



# Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling

Xueming Yan, Haiwei Xue, Shengyi Jiang & Ziang Liu

To cite this article: Xueming Yan, Haiwei Xue, Shengyi Jiang & Ziang Liu (2022) Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling, Applied Artificial Intelligence, 36:1, 2000688, DOI: [10.1080/08839514.2021.2000688](https://doi.org/10.1080/08839514.2021.2000688)

To link to this article: <https://doi.org/10.1080/08839514.2021.2000688>



© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 19 Nov 2021.



Submit your article to this journal [↗](#)



Article views: 3615



View related articles [↗](#)





View Crossmark data [↗](#)



Citing articles: 4 View citing articles [↗](#)

# Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling

Xueming Yan <sup>a</sup>, Haiwei Xue<sup>b</sup>, Shengyi Jiang <sup>a</sup>, and Ziang Liu<sup>c</sup>

<sup>a</sup>Guangzhou Key Laboratory of Multilingual Intelligent Processing & School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China; <sup>b</sup>School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China; <sup>c</sup>Faculty of Science, University of Alberta, Edmonton, Canada

## ABSTRACT


With the rapid development of social networks, more and more people express their emotions and opinions via online videos. However, most of the current research on multimodal sentiment analysis cannot do well with effective emotional fusion in multimodal data. To deal with the problem, we propose a multi-tensor fusion network with cross-modal modeling for multimodal sentiment analysis. In this study, the multimodal feature extraction with cross-modal modeling is utilized to obtain the relationship of emotional information between multiple modalities. Moreover, the multi-tensor fusion network is used to model the interaction of multiple pairs of bimodal and realize the emotional prediction of multimodal features. The proposed approach performs well in regression and different dimensions of classification experiments on the two public datasets CMU-MOSI and CMU-MOSEI.

## ARTICLE HISTORY

Received 8 July 2021  
Revised 18 October 2021  
Accepted 20 October 2021

## Introduction

With the popularity of social media, a short video has gradually become a mainstream form to convey personal feelings and opinions. These videos contain abundant image, audio, and text features, thus vividly conveying user sentiments. The information between the various modalities can be used to solve the problem of the scarcity of short text features through mutual auxiliary reference, and more accurate sentiment prediction results can be obtained. For example, the language content of the speaker may be ambiguous, but more precise sentiments can be acquired by combining facial expressions and voice tone, which can reduce the ambiguity caused by a single modality. Compared with unimodal emotion analysis, multimodal sentiment analysis can better perceive human affections through the combination of intonation, gesture, and micro-expression Liu and Cheng (2018) Patel, Hong, and Zhao (2016).

**CONTACT** Shengyi Jiang  [xueming126@126.com](mailto:xueming126@126.com)  Guangzhou Key Laboratory of Multilingual Intelligent Processing & School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510000, China

© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.  
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To model and integrate the multimodal features efficiently, most researches focus on some multimodal fusion approaches, which can be divided into fusion approaches based on concatenation features and fusion approaches on non-concatenation features. The multimodal features are spliced end to end to obtain the fusion result in concatenation-based approaches, but the simple multimodal concatenated features may cause the loss of the dynamic correlation information in sentiment analysis Kumar and Vepa (2020). Non-concatenation-based approaches can combine dynamic multimodal features and perform decision fusion in sentiment classification Chaturvedi et al. (2019). Some researchers Zadeh et al. (2017) proposed a tensor fusion network that can capture multimodal feature associations in the form of tensors and store modal dynamic information to achieve feature fusion between different modalities. However, traditional tensor fusion networks have the drawback of insufficient feature extraction and poor modal interaction capabilities Sahay et al. (2018).

In order to solve the problem, we propose a multi-tensor fusion network with cross-modal modeling (MTFN-CMM), which can predict more reasonable emotional intensity by enhancing the multi-modal feature fusion by obtaining more dynamic information of cross-modality. Initially, the textual, audio, and visual features are extracted from each utterance in the video separately, and then the context representation of multimodal features can be extracted with cross-modal modeling. Every two modalities are matched as a pair of cross-attention values, and three bimodal cross-attention values are selected to obtain the fine-grained bimodal relationship features. Further, the tensor fusion network with cross-modal modeling makes full use of the dynamic information of intra-modal relational and inter-modal interaction and obtains more accurate emotion prediction. The main contributions of the proposed approach are as follows:

- We proposed a multi-tensor fusion network with cross-modal modeling for multimodal sentiment analysis. Cross-modal modeling is used to extract the interaction relationship between the modalities, and then the dynamic information of cross-modality is fully utilized with a multi-tensor fusion network. The proposed multimodal features fusion approach can retain more cross-correlation information of multimodal and make the emotion prediction more accurate.
- Different from other fusion approaches Zadeh et al. (Zadeh, et al., 2018a), Poria et al. (2017), Li et al. (2021) and Xue et al. (2020), the proposed approach focuses on the fusion of the intra-modal and cross-modal features together rather than the concatenation of three unimodal features or the fusion of multiple bimodal features. To the best of our knowledge, the work is the first to represent non-concatenation multi-level cross-modal fusion with multi-tensor fusion network.

- We evaluate the proposed approach with a series of regression and classification experiments on the two public datasets CMU-MOSI and CMU-MOSEI. The experiment results show that our MTFN-HA approach outperforms other baseline approaches for multi-modal sentiment analysis on a series of regression and classification tasks.

The remainder of the paper is organized as follows: [Section 2](#) is a brief introduction of the related work. [Section 3](#) describes multi-tensor fusion network with cross-modal modeling for multimodal sentiment analysis in detail. [Section 4](#) provides our experiment results in comparison to recent representative methods and discussion. Finally, we conclude this paper in [Section 5](#).

## Related Work

With the spread of social media and short videos, multimodal sentiment analysis has an important impact on some fields, such as human communication comprehension Zadeh et al. (Zadeh, et al., 2018b), financial and political forecasting Xing, Cambria, and Welsch (2018), Ebrahimi, Yazdavar, and Sheth (2017), community detection Young et al. (2018), dialog systems Majumder et al. (2018), Zadeh et al. (2016), and so on. Unlike unimodal sentiment analysis, multimodal sentiment analysis needs to better perceive human emotions through a variety of ways such as intonation, gestures, and micro-expressions. Because the fusion of multimodal features makes multimodal sentiment analysis more complicated, it is necessary to comprehensively consider the intra-modal and intermodal dynamics in sentiment analysis Poria et al. (2016).

Most of the previous works focused on multimodal feature fusion for multimodal sentiment analysis. Concatenation-based feature fusion approaches are simply to concatenate multimodal features at the input section. Some researchers concatenate the multimodal features as an input for the concerned learning model. Zhou et al. (2016) proposed the Long-Short Term Hybrid Memory and Multi-head Attention mechanism to capture the most important semantic features. Kumar et al. Kumar and Vepa (2020) introduced a gated mechanism for attention in the binary classification of emotion tasks to reach higher accuracy. Multi-Head Attention Zadeh et al. (Zadeh, et al., 2018b) and Context-aware Interactive Attention Chauhan et al. (2019) are also applied to improve the sentiment intensity prediction for emotion classification. Besides, Li et al. (2021) introduced a hash feature network structure to concatenate and fuse the multimodal features, and the predicted sentiment label is inferred by calculating the hash value similarity. Although the concatenation-based feature fusion approaches can obtain the relevant features between different modalities with low complexity to a certain extent, it will lead to the loss of the high-dimensional multi-modal relationship information.

Some works on non-concatenated features fusion occur in recent years. Zadeh et al. (Zadeh, et al., 2018a) and Poria et al. (2017) propose different attention mechanisms separately to fuse the instantaneous features of multiple modalities in the perspective of time sequence. Chaturvedi et al. (2019) used deep learning to extract features from each modality and then projected them to a common effective space to speed up the classification performance. Tsai et al. (2019) introduced the multi-modal transformer to reduce the long-range dependencies between elements across modalities and capture the correlated cross-modal signals. In addition, tensor fusion is also a typical non-concatenated features fusion approach, which can quickly build a spatial relationship model from different source features. Zadeh et al. (2017) regarded the problem of multimodal sentiment analysis as modeling intramodal and intermodal dynamics and introduced a novel tensor fusion network that can combine multimodal dynamics information. Motivated by the work of Zadeh et al. (2017), Sahay et al. (2018) presented relational tensor network architecture to model the intermodal interactions for the sequence of segments in a video. The relational tensor network is regarded as a generalization of tensor fusion with multiple Bi-LSTM for multimodalities and an n-fold Cartesian product from modality embedding. These approaches can also fuse different modal features and can retain as much multimodal feature relationship information as possible, but it is easy to cause high-dimensional feature sparsity problems in the cross-modal fusion process.

As current multimodal sentiment analysis approaches are insufficiently prepared for feature extraction and modalities interaction to a certain extent, they might cause the loss of the critical emotional features and decrease the effectiveness of multimodal fusion due to noise interference in some modal features. Our work is related to a variant of the tensor method to enhance the modal feature fusion, which proposed the multi-tensor fusion network with cross-modal modeling to acquire more intramodal relational information and intermodal interaction and affects the inference effect of the final network model for multimodal sentiment analysis.

### **Multi-tensor Fusion Network with Cross-modal Modeling**

The multi-tensor fusion network with cross-modal modeling (MTFN-CMM) for multimodal sentiment analysis consists of three main components: extracting context-independent unimodal features, multimodal feature extraction with cross-modal modeling, and multi-tensor fusion network for the prediction of multimodal emotional intensity. To enhance the cross-modal feature fusion, the Bi-LSTM network and cross-attention mechanism are separately used to capture more intramodal relational information and intermodal

interaction, and a multi-level tensor fusion network is utilized to enhance the ability to acquire cross-modal features, improve the inference accuracy of the final result. We describe these components in detail as follows.

### **Extracting Context-Independent Unimodal Features**

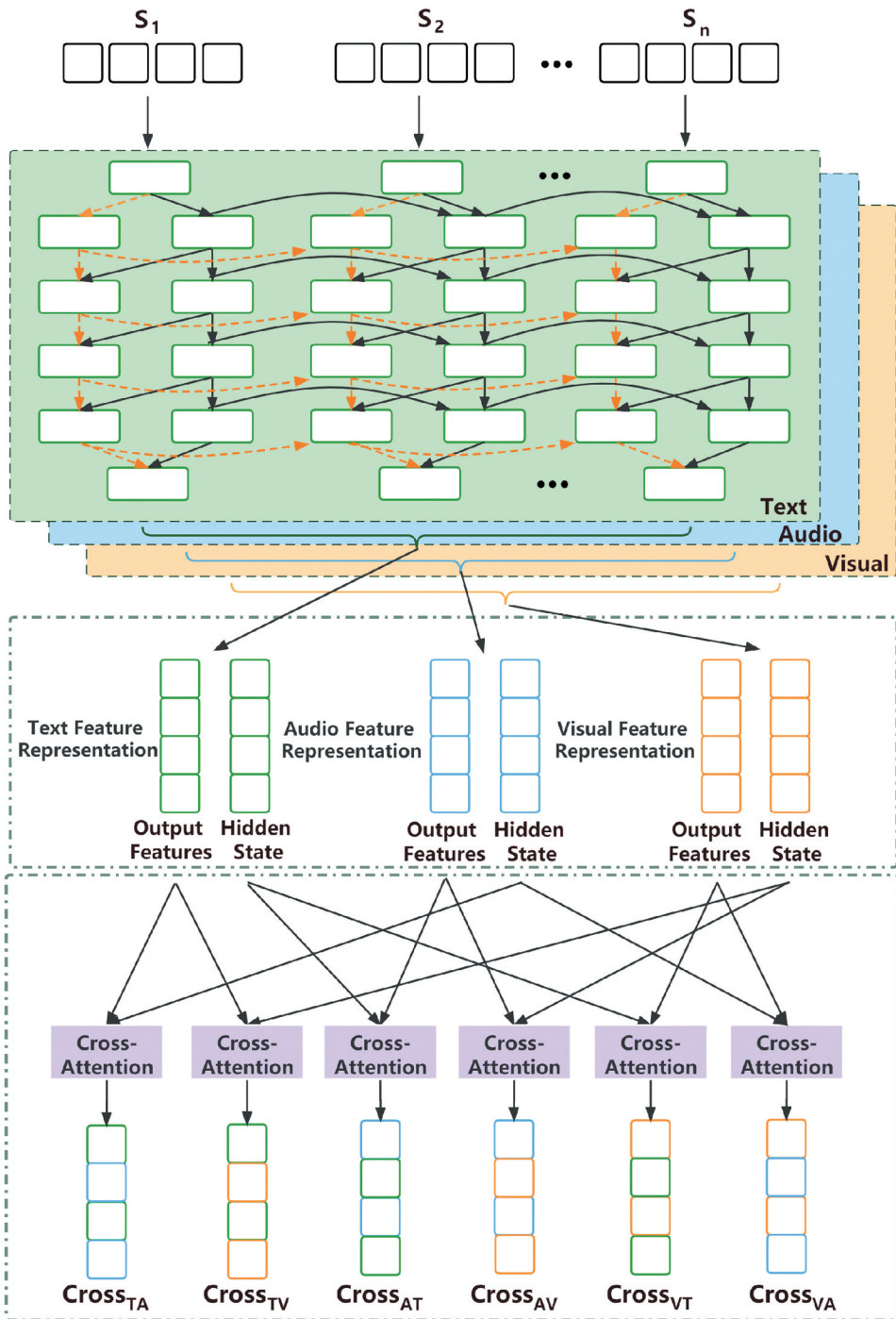
The unimodal features (the textual, audio, and visual features) are extracted from each utterance in the video separately. Initially, we use the FACET to extract the user's facial features in the video, the key feature of the face is extracted as a structured feature vector, which can be used to represent the basic facial features of affection of each frame and highlight the affected facial features. For audio, we use the COVAREP to extract low- and medium-level acoustic features. This tool is usually able to extract rich voice features including 12 Mel-Frequency Cepstral Coefficients (MFCCs), voice segment features, peak slope parameters, and maximum dispersion coefficients. For text, we directly use the Glove Pennington, Socher, and Manning (2014) to quickly obtain the word vector feature. Through the feature extraction work of these open-source tools, the unstructured video data can be transformed into structured vector features. Because the extracted feature dimensions are usually different and the timing between multiple modalities is not aligned, it is necessary to perform modal preprocessing timing alignment at the word level by way of manual annotation, and the alignment result can be input into the network framework in the next section.

### **Multimodal Feature Extraction with Cross-modal Modeling**

Figure 1 presents the process of multimodal feature extraction with cross-modal modeling. Firstly, we capture the long-distance context representation of the intra-modality related information effectively by Bi-LSTM network Plank, Søggaard, and Goldberg (2016), and then every two modalities can be embedded into a single value matrix of attention mechanism to enhance the salient features in the bimodal interaction and weaken the irrelevant features.

In order to capture the long-distance context representation of intra-modality effectively, we put the extracted features vector of each modality into the Bi-LSTM network to obtain the output state of the last layer and the characteristic state of the hidden layer. Taking the video modal feature sequence  $(S_1, S_2, \dots, S_n)$  as an example, the context representation result output  $(O_l \in R^{n \times d})$  can be defined as follows.

$$O_l, H_l = Bi - LSTM(S_1, S_2, \dots, S_n), l \in \{A, V, T\} \quad (1)$$



**Figure 1.** The process of multi-modal feature extraction with cross-modal modeling.

Where  $n$  is the length of feature sequence,  $d$  is the feature dimension, subscript  $l$  is the modality of the video,  $A$ ,  $V$  and  $T$  are audio, visual and text modalities, respectively, and  $H_l$  is the hidden layer state output of Bi-LSTM.

Further, we apply the cross-attention mechanism for bimodal embedding and fusion to capture the interaction characteristics of these pairs of modalities. Each bimodal feature is fused by the output features of one modal and a hidden state of another modal through a cross-attention mechanism. Taking visual and text modalities as an example,  $Cross_{VT}$  can be obtained by fusing the LSTM output features expressed by visual features and the LSTM hidden state expressed by text features through the cross-attention mechanism to capture the interactive information between two peaks. We capture the interaction characteristics of these two modalities by calculating  $Cross'_{VT}$  and  $Cross'_{TV}$  cross-attention value, the calculation method is described as follows.

$$Cross_{VT} = Cross'_{VT} O_V \quad (2)$$

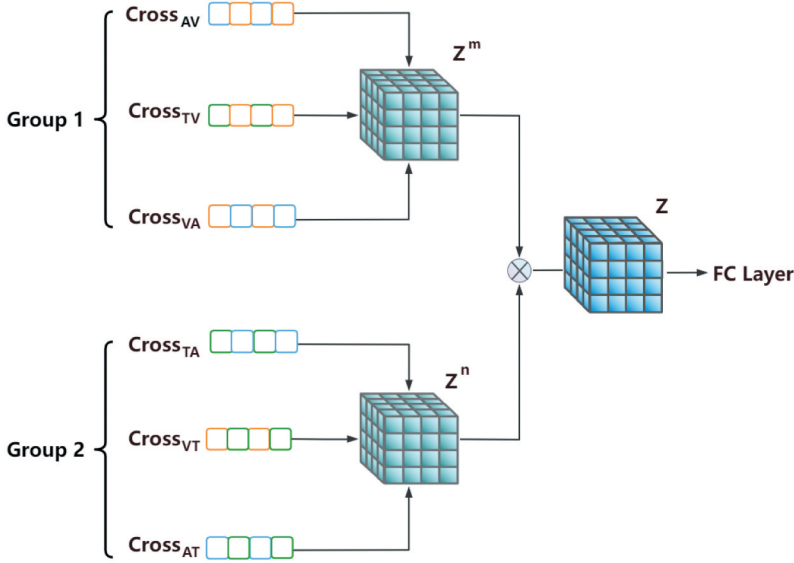
$$Cross_{TV} = Cross'_{TV} O_T \quad (3)$$

$$Cross'_{VT} = softmax\left(\frac{O_T H_V}{\sqrt{d_k}}\right) \quad (4)$$

$$Cross'_{TV} = softmax\left(\frac{O_V H_T}{\sqrt{d_k}}\right) \quad (5)$$

Where  $H_V$  and  $H_T$  are unimodal contexts from text and visual modalities, respectively, representing hidden layer state characteristics.  $\sqrt{d_k}$  is the normalized coefficients;  $O_V$  and  $O_T$  are unimodal contexts representing output features from text and visual modalities, respectively;  $Cross_{VT}$  represents cross-attention mechanism value matrix of visual modal features embedded with text context information.  $Cross_{TV}$  represents the value matrix of the cross-attention mechanism for the text modal features embedded in the visual context information. Similarly, we can obtain six groups of cross-attention value matrices ( $Cross_{TA}$ ,  $Cross_{TV}$ ,  $Cross_{AT}$ ,  $Cross_{AV}$ ,  $Cross_{VT}$ , and  $Cross_{VA}$ ), which can be generated by different bimodal combinations through the cross-attention mechanism. To compress the sparse features and make the bimodal spatial features converge intensively, the six groups of cross-attention value matrices are regarded as an input to a fully connected network block with “Dropout+Linear+ ReLU” respectively.





**Figure 2.** The overview of multi-tensor fusion network with three-peak modalities.

### **Multi-tensor Fusion Network with Three-peak Modalities**

Figure 2 demonstrates an overview of the multi-tensor fusion network with three-peak modalities. In the network framework model, we use multi-tensors to capture the interaction characteristics in three modalities and store the high-dimensional semantic connections across modalities, so that the final regression prediction can get better prediction values.

In order to better model the three-peak modal relationship, we divide six groups of cross-attention value matrices into two groups. To make the dynamic relationship modeling of the three models in tensor space, the cross-modal features of each group need to include three modalities of text, video, and audio at the same time. So, we choose  $(Cross_{AV}, Cross_{TV}, Cross_{VA})$  as a group, and  $(Cross_{TA}, Cross_{VT}, Cross_{AT})$  as the other group, instead of randomly dividing them into two groups.

Moreover, a multi-tensor fusion network can be obtained by combining the feature vectors of different modalities multiple times. The specific tensor is defined as follows.

$$Tensor_{bi}(Z^a, Z^b) = \begin{bmatrix} Z^a \\ \mathbf{1} \end{bmatrix} \otimes \begin{bmatrix} Z^b \\ \mathbf{1} \end{bmatrix} \quad (6)$$

$$Tensor_{tr}(Z^a, Z^b, Z^c) = \begin{bmatrix} Z^a \\ \mathbf{1} \end{bmatrix} \otimes \begin{bmatrix} Z^b \\ \mathbf{1} \end{bmatrix} \otimes \begin{bmatrix} Z^c \\ \mathbf{1} \end{bmatrix} \quad (7)$$

Where tensor fusion can be used for bimodal or multi-modal modes,  $\otimes$  represents the outer product between vectors. The multi-tensor fusion network is simple to integrate, and the shared feature subspace is often semantically invariant Xue et al. (2020). A series of tensor fusion of bimodal or multi-modal modes enables the network model to transfer feature knowledge from one model to another, and fully capture the dynamic relationship information between the modalities. Before multi-level tensor fusion, the “1” vector should be stitched onto the feature vectors of each modality, so that each modality can be modeled correctly in the end.  $Tensor_{bi}$  is used to capture the bimodal effect of two modalities, and  $Tensor_{tr}$  is used to capture the three-peak effect of three modalities.

$$Z = Tensor_{bi}(Z^m, Z^n) \quad (8)$$

$$Z^m = Tensor_{tr}(Cross_{AV}, Cross_{TV}, Cross_{VA}) \quad (9)$$

$$Z^n = Tensor_{tr}(Cross_{TA}, Cross_{VT}, Cross_{AT}) \quad (10)$$

Regarding the two obtained tensors as a new view, the two compressed tensors can be merged again to achieve hierarchical fusion.

After multi-level tensor fusion, a tensor with spatial relations is acquired and we can obtain a multimodal fusion tensor network with high-dimensional relations. The prediction layer uses function Sigmoid for normalization so as to facilitate the control of the output range. The regression prediction method is defined as follows.

$$I = 6 * Sigmoid(FC(Z; W_s)) - 3 \quad (11)$$

where  $I$  is the prediction result of affect intensity,  $W_s$  is the weight and  $Z$  is the total tensor.  $FC$  is a fully connected neural network used with the weight  $W_s$  conditioned on the total tensor  $Z$ , which contains two layers of dropout and two layers of activation unit ReLu, and be connected to the last prediction layer.

On the basis of the cross-modal modeling, the multi-tensor fusion network with three-peak modalities can dynamically realize multi-modal semantic combination during model training, and make full use of the multi-layer characteristics of deep neural networks.

## Experiment and Analysis

### The Datasets

To evaluate the effectiveness of our proposed approach, we compared MTFN-CMM with other baseline models on CMU-MOSI Zadeh et al. (2016) and CMU-MOSEI Zadeh et al. (2018c), two public benchmark multimodal

sentiment analysis datasets. The CMU-MOSI dataset contains YouTube videos from 93 different speakers, which covers 2199 opinions. There are 23.2 opinion clips per video on average. The average length of each opinion clip is about 4.2 seconds and the opinions are expressed totally in 26,295 words. Each opinion fragment correspondingly has an emotional label. Emotional intensity labels range from  $[-3, 3]$ . The number of discourse fragments used in the training set, test set, and verification set accounted for 229, 229, and 686, respectively. CMU-MOSEI is an emotion analysis dataset, which has 3,229 videos and 22,676 quotes from more than 1,000 online YouTube speakers. The number of utterances is about ten times as much as that of CMU-MOSI. The training, validation, and test sets consist of 16,216, 1,835, and 4,625 vocalizations, respectively.

### **Parameter Setting and Evaluation Metrics**

We extracted the unimodal features separately on the multimodal data sets. 1) Text feature extraction. We converted the text lines obtained from the SDK into the pre-trained Glove word embedding. Both annotated dataset word embeddings are 300-dimensional vectors. 2) Audio feature extraction. Covarep Degottex et al. (2014) was used to extract the low acoustic features with dimensions of 5 (CMU-MOSI) and 74 (CMU-MOSI), including 12 Mehr cepstrum coefficients (MFCC), etc. 3) Video feature extraction. Facet Stockli et al. (2018) was employed to extract facial expressions to represent the basic features and highlight emotional features of each frame. The sizes of extracted features were 20 (CMU-MOSI) and 35 (CMU-MOSI), respectively. Table 1 shows the hyperparameters in our network framework for the multimodal emotional intensity prediction task.

The regression and different dimensions of classification experiments are performed on the two public data sets CMU-MOSI and CMU-MOSEI. We chose the mean absolute error (*MAE*), Pearson correlation coefficient (*r*) as the model evaluation indexes of our regression experiment, and the accuracy  $A_2$  value and *F1* score as the model evaluation indexes of two-class classification experiment. Moreover, the accuracy  $A_5$  of five-class sentiment classification

**Table 1.** Hyperparameters of MTFN-CMM.

Hyperparameters	CMU-MOSI	CMU-MOSEI
Batch Size	24	8
Optimizer	AdamW	AdamW
Bi-LSTM Hidden Unit Size	128	128
Attention Dense Layer Unit Size	64	128
Dense Dropout	0.2	0.1
FC-Layer Unit Size	64	128
Output Dropout	0.3	0.1
Gradient Clip	1.0	1.0
Epochs	50	30

and the accuracy  $A_7$  of seven-class sentiment classification are regarded as the model evaluation indexes. If the average absolute error  $MAE$  is smaller, the correlation coefficient  $r$  is larger in the regression experiments, and  $F1$  and the accuracy  $A_i$  in different dimensions of classification experiments is higher, the performance of the model represented by the multimodal sentiment analysis is better.

### Comparison with Other Baseline Models

We compare the proposed approach on CMU-MOSI and CMU-MOSEI datasets with the following baseline model: Dynamic Fusion Graph (DFG) Zadeh et al. (2018c), Early Fusion – LSTM (EF-LSTM) Plank, Søgaard, and Goldberg (2016), Multi-attention recurrent network (MARN) Zadeh et al. (Zadeh, et al., 2018b), Multimodal Factorization Model (MFM) Tsai et al. (2018), Multimodal Cyclic Translation Network (MCTN) Pham et al. (2019), Context-Aware Interactive Attention (CIA) Chauhan et al. (2019), Tensor Fusion Network Zadeh et al. (2017) and Relational Tensor Network Sahay et al. (2018). The experiment results are shown in Table 2.

We observe that the proposed MTFN-CMM yields better performance against other baseline models for the regression and different dimensions of classification experiments on the two benchmark datasets. The experiment results in bold indicate the corresponding model obtained the best performance on the concerned indicator. For intensity prediction, our proposed MTFN-CMM obtains lesser mean absolute error  $MAE$  with high Pearson correlation scores  $r$ . In comparison to context-aware interactive attention (CIA), we yield approximately 0.02 and 0.09 points improvement in  $MAE$  with higher  $r$  on two benchmark datasets, respectively. In addition, our proposed approach achieves approximately 1.5 and 3 percentage higher  $F1$  scores with slightly higher accuracy on two benchmark datasets for two-class sentiment classification. This may be attributed to the use of the cross-modal modeling to strengthen the early bimodal feature extraction work, which

**Table 2.** Regression experimental results compared with other baselines.

Metric	CMU-MOSI					CMU-MOSEI					
	MAE	r	F1	A2	A7	MAE	r	F1	A2	A5	A7
DFG	–	–	–	–	–	0.710	0.540	77.0	76.9	45.1	45.0
EF-LSTM	1.023	0.622	75.2	75.4	32.4	0.642	0.616	77.9	76.2	44.7	43.9
MARN	0.968	0.625	77.0	77.1	34.7	–	–	–	–	–	–
MFM	0.951	0.662	78.1	78.1	36.2	–	–	–	–	–	–
MCTN	0.909	0.676	79.1	79.3	–	0.609	0.670	80.6	79.8	–	–
CIA	0.914	0.689	79.5	79.9	<b>38.92</b>	0.680	0.590	78.2	80.4	49.2	<b>50.1</b>
TFN	0.946	0.640	77.1	77.2	34.9	0.617	0.646	76.6	76.4	44.9	44.0
RTN	0.942	0.661	79.2	79.2	35.2	0.609	0.650	78.2	77.9	45.6	45.1
MTFN-CMM(ours)	<b>0.891</b>	<b>0.691</b>	<b>81.0</b>	<b>80.9</b>	<b>38.92</b>	<b>0.589</b>	<b>0.674</b>	<b>81.3</b>	<b>80.8</b>	<b>49.3</b>	<b>50.1</b>
T-test	0.0019	0.0002	0.005	0.0018	0.0003	0.0002	0.00001	0.036	0.037	0.00003	0.000001

enables the multi-tensor fusion network to establish more intra-modal and inter-modal connections. Compared with tensor fusion network (TFN) and relational tensor network (RTN) approaches, our proposed MTFN-CMM has certain advantages in regression and different dimensions of classification experiments. Especially, we observe approximately 3–5 percentage improvement in accuracy values on five-class and seven-class classification experiments. It indicates that MTFN-CMM effectively exploits this interaction for a three-peak modal relationship with multi-tensor fusion and our model can learn more cross-modal interaction among the multi-modalities than TFN and RTN models.

We perform a statistical significance test (t-test) on the experimental results, and observe that performance improvement in the proposed MTFN-CMM over other compared baseline models is significant with 95% confidence (i.e.,  $p$ -value  $< 0.05$ ).

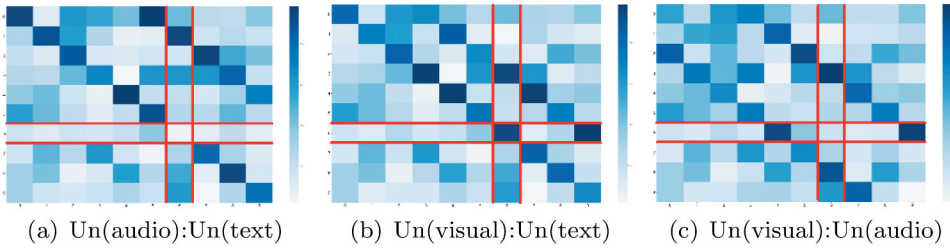
### Coupling Experiments with Multi-tensor Fusion

To further study the effectiveness of cross-modal fusion with multi-tensor network, we perform the coupling effects of multi-tensor fusion network (MTFN) on cross-modal modeling (CMM) with different attention mechanisms, such as self-attention mechanism (SM), multi-head attention mechanism (MM), and cross-attention mechanism (CM). The experimental results are observed in Table 3.

Through the experiments of different attention mechanisms in cross-modal modeling, it can be found that coupling multi-tensor fusion network with cross-modal modeling can basically obtain lesser mean absolute error  $MAE$  on two datasets CMU-MOSI and CMU-MOSEI. The multi-head attention mechanism achieving a better coupling experimental effect than the self-attention mechanism. However, compared with MTFN-CMM (SM) and MTFN-CMM (MM), our proposed MTFN-CMM can perform better in regression and classification experiments. This may be because that the cross-modal modeling with a cross-attention mechanism can capture more interactive information within and between the modalities, which is conducive to the multi-tensor fusion of the three-peak cross-modality. Besides, cross-modal modeling with a cross-attention mechanism focuses on the contextual

**Table 3.** Coupling experimental results on multi-tensor fusion network with cross-modal modeling with different attention mechanisms.

Metric	CMU-MOSI			CMU-MOSEI		
	MAE	r	F1	MAE	r	F1
MTFN	0.942	0.661	79.2	0.609	0.650	78.2
MTFN-CMM(SM)	0.918	0.612	78.6	0.605	0.670	77.2
MTFN-CMM(MM)	0.908	0.684	80.1	0.598	0.671	80.4
<b>MTFN-CMM(ours)</b>	<b>0.891</b>	<b>0.691</b>	<b>81.0</b>	<b>0.589</b>	<b>0.674</b>	<b>81.3</b>



**Figure 3.** An example of the cross attention weight matrix heat map of ten video segments.

information within cross-modality and facilitates hierarchical multi-tensor network fusion in different modalities. It is speculated that the multi-tensor fusion network may be better able to capture dynamic relationship information with bimodal or even multi-peak interaction from cross-modal modeling with the cross-attention mechanism.

### **Visualization Analysis of Cross-modal Modeling**

In order to further explore the influence of cross-modal modeling feature extraction in MTFN-CMM, we visually analyze the degree of cross-modal bimodal interaction of videos with audio modal noise interference in the process of multi-modal feature fusion.

As shown in Figure 3, we randomly intercept and select 10 video segments from adjacent contexts in the video, and display the weight matrix of the cross-attention mechanism in the cross-modal modeling process with the help of the heatmaps, where the cross-modal weight matrix is used to measure the bimodal relationship information. Figure 3a–c show the audio-text modal, visual-text modal, and visual-audio modal cross-attention weight matrices of 10 adjacent segments, respectively. By observing the color depth of the cross-attention value of the seventh segment, we found that lighter colors appeared in both Figure 3a and c, and darker colors appeared in Figure 3b. The results show that audio modalities capture opposite emotional features from text and visual modalities. At the same time, video and text capture highly similar emotional features. By comprehensively analyzing the cross-modal modeling feature extraction of the seventh segment information, it can be inferred that the audio modal in the multi-modal feature may have noise interference. Therefore, especially when there are a lot of noise features in the real video, cross-modal modeling feature extraction is very effective in improving the performance of multi-tensor fusion network feature fusion.

## Conclusion

We proposed a multi-tensor fusion network with the cross-modal modeling for multimodal sentiment analysis in this study, which can capture intra-modal dynamics and inter-modal interactions and can be used for multi-modal affective intensity prediction effectively. The performance of the proposed method is superior to the existing advanced methods and shows significant performance improvement on CMU-MOSI and CMU-MOSI datasets. The proposed approach only tries to retain coarse-grained modal fusion, and we plan to do some work on fine-grained modal fusion in the future, so as to further reduce the number of required parameters of the model and improve the accuracy of inference and prediction of emotional intensity Xi, Lu, and Yan (2020), Mittal et al. (2020) and Fs et al. (2020).

## Acknowledgments

Thanks to Yifeng Tan for revising the English grammar of the paper.

## Disclosure Statement

No potential conflict of interest was reported by the author(s).

## Funding

This research was supported in part by the Guangzhou Science and technology project [202102020878], the National Natural Science Foundation of China [62006053], the Special Innovation Project of Guangdong Education Department [2018KQNCX072].

## ORCID

Xueming Yan  <http://orcid.org/0000-0001-7809-3436>

Shengyi Jiang  <http://orcid.org/0000-0002-6753-474X>

## References

- Chaturvedi, I., R. Satapathy, S. Cavallari, and E. Cambria. 2019. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters* 125:264–70. doi:10.1016/j.patrec.2019.04.024.
- Chauhan, D. S., M. S. Akhtar, A. Ekbal, and P. Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), HongKong, 5647–57.

- Degottex, G., J. Kane, T. Drugman, T. Raitio, and S. Scherer. 2014. Covarepa collaborative voice analysis repository for speech technologies. In IEEE international conference on acoustics, speech and signal processing (ICASSP), 960–64. IEEE, Florence, Italy.
- Ebrahimi, M., A. H. Yazdavar, and A. Sheth. 2017. Challenges of sentiment analysis for dynamic events. *IEEE Intelligent Systems* 32 (5):70–75. doi:10.1109/MIS.2017.3711649.
- Fs, A., B. Jra, C. Ai, and A. Mg. 2020. Multimodal subspace support vector data description. *Pattern Recognition*, 110:107648.
- Kumar, A., and J. Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4477–81. IEEE, Spain.
- Li, Y., K. Zhang, J. Wang, and X. Gao. 2021. A cognitive brain model for multimodal sentiment analysis based on attention neural networks. *Neurocomputing* 430:159–73. doi:10.1016/j.neucom.2020.10.021.
- Liu, Y. J., and C. L. Cheng. 2018. A gesture feature extraction algorithm based on key frames and local extremum. *Computer Technology and Development* 28 (3):127–31.
- Majumder, N., D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based Systems* 161:124–33. doi:10.1016/j.knosys.2018.07.041.
- Mittal, T., U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence* 34:1359–67. doi:10.1609/aaai.v34i02.5492.
- Patel, D., X. Hong, and G. Zhao. 2016. Selective deep features for microexpression recognition. In The 23rd international conference on pattern recognition (ICPR), 2258–63. IEEE, Cancún, Mexico.
- Pennington, J., R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 1532–43.
- Pham, H., P. P. Liang, T. Manzini, L. P. Morency, and B. Pzos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:6892C6899. doi:10.1609/aaai.v33i01.33016892.
- Plank, B., A. Søgaard, and Y. Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. arXiv preprint arXiv:160405529.
- Poria, S., E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. P. Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In IEEE International Conference on Data Mining (ICDM), 1033C1038. IEEE, New Orleans, LA, USA.
- Poria, S., I. Chaturvedi, E. Cambria, and A. Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In The 16th IEEE international conference on data mining (ICDM), 439C448. IEEE, Barcelona, Spain.
- Sahay, S., S. H. Kumar, R. Xia, J. Huang, and L. Nachman. 2018. Multimodal relational tensor network for sentiment and emotion classification. arXiv preprint arXiv:180602923.
- Stockli, S., M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson. 2018. Facial expression analysis with affdex and facet: A validation study. *Behavior Research Methods* 50 (4):1446–60. doi:10.3758/s13428-017-0996-1.
- Tsai, Y. H. H., P. P. Liang, A. Zadeh, L. P. Morency, and R. Salakhutdinov. 2018. Learning factorized multimodal representations. arXiv preprint arXiv:180606176.



- Tsai, Y. H. H., S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference on Association for Computational Linguistics, 6558–61. NIH Public Access, Florence, Italy.
- Xi, C., G. Lu, and J. Yan. 2020. Multimodal sentiment analysis based on multi-head attention mechanism. In Proceedings of the 4th International Conference on Machine Learning and Soft Computing, New York NY United States, 34C39.
- Xing, F. Z., E. Cambria, and R. E. Welsch. 2018. Natural language based financial forecasting: A survey. *Artificial Intelligence Review* 50 (1):49–73. doi:10.1007/s10462-017-9588-9.
- Xue, H., X. Yan, S. Jiang, and H. Lai. 2020. Multi-tensor fusion network with hybrid attention for multimodal sentiment analysis. The International Conference on Machine Learning and Cybernetics (ICMLC), shenzhen, 169–74. IEEE.
- Young, T., E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA.
- Zadeh, A. B., P. P. Liang, S. Poria, E. Cambria, and L. P. Morency. 2018c. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* 1:2236–46.
- Zadeh, A., M. Chen, S. Poria, E. Cambria, and L. P. Morency. 2017. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:170707250.
- Zadeh, A., P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency. 2018a. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA.
- Zadeh, A., P. P. Liang, S. Poria, P. Vij, E. Cambria, and L. P. Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA.
- Zadeh, A., R. Zellers, E. Pincus, and L. P. Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31 (6):82–88. doi:10.1109/MIS.2016.94.
- Zhou, P., W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* 2:207–12.