BAFN: Bi-Direction Attention Based Fusion Network for Multimodal Sentiment Analysis

Jiajia Tang, Dongjun Liu, Xuanyu Jin, Yong Peng, *Member, IEEE*, Qibin Zhao, *Senior Member, IEEE*, Yu Ding, and Wanzeng Kong, *Member, IEEE*

***Abstract*— Attention-based networks currently identify their effectiveness in multimodal sentiment analysis. However, existing methods ignore the redundancy of auxiliary modalities. More importantly, existing methods only attend to top-down attention (static process) or down-top attention (implicit process), leading to the coarse-grained multimodal sentiment context. In this paper, during the preprocessing period, we first propose the multimodal dynamic enhanced block to capture the intra-modality sentiment context. This can effectively decrease the intra-modality redundancy of auxiliary modalities. Furthermore, the bi-direction attention block is proposed to capture fine-grained multimodal sentiment context via the novel bi-direction multimodal dynamic routing mechanism. Specifically, the bi-direction attention block first highlights the explicit and low-level multimodal sentiment context. Then, the low-level multimodal context is transmitted to a carefully designed bi-direction multimodal dynamic routing procedure. This allows us to dynamically update and investigate high-level and much more fine-grained multimodal sentiment contexts. The experiments demonstrate that our fusion network can achieve state-of-the-art performance. Notably, our model outperforms the best baseline on the metric 'Acc-7' with an improvement of 6.9%.**

***Index Terms*— Multimodal fusion network, multimodal sentiment analysis, attention mechanism.**

I. Introduction

MULTIMODAL sentiment analysis has raised increasing interest in artificial intelligence systems, which focuses on reaching the much more correct sentiment message via the integration of multiple sentimental modalities. Among these, the text [1], audio [2], and video [3] modalities are popularly utilized to analyze the related multimodal research [4], [5], [6], [7], [8], [9]. For instance, the multimodal sentiment analysis technique has already been applied to the interaction between the humanoid robot Pepper and patients, which allows for significant improvement of life quality of patients. Due to the consistency and complementarity among multiple sentiment modalities, capturing the joint representations indeed boosts the sentiment analysis performance. Consequently, the primary concern of the multimodal sentiment analysis task is to learn the much more sophisticated multimodal sentiment context among multiple modalities [10], [11], [12], [13], [14].

Recently, attention-based networks [15] have gained widespread attention for their significant performance in capturing the task-related context among various modalities in computer vision and NLP [16], [17], [18]. Existing attention-based networks consist of down-top attention and top-down attention based modules. EF-Net [19] employed the down-top attention based architecture to deal with image presentation, contributing to spatial context among distinct receptive areas of the image. In addition, McIntosh [20] proposed the down-top attention based fusion network to highlight cross-modality context between video and text via the implicit process. Compared to the above down-top attention based networks, MulT [21] and MMLGAN [22] presented the top-down attention based multimodal fusion framework. This can explore sentiment context among multiple modalities via the static process. Nevertheless, the aforementioned techniques totally ignore the redundancy of auxiliary modalities (audio and video). Note that, compared to text, the auxiliary modalities consist of many more redundancy messages [23], [24]. Intuitively, directly integrating text and the original auxiliary modalities into the joint representation may increase the difficulty of effectively reasoning about multimodal messages. More importantly, the top-down attention based models leverage the static method to simply investigate low-level explicit interactions at once. Additionally, the down-top attention based models focus on the implicit multimodal interactions, which fails to exploit explicit interactions. Note that the above explicit interactions are captured via the immediate calculation among the original input modalities. In contrast, the implicit interactions are captured by calculating the correlations between the original input modalities and the output information. Accordingly, existing methods can only capture

Manuscript received 27 April 2022; revised 31 July 2022 and 14 September 2022; accepted 22 October 2022. Date of publication 28 October 2022; date of current version 5 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U20B2074 and Grant U1909202, in part by the National Key Research and Development Program of China for Intergovernmental International Science and Technology Innovation Cooperation Project under Grant 2017YFE0116800, in part by the Key Research and Development Project of Zhejiang Province under Grant 2021C03001 and Grant 2021C03003, in part by JSPS KAKENHI under Grant 17K00326, and in part by the Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province under Grant 2020E10010. This article was recommended by Associate Editor R. Du. *(Corresponding author: Wanzeng Kong.)*

Jiajia Tang, Dongjun Liu, Xuanyu Jin, Yong Peng, and Wanzeng Kong are with the College of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: hdutangjiajia@163.com; liudongjun@hdu.edu.cn; jxuanyu599@163.com; yongpeng@hdu.edu.cn; kongwanzeng@hdu.edu.cn.)

Qibin Zhao is with the Center for Advanced Intelligence Project, RIKEN, Saitama 351-0198, Japan (e-mail: qibin.zhao@riken.jp).

Yu Ding is with Netease Fuxi AI Lab, NetEase, Hangzhou 310052, China (e-mail: dingyu01@corp.netease.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3218018.

Digital Object Identifier 10.1109/TCSVT.2022.3218018

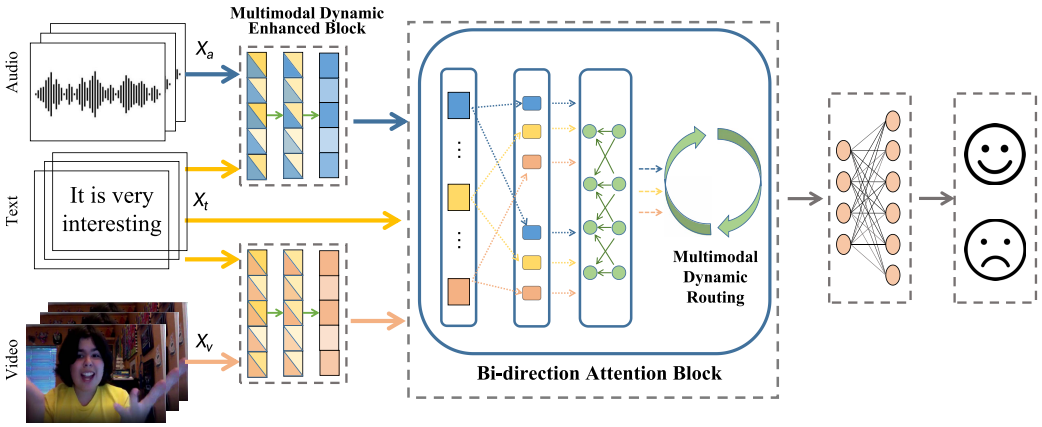


Fig. 1. BAFN: Initially, during the preprocess period, the multimodal dynamic enhanced block is utilized to dynamically decrease the intra-modality redundancy of auxiliary modalities(audio X_a, video X_v). Then, the bi-direction attention block is further proposed to exploit the much more fine-grained multimodal sentiment context.

the relatively coarse-grained multimodal sentiment context, leading to the great deterioration of task performance.

In this paper, during the preprocessing period, the multimodal dynamic enhanced block is first proposed to dynamically capture the intra-modality sentiment context. Due to the incorporation of guidance from the more discriminative text modality, the enhanced block indeed has the potential to effectively decrease the intra-modality redundancy of auxiliary modalities. Furthermore, the bi-direction attention block is proposed to obtain the fine-grained multimodality sentiment context via the novel bi-direction multimodal dynamic routing mechanism. Specifically, the bi-direction attention block first captures the explicit and low-level multimodality sentiment context via the static multimodal process. Then, the above low-level multimodal context is transmitted to the carefully designed multimodal dynamic routing procedure associated with multiple iterations. This naturally gives the multimodal fusion network the strong ability to dynamically update and investigate the much more fine-grained multimodal sentiment context. To the best of our knowledge, our model is the first dynamic multimodal sentiment fusion network that simultaneously focuses on the analysis of redundancy of auxiliary modalities, as well as the investigation of the much more fine-grained multimodal sentiment context. In addition, our proposed bi-direction attention based fusion network (BAFN) has demonstrated the superiority on two multimodal sentiment analysis benchmarks.

II. Related Work

The existing multimodal sentiment learning model consists of the following two leading lines:

A. Non-Attention Based Multimodal Learning

Recently, LSTM- and RNN-based techniques have drawn a surge of interest in multimodal sentiment analysis for their excellence in exploiting the temporal correlation from the sequence [25], [26], [27], [28], [29], [30], [31]. For instance, *BC-LSTM* [32] proposed bi-directional LSTM to highlight the sentiment properties among modality utterances. *RMFN* [33] utilized RNN to decompose the complex multimodal fusion process into several fusion substages, leading to the much more sophisticated multimodal intercorrelations. *MV-LSTM* [34] presented the multiview LSTM block to explicitly model the consistency and complimentary message among multiple modalities. Additionally, *Self-MM* [35] introduced the subtask analysis method to exploit the multimodal sentiment intercorrelations. *MFM* [36] factorized the multimodal joint distribution into the inter-modality and intra-modality sentiment pertinence. *ICCN* [37] applied the deep canonical correlation analysis (CCA) mechanism to retrieve the nonlinear and complex intercorrelations among various modalities. *ERLDK* [38] proposed a reinforcement learning module to perform the multimodal emotion recognition task. *The hybrid deep model* [39] utilized the deep DBN model to fuse audio and visual representations. In addition, tensor-based models have raised increasing interest due to the high-dimension properties. *TFN* [40] employed the tensor network to explicitly account for the unimodal, bimodal, and trimodal sentiment interactions. Based on *TFN*, *LMF* [41] further proposed modality-specific low-rank factors to deal with multiple modalities, which can significantly decrease the computational complexity of the multimodal learning model. However, the above networks fail to effectively explore the multimodal sentiment context from the long sequence, which may limit the expressive power of the learning model.

B. Attention Based Multimodal Learning

Compared to the aforementioned models, attention-based frameworks have demonstrated superiority in the analysis of long sequence representation [42], [43], [44], [45], [46], [47], [48]. *MMLGAN* [22] proposed a multimodal local-global attention network to integrate representations of different modalities, leading to a discriminative affective representation. *MARN* [49] leveraged the multi-attention block to simultaneously investigate multiple cross-modality sentiment contexts in each time step and then store the sentiment context in the hybrid memory block. *RAVEN* [50] applied the attention gating mechanism to learn a nonlinear combination between

the visual and acoustic modality, which brings forth the nonverbal shift vector. Similarly, *MAG* [5] introduced an attention gated memory to integrate the text and nonverbal cues into another vector, which is subsequently added to the text modality. Additionally, *MFN* [51] utilized three LSTMs to attend to each modality separately and employed a special attention mechanism called the delta-memory attention network to identify the cross-modality sentiment interactions. *MISA* [10] employed the distribution similarity block to calculate similar portions across all modalities and leveraged the self-attention mechanism to account for the multimodal sentiment interactions. *MulT* [21] proposed the directional pairwise cross-modality attention mechanism to capture the interactions between multimodal sequences across distinct time steps, and latently adapt the sequence from one modality to another. *MCTN* [52] assigned the cyclic consistency loss to the standard Transformer, which can ensure that the multimodal sentiment contexts retain maximal information from all modalities.

However, existing multimodal sentiment analysis networks neglect the redundancy of auxiliary modalities (audio and video), which increases the difficulty of effectively reasoning about multimodal sentiment context. More importantly, the aforementioned attention-based network only attempt to capture the low-level explicit or implicit multimodal interactions via the uni-directional attention mechanism (top-down or down-top attention). This fails to capture the much more fine-grained multimodal sentiment context among multiple modalities.

In recent years, several papers research on multimodal learning have been published [9], [22], [38], [39]. The paper [22] is the most closely relevant to our paper, which used the attention-based method to deal with the multimodal sentiment analysis task. *MMLGAN* [22] proposed the top-down attention based multimodal fusion framework to capture multimodal sentiment context. *MHMAN* [9] attends to the hierarchical multimodal fusion network for video question answering. *ERLDK* [38] and *The hybrid deep model* [39] focused on the non-attention based multimodal fusion network for multimodal sentiment analysis. Compared to previously published papers, our proposed network can simultaneously analyze the redundancy of auxiliary modalities, and investigate the much more fine-grained multimodal sentiment context.

III. Methodology

As shown in Figure 1, the proposed BAFN consists of two essential components: 1) during the preprocessing period, the multimodal dynamic enhanced module is leveraged to decrease the intra-modality redundancy of auxiliary modalities, and 2) the bi-direction attention block is further proposed to capture the much more fine-grained multimodal sentiment context.

A. Preliminaries

The two public multimodal sentiment analysis benchmarks consist of audio, video, and text. The utterance-level representation of the above modalities are represented as $\boldsymbol{X_a} \in \mathbb{R}^{T_a \times d_a}$, $\boldsymbol{X_v} \in \mathbb{R}^{T_v \times d_v}$, and $\boldsymbol{X_t} \in \mathbb{R}^{T_t \times d_t}$. $T_i (i \in \{a, v, t\})$ refers to the number of utterances, and the feature dimension is denoted as $d_i (i \in \{a, v, t\})$. Note that, the original benchmarks utilized the interval duration of each word utterance as a time step. Then, the benchmarks aligned audio and video by calculating the average value over the utterance interval of each word, i.e., $T_a = T_v = T_t$. Due to the properties of the dot product, we adopt the linear function to analyze $\{X_a, X_v, X_t\}$ for retrieving the same feature dimension d_i, i.e., $d_a = d_v = d_t$.

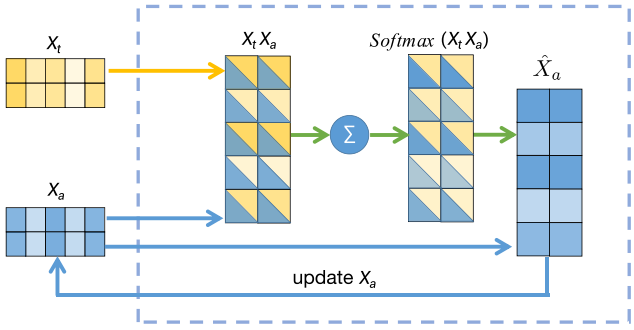


Fig. 2. Multimodal dynamic enhanced block. Initially, X_a and X_t are mapped into the cross-modality fusion space $X_a \cdot X_t$. Subsequently, the softmax function is utilized to exploit the cross-modality sentiment context coefficients. Then, the coefficients are applied to deal with the original X_a, leading to the much more discriminative modality representation $\hat{X}_a$.

B. Multimodal Dynamic Enhanced Block

During the preprocessing period, the multimodal dynamic enhanced block (Figure 2) is proposed to dynamically capture the intra-modality sentiment context of $\boldsymbol{X_a} \in \mathbb{R}^{T_a \times d_a}$ and $\boldsymbol{X_v} \in \mathbb{R}^{T_v \times d_v}$ with the help of the text modality $\boldsymbol{X_t} \in \mathbb{R}^{T_t \times d_t}$. This allows us to effectively decrease the redundancy in the auxiliary modalities (audio and video modality). Specifically, the enhanced block consists of M process heads, where each head includes N adaptive iterations. That is, different processing heads consist of a distinct number of iterations. Intuitively, the multihead mechanism allows for extracting the intra-modality sentiment context with the multispect view, yielding the comprehensive sentiment context. Additionally, the multiple iterations has the potential to dynamically update the intra-modality sentiment context, leading to the much more discriminative audio (video) modality.

For the single-head case, the intra-modality context ${\boldsymbol{X_a}_m}^{[N_m]}$ of the m-th head associated with N_m iterations is formulated as follows:

$$
\begin{aligned}
{\boldsymbol{X_a}_m}^{[N_m]} &= \text{Softmax}\left(\boldsymbol{X_a} \cdot \boldsymbol{X_t}\right)\boldsymbol{X_a}, N_m = 1 \\
{\boldsymbol{X_a}_m}^{[N_m]} &= \text{Softmax}\left(\sum_{i=1}^{N_m - 1} {\boldsymbol{X_a}_m}^{[i]} \cdot \boldsymbol{X_t}\right){\boldsymbol{X_a}_m}^{[N_m - 1]}, \quad N_m \geq 2.
\end{aligned}
\tag{1}
$$

During the first iteration, audio modality $\boldsymbol{X_a}$ and text modality $\boldsymbol{X_t}$ are explicitly mapped into the cross-modality fusion space $\boldsymbol{X_a} \cdot \boldsymbol{X_t}$. Subsequently, the softmax function is introduced to analyze $\boldsymbol{X_a} \cdot \boldsymbol{X_t}$, leading to the cross-modality sentiment context coefficients. Note that the softmax function attempts to compute an attention score matrix, where the (i,j)-th element of the matrix refers to the similarity between

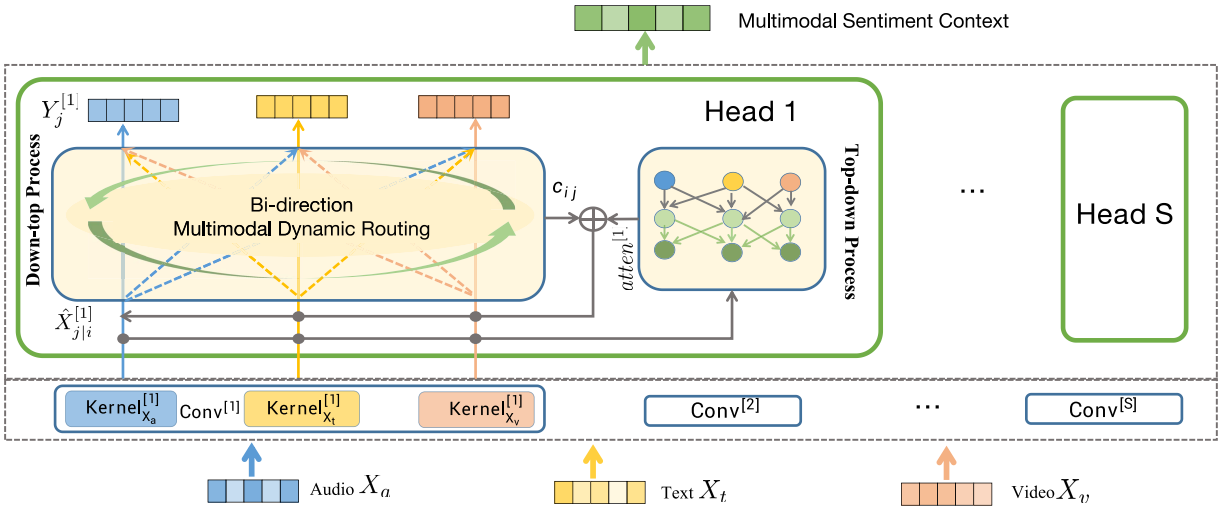


Fig. 3. Bi-direction attention block: The X_i and Y_j refer to the modality representation and multimodal sentiment context, respectively. The multimodal sentiment context $Y_j^{[s]}$ could be represented as the weighted sum of $\hat{X}_{j|i}^{[s]}$, with the help of the dynamic down-top coefficients $c_{ij}^{[s]}$ and the top-down multimodal sentiment context $atten^{[s]}$. This indeed gives the multimodal fusion model the strong ability to investigate the much more fine-grained multimodality sentiment context.

the i-th element of audio modality $\boldsymbol{X_a}$ and the j-th element of text modality $\boldsymbol{X_t}$. Therefore, the i-th element of ${\boldsymbol{X_a}_m}^{[N_m]}$ is a weighted summary of $\boldsymbol{X_a}$, with the weight value determined by the i-th row in the above attention score matrix. Similarly, many attention-based methods, such as *MISA* [10] also used the softmax function to analyze the cross-modality interactions. Compared to the other weighted averaging methods, the softmax function can effectively deal with the gradient exploding issue [53]. Hence, we use the softmax function to analyze the multimodal fusion message in our work. Indeed, the above sentiment context coefficients have the strong ability to determine how the utterances of the audio $\boldsymbol{X_a}$ are influenced by the utterances in the text $\boldsymbol{X_t}$. Then, the cross-modality sentiment context coefficients are applied to deal with the original audio modality $\boldsymbol{X_a}$, contributing to the much more discriminative sentiment properties of audio. In other words, the large coefficients provide us with the benefit of highlighting the text-related sentiment properties of audio $\boldsymbol{X_a}$, and the small coefficients attempt to peel off the impact of redundancy of audio $\boldsymbol{X_a}$. Due to the incorporation of guidance from the more discriminative modality (text), the above process indeed provides us with the strong ability to effectively decrease the redundancy of auxiliary modalities (audio $\boldsymbol{X_a}$ and video $\boldsymbol{X_v}$) [54], [55].

Then, the next iteration attempts to dynamically update the cross-modality fusion space $\boldsymbol{X_a} \cdot \boldsymbol{X_t}$ based on the output of the previous iteration ${\boldsymbol{X_a}_m}^{[N_m-1]}$. That is, the output of the previous iteration is leveraged to construct a new cross-modality fusion space ${\boldsymbol{X_a}_m}^{[N_m-1]} \cdot \boldsymbol{X_t}$ for the next iteration, leading to a much more compact cross-modality fusion space. Note that the enhanced process of $\boldsymbol{X_v}$ is similar to that of $\boldsymbol{X_a}$.

Taking the single-head enhanced block as a basis, the multihead enhanced network is further established to collect the multiway intra-modality sentiment context. Additionally, the convolution operation is introduced to analyze the multiway intra-modality sentiment context. This can further extract the essential interactions among distinct ${\boldsymbol{X_a}_m}^{[N_m]}$, leading to the much more discriminative modality $\hat{\boldsymbol{X_a}}$:

$$\hat{\boldsymbol{X_a}} = \text{Conv}\left(\text{concat}\left({\boldsymbol{X_a}_1}^{[N_1]}, \cdots, {\boldsymbol{X_a}_M}^{[N_M]}\right)\right) \tag{2}$$

where 'Conv' refers to the convolution operation, 'M' indicates the total number of process heads, and 'concat' denotes the concatenation operation.

C. Bi-Direction Attention Block

When the above multimodal dynamic enhanced process is finished, the bi-direction attention block is further proposed to capture the much more fine-grained multimodal sentiment context. With the help of the enhanced block, the much more discriminative auxiliary modalities can be transmitted to the bi-direction attention block. This significantly boosts the learning efficiency to effectively extract the multimodal sentiment context among multiple relatively discriminative modalities.

As shown in Figure 3, the bi-direction attention block comprises modality representations $\{X_i\}_{i=1}^{N_x}$ and multimodal sentiment context $\{Y_j\}_{j=1}^{N_y}$. For simplicity, we utilize the index 'i' to represent $\{a, v, t\}$. That is, $X_1 = X_a$, $X_2 = X_t$, and $X_3 = X_v$. X_i and Y_j indicate the input message and output message of the above attention block, respectively. N_x and N_y refer to the number of modality representations and multimodal sentiment context, respectively. Initially, we utilize the convolution operation to transform the original modality representation X_i into $\hat{X}_{j|i}$ with respect to Y_j. Then, the modality transformation $\hat{X}_{j|i}$ can be leveraged to generate the multimodal sentiment context Y_j. The above process is

formulated as follows:

$$\begin{aligned}\hat{X}_{j|i} &= \text{Conv}\left(X_i, kernel_i\right) \\ &= \text{sigmoid}\left(\sum X_i * kernel_i + bias_i\right).\end{aligned} \tag{3}$$

In addition, we extend the above single-head convolution transformation design to the multi-head case associated with varying convolution kernels. The multi-head mechanism indeed allows for multiway and comprehensive information flow between the modality transformation $\hat{X}_{j|i}$ and the multimodal sentiment context Y_j, where 's' refers to the specific head:

$$\begin{aligned}\hat{X}^{[s]}_{j|i} &= \text{Conv}^{[s]}\left(X_i, kernel^{[s]}_i\right) \\ &= \text{sigmoid}\left(\sum X_i * kernel^{[s]}_i + bias^{[s]}_i\right).\end{aligned} \tag{4}$$

Then, the modality transformation $\hat{X}^{[s]}_{j|i}$ and the multimodal sentiment context $Y^{[s]}_j$ are utilized to account for the implicit multimodal interaction space $b^{[s]}_{ij}$. The values of $b^{[s]}_{ij}$ are initialized as 0, and $Y^{[s]}_j$ is initialized with random value. The following formulation is leveraged to dynamically update $b^{[s]}_{ij}$. The above $b^{[s]}_{ij}$ and $\hat{X}^{[s]}_{j|i}$ are latently used to dynamically update the multimodal sentiment context $Y^{[s]}_j$:

$$b^{[s]}_{ij} = b^{[s]}_{ij} + \left(\hat{X}^{[s]}_{j|i} Y^{[s]}_j\right). \tag{5}$$

Subsequently, the dynamic multimodal routing procedure with N_y iterations is conducted to dynamically extract the multimodal sentiment context among multiple modalities. At each iteration, we leverage the dynamic down-top coefficients $c^{[s]}_{ij}$ to account for the down-top information flow between $\{X_i\}^{N_x}_{i=1}$ and $\{Y_j\}^{N_y}_{j=1}$, which is calculated based on the interaction space $b^{[s]}_{ij}$ initialized as 0. That is, $c^{[s]}_{ij}$ can be utilized to identify how each element of $Y^{[s]}_j$ is influenced by modality representations $\hat{X}^{[s]}_{j|i}$. Note that, the above procedure focuses on analyzing the interaction between $\hat{X}^{[s]}_{j|i}$ and $Y^{[s]}_j$, leading to implicit multimodal interactions among $\hat{X}^{[s]}_{j|i}$. The detailed procedure is formulated as follows:

$$\begin{aligned}\{c^{[s]}_{ij}\}^{N_y}_{j=1} &= \text{Softmax}\left(\left\{b^{[s]}_{ij}\right\}^{N_y}_{j=1}\right) \\ &= \frac{\exp\left(b^{[s]}_{ij}\right)}{\sum^{N_y}_{j=1}\exp\left(b^{[s]}_{ij}\right)}.\end{aligned} \tag{6}$$

Compared to the above dynamic down-top process, the top-down procedure tends to leverage the static method to simply investigate the explicit multimodal interactions among original input modalities at once. That is, the above explicit multimodal interactions are captured via the immediate calculation among modality transformations $\hat{X}^{[s]}_{j|i}$. Then, we attempt to measure the top-down multimodal sentiment context $atten^{[s]}$ of the s-th head as follows:

$$\begin{aligned}\hat{X}^{[s]} &= \text{concat}\left(\hat{X}^{[s]}_{j|i_1}, \ldots, \hat{X}^{[s]}_{j|i_{N_x}}\right) \\ atten^{[s]} &= \text{Softmax}\left(W_q \hat{X}^{[s]} W^T_k \left(\hat{X}^{[s]}\right)^T\right) W_v \hat{X}^{[s]}\end{aligned} \tag{7}$$

where 'concat' refers to the concatenation operation. W_q, W_k, and W_v are the transformation matrices.

It is important to note that, the top-down multimodal analysis process simply investigates the explicit multimodal interactions at once via the static method. That is, the above method can only capture the relatively simple interactions, and completely ignores the implicit multimodal interactions, leading to the relatively coarse-grained multimodal sentiment context. Additionally, the down-top multimodal analysis focuses on the implicit interactions, which completely ignores the explicit multimodal interactions. Intuitively, we directly leverage the static top-down method or the dynamic down-top method to analyze multiple modalities, which indeed fails to effectively capture the much more expressive multimodal sentiment context. Therefore, we integrate the top-down multimodal sentiment context $atten^{[s]}$ and the down-top coefficients $c^{[s]}_{ij}$ into the novel bi-direction dynamic coefficients $\left(atten^{[s]} + c^{[s]}_{ij}\right)$. This naturally allows for bi-direction dynamic multimodal fusion among $\hat{X}^{[s]}_{j|i}$. That is, our bi-direction dynamic coefficients allows for simultaneously analyzing the high-level explicit and implicit multimodal interactions, leading to the much more fine-grained multimodal sentiment context.

Specifically, the top-down explicit multimodal sentiment context $atten^{[s]}$ is further transmitted to our carefully designed dynamic multimodal routing procedure associated with multiple dynamic iterations. That is, the multimodal sentiment context $Y^{[s]}_j$ can be represented as the weighted sum of $\hat{X}^{[s]}_{j|i}$, with the help of the dynamic down-top coefficients $c^{[s]}_{ij}$ and the top-down explicit multimodal sentiment context $atten^{[s]}$. Due to the joint guidance from bidirectional multimodal fusion processes, the novel bi-direction dynamic coefficient $\left(c^{[s]}_{ij} + atten^{[s]}\right)$ naturally allows us to simultaneously capture implicit and explicit multimodal interactions at each iteration. Then, the multimodal sentiment context can be dynamically updated via multiple dynamic iterations, which allows for transmitting the coarse-grained multimodal sentiment context to the much more fine-grained one. This indeed gives the fusion model the strong ability to effectively extract the much more high-level and fine-grained multimodal sentiment context. The above procedure is formulated as follows:

$$Y^{[s]}_j = \sum_i \left(c^{[s]}_{ij} + atten^{[s]}\right) \hat{X}^{[s]}_{j|i}. \tag{8}$$

When the head 's' is set to 2, each modality can compute two corresponding multimodal sentiment contexts $Y^{[1]}_j$ and $Y^{[2]}_j$. Then, the above sentiment context can be further integrated into the modality-aware multimodal sentiment context 'con_a', 'con_v', and 'con_t' via a convolution operation. For instance, $con_a = conv\left(concat\left(Y^{[1]}_{j_a}, Y^{[2]}_{j_a}\right), kernel_a\right)$. Subsequently, all the modality-aware multimodal sentiment contexts are further merged into the output 'multimodality sentiment $context'$ via a convolution operation: 'multimodality sentiment $context' = conv\left(concat\left(con_a, con_v, con_t\right), kernel\right)$.

As mentioned previously, the convolution transformation is leveraged to analyze the X_i, which allows for the convolutional nonlinear representation. Accordingly, we use the

HingeLoss proposed by C. Baile [56] utilized to analyze nonlinear information for reducing the discrepancy among modality-aware multimodal sentiment contexts. Moreover, the standard cross-entropy loss [57] is also used to measure the task performance of the proposed model. The total loss is formulated as follows:

$$\begin{aligned} totalloss &= \sum HingeLoss\left(con_i, con_j\right) \\ &\quad + TaskLoss\left(rL, pL\right) \\ &= \sum max\left(0, 1 - \|D\left(con_i\right) - D\left(con_j\right)\|_2\right) \\ &\quad + rL \cdot log\left(pL\right) \end{aligned} \tag{9}$$

where 'D' refers to the variance function, 'rL' and 'pL' indicate the original label and predicted label respectively, $i, j \in \{a, v, t\}$, and $i \neq j$.

IV. Experiments Setups

A. Datasets

The CMU-MOSI (Multimodal Opinion Sentiment Intensity) dataset [59] comprises 2,199 video segments collected from 93 opinion videos of YouTube movie reviews. Each video comprises multiple opinion segments. Each segment of the video is manually annotated with the continuous sentimental label in the range of [−3, 3]. The value −3 indicates the strong negative sentiment, and the value 3 denotes the strong positive sentiment. The above dataset consists of 1,284 training, 229 validation, and 686 testing samples. The CMU-MOSEI dataset [58] is an extension of CMU-MOSI associated with many more utterance segments. This version is composed of 22,856 annotated video utterances (segments) from 5,000 opinion videos of YouTube movie reviews. Similarly, each segment is assigned with a specified sentiment in the range [−3, 3]. There are 16,326 segments in the training set, 1,871 segments in the validation set, and 4,659 segments in the testing set. Note that, the same speaker of the training set is not allowed to appear in the testing set, which benefits the model in exploiting the speaker-independent multimodal sentiment interactions.

B. Features and Alignment

For CMU-MOSI and CMU-MOSEI, we adopt the same method of MAG and MISA to extract the features of the specific modality. Specifically, the pretrained BERT [60] and XLNet [61] are utilized to exploit the corresponding textual representations. Note that the original benchmarks first leveraged the P2FA forced alignment model to align the text and audio at the phoneme level [62]. The benchmarks used the interval duration of each word utterance as a time step. Subsequently, the visual and audio are aligned by calculating the average value over the utterance interval of each word of the text modality [51].

C. Evaluation Metrics

In this paper, the following evaluation metrics are introduced to analyze task performance: mean absolute error (MAE), Pearson correlation (Corr), binary accuracy (Acc-2), F-Score (F1), and multiclass accuracy (Acc-7) ranging from -3 to 3. For all metrics except for MAE, a relatively higher value represents better task performance. Essentially, two distinct methods are proposed to measure Acc-2 and F1. 1) In the work of [49], the negative class is annotated with the label in the range of [−3, 0), while the range of the non-negative class is [0, 3]. 2) In contrast, in the work of [21], the ranges of the negative and positive classes are [−3, 0) and (0, 3], respectively. The marker -/- is employed to distinguish the distinct strategies, where the left-side value refers to 1) and the right-side value stands for 2).

D. Training and Implementation Details

For all baselines and our proposed BAFN model, the grid search is performed over the hyperparameters to select the model with the best validation classification or regression loss. The ranges of the essential hyperparameters are summarized as follows: head [1, 6], iteration [1, 7], and convolution kernel {3, 5, 7}. The training duration of the learning models is governed by an early-stopping strategy with patience of 20 to 30 epochs, and the Adam optimizer is introduced to the task. Additionally, our entire network is trained in an end to end way rather than in two stages, where the network learns all the steps between the initial input and the final output. Specifically, during the training period, the output message of the multimodal dynamic enhance block is directly transmitted to the bi-direction attention block. Note that all baselines are verified on the same benchmarks (CMU-MOSI and CMU-MOSEI). Thus, the splits of the training, validation, and testing trials are exactly the same for all baselines. In our work, we leveraged several CPUs and GPUs as the computing resources to conduct corresponding experiments. In detail, the CPUs are Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz, and the GPUs are GeForce RTX 3080. Additionally, PtTorch and Python are utilized to implement all our experiments.

E. Comparisons

We introduced the non-attention based multimodal learning and attention-based multimodal learning models as baselines. Non-attention based: Bi-directional LSTM (BC-LSTM) [32], RNN-based multistage fusion network (RMFN) [33], Multi-view LSTM (MV-LSTM) [34], Self-Supervised Multi-task Multimodal model (Self-MM) [35], Multimodal Factorization Model (MFM) [36], Interaction Canonical Correlation Network (ICCN) [37], Tensor Fusion Network (TFN) [40], Low-rank Multimodal Fusion (LMF) [41]. Attention-based: Multi-attention Recurrent Network (MARN) [49], Recurrent Attended Variation Embedding Network (RAVEN) [50], Multimodal Adaptation Gate (MAG) [5], Memory Fusion Network (MFN) [51], Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis(MISA) [10], Multimodal Transformer for Unaligned Multimodal Language Sequences (MulT) [21], Multimodal Cyclic Translation Network (MCTN) [52]. We also introduced the down-top attention based Capsule Network [16] as the baseline model.

TABLE I

PERFORMANCE OF BAFN (BERT) ON CMU-MOSI. NOTE THAT (BERT) MEANS THE TEXTUAL REPRESENTATION IS EXPLORED VIA BERT; ⊗ FROM [21]; △ FROM [37]

Models	CMU-MOSI				
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	Acc-7(↑)
BC-LSTM	1.079	0.581	73.9/-	73.9/-	28.7
MV-LSTM	1.019	0.601	73.9/-	74.0/-	33.2
$RMFN^{\otimes}$	0.922	0.681	78.4/-	78.0/-	38.3
$RAVEN^{\otimes}$	0.915	0.691	78.0/-	76.6/-	33.2
MFN	0.965	0.632	77.4/-	77.3/-	34.1
MARN	0.968	0.625	77.1/-	77.0/-	34.7
TFN	0.970	0.633	73.9/-	73.4/-	32.1
LMF	0.912	0.668	76.4/-	75.7/-	32.8
MulT	0.871	0.698	-/83.0	-/82.8	40.0
$MCTN^{\otimes}$	0.909	0.676	79.3/-	79.1/-	35.6
$MFM^{\otimes}$	0.951	0.662	78.1/-	78.1/-	36.2
Capsule Network (Bert)	0.762	0.778	83/86	83.4/86.1	39.5
$TFN(Bert)^{\triangle}$	0.901	0.698	-/80.8	-/80.7	34.9
$LMF(Bert)^{\triangle}$	0.917	0.695	-/82.5	-/82.4	33.2
ICCN (Bert)	0.860	0.710	-/83.0	-/83.0	39.0
MISA (Bert)	0.783	0.761	81.8/83.4	81.7/83.6	42.3
MAG (Bert)	0.712	0.796	84.2/86.1	84.1/86.0	-
Self-MM (Bert)	0.713	0.798	84.0/85.98	84.42/85.95	-
BAFN (Non-Enhanced) (Bert)	0.684	0.824	86.0/88.4	85.9/88.4	47.8
BAFN (Bert)	**0.669**	**0.833**	**86.5/89.1**	**86.5/89.1**	**49.2**

TABLE II

PERFORMANCE OF BAFN (XLNET) ON CMU-MOSI. NOTE THAT (X) MEANS THE TEXTUAL REPRESENTATION IS EXPLORED VIA XLNET; ◇ FROM [5]

Models	CMU-MOSI			
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)
TFN	0.970	0.633	73.9/-	73.4/-
MARN	0.968	0.625	77.1/-	77.0/-
MFN	0.965	0.632	77.4/-	77.3/-
RMFN	0.922	0.681	78.4/-	78.0/-
MulT	0.871	0.698	-/83.0	-/82.8
Capsule Network (X)	0.75	0.799	83.7/85.9	83.8/85.9
$TFN(X)^{\diamond}$	0.914	0.713	78.2/80.1	78.2/78.8
$MARN(X)^{\diamond}$	0.921	0.707	78.3/79.5	78.8/79.6
$MFN(X)^{\diamond}$	0.898	0.713	78.3/79.9	78.4/79.1
$RMFN(X)^{\diamond}$	0.901	0.703	79.1/81.0	78.6/80.0
$MulT(X)^{\diamond}$	0.849	0.738	87.9/84.4	80.4/83.1
MAG (X)	0.675	0.821	85.7/87.9	85.6/87.9
BAFN (Non-Enhanced) (X)	0.672	0.827	85.2/87.4	85.1/87.4
BAFN (X)	**0.661**	**0.836**	**86.6/88.8**	**86.5/88.8**

V. EXPERIMENTS RESULTS AND ANALYSIS

A. Performance Comparison With State-of-the-art Models

The performance of state-of-the-art baselines, our proposed BAFN and the ablation case BAFN (Non-Enhanced) are illustrated in the following tables. Note that BAFN (Non-Enhanced) refers to the case in which BAFN performs the multimodal learning task on the original modality data rather than the outputs of the multimodal dynamic enhanced block. The bottom rows in Table I, Table II, and Table III demonstrate the superiority and effectiveness of BAFN. Particularly, on the CMU-MOSEI benchmark, BAFN exceeds the previous best Self-MM (Bert) on the metric 'Corr' by a margin of 5.0%. Additionally, on the CMU-MOSI dataset, BAFN outperforms MISA (Bert) on the metric 'Acc-7' with an improvement of 6.9%. This implies that our proposed multimodal dynamic enhanced module and bi-direction attention block indeed provide us with the benefit of effectively exploiting the much more fine-grained and discriminative multimodal sentiment context among multiple modalities. Essentially, we can observe that BAFN obtains better results than the ablation case BAFN (Non-Enhanced). The observations signify the necessity of decreasing the intra-modality redundancy of auxiliary modalities (audio and video) before the multimodal fusion process.

B. Effect of the Bi-Direction Attention Block of BAFN

In this work, the bi-direction attention block is proposed to explore the much more fine-grained multimodal sentiment context among multiple modalities via the presented bi-direction multimodal dynamic routing mechanism. Therefore, we attempt to investigate how bi-direction attention affects the multimodal sentiment analysis performance. Specifically, the t-SNE method is utilized to provide the corresponding visualization for the multimodal fusion representations learned by the BAFN. The visualization of the binary classification task and multiclassification task are illustrated in Figure 4. For the binary classification task, the red points refer to the positive sentiment class, and the green points indicate the negative sentiment class. For the multiclassification task, the color of the points depends on the corresponding annotated sentimental labels. Additionally, the performance comparison is demonstrated in Table IV. It is interesting to find that, compared to the BAFN (without a bi-direction attention block), the multimodal fusion message becomes increasingly separable when the BAFN is associated with the bi-direction attention block. This implies that the bi-direction attention block is able to exploit much more fine-grained multimodal sentiment context, leading to the significant improvement

TABLE III
PERFORMANCE OF BAFN (BERT) ON CMU-MOSEI. NOTE THAT (BERT) MEANS THE TEXTUAL REPRESENTATION IS EXPLORED VIA BERT; ⊗ FROM [58]; △ FROM [37]

Models	CMU-MOSEI				
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	Acc-7(↑)
$MFN^{\otimes}$	-	-	76.0/-	76.0/-	-
$MV-LSTM^{\otimes}$	-	-	76.4/-	76.4/-	-
RAVEN	0.614	0.662	79.1/-	79.5/-	50.0
MCTN	0.609	0.670	79.8/-	80.6/-	49.6
MulT	0.580	0.703	-/82.5	-/82.3	51.8
Capsule Network (Bert)	0.581	0.80	83.8/86.4	84/86.3	48.6
$TFN(Bert)^{\triangle}$	0.593	0.700	-/82.5	-/82.1	50.2
$LMF(Bert)^{\triangle}$	0.623	0.677	-/82.0	-/82.1	48.0
$MFM(Bert)^{\triangle}$	0.568	0.717	-/84.4	-/84.3	51.3
ICCN (Bert)	0.565	0.713	-/84.2	-/84.2	51.6
MISA (Bert)	0.555	0.756	83.6/85.5	83.8/85.3	**52.2**
Self-MM (Bert)	0.530	0.765	83.79/85.23	83.74/85.3	-
BAFN (Non-Enhanced) (Bert)	0.563	0.806	85.3/86.9	85.2/86.8	49.9
BAFN (Bert)	**0.551**	**0.815**	**86.3/87.1**	**86.1/87.1**	51.3

TABLE IV
PERFORMANCE OF BAFN (WITHOUT ATTENTION BLOCK) AND BAFN (WITH ATTENTION BLOCK)

Models	CMU-MOSI				
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	Acc-7(↑)
BAFN (without attention block) (Bert)	0.762	0.778	83/86	83.4/86.1	39.5
BAFN (with attention block) (Bert)	0.684	0.824	86.0/88.4	85.9/88.4	47.8

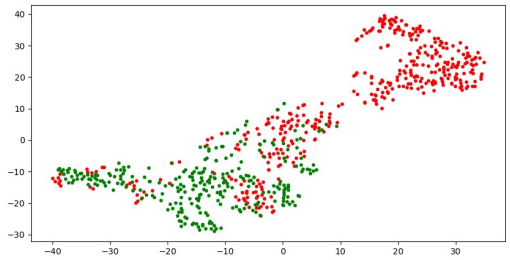
of classification efficiency and expressive capability in the multimodal sentiment analysis task. More importantly, we find that all the points are on a curve that is similar to the manifold structure. The above structure may demonstrate that the presented novel bi-direction attention mechanism can exploit the low-dimensional inherent data structure from the multimodal sentiment representative space. That is, our proposed bi-direction attention block has the potential to effectively discover the intrinsic sentiment portions among multiple modalities by simultaneously considering the explicit and implicit essential multimodal interactions.

C. Effect of Head and Convolution Kernel of The Bi-Direction Attention Block

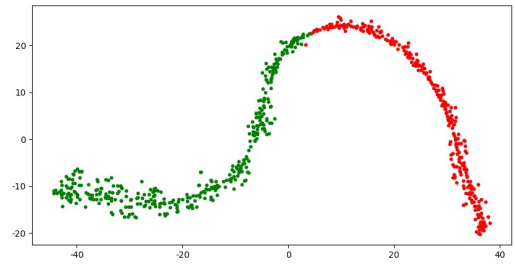
During the bi-direction multimodal dynamic routing process, the multihead mechanism and convolution operation are introduced to deal with the multiple modalities. Therefore, we are interested in measuring how varying heads and convolution kernel sizes affect the task performance of multimodal sentiment analysis. The head s varies from 2 to 6, and each head is associated with a corresponding convolution kernel is of the same size (3×3, 5×5, or 7×7). In Figure 5, BAFN can achieve good results with respect to the tested head and kernel. Notably, kernel_3×3-based settings reach the peak value at head 3, and kernel_5×5-based settings maximize prediction performance at head 4. This implies that the multihead strategy can give each head the strong ability to effectively exploit the multiway and comprehensive information flow between the modality transformation and the multimodal sentiment context. It is interesting to find that, compared to the kernel_3×3- and kernel_5×5-based settings, kernel_7×7-based settings receive the best performance at head 5. Actually, kernel_7×7 attempts to perform the multimodal sentiment analysis task within the relatively larger receptive field, which may lead to the lack of the much more fine-grained multimodal intercorrelations among multiple modalities to some extent. Therefore, kernel_7×7-based settings require more heads to highlight the much more comprehensive multimodal sentiment context. Indeed, the setting associated with too many heads may capture a similar multimodal interaction pattern within the same feature map, leading to information redundancy and poor task performance. In contrast, the setting that comprises too few heads may fail to effectively explore the sophisticated multimodal sentiment context. In conclusion, the above observations signify the necessity and effectiveness of the multihead mechanism and convolution operation in the multimodal sentiment analysis task.

D. Effect of The Multimodal Dynamic Enhanced Block of BAFN

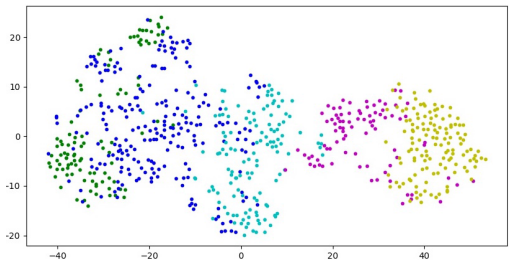
During the preprocess period, the multimodal dynamic enhanced module is utilized to decrease the intra-modality redundancy of auxiliary modalities (audio and video modality). Consequently, we attempt to examine how the multimodal sentiment analysis model behaves by considering the multimodal dynamic enhanced module. Moreover, we also investigate how the multimodal dynamic enhanced module affects the sentiment analysis performance of the MAG and MISA models. As shown in Figure 6, compared to the BAFN (without an enhanced block), the BAFN (with an enhanced block) achieves the relatively higher sentiment analysis performance. Similarly, compared to the MAG (without an enhanced block) and MISA (without an enhanced block), the MAG (with an enhanced block) and MISA (with an enhanced block) achieve better sentiment analysis performance. Intuitively, with the help of the multimodal dynamic enhanced, the much more



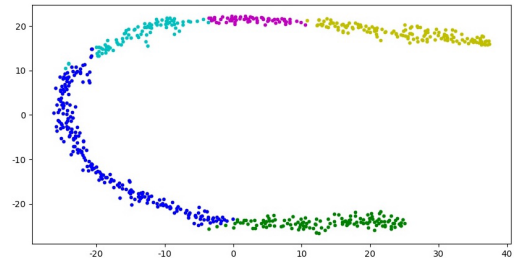
(a) BAFN (without bi-direction attention block) in binary classification task



(b) BAFN (with bi-direction attention block) in binary classification task



(c) BAFN (without bi-direction attention block) in multi-classification task



(d) BAFN (with bi-direction attention block) in multi-classification task

Fig. 4. t-SNE visualization of the multimodal fusion message learned by BAFN on CMU-MOSI.

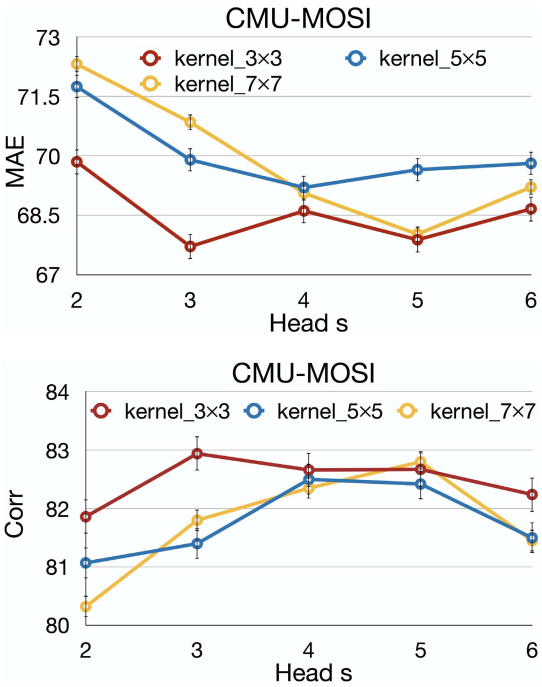


Fig. 5. Results of effect of head and convolution kernel of bi-direction attention block on CMU-MOSI.

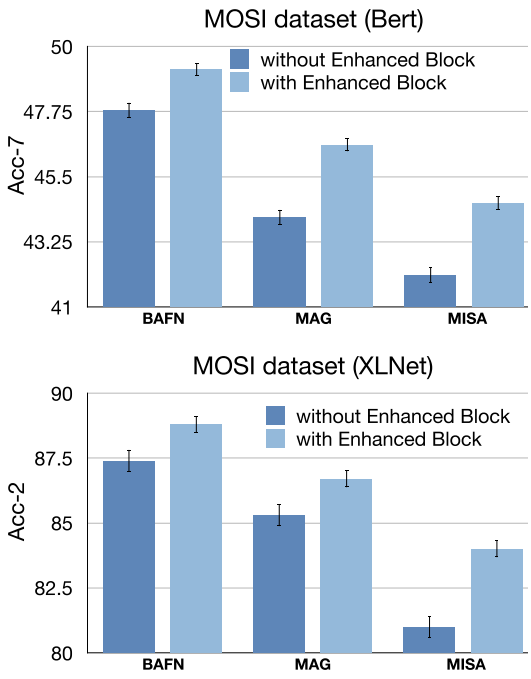


Fig. 6. Effect of the multimodal dynamic enhanced block of BAFN on CMU-MOSI.

discriminative auxiliary modalities can be transmitted to the following multimodal fusion block. This significantly boosts the learning efficiency to capture the multimodal sentiment context among multiple relatively discriminative modalities. In summary, the above observations demonstrate the necessity of leveraging the multimodal dynamic enhanced module to deal with the multiple modalities before the multimodal fusion procedure.

E. Effect of the Head of the Multimodal Dynamic Enhanced Block

In this work, the multimodal dynamic enhanced block is proposed to explicitly facilitate the intra-modality sentiment context from the auxiliary modalities. Specifically, the proposed multimodal dynamic enhanced block comprises M process heads. Therefore, we are interested in investigating how distinct process heads affect sentiment analysis performance. The process head varies from 1 to 6. The number of iterations of each head is set to 3. As shown in Figure 7, our proposed model can obtain fairly good performance with

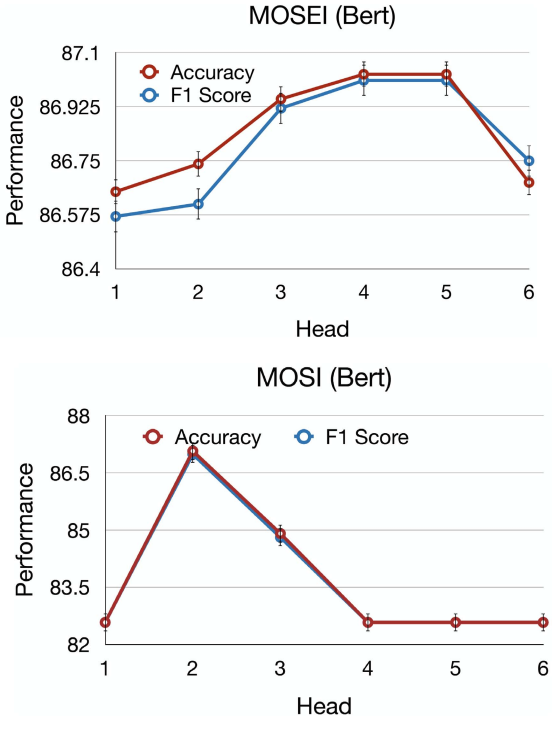


Fig. 7. Effect of the head of multimodal dynamic enhanced block on CMU-MOSI and MOSEI.

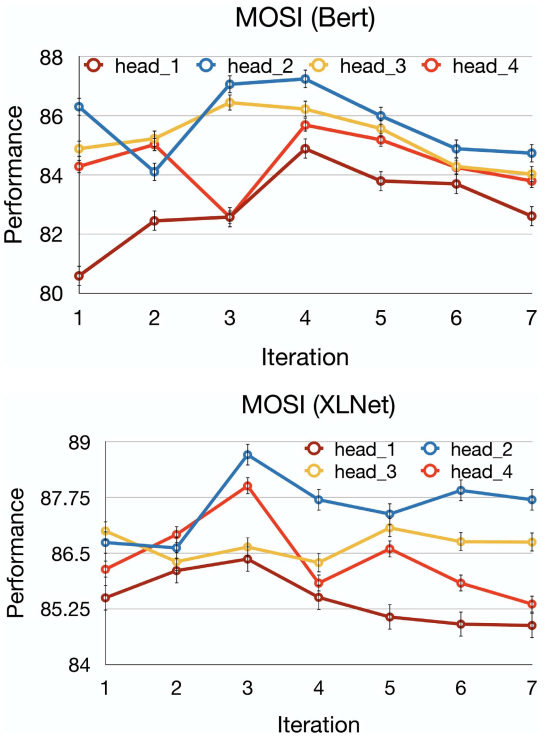


Fig. 8. Effect of the dynamic iteration of multimodal dynamic enhanced block on CMU-MOSI.

respect to the process heads. It is important to observe that, our model reaches the peak value at head 2 for the case of CMU-MOSI (Bert). Regarding the CMU-MOSEI (Bert), we can observe that the relatively higher performance is achieved at head 4. Compared to the CMU-MOSI dataset, the CMU-MOSEI dataset includes many more utterance segments. Therefore, the CMU-MOSEI dataset requires more process heads to highlight the multimodal sentiment context. Indeed, the multihead mechanism allows for exploiting the intra-modality sentiment context with the multispect view, yielding the comprehensive sentiment context. Accordingly, the proposed multihead enhanced strategy further boosts the expressive efficiency of the learning model. Additionally, the too-simple enhanced bock, which is comprised of too few heads (e.g., 1 head), may fail to effectively discover the comprehensive intra-modality sentiment context. In addition, the excessively complex enhanced block that consists of too many heads may investigate similar multimodal interaction messages, leading to information redundancy.

F. Effect of The Dynamic Iteration of the Multimodal Dynamic Enhanced Block

The proposed multimodal dynamic enhanced block comprises M process heads, and each head consists of N adaptive iterations. In this part, we attempt to analyze how various adaptive iterations affect the task performance. The number of adaptive iterations ranges from 1 to 7, and the number of heads varies from 1 to 4. In Figure 8, the head_i is used to denote the head number of the testing case. For instance, the head_2 refers to the testing case associated with 2 heads. The corresponding comparison of runtime for models with various heads

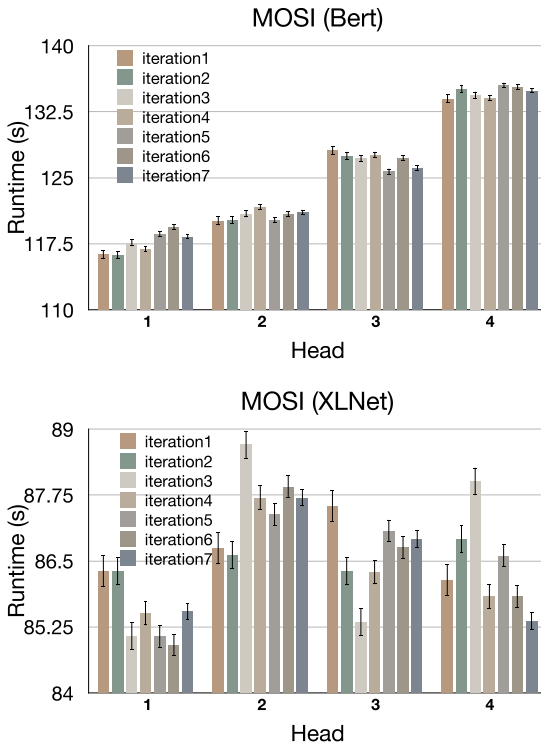


Fig. 9. The average runtime of the testing settings associated with multiple heads and multiple iterations on CMU-MOSI.

and differing numbers of iterations in each head is illustrated in Figure 9. Note that, each runtime is obtained by calculating the average value over 5 repeat experiments. As shown in Figure 8,

our proposed model can obtain fairly good performance with respect to the adaptive iterations. It is interesting to find that our model maximizes the task performance at adaptive iteration 4 of head_2 for the case of CMU-MOSI (Bert). For CMU-MOSI (XLNet), we can observe that the relatively better performance is received at adaptive iteration 3 of head_2. Intuitively, each adaptive iteration attempts to exploit the intra-modality context from the audio (video) modality via the more discriminative modality (text). Then, the multiple stacked iterations focus on dynamically updating or modifying the intra-modality sentiment context, leading to the much more discriminative audio (video) modality. Subsequently, the much more discriminative auxiliary modalities are transmitted to the bi-direction attention block, which significantly boosts the learning efficiency of the model. Note that, compared to the one-head case associated with multiple dynamic iterations, the multiple-head case can achieve better task performance.

VI. Conclusion

We propose a multimodal dynamic enhanced block to decrease the intra-modality redundancy of auxiliary modalities (audio and video modality) via the more discriminative modality (text) during the preprocessing period. This provides us with the benefit of effectively obtaining the more discriminative audio and video. Furthermore, the bi-direction attention block is proposed to capture the much more fine-grained multimodal sentiment context using the novel bi-direction multimodal dynamic routing mechanism. Note that our model exceeds the previous best baseline on the metric 'Acc-7' by a large margin of 6.9%. To the best of our knowledge, our model is the first dynamic multimodal sentiment fusion network that simultaneously considers the redundancy of auxiliary modalities and the investigation of fine-grained multimodal sentiment context.

References

[1] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Comput. Surveys*, vol. 50, no. 2, pp. 1–33, Mar. 2018.

[2] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Sentiment extraction from natural audio streams," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 8485–8489.

[3] A. A. L. Cunha, M. C. Costa, and M. A. C. Pacheco, "Sentiment analysis of YouTube video comments using deep neural networks," in *Proc. Int. Conf. Artif. Intell. Soft Comput. (ICAISC)*, May 2019, pp. 561–570.

[4] Q. T. Ain et al., "Sentiment analysis using deep learning techniques: A review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 424–433, Jan. 2017.

[5] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Jul. 2020, pp. 2359–2369.

[6] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces (ICMI)*, 2011, pp. 169–176.

[7] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 350–358.

[8] A. Kumar and G. Garg, "Sentiment analysis of multimodal Twitter data," *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 24103–24119, Sep. 2019.

[9] T. Yu, J. Yu, Z. Yu, Q. Huang, and Q. Tian, "Long-term video question answering via multimodal hierarchical memory attentive networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 931–944, Mar. 2021.

[10] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131.

[11] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification," *Int. J. Multimedia Inf. Retr.*, vol. 9, no. 2, pp. 103–112, Jun. 2020.

[12] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jul. 2019, pp. 371–378.

[13] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016.

[14] C. Yan et al., "STAT: Spatial-temporal attention mechanism for video captioning," *IEEE Trans. Multimedia.*, vol. 22, no. 1, pp. 229–241, Jun. 2019.

[15] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2018.

[16] H. Lin et al., "Dynamic context-guided capsule network for multimodal machine translation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1320–1329.

[17] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE Int. Conf. Comput. Vision. (ICCV)*, Oct. 2021, pp. 22–31.

[18] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Comput. Sci.*, vol. 10, no. 4, pp. 429–439, 2015.

[19] D. Gu et al., "Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021.

[20] B. McIntosh, K. Duarte, Y. S. Rawat, and M. Shah, "Visual-textual capsule routing for text-based video segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9942–9951.

[21] 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. 57th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2019, pp. 6558–6569.

[22] Y. Ou, Z. Chen, and F. Wu, "Multimodal local-global attention network for affective video content analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1901–1914, May 2021.

[23] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic- and visual-LSTM with channel-aware temporal convolution network," *IEEE/ACM Trans. Audio, Speech, Langauge Process.*, vol. 29, pp. 1424–1437, 2021.

[24] S. Seo, S. Na, and J. Kim, "HMTL: Heterogeneous modality transfer learning for audio-visual sentiment analysis," *IEEE Access*, vol. 8, pp. 140426–140437, 2020.

[25] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical multimodal LSTM for dense visual-semantic embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1881–1889.

[26] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 176–183.

[27] Y. Chen, J. Yuan, Q. You, and J. Luo, "Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 117–125.

[28] N. Xu and W. Mao, "MultiSentiNet: A deep semantic network for multimodal sentiment analysis," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 2399–2402.

[29] A. Agarwal, A. Yadav, and D. K. Vishwakarma, "Multimodal sentiment analysis via RNN variants," in *Proc. IEEE Int. Conf. Big Data, Cloud Comput., Data Sci. Eng. (BCD)*, May 2019, pp. 19–23.

[30] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Gener. Comput. Syst.*, vol. 115, pp. 279–294, Feb. 2021.

[31] Y. Zhang et al., "Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis," *Neural Netw.*, vol. 133, pp. 40–56, Jan. 2021.

[32] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Jan. 2017, pp. 873–883.

[33] P. P. Liang, Z. Liu, A. B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Conf. Empirical Methods Natural Langauge Process.*, Nov. 2018, pp. 150–161.

[34] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2016, pp. 338–353.
[35] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, May 2021, pp. 10790–10797.
[36] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," 2018, *arXiv:1806.06176*.
[37] 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. (AAAI)*, Apr.2020, pp. 8992–8999.
[38] 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.
[39] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2017.
[40] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. EMNLP*, Sep. 2017, pp. 1103–1114.
[41] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.
[42] Q.-T. Truong and H. W. Lauw, "VistaNet: Visual aspect attention network for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jul. 2019, pp. 305–312.
[43] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Trans. Audio, Speech, Langauge Process.*, vol. 28, pp. 429–439, 2020.
[44] G. Xiao et al., "Multimodality sentiment analysis in social Internet of Things based on hierarchical attentions and CSAT-TCN with MBM network," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12748–12757, Aug. 2021.
[45] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1008–1017.
[46] F. Huang, K. Wei, J. Weng, and Z. Li, "Attention-based modality-gated networks for image-text sentiment analysis," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 3, pp. 1–19, Aug. 2020.
[47] L.-F. Xie and X.-Y. Zhang, "Gate-fusion transformer for multimodal sentiment analysis," in *Proc. Int. Conf. Pattern Recognit. Artif. Intell. (ICPRAI)*, Oct. 2020, pp. 28–40.
[48] K. Yang, H. Xu, and K. Gao, "CM-BERT: Cross-modal BERT for text-audio sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 521–528.
[49] 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. AAAI Conf. Artif. Intell. (AAAI)*, Apr. 2018, pp. 5642–5649.
[50] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jul. 2019, pp. 7216–7223.
[51] 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. AAAI*, Feb. 2018, pp. 5634–5641.
[52] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jul. 2019, pp. 6892–6899.
[53] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural. Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.
[54] Z. Wang, Z. Wan, and X. Wan, "TransModality: An End2End fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, Apr. 2020, pp. 2514–2520.
[55] Y. Liu, X. Yu, Z. Chen, and B. Liu, "Sentiment analysis of sentences with modalities," in *Proc. Int. Conf. Inf. Knowl. Manage*, Oct. 2013, pp. 39–44.
[56] C. Bailer, K. Varanasi, and D. Stricker, "CNN-based patch matching for optical flow with thresholded Hinge embedding loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3250–3259.
[57] E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham, "Uses and abuses of the cross-entropy loss: Case studies in modern deep learning," 2020, *arXiv:2011.05231*.
[58] A. Bagher 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.
[59] 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. 2016.
[60] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[61] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural. Inf. Process. Syst.*, Jun. 2019, pp. 5754–5764.
[62] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, pp. 5687–5690, May 2008.

Jiajia Tang received the Ph.D. degree from Hangzhou Dianzi University, China, in 2022. She is currently a Lecturer with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. Her current research interests include tensor neural networks, deep learning, multimodal fusion, and affective computing.

Dongjun Liu is currently pursuing the master's degree with the School of Computer Science at Hangzhou Dianzi University, China. His research interests include brain–machine collaborative intelligence, multimodal learning, and transfer learning.

Xuanyu Jin received the B.S. degree from Zhejiang International Studies University, Hangzhou, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou. Her current research interests include tensor decomposition, tensor neural networks, and deep learning.

Yong Peng (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University in 2015. He is currently a Full Professor with the School of Computer Science and Technology, Hangzhou Dianzi University. He has published more than 30 SCI-indexed journal articles such as in IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING (TNSRE), IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS (TCDS), IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT (TIM), *Information Sciences*, *Neural Networks*, and *Knowledge-Based Systems*. His main research interests include machine learning, pattern recognition, and EEG-based brain–computer interfaces. He was awarded the President Prize of Chinese Academy Sciences in 2009 and the Third Prize of the Chinese Institute of Electronics in 2018.

Qibin Zhao (Senior Member, IEEE) received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2009. From 2009 to 2017, he was a Research Scientist with the RIKEN Brain Science Institute, Wako, Japan. He is currently a Unit Leader of tensor learning unit with RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan. He is also a Visiting Professor with the Saitama Institute of Technology, Fukaya, Japan. He is also a Visiting Associate Professor with the Tokyo University of Agriculture and Technology, Tokyo. He has authored or coauthored more than 100 papers in international journals and conferences and two monographs. His research interests include machine learning, tensor factorization and tensor networks, and computer vision. He serves as an Editorial Board Member for *Science China Technological Sciences*.

Yu Ding received the B.S. degree from Automation from Xiamen University, the M.S. degree in computer science from Pierre and Marie Curie University, France, and the Ph.D. degree in computer science from Telecom Paristech, Paris, France, in 2014. He is currently an artificial intelligence expert at Netease Fuxi AI Lab, Hangzhou, China. His research interests include deep learning, image and video processing, talking-head generation, animation generation, multimodal computing, affective computing, nonverbal communication (face, gaze, and gesture), and embodied conversational agent.

Wanzeng Kong (Member, IEEE) received the Ph.D. degree from the Department of Electrical Engineering, Zhejiang University, in 2008. He was a Visiting Research Associate with the Department of Biomedical Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA, from 2012 to 2013. He is currently a Full Professor and the Director of the Cognitive Computing and BCI Laboratory, School of Computer Science and Technology, Hangzhou Dianzi University. His current research interests include machine learning, pattern recognition, and cognitive computing.