

Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis

Yifeng Wang, Jiahao He, Di Wang^{*}, Quan Wang, Bo Wan, Xuemei Luo

Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, School of Computer Science and Technology, Xidian University, Xi'an, 710071, China

ARTICLE INFO

Communicated by D. Cavaliere

Keywords:

Multimodal sentiment analysis
Transformer
Multimodal fusion

ABSTRACT

Multimodal Sentiment Analysis (MSA) constitutes a pivotal technology in the realm of multimedia research. The efficacy of MSA models largely hinges on the quality of multimodal fusion. Notably, when conveying information pertinent to specific tasks or applications, not all modalities hold equal importance. Previous research, however, has either disregarded the importance of modalities altogether or solely focused on the importance of linguistic and non-linguistic modalities while neglecting the importance between non-linguistic modalities. To facilitate effective multimodal information fusion based on the relative importance of modalities, a novel multimodal fusion mode named Multimodal Transformer with Adaptive Modality Weighting (MTAMW) is proposed in this paper. Specifically, we introduce a multimodal adaptive weight matrix that allocates appropriate weights to each modality based on its contribution to sentiment analysis. Furthermore, a multimodal attention mechanism is introduced, utilizing multiple Softmax functions to compute attention weights, thereby efficiently fusion multimodal information via a single-stream Transformer. By meticulously considering the relative importance of each modality during the fusion process, more effective multimodal information fusion is achievable. Extensive experiments on benchmark datasets show that it is superior to or comparable to state-of-the-art methods on MSA tasks. The codes for our experiments are available at <https://github.com/Vamos66/MTAMW>.

1. Introduction

In recent years, with the unprecedented development of the mobile Internet, we have witnessed an explosive growth of multimodal data. Analyzing multimodal data helps us to understand the human world and benefits various applications such as smart retail [1], security monitoring [2], and intelligent question-answering [3]. Multimodal sentiment analysis (MSA) is a popular topic in the field of multimodal analysis, which aims to infer people's sentiments from multimodal information. In this paper, we mainly focus on identifying human sentiments in videos. There are three kinds of multimodal information in videos: linguistic(text), acoustic (sound), and visual (image) information.

Multimodal information fusion is the process of combining data from multiple sources or modalities to improve the accuracy, robustness, and overall quality of the resulting information [4]. Different modalities have unique characteristics [5], and the goal of MSA is to integrate multiple unimodal information into a unified and interrelated multimodal representation. Notably, multimodal fusion methods can be broadly classified into two principal categories: those adopting

equal weighting strategies and those emphasizing the significance of linguistic modality. These former methods [6–9] usually use a ternary symmetric architecture for multimodal fusion, with each modality having equal importance. The latter category of methods [10–12] typically captures information related to linguistic modality from non-linguistic (visual and acoustic) modalities and then fuses it with the linguistic model which will highlight the importance of linguistic modality in fusion. While existing models have achieved promising performance in MSA tasks, as several studies [7,8,13] have pointed out, the linguistic, acoustic, and visual modalities contribute differently to sentiment analysis. If the relative importance of the three modalities is not taken into account when fusing multimodal information, the performance of MSA models may degrade. Therefore, it is crucial to consider the importance of each modality in multimodal fusion to achieve optimal performance in MSA tasks.

Transformer [14] has been a tremendous success in various artificial intelligence domains, including natural language processing, computer vision, audio processing, etc [15]. In MSA tasks, the Transformer-Encoder and the Transformer-Decoder play essential roles in integrating

^{*} Corresponding author.

E-mail address: wangdi@xidian.edu.cn (D. Wang).

<https://doi.org/10.1016/j.neucom.2023.127181>

Received 14 April 2023; Received in revised form 1 October 2023; Accepted 26 December 2023

Available online 29 December 2023

0925-2312/© 2023 Elsevier B.V. All rights reserved.

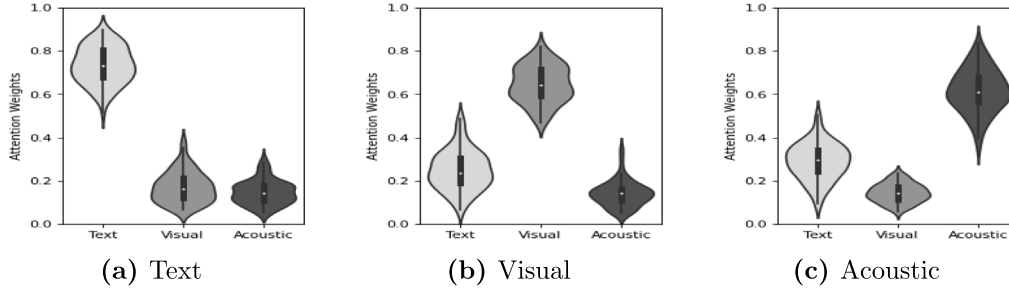


Fig. 1. Distributions of the attention weights of the single-stream Transformer model: (a), (b), and (c) represent the attention weight distributions from multimodal to text, visual, and acoustic modalities, respectively.

multimodal information. The Transformer-Encoder is used to fuse features representing the context of each modality. Such as, The Modality Invariant and Specific Representations (MISA) [8] uses Transformer-Encoder to fuse coherent and differential features of multiple modalities. However, this approach may not fully leverage the Transformer's capability to capture the long-range dependencies of multimodal features. On the other hand, the Transformer-Decoder is utilized to fuse sequential features from pairs of modalities. For example, the Multimodal Transformer (MulT) [7] uses six cross-modal Transformers to fuse modalities pairs (text–visual, text–acoustic, visual–text, visual–acoustic, acoustic–visual, acoustic–visual). It can well capture the long-range dependencies between features across modalities. However, as Multi utilizes six Transformers, it has a huge number of parameters and suffers from high computational burden. If linguistic, visual, and acoustic sequences are concatenated and fused by a single-stream Transformer-Encoder, the model parameters will be significantly reduced and the long-range dependencies between features of different modalities can be captured. When computing self-attention, the attention mechanism learns to weigh the importance of each feature vector in a sequence relative to other feature vectors in the same sequence [14]. In a multimodal environment where feature vectors belong to different modalities, the attention mechanism also learns to weigh the importance of each modality with respect to the other modalities. However, the feature vectors from different modalities are projected into separate spaces [16], making them less comparable and harder for the attention mechanism to weigh. This results in lower intermodal attention values compared to intramodal ones. Conversely, feature vectors of the same modality are more comparable, leading to higher intramodal attention values. Fig. 1's visualization of attention weight distribution confirms this. When using a single-stream Transformer encoder, most attention weights are allocated within the same modality. This biases the fusion model towards intra-modal information, undermining effective multimodal integration.

To address the aforementioned issue, we propose a novel multimodal fusion model named multimodal Transformer with adaptive modality weighting (MTAMW) in this paper. Within the MTAMW model, text, image, and sound sequences are concatenated and fed into a single-stream Transformer. A multimodal adaptive weight matrix is introduced to adaptively learn the importance of different modalities by performing element-wise product with an attentional weight matrix in the Transformer. In addition, the weight of each modality is different in different layers of Transformer, which adaptively adjusts according to attention value of each modality in each layer. To alleviate the problem that the intra-modality attention values are significantly larger than inter-modality attention values in calculating the self-attention of the concatenated sequences, we use multiple Softmax functions to calculate attention weights for each modality separately.

Our contributions can be summarized as follows:

- We introduce a novel multimodal adaptive weight matrix that assigns appropriate weights to each modality during the fusion process. This innovative approach enables us to conduct a more precise sentiment analysis by considering the unique contributions of each modality.

- A multimodal attention mechanism is tailored for fusing multimodal information, which successfully addresses the problem of over-focusing on intra-modality attention in calculating the self-attention of the concatenated sequences by using multiple Softmax functions to compute attention weights.
- Our contributions are validated through extensive experiments conducted on various benchmark datasets. The results of these experiments demonstrate the superior performance of our proposed method in comparison to existing state-of-the-art models.

2. Related work

In this section, we briefly overview some related work in multimodal sentiment analysis, and then introduce the attention mechanism in Transformer.

2.1. Multimodal sentiment analysis (MSA)

MSA requires the effective fusion of text, visual, and acoustic information, and finally uses the fusion of multimodal information to predict sentiment intensity [17]. The importance of the text modality to multimodal information was not overemphasized prior to the development of text pre-training models like BERT [18] and XL-Net [19]. Instead, these various modalities were handled equally. At that stage, multiple methods [20,21] were done by first using CNN [22], Recursive Neural Networks (RNN), or Long-Short Term Memory (LSTM) [23] to obtain temporal and spatial information for each modality, then concatenation, summing, or tensor product of the spatio-temporal information of the different modalities, and finally analyzing the sentiment. C-MKL [20] uses CNN to extract multimodal information, followed by multiple kernel learning for sentiment analysis. Later works have focused on multimodal fusion, and the Tensor fusion network (TFN) [6] has been developed by modeling intra-modality and inter-modality information so that information on the interactions between three modalities, unimodal, bimodal, and trimodal, can be captured. TFN has achieved good results in MSA, but when the number of modalities increases, there is an exponential increase in computational effort. To address the disadvantages of TFN in terms of poor computational efficiency, Low-rank Multimodal Fusion (LMF) [24] conducts multimodal fusion by decomposing the tensor and weights in parallel and using modality-specific low-rank factors. MulT [7] uses six cross-modal Transformer-Encoders to model all pairs of modalities, fuses multimodal information by focusing directly on low-level features in other modalities, subsequently uses the Self-attention mechanism to fuse features, and finally conducts sentiment analysis.

When text-based pre-training models like BERT [18], XL-Net [19], and GPT [25] were used in many downstream tasks of natural language processing, people started to introduce these pre-training models into MSA tasks. Because BERT can only be used for linguistic modalities, MAG-BERT [10] uses the component MAG, which converts the effect of non-linguistic information on linguistic information into an offset vector, and then sums the feature vectors of linguistic modalities and the

offset vectors, which are finally fed into the pre-trained model BERT. Finally, the experimental results of MAG-BERT confirm that pre-training models can contribute significantly to multimodal sentiment analysis. Similarly, TextMI [26] presents non-linguistic information (audio and visual modalities) as text, then concatenates them with textual modalities and finally feeds them into a language pre-training model for fine-tuning. However, text-based pre-training models like BERT can only be used to directly extract information from textual modalities, but not for non-textual modalities, which makes text modalities dominant in multimodal sentiment analysis. If the non-linguistic modality is ignored, however, it does not fully utilize multimodal information. Subsequently, MISA [8] uses the modality-invariant and -specific representations to project each modality into two different subspaces, which are used to learn commonalities between the different modal representations for reducing modal gaps and to learn features specific to each modality, and finally to complement feature fusion. Self-MM [9] uses self-supervised learning to obtain labels for each modality and then joint multimodal and unimodal learning (multitask learning) to learn coherence and discrepancy, respectively. SUGRM [27] automatically generates unimodal annotations through a unimodal label generation module, then recalibrates multimodal features, and finally uses the recalibrated features to jointly train multimodal and unimodal sentiment analysis. MMIM [11], to preserve task-relevant information during multimodal fusion, is integrated with the main task (MSA) by adding maximized mutual information between unimodal input pairs (inter-modal) and between multimodal fusion results and unimodal inputs Joint training. MUTA-Net [28] uses a modal-utterance-temporal attention network with multimodal sentiment loss for learning discriminative multi-relational representations. Different from previous work, our work aims to use a single-stream Transformer to fuse multimodal information. By introducing multimodal adaptive weight matrix in the Transformer and replacing the self-attention mechanism in the Transformer with our proposed multimodal attention mechanism to obtain multimodal representation.

2.2. Attention mechanism

The attention mechanism plays a crucial role in enabling a system to focus on task-critical information [29]. In the context of natural language processing, Bahdanau attention [30] introduces a groundbreaking approach that relieves the encoder from the burden of compressing all the source sentence information into a fixed-length vector. Moreover, it empowers the decoder with an attention mechanism, which allows the model to selectively emphasize relevant information and overcome the limitations of fixed-length representations. Consequently, the decoder can selectively retrieve the necessary information as needed. Extending this concept to visual tasks, Visual attention [31] introduces a visual attention module that leverages an attention mechanism. This module enables the model to selectively concentrate on different regions of an image while generating each word in an image caption. By incorporating visual attention, the model can effectively align image features with corresponding textual descriptions, enhancing the quality of image caption generation. In a broader context, the Transformer [14] completely abandons the traditional recurrent neural network structure and effectively uses attention mechanisms to capture information that is useful to the task, where there are two types of attention mechanisms used to capture information, namely multi-head attention and multi-head self-attention. The multi-head self-attention mechanism is used to capture useful information in the intra-sequence, while the multi-head attention mechanism is used to capture useful information in the inter-sequence.

In this work, text, visual, and audio modal sequences are concatenated and then fed into a single-stream Transformer-Encoder. However, in MSA tasks, the text modality carries more abundant sentiment information compared to the visual and acoustic modalities. This difference may cause models to excessively rely on text information when

predicting sentiment intensity, neglecting visual and acoustic cues, resulting in language bias [32]. Inspired by the multi-headed attention mechanism [14], multiple Softmax functions are used to calculate attention weights for each modality separately, to avoid the fusion model focusing on the intra-modality information and ignoring the inter-modality information. Additionally, it helps alleviate the issue of language bias. Coincidentally, our method of computation is similar to Cross Attention in the Multi-Modality Cross Attention (MMCA) Network [33]. The MMCA fuses image and text information by feeding the model text and image sequences with high-level semantic information and no semantic gaps, respectively, and then uses Cross Attention to learn token-level multimodal representations, and finally uses a mean pooling layer or a 1d-CNN layer to integrate the final representation of each modality. However, when MTAMW integrates multimodal representations, there is a semantic gap [34] in the multimodal information input to the Multimodal Transformer because the text representations obtained using BERT [18] have higher-level features, while the input visual and acoustic representations are lower-level features. If token-level multimodal representations are captured by Cross Attention as in MMCA, then low-level semantic information will introduce a large amount of noise, which is not conducive to multimodal fusion. The Image-Text Consistency Driven Multimodal Sentiment Analysis [35] to solve the problem of semantic gap and language bias during information fusion, a new image-text association model is proposed to examine the relationship between images and text, based on which a multimodal sentiment analysis method is derived by combining low-level visual features and different text features as rich features. Different from previous approaches the MTAMW introduces learnable embedding [36] $CLS-V$ and $CLS-A$ at the head of visual and acoustic representations respectively to bridge the semantic gap between low-level features and high-level features by learning high-level multimodal representations at the sentence level, as well as CLS with the same capability in the text, denoted as $CLS-T$, and the rest of the token is used to capture low-level and fine-grained semantic information.

3. Method

In this section, the problem formulation and some basic notations are initially introduced. Then, details of the proposed MTAMW method are given. Finally, the optimization algorithm is introduced.

3.1. Problem formulation

Multimodal sentiment analysis aims to detect the sentiment intensity of a multimodal signal. Typically, there are three kinds of signals contained in a multimodal signal X , that are, text (t), visual (v), and acoustic (a) sequences. These sequences are denoted as $X_m = \{x_m^1, x_m^2, \dots, x_m^{L_m}\}$, $m \in \{t, v, a\}$ respectively, where L_m is the sequence length of modality m . x_t^i , x_v^i , and x_a^i denote the i th word, the i th visual frame (image), and the i th acoustic frame, respectively.

For the text sequence X_t , in order to obtain contextual information from multimodal information, we add special tokens-[CLS] and [SEP] to its head and tail, respectively, and then use a pre-trained BERT-base-uncased tokenizer to convert it to the corresponding index sequence $U_t \in \mathbb{R}^{L'_t}$, where $L'_t = L_t + 2$. For non-textual modalities X_m , $m \in \{v, a\}$, after extracting visual and acoustic modality features using Facet¹ and COVAREP [37] respectively, a learnable embedding $CLS-m$, $m \in \{V, A\}$ is added before the feature of the first frame, as ViT [36] does. The final representation of non-textual modalities $U_m \in \mathbb{R}^{L'_m \times d_m}$, $m \in \{v, a\}$ is obtained, where d_m , $m \in \{v, a\}$ is the dimension of the visual and acoustic features, and the length $L'_m = L_m + 1$, $m \in \{v, a\}$.

Given a multimodal signal $X = \{U_t, U_v, U_a\}$, the goal of the proposed MTAMW model is to predict the sentiment intensity \hat{y} of X , where $\hat{y} \in [-3, +3]$.

¹ <https://imotions.com/>

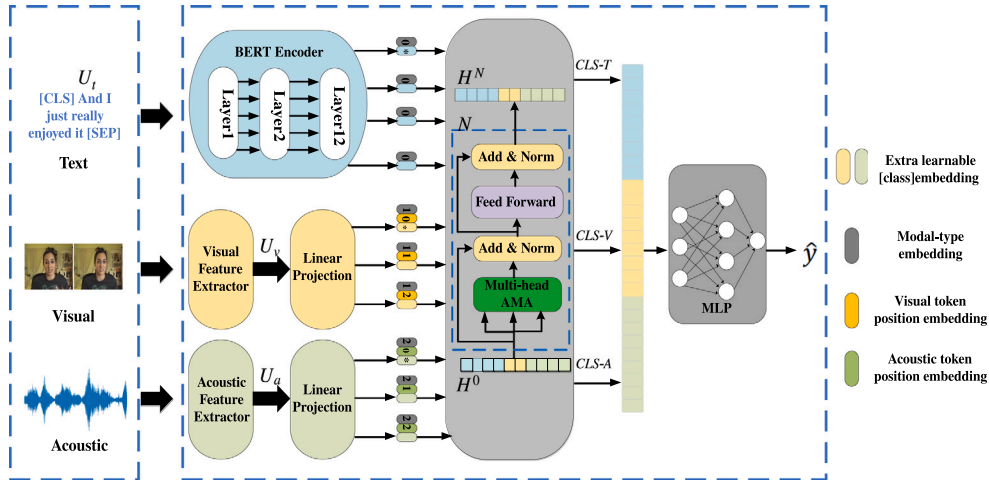


Fig. 2. Overall architecture of MTAMW network.

3.2. Overall architecture

As shown in Fig. 2, the raw input is first processed as a sequential representation vector using a fixed feature extractor for visual and acoustic data, while a pre-trained BERT [18] encoder is used for text. The text, visual, and acoustic representations are then fused using multimodal fusion and multimodal refusion. The multimodal fusion component of this focuses on enabling each modality to accept information from multiple modalities. The multimodal re-fusion component extracts the embedding of each modal head and then proceeds through a multilayer perceptron to obtain a more discriminative representation of the sentiment polarity and output the predicted sentiment intensity.

3.3. Representation extraction and concatenation

The representations of visual and acoustic modalities are vector sequences, however, the representations of textual modality are index sequences. To achieve superior fusion in the following steps, the pre-trained language model BERT is used to encode the input text index sequences, and the output of the last layer is used as the representation of the text modality. The representation of the resulting text modality is

$$H_t = \text{BERT}(U_t; \theta_t^{\text{BERT}}) \in \mathbb{R}^{L'_t \times d} \quad (1)$$

where H_t is the representation of the text modality, θ_t^{BERT} is the parameter of BERT and d is the feature dimension.

In order to concatenate the representations of the three modalities together and then feed them into a single-stream Transformer, the same feature dimensions need to be maintained for the text, visual, and acoustic modalities. Therefore, two fully connected layers are used to transform the feature dimensions of the visual and acoustic modalities.

$$\begin{aligned} H_v &= FC(U_v; \theta_v^{FC}) \in \mathbb{R}^{L'_v \times d} \\ H_a &= FC(U_a; \theta_a^{FC}) \in \mathbb{R}^{L'_a \times d} \end{aligned} \quad (2)$$

where H_v and H_a are representations of visual and acoustic modalities, respectively, and θ_v^{FC} and θ_a^{FC} are parameters of the fully connected layer used to change the dimension of the features of the visual and acoustic modalities, respectively.

After obtaining the representations H_m , $m \in \{t, v, a\}$ of each modality, the multimodal information needs to be fused using a single-stream Transformer by concatenating the representations of the three modalities. However, to make use of the order of the sequences, positional information [14] must be injected as the model contains no recursion and no convolution. Additionally, by concatenating the representations

of the three modalities, the modal-type information is lost, and modal-type information [38] must be injected to make use of it. Because the positional information is already injected after the text representation is obtained using BERT, only the modal-type information needs to be injected into the text representation, but for the visual and acoustic representations, we need to inject both positional and modal-type information into them.

$$H_t^0 = H_t + ME(m^{type}; \theta^M) \quad (3)$$

$$H_m^0 = H_m + ME(m^{type}; \theta^M) + PE(m^{pos}; \theta^P), m \in \{v, a\} \quad (4)$$

where H_m^0 , $m \in \{t, v, a\}$ denotes textual, visual, and acoustic representations that have been injected with positional and modal-type information, respectively. $ME(m^{type}; \theta^M) \in \mathbb{R}^{L'_m \times d}$, $m \in \{t, v, a\}$ computes the embedding of the modal-type index, and $PE(m^{pos}; \theta^P) \in \mathbb{R}^{L'_m \times d}$, $m \in \{v, a\}$ computes the embedding of the index of each position in the same modality. $m^{type} \in \mathbb{R}^{L'_m}$, $m \in \{t, v, a\}$, denote the modal-type indexes of the text, visual and acoustic modalities, respectively. $m^{pos} \in \mathbb{R}^{L'_m}$, $m \in \{v, a\}$ denote the position indexes in the visual and acoustic modalities. θ^M and θ^P denote the parameters for computing the modal-type embedding and for computing the sequence positional embedding, respectively.

Finally H_t^0 , H_v^0 and H_a^0 are concatenated to gain $H^0 \in \mathbb{R}^{L \times d}$, where $L = \sum_{m \in \{t, v, a\}} L'_m$. Then input H^0 into MTAMW.

$$H^0 = \text{Concat}[H_t^0; H_v^0; H_a^0] \quad (5)$$

3.4. Adaptive multimodal attention mechanism

After concatenating the representations of the three modalities, to enable each modality to receive information from multiple modalities, we design an efficient fusion method multimodal transformer with adaptive modality weighting, in which the attention mechanism performs an adaptive multimodal attention mechanism by modifying the Transformer's self-attention mechanism and then concatenating vectors representing the context of each modality for downstream prediction. In the self-attention mechanism of the i th layer MTAMW, given the weight matrices $W_i^Q \in \mathbb{R}^{d \times d_k}$, $W_i^K \in \mathbb{R}^{d \times d_k}$ and $W_i^V \in \mathbb{R}^{d \times d}$, the queries Q^i , keys K^i , and values V^i are calculated as shown below:

$$\begin{aligned} Q^i &= H^i W_i^Q \in \mathbb{R}^{L \times d_k} \\ K^i &= H^i W_i^K \in \mathbb{R}^{L \times d_k} \\ V^i &= H^i W_i^V \in \mathbb{R}^{L \times d} \end{aligned} \quad (6)$$

where H^i is a representation of the multimodal output of the i th layer MTAMW. In particular, H^0 is the original multimodal representation. The definitions of Q^i , K^i , and V^i are similar to those in self-attention [14].

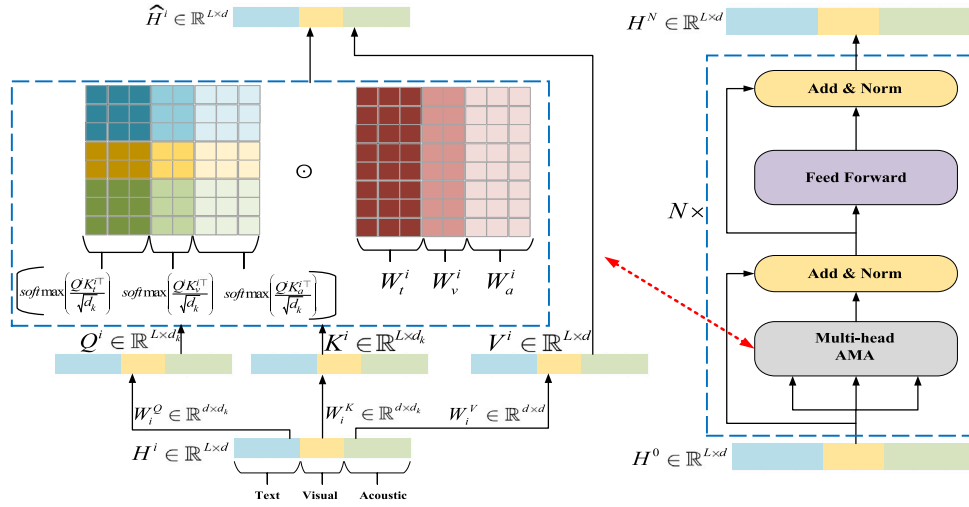


Fig. 3. Details of the MTAMW. The figure on the left shows the overall structure of adaptive multimodal attention mechanism. The figure on the right shows the overall structure of the multimodal transformer with adaptive modality weighting.

3.4.1. Multimodal attention mechanism

However, the feature vectors of different modalities exist in different feature spaces, making the intra-modal attention values significantly larger than the inter-modal attention values when calculating the self-attention of a concatenated sequence, thus the fusion model focuses only on the intra-modal information and ignores the inter-modal information. Inspired by the multi-headed attention mechanism allowing models to jointly attend to information from different representation subspaces at different locations [14]. Multiple Softmax functions were used to calculate the attentional weights for each modality separately, and the Multimodal Attention Mechanism (MAM) was eventually designed.

The core of MAM is to replace a single Softmax with multiple Softmax functions, one intra- and two inter-attention for each modality, thus avoiding the fusion model focusing on the intra-modality information and ignoring the inter-modality information. At the same time, the learnable embedding of each modality was used to learn the sentence-level multimodal representation, and the rest of the tokens provided useful information to the learnable embedding by capturing low-level and fine-grained semantic information, which effectively filtered the noise caused by semantic gaps.

We decompose $H^i W_i^K$ in Eq. (6) by the following equation and then obtain $K_m^i = H_m^i W_i^{K^m}$, $m \in \{t, v, a\}$.

$$\begin{aligned} K^i &= \text{Concat}[H_t^i W_i^{K^t}; H_v^i W_i^{K^v}; H_a^i W_i^{K^a}] \\ &= \text{Concat}[K_t^i; K_v^i; K_a^i] \in \mathbb{R}^{L \times d_k} \end{aligned} \quad (7)$$

where K_m^i , $m \in \{t, v, a\}$ denotes the keys of text, visual and acoustic modalities in the attention mechanism output of the i th layer MTAMW, respectively. H_m^i , $m \in \{t, v, a\}$ denotes the text, visual, and acoustic modal representations of the output of layer i th MTAMW, respectively. $W_i^{K^m}$, $m \in \{t, v, a\}$ denotes the weight matrices of text, visual and acoustic modalities in the attention mechanism of the i th layer MTAMW, respectively.

The multimodal attention $MAM(Q^i, K^i, V^i) \in \mathbb{R}^{L \times d}$ is presented as:

$$\begin{aligned} MAM(Q^i, K^i, V^i) &= \text{Concat}[\text{softmax}(\frac{Q^i K_t^{iT}}{\sqrt{d_k}}); \\ &\quad \text{softmax}(\frac{Q^i K_v^{iT}}{\sqrt{d_k}}); \\ &\quad \text{softmax}(\frac{Q^i K_a^{iT}}{\sqrt{d_k}})] V^i \end{aligned} \quad (8)$$

Specifically, the scaled (by $\sqrt{d_k}$) softmax in Eq. (8) calculates the attention weight matrix $\text{softmax}(\frac{Q^i K_m^{iT}}{\sqrt{d_k}}) \in \mathbb{R}^{L \times L'_m}$, $m \in \{t, v, a\}$, whose (x, y) -th entry measures the attention of the x th feature of the multimodal representation to the y th feature of m modality. Hence, the x th feature of the multimodal representation is a weighted summary of V^i , with the weight determined by the x th row in the concatenated attention weight matrix.

3.4.2. Adaptive weight matrix

Because sentiment-related information is not evenly distributed across modalities, the relative importance of modalities needs to be taken into account to effectively integrate multimodal information. However, Eq. (8) does not reflect the relative importance of modalities. Different weights can be assigned to each modality to indirectly reflect the relative importance of modalities. To adaptively adjust the relative importance of modalities, a multimodal adaptive weight matrix $W^i \in \mathbb{R}^{L \times L}$ is introduced:

$$W^i = \text{Concat}[w_t^i J_t; w_v^i J_v; w_a^i J_a] \quad (9)$$

where $J_m \in \mathbb{R}^{L \times L'_m}$, $m \in \{t, v, a\}$ denote the all-ones matrices used for scalar multiplication with the weights of the corresponding modalities, respectively. W^i has learnable variables w_m^i , $m \in \{t, v, a\}$, representing the weight of text, visual and acoustic modalities in i th layer MTAMW, respectively. An element-wise product of the adaptive weight matrix and the attention weight matrix can change the attention weights of modalities and hence reflect the relative importance of the modalities. Where w_t^i , w_v^i , and w_a^i can be used to adjust the weight of different modalities in the same layer of MTAMW, and $w_m^1, w_m^2, \dots, w_m^N$ can be used to adjust the weight of the same modality at different levels of MTAMW. The results after element-wise product with the adaptive weight matrix and attention weights are then matmul product with V^i to finally obtain Adaptive Multimodal Attention (AMA). The AMA details are shown in the figure on the left in Fig. 3. The AMA is calculated as follows:

$$\begin{aligned} AMA(Q^i, K^i, V^i) &= (\text{Concat}[\text{softmax}(\frac{Q^i K_t^{iT}}{\sqrt{d_k}}); \\ &\quad \text{softmax}(\frac{Q^i K_v^{iT}}{\sqrt{d_k}}); \\ &\quad \text{softmax}(\frac{Q^i K_a^{iT}}{\sqrt{d_k}})]) \odot W^i V^i \end{aligned} \quad (10)$$

where \odot denotes element-wise product.

After obtaining the weights of each modality in the i th layer MTAMW, the relative importance between the modalities can be calculated by the following:

$$\begin{aligned}
 RI(A^i) &= \frac{\mathbb{E}[w_a^i softmax(U_a)]}{\sum_{m \in \{t,v,a\}} \mathbb{E}[w_m^i softmax(U_m)]} \\
 &= \frac{\frac{\sum_{j=1}^{L_a'} w_a^j softmax(u_a^j)}{L_a'}}{\sum_{m \in \{t,v,a\}} \frac{\sum_{j=1}^{L_m'} w_m^j softmax(u_m^j)}{L_m'}} \\
 &= \frac{\frac{w_a^j}{L_a'}}{\sum_{m \in \{t,v,a\}} \frac{w_m^j}{L_m'}}
 \end{aligned} \quad (11)$$

where $RI(A^i)$ denotes the value of the relative importance of the acoustic modality in the i th layer of the MTAMW, and $\mathbb{E}[w_m^i softmax(U_m)]$ denotes the average attention score assigned to the m modality, $m \in \{t, v, a\}$, in the i th layer of the MTAMW. Since multiple Softmax functions are used to calculate the attentional scores, $\sum_{j=1}^{L_a'} softmax(u_m^j) = 1$.

3.5. Multimodal transformer with adaptive modality weighting

Based on the adaptive multimodal attention blocks, a multimodal transformer with adaptive modality weighting has been designed, which enables one modality to accept information from multimodal. The structure of the MTAMW is shown in the subfigure on the right in Fig. 3. The MTAMW consists of stacked blocks which include a multi-head Adaptive Multimodal Attention (AMA) layer and a Multilayer Perceptron (MLP). Formally, the multimodal representation H^0 is input, and then MTAMW computes feed-forwardly for $i = 1 \dots N$ layers:

$$\begin{aligned}
 &\text{Calculate } Q^{i-1}, K^{i-1}, \text{ and } V^{i-1} \\
 \hat{H}^i &= LN(AMA^{mul}(Q^{i-1}, K^{i-1}, V^{i-1}) + H^{i-1}) \\
 H^i &= LN(MLP(\hat{H}^i) + \hat{H}^i)
 \end{aligned} \quad (12)$$

where AMA^{mul} denotes the multi-head adaptive multimodal attention mechanism.

3.6. Re-fusion for sentiment information

To obtain sentiment-focused relevant features from the multimodal representations, the embedding $CLS-T$, $CLS-V$, and $CLS-A$ from the multimodal feature sequences output by the last layer of the MTAMW are concatenated, and they are fed into the multilayer perceptron to help learn more discriminative representations for different sentiment polarities.

$$\hat{y} = MLP(Concat[CLS-T; CLS-V; CLS-A]) \quad (13)$$

3.7. Train

The loss function to be optimized is the loss function for the MSA task only, with the following equation:

$$\mathcal{L} = MSE(\hat{y}, y) \quad (14)$$

where \hat{y} is the MTAMW model outputting the predicted value of sentiment intensity and y is the true value of sentiment intensity. Here MSE is denoted as mean squared error loss and is a common method used for regression tasks. We summarize the training algorithm in Algorithm 1.

4. Experiments

In this section, some details of the experiment, including the dataset, feature extraction, baselines, results, etc., will be presented.

Algorithm 1 Multimodal Transformer with Adaptive Modality Weighting

Input: Unimodal inputs text U_t , visual U_v , acoustic U_a , multimodal sentiment label y
Output: Prediction sentiment intensity \hat{y}

- 1: Initialize model parameters $M(\theta; x)$
- 2: **for** each training epoch **do**
- 3: **for** mini-batch in dataLoader **do**
- 4: $H_t = BERT(U_t)$ // Compute text features
- 5: $H_v = FC(U_v)$, $H_a = FC(U_a)$ // Transform the feature dimensions
- 6: $H^0 \leftarrow (H_t, H_v, H_a)$ by Eq. (3) ~ Eq. (5) // Compute multimodal sequence features
- 7: $H^N = MTAMW(H^0)$ // Compute multimodal information
- 8: $\hat{y} = MLP(CLS-T, CLS-V, CLS-A)$ // Re-fusion sentiment information and compute sentiment intensity
- 9: $\mathcal{L} = MSE(\hat{y}, y)$ // Compute the loss
- 10: Update all network parameters θ using BP algorithm
- 11: **end for**
- 12: **end for**

4.1. Datasets

We evaluated the MTAMW model on five publicly available multimodal sentiment analysis datasets, CMU-MOSI [39], CMU-MOSEI [40], and MOSI-ASR [41] (containing three datasets), respectively.

CMU-MOSI: This is an MSA dataset containing linguistic, visual, and acoustic modalities. The video clips were sourced from the Youtube website, using 93 videos uploaded by Youtubers on the website commenting on films. The videos were finally intercepted into 2199 video clips. Each clip was annotated with a sentiment intensity ranging from -3 to $+3$, indicating the polarity and intensity of the sentiment.

CMU-MOSEI: Similar to the CMU-MOSI dataset, the CMU-MOSEI dataset also contains linguistic, visual, and acoustic modalities. However, the size of the CMU-MOSEI dataset is much larger than that of CMU-MOSI. It contains 23453 annotated video clips from online video sharing sites covering 250 different topics and 1000 different speakers. The sentiment intensity of each video clip is also in the range of -3 to $+3$.

MOSI-ASR: When state-of-the-art models are deployed in the real world, their performance often experiences a significant drop [41]. The main reason behind this is that practical applications can only obtain textual outputs through automatic speech recognition (ASR) [42] models, which are prone to errors due to limitations in model capacity. In order to effectively evaluate the performance of models in real-world applications and further analyze this issue, SWRM [41] has constructed three real-world multimodal sentiment analysis datasets based on the existing CMU-MOS dataset. Specifically, three widely used ASR APIs, SpeechBrain [43], IBM,² and iFlytek,³ were employed to process the raw audio and obtain transcribed texts. The ASR outputs were then substituted for the manually transcribed texts in CMU-MOSI, with the visual and audio features remaining unchanged, resulting in three real-world datasets⁴: MOSI-SpeechBrain, MOSI-IBM, and MOSI-iFlytek, which were segmented in the same way as CMU-MOSI.

The split specifications of the five datasets are provided in Table 1.

4.2. Feature extraction

Language Features: For the extraction of language features, current approaches generally use word embeddings, either using GloVe [44] to

² <https://www.ibm.com/cloud/watson-speech-to-text>

³ <https://global.xfyun.cn/products/lfasr>

⁴ <https://github.com/albertwy/SWRM>

Table 1
Dataset statistics in CMU-MOSI, CMU-MOSEI, and MOSI-ASR.

| Split | CMU-MOSI | CMU-MOSEI | MOSI-ASR |
|------------|----------|-----------|----------|
| Train | 1284 | 16 326 | 1284 |
| Validation | 229 | 1871 | 229 |
| Test | 686 | 4659 | 686 |
| All | 2199 | 22 856 | 2199 |

generate word vectors or using hidden states from pre-trained language models to represent word vectors. In this paper, BERT is used to extract text features, and the BERT embedding size is 768.

Visual Features: For the visual modality, CMU-MOSI, CMU-MOSEI, and MOSI-ASR use Facet to extract the facial features of the commentators in the video. For the CMU-MOSI and MOSI-ASR datasets, Facet is used to extract visual features with 20-dimensions per frame. For the CMU-MOSEI dataset, Facet is used to extract visual features with 35-dimensions per frame.

Acoustic Features: For the acoustic modalities, both datasets use COVAREP to extract acoustic features. For the CMU-MOSI and MOSI-ASR datasets, COVAREP is used to extract acoustic features with 5-dimensions per frame. For the CMU-MOSEI dataset, COVAREP is used to extract acoustic features with 74-dimensions per frame.

To be fair for comparison, for the CMU-MOSI and CMU-MOSEI datasets we used the same features from MMIM and compared them with SOTA methods including MAG-BERT, Self-MM, and MMIM. The dataset provided by SWRM [41] was used for MOSI-ASR.

4.3. Baselines

Compare the MTAMW model with the following baselines:

- **TFN:** The Tensor Fusion Network [6] models intra modality and inter modality information through tensor outer product, so as to capture the interaction information between the three modes.
- **LMF:** The Low rank Multimodal Fusion [24] is an improvement on the TFN model. It uses the mode specific low order factor to fuse multimodal information by decomposing the high-order tensor.
- **MFM:** The Multimodal Factorization Model [45] decomposes the multimodal representation into two independent groups of factors, one for discriminant representation and the other for generative representation.
- **ICCN:** The Interaction Canonical Correlation Network [13] minimizes canonical loss between text–audio modal representation pairs and between text–visual modal representation pairs, ultimately enhancing the effectiveness of multimodal fusion.
- **Mult:** The Multimodal Transformer [7] models all pairs of modes by using the cross modal attention mechanism, and fuses multimodal information by directly focusing on low-level features in other modes.
- **MAG-BERT:** The Multimodal Adaptation Gate for BERT [18] uses the information of non linguistic modes to adjust the linguistic information of linguistic modes, and then inputs it into the pre training model BERT for fine adjustment.
- **MISA:** The Modality Invariant and Specific Representations [8] project each mode into two different subspaces, which are used to learn the commonalities between different modal representations, reduce modal differences and learn the unique features of each mode.
- **Self-MM:** The Learning Modality Specific Representations with Self Supervised Multi Task Learning [9] uses self supervised learning to obtain labels for each mode, and then combines multimodal and single-mode learning (multi task learning) to learn consistency and difference respectively.

- **MIMM:** The Improving Multimodal Fusion with Hierarchical Mutual Information Maximization [11] can effectively retain task related information during multimodal fusion by adding the maximum mutual information joint main task to the single mode input pair (between modes) and between multimodal fusion results and single mode input for training.
- **SWRM:** The Sentiment Word Aware Multimodal Refinement [41] uses the sentiment word aware multimodal refinement model to dynamically refine the erroneous sentiment words by leveraging multimodal sentiment clues.

4.4. Evaluation metrics

For the regression task the mean absolute error (MAE) was used to measure the mean absolute error between predicted and true values, and Pearson correlation (Corr) was also used to measure correlation. For the evaluation classification task, the effectiveness of the model in classifying sentiment was evaluated using seven-class accuracy (Acc-7) ranging from -3 to 3 , binary classification (Acc-2) and F1 score (F1) on both the (non-negative / negative) dataset and the (negative / positive) dataset.

4.5. Parameter settings

Adam was used as the optimizer for CMU-MOSI and CMU-MOSEI, with an initial learning rate of $1e-2$ for the parameters of the Adaptive Weight Matrix (AWM) and learning rates in $\{2e-5, 4e-5, 6e-5, 8e-5\}$ for the other parameters. For MOSI-ASR, Adam was used as the optimizer with an initial learning rate of $1e-3$ for the parameters of the AWM and learning rates in $\{2e-5, 4e-5, 6e-5, 8e-5\}$ for the other parameters. The number of layers of the MTAMW is $N = 3$. The sequence length L'_m , $m \in \{t, v, a\}$ is 50, 70 and 80, respectively. For a fair comparison, our model uses the same non-linguistic features as state-of-the-art methods.

5. Results and discussion

5.1. Comparison with baselines

Table 2 shows the results of the comparison of the datasets CMU-MOSI and CMU-MOSEI. Each task consists of an *Aligned* and an *Unaligned* version, depending on the *Data Setting*. To ensure fairness, the same dataset as MMIM was used. From Table 2 we can see that the results from MTAMW are better or more comparable than many of the baseline methods. Specifically, our model outperforms SOTA for all metrics except Acc-7 on CMU-MOSEI and in Acc-7, Acc-2 and F1 on both the negative/non-negative and positive/negative datasets on CMU-MOSI. These results provide preliminary evidence of the validity of our approach in the MSA task.

Table 3 shows that the results for MOSI-SpeechBrain, MOSI-IBM, and MOSI-iFlytek demonstrate that the performance of MTAMW on datasets with ASR Errors is greatly improved compared to SOTA, indicating that our proposed method can effectively filter the noise from ASR errors when obtaining multimodal sentence-level representations.

5.2. Impact of varying feature dimensions

We will investigate the impact of different feature dimensions when concatenating visual, textual, and acoustic features in the MTAMW model. We aim to determine how changing the number of feature dimensions affects model performance. To conduct our experiments, we follow a procedure in which, after extracting textual representations using Eq. (1), we apply a fully connected layer to adjust the dimensions of textual features to match the dimensions of visual and audio features obtained using Eq. (2). Subsequently, we concatenate these features. It is worth noting that this fully connected layer is not present in the

Table 2

Sentiment prediction results on CMU-MOSI and CMU-MOSEI. “–” indicates that the relevant value is not given in the corresponding reference. [◊] indicates that all models use BERT as a text encoder. The best results are highlighted in bold. ¹, ² and ³ indicate that the corresponding results are taken from the Refs. [8], [9] and [11]. _{*} indicates that models are under the same non-linguistic feature conditions.

| Model [◊] | CMU-MOSI | | | | | CMU-MOSEI | | | | | Data setting |
|------------------------------------|--------------|--------------|--------------|--------------------|--------------------|--------------|--------------|--------------|--------------------|--------------------|--------------|
| | MAE | Corr | Acc-7 | Acc-2 | F1-Score | MAE | Corr | Acc-7 | Acc-2 | F1-Score | |
| TFN ¹ | 0.901 | 0.698 | 34.9 | −/80.8 | −/80.7 | 0.593 | 0.700 | 50.2 | −/82.5 | −/82.1 | Unaligned |
| LMF ¹ | 0.917 | 0.695 | 33.2 | −/82.5 | −/82.4 | 0.623 | 0.677 | 48.0 | −/82.0 | −/82.1 | Unaligned |
| MF ¹ | 0.877 | 0.706 | 35.4 | −/81.7 | −/81.6 | 0.568 | 0.717 | 51.3 | −/84.4 | −/84.3 | Aligned |
| ICCN ¹ | 0.862 | 0.714 | 39.0 | −/83.0 | −/83.0 | 0.565 | 0.713 | 51.6 | −/84.2 | −/84.2 | Aligned |
| MuT ² | 0.861 | 0.711 | – | 81.5/84.1 | 80.6/83.9 | 0.580 | 0.703 | – | −/82.5 | −/82.3 | Unaligned |
| MAG-BERT ² | 0.731 | 0.789 | – | 82.5/84.3 | 82.6/84.3 | 0.539 | 0.753 | – | 83.8/85.2 | 83.7/85.1 | Aligned |
| MISA ¹ | 0.783 | 0.761 | 42.3 | 81.8/83.4 | 81.7/83.6 | 0.555 | 0.756 | 52.2 | 83.6/85.5 | 83.8/85.3 | Aligned |
| Self-MM ² | 0.713 | 0.798 | – | 84.00/85.98 | 84.42/85.95 | 0.530 | 0.765 | – | 82.81/85.17 | 82.53/85.30 | Unaligned |
| MAG-BERT ³ _* | 0.727 | 0.781 | 43.62 | 82.37/84.43 | 82.50/84.61 | 0.543 | 0.755 | 52.67 | 82.51/84.82 | 82.77/84.71 | Aligned |
| Self-MM ³ _* | 0.712 | 0.795 | 45.79 | 82.54/84.77 | 82.68/84.91 | 0.529 | 0.767 | 53.46 | 82.68/84.96 | 82.95/84.93 | Unaligned |
| MMIM ³ _* | 0.700 | 0.800 | 46.65 | 84.14/86.06 | 84.00/85.98 | 0.526 | 0.772 | 54.24 | 82.24/85.97 | 82.66/85.94 | Unaligned |
| MTAMW _* | 0.712 | 0.794 | 46.84 | 84.40/86.59 | 84.20/86.46 | 0.525 | 0.782 | 53.73 | 83.09/86.49 | 83.48/86.45 | Unaligned |

Table 3

Experimental results using MOSI-ASR datasets. [†] indicates that the corresponding result is from the Ref. [41].

| Model | MOSI-SpeechBrain | | | |
|----------------------|------------------|--------------|--------------------|--------------------|
| | MAE | Corr | Acc-2 | F1-Score |
| TFN(B) [†] | 1.156 | 0.485 | 68.98/69.51 | 68.95/69.57 |
| LMF(B) [†] | 1.174 | 0.487 | 68.86/69.36 | 68.88/69.48 |
| MuT(B) [†] | 1.090 | 0.547 | 71.78/72.74 | 71.70/72.75 |
| MISA [†] | 0.985 | 0.654 | 73.79/74.51 | 73.85/74.66 |
| Self-MM [†] | 0.910 | 0.672 | 73.67/74.85 | 73.72/74.98 |
| SWRM [†] | 0.906 | 0.675 | 74.58/75.70 | 74.62/75.82 |
| MTAMW | 0.884 | 0.686 | 76.82/78.66 | 76.69/78.61 |
| Model | MOSI-IBM | | | |
| | MAE | Corr | Acc-2 | F1-Score |
| TFN(B) [†] | 1.094 | 0.582 | 71.81/72.13 | 71.78/73.21 |
| LMF(B) [†] | 1.047 | 0.591 | 73.06/74.30 | 73.09/74.41 |
| MuT(B) [†] | 1.003 | 0.643 | 75.57/76.74 | 75.54/76.79 |
| MISA [†] | 0.912 | 0.713 | 76.97/78.08 | 76.99/78.17 |
| Self-MM [†] | 0.857 | 0.732 | 77.32/78.60 | 77.37/78.72 |
| SWRM [†] | 0.829 | 0.739 | 78.43/79.70 | 78.47/79.80 |
| MTAMW | 0.804 | 0.747 | 80.47/82.16 | 80.44/82.20 |
| Model | MOSI-iFlytek | | | |
| | MAE | Corr | Acc-2 | F1-Score |
| TFN(B) [†] | 1.070 | 0.565 | 71.95/72.62 | 72.01/72.76 |
| LMF(B) [†] | 1.066 | 0.595 | 71.98/72.35 | 72.03/72.49 |
| MuT(B) [†] | 0.898 | 0.681 | 77.32/78.75 | 77.05/78.56 |
| MISA [†] | 0.856 | 0.745 | 79.59/79.82 | 79.62/79.91 |
| Self-MM [†] | 0.788 | 0.758 | 80.26/81.16 | 80.26/81.20 |
| SWRM [†] | 0.784 | 0.760 | 80.47/81.28 | 80.47/81.34 |
| MTAMW | 0.773 | 0.771 | 83.09/83.23 | 83.31/83.32 |

original MTAMW model. While keeping all other model parameters constant, we solely manipulate the number of feature dimensions in the concatenated multimodal features. From Table 4, it can be observed that the model performs best when the feature dimension for multimodal concatenation is 768, which matches the dimension of textual features extracted by BERT.

5.3. Ablation study

To further study the influence of each component in the MTAMW. A very comprehensive ablation analysis was carried out on the dataset CMU-MOSEI. These are the following models obtained with modifications to the MTAMW:

- $W_{v,a} - W_t$: In the adaptive weight matrix of MTAMW, the same weights are used to measure the importance of the visual and acoustic modalities, and the importance of the text modality is measured using separate weights.

- $W_{t,a} - W_v$: In the MTAMW model, the same weights are used to measure the importance of text and acoustic modalities, and the importance of visual modalities is measured using separate weights.
- $W_{t,v} - W_a$: The same weights are used to measure the importance of text and visual modalities, and the importance of acoustic modalities is measured using separate weights.
- w/o AWM: Remove the Adaptive Weight Matrix components from the MTAMW model.
- rp MAM: Replacing the Multimodal Attention Mechanism with the self-attention mechanism.
- w/o all: Remove the AWM component from the MTAMW model and replace the MAM with the self-attention mechanism.
- rp CELoss: Replacing MSE loss with Cross-Entropy loss.

The results at different ablation settings are shown in Table 5. It can be noted that there is a significant drop in the performance of the model when the same weights are used by two modalities to measure their importance. In particular, when the AWM is removed, the three modalities are considered equally important, and the performance of the model decreases even more significantly. This demonstrates that textual, visual, and acoustic modalities contribute differently to sentiment information, further indicating the validity of the proposed AWM. It is also noted that using the self-attention mechanism to replace the MAM results in a significant drop in the performance of the model. Furthermore, when the AWM is removed from the model and the MAM is replaced with a self-attention mechanism, the model performs worse than when one of the components is changed. This effectively demonstrates the validity of the AWM and MAM components in the MTAMW model and provides further evidence of the applicability of the MTAMW model to the MSA task. Additionally, we compared the effects of using MSE loss and Cross-Entropy loss on the model's performance. The results clearly indicated that the model trained with MSE loss outperformed the model trained with Cross-Entropy loss. This finding suggests that MSE loss is more suitable for sentiment analysis tasks like MOSI and MOSEI, yielding better performance in multimodal sentiment analysis.

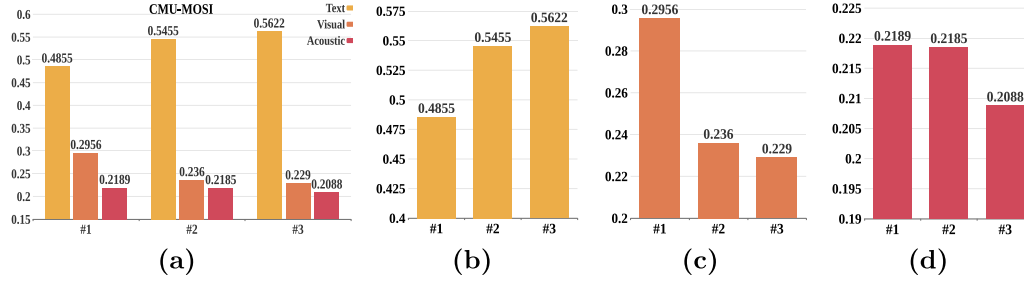
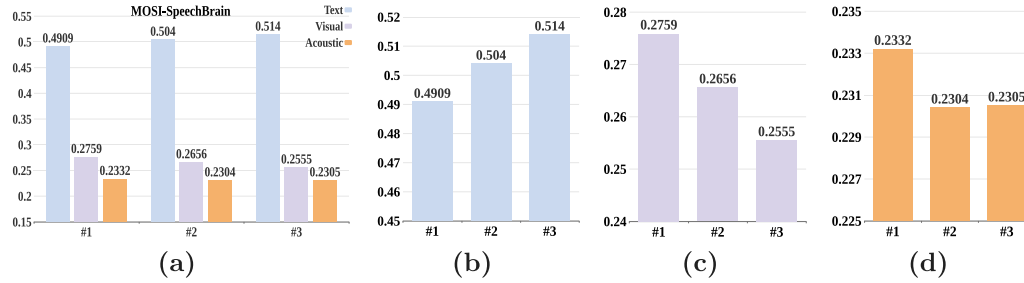
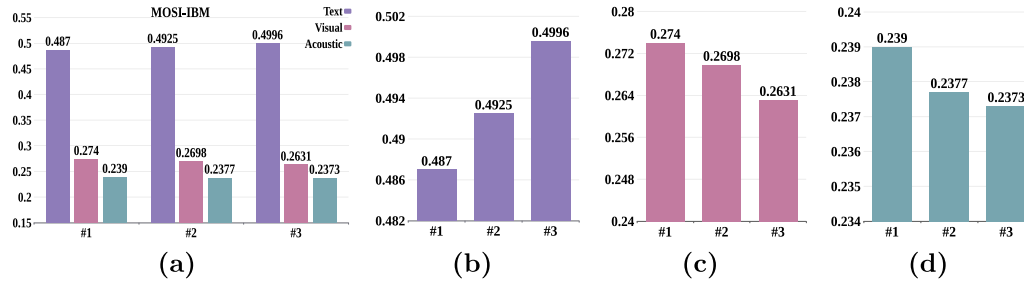
5.4. Visualization of relative importance between modalities

To obtain the relative importance between the modalities in the CMU-MOSI, MOSI-SpeechBrain, and MOSI-IBM datasets, The value of the relative importance of each modality in each layer of MTAMW was calculated by (11), and the results are shown in Figs. 4, 5, and 6. where (a) indicates the variation in the relative importance of text, visual, and acoustic modalities in the different layers of MTAMW. (b), (c) and (d) indicate the variation in the relative importance of the corresponding modalities in the different layers of MTAMW, respectively. The numbers (#) under each sub-picture indicate the corresponding number of

Table 4

Impact of different feature dimensions for model performance on CMU-MOSI and CMU-MOSEI.

| Dimension | CMU-MOSI | | | | | CMU-MOSEI | | | | |
|------------|--------------|--------------|--------------|--------------------|--------------------|--------------|--------------|--------------|--------------------|--------------------|
| | MAE | Corr | Acc-7 | Acc-2 | F1-Score | MAE | Corr | Acc-7 | Acc-2 | F1-Score |
| $d = 64$ | 0.748 | 0.791 | 45.04 | 82.50/84.60 | 82.48/84.64 | 0.538 | 0.777 | 52.65 | 81.65/85.25 | 82.12/85.21 |
| $d = 128$ | 0.732 | 0.792 | 44.75 | 82.79/85.06 | 82.67/85.01 | 0.537 | 0.778 | 53.12 | 80.27/85.55 | 80.99/85.63 |
| $d = 256$ | 0.722 | 0.793 | 45.33 | 83.09/85.51 | 82.95/85.46 | 0.533 | 0.777 | 52.92 | 82.38/85.82 | 82.77/85.76 |
| $d = 512$ | 0.728 | 0.795 | 45.87 | 83.57/ 85.82 | 83.55/85.77 | 0.531 | 0.780 | 53.34 | 82.59/86.18 | 83.02/86.15 |
| MTAMW(768) | 0.712 | 0.794 | 46.84 | 84.40/86.59 | 84.20/86.46 | 0.525 | 0.782 | 53.73 | 83.09/86.49 | 83.48/86.45 |

**Fig. 4.** Visualization of the importance of different modalities in CMU-MOSI.**Fig. 5.** Visualization of the importance of different modalities in MOSI-SpeechBrain.**Fig. 6.** Visualization of the importance of different modalities in MOSI-IBM.

layers in MTAMW. As can be noted in Figs. 4(a), 5(a), and 6(a), for all three datasets, the relative importance between modalities when fusing multimodal information for each layer of MTAMW is shown to be: text > visual > audio. This means that the text modality has more information related to sentiment, followed by the visual modality and finally the acoustic modality. Different modalities contribute differently to sentiment information because they convey information in different ways and capture different aspects of the expressed sentiment. Among other things, text can provide explicit information about the expressed sentiment through the use of emotive words, modifiers, and other linguistic features, and text may be more reliable as a source of emotional information. Visual content can convey emotion through facial expressions, body language, and other non-linguistic cues. Audio can convey emotion through the speaker's tone, pitch, and intensity, and audio may be more ambiguous as a source of sentiment analysis.

We also compared the relative importance of each modality in Fig. 4(a) with that of Figs. 5(a) and 6(a), and found that the relative

importance of the text modality in Fig. 4(a) was higher than that of 5(a) and 6(a), while the relative importance of both the visual and audio modalities in Fig. 4(a) was lower than that of 5(a) and 6(a). This is because the text acquired by ASR is accompanied by errors, making the quality of the text lower. From the variation of the relative importance of each modality in different layers of MTAMW in Fig. 4(b) to Fig. 4(d), Fig. 5(b) to Fig. 5(d), and Fig. 6(b) to Fig. 6(d), it can be noted that the relative importance of the text modality shows an increasing trend, while the relative importance of the visual and acoustic modalities basically shows a decreasing trend. This is because, in multimodal semantic information, the importance of a particular modality may vary depending on the level of semantic processing or analysis being performed. Also, this shows that the same modality is of different importance in different levels of multimodal information.

The above demonstrates that our model when fusing multimodal information can adaptively assign appropriate weights to each modality based on the sentiment-related information contained in each modality,




Table 5

Ablation study of MTAMW on CMU-MOSEI.

| Description | MAE | Corr | Acc-2 | F1-Score |
|-----------------|--------------|--------------|--------------------|--------------------|
| $W_{v,a} - W_t$ | 0.530 | 0.780 | 82.98/86.05 | 83.33/85.98 |
| $W_{t,a} - W_v$ | 0.530 | 0.777 | 81.76/85.77 | 82.25/85.76 |
| $W_{t,v} - W_a$ | 0.531 | 0.779 | 83.01/85.83 | 83.40/85.72 |
| w/o AWM | 0.545 | 0.775 | 80.62/85.39 | 81.25/85.43 |
| rp MAM | 0.544 | 0.766 | 82.74/85.75 | 83.09/85.66 |
| w/o all | 0.539 | 0.772 | 80.17/84.84 | 80.85/84.89 |
| rp CELoss | – | – | 82.97/85.25 | 83.29/85.24 |
| MTAMW | 0.525 | 0.782 | 83.09/86.49 | 83.48/86.45 |

Table 6

Case study for MTAMW on CMU-MOSEI.

| Modality | Example | Label | MTAMW |
|----------|-----------------------------------------------------------------------------------|-------|-------|
| T | It was boring | | |
| A |  | –2.33 | –2.39 |
| V | Smile and surprised | | |
| T | It's actually based on a novel (umm) | | |
| A |  | 0.00 | 0.00 |
| V | Bland expression | | |
| T | It's rated g so the whole family can enjoy it | | |
| A |  | 1.00 | 1.00 |
| V | Head down | | |

and can also adaptively adjust the relative importance of the modalities according to the change in the quality of the modalities. The importance of each modality in different layers of MTAMW can also be adaptively adjusted.

5.5. Case study

Table 6 displays the predicted and true values for the three cases, along with the corresponding input data (for the visual modality, it is illustrated literally). For the first case, if the sentiment is only analyzed from the textual modality, there is a high probability that a very negative sentiment will be predicted, but if the information from the visual modality is incorporated, the predicted intensity of the sentiment is more accurate. The first case shows that the MTAMW model can effectively use information from both textual and visual modalities to predict sentiment intensity. For the second case, our model can effectively use visual and acoustic information to complement the textual information, and the final predicted sentiment intensity is equal to the true sentiment intensity. As far as the third case is concerned, visual modality and acoustic modality cannot add useful information to text modality. If you only use text information to predict sentiment intensity, will get results that are very close to the truth. It can be seen that the predicted results of the MTAMW model are equal to the real results.

The results of these cases show that the MTAMW model can adaptively adjust the importance of different modalities for multimodal fusion.

5.6. Further analysis

We chose a case study for the visualization of the attention scores of the last layer in the MTAMW, and the results are shown in Fig. 7. Overall all modalities focus attention on textual and visual information, as acoustic information can only provide low-level semantic information like tone. For the text modality, attention was focused not only on useful textual information but also very accurately on the visual area

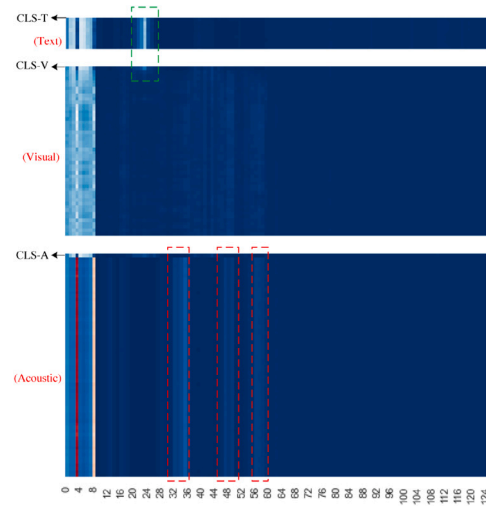


Fig. 7. Visualization of attention scores in the last layer of the MTAMW. The effective sequence length is 125, where 0–8 belong to the attention scores of the text modality, which are specified as ‘[CLS]’, ‘but’, ‘it’, ‘ ’ ’, ‘s’, ‘a’, ‘great’, ‘color’, ‘[SEP]’ after splitting. 9–58 are the attention scores of the visual modality. 59–124 are the attention scores of the acoustic modality. The visual information in the green dashed box is the expression of a smile. The visual interval in the red dashed box corresponds to the high tone of the speaker.

where the expression was a smile, as well as the $CLS-T$ effectively capturing multimodal information. For the visual modality, when acquiring textual information, $CLS-V$ was used mainly to capture useful textual information, while the rest of the visual tokens focused their attention on less informative tokens (“ ’ ” and “[SEP]”) in order not to bring in noise, and this was even more evident for the acoustic modality when acquiring textual information. The acoustic modality, when acquiring visual information, focuses its attention mainly on the visual intervals with a high tone. In general, the visual and acoustic modalities rely primarily on $CLS-V$ and $CLS-A$ to capture high-level semantic information, while the rest of the tokens capture useful low-level and fine-grained information. This case effectively demonstrates that MTAMW can be effective in focusing attention on useful information.

6. Conclusion

In this paper, MTAMW is proposed, in which the adaptive weight matrix can effectively adjust the importance of different modal information, and the multimodal attention mechanism can capture multimodal information very effectively. Comprehensive experiments on various datasets are conducted, followed by an ablation study, and the results validate the effectiveness of the proposed model. Representative examples are shown and the attention scores of one case are visualized to provide a deeper insight into how the model works. In future work, we will build an end-to-end multimodal learning network and explore the relationship between the relative importance of modalities.

CRedit authorship contribution statement

Yifeng Wang: Conceptualization, Methodology. **Jiahao He:** Software, Writing – original draft. **Di Wang:** Formal analysis, Writing – review & editing. **Quan Wang:** Project administration, Supervision. **Bo Wan:** Data curation, Methodology. **Xuemei Luo:** Visualization, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2022ZD0117103, in part by the National Natural Science Foundation of China under Grants 62072354, 62002272 and 62072355, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23084, in part by the by the Foundation of National Key Laboratory of Human Factors Engineering under Grant 6142222210101, in part by the Science and Technology Program of Guangzhou under Grant SL2022A04J00303, in part by a grant from the Youth Innovation Team of Shaanxi Universities under Grant 2023-CX-TD-08, and in part by the Shaanxi Qinchuangyuan “scientists+engineers” team under Grant 2023KXJ-040.

References

- [1] S.G. Dacko, Enabling smart retail settings via mobile augmented reality shopping apps, *Technol. Forecast. Soc. Change* 124 (2017) 243–256.
- [2] K. Song, T. Yao, Q. Ling, T. Mei, Boosting image sentiment analysis with visual attention, *Neurocomputing* 312 (2018) 218–228.
- [3] C.C. Green, B. Raphael, Research on intelligent question-answering system, *Tech. rep.*, Stanford Research Inst Menlo Park CA, 1967.
- [4] P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, *Multimedia syst.* 16 (6) (2010) 345–379.
- [5] B. Yang, B. Shao, L. Wu, X. Lin, Multimodal sentiment analysis with unidirectional modality translation, *Neurocomputing* 467 (2022) 130–137.
- [6] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.
- [7] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2019, NIH Public Access, 2019, p. 6558.
- [8] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.
- [9] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, (12) 2021, pp. 10790–10797.
- [10] W. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2020, NIH Public Access, 2020, p. 2359.
- [11] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9180–9192.
- [12] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 6–15.
- [13] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, (05) 2020, pp. 8992–8999.
- [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [15] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, *AI Open* (2022).
- [16] D. Lin, X. Tang, Inter-modality face recognition, in: *Computer Vision—ECCV 2006: 9th European Conference on Computer Vision*, Graz, Austria, May 7–13, 2006, *Proceedings, Part IV* 9, Springer, 2006, pp. 13–26.
- [17] Y. Li, K. Zhang, J. Wang, X. Gao, A cognitive brain model for multimodal sentiment analysis based on attention neural networks, *Neurocomputing* 430 (2021) 159–173.
- [18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
- [19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [20] S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2539–2544.
- [21] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: *2016 IEEE 16th International Conference on Data Mining (ICDM)*, IEEE, 2016, pp. 439–448.
- [22] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, *Handb. Brain Theory Neural Netw.* 3361 (10) (1995) 1995.
- [23] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780.
- [24] Z. Liu, Y. Shen, Efficient low-rank multimodal fusion with modality-specific factors, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2018.
- [25] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.
- [26] M. Kamrul Hasan, M. Saiful Islam, S. Lee, W. Rahman, I. Naim, M.I. Khan, E. Hoque, TextMI: Textualize multimodal information for integrating non-verbal cues in pre-trained language models, 2023, arXiv e-prints, arXiv:2303.
- [27] Y. Hwang, J.-H. Kim, Self-supervised unimodal label generation strategy using recalibrated modality representations for multimodal sentiment analysis, in: *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 35–46.
- [28] Z. Tang, Q. Xiao, X. Zhou, Y. Li, C. Chen, K. Li, Learning discriminative multi-representation representations for multimodal sentiment analysis, *Inform. Sci.* 641 (2023) 119125.
- [29] P.J. Burt, Attention mechanisms for vision in a dynamic world, in: *9th International Conference on Pattern Recognition*, IEEE Computer Society, 1988, pp. 977–978.
- [30] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.
- [31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: *International Conference on Machine Learning*, PMLR, 2015, pp. 2048–2057.
- [32] G. Kv, A. Mittal, Reducing language biases in visual question answering with visually-grounded question encoder, in: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII* 16, Springer, 2020, pp. 18–34.
- [33] X. Wei, T. Zhang, Y. Li, Y. Zhang, F. Wu, Multi-modality cross attention network for image and sentence matching, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10941–10950.
- [34] Y. Pang, Y. Li, J. Shen, L. Shao, Towards bridging semantic gap to improve semantic segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4230–4239.
- [35] Z. Zhao, H. Zhu, Z. Xue, Z. Liu, J. Tian, M.C.H. Chua, M. Liu, An image-text consistency driven multimodal sentiment analysis approach for social media, *Inf. Process. Manage.* 56 (6) (2019) 102097.
- [36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: *International Conference on Learning Representations*, 2020.
- [37] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP—A collaborative voice analysis repository for speech technologies, in: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 960–964.
- [38] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 5583–5594.
- [39] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, arXiv preprint arXiv:1606.06259.
- [40] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
- [41] Y. Wu, Y. Zhao, H. Yang, S. Chen, B. Qin, X. Cao, W. Zhao, Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors, 2022, arXiv preprint arXiv:2203.00257.
- [42] D. Yu, L. Deng, Automatic Speech Recognition, vol. 1, Springer, 2016.
- [43] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawlatiabad, A. Heba, J. Zhong, et al., SpeechBrain: A general-purpose speech toolkit, 2021, arXiv preprint arXiv:2106.04624.
- [44] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
- [45] Y.-H.H. Tsai, P.P. Liang, A. Zadeh, L.-P. Morency, R. Salakhutdinov, Learning factorized multimodal representations, in: *International Conference on Representation Learning*, 2019.



Yifeng Wang received the Ph.D. degree in computer science and technology from Xidian University, Xi'an, China, in 2009. He is currently an Associate Professor with the School of Computer Science and Technology, Xidian University. His research interests focus on machine learning and computer vision.



Jiahao He received the B.S. degree in intelligent science and technology from Central South University, Changsha, China, in 2021. He is currently pursuing the M.S. degree at Xidian University. His research interests focus on machine learning and computer vision.



Di Wang received the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2016. She is currently an Associate Professor in the School of Computer Science and Technology at Xidian University. Her research interests include machine learning and multimedia information retrieval. In these areas, she has published several scientific articles in refereed journals including the IEEE TPAMI, TIP, TMM, TCYB and TCSVT, and conferences including the SIGIR and IJCAI.



Quan Wang received the B.Sc., M.Sc., and Ph.D. degrees in computer science and technology from Xidian University, Xi'an, China. He is currently a Professor in the School of Computer Science and Technology at Xidian University. His current research interests include input and output technologies and systems, image processing and image understanding.



Bo Wan received the B.S., M.S., and Ph.D. degrees from Xidian University, Xi'an, Shaanxi, China. He is currently a Professor with the School of Computer Science and Technology, Xidian University. His current research interests include input/output technologies and systems, human computer interaction, and cloud computing.



Xuemei Luo received the Ph.D. degree in computer system structure from Xidian University, Xi'an, China, in 2012. She is currently a Lecturer with the School of Computer Science and Technology, Xidian University. Her research interests include color management, graphics and image processing, and machine learning.