Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Modal Emotion Recognition Using Tensor Decomposition Fusion and Self-Supervised Multi-Tasking

Jiawei Zhu
Shanghai University
Xianxun Zhu
Shanghai University
Shoujin Wang
University of Technology Sydney
Tao Wang
Shanghai University
Jingze Huang
Shanghai University
Rui Wang
rwang@shu.edu.cn

Shanghai University

Research Article

Keywords: Self-supervised, multi-tasking, Emotion recognition, Multimodal

Posted Date: February 6th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3916468/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at International Journal of Multimedia Information Retrieval on September 3rd, 2024. See the published version at https://doi.org/10.1007/s13735-024-00347-3.

Multi-Modal Emotion Recognition Using Tensor Decomposition Fusion and Self-Supervised Multi-Tasking

Jiawei Zhu[1], Xianxun Zhu[1], Shoujin Wang[2], Tao Wang[1], Jingze Huang[1] and Rui Wang[1*]

[1]School of Communication and Information Engineering, Shanghai University, No. 99, Shangda Road, Shanghai, 200444, China.

[2]Data Science Institute, University of Technology Sydney, Sydney, 2007, Australia.

*Corresponding author(s). E-mail(s): rwang@shu.edu.cn;

Abstract

With technological advancements, we can now capture rich dialogue content, tones, textual information, and visual data through tools like microphones, the internet, and cameras. However, relying solely on a single modality for emotion analysis often fails to accurately reflect the true emotional state, as this approach overlooks the dynamic correlations between different modalities. To address this, our study introduces a multimodal emotion recognition method that combines tensor decomposition fusion and self-supervised multi-task learning. This method first employs Tucker decomposition techniques to effectively reduce the model's parameter count, lowering the risk of overfitting. Subsequently, by building a learning mechanism for both multimodal and unimodal tasks and incorporating the concept of label generation, it more accurately captures the emotional differences between modalities. We conducted extensive experiments and analyses on public datasets like CMU-MOSI and CMU-MOSEI, and the results show that our method significantly outperforms existing methods in terms of performance. The related code is open-sourced at https://github.com/ZhuJw31/MMER-TD.

Keywords: Self-supervised, multi-tasking, Emotion recognition, Multimodal

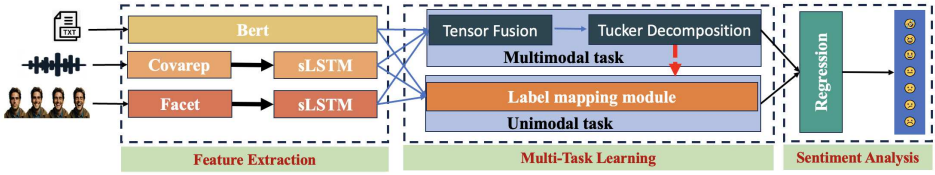


Fig. 1: Multitask Learning Framework for Multimodal Sentiment Analysis.

1 Introduction

In the era of big data, the ubiquitous information in our lives has become a vital channel for acquiring knowledge [1]. Modern technologies like microphones, the internet, and cameras enable us to capture dialogue content, tones, textual information, and visual data. In the field of emotion recognition, relying solely on a single modality (such as text, sound, or vision) for emotion analysis may not accurately reflect the actual emotional state [2]. For instance, the phrase "I'm fine" may appear positive if analyzed merely from its textual context. However, if combined with the speaker's tone and context, the same phrase might convey sadness, thus expressing negative emotion. This illustrates how traditional single-modality emotion recognition methods, due to their one-sidedness in information, often yield limited analysis results. In contrast, multimodal emotion recognition, by comprehensively analyzing various modalities like voice, image, and text, can capture and interpret an individual's emotional state more holistically. Therefore, multimodal emotion recognition has become one of the hot topics in current research [3, 4]. In the field of multimodal emotion analysis, researchers are extensively dedicated to developing integrated fusion mechanisms to combine data from text, vision, and acoustics, aiming for more precise and comprehensive emotional interpretation. Current research progress showcases a variety of strategies and technological approaches. For instance, early fusion strategies begin integrating information from different modalities in the initial stages of data processing, such as using LSTM networks to link features of each modality for key sequence information extraction [5, 6]. Late fusion, on the other hand, focuses on independently analyzing the intrinsic characteristics of each modality before integrating cross-modal data at the final stage, with the advantage of preserving the independence of each modality's data. Hybrid fusion methods combine the strengths of both early and late fusion, considering both initial modal information and fully utilizing later data representations, thus showing strong competitiveness in multimodal emotion analysis. Representation-based approaches emphasize effective representation of modal data before fusion, such as independently extracting features of each modality through pre-trained technologies like BERT and the Facial Action Coding System [7]. Pre-trained model methods, like SentiBERT and CM-BERT [8, 9], adapt to specific multimodal emotion analysis tasks by fine-tuning large-scale models and datasets. Multi-channel attention-based methods, such as MHSAN and

MA-RNN [10, 11], highlight the correlations and importance between different modalities and their internal discourse through attention mechanisms. Quantum theory methods, as the latest attempt, like the QMN model, use quantum interference theory to simulate the interactions between different modalities, presenting a unique perspective and potential [12, 13].

Despite significant progress in the field of multimodal emotion recognition, existing methods still face a series of challenges. Firstly, methods based on tensor fusion often come with a large number of parameters, which not only increases the computational burden but also leads to a higher risk of model overfitting. Secondly, a common issue in multimodal emotion analysis research is the overreliance on limited datasets. Existing datasets usually provide uniform multimodal annotations, but this approach fails to fully capture the independent emotional characteristics of each modality, limiting the accuracy of capturing differences across modalities. To address these issues, our study has developed an innovative multimodal emotion recognition method that combines tensor decomposition fusion with self-supervised multi-task learning. The main innovations and contributions of our study are as follows:

- **Introduction of Tucker Decomposition:** To tackle the large parameter issue inherent in traditional tensor fusion, we employed Tucker decomposition technology, effectively reducing the model parameters and thus lowering the risk of overfitting.
- **Construction of a Multi-task Learning Mechanism:** Addressing the limitation of single annotation, we introduced the concept of label generation. By constructing a learning mechanism for both multimodal and unimodal tasks, our method captures the emotional differences between modalities more precisely.
- **Extensive experiments and analyses** were conducted on public datasets such as CMU-MOSI and CMU-MOSEI. The results demonstrate that our method significantly outperforms existing methods in terms of performance. We have made the related code open source.

2 Methodology

In this study, we propose a multimodal emotion recognition method based on tensor decomposition fusion and self-supervised multi-task learning. As illustrated in Figure 1, we first extract features from each unimodal data; subsequently, these features are integrated using a tensor fusion method, and dimensionality reduction is applied to the fused data using Tucker decomposition. Following this, we conduct multimodal emotion recognition tasks and employ a label generation module for unimodal tasks to perform comprehensive prediction and analysis.

2.1 UNIMODAL FEATURE EXTRACTION

In multimodal emotion recognition research, the extraction of features from individual modalities is crucial for the subsequent fusion process. Specifically, for text-based emotion recognition, we employed the BERT (Bidirectional Encoder Representations from Transformers) model for word vector representation in texts. BERT, a model based on the Transformer architecture, generates more precise word vector representations through its understanding of bidirectional context [8]. To extract comprehensive audio features, we chose the COVAREP toolkit [14]. COVAREP is an open-source audio analysis toolkit capable of extracting a range of critical vocal features, such as pitch and formant frequencies, essential for understanding the emotional variations in speech. Additionally, to delve deeper into the contextual information of speech, we utilized the SLSTM technique [15]. SLSTM captures subtle variations and long-term dependencies in speech signals through time-series modeling, providing a more comprehensive approach to speech emotion recognition [16].

For facial expression feature extraction, we used the FACET software [17]. FACET is an advanced facial expression analysis tool that can identify and quantify various micro-expressions and emotional features on the face. It extracts key facial expression features by analyzing movements of facial muscles. Finally, to capture the continuity of facial expressions across frames, we employed the SLSTM model. SLSTM, by temporally modeling facial expressions in consecutive frames, effectively recognizes and understands the dynamic changes in facial expressions.

2.2 MULTIMODAL TASK

2.2.1 Tensor fusion

In multimodal tasks, the Tensor Fusion Network [18] excels at modeling the interactions between three distinct modalities, accounting for intermodal dynamics. To facilitate this, a third-order Cartesian product is employed, formally defined as:

$$\mathbf{z}_m = \mathbf{f}'_a \otimes \mathbf{f}'_t \otimes \mathbf{f}'_v \tag{1}$$

where $\mathbf{f}'_a$, $\mathbf{f}'_t$, and $\mathbf{f}'_v$ are the augmented dimensionality representations of the extracted features $\mathbf{f}_a$, $\mathbf{f}_t$, and $\mathbf{f}_v$, respectively, and $\otimes$ denotes the outer product operation. The tensor $\mathbf{z}_m$ encompasses all points in the Cartesian space, encapsulating all possible permutations of acoustic, textual, and visual modalities. The process is shown in Figure 2.

The individual modality features are represented by $\mathbf{f}_a\mathbf{f}_t\mathbf{f}_v$, while the sections, $\mathbf{z}_a \otimes \mathbf{z}_t$, $\mathbf{z}_a \otimes \mathbf{z}_v$, and $\mathbf{z}_t \otimes \mathbf{z}_v$, denote the pairwise intermodal interactions. Prior to the outer product operation, an expansion of dimensions is required, as shown by the section, $\mathbf{z}_a \otimes \mathbf{z}_t \otimes \mathbf{z}_v$, which signifies the interaction among all three modalities.

The augmentation process for any two modalities prior to the outer product is as follows:

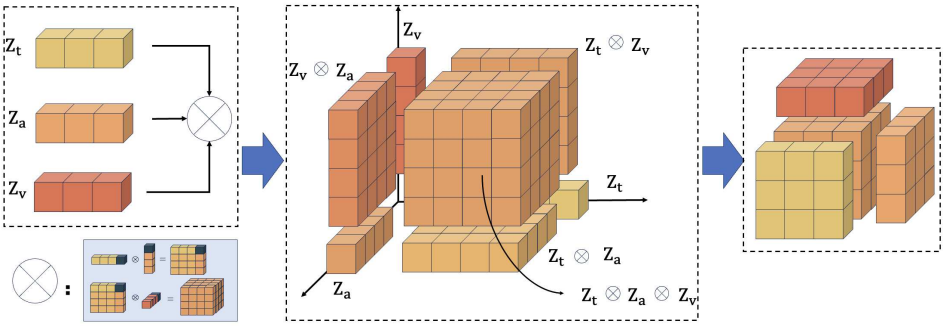


Fig. 2: Flowchart of Multimodal Feature Fusion and Tensor Decomposition Process.

$$\mathbf{z}_m = \begin{bmatrix} \mathbf{z}'_v \\ \mathbf{z}'_t \\ \mathbf{z}'_a \end{bmatrix} = \begin{bmatrix} \mathbf{f}_v \otimes \mathbf{1} \\ \mathbf{f}_t \otimes \mathbf{1} \\ \mathbf{f}_a \otimes \mathbf{1} \end{bmatrix} \tag{2}$$

This expansion allows for the computation of correlations between modalities and preserves intrinsic information within each modality.

2.2.2 Tucker decomposition

Due to the high dimensionality of $\mathbf{z}_m$ resulting from the outer product of three vectors, substantial computational resources are required. Tucker decomposition, akin to a higher-order Principal Component Analysis (PCA), addresses this by decomposing a high-dimensional tensor into a core tensor $\mathbf{G}$ and a set of factor matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}, \ldots$:

$$\mathbf{z}_m = \mathbf{G} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \times_3 \mathbf{A}^{(3)} \tag{3}$$

where $\times_n$ denotes the n-mode product. The factor matrices are orthogonal, and the core tensor $\mathbf{G}$ encapsulates the principal interactions and serves as a compressed version of the original tensor [19].

The fused tensor $\mathbf{z}_m$ is decomposed into three matrices $\mathbf{W}_a$, $\mathbf{W}_v$, and $\mathbf{W}_t$, representing the acoustic, visual, and textual components, respectively, alongside a core matrix $\mathbf{T}_c$:

$$\mathbf{T}_c = (\mathbf{G} \times_1 \mathbf{W}_a) \times_2 \mathbf{W}_v \times_3 \mathbf{W}_t \tag{4}$$

The core matrix $\mathbf{T}_c$ not only encapsulates the essential content from the acoustic, visual, and textual components but also their intrinsic interrelations, while maintaining a lower data volume and dimensionality compared to the original fusion tensor $\mathbf{z}_m$.

2.3 UNIMODAL TASK

In the evaluation of multimodal fusion algorithms, performance metrics play a critical role. These metrics enable the assessment of the efficacy of the algorithms in various contexts. The Tucker decomposition, in particular, is leveraged to transform the high-dimensional fusion tensor into a core tensor and associated matrices, significantly reducing the complexity of the multimodal data. As illustrated in Figure 3, The unimodal Latent Geometric (LG) framework is applied to extract the most informative features from each modality, subsequently represented by the core tensor C [20]. The core tensor C is defined as :

$$C_p = \frac{\sum_{i=1}^{N} I(v_i(s) > 0) \cdot f_p^g \cdot v_i}{\sum_{i=1}^{N} I(v_i(s) > 0)} \tag{5}$$

$$C_m = \frac{\sum_{i=1}^{N} I(v_i(s) < 0) \cdot f_m^g \cdot v_i}{\sum_{i=1}^{N} I(v_i(s) < 0)} \tag{6}$$

where I is an indicator function, $v_i(s)$ is the i-th scalar support vector, f_p^g and f_m^g are the positive and negative class centroids of the data distribution.

The task involves the optimization of a trade-off between multimodal integration and the uniqueness of each modality. This balance is achieved through the evaluation of the distance metrics D_p and D_m, which are calculated as follows:

$$D_p^i = \left\| F - C_p^i \right\|_2^2 \tag{7}$$

$$D_m^i = \left\| F - C_m^i \right\|_2^2 \tag{8}$$

where F is the feature vector, and $\|\cdot\|_2^2$ denotes the L2 norm.

Furthermore, a balance is sought between the preservation of modality-specific characteristics and the fusion process. This is expressed through the balance coefficient α, determined by:

$$\alpha_i = \frac{D_n^i - D_p^i}{D_p^i + \epsilon} \tag{9}$$

where D_n^i is the norm distance, D_p^i is the positive class distance, and ϵ is a small constant to prevent division by zero. In the context of error minimization, the enhancement of the core tensor aims to address the bias inherent in the fusion process. The adjusted bias E is computed for each modality using the following equation:

$$E = \frac{y_m \cdot \alpha_s + y_m + \alpha_m}{2 \cdot \alpha_m} \tag{10}$$

where y_m is the modality-specific performance measure, and α_s, α_m are the scaling factors.

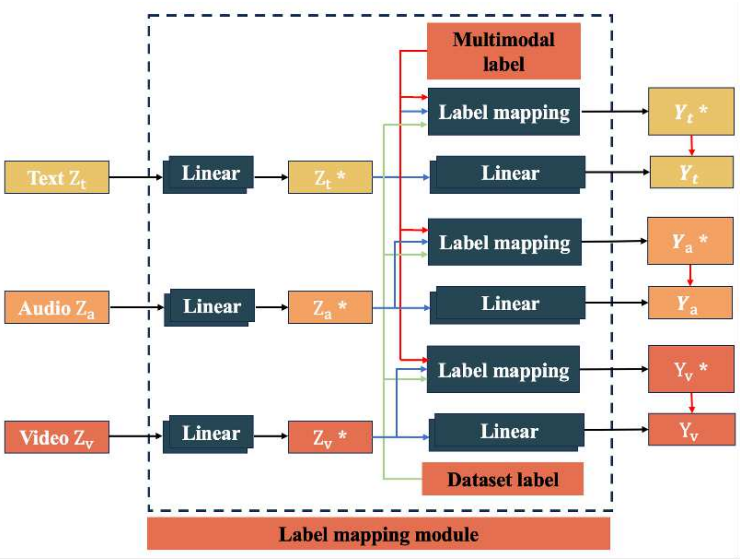


Fig. 3: Diagram of Multimodal Data Label Mapping Module Structure.

The optimization process addresses the challenge of overfitting while maintaining the integrity of the individual modalities. The optimization function δ_s is defined by the equation:

$$\delta_s = \frac{\alpha_s - \alpha_m \cdot y_m + \alpha_m}{2 \cdot \alpha_m} \tag{11}$$

This function seeks to minimize the error across the modalities while preserving the high-dimensional structure of the data. It is instrumental in achieving an optimal balance between the core tensor and the individual modality features. For the optimization of the multimodal fusion process, the scaling function s plays a crucial role, enabling the adjustment of the contribution of individual modality performance measures y_m within the fusion scheme. The scaling function s is defined for each modality i as:

$$y_s^{(i)} = \begin{cases} \frac{y_m}{i-1} - \frac{(i-1)}{i+1} y_s & \text{for } i = 1 \\ \frac{y_m}{i+1} - \frac{2}{i+1} y'_s & \text{for } i > 1 \end{cases} \tag{12}$$

The scaling function s ensures the equitable contribution of each modality, balancing the multimodal integration with modality-specific characteristics. The optimization process employs a soft version of the max operator to maintain a balance between the fusion of modalities and the preservation of modality-specific information. This optimization is realized by minimizing the loss function L, which is computed as:

$$L = \frac{1}{N} \sum_{i=1}^{N} \left(\sigma(y_{im} - y_{il}) + \sum_k w_k^i \left\| y_k^i - y_s^{(i)} \right\| \right) \tag{13}$$

where σ is the sigmoid function, y_{im} and y_{il} represent the maximum and minimum performance measures of the modalities, w_k^i are the weights assigned to each modality difference, and y_k^i is the performance measure for modality k. The weights w_s^i are adjusted via a hyperbolic tangent function to control the impact of each modality:

$$w_s^i = \tanh(y_s^i - y_m^i) \tag{14}$$

This function ensures that the optimization process not only integrates the strengths of each modality but also preserves their unique contributions to the overall fusion result.

3 DATASETS AND EVALUATION METHODS

3.1 Datasets

The CMU-MOSI [21]dataset offers multimodal data from 93 videos, including text, visual gestures, and audio-visual features. It features subjective, emotion-intensity annotations for 2198 utterances, aligned with words and acoustic elements, scored from -3 (strongly negative) to +3 (strongly positive).

The CMU-MOSEI [22] dataset includes 23,453 annotated segments from 5,000 videos featuring 1,000 different speakers and covering 250 unique topics.

3.2 Evaluation Criteria

To evaluate the performance of our multimodal fusion algorithm, we utilized the F1 Score ($F1$-$Score$), also known as the F1 Measure, and an accuracy metric called Acc-2. These are defined as follows:

$$F1 = \frac{2TP}{2TP + FN + FP} \tag{15}$$

$$Acc_i = \frac{TP_i}{TP_i + FP_i} \tag{16}$$

Here, TP denotes the number of true positive observations, FP represents the number of false positive observations, and FN stands for the number of false negative observations.

Moreover, we considered the F1 Score as a harmonic mean of precision and recall, with Acc-2 being a binary accuracy excluding the neutral class; i.e., a score of 0. Additionally, the Mean Absolute Error (MAE) and the Correlation Coefficient ($Corr$) were employed for a more nuanced understanding of the results:

$$MAE(x, h) = \frac{1}{m} \sum_{i=1}^{m} \|h(x_i) - y_i\| \tag{17}$$

We evaluated metrics such as the F1 score, Acc, and Corr, while computational costs were gauged using the MAE and the number of parameters , and parameter count.

These metrics and cost measures provide a comprehensive view of the algorithm's performance and its operational demands.

4 EXPERIMENT AND ANALYSIS

4.1 Baseline

Io thoroughly evaluate the overall performance of our model, we conducted a fair comparison with leading multimodal emotion recognition models.

Tensor Fusion Network (TFN) [23] : TFN leverages the Cartesian product to calculate tensors from each modality, effectively capturing unimodal, bimodal, and trimodal interaction information.

Low-rank Multimodal Fusion (LMF) [24] : As an advancement over TFN, LMF aims to reduce computational complexity and enhance efficiency through low-rank tensor factorization. Specifically, the model employs a low-rank decomposition approach to break down multimodal tensors into smaller ones and utilizes a gating mechanism to control the significance of each decomposed tensor, enabling effective fusion of multimodal information.

Multimodal Factorization Model (MFM) [25] : MFM captures interactions between modalities using factorization, exhibiting flexible generative capabilities by adjusting independent factors and reconstructing missing modalities. It represents each modality as a set of factors whose interactions produce joint features, which are then reconstructed for missing modalities using deep neural networks.

Modality-Invariant and Modality-Specific Representations (MISA) [26] : MISA learns and fuses representations of different modalities by integrating distributional similarity, orthogonal loss, reconstruction loss, and task prediction loss. Applicable to a variety of multimodal data tasks such as audio-visual speech synthesis and emotion recognition, this approach focuses on harmonizing modality-invariant features with modality-specific nuances.

Self-Supervised Multimodal Learning Framework (Self-MM) [27] : Employed for multimodal emotion analysis, Self-MM adopts self-supervised tasks to learn representations of different modalities. The framework assigns unimodal tasks with auto-generated labels to each modality to learn effective modality-specific features. These representations are then fused to train classifiers for the emotion analysis task. Self-MM has achieved leading results on the MOSI and MOSEI datasets, demonstrating the efficacy of self-supervised learning in multimodal tasks.

Modality-Adaptive Gate BERT (MAG-BERT) [28] : Integrating a modality-adaptive gate into the BERT model, MAG-BERT optimizes the

fusion process among different modalities. It adjusts the fused modality representation based on modality-specific information, effectively learning cross-modal interactions while preserving the unique characteristics of each modality. MAG-BERT has shown superior performance across various multimodal tasks.

4.2 Experimental Setup

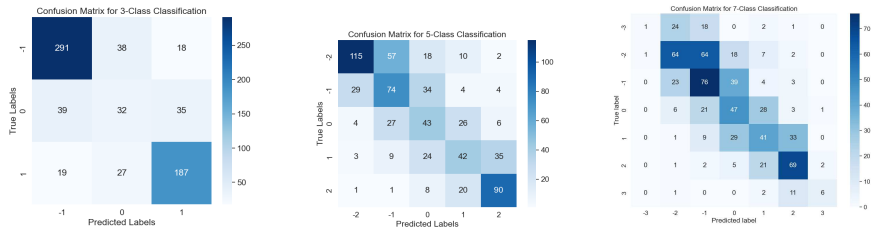
For experiments on the MOSI dataset, a batch size of 64 was employed, with Adam as the optimizer, and the L1 Loss function was utilized. The learning rates were set at 5e-5 for text, 1e-3 for audio, and 1e-4 for video modalities. In the LSTM network, audio and video features had hidden layer sizes of 32 and 64, respectively. An early stopping criterion was implemented after eight iterations, and updates were scheduled every four epochs. For the MOSEI dataset, the batch size was reduced to 32, the audio learning rate was increased to 5e-3, and the video LSTM hidden layer size was decreased to 32, with all other parameters maintained as consistent. Hardware-wise, the experiments ran on a server equipped with an i7-7700 CPU and a Geforce GTX 1080Ti GPU, supported by 32GB of RAM.

4.3 Comparative Overall Performance

To benchmark the results of our designed model, Table 1 presents a comparative analysis of our proposed model against baseline models on the MOSI dataset. In assessing the performance across various models on MOSI, our MMER-TD (Ours) model has demonstrated significant advantages in multiple metrics. Specifically, it achieved 83.53 % in binary accuracy and 83.41 % in F1 Score, surpassing other models. Furthermore, it exhibited commendable performance in Mean Absolute Error, recording a score of 71.710, indicative of high predictive precision. The model also scored the highest in correlation at 79.210, suggesting a strong agreement with the true outcomes. In the more stringent multi-class accuracy tests, our model attained high scores of 52.190 % and 45.480 %, respectively.

Table 1: Comparison of Multimodal Emotion Recognition Models on the MOSI Dataset.

Model	Acc-2	F1(Has/Non)	MAE	Corr
TFN	—/80.8	—/80.7	90.100	69.800
LMF	—/82.5	—/82.4	91.700	69.500
MFN	77.3/—	77.4/—	96.500	63.200
MISA	80.79/82.1	80.77/82.03	80.400	76.400
Selfmm	82.88/84.6	82.78/84.64	72.080	79.550
MAG–Bert	82.5/84.0	82.4/84.0	**71.200**	**79.600**
Ours	**83.53/85.52**	**83.41/85.47**	71.710	79.210



(a) 3-Class Classification. (b) 5-Class Classification. (c) 7-Class Classification.

Fig. 4: Comparative Confusion Matrices for Multiclass Classification in MOSI.

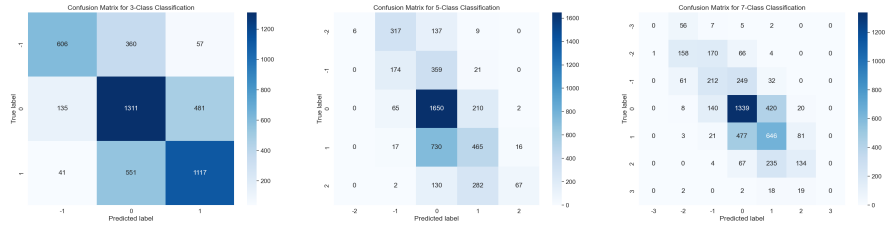
Table 2: Comparison of Multimodal Emotion Recognition Models on the MOSEI Dataset.

Model	Acc-2	F1(Has/Non)	MAE	Corr
TFN	—/82.5	—/82.1	59.300	70.000
LMF	—/82.0	—/82.1	62.300	67.700
MFN	76.0/—	76.0/—	—	—
MISA	82.59/84.23	82.67/83.97	56.800	72.400
Selfmm	82.63/ **84.97**	82.89/**84.83**	53.220	**76.590**
MAG–BERT	81.3/84.8	81.7/84.7	56.700	74.200
Ours	**82.84**/84.81	**83.01**/84.6	**53.160**	75.980

As shown in Table 2, Ours model presented robust performance across several key indicators in the MOSEI dataset comparison. Notably, in accuracy and F1 Score, our model recorded impressive scores of 82.84%/84.81% and 83.01%/84.6%, respectively, indicating superior classification effectiveness. The model's MAE of 53.160 was competitive with the best performances, reflecting accurate prediction capabilities. It also achieved a high correlation coefficient of 75.980, demonstrating consistency in predictions. Facing the challenges of multi-category accuracy, our model continued to lead with scores of 55.500% and 53.880%, further underscoring its exceptional performance in emotion analysis tasks. In summary, MMER-TD (Ours) consistently exhibited superior performance across multiple crucial metrics.

4.4 Classification Recognition Results

To delve into the confusion matrices of our model across two datasets, refer to Figures 4 and 5. These matrices display the performance across different classification tasks within the MOSEI dataset. Confusion matrices are pivotal for assessing the performance of classification models, depicting the relationship between actual and predicted categories. In Figure 4a for the 3-class classification, most of the true "0" labels are correctly predicted as "0", indicated



(a) 3-Class Classification. (b) 5-Class Classification. (c) 7-Class Classification.

Fig. 5: Comparative Confusion Matrices for Multiclass Classification in MOSEI.

by the darkest area of the matrix, suggesting a majority count. However, the model tends to misclassify negative "-1" and positive "1" emotions as "0", with misclassifications of "-1" being notably frequent. Figure 4b for the 5-class classification illustrates a decline in discrimination ability as the number of categories increases. Although predictions for the "0" category remain relatively accurate, errors for other emotional categories rise, particularly for "-1" and "1". Figure 4c for the 7-class classification further demonstrates a decrease in classification performance with an increase in category count. In this more complex classification task, the model struggles to discern subtler emotional differences, especially around the neutral category.

Figure 5a presents the 3-class confusion matrix, where the horizontal axis represents predicted labels and the vertical axis denotes true labels, with positive "1", negative "-1", and neutral "0" emotions. We observe that predictions for positive "1" and negative "-1" emotions are relatively precise, with the recognition of negative emotions being more pronounced, indicating the model's better performance in distinguishing extreme emotions. In Figure 5b for the 5-class classification, as the number of categories grows, the model's identification of neutral "0" emotions predominates, but the predictive dispersion for other emotional categories (-2, -1, 1, 2) increases, suggesting a decrease in performance for differentiating a greater number of emotional categories. Figure 5c for the 7-class classification further reveals the escalating difficulty in predictions with the addition of emotional categories. For emotions beyond neutral "0", especially for the peripheral categories (-3, -2, 2, 3), the accuracy of the model's predictions further declines, highlighting the challenges in recognizing subtle emotional variations.

From these confusion matrices, we observe that as the number of categories increases, the classification difficulty escalates, and the model's predictive accuracy diminishes, particularly for categories with weaker emotional intensity or those close to neutral. This suggests that enhancing the ability to recognize subtle emotional changes poses a challenge in emotion analysis tasks. Moreover, these matrices can guide us in understanding where the model needs

refinement, such as by improving its capability to recognize edge-case emotional categories or by providing more training data to enhance predictions for specific emotions.

4.5 Ablation Studies

4.5.1 Modality Ablation Study

Table 3: Comparison of Multimodal Emotion Recognition Tasks on the MOSI Dataset.

Tasks	Has0_acc_2	Non0_acc_2	Has0_F1_score	Non0_F1_score	Mult_acc_5	Mult_acc_7	MAE	Corr	Loss
M	82.22	84.15	82.12	84.11	52.92	46.79	72.68	79.12	72.58
M+V	82.94	84.91	82.87	84.90	51.60	45.34	72.07	79.59	71.89
M+A	82.94	84.60	82.92	84.63	52.04	44.75	71.24	79.46	70.92
M+T	82.36	84.45	82.27	84.43	**55.10**	47.96	**70.72**	79.35	**70.48**
M+V+A	82.36	83.99	82.30	83.98	54.81	**48.40**	71.68	78.58	71.24
M+V+T	82.65	84.30	82.63	84.33	51.90	44.90	71.99	79.48	72.10
M+T+A	83.09	85.21	82.97	85.16	54.23	46.94	71.69	79.21	71.33
M+V+A+T	**83.53**	**85.67**	**83.42**	**85.63**	52.33	46.50	71.71	**79.59**	71.89

4.5.2 Modality Ablation Study

Table 4: Comparison of Multimodal Emotion Recognition Tasks on the MOSEI Dataset.

Tasks	Has0_acc_2	Non0_acc_2	Has0_F1_score	Non0_F1_score	Mult_acc_5	Mult_acc_7	MAE	Corr	Loss
M	82.41	84.81	82.68	**84.67**	54.36	52.66	54.34	75.52	54.41
M+V	78.77	83.40	79.44	83.42	55.50	53.85	53.83	75.82	53.82
M+A	78.18	83.50	78.94	83.57	54.98	53.23	54.28	75.65	54.33
M+T	77.14	82.70	77.94	82.78	55.03	53.41	**52.99**	**76.54**	**52.99**
M+V+A	78.25	83.93	79.00	84.04	55.46	53.71	54.14	75.75	54.13
M+V+T	79.14	83.62	80.06	83.67	**55.56**	53.74	53.42	76.41	53.44
M+T+A	82.68	84.75	82.96	84.63	55.05	53.39	53.38	76.24	53.38
M+V+A+T	**82.84**	**84.81**	**83.01**	84.60	55.50	**53.88**	53.16	75.98	53.16

The modality ablation experiments aim to assess the specific contributions of different modalities within a multimodal learning framework.

In Table 3, detailing the modality ablation experiments on the MOSI dataset, we observe the performance of various modality combinations across multiple performance metrics. These metrics include binary accuracy (Has0_acc_2, Non0_acc_2), F1 scores (Has0_F1_score, Non0_F1_score), multi-class accuracy (Mult_acc_5, Mult_acc_7), Mean Absolute Error, correlation, and loss. The results demonstrate that combinations of multiple modalities generally yield superior performance across most metrics, particularly notable in Non0_acc_2, Non0_F1_score, and Mult_acc_7. This highlights the complementary nature of each modality, enhancing the model's capability to recognize complex emotional states.

Similarly, Table 4 presents the performance comparison of different modality combinations for the MOSEI dataset.In the context of the MOSEI dataset,

models integrating multiple modalities consistently surpass those relying on single modalities, particularly in metrics like Mult_acc_5 and Correlation. This underscores the substantial enhancement in performance achieved through multimodal fusion..

The data from these tables suggest that in emotion analysis tasks, integrating visual, auditory, and textual information provides richer contextual cues, thereby improving the predictive accuracy and robustness of the model. This multimodal fusion approach is vitally significant for developing efficient and accurate emotion analysis tools. It also suggests that researchers designing multimodal emotion analysis models should consider the characteristics and contributions of each modality to better capture and understand human emotional expressions.

4.5.3 Model ablation studies.

Table 5: Comparison of Multimodal Emotion Recognition Models on the MOSI Dataset.

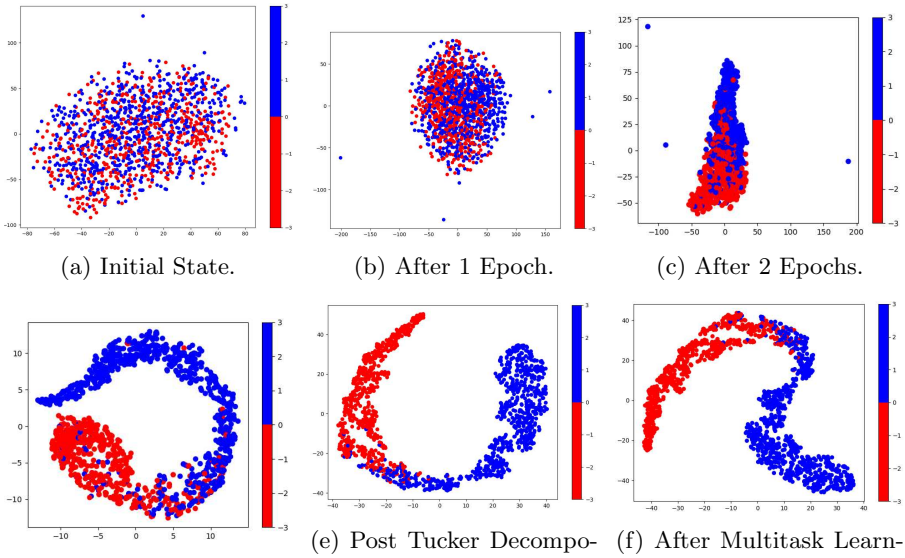
Model	Acc-2	F1(Has/Non)	MAE	Corr	Mult_acc_5	Mult_acc_7	Param
Baseline	80.61/81.86	80.67/81.97	89.890	73.450	45.480	39.800	101.52M
Baseline + TK	81.2/82.77	81.2/82.83	74.410	78.400	51.900	**46.360**	75.36M
Baseline + LG	**83.67**/85.21	**83.66**/85.24	73.580	79.140	49.850	44.310	141.09M
Baseline + TK +LG	83.53/**85.52**	83.41/**85.47**	**71.710**	**79.210**	**52.190**	45.480	**85.85M**

Table 6: Comparison of Multimodal Emotion Recognition Models on the MOSEI Dataset.

Model	Acc-2	F1(Has/Non)	MAE	Corr	Mult_acc_5	Mult_acc_7	Param
Baseline	77.14/82.39	78.12/82.62	57.430	72.690	53.470	51.920	101.52M
Baseline + TK	78.39/82.61	79.14/82.69	53.540	76.280	55.430	53.780	75.36M
Baseline + LG	**82.95**/84.78	**83.24**/84.25	53.670	**76.510**	54.960	53.220	141.09M
Baseline + TK +LG	82.84/**84.81**	83.01/**84.6**	**53.160**	75.980	**55.500**	**53.880**	**85.85M**

To investigate the specific contributions of the TK and LG modules to model performance, this paper presents a detailed analysis of a series of ablation experiments conducted on the MOSEI and MOSI datasets. These results are presented in tables, comparing key performance metrics of the model under different configurations, including binary classification accuracy, F1 score), mean absolute error, correlation, multi-class accuracy, and the number of model parameters (Param).

The results clearly demonstrate that the introduction of the TK and LG modules significantly enhances the model's performance on most metrics.



(a) Initial State. (b) After 1 Epoch. (c) After 2 Epochs.

(d) Post Tensor Fusion. (e) Post Tucker Decomposition. (f) After Multitask Learning.

Fig. 6: Stages of Model Training and Transformation: From Initial State to Multitask Learning via Tensor Fusion and Tucker Decomposition.

This improvement is evident in both MOSEI and MOSI datasets, particularly notable in the MOSEI dataset. Specifically, the incorporation of the TK and LG modules resulted in enhancements in Acc-2, F1, and Mult_acc_5/7 metrics, affirming their positive role in the model, effectively boosting overall performance.

For instance, in the MOSI dataset (as seen in Table 5), the model showed improvements in Acc-2, F 1, MAE, and Corr metrics after the integration of TK and LG modules. Similarly, in the MOSEI dataset (refer to Table 6), the same configuration also demonstrated improvements in these key metrics, especially in Acc-2 and F1 scores, which increased by nearly 2 percentage points. Notably, despite the enhanced performance, the number of model parameters decreased, indicating that the inclusion of TK and LG modules not only improved the model's performance but also its parameter efficiency.

In summary, the experiments conclusively show that the incorporation of the TK and LG modules significantly enhances the overall performance of the model in sentiment analysis on the MOSI and MOSEI datasets, while also improving parameter efficiency. Therefore, the TK and LG modules represent a beneficial improvement to the sentiment analysis model, enhancing its performance while maintaining efficient parameter configuration.

Table 7: Analysis of Movie Sentiments.

Case	Label	M	T	A	V
and he was still boring	-1.8	-1.6802	-1.7494	-0.4231	-0.4963
um yeah the special effects were ok	0	1.1852	2.0493	-0.6968	-0.4917
nothing that's worth going to theaters to see	-2.6	-0.979	-1.0471	-0.6644	-0.495
and yeah i think the only character who actually showed their motivation behind anything they did was the villain	-1.8	-0.5788	-0.441	-0.4388	-0.4917
and then Cameron Diaz who's supposed to be plucky i guess but she's just kind of useless	-1.8	-1.3343	-1.3906	-0.6374	-0.4928
hi I'm pretty I have a giant smile I'm supposed to know things um walk off screen	-0.8	0.5021	1.8813	-0.6185	-0.4908
um yeah the characters didn't care whether they lived or died	-2.4	-1.1519	-1.163	-0.6402	-0.4909
Kato was the only one that was any good whatsoever	-0.6	-1.2484	-1.1906	-0.5077	-0.4926
he was the only character that was slightly interesting	-0.8	1.0822	2.1148	-0.4484	-0.4898
who seemed to have genuine affection for other human beings	0.6	0.7664	2.3568	-0.6645	-0.4914
oh my gosh bad movie	-2.8	-1.8624	-1.7539	-0.5705	-0.4885
i wish i weren't	-0.8	-1.429	-1.4384	-0.6019	-0.4956
really bad movie	-2.6	-1.5202	-1.4861	-0.6417	-0.4979
but basically Seth Rogen brought none of the charm that he normally does to his roles	-2.2	-1.3087	-1.4014	-0.7262	-0.4899

4.6 Visualization

To further illustrate the regression process and superiority of the model, Figure 6a displays the initial distribution of the dataset, where red and blue points may represent different categories or distinct characteristics of data points. Figures 6b and 6c for epoch 1 and epoch 2: These figures depict the state of the data after 1 and 2 training epochs, respectively. As training progresses, data points begin to rearrange in a certain pattern, indicating that the model is learning to differentiate between various categories or features. Figure 6d Tensor Fusion: This figure represents the state where text, images, and audio are fused together in a multimodal learning context. Tensor fusion is commonly used to combine different types of data to extract richer features. Figure 6e Tucker Decomposition: This diagram showcases the tensor decomposition performed using Tucker decomposition technique. Tucker decomposition is a multilinear algebraic tool for breaking down multidimensional arrays (tensors) to discover underlying structures and reduce the dimensionality of data. Figure 6f Final: The last image represents the state after training is completed. At this point, data points are clearly separated by category or characteristic, indicating that the model can effectively distinguish between different data points.

4.6.1 Case Analysis

In the course of our experiments, we observed that the textual modality contributed most significantly among the three modalities. As illustrated in the accompanying figure, an analysis of the table reveals that the emotional labels of certain sentences align with the scores in the "M" column, derived from multimodal analysis. For instance, consider the sentence "um yeah the special effects were ok," with an actual label of "0" indicating a neutral emotion. The "M" column scores it as "1.1852," which is close to neutral, suggesting that the text modality's analysis in this case coincides with the actual emotion. As shown in Table 7. In other cases, such as "because I truly love an action flick, action comedy flick even better, right," the actual label is "1.6," denoting a positive emotion. The "M" column score is also positive at "2.2367," indicating that the textual modality's analysis aligns with the actual emotion once again, which demonstrates the high accuracy of emotional analysis in the text modality. However, for a more comprehensive evaluation of the model's performance, we should also consider the analysis results from the "T," "A," and "V" columns, as multimodal analysis provides a richer and more accurate emotional recognition by integrating various types of data.

5 CONCLUSION

This study introduces a multimodal emotion recognition method that synergizes tensor decomposition fusion with self-supervised multitask learning. By employing Tucker decomposition, the model significantly reduces its parameter count, mitigating the risk of overfitting. Concurrently, it establishes a learning mechanism for both multimodal and unimodal tasks, incorporating the concept of label generation to capture the emotional variances across different modalities with greater precision. Experiments on public datasets such as CMU-MOSI and CMU-MOSEI have demonstrated notable performance improvements with this method. The approach presented in this research markedly excels in the domain of multimodal emotion recognition. By amalgamating tensor decomposition fusion with self-supervised multitask learning, this method not only enhances the accuracy of emotion recognition but also simplifies the model complexity, showcasing tremendous potential for practical applications. With the continuous evolution and refinement of technology, this multimodal emotion recognition method is anticipated to play a significant role in the field of artificial intelligence, especially in applications related to human-computer interaction and social media analysis.

Acknowledgments. This work was supported by National Natural Science Foundation of China(61771299)

Declarations

- Conflict of interest/Competing interests: There is no conflict of interest in our work.

- Availability of data: Data openly available in a public repository. The data that support the findings of this study are openly available in Mosei and Mosi at http://immortal.multicomp.cs.cmu.edu/.
- Authors' contributions:Zhu Jiawei, Wang Rui and Zhu Xianxun contributed the central idea, analyzed most of the data and wrote the first draft of the paper. The rest of the authors devoted themselves to perfecting these ideas, made additional analysis, and finally completed this paper.

References

[1] Khang, A., Gupta, S. K., Rani, S., & Karras, D. A. (Eds.). (2023). Smart Cities: IoT Technologies, big data solutions, cloud platforms, and cybersecurity techniques. CRC Press.

[2] Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2023). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. Expert Systems with Applications, 121692.

[3] Pan, B., Hirota, K., Jia, Z., & Dai, Y. (2023). A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. Neurocomputing, 126866.

[4] Liu, W., Qiu, J. L., Zheng, W. L., & Lu, B. L. (2019). Multimodal emotion recognition using deep canonical correlation analysis. arXiv preprint arXiv:1908.05349.

[5] Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. Neural Computation, 32(5), 829-864.

[6] Wang, Q., Wang, J., Quan, X., Feng, F., Xu, Z., Nie, S., ... & Liu, D. (2023, July). Mustie: Multimodal structural transformer for web information extraction. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2405-2420).

[7] Yang, H., Yin, L., Zhou, Y., & Gu, J. (2021). Exploiting semantic embedding and visual feature for facial action unit detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(pp. 10482-10491).

[8] Yin, D., Meng, T., & Chang, K. W. (2020). Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. arXiv preprint arXiv:2005.04114.

[9] Yang, K., Xu, H., & Gao, K. (2020, October). Cm-bert: Cross-modal bert for text-audio sentiment analysis. In Proceedings of the 28th ACM

international conference on multimedia (pp. 521-528).

[10] Park, G., Han, C., Yoon, W., & Kim, D. (2020). MHSAN: multi-head self-attention network for visual semantic embedding. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1518-1526).

[11] Kim, T., & Lee, B. (2020, June). Multi-attention multimodal sentiment analysis. In Proceedings of the 2020 International Conference on Multimedia Retrieval(pp. 436-441).

[12] Levy, R., Luo, D., & Clark, B. K. (2024). Classical shadows for quantum process tomography on near-term quantum computers. Physical Review Research, 6(1), 013029.

[13] Jones, L. O., Mosquera, M. A., Schatz, G. C., & Ratner, M. A. (2020). Embedding methods for quantum chemistry: Applications from materials to life sciences. Journal of the American Chemical Society, 142(7), 3281-3295.

[14] Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014, May). COVAREP—A collaborative voice analysis repository for speech technologies. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 960-964). IEEE.

[15] Yuan, X., Li, L., & Wang, Y. (2019). Nonlinear dynamic soft sensor modeling with supervised long short-term memory network. IEEE Transactions on Industrial Informatics, 16(5), 3168-3176.

[16] Li, J., Zhang, X., Li, F., Duan, S., & Huang, L. (2024). Acoustic-articulatory emotion recognition using multiple features and parameter-optimized cascaded deep learning network. Knowledge-Based Systems, 284, 111276.

[17] Lee, C., Kim, S., Han, D., Yang, H., Park, Y. W., Kwon, B. C., & Ko, S. (2020, April). GUIComp: A GUI design assistant with real-time, multi-faceted feedback. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-13).

[18] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.

[19] Malik, O. A., & Becker, S. (2018). Low-rank tucker decomposition of large tensors using tensorsketch. Advances in Neural Information Processing Systems, 31.

[20] Tellamekala, M. K., Amiriparian, S., Schuller, B. W., André, E., Giesbrecht, T., & Valstar, M. (2023). COLD fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[21] AmirZadeh, Rowan Zellers, Eli Pincus, & Louis-Philippe Morency. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. CoRR, abs/1606.06259, 2016.

[22] Amir Zadeh. CMU-MOSEI dataset. http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/, 2018. Accessed: 2018.

[23] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.

[24] Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064.

[25] Tsai, Y. H. H., Liang, P. P., Zadeh, A., Morency, L. P., & Salakhutdinov, R. (2018). Learning factorized multimodal representations. arXiv preprint arXiv:1806.06176.

[26] Hazarika, D., Zimmermann, R., & Poria, S. (2020, October). Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1122-1131).

[27] Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021, May). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 12, pp. 10790-10797).

[28] Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L. P., & Hoque, E. (2020, July). Integrating multimodal information in large pretrained transformers. In Proceedings of the Conference. Association for Computational Linguistics. Meeting (Vol. 2020, p. 2359). NIH Public Access.