TensorFormer: A Tensor-Based Multimodal Transformer for Multimodal Sentiment Analysis and Depression Detection

Hao Sun, Yen-Wei Chen, *Member, IEEE*, and Lanfen Lin, *Member, IEEE*

Abstract—Sentiment analysis is an important research field aiming to extract and fuse sentimental information from human utterances. Due to the diversity of human sentiment, analyzing from multiple modalities is usually more accurate than from a single modality. To complement the information between related modalities, one effective approach is performing cross-modality interactions. Recently, Transformer-based frameworks have shown a strong ability to capture long-range dependencies, leading to the introduction of several Transformer-based approaches for multimodal processing. However, due to the built-in attention mechanism of the Transformers, only two modalities can be engaged at once. As a result, the complementary information flow in these Transformer-based techniques is partial and constrained. To mitigate this, we propose, TensorFormer, a tensor-based multimodal Transformer framework that takes into account all relevant modalities for interactions. More precisely, we first construct a tensor utilizing the features extracted from each modality, assuming one modality is the target while the remaining tensors serve as the sources. We can generate the corresponding interacted features by calculating source-target attention. This strategy interacts with all involved modalities and generates complementing global information. Experiments on multimodal sentiment analysis benchmark datasets demonstrated the effectiveness of TensorFormer. In addition, we also evaluate TensorFormer in another related area: depression detection and the results reveal significant improvements when compared to other state-of-the-art methods.

Index Terms—Depression detection, modality interactions, multimodal learning, sentiment analysis, transformer

1 Introduction

With the rapid growth of social media in recent years, there is a huge increase in the amount of user-generated content on social media, such as videos. Automatic sentiment analysis from different modalities is beneficial to human-computer interaction and has attracted a great deal of research interest [1], [2]. Multimodal sentimental analysis (MSA) aims to capture and integrate sentimental information from different relevant modalities to predict the sentimental state or tendency toward the speakers. Typically, there are three modalities we can obtain from videos: visual, acoustic, and textual.

However, the subtle sentiment information of different modalities is not always identical. For example, when a boy says "It is raining" with a happy tone and peaceful expression, we cannot infer his sentiment from visual or textual information because they are neutral, but we can infer his sentiment from his acoustic tone because it's relatively positive. Therefore, it is important to design a fusion scheme that can leverage the information spread across different modalities.

In multimodal learning, it is proven that the information learned from different modalities should be integrated to describe the semantics more holistically [3]. Due to the subtlety of human psychological state, the comprehensive information exchange can better ensure the complementarity and consistency of multimodal semantics. As in the acoustic sense in the previous example, complementarity means that a modality can provide information that is missing in other modalities. The consistency indicates that the learned information should refer to the same semantics. For example, it is not suitable that one modality discusses the sentiment tendency while other modalities focus on gender. At present, one efficient approach to ensure the comprehensive information exchange is to design structures that perform interactions between modalities [4], [5], [6]. For this purpose, many researchers have used Transformer's Query-Key-Value (QKV) operations to perform cross-modality interactions because of its impressive progress in natural language processing (NLP) and computer vision (CV). For instance, Delbrouck et al., [7] proposed a Transformer-based joint-encoding scheme to encode dual modalities with a single encoder whereas Han et al., [8] used Transformer's self-attention mechanism to blend multimodal features using text-acoustic and text-visual modality pairs. However, the Transformers' design structure has certain limitations, particularly the fact that it allows only two modality interactions at once during cross-modal interactions, as shown in Fig. 1a. Thus, to fully exchange multimodal information, C_N^2 bi-modality interactions are

- *Hao Sun and Lanfen Lin are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China. E-mail: {sunhaoxx, llf}@zju.edu.cn.*
- *Yen-Wei Chen is with the College of Information Science and Engineering, Ritsumeikan University, Shiga 603-8577, Japan. E-mail: chen@is.ritsumei.ac.jp.*

Manuscript received 24 May 2022; revised 22 December 2022; accepted 27 December 2022. Date of publication 30 December 2022; date of current version 29 November 2023.
(Corresponding author: Lanfen Lin.)
Recommended for acceptance by E. Cambria.
Digital Object Identifier no. 10.1109/TAFFC.2022.3233070

1949-3045 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

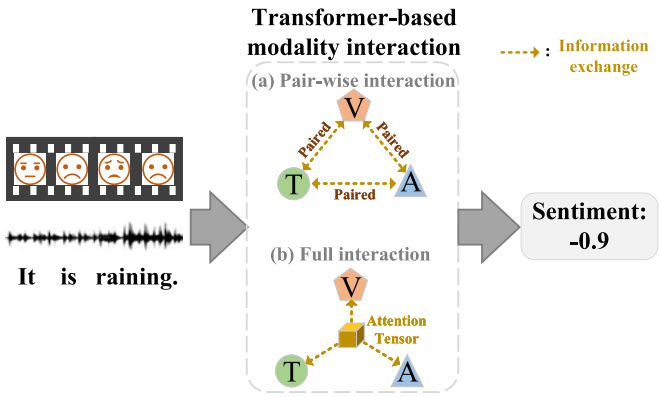


Fig. 1. Different transformer-based modality interaction modes. Sentiment analysis is used as an example. For pair-wise interaction, the information exchange is encountered in each pair while the remaining modality is ignored. For full- interaction, the information exchange can take all involved modalities into account at meanwhile.

required when processing N modalities. This multi-step processing has two obvious disadvantages: the first is the lack of scalability, and the second is the insufficient information exchange because of the neglect of remaining modalities. These two shortcomings will further affect the accuracy of mental state prediction.

To address this problem, we propose a tensor-based multimodal Transformer called *TensorFormer*, which allows all concerned modalities to exchange all relevant information simultaneously. A global cross-attention module and a parallel feed-forward module are included in our TensorFormer. Different from the conventional self-attention mechanism, we introduce an attention tensor to perform comprehensive cross-modality interactions in the global cross-attention module, adapting to the characteristics of multimodal data processing. The intuitive scheme is shown in Fig. 1b. Specifically, we first construct the attention tensor using all involved (N) modalities. Then each modality is chosen as the target while the remaining $(N-1)$ modalities in the attention tensor are treated as the corresponding source. We can obtain interacted features with comprehensive information by calculating the target-source attention. Additionally, the global interaction can be carried out more concisely by mathematical derivations of the conventional QKV mode. The parallel feed-forward module is composed of several multilayer perceptrons (MLP) for each modality, aiming to process the channel-wise features parallelly. Apart from that, we can stack TensorFormer blocks for multiple layers to process more complex information, similar to other Transformer-based methods [9]. During the learning process, TensorFormer can also be trained in an end-to-end manner. Moreover, as we aim to perform global cross-modality interaction simultaneously, TensorFormer is designed with good scalability, making it simple to extend to more complex scenarios where there are more than three modalities.

We evaluate TensorFormer using two widely used MSA benchmark datasets, CMU-MOSI and CMU-MOSEI. As revealed by Qureshi et al., [10], due to the close relationship between sentiment analysis and video-based depression diagnosis, many common methods apply to both fields. Thus we conduct further experiments on the depression detection dataset, AVEC2019. The experimental results demonstrate TensorFormer can reach better performance compared with other state-of-the-art methods on almost all metrics. The ablation analysis also indicates the effectiveness of TensorFormer. Our contributions can be summarized as follows:

- We propose TesnorFormer, a novel tensor-based multimodal Transformer for MSA and depression detection. Compared with previous multimodal works, TensorFormer can exchange global cross-modality information more effectively.
- We formulate an attention tensor, a tensor-based cross-modality attention mechanism. Attention tensor not only considers the characteristics of different modalities but also performs interactions with all concerned modalities at the same time. We formulate an attention tensor, a tensor-based cross-modality attention mechanism.
- The results of experiments on the widely used benchmark datasets demonstrate the effectiveness of TensorFormer, outperforming the state-of-the-art approaches.

The rest of the paper is structured as follows: In Section 2, we describe related works including multimodal data interactions and methods used in sentiment analysis and depression detection. In Section 3, we introduce our proposed approaches in detail including the construction of attention tensor and the structure of TensorFormer. Section 4 discusses our experimental details on three different datasets. Section 5 details our results, comparisons with other methods, and some further analysis, which demonstrate the effectiveness of our proposed methods. The conclusions are presented in Section 6 with some future work plans.

2 Related Works

Research in multimodal sentiment analysis and depression detection mainly focuses on extracting effective features from sequential multimodal data and integrating them to predict comprehensive sentiment or depression tendencies. The majority of current approaches apply deep neural networks [11], [12], [13] or analyze the mental states on the utterance level [14].

2.1 Multimodal Sentiment Analysis and Depression Detection

Most sentiment and depression analysis tasks rely on spoken words and human utterances. In earlier times sentiment analyses are mainly based on syntactic representations of text, in which sentiment is explicitly expressed [15], [16], [17]. But Poria et al., [18] explored several different architectures and demonstrated that multimodal learning yields superior results to unimodal learning. The majority of current multimodal sentiment analysis benchmarks focus on user-generated videos, in which there are at least three modalities [19], [20], [21], [22], [23]. The challenges of multimodal sentiment analysis are in extracting sentimental features effectively and ensuring their complementarity and consistency. Works in this area usually focus on exchanging information between the extracted features and fusing them

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

before predictions [4], [24], [25], [26]. Nowadays with the growth of computing hardware, people attempt to predict another long-term human mind state: depression level [27], [28], [29]. Compared to sentiments, depression level is more obscure and needs longer sequences to diagnose, which brings another challenge. To this end, some works like [30] attempted to employ Transformer to process the longer sequential data.

2.2 Multimodal Data Interaction

Multimodal data interaction is an important approach for exchanging information across different modalities. There are many different modality interaction strategies designed for sentiment analysis and depression detection. For example, Cambria et al., [31] proposed sentic blending that could enable continuous multimodal semantics and sentics. And Tensor Fusion Network (TFN) is designed by Zadeh et al., [4] to fuse cross-modality information from all three modalities (textual, acoustic, and visual modality) using the Cartesian product. To process the sequential signals, Zadeh et al., [24] followingly proposed a Memory Fusion Network which can process the involved signals in a Recurrent Neural Network (RNN) manner. To obtain the correlation between modalities, some conventional approaches use attention mechanisms. Chen et al., [25] developed a Sentimental Words Aware Fusion Network (SWAFN) to calculate coattention scores between text and other modalities. Deng et al., [32] proposed a deep dense fusion network with multimodal residual (DFMR) to integrate multimodal information in a paired manner. Han et al., [33] also designed a hierarchical attention network to identify the relative importance of multimodal features at different context levels. Some methods aim to use alternative learning strategies for sentiment analysis. Peng et al., [34] used modality distillation, while works like [35], [36], [37] used word-level reinforcement learning and domain knowledge to carry out modality interactions. Most recently, Hazarika et al., [38] developed a framework to learn modality-invariant and modality-specific representations.

Other methods have been proposed to address specific multimodal interaction problems in particular scenarios, in addition to general multimodal data interactions. For example, Hazarika et al., [39] emphasized the role of inter-speaker dependency to classify emotions during conversations whereas Tu et al., [40] argued that it is important to analyze the context and common-sense knowledge for multimodal data interactions. Chen et al., [41] addressed the issue that nonverbal cues like facial expressions are indicative of depressive disorders, and proposed a chained-fusion mechanism to jointly learn facial appearance, showing that sequential fusion provides a clear probabilistic perspective of the model correlation in multimodal interactions.

However, these methods are limited due to the lack of the ability to capture long-term dependencies between different modalities (e.g., the forgetting issue in RNN and LSTM). The analysis of mental states is greatly hampered by the ineffective interaction of different modalities over long temporal multimodal series, which has attracted a lot of attention.

2.3 Transformer-Based Data Interaction

Recently with the great success of Transformer in NLP and CV, many Transformer-based methods are proposed to perform cross-modality interaction. Compared with previous methods, Transformer is more effective for capturing long-term correlation between different modalities. Delbrouck et al., [7] proposed, for example, a Transformer-based joint-encoding (TBJE) that relies on a modular co-attention layer to jointly encode one or two modalities. Tsai et al., [42] proposed a Multimodal Transformer (MulT) for pair-wise cross-model attention. To execute latent signal interactions, Sahay et al., [43] proposed a low-rank version of MulT (LMF-MulT). Moreover, Wang et al., [44] introduced a TransModality, an end-to-end sophisticated structure that can incorporate information from the source modality to the target modality. Han et al., [8] treated text as the main modality while using acoustic and visual modalities to enhance the textual information in a symmetric structure. Aside from these methods, some other researchers also introduced theories in perception literature. For example, Bao et al., [45] implemented the Global Workspace Theory and combined long short-term memory networks (LSTM) [46] and Transformer to simulate the human perception process.

Most Transformer-based approaches adopted the traditional QKV self-attention operations to perform the bi-modality interactions. More specifically, these approaches first set the main modality features as Q and set auxiliary modality features as K and V, then perform the cross-modality interaction

$$O = Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V, \tag{1}$$

where d_k is a scaled factor and O is the interacted features. The QKV operation is effective for exchanging information between two modalities. But one of the limitations is that Transformer cannot exchange information among three or more modalities simultaneously. This limitation causes insufficient information exchange among different modalities, which can reduce prediction accuracy. Our approach is also based on Transformer, but with some improvements for global multimodal processing with three or more modalities.

3 Approaches

In the task of multimodal sentiment analysis or depression detection, there are generally three modalities: t(textual), a(acoustic), and v(visual). The three modalities are 2-D sequences represented as $X_m \in R^{T_m \times d_m}$ where T_m and d_m are the sequential lengths and feature vector size of modality $m \in \{t, a, v\}$. The target is to perform the cross-interactions between X_t, X_a, and X_v and determine the underlying sentiment or depression intensity.

3.1 Overall Pipeline

The overview of our pipeline for multimodal sentiment analysis and depression detection is shown in Fig. 2. We first extract the features from involved modalities and then accomplish the comprehensive and global cross-modality interactions by TensorFormer. To generate contextualized word embeddings X_t from raw sentences in textual modality, we use the pre-trained model, BERT [47] as the feature extractor.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

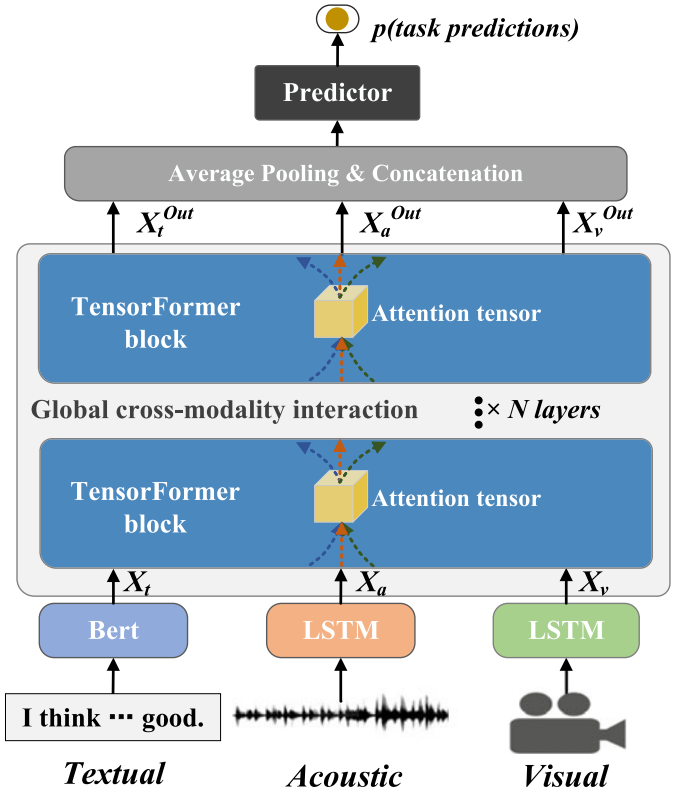


Fig. 2. The overview of our pipeline for multimodal sentiment analysis and depression detection with TensorFormer.

We use bidirectional LSTM [46] to capture the internal dependency of acoustic and visual modalities and use the hidden states as extracted embeddings X_a and X_v. The extracted features are then fed into N layer TensorFormer blocks, which completely exchange information between modalities in a tensor-based manner. The processed multimodal features $X_m^{Out}(m \in \{t, a, v\})$ are then average pooled and concatenated to generate the final predictions.

3.2 Global Cross-Modality Interaction

To perform the comprehensive cross-modality interactions, we design TensorFormer, whose basic block structure is shown in Fig. 3a. For generality and scalability, we represent input modality features as X_α, X_β and X_γ to illustrate the structure of TensorFormer where $X_m \in R^{T_m \times d_m}$ and $m \in \{\alpha, \beta, \gamma\}$. Before the information exchange, linear transformations $W_{P_m}(m \in \{\alpha, \beta, \gamma\})$ is adopted to project the features into the same dimension X_m^P. There are two main modules in the TensorFormer block, the *global cross-attention* module, and the *parallel feed-forward* module. The global cross-attention module aims to perform the full modality interaction. The core component of this module is *attention tensor* with its corresponding queries that are designed to calculate cross-modality attention scores comprehensively and simultaneously. The parallel feed-forward module is composed of several branches to process the channel-wise information for corresponding modalities.

3.2.1 Attention Tensor

The key idea of TensorFormer's global cross-attention mechanism is the attention tensor, a tensor containing important and relevant information from all modalities. To generate the attention tensor, we first perform average pooling per sequential point on input features $X_m^P = [x_m^1, x_m^2, \ldots, x_m^{T_m}]$ to obtain the expected features $\bar{X}_m^P = [\mathbb{E}[x_m^1], \mathbb{E}[x_m^2], \ldots, \mathbb{E}[x_m^{T_m}]]$ where $\bar{X}_m^P \in R^{T_m \times 1}$ and $m \in \{\alpha, \beta, \gamma\}$. Then we apply the Cartesian product to generate the attention tensor

$$A = \{(x_\alpha, x_\beta, x_\gamma) \,|\, x_\alpha \in \bar{X}_\alpha^P, x_\beta \in \bar{X}_\beta^P, x_\gamma \in \bar{X}_\gamma^P\}, \tag{2}$$

which is mathematically equivalent to a differentiable outer product among X_α, X_β and X_γ [4]

$$A = \bar{X}_\alpha^P \otimes \bar{X}_\beta^P \otimes \bar{X}_\gamma^P. \tag{3}$$

To put the implementation of Equation 3 more precisely, we first perform outer product between two random modality features (take $\bar{X}_\alpha^P$ and $\bar{X}_\beta^P$ for example), the flattened results $u_{\alpha\beta}$ will be multiplied to $\bar{X}_\gamma^{P\top}$ to generate flattened attention tensor $v_{\alpha\beta,\gamma}$, then we unfold $v_{\alpha\beta,\gamma}$ to obtain the final attention tensor A

$$\begin{aligned} u'_{\alpha\beta} &= \bar{X}_\alpha^P \otimes \bar{X}_\beta^{P\top} \in R^{T_\alpha \times T_\beta}, \\ u_{\alpha\beta} &= Flatten(u'_{\alpha\beta}) \in R^{(T_\alpha T_\beta)\times 1}, \\ v_{\alpha\beta,\gamma} &= u_{\alpha\beta} \otimes \bar{X}_\gamma^{P\top} \in R^{(T_\alpha T_\beta)\times T_\gamma}, \\ A &= Unfold(v_{\alpha\beta,\gamma}) \in R^{T_\alpha \times T_\beta \times T_\gamma}. \end{aligned} \tag{4}$$

In tensor A, every score $A^{i,j,k}$ represents the attention score among ith element of modality α, jth element of modality β and kth element of modality γ, and is calculated as

$$A^{i,j,k} = \bar{X}_\alpha^{Pi} \times \bar{X}_\beta^{Pj} \times \bar{X}_\gamma^{Pk}. \tag{5}$$

3.2.2 Attention for Global Cross-Modality Interaction

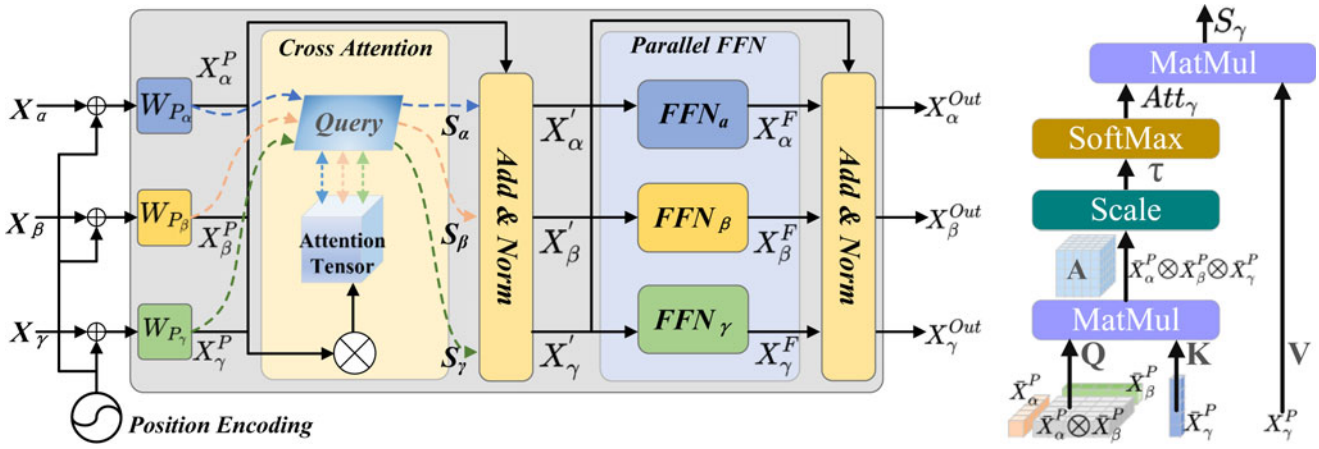
The attention tensor incorporates comprehensive information from all involved modalities. Each modality can inquire about the attention score from A, which contains the information of all other modalities. Here, we first present detailed illustrations in the conventional QKV manner, then derive a more straightforward demonstration of the query procedure. To have a better correspondence with Equation (4), we take X_γ^P as an example to perform the query, as shown in Fig. 3b. During the procedure, X_γ^P can get the complemented attention score from X_α^P and X_β^P. Before a query operation, Q, K and V are first created as

$$\begin{aligned} Q &= u_{\alpha\beta} \in R^{(T_\alpha T_\beta)\times 1}, \\ K &= \bar{X}_\gamma^P \in R^{T_\gamma \times 1}, \\ V &= X_\gamma^P \in R^{T_\gamma \times d_\gamma}. \end{aligned} \tag{6}$$

Then we employ the cross-attention mechanism to calculate the attention scores and weighted summary of modality γ

$$\begin{aligned} S_\gamma &= softmax_\gamma\left(\frac{QK^\top}{\tau}\right)V \\ &= softmax_\gamma\left(\frac{u_{\alpha\beta}\bar{X}_\gamma^{P\top}}{\tau}\right)X_\gamma \\ &= softmax_\gamma\left(\frac{A}{\tau}\right)X_\gamma \\ &= Att_\gamma X_\gamma, \end{aligned} \tag{7}$$

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.



(a) The structure of TensorFormer block. ⊗ means the Cartesian product. The *Parallel FFN* means *parallel feed-forward* module.

(b) Cross-modality interaction.

Fig. 3. The overview structure of TensorFormer Block and the construction of attention tensor. In the example three modalities are used for the illustration.

where $Att_\gamma \in R^{T_\alpha \times T_\beta \times T_\gamma}$ is attention score, $S_\gamma \in R^{T_\alpha \times T_\beta \times d_\gamma}$ is weighted summary of modality γ, τ is a unified regularization coefficient, $softmax_\gamma$ is softmax operation on γ axis of $A \in R^{T_\alpha \times T_\beta \times T_\gamma}$ which can be further formulated as

$$softmax_\gamma(A) = \frac{\sum_{i=1}^{T_\alpha}\sum_{j=1}^{T_\beta} e^{A^{i,j,k}}}{\sum_{i=1}^{T_\alpha}\sum_{j=1}^{T_\beta}\sum_{k'=1}^{T_\gamma} e^{A^{i,j,k'}}}, k = 1, \ldots, T_\gamma \quad (8)$$

Therefore, according to Equation (7), we can obtain more direct and efficient expressions for each modality's query

$$\begin{aligned} S_\alpha &= Att_\alpha X_\alpha = softmax_\alpha(A/\tau)X_\alpha^P, \\ S_\beta &= Att_\beta X_\beta = softmax_\beta(A/\tau)X_\beta^P, \\ S_\gamma &= Att_\gamma X_\gamma = softmax_\gamma(A/\tau)X_\gamma^P, \end{aligned} \quad (9)$$

which have great significance for multimodal information fusion. On the one hand, the operations in Equation (9) allow us to perform the multimodal attention query a more concise tensor manner. On the other hand, when querying the attention tensor, the modality that executes the query will see the information of all other modalities at the same time. Finally, the weighted summary S_m will be average-pooled and added to the origin residual X_m^P

$$X'_m = avgpooling(S_m) + X_m^P, \quad (10)$$

where $S_m \in R^{T_{m_1} \times T_{m_2} \times d_m} (m_1, m_2 \in \{M \setminus m\})$, M is the set of all modality kinds and average pooling is performed across T_{m_1} and T_{m_2} axes. $X'_m \in R^{T_m \times d_m}$ is the queried X_m with information from other modalities. It should be mentioned that the average-pooled S_m is in shape R^{1*1*d_m}. To add with $X_m^P \in R^{T_m * d_m}$, we first squeezed S_m to R^{1*T_γ} and duplicated it to $S'_m \in R^{T_m * d_m}$ along the first axis. Then $S'_m \in R^{T_m * d_m}$ and $X_m^P \in R^{T_m * d_m}$ can be added directly.

3.3 Parallel Feed-Forward

The *parallel feed forward* structure of the TensorFormer block is composed of three parallel branches for respective modalities (α, β, and γ). For each modality m, the feed-forward module FFN_m aims to process features X'_m in channel-wise and is composed of two linear transformations

$$X_m^F = FFN_m(X'_m) = W_{F_m^2} W_{F_m^1} X'_m, \quad (11)$$

where $W_{F_m^1} \in R^{d_{hid} \times d_m}$ and $W_{F_m^2} \in R^{d_m \times d_{hid}}$. The processed information X_m^F is then added to the residual X'_m to produce TensorFormer's final output $X_m^{Out} \in R^{T_m \times d_m}$

$$X_m^{Out} = X'_m + X_m^F. \quad (12)$$

3.4 More Modality Cases

TensorFormer can also handle more modality cases well. Suppose there are M modalities $X_1^P, X_2^P, \ldots, X_M^P$, we can construct the attention tensor in the same approach

$$A = \bar{X}_1^P \otimes \bar{X}_1^P \otimes \ldots \otimes \bar{X}_M^P. \quad (13)$$

$A \in R^{T_1 \times T_2 \times \ldots \times T_M}$ is the tensor with attention scores from all M sequences. Then we query the weighted summary of each modality

$$\begin{aligned} S_1 &= softmax_1(A/\tau)X_1^P, \\ S_2 &= softmax_2(A/\tau)X_2^P, \\ &\ldots \\ S_M &= softmax_M(A/\tau)X_M^P. \end{aligned} \quad (14)$$

The overall process is shown in Fig. 4. The residual operation and the feed-forward procedure can also be represented by Equations (10), (11), and (12). As a result, we can say, our proposed TensorFormer has excellent scalability and is simple to adapt to more than three modalities.

3.5 Loss Function in Training

The task of multimodal sentiment analysis is a regression problem. To predict the sentiment intensity, we employ mean absolute error (MAE) as a loss function to guide the learning process which can be shown as

$$\mathcal{L}_{MAE} = \frac{1}{D}\sum_{i=1}^{D} |y_i - p_i|, \quad (15)$$

where D is the number of samples, y_i and p_i are the ground truth label and prediction of ith sample respectively. We

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

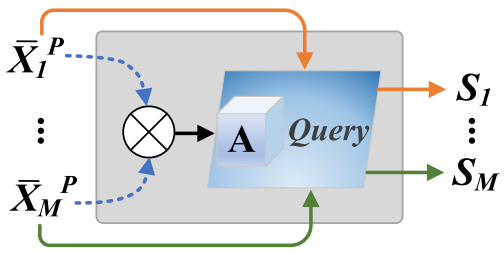


Fig. 4. The construction and query procedure of global cross-modality attention with more modalities.

choose MAE as our loss function since we observe it performs better than other functions during training, such as mean squared error (MSE).

For depression detection, we employ the Concordance Correlation Coefficient (CCC) loss as the cost function

$$\mathcal{L}_{CCC} = 1.0 - \frac{2C_{py}}{C_p^2 + C_y^2 + (\bar{p} - \bar{y})^2}, \tag{16}$$

where C_p and C_y are the variance of predictions and labels, C_{py} is the covariance between them. CCC is very suitable for estimating depressions [27], [28], because it can not only calculate precise correlation and accuracy but also is not unbiased by changes in scale and location [48]. The CCC ranges from -1 to 1, with 1 denoting the ideal positive correlation and -1 denoting the completely negative correlation.

4 Experiments

4.1 Datasets

We use two popular MSA benchmark datasets and one depression detection dataset to evaluate the effectiveness of our TensorFormer. For each timestamp, the datasets involved text-aligned multimodal features (textual, acoustic, and visual).

- *CMU-MOSI:* The CMU-MOSI [19] dataset is a common and popular dataset for multimodal sentiment analysis. The dataset is collected from YouTube where speakers express their opinions on certain topics. There are 1,283 utterances for training with 229 utterances for validation and 686 utterances for the test. Each utterance is annotated with a number in range $[-3, 3]$, indicating the strength of negative (< 0) or positive (> 0) sentiments.
- *CMU-MOSEI:* The CMU-MOSEI [20] dataset is an enlarged version of the CMU-MOSI [19] dataset. The videos are also collected from the Internet and annotated similarly to CMU-MOSI [19]. Specifically, there are 16,315 utterances for training with 1,817 utterances for validation and 4,654 utterances for the test.
- *AVEC2019:* The AVEC2019 DDS dataset [50] is collected from clinical interviews. Each sample is an audiovisual recording of a patient. The interviews are conducted by a virtual agent with human interferences. Different from CMU-MOSI and CMU-MOSEI, each modality of this dataset provides several different kinds of features. For example, acoustic modality comprises of Mel frequency cepstrum coefficient (MFCC), eGeMaps, and deep features extracted by VGG [54] and DenseNet [55]. The samples are annotated by PHQ-8 scores in the interval $[0, 24]$; a higher PHQ-8 score indicates a more serious tendency toward depression. There are 163 training samples, 56 validation samples, and 56 test samples in this benchmark dataset. In our experiments, we choose MFCC as acoustic features and facial action units as visual features.

4.2 Preprocessing and Metrics

To ensure fair competition with other baselines, we first process the data following prior works. Different methods are employed for different kinds of modalities.

- *Textual:* As the benchmark datasets provide raw sentences, we utilize BERT [47] pre-trained model to extract word embeddings. Each sentence is first concatenated with two special tokens: $[CLS]$ at the head and $[SEP]$ at the tail. The sentences are then tokenized and fed to the BERT pre-trained encoder, which extracts 768 dimensions of hidden states (d_t). On CMU-MOSEI and AVEC2019 datasets, we keep the first 100 words of each sentence and discard the rest to optimize performance.
- *Acoustic:* For acoustic modality of CMU-MOSI and CMU-MOSEI, COVAREP framework [56] is employed to generate the 74-dimensional features (d_a). The acoustic features include Mel-frequency cepstral coefficients, pitch, glottal source parameters, and other sentiment-related features.
- *Visual:* Facet toolkit is used to create visual features based on the Facial Action Coding System (FACS) [57] for CMU-MOSI and CMU-MOSEI datasets. The visual features involve facial action units and face poses. The dimension (d_v) of visual features is 47 in CMU-MOSI and 35 in CMU-MOSEI.

We provide several consensus metrics for comparison with different approaches following the previous works. For the MSA regression task, we use MAE and Pearson correlation coefficient (Corr) as measurements. The MAE can be calculated by

$$MAE = \frac{1}{D}\sum_{i=1}^{D} |y_i - p_i|, \tag{17}$$

where D is the number of test samples. Whereas the Corr can be calculated as

$$Corr = \frac{cov(Y, P)}{\sigma_Y \sigma_P} = \frac{E[(Y - \mu_Y)(P - \mu_P)]}{\sigma_Y \sigma_P}, \tag{18}$$

where Y and P are vectorized labels and predictions, σ is the standard deviation, μ is the expectation, and $cov(Y, P)$ is the covariance between Y and P. Furthermore, the binary metrics (Acc-2, F-Score) and seven-class metrics (Acc-7) are derived and calculated from the sentimental scores to coarsely estimate sentiment level. Specifically, binary metrics are calculated from positive/negative classes which are assigned for > 0 and < 0 sentiment intensity. The seven-class accuracy is computed from rounded sentiment tendency ranging in [-3, 3] (e.g., 2.8 is in class-3 and 0.1 is in class-1). For the depression detection task, we employ CCC

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
The Results on MSA Benchmark Dataset, CMU-MOSI

Models	TF- based	CMU-MOSI				
		MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
TFN [4]		0.970	0.633	73.9	73.4	32.1
MFN [24]		0.965	0.632	77.4	77.3	34.1
ICCN [49]		0.862	0.714	83.0	83.0	39.0
SWAFN [25]		0.880	0.697	80.2	80.1	40.1
MulT [42]	✓	0.871	0.698	83.0	82.8	40.0
LMF-MulT [43]	✓	0.957	0.681	78.5	78.5	34.0
MAT [7]	✓	-	-	-	80.0	-
MNT [7]	✓	-	-	-	80.0	-
MISA [38]	✓	0.817	0.748	82.1	82.0	41.4
BBFN [8]	✓	0.776	0.755	84.3	84.3	45.0
TensorFormer(Ours)	✓	**0.753**	**0.771**	**85.5**	**85.5**	**46.2**

The TF-based *denotes whether the methods are based on Transformer.*

and rooted mean squared error (RMSE) as comparison metrics following previous works. The CCC is represented as

$$CCC = \frac{2C_{py}}{C_p^2 + C_y^2 + (\bar{p} - \bar{y})^2}, \tag{19}$$

where the notations are the same as Equation (16), and the RMSE is represented as

$$RMSE = \sqrt{\frac{1}{D}\sum_{i=1}^{D}(y_i - p_i)^2}. \tag{20}$$

4.3 Experimental Setup

During our training process, we set $M = 3$ as we have three modalities. We find that TensorFormer is so effective that ($N = 2$) stacked layers can reach state-of-the-art performance, and we have discussed the choice of N in Section 5.2. We set the common projected dimension of involved modalities to 128 ($d_m = 128$) before interactions. In the weighted summary procedure, we set the scaled factor τ to $\sqrt{d_m}$ following origin Transformer structures. And we set d_{hid} to 128 in the feed-forward structures. Adam optimizer [58] is adopted in our framework. We set the learning rate to 0.004 and drop it by 0.1 every 50 training epochs. In the meantime, the learning rate of BERT was set to 0.01 times that of other parts. The models are implemented using the PyTorch framework and experiments are conducted on two Nvidia RTX 3090 GPUs.

4.4 Baselines

To prove the effectiveness of our TensorFormer, we compare our experimental results with several state-of-the-art methods. For the sentiment analysis, TFN [4] is also a tensor-based approach that employs the Cartesian product to obtain the overall description of involved modalities. MFN [24] uses LSTM-related structures to process the temporal information from three modalities at the same time. When fusing the modalities, ICCN [49] applies mathematical measurements to constrain the correlation between modalities. MulT [42] with its low-rank version LMF-MulT [43] both discard the traditional alignment for different modalities, but employs the stacked Transformer to expand the available temporal frames for alignment. MAT and MNT [7] employ the self-attention mechanism of Transformer to process cross-modality information with different normalization operations. SWAFN [25] calculates the cross-modal coattention of text-visual and text-acoustic pairs to interact between different modalities. MISA [38] learns modality-invariant and modality-specific representations by projecting them into multiple subspaces. BBFN [8] learns

TABLE 2
The Results on One MSA Benchmark, CMU-MOSEI

Models	TF- based	CMU-MOSEI				
		MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
TFN [4]		0.593	0.700	82.5	82.1	50.2
MFN [24]		-	-	76.0	76.0	-
ICCN [49]		0.565	0.713	84.2	84.2	51.6
MulT [42]	✓	0.580	0.703	82.5	82.3	51.8
LMF-MulT [43]	✓	0.620	0.668	80.8	81.3	49.3
MAT [7]	✓	-	-	82.0	82.0	-
MNT [7]	✓	-	-	80.5	80.5	-
MISA [38]	✓	0.557	0.748	84.9	84.8	51.7
BBFN [8]	✓	0.529	0.767	86.2	86.1	54.8
TensorFormer(Ours)	✓	**0.517**	**0.779**	**86.9**	**86.8**	**55.1**

The TF-based *means whether the methods are based on Transformer.*

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
The Results on Depression Detection Dataset, AVEC2019

Models	TF-based	AVEC2019	
		CCC(↑)	RMSE(↓)
Baseline [50]		0.111	6.37
EF [51]		0.344	-
Bert-CNN & Gated-CNN [28]		0.403	6.11
Multi-scale Temporal Dilated CNN [52]		0.430	4.39
Hierarchical BiLSTM [53]		0.442	5.50
Adaptive Fusion Transformer [30]	✓	0.443	5.61
TensorFormer(Ours)	✓	**0.493**	**4.31**

The TF-based *means whether the methods are based on Transformer.*

TABLE 4
The Studies of Paired T-Test Analysis With Our Approaches on the CMU-MOSEI Dataset

Other methods versus TensorFormer	p-value
TFN [4]	0.002
MFN [24]	0.016
MulT [42]	0.024
BBFN [8]	0.007

complementary information using two symmetric branches that are also based on Transformer structures. For depression detection, the baseline [50] method uses late fusion and directly averages the final predictions from involved modalities. To adaptively fuse the final predictions, Sun et al., [30] propose the adaptive fusion Transformer networks. EF [51] introduces simple linguistic and word-duration features to estimate the depression level. Bert-CNN & Gated-CNN [28] are designed with the gate mechanism to fuse the information attained from textual, acoustic, and visual signals. Multi-scale Temporal Dilated CNN [52] uses dilated CNN to extract multimodal features with a larger receptive field to process longer sequences. Hierarchical BiLSTM [53] applies the hierarchical biLSTM to capture the sequential information in a pyramid-like structure.

5 Results and Analysis

5.1 Results Summary

Our experimental results have been shown in Tables 1, 2, and 3. Some experimental settings for these results are covered in Section 5.2. According to the results, we can get the MAE of 0.753 and 0.517 on the CMU-MOSI and CMU-MOSEI datasets, and the CCC of 0.493 on the AVEC2019 dataset. On almost all metrics, our TensorFormer delivers state-of-the-art performance. Transformer-based methods perform comparatively better in comparison to non-Transformer approaches, showing the effectiveness of Transformer for multimodal processing. Our TensorFormer can reach better performance contrast to some other Transformer-based methods like MNT [7], MAT [7], and BBFN [7], which proves that our proposed global attention tensor surpasses traditional paired cross-attention QKV methods. Our approach significantly outperforms TFN [4] and MulT [42], the previous two tensor-based methods, on CMU-MOSI and CMU-MOSEI benchmarks. Due to the limitation of our hardware, we only keep the first 100 sequential frames for the CMU-MOSEI dataset, which is the possible reason for the relatively low performance improvement on this benchmark. On the depression detection benchmark, our TensorFormer achieves a significant performance compared with other approaches.

In addition, we use the IBM SPSS Statistics to conduct the t-test on CMU-MOSEI in Table 4. The statistical evaluation shows that our TensorFormer has a statistical difference (p-values < 0.05) with other experiments. The outcome also demonstrates the efficacy of our strategy.

TABLE 6
The Ablation Studies of N's Selection on the CMU-MOSEI Dataset

N	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
1	0.601	0.677	81.2	81.1	48.7
2	**0.517**	**0.779**	**86.9**	**86.8**	**55.1**
3	0.533	0.752	83.9	83.6	53.3
4	0.561	0.741	83.0	82.9	51.0

5.2 Ablation Study

We conduct an ablation study on the selection of N to determine the most effective structure. The experiments are conducted on the CMU-MOSEI dataset and corresponding results are shown in Table 6. At most, we can set $N = 4$ due to our GPU capability. We find that choosing $N = 2$ maximizes the performance. When N is set to more than 2, we observe that learning overfits during training. Thus, N is set to 2 in the following experiments.

TABLE 5
The Ablation Studies of TensorFormer Performed on CMU-MOSI Dataset

Model	Modality	AT	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
Model1	T		0.861	0.649	74.2	74.2	32.1
Model2	V		1.201	0.471	63.7	61.1	15.7
Model3	A		0.949	0.523	68.6	68.6	27.3
Model4	T+A	✓	0.800	0.729	78.5	77.7	39.0
Model5	T+V	✓	0.827	0.711	76.5	76.5	37.0
Model6	A+V	✓	0.869	0.653	74.5	74.8	32.1
Model7	T+A+V		0.842	0.701	78.5	77.9	34.1
Model8 (TensorFormer)	T+A+V	✓	**0.753**	**0.771**	**85.5**	**85.5**	**46.2**

AT *means the attention tensor module.*

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

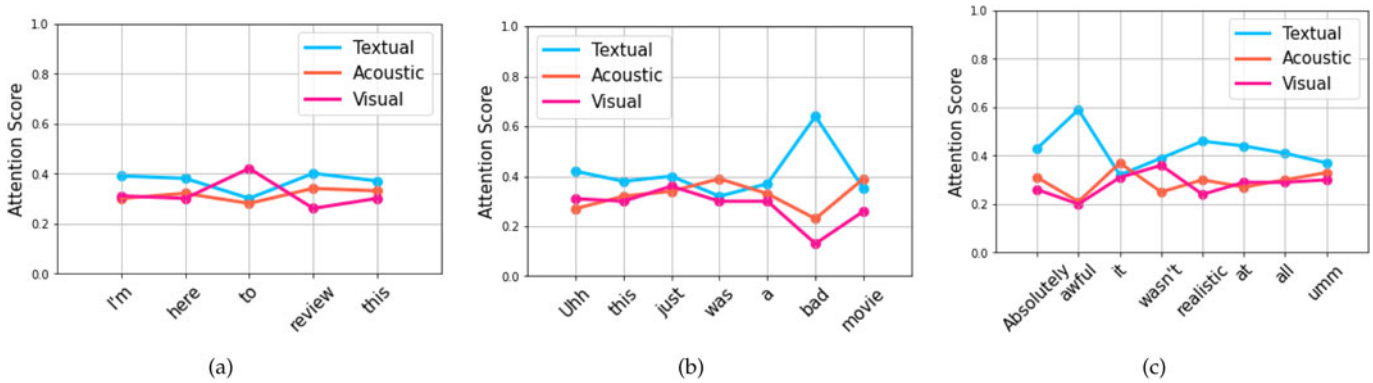


Fig. 5. The visualization of attention scores for three samples from the CMU-MOSEI dataset.All modalities are aligned by the spoken words. We visualize the corresponding attention weights of three modalities based on each word.

To further prove the effectiveness of our TensorFormer and its components, we perform some other ablation studies on CMU-MOSI dataset shown in Table 5. For models that have no attention tensor (e.g., Model1 and Model7), we leave the cross-attention module empty and use the input as the output directly. We begin with an ablation study on the affected modalities. We observe from Model1, 2, and 3 that textual modality has richer information, whereas the impact of the visual modality on sentiment analysis is not very strong. The effective pre-trained Bert model [47] may be the cause of the significant influence of textual modality. We can conclude from the comparison of results of Model1, 4, 5, 6, and 8, that multimodal learning is essential to sentiment analysis. The performance increases as the participating modalities interact more efficiently. To confirm the efficacy of our proposed attention tensor, we perform comparison studies between Model7 and Model8. The results show that our attention tensor is significantly helpful in improving performance.

5.3 Attention Visualization

To intuitively examine our global cross-modality attention approach, we visualize attention scores across different modalities. Specifically, we save the interacted features S_t, S_a, and S_v in several examples and normalize them at each timestamp. The visualized results are shown in Fig. 5. The visualization samples are included in the CMU-MOSEI dataset. As seen in the figure, the textual modality has higher attention weights than the other two modalities, which implies that the strong Bert extractor may be the reason. Additionally, the textual modality has substantially larger weights for affective words (such as *bad* and *awful*) than other neutral terms (e.g., *just*, *a*, and *this*), as shown in Figs. 5b and 5c. In certain conditions, the acoustic modality gets higher weights (e.g., *movie* and *it*), we believe the cause is that speakers are speaking in emotional tones. The visualization of S_m also illustrates the complementarity of involved modalities and evaluates the effectiveness of our TensorFormer from another viewpoint.

6 Conclusion

In this paper, we have proposed a Multimodal Transformer (TensorFormer) to process the multimodal data for sentiment analysis. The key idea of TensorFormer is a novel attention mechanism: global attention tensor. Using TensorFormer, we can perform comprehensive interaction of information from all relevant modalities, ensuring their complementarity and consistency. Our experimental outcomes and ablation experiments have proven TensorFormer's efficacy for sentiment analysis. TensorFormer can handle three modalities and be applied to additional modalities because it is designed with flexibility and scalability. Theoretically, TensorFormer is a generalized multimodal data processing framework that can be applied to any downstream tasks. We intend to apply TensorFormer in other multimodal research areas in the future. Moreover, we plan to conduct some case studies in real scenes, so as to make it suitable for different scenarios. TensorFormer is efficient in terms of time but not in terms of space usage. The space complexity of TensorFormer is $O(n^M)$ where n is the sequential length and M is the number of modalities.

Acknowledgments

Authors would like to thank Dr. Rahul Kumar Jain of Ritsumeikan University, Japan for his kind English proof and correction.

References

[1] E. Cambria, S. Poria, A. Hussain, and B. Liu, "Computational intelligence for affective computing and sentiment analysis [guest editorial]," *IEEE Comput. Intell. Mag.*, vol. 14, no. 2, pp. 16–17, May 2019.

[2] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," in *A Practical Guide to Sentiment Analysis*, Berlin, Germany: Springer, 2017, pp. 1–10.

[3] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[4] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[5] F. Huang, S. Zhang, J. Zhang, and G. Yu, "Multimodal learning for topic sentiment analysis in microblogging," *Neurocomputing*, vol. 253, pp. 144–153, 2017.

[6] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell. 30th Innov. Appl. Artif. Intell. Conf. 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 5642–5649.

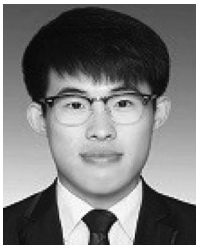
[7] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Proc. 2nd Grand-Challenge Workshop Multimodal Lang.*, 2020, pp. 1–7.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

[8] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interaction*, 2021, pp. 6–15.
[9] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–22.
[10] S. A. Qureshi, G. Dias, M. Hasanuzzaman, and S. Saha, "Improving depression level estimation by concurrently learning emotion intensity," *IEEE Comput. Intell. Mag.*, vol. 15, no. 3, pp. 47–59, Aug. 2020.
[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.
[12] A. Hussain, E. Cambria, S. Poria, A. Hawalah, and F. Herrera, "Information fusion for affective computing and sentiment analysis," *Inf. Fusion*, vol. 71, pp. 97–98, 2021. [Online]. Available: http://researchrepository.napier.ac.uk/Output/2763759
[13] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2022.3204972.
[14] K. He, R. Mao, T. Gong, C. Li, and E. Cambria, "Meta-based self-training and re-weighting for aspect-based sentiment analysis," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2022.3202831.
[15] Y. Miura, S. Sakaki, K. Hattori, and T. Ohkuma, "TeamX: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data," in *Proc. 8th Int. Workshop Semantic Eval.*, 2014, pp. 628–632.
[16] M. Hagen, M. Potthast, M. Büchner, and B. Stein, "Webis: An ensemble for twitter sentiment detection," in *Proc. 9th Int. Workshop Semantic Eval.*, 2015, pp. 582–589.
[17] E. Cambria, A. Hussain, and A. Vinciarelli, "Affective reasoning for big social data analysis," *IEEE Trans. Affective Comput.*, vol. 8, no. 4, pp. 426–427, Fourth Quarter 2017.
[18] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.
[19] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.
[20] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[21] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.
[22] L. Stappen et al., "The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," in *Proc. 2nd Multimodal Sentiment Anal. Challenge*, 2021, pp. 5–14.
[23] L. Stappen et al., "MuSe-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox," in *Proc. 2nd Multimodal Sentiment Anal. Challenge*, 2021, pp. 75–82.
[24] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. 32nd AAAI Conf. Artif. Intell. 30th Innov. Appl. Artif. Intell. Conf. 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 5634–5641.
[25] M. Chen and X. Li, "SWAFN: Sentimental words aware fusion network for multimodal sentiment analysis," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1067–1077.
[26] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, 2018.
[27] J. Joshi et al., "Multimodal assistive technologies for depression diagnosis and monitoring," *J. Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.
[28] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of BERT-CNN and gated CNN representations for depression detection," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, 2019, pp. 55–63.
[29] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, 2019, pp. 81–88.
[30] H. Sun et al., "Multi-modal adaptive fusion transformer network for the estimation of depression level," *Sensors*, vol. 21, no. 14, 2021, Art. no. 4764.
[31] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics," in *Proc. IEEE Symp. Comput. Intell. Hum.-Like Intell.*, 2013, pp. 108–117.
[32] H. Deng, P. Kang, Z. Yang, T. Hao, Q. Li, and W. Liu, "Dense fusion network with multimodal residual for sentiment classification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.
[33] S. Han, R. Mao, and E. Cambria, "Hierarchical attention network for explainable depression detection on twitter aided by metaphor concept mappings," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 94–104.
[34] W. Peng, X. Hong, and G. Zhao, "Adaptive modality distillation for separable multimodal sentiment analysis," *IEEE Intell. Syst.*, vol. 36, no. 3, pp. 82–89, May/Jun. 2021.
[35] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interaction*, 2017, pp. 163–171.
[36] H. Peng, Y. Ma, S. Poria, Y. Li, and E. Cambria, "Phonetic-enriched text representation for chinese sentiment analysis with reinforcement learning," *Inf. Fusion*, vol. 70, pp. 88–99, 2021.
[37] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1034–1047, Mar. 2022.
[38] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[39] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. Assoc. Comput. Linguistics*, 2018, pp. 2122–2132.
[40] G. Tu, J. Wen, C. Liu, D. Jiang, and E. Cambria, "Context- and sentiment-aware networks for emotion recognition in conversation," *IEEE Trans. Artif. Intell.*, vol. 3, no. 5, pp. 699–708, Oct. 2022.
[41] Q. Chen, I. Chaturvedi, S. Ji, and E. Cambria, "Sequential fusion of facial appearance and dynamics for depression recognition," *Pattern Recognit. Lett.*, vol. 150, pp. 115–121, 2021.
[42] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.
[43] S. Sahay, E. Okur, S. H. Kumar, and L. Nachman, "Low rank fusion based transformers for multimodal sequences," in *Proc. 2nd Grand-Challenge Workshop Multimodal Lang.*, 2020, pp. 29–34.
[44] Z. Wang, Z. Wan, and X. Wan, "TransModality: An End2End fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, 2020, pp. 2514–2520.
[45] C. Bao, Z. Fountas, T. Olugbade, and N. Bianchi-Berthouze, "Multimodal data fusion based on the global workspace theory," in *Proc. Int. Conf. Multimodal Interaction*, 2020, pp. 414–422.
[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
[48] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, pp. 255–268, 1989.
[49] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.
[50] F. Ringeval et al., "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, 2019, pp. 3–12.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.

[51] H. Kaya et al., "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, 2019, pp. 27–35.
[52] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, "Multi-modality depression detection via multi-scale temporal dilated CNNs," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, 2019, pp. 73–80.
[53] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, 2019, pp. 65–71.
[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[55] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[56] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP–A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.
[57] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford Univ. Press, 1997.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–15.



Hao Sun received the BEng degree from the Harbin Institute of Technology, in 2016. Now he is working toward the PhD degree with the College of Computer Science and Technology, Zhejiang University, China. His research interests include computer vision, video processing, multimodal learning and sentiment analysis.

Yen-Wei Chen (Member, IEEE) received the BE degree from Kobe University, Kobe, Japan, in 1985, the ME and DE degrees both from Osaka University, Osaka, Japan, in 1987 and 1990, respectively. From 1991 to 1994, he was a research fellow with the Institute for Laser Technology, Osaka. From October 1994 to March 2004, he was an associate professor and a professor with the Department of Electrical and Electronic Engineering, University of the Ryukyus, Okinawa, Japan. He is currently a professor with the College of Information Science and Engineering, Ritsumeikan University, Kyoto, Japan. He is also a chair professor with the College of Computer Science and Technology, Zhejiang University, China. He is an associate editor of *International Journal of Image and Graphics (IJIG)* and an editorial board member of the *International Journal of Knowledge based and Intelligent Engineering Systems*. His research interests include pattern recognition, image processing and machine learning. He has published more than 200 research.

Lanfen Lin (Member, IEEE) received the BS and PhD degrees in aircraft manufacture engineering from Northwestern Polytechnical University, in 1990, and 1995 respectively. She held a postdoctoral position with the College of Computer Science and Technology, Zhejiang University, China, from January 1996 to December 1997. Now she is a full professor and the vice director of the Artificial Intelligence Institute in Zhejiang University. Her research interests include medical image processing, Big Data analysis, data mining, and so on.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on August 17,2024 at 14:22:50 UTC from IEEE Xplore. Restrictions apply.