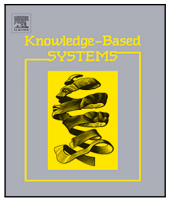
Knowledge-Based Systems 285 (2024) 111346

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis

Jiehui Huang [a], Jun Zhou [a], Zhenchao Tang [a], Jiaying Lin [a], Calvin Yu-Chian Chen [a,b,c,d,e,*]

[a] Artificial Intelligence Medical Research Center School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518107, Guangdong, China
[b] AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, 518055, Guangdong, China
[c] School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen, 518055, Guangdong, China
[d] Department of Medical Research, China Medical University Hospital, Taichung, 40447, Taiwan
[e] Department of Bioinformatics and Medical Engineering, Asia University, Taichung, 41354, Taiwan

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Transformer
Modality binding learning
Modality-invariant
Multimodal fusion

ABSTRACT

Multimodal emotion analysis is an important endeavor in human–computer interaction research, as it enables the accurate identification of an individual's emotional state by simultaneously analyzing text, video, and sound features. Although current emotion recognition algorithms have performed well using multimodal fusion strategies, two key challenges remain. The first challenge is the efficient extraction of modality-invariant and modality-specific features prior to fusion, which requires deep feature interactions between the different modalities. The second challenge concerns the ability to distinguish high-level semantic relations between modality features. To address these issues, we propose a new modality-binding learning framework and redesign the internal structure of the transformer model. Our proposed modality binding learning model addresses the first challenge by incorporating bimodal and trimodal binding mechanisms. These mechanisms handle modality-specific and modality-invariant features, respectively, and facilitate cross-modality interactions. Furthermore, we enhance feature interactions by introducing fine-grained convolution modules in the feedforward and attention layers of the transformer structure. To address the second issue, we introduce CLS and PE feature vectors for modality-invariant and modality-specific features, respectively. We use similarity loss and dissimilarity loss to support model convergence. Experiments on the widely used MOSI and MOSEI datasets show that our proposed method outperforms state-of-the-art multimodal sentiment classification approaches, confirming its effectiveness and superiority. The source code can be found at https://github.com/JackAILab/TMBL.

1. Introduction

Recently, human–computer interaction has become increasingly important, and accurately recognizing emotions is a crucial aspect of this process. Thanks to the rapid development of the internet industry, emotion analysis has expanded to include not only single-modal data like text [1,2], but also multimodal data such as video, audio, and text. This shift focuses on gaining a more accurate understanding of human emotions [3]. Multimodal data, such as video, audio, and text, provides a more accurate analysis of emotional states due to the complementary information between multiple data modalities. Therefore, the key to multimodal emotion analysis tasks is designing effective multimodal fusion approaches to integrate various data modalities [4]. Maintaining consistency and differential information for each modality is crucial for multimodal emotion analysis models [5].

First, to effectively integrate multiple modalities of data, previous research efforts have focused on using early fusion methods to combine different modal features [6–9]. However, these early fusion methods fail to capture fine-grained information between features [10]. Therefore, researchers have proposed several other approaches to fuse element-level features, including methods based on recurrent neural networks [11,12], attention-based methods [12–14], and multimodal perceptual word embeddings [15,16]. An important reason for the success of these methods is their effective exploration of fine-grained interactions between features and thorough exploration of correlations between different modalities. However, these methods are limited by

* Corresponding author at: Artificial Intelligence Medical Research Center School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518107, Guangdong, China.
E-mail address: cy@pku.edu.cn (C.Y.-C. Chen).

https://doi.org/10.1016/j.knosys.2023.111346
Received 22 August 2023; Received in revised form 22 November 2023; Accepted 26 December 2023
Available online 28 December 2023

their specific structural designs, which require domain expertise and lack strong generalizability in practical applications [17,18].

The Transformer model has achieved great success with different modal data and has shown strong potential in different multimodal tasks [19–21]. Several Transformer-based models have been attempted to facilitate better attentional interaction across modalities [18,22–25]. Existing transformer-based models focus on the textual modality and consider the correlation between textual-based modalities, [18,26], or they use multiple parallel structures to incorporate pairwise interactions between the three modalities, [22,23]. However, few models consider the coupling of all three modalities simultaneously, which may be a more promising approach for modal fusion. In addition, existing Transformer models for multimodal emotion analysis often overlook fine-grained feature exploration, which limits the effectiveness of modality data interaction.

Second, preserving the differences between modalities and exploring the similarities between modalities are key challenges in multimodal sentiment analysis [23,27–30]. To capture modality-specific features, separate network structures are generally used to extract their respective features [8,28,31], accompanied by the design of corresponding dissimilarity loss functions. To capture modality-invariant features, multiple modalities are typically input into a shared network structure [28,32], accompanied by the design of corresponding similarity loss functions. However, existing models mostly focus on extracting modality-invariant features, such as modality polarity [32] or consistency in temporal distribution [23], without fully considering the importance of modality-specific features. Moreover, even when a model considers both modality-invariant and modality-specific features, there is often insufficient distinction between them, which may result in the model leaning towards one modality to gain more information, thus undermining the robustness of the model.

Therefore, to address the above challenges more effectively, this paper proposes an optimized Transformer model for stable learning of multimodal emotion features. Specifically, we first design a Transformer network structure with finer-grained feed-forward and attention layers. In addition, this Transformer model can adaptively adapt to the number of input modalities, allowing it to learn up to three modal data types simultaneously. Second, inspired by the CLIP model [19], we develop a modality binding mechanism that efficiently combines the shared modality features, which are then collectively input into the Transformer model to obtain modality-invariant feature representations. Third, to avoid confusion between modality-invariant and modality-specific features during the learning process, we introduce classification tokens and position tokens into the feature data before entering the modality fusion module to ensure clarity. We jointly train the entire model using similarity, dissimilarity, and prediction loss functions. In summary, our contributions can be summarized as follows:

(1) In this study, we develop a binding learning mechanism to facilitate cross-modal feature interaction. Specifically, we design a Transformer model capable of handling bimodal and trimodal features, enabling it to capture and utilize bound modal features more effectively.

(2) In our model design work, we also introduce an optimized transformer structure that can effectively filter out inconsistent noise between bonded modes. This is achieved by incorporating fine-grained convolution operations into the attention and feed-forward layers of the model.

(3) To distinguish modality-invariant features from modality-specific features and improve the model's understanding of high-level semantic relationships between modalities, we introduce CLS and PE feature vectors in the modality fusion process. Furthermore, experiments on mainstream datasets show that our proposed model achieves state-of-the-art performance, and the visualization results further verify the effectiveness of the modal fusion mechanism.

2. Related work

In this paper, we aim to improve the Transformer model-based multimodal sentiment analysis system and give some better insights into modality fusion. The research focuses on how to simultaneously conduct interactions across the three modalities and achieve efficient modality fusion. In this section, we discuss related Transformer models and multimodal fusion schemes.

2.1. Multimodal sentiment analysis

With the rapid development of smart devices, a large amount of user data has been accumulated on social media, and the research on multimodal sentiment analysis has become more and more in-depth [33]. Previous multimodal sentiment analysis methods mainly focus on early fusion and late fusion. Early fusion first extracts the features of each mode separately, builds their joint representation and then uses a classifier to classify the joint representation for sentiment. For example, Poria et al. [7] proposed a parallelizable decision-level data fusion method, after extracting trimodal features, these combined feature vectors were trained using a multi-kernel learning (MKL) algorithm. Yu et al. [8] proposed a self-supervised multi-task multi-modal (Self-MM) sentiment analysis network, which additionally designed a label generation module using a self-supervised learning strategy. Then three independent unimodal tasks are additionally introduced on top of the earlier fusion methods to better learn the consistency and difference of the modalities. In the later stage of fusion, the final prediction result is obtained by means of weighted average or majority voting on the analysis results of each mode. For example, Zadeh et al. [9] propose a Multi-Attention Recurrent Network (MARN) that takes a majority vote for classification tasks and predicts the expected label for regression tasks. Mao [34] introduced the interactive platform Robust-MSA, enabling users to observe the effect of modal noise and suggest defensive measures. This can assist researchers in their analysis of model performance on real-world datasets more effectively.

However, low-level interactions between different modalities are ignored since both early fusion and late fusion cannot fully explore cross-view dynamic interactions between modalities [1]. Therefore, researchers have proposed many fine-grained fusion schemes, including tensor-based fusion [35,36], word-level fusion [9,15], translation-based fusion [21,37,38], context-based fusion [39,40], quantum-based fusion [41], and feature space manipulation-based fusion [28]. Notably, these feature fusion schemes aim to promote better interactions between modalities. Due to the successful application of the Transformer model in the multimodal field, it is considered to be a very potential multimodal fusion framework. This paper starts from the fusion scheme based on the function of space operation and uses an optimized Transformer network structure, which can effectively improve the robustness of the multimodal analysis system.

2.2. Transformer-based multimodal interactions

The Transformer model has achieved great success in both the text domain and the visual domain, and more and more unified Transformer models have recently appeared for their respective downstream tasks [19,42,43]. The Transformer-based self-attention mechanism can capture more global information than traditional attention methods, and it can naturally capture the interaction between multimodal data [23]. In the field of multimodal sentiment analysis, the Transformer model also shows great potential. For example, Wang et al. [44] proposed a Transformer model based on modality translation, which can efficiently perform end-to-end fusion between modalities. Yu et al. [43] proposed a Hierarchical Interactive Multimodal Transformer (HIMT) model to extract salient features with semantic concepts from images via object detection methods. Features extracted from images eliminate the semantic gap between text and image

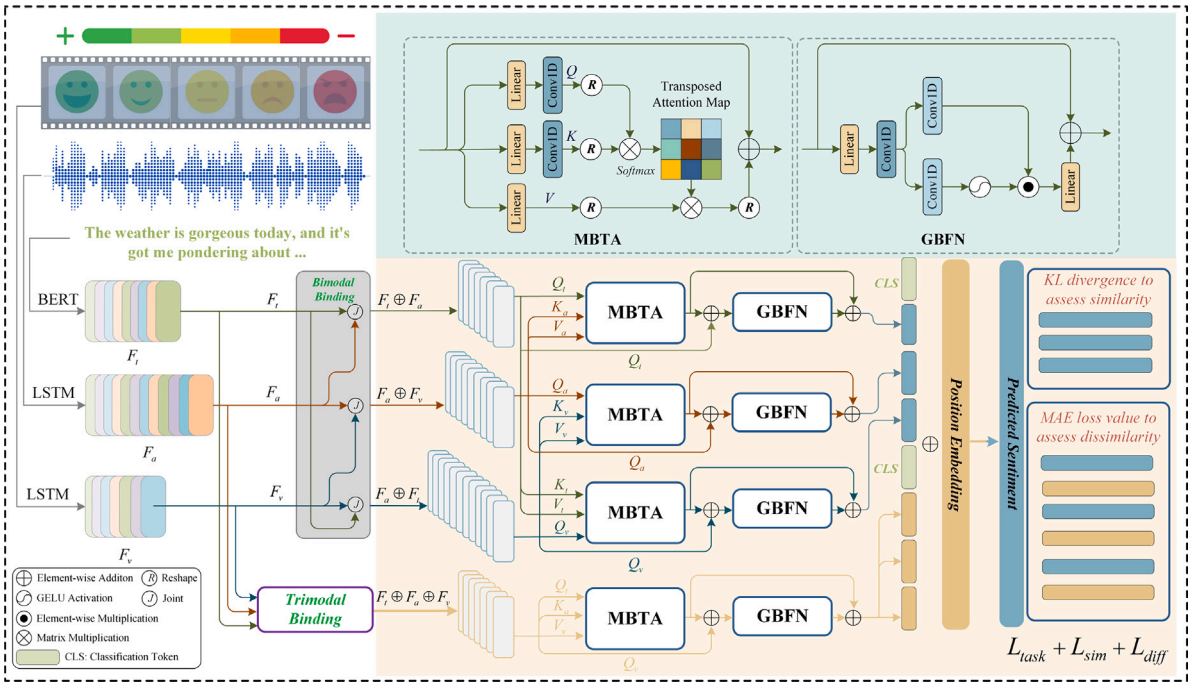


Fig. 1. The whole framework of our proposed TMBL, where the optimized transformer is composed of MBTA and GBFN. The predicted features will be distinguished by classification tokens and the modality positions are learnable.

representations, effectively accomplishing the image-text interaction. Sun et al. [22] proposed an efficient multimodal Transformer with double-layer feature restoration (EMT-DLFR), aiming at solving the missing modality problem in multimodal sentiment analysis tasks. In the literature [23], the researchers proposed a Transformer model based on the gated inter-modal attention mechanism and used a time-invariant feature fusion method based on spatial manipulation. The model effectively filters the heterogeneity between modalities and uses KL divergence to constrain the similarity of the modalities. Wang et al. [18] took the text as the main modality and proposed a new method called Text Enhanced Transformer Fusion Network (TETFN).

The above-mentioned Transformer-based models have achieved excellent performance. However, most of the models are designed to more effectively interact with the two modalities [18,23,43] while the interaction of the three modalities is rarely considered. Besides, the existing Transformer-based model is not sufficient for the extraction of each mode feature in one framework. The Transformer model proposed in this paper can extract three modal features at the same time, and we have also modified the feedforward layer and attention layer of the Transformer model so that the model can extract more fine-grained modal features.

2.3. Multimodal fusion strategies

In related work on multimodal sentiment analysis, it is shown that the fusion of different modalities is a key issue in multimodal sentiment analysis [45,46]. Among them, feature fusion based on spatial manipulation [28] is a promising fusion method, which focuses on mapping features to feature space after feature extraction, and learning the relationship between features through a series of mathematical analyses or operations [1].

For example, Hardoon [47] et al. propose a general method for learning semantic representations of web images and their associated texts using kernel canonical correlation analysis. Hazarika et al. [28] project each modality to a modality-invariant and modality-specific subspace, effectively accomplishing modality fusion. Mai et al. [48] performed both intra-modal/inter-modal contrastive learning and semi-contrastive learning, effectively exploring cross-modal interactions in the feature space and reducing the modality gap. Yue [30] presented the Knowledge Fusion Network, which efficiently employs the ConceptNet knowledge base to amalgamate prior knowledge and incorporates similarity analysis of both text and image modalities to enhance the sarcasm detection model's performance.

However, existing modality fusion methods are insufficient for modality-invariant and modality-specific considerations, and few consider the semantic order among different modality features. Interestingly, a good modality order can help the model better understand the meaning of the extracted features, thereby improving the robustness of the model.

3. Proposed model

As shown in Fig. 1, our proposed model framework first classifies the modality features into two types, modality-invariant and modality-specific. Note that modality-specific features refer to the inherent feature information of each modality data itself, and modality-invariant features refer to feature information shared by all modalities, which usually requires projecting the modality data on the same hidden layer dimension first. Specifically, for modality-specific feature data, we use the cross-modal strategy to complete the interaction of modality-specific features, and for modality-invariant features, we design a modal binding learning strategy to complete the interaction between modalities. Secondly, we use an optimized All-in-One Transformer architecture to better capture the characteristics between these interaction modalities. The optimized Transformer is composed of **M**ulti-head **B**inding **T**ransposed **A**ttention (MBTA) and the **G**ated **B**inding **F**eed-Forward **N**etwork (GBFN). Thirdly, we have implemented the inclusion of a CLS Token and Position Embedding to enhance the model's capability for identifying modality types and contexts. As a result, the loss function for the modality ensemble has been separated into three parts, including the classification loss for emotion prediction, modality-invariant similarity loss, and modality-specific dissimilarity loss.

3.1. Feature extraction and binding strategies

We define multimodal utterance data as U, which includes video, text, and acoustic as U_v, U_t, and U_a, respectively. Here, $\mathbf{U}_v \in \mathbb{R}^{L_v \times d_v}$, $\mathbf{U}_t \in \mathbb{R}^{L_t \times d_t}$, and $\mathbf{U}_a \in \mathbb{R}^{L_a \times d_a}$, L_u $(u = v, t, a)$ define the length of the utterance, d_u $(u = v, t, a)$ define the feature dimensions. To obtain an effective multimodal feature embedding, the video (U_v) and acoustic (U_a) components will undergo initial processing through COVAREP [49] and FACET [50] to extract $\mathbf{V}_t$ and $\mathbf{A}_t$ feature embeddings, respectively.

In order to better complete the feature fusion of the modality, we follow the feature extraction method in the literature [28]. Formally, for the modal feature embedding of video V_t and acoustic A_t will then be processed by vLSTM and aLSTM, respectively. In particular, the text (U_t) will be processed by pre-trained BERT [51] to directly obtain the textual feature F_t. Note that the data features extracted using LSTM can better extract data context features, and the pre-trained BERT used can effectively extract text features [28]. Specifically, video features and acoustic features are obtained by formula (1) and formula (2), respectively. Here, θ_{vLSTM} and θ_{aLSTM} represent the trainable LSTM parameters for video data and acoustic data. Text features are obtained by formula (3). Here θ_{bert} is the parameter of the pre-trained BERT model, and T_t represents the original text data. The Bert model used consists of 12 stacked Transformer layers [51].

$$\mathbf{F}_v = \mathrm{vLSTM}\left(V_t; \theta_{vLSTM}\right) \tag{1}$$

$$\mathbf{F}_a = \mathrm{aLSTM}\left(A_t; \theta_{aLSTM}\right) \tag{2}$$

$$\mathbf{F}_t = BERT\left(U_t; \theta_{bert}\right) \tag{3}$$

where $\mathbf{F}_v \in \mathbb{R}^{d_v}$, $\mathbf{F}_t \in \mathbb{R}^{d_t}$, and $\mathbf{F}_a \in \mathbb{R}^{d_a}$.

Additionally, to fully leverage the potential of intermodal interaction, we have developed binding strategies, namely Bimodal Binding and Trimodal Binding, for the extracted modal feature embeddings. (1) **Bimodal Binding**: Bimodal Binding combines non-aligned modalities that share modality-invariant features, feeding them into a redesigned Transformer for cross-modal learning. The detailed description of the Transformer structure will be provided in the subsequent section. (2) **Trimodal Binding**: in Trimodal Binding, as depicted in Fig. 2, we project both modalities onto the same hidden layer dimension. Subsequently, we employ a method akin to the CLIP model [19] to derive a cross-modal weight matrix A. However, unlike the CLIP model, the matrix $A = \mathrm{Pro}\left(F_v\right) @ \mathrm{Pro}\left(F_t\right)$ obtained in our approach strengthens the interconnection between modalities, rather than being directly utilized for the classification task. As shown in formula (4), we use the trimodal binding strategy to transform the modal features into a more closely related feature matrix F_v', F_t', F_a':

$$\begin{aligned}
\mathbf{F}_v' &= F_v * \mathrm{LN}\left(\mathrm{Pro}\left(F_v\right) @ \mathrm{Pro}\left(F_t\right)\right) \\
\mathbf{F}_t' &= F_t * \mathrm{LN}\left(\mathrm{Pro}\left(F_t\right) @ \mathrm{Pro}\left(F_v\right)\right) \\
\mathbf{F}_a' &= F_a * \mathrm{LN}\left(\mathrm{Pro}\left(F_a\right) @ \mathrm{Pro}\left(F_t\right)\right)
\end{aligned} \tag{4}$$

where Pro represents the projection linear layer, @ represents matrix multiplication, LN represents layer normalization, $\mathbf{F}_v'$, $\mathbf{F}_t'$, and $\mathbf{F}_a' \in \mathbb{R}^{d_n}$.

3.2. Binding learning transformer module

In Fig. 1, our Transformer model comprises two main components: MBTA and GBFN. The MBTA module enhances the traditional Transformer model by incorporating an additional layer of convolution modules, enabling more precise feature capture. Moreover, to facilitate cross-modal attention, Q and V are derived from different modal features, leading to the creation of a cross-modal attention matrix. In the GBFN module, the features are divided into two distinct sets, and the modalities' disparities are captured by leveraging an activation function. This approach enables better integration of features from both modalities. The subsequent two subsections provide detailed explanations of the structure of these two enhanced Transformer modules.

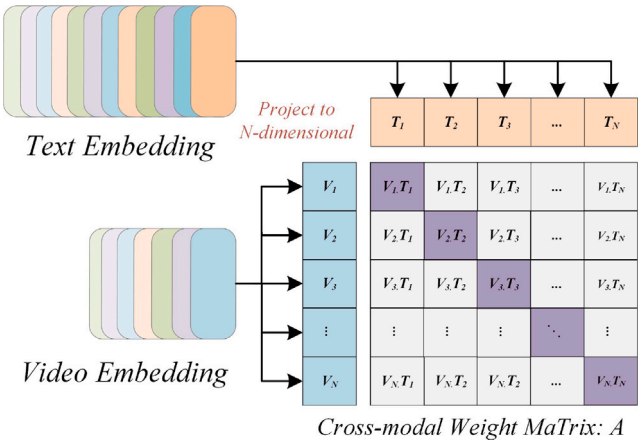


Fig. 2. Illustrate a three-modal binding strategy, exemplifying the use of text mode and video mode.

3.2.1. Multi-head binding transposed attention

We introduce a modality-bound self-attention layer to enhance our model's ability to encode the three distinct sources of information. To capture the order information of the input data, we incorporate position embeddings into the low-level representations of each modality using the formula (5). This enables our model to better comprehend the sequential nature of the input information.

$$h_u^0 = LN(F_u + PE_u) \tag{5}$$

where F_u and $PE_u \in \mathbb{R}^{d_u}$ represents the feature and positional embedding of each modality. LN represents layer normalization and h_u^0 denotes the modality's initial state within the Transformer network, which will then be passed on to the subsequent layers of the network for further processing.

For each modality, we employ a linear projection layer to obtain the corresponding queries (Q), keys (K), and values (V) within the multi-head attention layer. To illustrate this concept, we examine the scenario of bimodal text and sound binding learning, as exemplified in the formula (6).

$$\begin{aligned}
V_a &= \mathrm{h}_\mathrm{a}^0 W_{V_a} \\
Q_t &= \mathrm{Cov}_{qk}\left(\mathrm{h}_\mathrm{t}^0 W_{Q_t}\right) \\
K_a &= \mathrm{Cov}_{qk}\left(\mathrm{h}_\mathrm{a}^0 W_{K_a}\right)
\end{aligned} \tag{6}$$

where the dimensions of W_{Q_t}, W_{K_a}, and W_{V_a} are denoted as $\mathbb{R}^{d_t \times d_k}$, $\mathbb{R}^{d_a \times d_k}$, and $\mathbb{R}^{d_a \times d_g}$, respectively. Here, d_g represents the output dimension of the attention layer. The Cov qk convolutional layer contributes parameters to Q and K, with both its input and output channels set to d_k. Its primary objective is to facilitate enhanced information sharing within the same modality, thereby promoting more nuanced feature interaction.

It is important to note that in dual-modal binding learning, K and V are sourced from the same modality, while Q is sourced from a different modality. This approach, consistent with previous work [18], effectively facilitates cross-modal attention retrieval. However, in the case of three-modal binding learning, we redefine the modality relationship by incorporating Q, K, and V from all three modalities to enable a more comprehensive interaction between modal features. Subsequently, utilizing the formula (7), we finalize the information fusion process within the attention layer. $Z_{a \to t}$ denotes the interactive feature between the text and voice modes.

$$Z_{a \to t} = \mathrm{softmax}\left(\frac{Q_t K_a^T}{\sqrt{d_k}}\right) V_a \tag{7}$$

3.2.2. Gated Binding Feed-Forward Network

In contrast to prior research [23], we initially extract the interaction features Z, and input them to the feed-forward layer using the formula (8). This approach is motivated by the desire to enhance the model's discrimination ability, allowing it to discern information features characterized by high semantic density.

$$Z_1^i, Z_2^i = \mathrm{Cov}_{up}\left(Z^i W_{Z^i}\right) \tag{8}$$

where W_{Z^i} signifies the weight matrix of linear layer projection layer. Cov_{up} refers to an upsampling convolution operation that doubles the feature dimension of the interaction feature Z^i. Z_1^i and Z_1^i denote two interaction features obtained after expanding the feature dimension and subsequently separating them. $W_{Z^i} \in \mathbb{R}^{d_g \times d_h}$, $Z^i \in \mathbb{R}^{d_g}$, $Z_{1/2}^i \in \mathbb{R}^{d_h}$, i represents the number of stacked transformer layers.

Subsequently, we utilize the formula (9) to merge the two separated features Z, while employing a residual connection to facilitate the information transfer within the feed-forward layer.

$$O^i = Z^i + \left(\mathrm{ReLu}\left(Z_1^i\right) * Z_2^i\right) W_{O^i} \tag{9}$$

where O^i represents the output results of Transformer, $W_{O^i} \in \mathbb{R}^{d_h \times d_g}$, $O^i \in \mathbb{R}^{d_g}$.

3.2.3. Modal fusion

It is worth noting that the Transformer model incorporates both three-modal binding learning and dual-modal binding learning, which correspond to the modality-invariant feature O_I and modality-specific feature O_S, respectively. To enhance the model's comprehension of the sequential relationship between modalities and distinguish specific modalities from invariant ones, we introduce learnable parameters CLS and PE in the modality fusion section. As depicted in formula (10), we successfully integrate all modality features to obtain the final feature matrix O_F.

$$O_F = \left(O_I \oplus CLS \oplus O_S \oplus CLS\right) + PE \tag{10}$$

Lastly, we employ convolutional modules and apply average pooling to project all features onto their respective categorical dimensions.

3.3. Training strategy

Building upon [28], our proposed model incorporates three distinct losses. The task loss, denoted as formula (11), aims to enhance sentiment classification accuracy. The similarity loss, formulated as formula (12), facilitates the learning of modality-invariant features. Furthermore, the discriminative loss, as defined by the formula (13), promotes the differentiation of modality-specific features.

$$L_{\text{task}} = \frac{1}{N}\sum_{i=1}^{N} \| \text{ pred }^i - y^i \|_2^2 \tag{11}$$

$$L_{\text{sim}} = \frac{1}{3}\sum_{\substack{(m_1,m_2)\in \\ \{(l,a),(l,v) \\ ,(a,v)\}}} \left\{1 - \text{sim}\left(O_{m_1}^I, O_{m_2}^I\right)\right\} \tag{12}$$

$$L_{\text{diff}} = \sum_{\substack{m\in \\ \{l,v,a\}}} \left\| O_m^{I^T} O_m^S \right\|_F^2 + \sum_{\substack{(m_1,m_2)\in \\ \{(l,a),(l,v) \\ ,(a,v)\}}} \left\| O_{m_1}^{S^T} O_{m_2}^S \right\|_F^2 \tag{13}$$

where pred ^{i} denotes the sentiment strength predicted by the model, while y^i corresponds to the true label. O_m^I refers to a feature that remains invariant across different modalities, while O_m^S represents a modality-specific feature. The notation $\|.\|_F^2$ represents the squared Frobenius norm, and sim represents the cosine similarity function. Hence, the overall loss L_{total} of the model is defined by Eq. (14):

$$L_{\text{total}} = L_{\text{task}} + \lambda_1 L_{\text{sim}} + \lambda_2 L_{\text{diff}} \tag{14}$$

Table 1
Dataset statistics in MOSI and MOSEI.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16 326	1871	4659	22 856

4. ExperimentResult

In our experimental evaluations, we utilize two publicly available datasets: CMU-MOSI [52] and CMU-MOSEI [53]. These datasets are shown in Table 1 and notable for providing word-aligned multimodal signals encompassing linguistic, visual, and auditory modalities for each utterance.

4.1. Experment setting

4.1.1. Datasets

The CMU-MOSI dataset, known as the Multimodal Corpus of Sentiment Intensity, is a widely used benchmark in multimodal sentiment analysis. It comprises 2199 viewpoint video clips, where each clip is annotated with a sentiment score ranging from −3 to 3, representing strong negative to positive sentiment. The dataset includes comprehensive annotations for subjectivity, sentiment strength, visual features at the frame and clip level, as well as audio features at the millisecond level.

The CMU-MOSEI dataset is the largest dataset for multimodal sentiment analysis and emotion recognition. It comprises over 23,500 sentence utterance videos collected from more than 1000 YouTube speakers. This dataset ensures gender balance and includes sentences randomly selected from diverse topics and monologue videos. The videos have been transcribed and accurately punctuated. The MOSEI dataset builds upon the MOSI dataset by offering increased diversity in terms of utterances, samples, speakers, and topics (see Table 1).

4.1.2. Evaluation metrics

We have designed two evaluation tasks: regression and classification. For regression tasks, we report the mean absolute error (MAE) and the Pearson correlation coefficient (Corr) as performance metrics. Regarding classification tasks, we employ the F1-Score, binary accuracy (Acc-2), and seven-level accuracy (Acc7) as evaluation measures. To comprehensively assess the robustness of the models, we compute Acc-2 and F1-Score in two ways: negative/nonnegative (including zero values) [35] and negative/positive (excluding zero values) [6]. To represent these two metrics, we use the segment notation −/−, where the left score denotes the negation/non-negation condition, and the right scores correspond to negative/positive classifications. Note that a single Acc metric covers two evaluation categories (negative/non-negative and negative/positive), and the latter category represents the model's maximum performance capacity. Therefore, the method that incorporates the optimal negative/positive data values is considered the most effective approach. Except for MAE, higher values indicate better performance.

4.1.3. Implementation details

Our models are trained using the Adam optimizer with a model learning rate of 1×10^{-4}, the total training loss function $\lambda_1 = \frac{1}{2}$, $\lambda_2 = \frac{1}{2}$. The embedding size of the BERT model is set to 300, and the hidden layer dh of the modality map in Transformer is set to 256. For classification and regression tasks, the learning rate is multiplied by 0.5 after every 50 training epochs. Our network is implemented using the PyTorch framework and a single RTX 3090 GPU.

4.2. Baselines

Graph-MFN [53]. This method introduces the Dynamic Fusion Graph, a novel technique for multimodal fusion based on graph fusion modeling. **RAVEN** [54]. The Recurrent Participation Variation Embedding Network models the fine-grained structure of nonlinguistic subword sequences and dynamically transfers word representations based on nonlinguistic cues. **MCTN** [37]. This approach proposes Low-rank Multimodal Fusion as a method for learning robust joint representations through transitions between modalities. **CIA** [55]. CIA learns intermodal interactions between participating modalities via an autoencoder mechanism. It employs a context-aware attention module to exploit the correspondence between adjacent utterances. **MulT** [21]. MulT introduces a multimodal transformer architecture that uses directed pairwise cross-attention to focus on interactions between multimodal sequences at different time steps, facilitating the transformation of one modality into another state. **TFN** [35]. TFN formulates the multimodal sentiment analysis problem as modeling intra- and inter-modal dynamics. **LMF** [36]: An improvement on TFN, LMF proposes low-rank modeling of TFN tensors. **MFM** [56]. MFM learns discriminative and generative representations for each modality, utilizing the former for classification and the latter to learn modality-specific generative features. **ICCN** [57]. Interaction Canonical Correlation Network improves multimodal sentiment analysis by learning the correlation between all three modalities through Deep Canonical Correlation Analysis. **MISA** [28]. The model projects representations into modality-specific and modality-invariant spaces, learning distribution similarity, orthogonality loss, reconstruction loss, and task prediction loss. **MMIM** [58]. MMIM hierarchically maximizes the mutual information between unimodal input pairs, multimodal fusion results, and unimodal inputs to maintain task-related information through multimodal fusion. **HyCon** [48]. HyCon performs intra-/inter-modal contrastive learning and semi-contrastive learning simultaneously. **TETFN** [18]. Text-Enhanced Transformer Fusion Network learns text-oriented pairwise cross-modal mappings to efficiently preserve inter- and intra-modal relations. **BC-LSTM** [39]. This LSTM-based model enables utterances to capture contextual information from their surroundings in the same video, thus facilitating the classification process. **MV-LSTM** [59]. MV-LSTM employs a multi-view LSTM to explicitly capture both view-specific and cross-view interactions over time and structured output. **MAG-Bert** [60]. MAG-Bert integrates aligned non-linguistic information into Bert's textual representations through Bert's Multimodal Adaptation Gate.

4.3. Quantitative results

The results of multimodal sentiment analysis on the CMU-MOSEI dataset are presented in Table 2. We analyze the advantages of our model from three perspectives. Firstly, we compare our model with Transformer-based models, such as MulT and TETFN, and find that our proposed model outperforms them in various indicators. Specifically, the TETFN model achieves impressive results in A2 and MAE, which are comparable to our own performance. However, the performance of the TETFN model on Corr and F1 index is not ideal. By contrast, our three-modal binding learning approach, utilizing the Transformer model framework, obtains impressive outcomes on the majority of metrics. TETFN outperforms us slightly on the A2 metric in differentiating negative from non-negative sentiment. However, TETFN's lower performance on other metrics highlights its limitations. This could be due to TETFN's text-based learning approach, which relies heavier on the quality of the text in the dataset for model diagnosis, making this text-centered, dual-modal learning approach less efficient. In contrast, our model uses tri-modal binding learning and is anticipated to produce more resilient outcomes. Additionally, MulT neglects finer-grained modal interactions, leading to suboptimal performance.

Secondly, we compare our model to other models utilizing modality-invariant and modality-specific learning strategies, including CIA, TFN, MFM, MISA, and HyCon. Our model achieves state-of-the-art (SOTA) performance on most metrics, while MISA and HyCon exhibit unstable performance, highlighting the superiority of our unified Transformer model structure.

Table 2
Performance comparison between TMBL and previous models on CMU-MOSEI dataset. The top three results are highlighted in boldface, and the best results are additionally underlined. In Acc-2 and F1-Score, the left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive".

Model	A2 (↑)	A7 (↑)	MAE (↓)	Corr (↑)	F1 (↑)
$Graph-MFN$[a] [53]	76.9/–	45.0	0.710	0.540	77.0/–
$RAVEN$[a] [54]	79.1/–	50.0	0.614	0.662	79.5/–
$MCTN$[a] [37]	79.8/–	49.6	0.609	0.670	80.6/–
CIA[a] [55]	80.4/–	50.1	0.680	0.590	78.2/–
$MulT$[a] [21]	–/82.5	51.8	0.580	0.703	–/82.3
TFN[a] [35]	–/82.5	50.2	0.593	0.700	–/82.1
LMF[a] [36]	–/82.0	48.0	0.623	0.677	–/82.1
MFM[a] [56]	–/84.4	51.3	0.568	0.717	–/84.3
$ICCN$[a] [57]	–/84.2	51.6	0.565	0.713	–/84.2
$MISA$[a] [28]	**83.6/85.5**	52.2	0.555	0.756	**83.8/85.3**
$MMIM$[a] [58]	82.2/86.0	**54.2**	**0.526**	**0.772**	82.7/85.9
$HyCon$[a] [48]	–/85.4	**52.8**	0.601	**0.778**	–/85.6
$TETFN$[b] [18]	**84.25/85.18**	–	**0.551**	0.748	**84.18/85.27**
Ours	**84.23/85.84**	**52.4**	**0.545**	**0.766**	**84.87/85.92**

[a] Means the results provided by AOBert [29].
[b] From TETFN [18].

Finally, we compare our model to various baseline mode fusion methods. Although MMIM performs excellently on the A7 and MAE indicators, other mode fusion methods fail to showcase their advantages effectively. In contrast, our model achieves more stable and discriminative performance on the A2 and F1 indicators, demonstrating its ability to fully extract modal features.

Next, we evaluate the model performance on the MOSI dataset, as shown in Table 3. Our models continue to achieve state-of-the-art performance on the F1 index and also demonstrate competitive performance on the A2 index. Notably, our model maintains a high level of accuracy in judging "negative/non-negative" and "negative/positive" sentiments on both datasets, as indicated by our F1 and A2 indicators, confirming its capability to capture fine-grained features. The proposed Transformer framework has increased potential, effectively enhancing the generalization capability of the Transformer model in the domain of multi-modal sentiment analysis. Despite being inferior to TETFN, our primary objective is to achieve outstanding robustness, and as such our model can provide effective A7 performance. Notably, the TETFN model performs considerably weaker on the MOSEI dataset than on the MOSI dataset. This may suggest overfitting of the TETFN model on the MOSI dataset. Expanding the MOSEI data samples does not offer any benefits to the TETFN model. It indicates that the TETFN model attempts to enhance performance by sacrificing robustness. Conversely, the TMBL model developed in this study uses modality binding learning, which results in more durable performance improvement.

Furthermore, we run the proposed method and the two state-of-the-art methods five times under different random number seeds and take the average as the final result. This ensures a more robust evaluation and comparison of our model's performance against existing methods.

4.4. Ablation study

4.4.1. Role of fine-grained design

In Table 4, we conduct experiments by removing the MBTA and GBFN modules individually to assess their impact on performance. The "Optimized Transformer" refers to our model with both MBTA and GBFN modules, while "Transformer" denotes the optimized Transformer without these additional modules. Firstly, we observe that the combined use of MBTA and GBFN modules yields the best performance, suggesting that these two modules complement each other and enhance

Table 3
Performance comparison between TMBL and previous models on CMU-MOSI dataset. The top three results are highlighted in boldface, and the best results are additionally underlined. In Acc-2 and F1-Score, the left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive".

Model	A2 (↑)	A7 (↑)	MAE (↓)	Corr (↑)	F1 (↑)
$BC-LSTM$[a] [39]	73.9/–	28.7	1.079	0.581	73.9/–
$MV-LSTM$[a] [59]	73.9/–	33.2	1.019	0.601	74.0/–
$RAVEN$[a] [54]	76.6/–	33.2	0.915	0.691	76.6/–
$MCTN$[a] [37]	79.1/–	35.6	0.909	0.676	79.1/–
CIA[a] [55]	79.8/–	38.9	0.914	0.689	–/79.5
$MulT$[a] [21]	–/83.0	**40.0**	0.871	0.698	–/82.8
TFN[a] [35]	73.9/–	32.1	0.970	0.633	73.4/–
LMF[a] [36]	76.4/–	32.8	0.912	0.668	75.7/–
MFM[a] [56]	78.1/–	36.2	0.951	0.662	78.1/–
$ICCN$[a] [57]	–/83.0	**39.0**	0.862	0.714	–/83.0
$MISA$[a] [28]	81.8/83.4	**42.3**	**0.783**	0.761	81.7/83.6
$MAG-Bert$[b] [60]	**82.42/84.15**	–	**0.734**	**0.789**	**82.45/84.13**
$TETFN$[b] [18]	**84.05/86.10**	–	**0.717**	**0.800**	**83.83/86.07**
Ours	**81.78/83.84**	36.3	0.867	**0.762**	**82.41/84.29**

[a] Means the results provided by AOBert [29].

[b] From TETFN [18].

Table 4
Affection of fine-grained design transformer on MOSEI.

	A2	A7	F1
Optimized transformer	**84.23/85.84**	**52.4**	**84.87/85.92**
MBTA (–)	82.97/84.35	50.6	83.56/85.22
GBFN (–)	84.10/85.20	50.1	84.26/85.21
Transformer	82.10/84.20	49.3	82.62/84.33

Table 5
Affection of modality binding strategy.

	A2	A7	F1
Modality binding	**84.23/85.84**	**52.4**	**84.87/85.92**
Bimodal binding (–)	83.67/85.15	50.7	83.67/85.30
Trimodal binding (–)	84.21/85.32	51.1	84.32/85.42
Modality Free	79.12/81.42	48.1	80.17/82.63

the overall model performance. Secondly, we note that the removal of the MBTA structure has the most significant impact on the Transformer model, indicating that the attention mechanism's design effectively improves the feature interaction among modalities during multimodal fusion.

4.4.2. Role of modality binding

Table 5 presents a comparison of the effects of bimodal and trimodal binding mechanisms. "Modality Free" refers to the scenario where all modalities are independently processed by the Transformer, and interaction occurs only during feature fusion. It is evident that when the modalities are not bound, the model's performance experiences a significant drop, particularly in A7, where it only reaches 48.1%. This underscores the crucial role of interaction during modal feature extraction. On the other hand, our designed bimodal and trimodal mechanisms effectively facilitate deep-level interactions among the modalities, leading to improved performance.

4.4.3. Role of CLS and PE

To assess the significance of the modal order relationship, we conducted an analysis of the role of CLS (Classification Token) and PE (Positional Encoding). As shown in Table 6, the inclusion of PE leads to improved performance in A2, improving the model's ability to discriminate between negative/non-negative and negative/positive emotions. This observation highlights the sensitivity of the model in emotion recognition. On the other hand, the inclusion of CLS significantly improves the performance of A7 and significantly improves the indicators of A2 and F1. Furthermore, when both CLS and PE are added to the model at the same time, the performance of the model is further

Table 6
Affection of modality position learning strategy.

	A2	A7	F1
Modality distinction	**84.23/85.84**	**52.4**	**84.87/85.92**
CLS (–)	84.31/84.90	50.0	84.00/85.12
PE (–)	84.23/85.70	51.4	83.78/85.80
No distinction	81.32/84.67	49.4	81.67/84.53

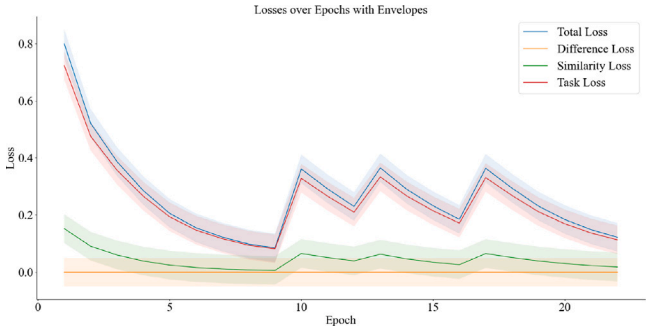


Fig. 3. During the training process, the loss value changes and comprises several components: difference loss, similarity loss, classification loss, and total loss.

Table 7
Comparation of different loss function.

	A2	A7	F1
SIM+Diff+TASK	**84.23/85.84**	**52.4**	**84.87/85.92**
SIM (–)	84.49/84.87	50.4	84.53/84.96
Diff (–)	84.08/85.15	49.8	84.19/85.25

improved. This convincingly demonstrates the effectiveness of the CLS and PE structures we have designed, as they improve the model's understanding of the order of the modalities, resulting in more refined emotion recognition capabilities.

4.4.4. Role of learning strategy

To investigate the efficacy of the three designed loss functions and quantify the extent to which the model learns modality invariance and specific representation, we visualized the changes in the loss function values during the training process. Fig. 3 illustrates the task loss, which corresponds to the model's classification loss (Eq. (13)), showing a consistent changing trend with the overall loss of the model. Over the course of training, both the similarity loss and difference loss gradually tend towards 0, affirming the effectiveness of our designed modality binding.

Furthermore, Table 7 presents the verification of the impact of similarity loss and difference loss on the model. It is evident that the promotion effect of similarity loss surpasses that of difference loss, which aligns with the findings in Fig. 3. Specifically, in the early stages of training, the similarity of the three modal features is more pronounced, while the difference features are relatively smaller. The difference loss effectively aids the model in capturing fine-grained distinctions between the modal features. As a result, the performance gap between negative/non-negative and negative/positive in A2 and F1 is only 0.38% and 1.02%, respectively.

4.4.5. Visualizing representations

Fig. 3 illustrates the performance of the regularization loss during training. However, equally important is the assessment of the generalization capability of these features. To achieve this, we performed t-SNE visualizations of both modality-invariant and modality-specific features using samples from the test set. As shown in Fig. 4, our modal binding mechanism efficiently aligns all modalities with the text modality, thereby reducing the semantic gap between them. This observation

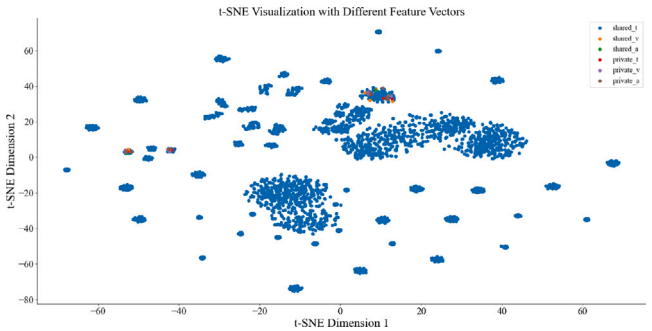


Fig. 4. Visualization of modality-invariant and subspace-specific features in the test set of the MOSEI dataset using t-SNE projections.

reaffirms the importance of the textual modality, as emphasized in prior studies [18,29]. Furthermore, it provides additional evidence of the success of our modality-bound learning model.

5. Conclusion

In this study, we present a novel multi-modal learning framework that utilizes a Transformer structure to facilitate simultaneous feature interactions among three modalities. This framework effectively captures both modality-invariant and modality-specific features while uncovering their inherent relationships for binding learning. To mitigate noise and establish modality order, we introduce CLS and PE learning strategies. Additionally, we enhance the Transformer structure by incorporating convolutional networks in the attention and feed-forward layers, enabling the model to capture finer-grained features. Furthermore, we leverage similarity loss and difference loss to enhance the robustness of the model's representations.

Our proposed method outperforms existing approaches, achieving state-of-the-art results in multimodal sentiment classification tasks across multiple datasets. Feature visualization and ablation studies confirm the effectiveness and robustness of various components. We believe that our proposed structure and method hold potential for broader applications in other fields related to multimodal interaction and fusion.

The current limitations of the model mainly focus on room for improvement in the analysis performance of small sample data, and more analysis tasks need to be expanded to further test the effectiveness of modal binding learning. In the future, we will use the modality binding learning model for multi-modal disease diagnosis tasks and other multi-modal work such as text-to-image generation. In addition, we plan to explore hinting techniques combined with large language models for analyzing other sentiment analysis tasks such as sarcasm detection to further enhance the accuracy and generalization capabilities of the model.

CRediT authorship contribution statement

Jiehui Huang: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Data curation, Conceptualization. **Jun Zhou:** Writing – review & editing, Investigation, Methodology, Formal analysis. **Zhenchao Tang:** Writing – review & editing, Visualization, Supervision, Methodology. **Jiaying Lin:** Visualization, Validation, Investigation. **Calvin Yu-Chian Chen:** Writing – review & editing, Validation, Funding acquisition, Conceptualization.

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 by the National Natural Science Foundation of China (Grant No. 62176272), Research and Development Program of Guangzhou Science and Technology Bureau, China (No. 2023B01J1016), and Key-Area Research and Development Program of Guangdong Province, China (No. 2020B1111100001).

References

[1] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, Inf. Fusion 95 (2023) 306–325.
[2] R. Das, T.D. Singh, Multimodal sentiment analysis: A survey of methods, trends and challenges, ACM Comput. Surv. (2023).
[3] R. Kaur, S. Kautish, Multimodal sentiment analysis: A survey and comparison, in: Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines, IGI Global, 2022, pp. 1846–1870.
[4] 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.
[5] Q.-T. Truong, H.W. Lauw, Vistanet: Visual aspect attention network for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 305–312.
[6] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web, in: Proceedings of the 13th International Conference on Multimodal Interfaces, 2011, pp. 169–176.
[7] 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.
[8] 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, 2021, pp. 10790–10797.
[9] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
[10] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, D. Zhang, Multimodal emotion recognition with temporal and semantic consistency, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 3592–3603.
[11] M. Chen, S. Wang, P.P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 163–171.
[12] A. Aslam, A.B. Sargano, Z. Habib, Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks, Appl. Soft Comput. (2023) 110494.
[13] Y. Du, Y. Liu, Z. Peng, X. Jin, Gated attention fusion network for multimodal sentiment classification, Knowl.-Based Syst. 240 (2022) 108107.
[14] J. Tang, D. Liu, X. Jin, Y. Peng, Q. Zhao, Y. Ding, W. Kong, BAFN: Bi-direction attention based fusion network for multimodal sentiment analysis, IEEE Trans. Circuits Syst. Video Technol. (2022).
[15] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
[16] P.P. Liang, Z. Liu, A. Zadeh, L.-P. Morency, Multimodal language analysis with recurrent multistage fusion, 2018, arXiv preprint arXiv:1808.03920.
[17] Z. Yu, J. Wang, L.-C. Yu, X. Zhang, Dual-encoder transformers with cross-modal alignment for multimodal aspect-based sentiment analysis, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2022, pp. 414–423.
[18] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognit. 136 (2023) 109259.
[19] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[20] P. Xu, X. Zhu, D.A. Clifton, Multimodal learning with transformers: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[21] 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.
[22] L. Sun, Z. Lian, B. Liu, J. Tao, Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2023).
[23] H. Sun, J. Liu, Y.-W. Chen, L. Lin, Modality-invariant temporal representation learning for multimodal sentiment classification, Inf. Fusion 91 (2023) 504–514.
[24] F. Wang, S. Tian, L. Yu, J. Liu, J. Wang, K. Li, Y. Wang, TEDT: Transformer-based encoding–decoding translation network for multimodal sentiment analysis, Cogn. Comput. 15 (1) (2023) 289–303.
[25] H. Sun, Y.-W. Chen, L. Lin, TensorFormer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection, IEEE Trans. Affect. Comput. (2022).
[26] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, X. Gao, Cross-modal enhancement network for multimodal sentiment analysis, IEEE Trans. Multimed. (2022).
[27] S. Rahmani, S. Hosseini, R. Zall, M.R. Kangavari, S. Kamran, W. Hua, Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects, Knowl.-Based Syst. 261 (2023) 110219.
[28] 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.
[29] K. Kim, S. Park, AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis, Inf. Fusion 92 (2023) 37–45.
[30] T. Yue, R. Mao, H. Wang, Z. Hu, E. Cambria, KnowleNet: Knowledge fusion network for multimodal sarcasm detection, Inf. Fusion 100 (2023) 101921.
[31] A. Ando, R. Masumura, A. Takashima, S. Suzuki, N. Makishima, K. Suzuki, T. Moriya, T. Ashihara, H. Sato, On the use of modality-specific large-scale pre-trained encoders for multimodal sentiment analysis, in: 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2023, pp. 739–746.
[32] H. Lin, P. Zhang, J. Ling, Z. Yang, L.K. Lee, W. Liu, PS-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis, Inf. Process. Manage. 60 (2) (2023) 103229.
[33] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion (2022).
[34] H. Mao, B. Zhang, H. Xu, Z. Yuan, Y. Liu, Robust-MSA: Understanding the impact of modality noise on multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 16458–16460.
[35] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, 2017, arXiv preprint arXiv:1707.07250.
[36] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, 2018, arXiv preprint arXiv:1806.00064.
[37] H. Pham, P.P. Liang, T. Manzini, L.-P. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6892–6899.
[38] Q. Zhang, L. Shi, P. Liu, Z. Zhu, L. Xu, ICDN: integrating consistency and difference networks by transformer for multimodal sentiment analysis, Appl. Intell. (2022) 1–14.
[39] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 873–883.
[40] D. Ghosal, M.S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, P. Bhattacharyya, Contextual inter-modal attention for multi-modal sentiment analysis, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3454–3466.
[41] Y. Zhang, D. Song, X. Li, P. Zhang, P. Wang, L. Rong, G. Yu, B. Wang, A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis, Inf. Fusion 62 (2020) 14–31.
[42] R. Hu, A. Singh, Unit: Multimodal multitask learning with a unified transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1439–1449.
[43] J. Yu, K. Chen, R. Xia, Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022).
[44] Z. Wang, Z. Wan, X. Wan, Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis, in: Proceedings of the Web Conference 2020, 2020, pp. 2514–2520.
[45] S. Sahay, E. Okur, S.H. Kumar, L. Nachman, Low rank fusion based transformers for multimodal sequences, 2020, arXiv preprint arXiv:2007.02038.
[46] J. Ye, J. Zhou, J. Tian, R. Wang, J. Zhou, T. Gui, Q. Zhang, X. Huang, Sentiment-aware multimodal pre-training for multimodal sentiment analysis, Knowl.-Based Syst. 258 (2022) 110021.
[47] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[48] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022).
[49] 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.
[50] E.L. Rosenberg, P. Ekman, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS), Oxford University Press, 2020.
[51] J.D.M.-W.C. Kenton, L.K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NaacL-HLT, Vol. 1, 2019, p. 2.
[52] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88.
[53] 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.
[54] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.-P. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7216–7223.
[55] D.S. Chauhan, M.S. Akhtar, A. Ekbal, P. Bhattacharyya, Context-aware interactive attention for multi-modal sentiment and emotion analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5647–5657.
[56] Y.-H.H. Tsai, P.P. Liang, A. Zadeh, L.-P. Morency, R. Salakhutdinov, Learning factorized multimodal representations, 2018, arXiv preprint arXiv:1806.06176.
[57] 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, 2020, pp. 8992–8999.
[58] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, 2021, arXiv preprint arXiv:2109.00412.
[59] S.S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, R. Goecke, Extending long short-term memory for multi-view structured learning, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part VII 14, Springer, 2016, pp. 338–353.
[60] 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.