into Equation (2) to get the following a simplified low-rank tensor representation:

$$\begin{aligned} h = W \cdot Z &= \left(\sum_{i=1}^{r} \otimes_{n=1}^{N} w_n^{(i)} \right) \cdot \left(\otimes_{n=1}^{N} z_n \right) \\ &= \sum_{i=1}^{r} \left(\otimes_{n=1}^{N} w_n^{(i)} \cdot \otimes_{n=1}^{N} z_n \right) = \bigwedge_{n=1}^{N} \left(\sum_{i=1}^{r} w_n^{(i)} \cdot z_n \right). \end{aligned} \tag{7}$$

We made a series of derivation changes in the above formula and finally turned the model calculated from an exponentially complex model into a linear model, where $\bigwedge_{n=1}^{N} x_n$ denotes the product of elements in the order of tensors: $\bigwedge_{n=1}^{N} x_n = x_1 \circ x_2 \circ \cdots \circ x_N$.

Compared with the original tensor representation method, the low-rank multimodal fusion method improves the calculation speed and reduces the complexity of the model. However, only a simple outer product operation is performed for each single mode, which largely ignores the correlation between each single mode and loses the global uniformity.

2.3. Attention Mechanism. Neural networks equipped with attention have parallelizable computation, lightweight structure, and the ability to capture both long-range and local dependencies. The core of the attention mechanism method is to measure the correlation between z_n and q. A compatibility function $g(z_n, q)$ generates score k, which can reflect the dependency between z_n and q. The score is converted into a probability by function softmax, and finally, the probability is used as a weight.

$$k = [g(z_n, q)]_{n=1}^{N}, \tag{8}$$

$$p(y|z, q) = \text{softmax}(k), \tag{9}$$

$$s = \sum_{n=1}^{N} p(y = n|z, q) \cdot z_n, \tag{10}$$

where k is represented as a vector of n correlation scores. By applying k to the function softmax, we get a probability distribution about attention $p(y|z, q)$. And s is the output vector for query q.

In the attention mechanism, choosing different compatibility functions $g(z_n, q)$ will have different experimental results. The different compatibility functions also directly lead to various categories of attention mechanisms. In this paper, the attention mechanism of our method uses the dot product attention compatibility function as follows:

$$g(z_n, q) = \left\langle w^{d_1} z_n, w^{d_2} q \right\rangle, \tag{11}$$

where w^{d_1}, w^{d_2} are learnable parameters, $\langle \bullet, \bullet \rangle$ denotes the inner product.

3. Our Methods

3.1. Overview. The method proposed in this paper is an improvement to the low-rank multimodal fusion method and an effective improvement to the input modal based on the low-rank multimodal fusion method. We propose a novel self-attention mechanism and apply it between input modalities to improve the correlation and local dependence among various modalities. We pay more attention to the improvement of the self-mode, so we choose to use the self-attention mechanism instead of the traditional attention mechanism model. Since our model does not introduce redundant parameters, our model maintains a low complexity while improving accuracy. In addition, our self-attention module uses parallel computing, which makes the calculation speed greatly improved compared with the traditional attention mechanism model. Compared with the model using traditional attention mechanism, our model has lower complexity and faster running speed.

3.2. Network Architecture. The overall framework of our network model is shown in Figure 1. Our model network is composed of three parts, namely, the extraction module, fusion module, and classification module. The fusion module is the core part of our model, which is what we will focus on next. The task of the extraction module is to transform the unimodal inputs x_v, x_a, and x_l into unimodal representation s z_v, z_a, and z_l through the subnetworks f_v, f_a, and f_l. The unimodal representation obtained by the extraction module is expressed in the form of tensor, which is more convenient for the following calculation. And the fusion module contains a self-attention module for each unimodal representation. The unimodal representation enters the fusion module and generates a unimodal representation with new weights through a self-attention mechanism. Observing our network model, we do not need to directly calculate the input tensor Z, we first decompose z_v, z_a, and z_1 in low rank to get z_v, z_a, and z_1,then assign the corresponding weights W_v, W_a, W_1 to each factor, and finally sum them with the weights, which greatly reduces the complexity of our model and reduces the calculation pressure. Finally, the input tensor passing through the self-attention module generates the output tensor in the fusion module, which is the final output result that can be used for classification.

3.3. Self-Attention Module. Since each unimodal has different information, the purpose of multimodal fusion is to make full use of the complementary information of multimodal data. We note that the self-attention module also has the ability to capture the global and local connections, so the most prominent part of our contribution in this article is to propose the introduction of the self-attention module into multimodal fusion. In the self-attention module, we use a different output vector calculation method than the traditional attention mechanism. This new method can perfectly meet the requirements of our simultaneous input of multiple tasks and realize parallel computing. The self-attention model we proposed is a weighted self-attention in proportion, which includes multitask self-attention. Our self-attention model formula is as follows:

$$s = v\text{softmax}\left(\frac{q^T k}{\sqrt{d_q}}\right)^T, s = [s_1, s_2, \cdots, s_n] \in \mathbb{R}^{d_{i\times n}}. \tag{12}$$

The three parameters q, k, and v of Equation (12) all conform to this equation $q, k, v = \varphi^{q,k,v}(z_n)$, which means that all three input parameters come from the same source, where $v \in \mathbb{R}^{d_{i\times n}}, k \in \mathbb{R}^{d_{n\times i}}, q \in \mathbb{R}^{d_{i\times n}}$. For the multitask attention mechanism, the input will be projected into multiple subspaces. This parameter uniformly scales the dot product attention to be embedded in each subspace.

Since $s = [s_1, s_2, \cdots, s_n] \in \mathbb{R}^{d_{i\times n}}$ is a series of output vectors of q, k, and v, therefore, s is a series of output vectors of z_n, we derive the following equation:

$$z'_n = s_i, z'_n \in \mathbb{R}^{d_n}, \tag{13}$$

where z'_n is the new unimodal representation generated by the self-attention module. In this way, the self-attention between our single modes is completed. Bring Equation (13) into Equation (7) and simplify it as follows:

$$h = W \cdot Z = \left(\sum_{i=1}^{r} \otimes_{n=1}^{N} w_n^{(i)}\right) \cdot \left(\otimes_{n=1}^{N} z'_n\right) = \bigwedge_{n=1}^{N} \left(\sum_{i=1}^{r} w_n^{(i)} \cdot z'_n\right). \tag{14}$$

It can be seen from formula (14) that the formula is consistent with the model we have shown. First, superimpose each weighting factor, then do element product between each single module.

Since we are merging multiple tasks at the same time, we will show below that when $n = 2$, our formula will expand to formula (15):

$$h = \bigwedge_{n=1}^{2}\left(\left(\sum_{i=1}^{r} w_1^{(i)} \otimes w_2^{(i)}\right) \cdot z_n'\right) = \left(\sum_{i=1}^{r} w_1^{(i)} \cdot z_1'\right) \circ \left(\sum_{i=1}^{r} w_2^{(i)} \cdot z_2'\right). \tag{15}$$

In this way, we can appropriately expand the formula according to the actual situation. It can be seen that the proposed method has high adaptability and can be flexibly applied in various tasks. In our self-attention module, we can see that our new input representation represents multiple tasks applied to multimodal fusion. And our self-attention module uses parallel computing to improve the accuracy of the model while maintaining a high speed of model calculation.

3.4. Training Loss. Our model adopts the mean absolute error (MAE) as our loss function. MAE is the average value of absolute error, which can better reflect the actual situation of classification error and can also reflect the classification performance of our model.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left|h_i - h_i^P\right|}{n}, \tag{16}$$

where h_i is the output tensor we got through our model and h_i^P is the classified value. n represents the total number of our training samples.

4. Experiment

4.1. Experimental Environment. Our tensor representation method is generally based on a tensor fusion network, but the biggest difference from this network is that our method uses a self-attention mechanism in the MF module. In the experiment, we compared our method with some of the latest multimodal fusion methods. Our experiment environment is 2080Ti*2 Graphic Processing Unit (GPU), 32 G memory, 12 Intel(R) Xeon(R) W-2133 CPU @ 3.60 GHz. Our model training and testing are completed on CONDA 4.8.3, python3.7.7, and pytorch 1.5.0.

4.2. Data sets. We conduct our experiments on multimodal data sets; they are CMU-MOSI [19], IEMOCAP [20], and POM [6]. These data sets provide data for sentiment analysis, speaker feature recognition, and emotion recognition. The goal of our experiment is to identify the speaker's emotions through these verbal or nonverbal behaviors.

4.2.1. IEMOCAP. This IEMOCAP data set is a collection of 151 recorded dialogue videos; each dialogue video has two speakers, so the entire data set has a total of 302 videos. Each video is marked with 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral).

4.2.2. POM. The POM data set consists of 903 movie review videos, and the speakers of each video are marked with confidence, enthusiasm, and other characteristics.

4.2.3. CMU-MOSI. 93 movie review videos on YouTube make up the data set CMU-MOSI. Each video contains multiple opinion segments, and each segment has tags about the opinion on sentiment.

4.3. Comparisons. We compare the results of our method with the following baselines and state-of-the-art models: support vector machines (SVM) [21], deep fusion model (DF) [11], bidirectional contextual LSTM (BC-LSTM) [13], multi-view LSTM (MVLSTM) [22], low-rank multimodal fusion network (LMF) [23], and hierarchical polynomial fusion network (HPFN) [24]. We report mean absolute error (MAE) and accuracy of classification, F1 score, and Pearson's correlation (Corr).

4.4. Evaluation Metrics. We report four evaluation metrics as used by our multiple task: F1-emotion, accuracy Acc-k where k is the number of classes, mean absolute error (MAE), and Pearson's correlation (Corr). Among those metrics, F1-emotion is the score of the model under different emotions; as a statistical measure of the accuracy of a binary classification model, it can be viewed as a weighted average of the model accuracy and recall with a maximum of 1and a minimum of 0. Accuracy (Acc) is defined as the percentage of the total sample that classifies the correct result. Mean absolute error (MAE) reflects the classification performance as we reported before. Pearson's correlation (Corr) considers the degree of correlation among variables.

4.5. Implementation Details. We use the CANDECOMP/-PARAFAC(CP) decomposition format as the tensor networks in our experiments as we mentioned in Equation (6). Following LMF, we choose the candidate CP ranks {1, 4, 8, 16}.The result in different ranks are reported in Figure 2.

4.6. Experimental Data Analysis. We compare our method's performance on the three tasks of sentiment analysis, speaker feature recognition, and sentiment recognition with the previous models with excellent performance. The results are shown in Table 1. In all data sets, our approach can produce competitive and consistent results across metrics such as F1, Corr, Acc, and MAE.

On the emotion recognition task, our model got the highest score on three emotions scored by F1. The results verify that our method outperforms other traditional method and is close to the state-of-the-art approaches.

On the multimodal personality trait recognition task, our model also achieved competitive results. Although LMF achieved a high score on the ACC indicator, our score is only 0.1 less than the LMF score.

On the multimodal sentiment analysis task, our model performs very well on performance indicators Corr and Acc-2. Nonetheless, our method scored only 0.057 less than the highest on MAE. All in all, our method perfectly completes the multimodal sentiment regression task.

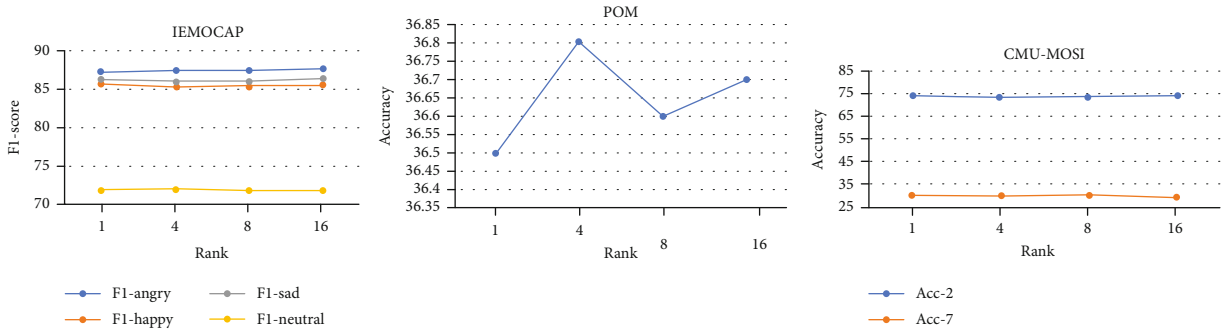

Figure 2: Results for recognition in different rank on IEMPCAP, POM, and CMU-MOSI.

Table 1: Results for emotion recognition on IEMPCAP, personality trait recognition on POM, and sentiment analysis on CMU-MOSI.

Data set	IEMOCAP				POM			CMU-MOSI				
Metric	F1-happy	F1-sad	F1-angry	F1-neutral	MAE	Corr	Acc	MAE	Corr	Acc-2	F1	Acc-7
SVM	81.5	78.8	82.4	64.9	0.887	10.4	33.9	1.864	0.057	50.2	50.1	17.5
DF	81.0	81.2	65.4	44.0	0.869	14.4	34.1	1.143	0.518	72.3	72.1	26.8
BC-LSTM	81.7	81.7	84.2	64.1	0.840	27.8	34.8	1.079	0.581	73.9	73.9	28.7
MV-LSTM	81.3	74.0	84.3	66.7	0.891	27.0	34.6	1.019	0.601	73.9	*74.0*	33.2
LMF	85.2	85.8	87.4	71.7	0.837	32.3	*36.8*	1.071	0.571	73.3	73.3	30.0
HPFN	*85.7*	86.2	*87.8*	*71.9*	0.840	35.6	36.7	*0.975*	0.601	73.0	73.1	*35.1*
Our methods	*85.7*	*86.3*	87.6	*71.9*	*0.834*	*52.3*	36.7	1.032	*0.610*	*74.1*	73.0	30.5

Best results are italicized.

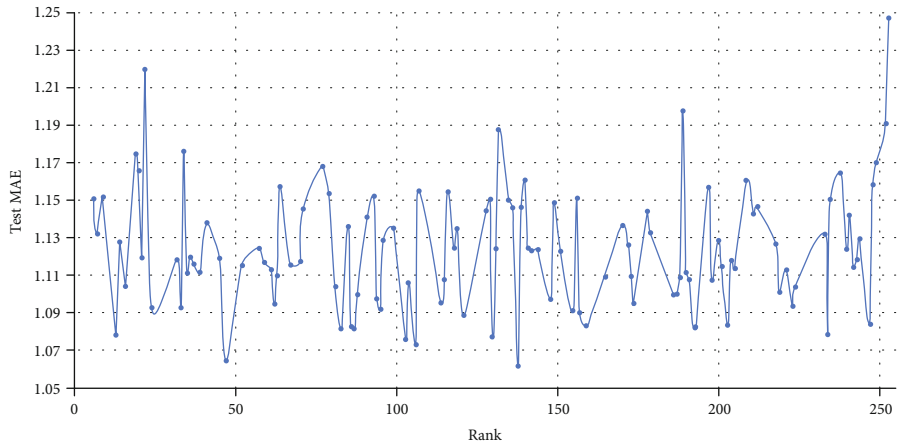

Figure 3: The effect of different level settings on the model performance: as the level increases, the results do not change significantly.

4.7. Influence of Rank Setting. In the experiment, the parameters changed by the actual situation often have a great influence on the experimental results. The different rank settings in our model will indeed affect the experimental results. In order to prove that our model can stand out in various tasks and has a high adaptability, we propose a new experiment, in which, we constantly set different values for rank to observe the effect of changes in rank on the experimental results.

In the experiment of the influence of rank on the experimental results, our other parameters are set as follows: the dropout of audio and video are both set to 0.2, and the text dropout is set to 0.5. For some other parameters, the learning rate is set to 0.001, batch size is set to 32, and the weight decay is set to 0.01.

To evaluate the impact of different level settings on our model, we measured the performance change of MAE in the CMU-MOSI data set while changing the number of levels. The results are shown in Figure 3. We have observed that although the rank value is constantly increasing, our training results have remained stable. Therefore, it can be seen that our model is not sensitive to rank, no matter what the rank is, the performance of our model can always remain stable. In some cases where the rank value is high, our model can still be adapted and used.

5. Conclusion

In this paper, we propose a multipeak fusion method based on a self-attention mechanism. This method uses a low-rank tensor representation, and the attention mechanism is used in tensor representation to improve the correlation between multiple representations. Our method achieves competitive results in different multimodal fusion tasks in different data sets. Our method reduces the complexity of the parameters while also reducing the measurement complexity. It is a novel attempt to apply the attention mechanism to multimodal fusion, and it shows higher efficiency and better performance on different downstream tasks. The application of the attention mechanism makes our model have higher classified ability under the premise of few parameters and high efficiency. In the experiment, our method performs better than the multimodal fusion method represented only by low-rank tensor.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61701259.

References

[1] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 1989.

[2] S. L. Chen, S. T. Huang, M. Tsutomu, and N. Ryohei, "Multimodal human emotion/expression recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, April 1998.

[3] W. Weninger and K. Schuller, "Youtube movie reviews: in, cross, and open-domain sentiment analysis in an audiovisual context," *Asbury Theological Seminary*, vol. 28, pp. 46–53, 2013.

[4] L. C. D. Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. No.97TH8237)*, Singapore, Singapore, September 1997.

[5] Y. Attabi and P. Dumouchel, "Anchor models and WCCN normalization for speaker trait classification," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, pp. 522–525, Portland, OR, USA, September 2012.

[6] S. Park, H. Shim, M. Chatterjee, K. Sagae, and L. Morency, "Computational analysis of persuasiveness in social multimedia," in *ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 50–57, New York, NY, USA, November 2014.

[7] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain, December 2016.

[8] H. Wang, A. Meghawat, L. Morency, and E. Xing, "Select-additive learning: improving cross-individual generalization in multimodal sentiment analysis," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, China, 2016.

[9] L. P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: harvesting opinions from the web," in *ICMI '11: Proceedings of the 13th international conference on multimodal interfaces*, New York, NY, USA, November 2011.

[10] T. Wortwein and S. Scherer, "What really matters — an information gain analysis of questions and reactions in automated PTSD screenings," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 15–20, San Antonio, TX, USA, October 2017.

[11] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrusaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *ICMI '16: Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, October 2016.

[12] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, http://arxiv.org/1606.01847.

[13] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, http://arxiv.org/1707.07250.

[14] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, and L. P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018.

[15] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6450–6458, Honolulu, HI, USA, 2017.

[16] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, 2018.

[17] H. Ge, Z. Yan, W. Yu, and L. Sun, "An attention mechanism based convolutional lstm network for video action recognition," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 20533–20556, 2019.

[18] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2526–2530, Calgary, AB, Canada, April 2018.

[19] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[20] C. Busso, M. Bulut, C.-C. Lee et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] S. S. Rajagopalan, L. P. Morency, T. Baltrušaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *European Conference on Computer Vision*, Amsterdam, Netherlands, 2016.

[23] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, http://arxiv.org/1806.00064.

[24] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019.