Dynamically Shifting Multimodal Representations via Hybrid-Modal Attention for Multimodal Sentiment Analysis

Ronghao Lin and Haifeng Hu, *Member, IEEE*

***Abstract*—In the field of multimodal machine learning, multimodal sentiment analysis task has been an active area of research. The predominant approaches focus on learning efficient multimodal representations containing intra- and inter-modality information. However, the heterogeneous nature of different modalities brings great challenges to multimodal representation learning. In this article, we propose a multi-stage fusion framework to dynamically fine-tune multimodal representations via a hybrid-modal attention mechanism. Previous methods mostly only fine-tune the textual representation due to the success of large corpus pre-trained models and neglect the inconsistency problem of different modality spaces. Thus, we design a module called the Multimodal Shifting Gate (MSG) to fine-tune the three modalities by modeling inter-modality dynamics and shifting representations. We also adopt a module named Masked Bimodal Adjustment (MBA) on the textual modality to improve the inconsistency of parameter spaces and reduce the modality gap. In addition, we utilize syntactic-level and semantic-level textual features output from different layers of the Transformer model to sufficiently capture the intra-modality dynamics. Moreover, we construct a Shifting HuberLoss to robustly introduce the variation of the shifting value into the training process. Extensive experiments on the public datasets, including CMU-MOSI and CMU-MOSEI, demonstrate the efficacy of our approach.**

***Index Terms*—Multi-stage fusion framework, intra- and inter-modality dynamics, multimodal representations shifting, hybrid-modal attention.**

I. Introduction

With the spread of modern technology and the web, increasingly many people communicate and socialize on various social media platforms such as Twitter and Sina Weibo. Sharing personal moments or expressing personal opinions through text, audio, images and videos on the network has become a very popular social activity [1], [2], [3]. These moments usually come with rich emotions and sentiments tendencies, which produce large amounts of multimodal sentiment data and information. For governments, these multimodal data are crucial to acquire the public's opinion tendency toward policies, while for enterprises, these data are necessary to know consumers' preferences toward products. In the field of natural language processing (NLP), using multimodal data to understand people's sentiments has become an increasingly growing research area called multimodal sentiment analysis (MSA) [4], [5].

Due to the wide range of applications and essential needs, MSA has been deeply studied in recent years [6], [7], [8], [9], [10]. These works basically focus on three modalities: textual modality (spoken language), acoustic modality (audio) and visual modality (video). In the process of communication, people simultaneously utilize these three modalities to understand the attitude and emotion of one another. An example of utterance is shown in Fig. 1, which illustrates the unimodal, bimodal and trimodal interactions to predict the sentiment. When the speaker says, "That was sick!", due to the negative connotation of "sick", only using the text information may infer negative sentiment. The same utterance with a loud voice indicated in the acoustic modality remains ambiguous through bimodal interaction. Nevertheless, when the speaker says the utterance with an excited and smiling face, as shown in the visual modality, it can be positively perceived. Moreover, the loud voice increases the sentiment to strongly positive, which is closest to the real sentiment (annotated as 2.2) of the speaker. The example shows that trimodal interaction can maximize mutual information and most accurately express sentiment, and multiple modalities are complementary and indispensable to one another.

The interaction of multiple modalities can be divided into intra-modality and inter-modality interaction. Among the deep learning methods, the processing mechanism of the human brain is simulated using neural network architecture [11]. Similarly, researchers would like to propose models that can handle multiple modalities as the human brain does, i.e., to concurrently model the intra-modality and inter-modality dynamics [8], [12], [13]. How to precisely represent each modality and properly fuse different modalities have become two key challenges of MSA.

The first challenge is to learn an discriminative and robust modality-specific representation that captures characteristic features of each modality. Nevertheless, only extracting low-level features such as the Mel-frequency cepstral coefficients (MFCCs) features [14] for acoustic modality or global

Manuscript received 23 March 2023; revised 6 June 2023 and 6 July 2023; accepted 4 August 2023. Date of publication 9 August 2023; date of current version 2 February 2024. This work was supported by the National Natural Science Foundation of China under Grant 62076262. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Ichiro Ide. *(Corresponding author: Haifeng Hu.)*

The authors are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: linrh7@mail2.sysu.edu.cn; huhaif@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TMM.2023.3303711

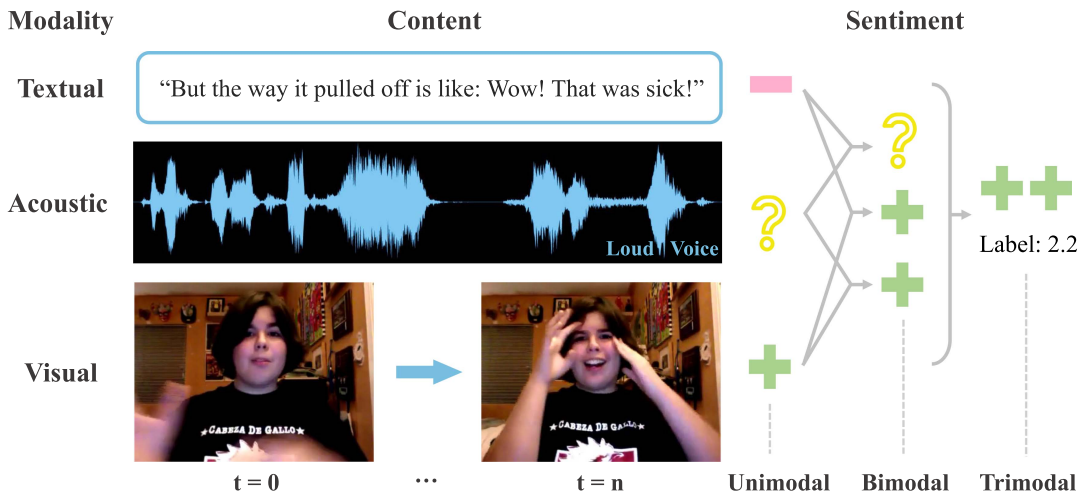


Fig. 1. Example of utterance containing unimodal, bimodal, and trimodal interaction for sentiment prediction. The example indicates that trimodal interaction can predict sentiment most accurately.

vectors (Glove) [15] for word representation is not sufficient for the model. Due to the high sensitivity to heterogeneous data distributions, the low-level representation must be further processed into a higher level representation [16]. In recent years, recurrent neural networks (RNNs) have achieved superior performance in extracting features for time sequential data such as audio [17] and video [18]. However, traditional RNNs will suffer from long-term dependency problem because of exploding and vanishing gradients problem [19], [20], [21]. To improve the problem of long-term dependencies, gating mechanisms are proposed for RNNs such as long short-term memory (LSTM) [20], [22] networks and gated recurrent unit (GRU) [23], [24] networks. In the following work, we use these RNN variants to learn time-dependent representations of visual modality and acoustic modality.

For the textual modality, the Transformer [25] architecture has been successfully utilized to learn contextual textual representations in NLP, such as bidirectional encoder representations from transformers (BERT) [26] and XLNet [27] based on Transformer-XL [28]. However, as stated in previous works [26], [29], the excellent performance in textual presentations of Tranformer-based models largely comes from pre-training on a large-scale corpus. Illustrating ideas from transfer learning [30], Transformer models must be fine-tuned for different specific downstream tasks to stimulate their true potential [31]. This pre-trained model with fine-tuning paradigm is increasingly popular in research [32], [33]. In this article, we also use this paradigm to fine-tune the pre-trained textual representation by capturing the inter-modality dynamics with other modalities for MSA task. Most MSA methods [34], [35], [36] directly apply pre-trained Transformer model to extract textual representations for multimodal fusion processes which lack explicit exploration on the intra-modality dynamics. Thus, our pre-trained textual representation consists of various granularity features that are output from various layers of the Transformer model, which further exploits the intra-modality dynamics of textual representations.

For the second challenge, inter-modality dynamics are usually captured through the fusion process among different modalities [36]. Previous multimodal fusion methods can be divided into early fusion (feature-level), late fusion (decision-level) and hybrid fusion (both feature-level and decision-level) [37]. First, early fusion methods [8], [9], [38], [39] integrate multimodal features by various concatenating methods at the input level. However, because the parameter space of different modalities is inconsistent, early fusion may limit capabilities in learning the intra-modality dynamics and result in overfitting since the modalities are fused at an early stage. Meanwhile, late fusion methods [35], [40], [41], [42] independently train multimodal data on individual unimodal models and fine-tune by performing decision voting on the output layer. Although late fusion will not suffer from the inconsistency of different modality parameter spaces, it may neglect low-level interactions among the modalities and hinder the model from learning the effective inter-modality dynamics. Hybrid fusion methods [43], [44], [45] perform multimodal fusion at input and output levels to exert the advantages of both early and late fusion methods.

To address the aforementioned challenges, guided by the idea of hybrid fusion, we propose the hybrid-modal attention mechanism to efficiently exploit the intra- and inter-modality dynamics. Specifically, the hybrid-modal attention mechanism can be considered a multi-stages fusion process. First, we construct a Multimodal Shifting Gate (MSG) to fine-tune the representations of different modalities by shifting the representation through attention with the other two modalities. Second, for the textual modality, we use MSG to shift the low-level features extracted by the first few layers of the Transformer model. Then, we feed it back into the remaining layers of the Transformer model to obtain the final textual representation, which contains rich semantic information. Here, we utilize different granularity features of the Transformer model to capture sufficient intra-modality dynamics. Next, we design a Masked Bimodal Adjustment (MBA) module on the final textual representation to reduce the modality gaps between the textual modality

and the other two modalities. Doing so can alleviate the inconsistency of the representation space and separately obtain two robust textual representations based on acoustic and visual modalities. Then, for acoustic and visual representations, we feed them and the two textual representations into MSG and attain the final representation of acoustic and visual modalities. Lastly, we concatenate the representations of the three modalities for the sentiment prediction task.

In conclusion, we propose a novel multi-stage fusion framework to dynamically shift multimodal representations via a hybrid-modal attention mechanism. The main contributions of our work are stated below:

- We present a module to learn the inter-modality dynamics among multiple modalities, which we call the Multimodal Shifting Gate (MSG). Unlike previous methods [33], [46], which only considered the representation of textual modality, we fine-tune the textual, acoustic and visual modalities in the model by shifting the representations and concatenate the final representations of three modalities to predict the sentiment for the MSA task.
- Because each layer of the Transformer model outputs various granularity features of the textual representation, we explicitly utilize low-level syntactic features and high-level semantic features to sufficiently model the intra-modality dynamics of the textual modality. Concretely, we adopt MSG to the output of the shallow layer of the Transformer model for the textual modality and to the output of the whole Transformer model for acoustic and visual modalities.
- We design a module named Masked Bimodal Adjustment (MBA) on high-level textual features to learn two robust textual representations that focus on the acoustic or visual modality. The MBA module is applied to concentrate on the common sentiment information in the bimodal representation and filter out irrelevant high-level textual features. It can effectively reduce the impact of the inconsistency problem of different representation spaces and reduce the modality gap.
- We construct a Shifting HuberLoss for the shifting value between multiple modalities. The Huber loss can decrease the punishment degree of outliers to avoid the interference of the noise inside acoustic and visual modalities and enhance the robustness of measuring the shifting value at the batch-level. The Shifting HuberLoss and the original label regression loss constitute the total loss.
- We conduct extensive experiments on both CMU-MOSI [47] and CMU-MOSEI [48] datasets to demonstrate the superiority of our method compared with state-of-the-art methods in the MSA task.

II. Related Work

A. Multimodal Sentiment Analysis

Multimodal sentiment analysis (MSA) has become a significant research area in natural language processing (NLP) that integrates information from different modalities to predict sentiment intensity. Previous methods mainly focus on multimodal fusion after modality representation learning. Tenser fusion network (TFN) [8] adopts outer product between unimodal representations to construct a tensor fusion network for the multimodal representations. This model belongs to early fusion methods which fuse features of different modalities by various ways of concatenating at the input level. However, early fusion methods limit capabilities in learning intra-modality dynamics and tend to overfit. A contextual LSTM network [41] learns the utterance-level contextual information in the same video to capture emotional characteristics. This model pertains to late fusion methods which integrate different modalities at the prediction level while may neglect low-level interactions between the modalities and prevent the model from learning effective inter-modality dynamics. To combine the advantages of the above fusion methods, hybrid fusion methods perform multimodal fusion at the feature-level and prediction-level. For example, multimodal transformer (MulT) [44] implicitly adapts representations stream from one modality to another based on cross-modality interactions across different time steps. Guided by the idea of hybrid fusion, we propose the hybrid-modal attention mechanism to capture the intra- and inter-modality dynamics in multiple modalities.

Specifically, both recurrent attended variation embedding network (RAVEN) [46] and multimodal adaption gate (MAG-BERT/MAG-XLNet) [33] generate multimodal shifted vectors attending to nonverbal information and utilize the vectors to shift word representations. Drawn lessons from these methods, we design a module named Multimodal Shifting Gate (MSG) to fine-tune representations based on various modality features. Different from the methods only focus on shifting word representations, we extend MSG module for each modality to sufficiently capture inter-modality dynamics. Moreover, previous fusion methods may suffer from the problem of the inconsistency of parameter spaces among multiple modalities. To mitigate the impact of the inconsistency of different modality spaces and reduce the disparities of modalities, we present a Masked Bimodal Adjustment (MBA) module on textual modality with the other modalities to learn robust textual representations.

B. Pre-trained Language Model

Contextual language representation models pre-trained on large text corpora have been demonstrated as state-of-the-art methods to learn textual representations in many NLP tasks. In these pre-trained language models, Transformer [25] is a popular and effective architecture, which is presented as an encoder-decoder paradigm to model sequential textual data. The superior performance of Transformer architecture originates from the multi-head self-attention module which attends to each word in a sentence by interactions with other words. To extract more relations among words, bidirectional encoder representations from transformers (BERT) [26] learn textual features from masked-out tokens in the sequence. In our work, firstly, the input embedder of BERT is utilized to generate token embedding, segment embedding, and position embedding. Then, these input embeddings are applied to multiple encoder layers consisting of a multi-head attention layer and a normalized residual connected

feed-forward layer. Lastly, a special [CLS] token is appended in the first place of the sentence as textual embedding to predict the final sentiment scores after an affine transformation.

Recently, Transformer-XL architecture [28] improves the performance of Transformer by capturing more long-range dependencies and enhancing model's prediction capability. Specifically, Transformer-XL architecture interacts context information among the segments with a recurrence mechanism and reuses hidden states with relative positional encoding avoiding temporal confusion. Compared to auto-encoder model BERT, XLNet [27] based on the Transformer-XL architecture is an auto-regressive model which promotes the independence among the masked out tokens and captures unidirectional context. To further show the effectiveness of our methods, we additionally utilize XLNet as the pre-trained language model to learn the textual representations in the framework.

Specifically, the output of each layer of the Transformer model contains various granularity features of the textual modality. The front layers output more low-level syntactic features while the last layers output more high-level semantic features [33], [49], [50]. Different from previous methods [32], [34], [35], [36] which only use the whole pre-trained Transformer model to learn textual representation, we utilize the outputs of different layers of the Transformer model to process different levels of textual features. Due to this, we can learn more intra-modality dynamics and attain efficient modality-specific representations to improve the performance of the proposed framework.

C. Recurrent Neural Network

Recurrent neural networks (RNNs) are designed to process various lengths of the time-sequential data by neurons with self-feedback mechanism, which have achieved considerable improvement in many tasks. To further address the issues of exploding and vanishing gradients in parameters update, long short-term memory (LSTM) [20], [22] networks introduce input gates, forget gates, and output gates into RNNs, which is broadly used to learn visual and acoustic representations in MSA task [35], [51]. Moreover, gated recurrent unit (GRU) [23], [24] networks are proposed to simplify the gating mechanism into update gates and reset gates. Besides, compared with single-directional RNNs, bi-directional RNNs increase another network layer to transmit information in the reverse order of time to obtain context information. In recent MSA research, [52] and [53] adopt LSTM to transform encoded sequences of different modalities into context-dependent hidden states and respectively win the MuSe 2020 [54] and 2021 [55] challenges, which demonstrate the powerful learning ability of LSTM. Similarly, we utilize these RNN variants with different direction settings to learn time-dependent modality-specific representations from low-level visual and acoustic features.

III. Proposed Architecture

In this section, we describe the proposed framework in detail. Firstly, we make the definition of the MSA task. Secondly, we introduce the modality-specific representation learning methods. For acoustic and visual modalities, we utilize RNN variants such as LSTM and GRU to extract their modality features. For textual modality, we adopt different layers of Transformer models such as BERT and XLNet to extract different levels of features, including low-level syntactic features and high-level semantic features. Then, we explain the multi-stage fusion framework, which aims at dynamically shifting multimodal representations via a hybrid-modal attention mechanism, as shown in Fig. 2. Specifically, we shift textual, acoustic, and visual modalities by Multimodal Shifting Gate (MSG) to learn more robust and informative multimodal representations. Especially for textual modality, MSG is used both in low- and high-level representation to capture sufficient intra-modality dynamics. For acoustic and visual modalities, we apply Masked Bimodal Adjustment (MBA) on high-level textual modality before feeding it into MSG to learn acoustic modality and visual modality representations. The MBA module can alleviate the problem of inconsistency of different representation spaces among different modalities and reduce the modality gap. Next, we concatenate representations of three modalities to obtain the final multimodal representation. Lastly, we construct a Shifting HuberLoss to take the shifting value into consideration, which is used with label regression loss to train the whole framework to predict the sentiment of the corresponding utterances.

A. Task Definition

The MSA task is to analyze sentiment with multimodal data of an utterance [56] by scoring the sentiment intensity. As shown in Fig. 2, the input to the model consists of textual, acoustic, and visual modalities, which can be represented as $T \in \mathbb{R}^{N_t \times d_t}$, $A \in \mathbb{R}^{N_a \times d_a}$ and $V \in \mathbb{R}^{N_v \times d_v}$ respectively. Here N_m denotes the sequence length of modality $m \in \{t, a, v\}$ and d_m denotes the respective feature dimensions. The output of the model is the predictive sentiment score $\hat{y} \in \mathbb{R}$, which is a continuous intensity variable in valence-arousal space, where the valence represents positive and negative feelings while the arousal represents the degree of sentiment [57]. In order to be consistent with previous MSA researches [33], [35], [51], here we represent the sentiment score simply in one dimension, where positive (>0) and negative (<0) values represent the valence and the value size from 0 to 3 in a linear scale represents the arousal.

B. Modality-specific Representation Learning

Firstly, we adopt low-level feature extraction to learn modality-specific representation for each modality. Following [8], [33], [35], [46], in preprocessing stage, we utilize pre-trained ToolKits to extract initial feature vectors A and V for acoustic and visual modalities. The details of preprocessing will be explained in Section IV. Then due to the strong time-dependent learning capability of RNNs, we utilize RNN variants to capture timing characteristics of acoustic and visual modalities, whose end-state hidden representations can be formulated as:

$$F_a = RNN(A; \theta_a^{RNN}) \tag{1}$$

$$F_v = RNN(V; \theta_v^{RNN}) \tag{2}$$

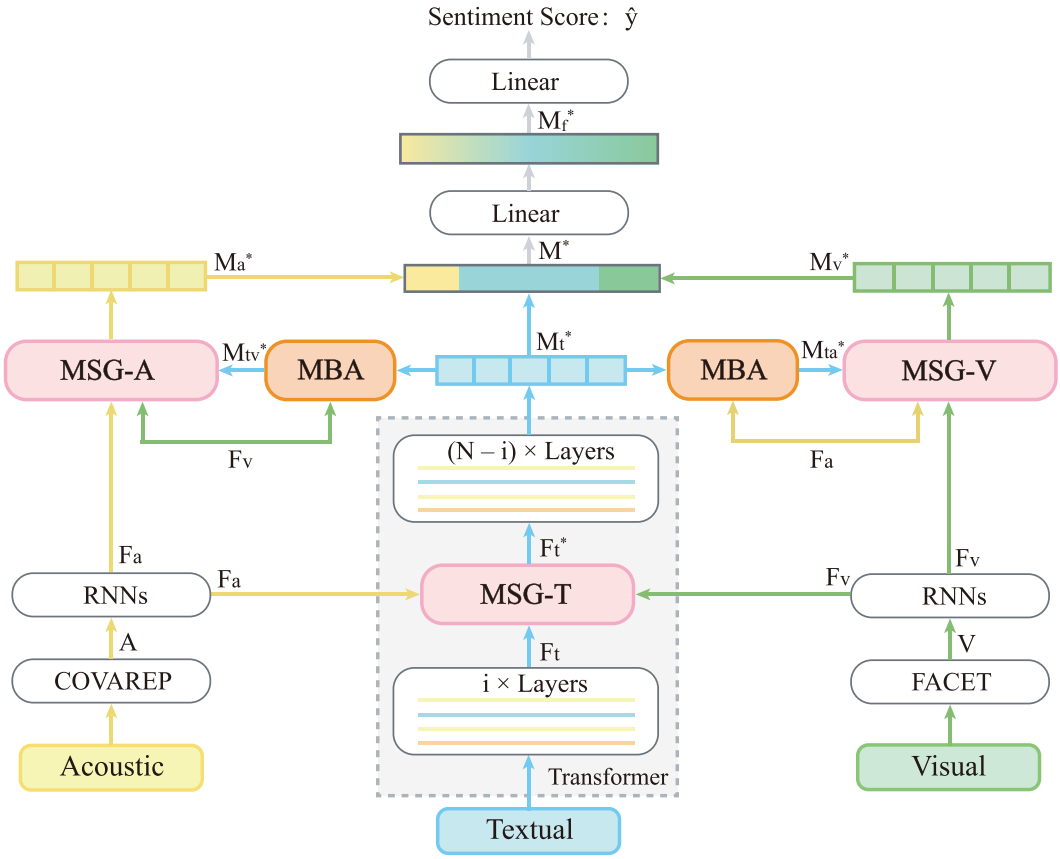


Fig. 2. Illustration of our proposed multi-stage fusion framework via the hybrid-modal attention mechanism, which mainly contains two components, Multimodal Shifting Gate (MSG) and Masked Bimodal Adjustment (MBA) to sufficiently capture intra- and inter-modality dynamics and dynamically shift multimodal representations.

where θ_a and θ_v denote the parameters for acoustic modality and visual modality, respectively. Here RNN represents four kinds of RNN variant networks, which are Long Short-Term Memory [20] with single direction (sLSTM) and bi-direction (bLSTM) and Gated Recurrent Unit [24] with single direction (sGRU) and bi-direction (bGRU) [58]. We utilize different networks to demonstrate the generalization of our proposed framework whose experimental results are shown in Section IV. For sLSTM and sGRU, F_a and F_v are $\in \mathbb{R}^{N_m \times d_m}$ while for bLSTM and bGRU, F_a and F_v are $\in \mathbb{R}^{N_m \times 2d_m}$. For brevity, we take sLSTM as an example to make the following deduction.

Then for textual modality, since the great success of the pre-trained language model, we use the pre-trained Transformer-based models such as bidirectional encoder representations from transformers (BERT) [26] and XLNet [27] to extract textual representation, which can be formulated as:

$$F_t = Transformer\left(T; \theta_t^{Transformer}\right) \in \mathbb{R}^{N_t \times d_t} \tag{3}$$

where θ_t represents the parameters for textual modality in the Transformer model. It is worth noting that F_t denotes syntactic-level textual representation which is the output of former i layers of the Transformer model. We extract this low-level syntactic feature for fine-tuning and then feed it back into the latter $(N - i)$ layers of the Transformer model to learn the final semantic-level textual representation. Compared to prior methods [34], [35], [36] which only utilize the final output of the Transformer model, we fine-tune both syntactic- and semantic-level textual representations to capture intra-modality dynamics sufficiently.

C. Hybrid-modal Attention Mechanism

After extracting modality-specific representations, we propose the hybrid-modal attention mechanism to fine-tune them. As shown in Fig. 2, the mechanism mainly contains two components, Multimodal Shifting Gate (MSG) and Masked Bimodal Adjustment (MBA), which can sufficiently fine-tune the original modality-specific representations and learn the inter-modality dynamics among three modalities.

1) Multimodal Shifting Gate (MSG): The MSG can integrate the other modalities into the original representation in the feature space and obtain the final multimodal-shifted representations. Different from [33] and [46] which only focus on shifting word representations, we extend MSG module for each modality to capture inter-modality dynamics more sufficiently. The module can be divided into three kinds according to different input modalities, which are MSG-T, MSG-A, and MSG-V correspondingly. Without losing generality, we take MSG-T as an example to explain the working principle of MSG.

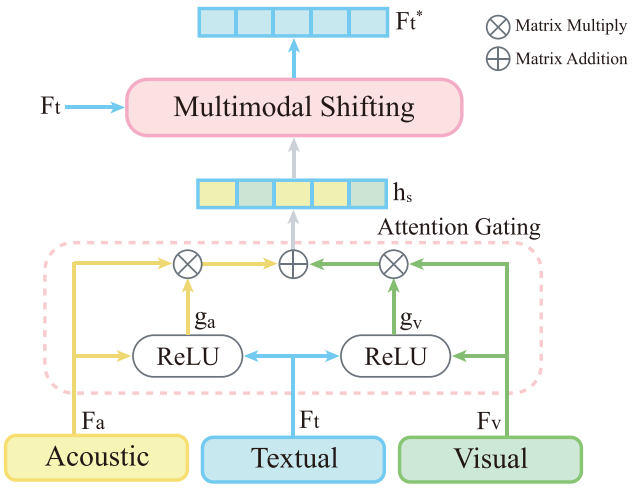


Fig. 3. Illustration of Multimodal Shifting Gate for textual modality (MSG-T). The attention gating focus on the inter-dynamics among three modalities and generate the shifting vector, which is combined with textual modality to obtain multimodal shifted textual representation.

As shown in Fig. 3, the input to MSG-T is the original representations of textual, acoustic, and visual modalities, which can be represented as triplet (F_t, F_a, F_v).

Firstly, the textual modality is concatenated with the other two modalities as bimodal factors $[F_t; F_a]$ and $[F_t; F_v]$ to produce two gating vectors g_a, g_v through a $ReLU$ activation layer, which can be formulated as:

$$g_a = ReLU(W_{ga}[F_t; F_a] + b_{ga}) \tag{4}$$

$$g_v = ReLU(W_{gv}[F_t; F_v] + b_{gv}) \tag{5}$$

where W_{ga} and W_{gv} are weight matrices for acoustic gating vector and visual gating vector and b_{ga} and b_{gv} are corresponding scalar biases. The attention gates focus on bimodal relevant information between textual modality and acoustic or visual modality.

Then, we calculate the shifting vector h_s by fusing representations of acoustic and visual modalities multiplied with the gating vectors, which can be formulated as:

$$h_s = g_a \cdot (W_a F_a) + g_v \cdot (W_v F_v) + b_h \tag{6}$$

where W_a and W_v are weight matrices for acoustic and visual representations and b_h is the bias vector.

Lastly, we weigh the shifting vector h_s and integrate it with the original textual representation to dynamically shift the textual representation. The multimodal shifted textual representation F_t^* can be formulated as:

$$F_t^* = F_t + \eta h_s \tag{7}$$

where η is an adaptive scaling factor aiming at constraining the effect of the shifting vector h_s within a desirable range, which is formulated as:

$$\eta = \min\left(\frac{\|F_t\|_2}{\|h_s\|_2 + \xi}\mu, 1\right) \tag{8}$$

where $\|F_t\|_2$ and $\|h_s\|_2$ denote the L_2 norm of F_t and h_s vectors, respectively. μ is a learnable parameter that is initialized as a threshold hyper-parameter determined by cross-validation and ξ is equal to 10^{-6} to avoid the denominator being 0.

As the output of MSG-T, the multimodal shifted textual representation F_t^* presently has fused information from its accompanying nonverbal contexts and contained the inter-modality dynamics between textual modality and the others. As for MSG-A and MSG-V, the learning processes are the same as MSG-T while the output are multimodal shifted acoustic representation M_a^* and multimodal shifted visual representation M_v^*, respectively.

2) Masked Bimodal Adjustment (MBA) on Textual Modality: As mentioned before, we insert MSG-T inside the transformer model to utilize different levels of features on textual modality aiming at sufficiently capturing intra-modality dynamics. Since we get the multimodal shifted textual representation F_t^*, we need to feed it back into the transformer model for the remaining layers to learn the final textual representations M_t^*. However, directly doing so may cause the inconsistency of parameter space with the other modalities due to the high-level semantic features which are processed by multi-layers in M_t^*. Previous methods such as [33] and [35] have not considered the inconsistency problem and directly fuse the representations of three modalities. Diverse from these methods, we design the MBA module for textual modality with the other modalities to alleviate the impact of the problem of modality space inconsistency. Furthermore, another reason for the inconsistency is the inherent sentiment variance among different modalities, causing the modality gap between each representation. The MBA module combines the information of bimodal can intuitively reduce the modality gap concurrently. Without losing generality, we take MBA between textual modality and visual modality as an example to explain how we construct the MBA module.

As shown in Fig. 4, firstly, following [44] and [32], to ensure each element in the representations has enough awareness of its neighborhood elements, we adopt a 1D temporal convolutional layer to both modalities, represented as:

$$M_t^{*\prime} = Conv1D(M_t^*, k_t) \in \mathbb{R}^{N_t \times d} \tag{9}$$

$$F_v^{\prime} = Conv1D(F_v, k_v) \in \mathbb{R}^{N_v \times d} \tag{10}$$

where k_t and k_v represent the size of convolutional kernels for textual and visual modality, and d is a common dimension that keeps the dimension of textual and visual modalities the same in the parameter space. Due to the significantly higher feature dimension of M_t^* than F_v, after the convolution layer, the value of $M_t^{*\prime}$ in the feature space becomes larger and larger than $F_v^{\prime}$ during the training process. So we scale $M_t^{*\prime}$ and $F_v^{\prime}$ to $\overline{M_t^*}$ and $\overline{F_v}$ respectively to normalise representations and prevent this situation by:

$$\overline{M_t^*} = \frac{M_t^{*\prime}}{\sqrt{\|M_t^{*\prime}\|_2}} \tag{11}$$

$$\overline{F_v} = \frac{F_v^{\prime}}{\sqrt{\|F_v^{\prime}\|_2}} \tag{12}$$

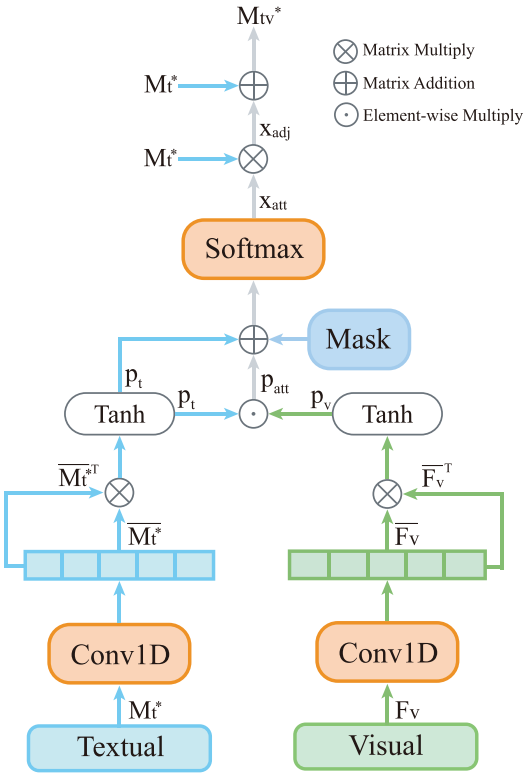


Fig. 4. Illustration of Masked Bimodal Adjustment (MBA) on textual modality with visual modality. Utilizing bimodal attention matrix between two modalities and the mask for textual modality, MBA can obtain more robust textual representations based on the other modality.

where $\|M_t^{*\prime}\|_2$ and $\|F_v'\|_2$ denote the L_2 norm of $M_t^{*\prime}$ and F_v' vectors.

Then, we utilize $\overline{M_t^*}$ and $\overline{F_v}$ as queries while their transposed as keys to obtain their dot product which contains the intra-modality features. To further interact the features in textual modality and visual modality, we project both the dot products into an attention space followed by a $Tanh$ activation function, which can be formulated as:

$$p_t = Tanh(W_{pt}(\overline{M_t^*} \cdot \overline{M_t^*}^T) + b_{pt}) \tag{13}$$

$$p_v = Tanh(W_{pv}(\overline{F_v} \cdot \overline{F_v}^T) + b_{pv}) \tag{14}$$

where W_{pt} and W_{pv} are weight matrices for the dot products of textual and visual modality and b_{pt} and b_{pv} are corresponding scalar biases. After the non-linear activation layer, we multiply p_t and p_v in element-wise to learn the visual-specific attention weight p_{att} of the textual modality, formulated as:

$$p_{att} = p_t \odot p_v \tag{15}$$

where $\odot$ denotes the element-wise multiplication between matrices.

If we only use p_{att} for attention value, the effect of textual modality will be significantly weakened due to the sparsity of visual modality. Hence, we sum p_{att} and p_t for effective visual-informed soft attention, which can further reduce the difference between the textual modality space and the visual modality space. Next, we feed the summation into a Softmax function to obtain the attention matrix x_{att}. Moreover, to remove the interference of the padding part of textual representation, we introduce a mask matrix using $-\infty$ on the padding position and 0 on the token position so that the attention value on the padding position will turn into 0 after Softmax function. The masked attention matrix x_{att} can be represented as:

$$x_{att} = Softmax(p_t + p_{att} + mask) \tag{16}$$

Finally, we multiply x_{att} and M_t^* to attain masked adjustment matrix x_{adj} and then add x_{adj} and M_t^* for the textual representation M_{tv}^* based on the bimodal interaction with visual modality, formulated as:

$$M_{tv}^* = M_t^* + x_{adj} = M_t^* + x_{att} \cdot M_t^* \tag{17}$$

The representation M_{tv}^* is more robust due to the masked bilinear attention [59] shown in (15)–(16) providing tight interaction between textual and visual modality. As for MBA on textual modality with acoustic modality, the output turns into representation M_{ta}^* with the same process. The final masked attention matrix x_{att} makes the model focus on the most critical tokens of textual modality which share common sentiment features with visual and acoustic modalities. Besides, it can filter out other irrelevant high-level semantic features contained in the textual representation. Aiming at reducing the inconsistency problem of modality space, M_{tv}^* and M_{ta}^* lastly replace textual representation M_t^* to be fed into MSG-A and MSG-V, respectively.

3) Multi-stage Fusion: As shown in Fig. 2, combining MSG and MBA as the hybrid-modal attention mechanism in the proposed framework, we fuse different modalities in multiple stages to sufficiently capture intra- and inter-modality dynamics.

Firstly, MSG is proposed to shift the origin representation (F_t, F_a, F_v) by the attention gating to obtain the final multimodal representations (M_t^*, M_a^*, M_v^*) for three modalities. Specifically, MSG-T is used inside the Transformer model to learn intra-modality dynamics from syntactic-level features to semantic-level features for the textual modality.

Secondly, MBA is designed to deal with the problem of modality space inconsistency and learn two robust textual representations (M_{ta}^*, M_{tv}^*) based on visual and acoustic modality separately, which also contains multimodal fusion process to get the masked attention matrix.

Thirdly, we concatenate three multimodal representations (M_t^*, M_a^*, M_v^*) into the final representation M^*, and then we pass it through a fully connected layer to further fuse three representations and effectively interact them. Finally, we attain the representation M_f^* and empirically feed the first embedding ([CLS]) as the whole utterance representation into a linear regression layer for the sentiment score $\hat{y}$.

D. Loss Construction

Unlike previous methods [33], [46] which only utilize label regression loss in the linear regression layer for training, we construct a Shifting HuberLoss with it to compose the total loss for the whole framework.

1) Label Regression Loss: We use Mean Squared Error (MSE) as the loss function for the label regression problem to

measure the difference between predicted sentiment score $\hat{y}$ and ground truth score y, which can be formulated as:

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2 \tag{18}$$

where N denotes the sequence length of the sentence.

2) Shifting HuberLoss: Because the shifted representation is more informative than the original representation, the shifting vector h_s can effectively represent the variation for three modalities in the process of fine-tuning. Specially, in consideration of the noise in acoustic and visual modalities, we utilize Huber Loss to reduce the impact of outliers and obtain a more robust shifted representation. The Shifting HuberLoss is formulated as:

$$\mathcal{L}^{m}_{shift} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{2}{h_s^{(i)}}^2, \text{for}|h_s^{(i)}| \le \delta \\ \delta\cdot(|h_s^{(i)}| - \frac{1}{2}\delta), \text{otherwise}\end{cases} \tag{19}$$

where m denotes different modalities and δ is a hyper-parameter. When the absolute value of the corresponding element $h_s^{(i)}$ is less than δ, the formulation of the Shifting HuberLoss is a square error function while when $h_s^{(i)}$ is larger than δ, the function becomes a linear error function. Doing so can reduce the punishment degree of outliers and then enhance the robustness to the noise.

3) Total Loss: The total loss consists of Label Regression Loss and Shifting HuberLoss, which can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \alpha\mathcal{L}^{t}_{shift} + \beta\mathcal{L}^{a}_{shift} + \gamma\mathcal{L}^{v}_{shift} \tag{20}$$

where α, β, and γ are hyper-parameters measuring the contribution of different loss terms to obtain superior performance in the multimodal sentiment analysis task.

IV. Experiment

In this section, we first describe two public experimental datasets. Then, we provide preprocessing computational descriptors to extract features for three modalities. Next, we introduce state-of-the-art MSA methods as baselines and the calculation principles of evaluation metrics for comparison. Then, we present the experiment implementation details for reproduction. Lastly, we report the experiment results of the proposed methods compared with the baselines and further analyze the results.

A. Datasets

We evaluate the proposed method on the following datasets:

CMU-MOSI [47] is a multimodal sentiment dataset focusing on MSA task, consisting of 2,199 opinion video segments with a total of 26,295 words in the utterances excerpted from 93 YouTube movie reviews. The CMU-MOSI dataset is annotated with labels for sentiment intensity in the range of -3 to +3, where the positive values indicate positive sentiment and vice versa.

CMU-MOSEI [48] is a large dataset of multimodal sentiment analysis and emotion recognition, consisting of 23,453 monologues videos utterances covering 250 distinct topics from 1,000 YouTube speakers. The utterances in CMU-MOSEI are randomly chosen from various movie review topics, annotated with sentiment scores in the range of [-3, +3] and six different emotion values.

B. Preprocessing

Referring to prior methods [12], [33], [39], [60], we adopt the following computational descriptors for three modalities.

Textual: [33] transcribes the videos using YouTube API followed by manual correction. We directly utilize the same textual descriptors to obtain textual features with 768 dimensions.

Acoustic: we use the acoustic analysis framework COVAREP [61] to extract 74-dimensional low-level acoustic features including 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking, speech polarity, glottal source parameters, spectral envelope, peak slope parameters, and maxima dispersion quotients, etc.

Visual: we employ the facial expression analysis toolkits Facet [62] to extract facial features including facial landmarks, action units, head pose, gaze tracking and histogram of oriented gradients (HOG) features, etc. Specially, the dimension of visual features is 47 for CMU-MOSI dataset and 35 for CMU-MOSEI dataset.

The acoustic and visual features are extracted from each utterance of the full video clip at 100 Hz and 30 Hz, respectively. Besides, there are co-occurring acoustic and visual features between individual words, which are averaged across each feature for each word. Following the convention formed by [63], we align three modalities for each word in the utterance using forced alignment [64].

C. Baselines

We compare the performance of the proposed method with various state-of-the-art MSA methods across multiple metrics. For fair comparison, we reproduce the best results of corresponding baselines by running hyper-parameters grid search based on various language models, including Glove [15], BERT [26] and XLNet [27] as shown in Tables I and II.

TFN (Tenser Fusion Network) [8] introduces multi-dimensional tensors to capture unimodal, bimodal, and tri-modal interactions explicitly which is indicated as intra- and inter-dynamics.

LMF (Low-rank Multimodal Fusion) [39] performs efficient multimodal fusion utilizing low-rank tensors to drastically reduce computational complexity while improve performance.

MFN (Memory Fusion Network) [42] leverages LSTM functions to learn view-specific interactions for each modality and discovers the cross-view interactions by a delta-memory attention network, which are summarized by multi-view gated memory eventually.

MARN (Multi-attention Recurrent Network) [9] presents a multi-attention recurrent network to discover cross-modality interaction through time and stores them in long-short term hybrid memory.

MFM (Multimodal Factorization Model) [13] learns modality-specific generative factors and multimodal discriminative factors to factorize representation for interpretability of the interactions which influence multimodal learning.

TABLE I
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND BASELINES ON CMU-MOSI DATASET

Methods	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
Glove					
TFN [8]	32.2	76.4	76.3	1.017	0.604
LMF [39]	30.6	73.8	73.7	1.026	0.602
MFN [42]	32.1	78.0	76.0	1.010	0.635
MARN [9]	34.7	77.1	77.0	0.968	0.625
MFM [13]	36.2	78.1	78.1	0.951	0.662
RAVEN [46]	33.8	78.8	76.9	0.968	0.667
MulT [44]	33.6	79.3	78.3	1.009	0.667
MCTN [65]	-	79.3	79.1	0.909	0.676
QMF [66]	35.5	79.7	79.6	0.915	0.696
BERT					
TFN [8]	33.7	80.2	80.1	0.926	0.671
LMF [39]	32.7	80.1	80.0	0.916	0.696
MFN [42]	34.2	80.0	80.0	0.939	0.676
MFM [13]	33.3	80.0	80.1	0.934	0.673
MulT [44]	35.0	80.5	80.5	0.916	0.696
MISA [51]	42.8	82.3	82.2	0.803	0.750
ICCN [34]	39.0	83.1	83.0	0.862	0.714
MAG-BERT [33]	42.9	84.1	84.1	0.781	0.769
CM-BERT [32]	43.8	84.3	84.2	0.759	0.787
Self-MM [35]	**45.3**	84.6	84.6	**0.725**	**0.790**
Proposed (BERT)	**45.3**	**85.7**†	**85.6**†	0.748	0.782
XLNet					
TFN [8]	34.2	80.4	80.4	0.930	0.670
LMF [39]	35.9	80.2	80.3	0.910	0.699
MFN [42]	36.1	80.2	80.1	0.947	0.674
MFM [13]	34.1	80.3	80.4	0.940	0.667
MulT [44]	35.5	80.8	80.6	0.905	0.701
MISA [51]	43.4	83.8	83.9	0.760	0.786
Self-MM [35]	44.3	84.9	84.9	0.710	0.810
MAG-XLNet [33]	44.4	85.1	85.2	0.740	0.804
Proposed (XLNet)	**46.1**†	**87.0**†	**87.0**†	**0.696**†	**0.816**†

The best results are marked in bold on BERT and XLNet, and † means the corresponding result is significantly better than the state-of-the-art with p-value < 0.05 based on paired t-test.

TABLE II
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND BASELINES ON CMU-MOSEI DATASET

Methods	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
Glove					
TFN [8]	49.8	79.4	79.7	0.610	0.671
LMF [39]	50.0	80.6	81.0	0.608	0.677
MFN [42]	49.1	79.6	80.6	0.618	0.670
RAVEN [46]	50.2	79.0	79.4	0.605	0.680
MulT [44]	48.2	80.2	80.5	0.638	0.659
QMF [66]	47.9	80.7	79.8	0.640	0.658
BERT					
TFN [8]	52.2	82.6	82.3	0.644	0.748
LMF [39]	52.0	83.7	83.8	0.625	0.759
MFN [42]	51.1	84.0	83.9	0.644	0.749
MFM [13]	50.8	83.4	83.4	0.665	0.740
MulT [44]	52.1	84.0	83.9	0.620	0.759
MISA [51]	51.0	84.6	84.1	0.591	0.761
ICCN [34]	51.6	84.2	84.2	**0.565**	0.713
MAG-BERT [33]	51.9	85.0	85.0	0.602	0.778
Self-MM [35]	**53.2**	84.8	84.9	0.594	0.765
Proposed (BERT)	52.8	**85.4**†	**85.4**†	0.583	**0.787**†
XLNet					
TFN [8]	52.4	82.9	82.5	0.656	0.755
LMF [39]	51.9	83.3	83.2	0.653	0.758
MFN [42]	51.5	83.3	83.2	0.667	0.751
MFM [13]	50.4	84.1	84.1	0.661	0.742
MulT [44]	52.3	84.3	84.2	0.642	0.761
MISA [51]	50.6	85.0	85.0	0.603	0.779
Self-MM [35]	52.1	85.6	85.6	0.584	0.789
MAG-XLNet [33]	51.3	85.8	85.9	0.593	0.790
Proposed (XLNet)	**52.6**†	**86.3**†	**86.3**†	**0.581**	**0.795**†

The best results are marked in bold on BERT and XLNet, and † means the corresponding result is significantly better than the state-of-the-art with p-value < 0.05 based on paired t-test.

RAVEN (Recurrent Attended Variation Embedding Network) [46] is a word-level RNN-based fusion approach, which models language by shifting word representations based on nonverbal behaviors including vocal patterns and facial expressions.

MulT (Multimodal Transformer) [44] implicitly adapts representations stream from one modality to another attending to cross-modality interactions across distinct time steps.

MCTN (Multimodal Cyclic Translation Network) [65] learns joint-representations by translating from source modality to target modality and adopts cycle consistency loss to ensure the translation retaining the maximal information.

MISA (Modality-Invariant and -Specific Representations) [51] projects each modality to modality-invariant subspace to learn representations with commonalities inside modalities and modality-specific subspace to extract characteristic features.

MAG-BERT/MAG-XLNet (Multimodal Adaption Gate) [33] designs an attachment gate to generate a shift conditioned on nonverbal modalities to the internal textual representation of pre-trained BERT [26] and XLNet [27].

ICCN(Interaction Canonical Correlation Network) [34] learns multimodal embeddings by capturing correlations between three modalities based on deep canonical correlations analysis.

CM-BERT (Cross-Modal BERT) [32] fine-tunes the pre-trained BERT model focusing on the interaction of textual and acoustic modality by adjusting the weight of words.

QMF (Quantum-inspired Multimodal Fusion) [66] models complicated interactions and correlations among modalities by superposition and entanglement inspired by quantum theory.

Self-MM (Self-Supervised Multi-Task Learning) [35] generates independent unimodal labels by self-supervised learning strategy and guides multiple subtasks to focus on samples with large difference between modality supervisions.

D. Evaluation Metrics

To completely demonstrate the superiority of the proposed method, we conduct experiments compared with the above baselines on two kinds of tasks, including classification tasks with seven-class accuracy (Acc7), binary accuracy (Acc2), and F1-Score (F1) as metrics and regression tasks with Mean Absolute Error (MAE) and Pearson correlation (Corr) as metrics. The computation of evaluation metrics is consistent with [33].

Specifically, Acc7 is used in seven classes classification ranging from -3 to +3 while Acc2 is calculated by negative/positive ($< 0/> 0$) classes for the sentiment scores following [44]. F1 is interpreted as the weighted average of precision and recall, which are formulated as:

$$F1 = 2\sum_{i=1}^{C} \frac{precision_i \cdot recall_i}{precision_i + recall_i} \tag{21}$$

where $precision_i = \frac{TP}{TP+FP}$, $recall_i = \frac{TP}{TP+FN}$ and i denotes category i in the classification task with C categories. TP, FP, and FN denote the number of true positive samples, false positive samples, and false negative samples, respectively.

In the regression task, MAE and Corr are calculated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y^{(i)} - \hat{y}^{(i)}| \tag{22}$$

$$\begin{aligned} Corr &= \frac{cov(y^{(i)}, \hat{y}^{(i)})}{\sigma_{y^{(i)}}\sigma_{\hat{y}^{(i)}}} \\ &= \frac{\sum_{i=1}^{N}(y^{(i)} - \overline{y^{(i)}})(\hat{y}^{(i)} - \overline{\hat{y}^{(i)}})}{\sqrt{\sum_{i=1}^{N}(y^{(i)} - \overline{y^{(i)}})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}^{(i)} - \overline{\hat{y}^{(i)}})^2}} \end{aligned} \tag{23}$$

where N denotes the total numbers of sentiment labels, $y^{(i)}$ and $\hat{y}^{(i)}$ denotes ground truth and predicted sentiment scores, respectively. Except for MAE, higher metrics values represent better performance.

E. Experiment Details

The proposed model is trained on a single GTX 1080Ti GPU. For optimization, we apply AdamW [67] as the optimizer with a linear warmup learning rate scheduler where the highest learning rate is 2e-5. The batch sizes of train, validation, and test set on both datasets are set as 48, 128 and 128. The number of training epochs is set to 40 for CMU-MOSI and 20 for CMU-MOSEI. We perform a grid-search on the validation set to find the best hyper-parameters. The best epoch is considered the one when the total task loss reaches the minimum on the validation set. Moreover, we run the proposed model five times following [68] under the same setting of hyper-parameters and present the average performance as the final test results with the relative standard deviation $< 0.7\%$. In Tables I and II, the best results are marked in bold with blue color on BERT while red color on XLNet, and † means the corresponding result is significantly better than the state-of-the-art with p-value < 0.05 based on paired t-test. Specifically, for the results of the proposed model conducted on BERT, the state-of-the-art method is Self-MM [35], while for the ones based on XLNet, the state-of-the-art method is MAG-XLNet [33].

F. Results and Analysis

1) Results on CMU-MOSI Dataset: We compare the proposed method with other baselines on CMU-MOSI dataset as shown in Table I. We can observe that Self-MM [35] based on BERT [26] and MAG-XLNet [33] based on XLNet [27] outperform other baseline methods in general. Compared with these two state-of-the-art methods, the proposed method achieves the best results on most metrics, respectively. Specifically, based on BERT, the proposed method outperforms Self-MM by 1.1% on Acc2 and 1% on F1. On MAE and Corr, the proposed method respectively achieves 0.748 and 0.782 which are competitive with Self-MM. Based on XLNet, the proposed method outperforms MAG-XLNet by 1.7% on Acc7, 1.9% on Acc2, 1.8% on F1, -0.044 on MAE and 0.012 on Corr. The result shows that the proposed method is superior to the baselines on CMU-MOSI dataset.

2) Results on CMU-MOSEI Dataset: We conduct experiments on CMU-MOSEI dataset to further compare the proposed method and the baselines. As shown in Table II, based on BERT, the proposed method outperforms Self-MM by 0.6% on Acc2, 0.5% on F1, and 0.022 on Corr. While based on XLNet, the proposed method outperforms MAG-XLNet by 1.3% on Acc7, 0.5% on Acc2, 0.4% on F1, -0.012 on MAE and 0.005 on Corr. Specially, the proposed method based on BERT attains 52.8% on Acc7 which is close to the state-of-the-art performance. Although the improvement ranges on CMU-MOSEI are more difficult than the ones on CMU-MOSI due to the larger scale of CMU-MOSEI, the proposed method still outperforms the baselines on most metrics. These results demonstrate the superiority of the proposed method, indicating the effectiveness of the multi-stages fusion framework and the hybrid-modal attention mechanism.

3) Ablation Study on Different Components: To better show the contribution of each component in the proposed framework, we perform ablation study on CMU-MOSI and CMU-MOSEI datasets by gradually adding different components in the proposed method as shown in Table III. Compared with the result of the model with no proposed modules, the performance of the model with the proposed modules on different metrics is improved in varying degrees, which is further discussed in the following.

Firstly, to explore the effectiveness of MSG module on multiple modalities, we use MSG on different modalities and combine the outputs in various ways. It is shown that the results of fine-tuning textual, acoustic and visual modalities meanwhile outperform the ones which only fine-tune a few of them. Especially compared with using MSG-T only, using MSG-T & A&V achieves 2.1%, 0.6% and 0.6% improvement on Acc7, Acc2, and F1 on MOSI while 0.6% and 0.5% improvement on Acc2 and F1 on MOSEI, respectively. A conclusion can be reached that utilizing MSG on three modalities and concatenating their outputs can maximize the information contained in each modality.

Comparing the result of MSG-T & A with MSG-T & A&V, we can see that the performance on MAE and Corr drops a little after introducing visual modalities. The same situation exists in the results on MAE and Corr of only MSG-T and MSG-T & A. We suppose the main reason is that there is noise information contained in acoustic and visual modalities which may be introduced by mistake due to the inconsistency of parameter space of different modalities. To reduce the impact of this problem, we further introduce MBA module with MSG-T & A&V as shown

TABLE III
ABLATION STUDY ON DIFFERENT COMPONENTS FOR THE PROPOSED METHOD BASED ON BERT ON CMU-MOSI AND CMU-MOSEI DATASET

Component	CMU-MOSI					CMU-MOSEI				
	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
no proposed modules	40.3	82.5	82.6	0.810	0.766	49.3	83.4	83.4	0.632	0.771
only MSG-T	42.0	84.3	84.2	0.772	0.779	50.4	84.3	84.4	0.614	0.777
MSG-T&A	44.5	83.8	83.9	0.778	0.775	51.7	84.6	84.7	0.622	0.775
MSG-T&V	43.1	84.0	84.0	0.790	0.764	49.5	83.8	83.9	0.623	0.774
MSG-T&A&V	44.1	84.9	84.8	0.779	0.776	50.9	84.9	84.9	0.617	0.775
MSG-T&A&V + MBA	45.1	85.2	85.2	0.752	0.780	51.9	85.2	85.0	0.593	0.779
Proposed (BERT)	**45.3**	**85.7**	**85.6**	**0.748**	**0.782**	**52.8**	**85.4**	**85.4**	**0.583**	**0.787**

The best results are highlighted in bold.

TABLE IV
ABLATION STUDY ON DIFFERENT MODALITIES FOR THE PROPOSED METHOD BASED ON BERT ON CMU-MOSI DATASET

Modality	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
only A	16.1	53.9	54.2	1.477	0.049
only V	15.8	54.1	54.3	1.492	0.034
only T	41.8	84.3	84.3	0.782	0.774
only A+V	16.1	54.7	54.9	1.483	0.073
only T+A	43.2	85.3	85.3	0.764	0.778
only T+V	44.7	85.2	85.2	**0.740**	**0.782**
Proposed (BERT)	**45.3**	**85.7**	**85.6**	0.748	**0.782**

The best results are highlighted in bold.

in Table III. The result of MSG-T & A&V with MBA outperforms the one without MBA by 1.0% on Acc7, 0.3% on Acc2, 0.4% on F1, −0.027 on MAE and 0.004 on Corr on MOSI while 1.0% on Acc7, 0.3% on Acc2, 0.1% on F1, −0.024 on MAE and 0.004 on Corr on MOSEI, which validate the effectiveness of MBA module on textual modality.

Lastly, to verify the necessity of conducting the Shifting Huberloss, we compare the results from the last two rows in Table III whose difference is introducing the Shifting Huberloss or not. It can be observed that using Shifting Huberloss can improve performance on all metrics, especially reduce MAE value which is most affected by outliers caused by data noise. The result indicates that Shifting Huberloss can take shifting values into consideration and further enhance the robustness to noise information contained in modalities.

4) Ablation Study on Different Modalities: To evaluate the effectiveness of capturing inter-modality information across modalities, we conduct an ablation study on different combinations of input modalities for the proposed method. As shown in Table IV, the modality-specific features extracted in single modality reach optimal performance, especially for textual modality as the dominant modality. However, introducing interaction with another modality by the proposed MSG and MBA brings better performance comparing the bimodal circumstance with the unimodal one, indicating the importance of cross-modality features. The ablation experiment results further show that MSG and MBA are productive in exploring inter-modality dynamics among various modalities.

5) Results for MSG-T Inserted into Different Encoder Layers of Transformer Models: Since various granularity features are learned from various encoder layers of the Transformer model

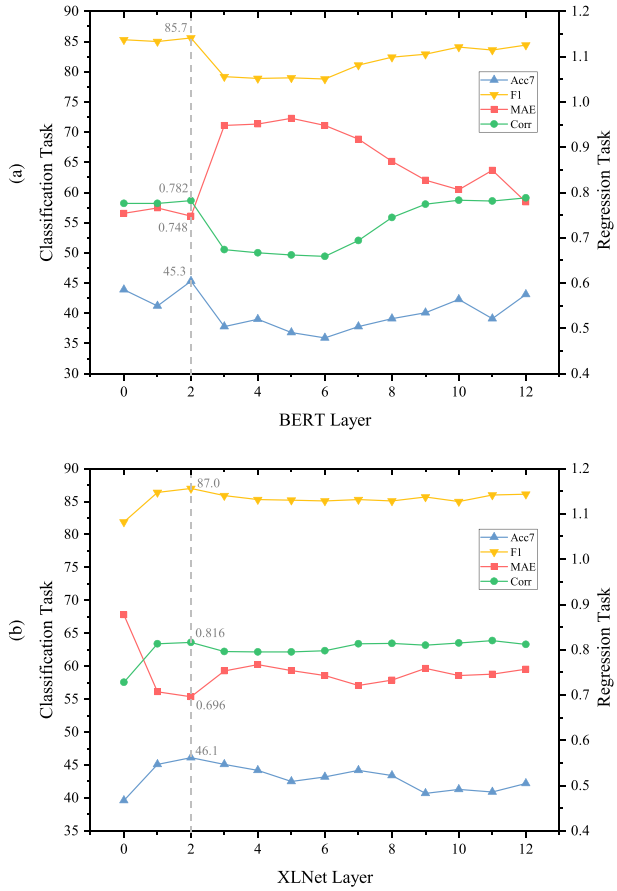
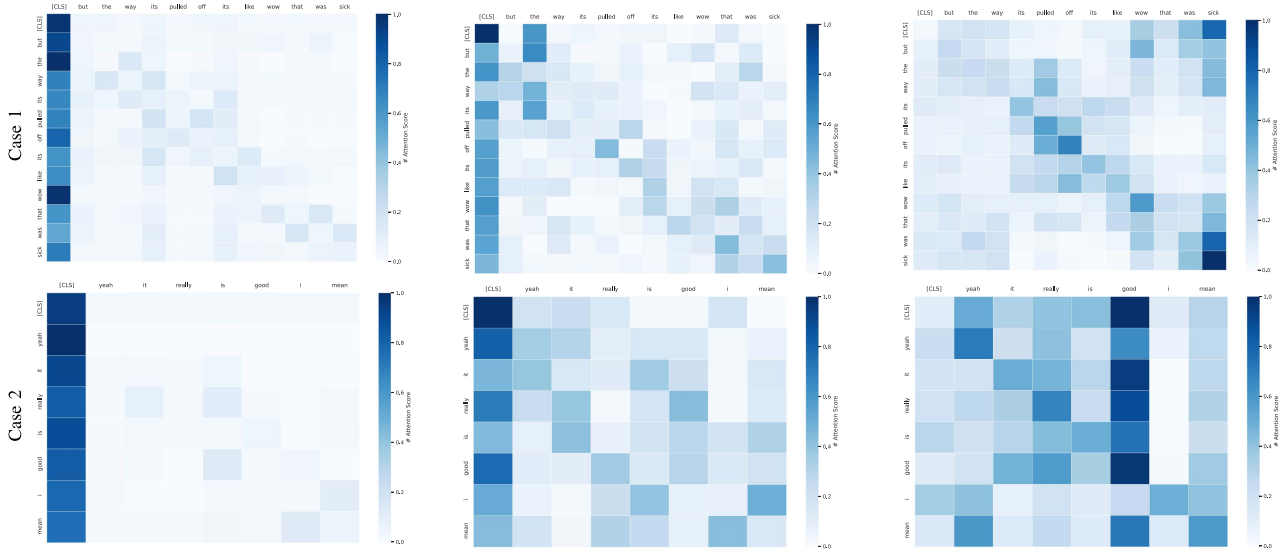


Fig. 5. Experiment results for MSG-T Inserted into Different Layers of the (a) BERT and (b) XLNet pre-trained model.

[33], [49], [50], we conduct experiments to insert MSG-T into different layers of the Transformer model. Specifically, we apply MSG-T at the embedding layer and at different encoder layers of BERT and XLNet to verify whether the proposed framework has captured the most relevant intra-modality dynamics inside textual modality with the other modalities.

As shown in Fig. 5(a) and (b), inserting MSG-T into the second layer of both BERT and XLNet achieves the best performance compared to the other layers. Firstly, the outputs of the first layer of BERT and XLNet are low-level feature information which is not enough for the latter hybrid attention with the other modalities. Next, the higher layers of BERT and XLNet extract more abstract linguistic features about the syntactic and



(a) 2^{nd}-layer of BERT (b) 3^{rd}-layer of BERT (c) 12^{th}-layer of BERT

Fig. 6. Visualization of the self-attention matrices in different encoder layers of BERT in the proposed method, including the (a) second, (b) third, and (c) last layer. The subfigures above contain two examples from CMU-MOSI dataset.

semantic structure. Besides, the acoustic and visual representations in the proposed method are corresponding to each word in the utterance which make shifting the textual representation on the later layers of BERT and XLNet is more difficult due to the high-level abstract information. After experimenting, we observe that the second layers of both BERT and XLNet are most suitable for the application of MSG-T.

Furthermore, the variation of results on different layers of BERT is larger than on the ones of XLNet and the drop of performance is more significant at the middle layers of BERT. The result indicates that un-directional context information in each token learned by XLNet captures sufficient intra-modality dynamics which can be utilized for better integration of the multimodal representations in the proposed framework.

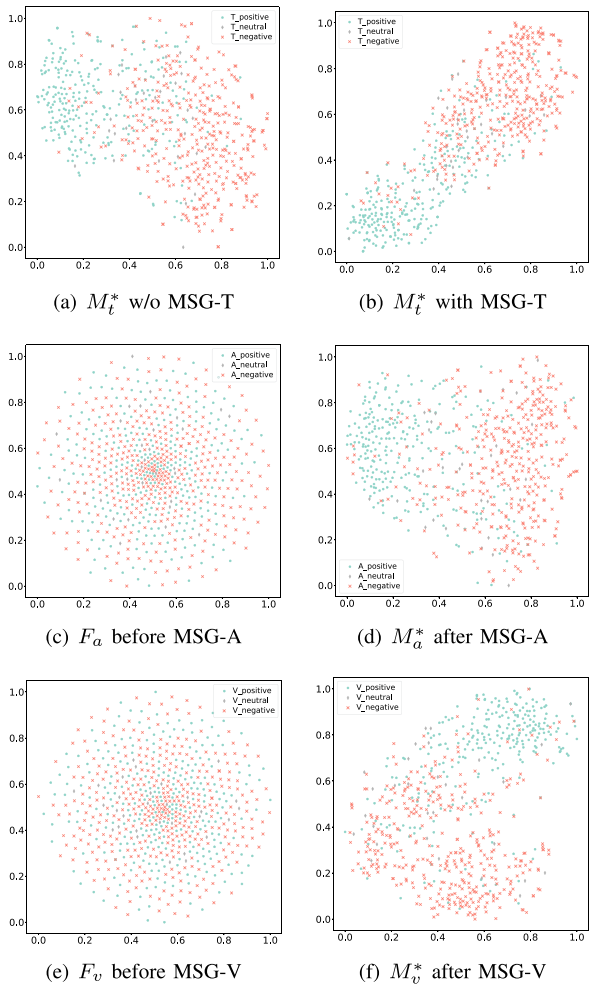
G. Further Discussions

For diving into our model to explore its performance in the MSA task, we analyze and discuss more experiments of the proposed model in the following.

1) Visualization of Attention Matrices in Different Encoder Layers of Transformer Models: As shown in Fig. 6, we visualize the self-attention matrices from the second, third, and last encoder layers of BERT to illustrate the syntactic and semantic features extracted by different layers of the Transformer model. Specifically, in Fig. 6(a) and (b), [CLS] token as the query in the attention mechanism reaches larger weights with all other tokens in the utterance, meaning that front Transformer layers concentrate on the textual features from every token. Besides, the diagonal of the attention matrices shows that the second higher values of attention weights for most tokens are the neighboring words, indicating that the n-gram phrase-level features are extracted as part of the syntactic features in the front layers. In addition, the visualization implicitly explains the reason for the superior performance of the MSG module inserted between the second and third Transformer layers where appear the alternation of token-level and phrase-level features suitable for the shifting interaction among different modalities. Moreover, Fig. 6(c) as the attention matrices in the last layer indicates that the Transformer model finally captures the semantic features for each utterance due to the largest attention weights assigned to the most meaningful and sentiment-related words. For example, "wow that was sick" in Case 1 and "yeah it really is good" in Case 2 respectively obtain the highest values of attention weights, denoting that the model has understood the semantic meaning of the sentence and learned to reason the sentiment according to corresponding sentiment-related words. The visualization results of various granularity features from diverse Transformer layers remain consistent with previous work [33], [49], [50]

2) Visualization of Representations with MSG: To illustrate the influence of MSG on different unimodal representations, we visualize the representations with and without MSG for textual, acoustic, and visual modalities in the embedding space. For textual modality in Fig. 7(a) and (b), the output textual representations M_t^* of the last BERT layer with inserted MSG-T are more compact and form more discriminative clusters for various sentiment classes compared to the one without MSG-T, showing the effectiveness of shifting in the fine-tuning of textual modality through the other modalities. As for acoustic and visual modalities, Fig. 7(c) and (e) denote the low-level modality-specific features F_a and F_v extracted by RNN, far from being utilized to predict sentiment. Nevertheless, after shifting fine-tuned in MSG-A and MSG-V, M_a^* and M_v^* contain more semantics information brought by textual modality and achieve better representation ability for utterance with different sentiments.

3) Visualization of Attention Weight p_{att} in MBA: The core of MBA is the modality-specific attention weight p_{att} learned by acoustic and visual modalities interacting with textual modality.



(a) M_t^* w/o MSG-T (b) M_t^* with MSG-T

(c) F_a before MSG-A (d) M_a^* after MSG-A

(e) F_v before MSG-V (f) M_v^* after MSG-V

Fig. 7. T-SNE [69] visualization of textual, acoustic and visual representations with and without MSG based on BERT in the embedding space on the testing set of CMU-MOSI, where "green dot", "gray damond" and "red x" denotes positive, neutral and negative sentiment respectively.

We visualize the attention weight p_{att} for acoustic and visual modalities with different sentiments in Fig. 8. For example, satisfied and emphasized tone saying "really really love" in acoustic modality causes the textual modality tends to raise positive sentiment intensity on the corresponding tokens and filter the irrelevant high-level features. Similarly, a squeezed and frowned face in visual modality makes textual modality focus more on the negative expression of "didn't really". The visualization of both acoustic- and visual-specific attention weight p_{att} demonstrate that MBA is capable of effectively reducing the modality gap and learning robust textual representations based on other modalities.

4) Function of Shifting HuberLoss: The Shifting HuberLoss is constructed to avoid the interference of the noise inside acoustic and visual modalities and enhance the robustness of measuring the shifting value, whose main contribution to the proposed model is to stabilize the training process of MSG. In (7), the magnitude of the shifted feature h_s has been controlled by the adaptive scaling factor η for each utterance sample, which can be seen as a monitor at the sample-level. However, at the batch-level, there is no monitor to control the shifting value which may cause the shifting value from different batches to vary largely and make the process of training and the parameters of the model unstable. The Shifting HuberLoss is presented for each batch from a global perspective to control the shifted feature not fluctuate too excessively and makes the training process and the parameters more stable. To further verify this conclusion, we conduct experiments on the proposed model with and without Shifting HuberLoss on CMU-MOSI based on BERT and visualize the performance variation as training proceeds. As shown in Fig. 9, the performance improvement of the proposed model with Shifting HuberLoss is more stable and smoother than the one without Shifting HuberLoss.

TABLE V
ABLATION STUDY ON DIFFERENT COMPONENTS FOR THE PROPOSED METHOD BASED ON XLNET ON CMU-MOSI DATASET

Component	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
no proposed modules	41.6	84.0	84.1	0.787	0.790
MSG-T&A&V	43.8	85.2	85.2	0.732	0.798
MSG-T&A&V + MBA	45.4	86.3	86.2	0.703	0.813
Proposed (XLNet)	**46.1**	**87.0**	**87.0**	**0.696**	**0.816**

The best results are highlighted in bold.

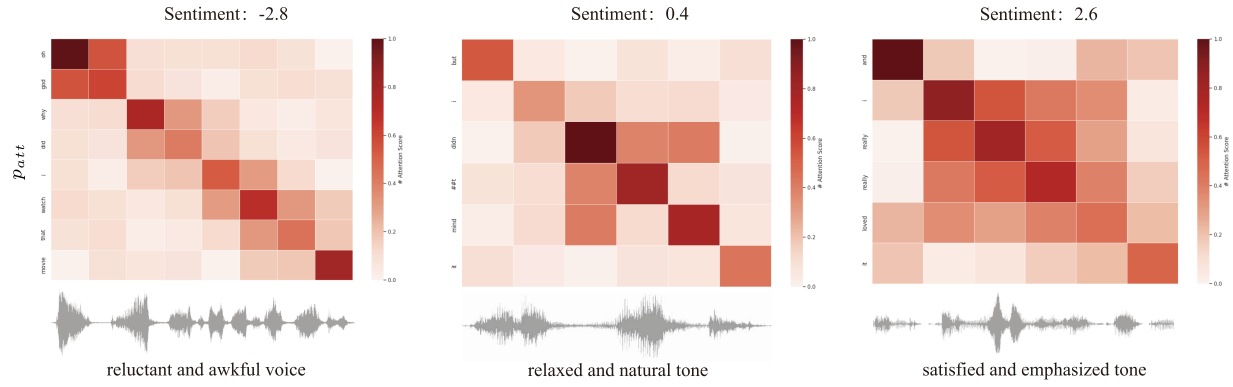
TABLE VI
PERFORMANCE COMPARISON FOR THE PROPOSED METHOD WITH DIFFERENT FEATURE EXTRACTION MODELS FOR THE MODALITIES ON CMU-MOSI DATASET

T	A&V	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
BERT	sLSTM	**45.3**	**85.7**	**85.6**	0.748	0.782
	bLSTM	43.8	85.3	85.4	0.758	0.773
	sGRU	44.1	85.2	85.3	**0.747**	0.780
	bGRU	43.9	85.2	85.2	0.757	**0.783**
XLNet	sLSTM	**46.1**	**87.0**	**87.0**	**0.696**	**0.816**
	bLSTM	45.3	86.5	86.5	0.725	0.802
	sGRU	45.1	86.8	86.8	0.719	0.803
	bGRU	45.7	86.5	86.5	0.717	0.806

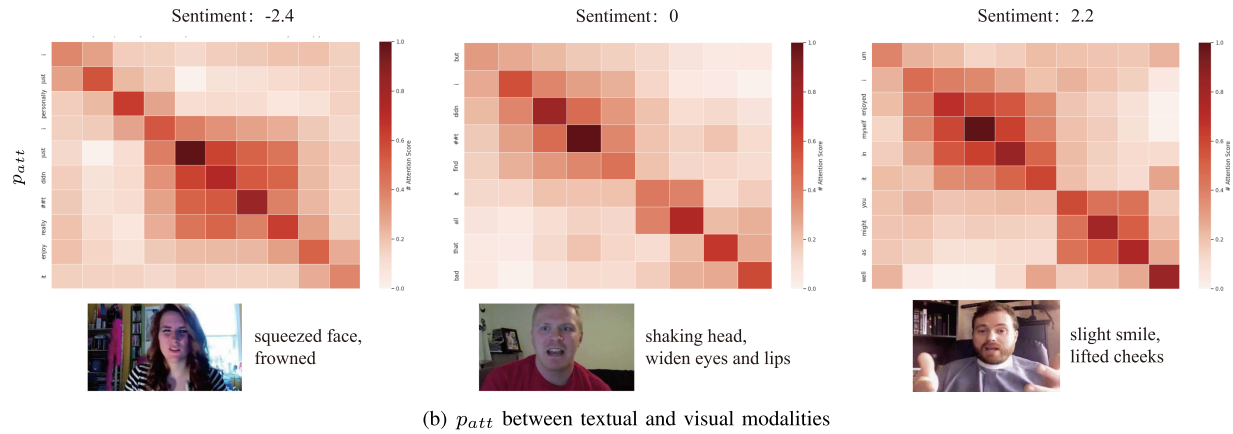
The best results are highlighted in bold.

5) Ablation Study on XLNet: To further evaluate the effectiveness of our model on XLNet, we have conducted an ablation study on CMU-MOSI dataset using XLNet as the textual representation encoder. As shown in Table V, the performance of the model with the proposed modules on different metrics is improved in varying degrees. The experiment results demonstrate the effectiveness of MSG/MBA and the Shifting HuberLoss on XLNet.

6) Results with Different Models on Feature Extraction for Three Modalities: To further verify the generalization of the proposed framework, we utilize different Transformer models including BERT and XLNet to extract textual representation and different RNN variants to extract acoustic and visual representations. As shown in Table VI, RNN variant networks contain Long Short-Term Memory [20] with single direction (sLSTM) and bi-direction (bLSTM) and Gated Recurrent Unit [24] with single direction (sGRU) and bi-direction (bGRU) [58]. The result of using BERT and XLNet with different RNN variants mostly achieves similar performance with each other, which



(a) p_{att} between textual and acoustic modalities



(b) p_{att} between textual and visual modalities

Fig. 8. Visualization of attention weight matrices p_{att} in MBA for both textual-acoustic and textual-visual modalities. The subfigures above contain six examples with positive, neutral and negative sentiments respectively from CMU-MOSI dataset.

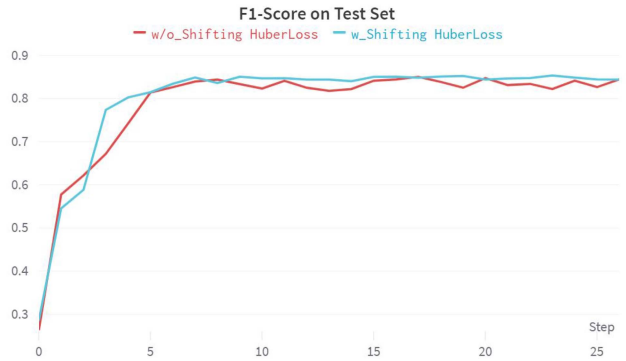


Fig. 9. Visualization of F1-Score variation as training proceeds with/without Shifting HuberLoss on CMU-MOSI based on BERT.

means that the proposed framework can remain stable utilizing different representation extraction methods. However, there are several differences in the performance of different models. When using BERT to extract textual representation, the best performance on different metrics is achieved by different RNN variants. For Acc7, Acc2, and F1, using BERT with sLSTM achieves the highest value $45.3\%, 85.7\%, 85.6\%$ which is better than the other models. While for MAE and Corr, using BERT with sGRU and with bGRU achieve the best performance 0.747 and 0.783. We suppose the reason is that textual representation learned by BERT is more sensitive to the feature extraction direction of sentences due to the characteristic of the auto-encoder method.

Compared with BERT, XLNet learns un-directional context for each token and captures more long-range dependencies in the textual representation. As shown in Table VI, the proposed method using XLNet achieves more stable and better performance on all metrics. Especially, using XLNet with sLSTM achieves the best performance among all models, which are 46.1%, 87.0% and 87.0% on Acc7, Acc2, and F1 while 0.696 and 0.815 on MAE and Corr, respectively.

Although there are slight differences when using BERT or XLNet different RNN variants to extract features of multiple modalities, the proposed method with all variants still achieves state-of-the-art performance compared to the baselines. The results further demonstrate the effectiveness and generalization of the proposed framework.

V. Conclusion

In this article, we propose a novel multi-stage fusion framework to dynamically shift representations for three modalities via the hybrid-modal attention mechanism. Firstly, we design a module called Multimodal Shifting Gate (MSG) to fine-tune the

representations by learning inter-modality dynamics. Then, we adopt a module named Masked Bimodal Adjustment (MBA) on textual modality to reduce the influence of inconsistent parameter space and reduce the modality gap. Besides, we leverage the syntactic-level and semantic-level textual representations of the Transformer model to capture sufficient intra-modality dynamics. Furthermore, we construct a Shifting HuberLoss to improve the robustness of measuring the shifting value. Lastly, we conduct extensive experiments based on different RNN variants and pre-trained Transformer models to demonstrate the efficacy of the proposed model.

REFERENCES

[1] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, Apr. 2019.

[2] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, pp. 2008–2020, 2015.

[3] Q. Fang, C. Xu, J. Sang, M. S. Hossain, and G. Muhammad, "Word-of-mouth understanding: Entity-centric multimodal aspect-opinion mining in social media," *IEEE Trans. Multimedia*, vol. 17, pp. 2281–2296, 2015.

[4] A. Zadeh, P. P. Liang, L.-P. Morency, S. Poria, E. Cambria, and S. Scherer, "Proceedings of grand challenge and workshop on human multimodal language (challenge-HML)," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, Melbourne, Australia, 2018.

[5] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132, Jan.–Mar. 2023.

[6] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. 2015 Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2539–2544.

[7] 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.

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

[9] A. Zadeh et al., "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.

[10] X. Guo, W.-K. A. Kong, and A. C. Kot, "Deep multimodal sequence fusion by regularized expressive representation distillation," *IEEE Trans. Multimedia*, early access, Jan. 13, 2022, doi: 10.1109/TMM.2022.3142448.

[11] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[12] P. P. Liang, Z. Liu, A. Bagher Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 150–161. [Online]. Available: https://aclanthology.org/D18-1014

[13] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rygqqsA9KX

[14] O. Abdel-Hamid et al., "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[16] S. Mai, H. Hu, and S. Xing, "A unimodal representation learning and recurrent decomposition fusion structure for utterance-level multimodal embedding learning," *IEEE Trans. Multimedia*, vol. 24, pp. 2488–2501, 2022.

[17] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 112–118.

[18] Q.-T. Truong and H. W. Lauw, "VistaNet: Visual aspect attention network for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 305–312.

[19] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] S. Hochreiter et al., "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, S. C. Kremer and J. F. Kolen, Eds., IEEE Press, 2001.

[22] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[23] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in*Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[24] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, 2014.

[25] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[27] Z. Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.

[28] Z. Dai et al., "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2978–2988. [Online]. Available: https://aclanthology.org/P19-1285

[29] T. B. Brown et al., "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[30] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[31] X. Han et al., "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[32] K. Yang, H. Xu, and K. Gao, "CM-BERT: Cross-modal BERT for text-audio sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 521–528.

[33] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2020, pp. 2359–2369.

[34] 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.

[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.*, 2021, pp. 10790–10797.

[36] W. Han et al., "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interaction*, 2021, pp. 6–15.

[37] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[38] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 949–954.

[39] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[40] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.

[41] S. Poria et al., "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[42] A. Zadeh et al., "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[43] H. Pham, T. Manzini, P. P. Liang, and B. Poczós, "Seq2Seq2Sentiment: Multimodal sequence to sequence models for sentiment analysis," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, 2018, pp. 53–63.
[44] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2019, pp. 6558–6569.
[45] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 4730–4738.
[46] Y. Wang et al., "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.
[47] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, *arXiv:1606.06259*.
[48] 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.
[49] Z. J. Wang, R. Turko, and D. H. Chau, "Dodrio: Exploring transformer models with interactive visualization," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.: Syst. Demonstrations*, 2021, pp. 132–141. [Online]. Available: https://aclanthology.org/2021.acl-demo.16
[50] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3651–3657.
[51] 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.
[52] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu, "Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism," in *Proc. 1st Int. Multimodal Sentiment Anal. Real-Life Media Challenge Workshop*, 2020, pp. 27–34.
[53] L. Sun et al., "Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model," in *Proc. 2nd Multimodal Sentiment Anal. Challenge*, 2021, pp. 15–20. [Online]. Available: https://doi.org/10.1145/3475957.3484456
[54] L. Stappen et al., "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *Proc. 1st Int. Multimodal Sentiment Anal. Real-life Media Challenge Workshop*, 2020, pp. 35–44.
[55] L. Stappen et al., "The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," in *Proc. 2nd Int. Multimodal Sentiment Anal. Challenge Workshop*, 2021, pp. 5–14.
[56] D. Olson, "From utterance to text: The bias of language in speech and writing," *Harvard Educ. Rev.*, vol. 47, no. 3, pp. 257–281, 1977.
[57] L.-C. Yu et al., "Building chinese affective resources in valence-arousal dimensions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2016, pp. 540–545.
[58] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[59] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling, in *Proc. Int. Conf. Learn. Representations*, 2017.
[60] Y. Gu et al., "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2018, pp. 2225–2235.
[61] 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.
[62] iMotions, "Facial expression analysis," 2017. [Online]. Available: https://imotions.com/
[63] M. Chen et al., "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interaction*, 2017, pp. 163–171.
[64] J. Yuan et al., "Speaker identification on the scotus corpus," *J. Acoustical Soc. Amer.*, vol. 123, no. 5, 2008, Art. no. 3878.
[65] 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.*, 2019, pp. 6892–6899.
[66] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, "Quantum-inspired multimodal fusion for video sentiment analysis," *Inf. Fusion*, vol. 65, pp. 58–71, 2021.
[67] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7
[68] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song, "What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis," *Inf. Fusion*, vol. 66, pp. 184–197, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253520303675
[69] L. van der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.