Information Fusion 101 (2024) 101973

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Modality translation-based multimodal sentiment analysis under uncertain missing modalities

Zhizhong Liu [a], Bin Zhou [a], Dianhui Chu [b,*], Yuhang Sun [a], Lingqiang Meng [a]

[a] The school of Computer and Control Engineering, Yantai University, Yantai, 264005, China
[b] College of Computer Science and Technology, Harbin Institute of Technology, Weihai, 264209, China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Uncertain missing modalities
Modality translation
Transformer

ABSTRACT

Multimodal sentiment analysis (MSA) with uncertain missing modalities poses a new challenge in sentiment analysis. To address this problem, efficient MSA models that consider missing modalities have been proposed. However, existing studies have only adopted the concatenation operation for feature fusion while ignoring the deep interactions between different modalities. Moreover, existing studies have failed to take advantage of the text modality, which can achieve better accuracy in sentiment analysis. To tackle the above-mentioned issues, we propose a modality translation-based MSA model (MTMSA), which is robust to uncertain missing modalities. First, for multimodal data (text, visual, and audio) with uncertain missing data, the visual and audio are translated to the text modality with a modality translation module, and then the translated visual modality, translated audio, and encoded text are fused into missing joint features (MJFs). Next, the MJFs are encoded by the transformer encoder module under the supervision of a pre-trained model (transformer-based modality translation network, TMTN), thus making the transformer encoder module produce joint features of uncertain missing modalities approximating those of complete modalities. The encoded MJFs are input into the transformer decoder module to learn the long-term dependencies between different modalities. Finally, sentiment classification is performed based on the outputs of the transformer encoder module. Extensive experiments were conducted on two popular benchmark datasets (CMU-MOSI and IEMOCAP), with the experimental results demonstrating that MTMSA outperforms eight representative baseline models.

1. Introduction

Sentiment analysis has been a popular research topic in machine learning and natural language processing for many years [1,2]. It aims to understand and interpret human sentiment through different modalities (e.g., text, voice tones, or facial expressions). Recently, automatic and accurate sentiment analysis has been shown to play a critical role in natural human–computer interactions [1,3], group decision-making systems [4], opinion mining [5], and decision making [6,7]. With the popularity of online video platforms (e.g., YouTube, Twitter, and Weibo), an increasing number of users are willing to express their emotions and opinions by posting videos. To effectively recognize the affective orientation of these videos, multimodal sentiment analysis (MSA) has been proposed and attracted increasing attention. For example, given a monologue video, the target of MSA is to detect the sentiment involved by leveraging multiple input modalities, including textual, auditory, and visual modalities [8].

Compared with single-modality data, multimodalities can represent different aspects of emotion and provide complementary information, which can significantly enhance the accuracy of sentiment analysis [9,10]. In the past few years, some effective MSA models have been proposed based on different techniques, such as recurrent neural networks [11], transformers [12,13], and graph convolutional neural networks [14,15]. Existing MAS research has achieved abundant results and promoted the rapid development of emotion recognition technology.

However, most MSA models assume that all modalities (textual, auditory, and visual) are always available [16]. Meanwhile, in real-world applications, uncertain missing modalities often occur due to some uncontrollable factors [17]. For example, as shown in Fig. 1, some visual information is missing when the camera is turned off or blocked; certain speech content cannot be obtained when the user is silent; or the voice and text are missing due to device errors. Therefore, the assumption that all modalities are always available cannot be held

* Corresponding author.
E-mail addresses: zhizhongliu@ytu.edu.cn (Z. Liu), 1073237135@s.ytu.edu.cn (B. Zhou), cdh@hitwh.edu.cn (D. Chu), 202200358047@s.ytu.edu.cn (Y. Sun), 1782356223@s.ytu.edu.cn (L. Meng).

https://doi.org/10.1016/j.inffus.2023.101973
Received 21 April 2023; Received in revised form 6 August 2023; Accepted 8 August 2023
Available online 12 August 2023

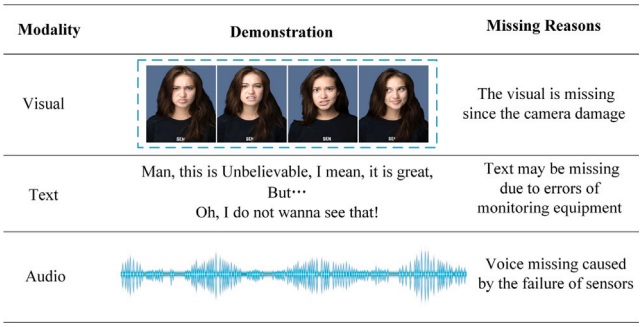


Fig. 1. Uncertain missing modalities in MSA (taking the missing visual modality as an example, which is marked with blue dashed boxes.).

in many real-world scenarios. Thus, most existing MSA models fail under uncertain missing modalities and solving the problem of MSA with uncertain missing modalities becomes a critical issue.

Recently, to address the aforementioned issues, studies on MSA with missing modalities have proposed several effective methods, which can be generally categorized into two groups: generative and joint learning methods. On the one hand, generative methods aim to generate new data that match the observed distributions. For example, Tran et al. [18] proposed a cascaded residual auto-encoder (CRA) that models the correlation between different modalities by stacking residual auto-encoders and then uses it to impute missing data. Cai et al. [19] designed a 3D encoder–decoder network to capture the relationship between different modalities and leveraged an auxiliary adversarial loss to make available modalities generate missing ones.

On the other hand, joint learning methods attempt to learn latent representations from observed ones. For example, Zhao et al. [20] proposed a unified model named Missing Modality Imagination Network (MMIN). MMIN learns robust joint multimodal representations and can predict the representation of any missing modality given the available modalities under different missing modality conditions. Zeng et al. [21] proposed a Tag-Assisted Transformer Encoder (TATE) network to address the problem of uncertain missing modalities. Although some remarkable methods have been proposed for MSA with missing modalities, they suffer from some shortcomings, which can be summarized as follows:

- Existing works just adopt the concatenation operation to implement feature fusion, which cannot consider deep interactions between features of different modalities.
- Existing works fail to fully utilize the text modality, which always achieves the best sentiment analysis accuracy among the three modalities.
- When dealing with uncertain missing modalities, existing works should consider all the situations of missing modalities and then handle each situation separately, which significantly increases the complexity of MSA.

To address the above-mentioned issues, this study proposes a modality translation network for MSA under uncertain missing modes (MTMSA). First, for multimodal data (textual, visual, and auditory) with uncertain missing modalities, the visual and auditory modalities are translated into the text modality using a modality translation module, while the text modality is encoded by the transformer encoder. The translated visual modality, translated audio, and encoded text are then fused into the missing joint features (MJFs). Next, the MJFs are encoded by the transformer encoder module with the supervision of a pre-trained model, transformer-based modality translation network (TMTN), which is trained with complete modalities (TMTN), thus enabling the transformer encoder module to produce joint features of uncertain missing modalities approximating those of complete modalities. Simultaneously, the encoded MJFs are input into the transformer decoder module, which is used to guide the transformer encoder module to learn the long-term dependencies between different modalities. Finally, sentiment classification is performed based on the outputs of the transformer encoder module. The main contributions of this study are summarized as follows:

- To capture the deep interaction between different modalities and utilize the text modality, we propose to translate the visual and auditory modalities into the textual modality with a modality translation module, which not only can improve the quality of the visual and auditory modalities through deep interaction, but also can fill the missing modalities through modality translation.
- To handle uncertain missing modalities in MSA, we apply a pre-trained model to supervise the transformer encoder module to generate joint features of uncertain missing modalities approximating those of complete modalities. This approach eliminates the need for the model to determine which modality/modalities is/are missing and can reduce the complexity of problem solving.
- Based on two popular benchmark datasets (CMU-MOSI and IEMOCAP), we conduct extensive experiments to verify the performance of our proposed model, MTMSA, with experimental results demonstrating that MTMSA outperforms eight baseline models.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 describes the proposed model. Section 4 presents the experimental evaluation and results. Finally, Section 5 summarizes the conclusions of this study and explores future research directions.

2. Related work

This work is related to the topics of multimodal sentiment analysis and multimodal sentiment analysis with missing modalities. In the following sections, we will introduce the related works about these two topics.

2.1. Multimodal sentiment analysis research

Multimodal sentiment analysis can boost the performance of sentiment analysis [2] by mining and integrating emotional information from multimodal data. Early multimodal sentiment analysis is usually implemented by feature classification. Arunkumar et al. [22] proposed a SVM with particle swarm optimisation (SVM-PSO) to infer opinion mining outputs, and verified that this method outperforms other classifiers. With the rapid development of deep neural network, multimodal sentiment analysis methods based on deep learning have achieved many remarkable results. Mahendhiran et al. [3] proposed a CLSA CapsNet for MSA, which combines concept level sentiment analysis and natural language concepts extracting, and then applies the capsule networks to interpret and analyze them. Experimental results proved that this method can achieve high accuracy of MSA.

Through effective fusion of multimodal sentiment data, it can improve the accuracy of sentiment analysis [23,24]. To learn the joint representations of multimodal data, three fusion strategies are commonly used. (1) Early Fusion: which links each modality together and then input the mixed data to a predictive model. Sun et al. [16] proposed a gated inter-modality attention mechanism to perform modality interactions and filter inconsistencies from multiple modalities in an adaptive manner. (2) Late Fusion: which builds a separate model for each modality and obtains joint features by combining the decisions of the models. Zheng et al. [25] designed a feature extraction scheme and a matching model structure for each modality separately, and used a late fusion method to fuse all features. (3) Hybrid Fusion: which fuses features by combining early fusion and late fusion. Mai et al. [26] proposed a HyCon model for hybrid contrastive learning

of trimodal representation. The model utilizes both intra-modal/inter-modal contrastive learning and semi-contrastive learning to thoroughly explore cross-modal interactions and reduce modal discrepancies.

Inspired by machine translation, some research applies the encoder–decoder structure for MSA and proposes some effective MSA methods based on modality translation. Mai et al. [27] proposed an adversarial encoder–decoder framework to convert the distribution of source modalities to the distribution of target modalities. Yang [8] proposed a multimodal translation framework that improves the quality of visual and audio features by converting them into text features extracted by BERT. Wang et al. [28] proposed an end-to-end translation network based on transformer, which utilizes transformer to perform convert between modalities, and uses forward and backward translation to capture correlations between multimodal features. However, above MSA models have been proposed under the assumption that all the modalities are always available, and these models will fail when some modalities are missing.

2.2. MSA with missing modalities

Currently, some research considering modalities missing has been carried out in multimodal machine learning and MSA, and some wonderful results have been achieved, which can be divided into two categories: generative methods [29–33] and joint learning methods [12,17,34–38].

Generative Methods: which generate new data with similar distributions to the available data by analyzing the available data. Kingma et al. [29] designed a variational auto-encoder (VAE) for efficient approximate posterior inference using ancestry sampling. Shang et al. [30] proposed a view imputation method that identifies the mapping between different views through a Generative Adversarial Network (GAN), and adopted a multimodal denoising auto-encoder to reconstruct missing views from the output of the GAN. Zhou et al. [31] proposed a deep neural network based on end-to-end feature augmentation generation and multi-source correlation. The feature augmentation generator exploits the available patterns to generate 3D feature augmentation images representing missing patterns. Moreover, Zhang et al. [32] proposed a cross-partial multi-view network that imputes missing views by learning latent multi-view representations and introducing adversarial strategies.

Joint Learning Methods: which exploit the interaction between different modalities to learn joint representations [34]. Han et al. [35] proposed a joint training approach that implicitly fuses multimodal information from auxiliary modalities, which effectively improves the multimodal emotion recognition performance. Zhang et al. [12] proposed an integrating consistency and difference networks to address the missing modality problem, which maps other modalities to the target modality through a cross-modal transformer to solve the problem missing modality. Luo et al. [17] proposed a multimodal reconstruction and align net to tackle the missing modality problem, which guides the reconstruction of missing modality features by introducing multimodal embeddings and missing index embeddings. Pham et al. [36] proposed a method to learn robust joint representation by the cyclic transformation between source and target modalities. Recently, Yuan et al. [37] utilized a transformer-based extractor to extract intra-modal and inter-modal relations, and used the extractor to supervise the reconstruction of missing modalities. Wei et al. [38] proposed a separable multimodal learning method to address the missing modality problem by capturing complementary information between modalities. The summary of the above related works are presented in Table 1.

Although above research has achieved wonderful results, they ignore the negative impact of poor-quality modalities on the performance of models. Moreover, existing models need to consider which modalities are missing in different situations, this increases the complexity of models.

3. Methodology

In this section, we first define the research problem, present an overview of our proposed model, and finally describe the key modules of our proposed model in detail.

3.1. Problem definition and notations

Assume that the multimodal data for sentiment analysis contain three modalities: $P = [X_v, X_a, X_t]$, where X_v, X_a and X_t denote the visual, auditory, and textual modalities, respectively. Without loss of generality, we use X_M^m to represent any missing modality, where $M \in \{v, a, t\}$. For instance, when the visual modality is absent, the multimodal data are denoted as $\{X_v^m, X_a, X_t\}$. When the visual and auditory modalities are missing, the multimodal data are denoted as $\{X_v^m, X_a^m, X_t\}$. The problem studied in this study can be defined as the accurate identification of a user's sentiment based on multimodal data $\{X_v^m, X_a, X_t\}$ with uncertain missing modalities. For simplicity, in the following sections, we use $\{X_v^m, X_a, X_t\}$ to represent multimodal data with uncertain missing modalities.

3.2. Model overview

To solve the problem of MSA with uncertain missing modalities, we propose a modality translation-based MSA model (MTMSA), the structure of MTMSA is illustrated in Fig. 2. The workflow of MTMSA is as follows: (1) Input multimodal data $\{X_v^m, X_a, X_t\}$ into the pre-trained TMTN model, and $\{X_v^m, X_a, X_t\}$ is encoded by the pre-trained TMTN. Meanwhile, (2) multimodal data $\{X_v^m, X_a, X_t\}$ are input into the TMTN model. In TMTN, the text modality is encoded with the transformer encoder, then the visual and encoded text are input into the modality translation module to translate the visual modality to textual modality. Meanwhile, the audio and the encoded text are input into another modality translation module to translate audio to text. Next, the translated visual, translated auditory, and encoded textual modalities are fused into the missing joint features (MJFs).

Then, MJFs are input into the transformer encoder module, which is supervised by the pre-trained TMTN to make the MJFs of uncertain missing modalities approximate those of complete modalities. The encoded MJFs are input into the transformer decoder module to learn the long-term dependencies between different modalities. Finally, sentiment classification is performed based on the outputs of the transformer encoder module. In the following sections, we introduce the transformer and then describe the key modules of TMTN in detail.

3.3. Transformer

Recently, the transformer model [39] has achieved great success in many artificial intelligence fields and demonstrated its superior performance. The key concepts of the transformer are as follows: Given the input X, we define the queries as $Q = XW_Q$, keys as $K = XW_K$, and values as $V = XW_V$, where $W_Q \in \mathbb{R}^{d\times d}$, $W_K \in \mathbb{R}^{d\times d}$ and $W_V \in \mathbb{R}^{d\times d}$ are weight matrices. The key component of multi-head dot-product attention is formalized in Eq. (1) as:

$$\begin{aligned}\text{Attention}(Q,K,V) &= \text{Soft}\max\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \\ &= \text{Softmax}\left(\frac{XW_QW_K^{\mathrm{T}}X^{\mathrm{T}}}{\sqrt{d_k}}\right)XW_V\end{aligned}. \tag{1}$$

Since the multi-head attention mechanism has multiple attention heads simultaneously, it can capture information from different subspaces. Therefore, to learn the expression of multiple semantics in multiple modalities, we use the multi-head attention mechanism to

Table 1
Summary of related works about missing modalities.

Category	Model	Adopted technique	Studied problem	Modality missing	Advantages	Disadvantages
Generative	MFNet [31]	Encoder and Decoder	Brain Tumor Segmentation	Visual modality missing	Utilizes the available modalities to generate 3D feature-enhanced image representing the missing modality	Only considers the visual modality and cannot be used for MSA
Generative	CPM-Nets [32]	GANs	Multi-view Learning	Arbitrary view-missing	Jointly exploits all samples and views and is flexible for arbitrary view-missing patterns	Only considers the visual modality and cannot be used for MSA
Generative	EDDN [19]	Encoder and Decoder	Image Generation	Visual modality missing	Can complete the missing modality without the category label information as an input	Only considers the visual modality and cannot be used for MSA
Generative	CRA [18]	Autoencoder	Missing Modalities Imputation	Uncertain missing	Provides a data imputation framework that leverages strengths of residual learning and autoencoder networks	No evidence can prove whether this model can be used for MSA
Generative	VIGAN [30]	GANs and DAE	Missing View Problem	Uncertain missing	Enables the knowledge integration for domain mappings and view correspondences to effectively recover the missing view	No evidence can prove whether this model can be used for MSA
Joint learning	ICDN [12]	CNN and Transformer	Multimodal Sentiment Analysis	Content Missing within a Modality	Utilizes a cross-modal Transformer to map alternative modalities to the target modality thus to solve the issue of missing modalities	Cannot be applied for MSA with uncertain missing modalities
Joint learning.	MRAN [17]	MLP and GloVe	Multimodal Sentiment Analysis	Uncertain Missing Modalities	Utilizes the superiority of text modality to increase the robustness of the missing modality problem in MSA	Lacks of deep semantic interaction between modalities
Joint learning.	MMIN [20]	CRA	Multimodal Sentiment Analysis	Uncertain Missing Modalities	Can predict the representation of any missing modality given available modalities under different missing modality conditions	Lacks of utilizing the superiority of the text modality
Joint learning	TATE [21]	Transformer	Multimodal Sentiment Analysis	Uncertain Missing Modalities	Utilizes the pre-trained network trained with full modalities to supervise the encoded vectors	Lacks of utilizing the superiority of the text modality
Joint learning	MCTN [36]	RNN	Multimodal Sentiment Analysis	Uncertain Missing Modalities	Provides a method of learning joint representations using only the source modality as input	Lacks of utilizing the superiority of the text modality
Joint learning	TFR-Net [37]	Transformer	Multimodal Sentiment Analysis	Content Missing within a Modality	Improves the robustness of models for the random missing in non-aligned modality sequences	Lacks of utilizing the superiority of the text modality
Joint learning	ESMLM [38]	Tensor fusion, Tucker decomposition, and Knowledge distillation	Multimodal Sentiment Analysis	Uncertainty of missing two modalities	Can successfully capture from even missing modalities the intra- and inter- modality interaction dynamics	Lacks of utilizing the superiority of the text modality

extract information in different semantic spaces of each modality. The multi-head attention mechanism is given by Eq. (2) as follows:

$$\begin{aligned} E_M &= \text{MultiHead}\,(Q,K,V) \\ &= \text{Concat}\left(head_1, head_2, \ldots, head_h\right) W_O \end{aligned}. \tag{2}$$

where $W_O \in \mathbb{R}^{d\times d}$ is a weight matrix, and h is the number of heads. The ith $head_i$ is calculated using Eq. (3) as follows:

$$head_i = \text{Attention}\left(XW_Q^i, XW_K^i, XW_V^i\right). \tag{3}$$

where $W_Q^i \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$, $W_K^i \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$, and $W_V^i \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$ are the ith weight matrices of the query, key, and value, respectively.

3.4. Modality translation module

Existing research has shown that MSA always obtains the best analysis results based on the text modality. That is, the accuracy of text-based sentiment analysis is approximately 70%–80%, whereas that of video- or audio-based sentiment analysis is approximately 60%–70% [13]. Inspired by the above results, we propose translating the visual and auditory modalities to the text modality using a modality translation module, which can make the visual and auditory modalities approximate the text modality, thus enhancing the quality of multimodal features and improving the effect of multimodal sentiment analysis. Moreover, this modality translation module can fills the visual and auditory modalities when the visual or/and auditory modality is/are missing. The framework of the modality translation module is illustrated in Fig. 3, and the calculation process of this module is described as follows:

First, we feed the sequence of each modality into a fully connected layer for dimensional transformation, and each modality is transformed into $X_v^m \in \mathbb{R}^{l_v\times d}$, $X_a \in \mathbb{R}^{l_a\times d}$, $X_t \in \mathbb{R}^{l_t\times d}$. In the remainder of this paper, we use $l(\cdot)$ and $d(\cdot)$ to represent the sequence length and feature dimension, respectively. Then a transformer encoder is used to extract the context features of each modality. The process of updating the modality representations can be formulated according to Eqs. (4)–(6) as follows:

$$E_{vm} = \text{MultiHead}\left(X_v^m, X_v^m, X_v^m\right) \tag{4}$$

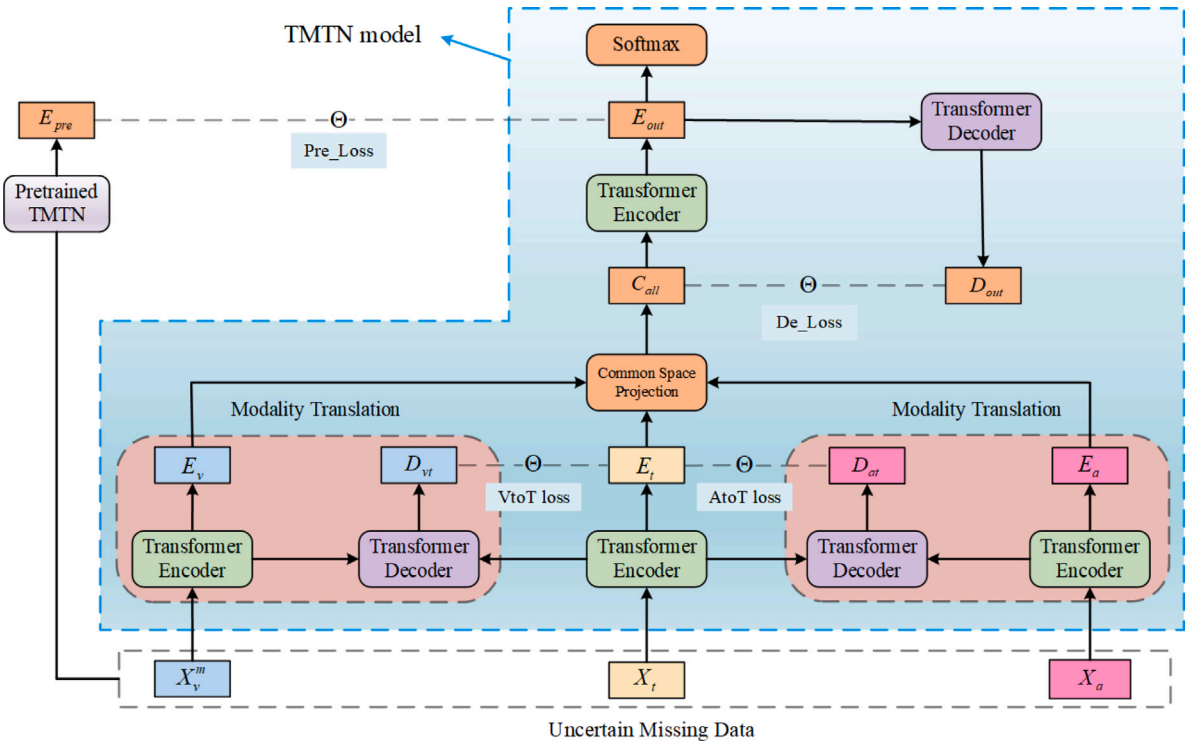


Fig. 2. The structure of MTMSA.

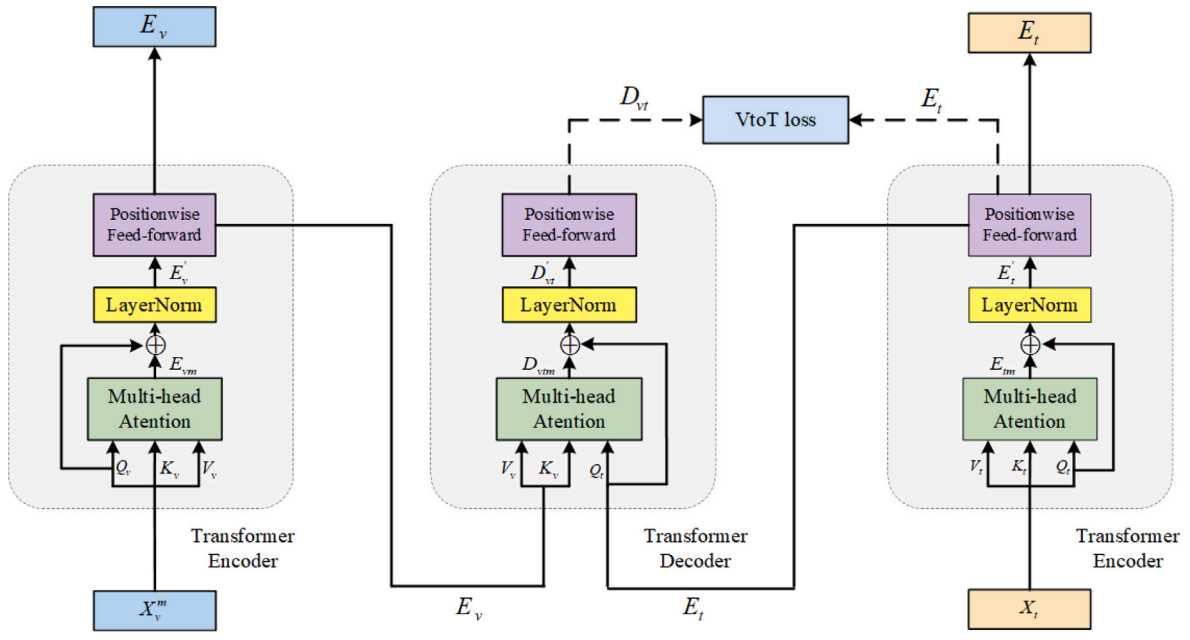


Fig. 3. The specific process of modality translation.

$$E_{am} = \text{MultiHead}\left(X_a, X_a, X_a\right) \tag{5}$$

$$E_{tm} = \text{MultiHead}\left(X_t, X_t, X_t\right). \tag{6}$$

where $E_{vm} \in \mathbb{R}^{l_v \times d}$, $E_{am} \in \mathbb{R}^{l_a \times d}$, and $E_{tm} \in \mathbb{R}^{l_t \times d}$. X_v^m, X_a, and X_t represent the missing visual, auditory, and textual modalities, respectively. Since the transformer encoder adopts the multi-head self-attention mechanism, Q, K, and V in the attention mechanism formula are the same.

Next, we add a residual connection for the extracted features of each modality and apply the layernorm layer for normalization. The calculation process is illustrated in Eqs. (7)–(9) as follows:

$$E_v = \text{Layernorm}\left(X_v^m + E_{vm}\right) \tag{7}$$

$$E_a = \text{Layernorm}\left(X_a + E_{am}\right) \tag{8}$$

$$E_t = \text{Layernorm}\left(X_t + E_{tm}\right). \tag{9}$$

Then, the normalized unimodal features are input into the position-wise feed-forward sublayer for linear transformation, thereby completing the encoding of the three unimodal data types. This process is illustrated in Eqs. (10)–(12) as follows:

$$E_v = \text{Relu}\left(E_v W_{vl}^1 + b_{vl}^1\right) W_{vl}^2 + b_{vl}^2 \tag{10}$$

$$E_a = \text{Relu}\left(E_a W_{al}^1 + b_{al}^1\right) W_{al}^2 + b_{al}^2 \tag{11}$$

$$E_t = \text{Relu}\left(E_t W_{tl}^1 + b_{tl}^1\right) W_{tl}^2 + b_{tl}^2. \tag{12}$$

where W_{vl}, W_{al}, and W_{tl} are the weight matrices, and b_{vl}, b_{al}, b_{tl} denote the learnable biases.

After obtaining the encoding of the visual and textual modalities, the transformer encoder is supervised by the transformer decoder to make the visual modality E_v or auditory modality E_a generated by the encoding module close to the text modality E_t, that is, the encoder is guided to convert the features of visual or auditory modalities to those of the textual modality.

Specifically, when translating the visual or auditory modalities into the textual modality, the encoding of the visual modality (or of the auditory modality) and the encoding of the textual modality are used as the input of the decoder. Then, the encoded textual modality is used as the query of the multi-head attention mechanism, and the encoded visual modality features (or encoded auditory modality features) are decoded as the key and value of the multi-head attention mechanism. The calculation process is described in Eqs. (13) and (14) as follows:

$$D_{vtm} = \text{MultiHead}\left(E_t, E_v, E_v\right) \tag{13}$$

$$D_{atm} = \text{MultiHead}\left(E_t, E_a, E_a\right). \tag{14}$$

Similar to the encoder, we add a residual connection and layernorm layer to the multi-head attention computation and inject a position-wise feed-forward sublayer to work as a decoder layer. Therefore, the updated modality representations can be calculated using Eqs. (15)–(18) as follows:

$$D_{vt} = \text{Layernorm}\left(E_t + D_{vtm}\right) \tag{15}$$

$$D_{at} = \text{Layernorm}\left(E_t + D_{atm}\right) \tag{16}$$

$$D_{vt} = \text{Relu}\left(D_{vt} W_{vtl}^1 + b_{vtl}^1\right) W_{vtl}^2 + b_{vtl}^2 \tag{17}$$

$$D_{at} = \text{Relu}\left(D_{at} W_{atl}^1 + b_{atl}^1\right) W_{atl}^2 + b_{atl}^2. \tag{18}$$

where W_{vtl} and W_{atl} are the weight matrices, and b_{vtl} and b_{atl} are the biases.

3.5. Common space projection

After treatment by the encoder module, the three modality features are linearly transformed to obtain the self-correlation common space of each modality [21], which is subsequently concatenated into the MJFs. The advantages of this method are as follows: First, a weight matrix is jointly trained by two modalities, and the interaction information between the two modalities is preserved in the weight. Second, when the missing modality features approach the complete modality features, it only needs to focus on the overall joint features. Thus, regardless of which modality is missing, it can approximate the features of the complete modalities. The calculation process for the common space projection can be described by Eqs. (19)–(21) as follows:

$$C_v = \left[W_{va} E_v \| W_{vt} E_v\right] \tag{19}$$

$$C_a = \left[W_{va} E_a \| W_{ta} E_a\right] \tag{20}$$

$$C_t = \left[W_{vt} E_t \| W_{ta} E_t\right]. \tag{21}$$

where W_{va}, W_{vt}, and W_{ta} are weight matrices, and $\|$ denotes the concatenation operation.

Then, we concatenate all the common vectors to obtain the common joint representation C_{all}. Since C_{all} is obtained by concatenating the uncertain missing modalities, it is the missing joint feature in the text and the calculation process can be described by Eq. (22) as follows:

$$C_{all} = \left[C_v \| C_a \| C_t\right]. \tag{22}$$

3.6. Transformer encoder-decoder module

To effectively model the long-term dependency of information among different modalities, we use a transformer encoder–decoder to capture the dependency information between joint features. The missing joint feature C_{all} is used as the input of the encoder, and the output E_{out} is obtained after encoding, and the calculation process can be described by Eqs. (23)–(25) as follows:

$$E_{allm} = \text{MultiHead}\left(C_{all}, C_{all}, C_{all}\right) \tag{23}$$

$$E_{out} = \text{Layernorm}\left(C_{all} + E_{allm}\right) \tag{24}$$

$$E_{out} = \text{Relu}\left(E_{out} W_e^1 + b_e^1\right) W_e^2 + b_e^2. \tag{25}$$

where the input of the query, key, and value is C_{all}, W_e^1 and W_e^2 are two weight matrices, and b_e^1 and b_e^2 are two learnable biases.

Similarly, considering the output E_{out} of the encoder as the input of the decoder, the representation of the decoded output D_{out} can be calculated using Eqs. (26)–(28) as follows:

$$D_{outm} = \text{MultiHead}\left(E_{out}, E_{out}, E_{out}\right) \tag{26}$$

$$D_{out} = \text{Layernorm}\left(E_{out} + D_{outm}\right) \tag{27}$$

$$D_{out} = \text{Relu}\left(D_{out} W_d^1 + b_d^1\right) W_d^2 + b_d^2. \tag{28}$$

where W_d^1 and W_d^2 are the parameter metrics, b_d^1 and b_d^2 are the biases, and Relu is the activation function.

Finally, the decoder loss is calculated between the output E_{out} of the joint feature encoder and the output D_{out} of the decoder. The decoder loss is described in detail in the next section.

3.7. Training objective

For the proposed model, MTMSA, the training target is to minimize the overall loss of the model. The overall loss function of the MTMSA model consists of the classification loss $(\mathcal{L}_{cls})$, pre-trained loss $(\mathcal{L}_{pretrain})$, decoder loss $(\mathcal{L}_{de})$, modality translation losses $\mathcal{L}_{AtoT}$ and $\mathcal{L}_{VtoT}$, which is defined in Eq. (29), as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{pretrain} + \lambda_2 \mathcal{L}_{de} + \lambda_3 \mathcal{L}_{VtoT} + \lambda_4 \mathcal{L}_{AtoT}. \tag{29}$$

where λ_1, λ_2, λ_3, and λ_4 are the weights of the four local losses. In the following sections, each loss function is described in detail.

For loss functions, Kullback–Leibler (KL) divergence is usually used to calculate the difference between two probabilities. However, KL divergence is not symmetrical; therefore, we adopt the Jensen–Shannon (JS) divergence to calculate the loss. The KL divergence and JS divergence are described in Eqs. (30) and (31), respectively, as follows:

$$D_{KL}(p\|q) = \sum_{i=1}^{N} p(x_i) \cdot \frac{p(x_i)}{q(x_i)} \tag{30}$$

$$JS(p\|q) = \frac{1}{2} D_{KL}(p\|q) + \frac{1}{2} D_{KL}(q\|p) \quad . \tag{31}$$

where p and q are two probability distributions.

(1) Pre-trained loss $(\mathcal{L}_{pretrain})$: It is used to approximate the MJFs to the complete joint features. Therefore, we calculate the JS divergence between the output of the pre-trained model (E_{pre}) and that of the transformer encoder (E_{out}). The structure of the pre-trained model is the TMTN model, as shown in Fig. 2, and the pre-trained model is trained with the complete modality. The pre-trained loss is defined as Eq. (32) as follows:

$$\begin{aligned}\mathcal{L}_{pretrain} &= JS\left(E_{out}\|E_{pre}\right)\\ &= \frac{1}{2}D_{KL}\left(E_{out}\|E_{pre}\right)+\frac{1}{2}D_{KL}\left(E_{pre}\|E_{out}\right).\end{aligned} \tag{32}$$

(2) Decoder loss $(\mathcal{L}_{de})$: It is used to supervise the common joint reconstruction. Therefore, we calculate the JS divergence loss between the output of the transformer decoder (D_{out}) and the updated common joint representations (C_{all}). The decoder loss is defined by Eq. (33) as follows:

$$\begin{aligned}\mathcal{L}_{de} &= JS\left(D_{out}\|C_{all}\right)\\ &= \frac{1}{2}D_{KL}\left(D_{out}\|C_{all}\right)+\frac{1}{2}D_{KL}\left(C_{all}\|D_{out}\right).\end{aligned} \tag{33}$$

(3) Modality translation loss $(\mathcal{L}_{AtoT}$ *and* $\mathcal{L}_{VtoT})$: Since the translation method is used to translate the visual and auditory modalities to the textual modality. Therefore, we calculate the JS divergence loss between the modality translation decoder output $(D_{at}$ and $D_{vt})$ and modality translation encoder representations (E_t). The modality translation loss functions are defined in Eq. (34), and Eq. (35) as follows:

$$\begin{aligned}\mathcal{L}_{AtoT} &= JS\left(D_{at}\|E_t\right)\\ &= \frac{1}{2}D_{KL}\left(D_{at}\|E_t\right)+\frac{1}{2}D_{KL}\left(E_t\|D_{at}\right)\end{aligned} \tag{34}$$

$$\begin{aligned}\mathcal{L}_{VtoT} &= JS\left(D_{vt}\|E_t\right)\\ &= \frac{1}{2}D_{KL}\left(D_{vt}\|E_t\right)+\frac{1}{2}D_{KL}\left(E_t\|D_{vt}\right).\end{aligned} \tag{35}$$

(4) Classification loss $(\mathcal{L}_{cls})$: For the final classification module, we feed E_{out} into a fully connected network with the softmax activation function to calculate the prediction score $\hat{y}$, which is illustrated in Eq. (36) as follows:

$$\hat{y} = \mathrm{softmax}\left(W_c E_{\mathrm{out}} + b_c\right). \tag{36}$$

where W_c and b_c are the weights and biases, respectively. Here, we employ the standard cross-entropy loss function for classification, which is defined in Eq. (37) as follows:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{n=1}^{N} y_n \log \hat{y}_n. \tag{37}$$

where N denotes the number of samples, y_n is the true label of the nth sample, and $\hat{y}_n$ is the predicted label. Moreover, the entire calculation process of the MTMSA model is described in Algorithm 1, where the pre-trained TMTN is trained with complete multi modalities.

Algorithm 1 :Modality Translation-Based Multimodal Sentiment Analysis

Input: Multimodal data with uncertain missing modalities: $[X_v^M, X_a, X_t]$, where $X_m^M \in \mathbb{R}^{l_m \times d_m}, m \in \{v,a,t\}$, M denotes a missing modality.
Output: The predicted sentiment category $\hat{y}$.

$X_m \leftarrow \mathrm{Dense}(X_m), X_m \in \mathbb{R}^{l_m \times d}, m \in \{v,a,t\}$
Phase I. In the modality translation module
    Encoder: Produce the encoded representation E_v,E_a,E_t of each modality according to Eqs. (4)–(12)
    $E_v \leftarrow \mathrm{encoder}\left(X_v^M, X_v^M, X_v^M\right)$
    $E_a \leftarrow \mathrm{encoder}\left(X_a, X_a, X_a\right)$
    $E_t \leftarrow \mathrm{encoder}\left(X_t, X_t, X_t\right)$
    Decoder: Produce the decoded representation D_{vt}, D_{at} according to Eqs. (13)–(18)
    $D_{vt} \leftarrow \mathrm{decoder}\left(E_t, E_v, E_v\right)$
    $D_{at} \leftarrow \mathrm{decoder}\left(E_t, E_a, E_a\right)$
    Calculate the translation losses using Eqs. (33) and (34)
    $\mathcal{L}_{AtoT} \leftarrow JS\left(D_{at}||E_t\right), \mathcal{L}_{VtoT} \leftarrow JS\left(D_{vt}||E_t\right)$
Phase II. In the common space projection module
    Calculate the self-related common space representation C_v, C_a, C_t of each modality according to Eqs. (19)–(21)
    Calculate the missing joint features using Eq. (22)
    $C_{all} \leftarrow \left[C_v||C_a||C_t\right]$
Phase III. In the transformer encoder–decoder module:
    Calculate the joint modal feature of the encoder output (E_{out}) and the decoder output (D_{out}) by Eqs. (23)–(28)
    $E_{out} \leftarrow \mathrm{encoder}\left(C_{all}, C_{all}, C_{all}\right)$
    $D_{out} \leftarrow \mathrm{decoder}\left(E_{out}, E_{out}, E_{out}\right)$
    Calculate the loss between decoder output D_{out} and missing joint modalities C_{all} by Eq. (33)
    $\mathcal{L}_{de} \leftarrow JS\left(D_{out}||C_{all}\right)$
Phase IV. Supervise the transformer encoder using the pre-trained module
    Calculate the loss between the output of the pre-trained model (E_{pre}) and the output of the transformer encoder (E_{out}) using Eq. (32)
    $\mathcal{L}_{pretrain} \leftarrow JS\left(E_{out}||E_{pre}\right)$
Phase V. Predict the sentiment category
    Train the whole model using Eq. (29)
    Calculate the prediction score $\hat{y}$ using Eq. (36)
    $\hat{y} \leftarrow \mathrm{softmax}\left(W_c E_{\mathrm{out}} + b_c\right)$
    Return $\hat{y}$
End

4. Experiments

To verify the performance of the proposed model, we conducted extensive experiments using two popular datasets: The Carnegie Mellon University Multimodal Opinion Sentiment and Intensity (CMU-MOSI) [40] and Interactive Emotional Dyadic Motion Capture (IEMOCAP) [41] datasets. In the following sections, we first describe the two public benchmark datasets and data preprocessing, then introduce the experimental settings and eight baseline models, and finally present the experimental results.

4.1. Benchmark datasets

To the best of our knowledge, most MSA studies verify model performance based on the public datasets CMU-MOSI and IEMOCAP. Therefore, we adopted these two datasets as the benchmark datasets. The details of the two datasets and the feature extraction process are as follows:

CMU-MOSI: The CMU-MOSI dataset contains 2199 short monologue video clips extracted from 93 YouTube movie review videos. Each sample in the dataset is annotated using an emotional score ranging from −3 to 3.

IEMOCAP: The IEMOCAP dataset is a widely used multimodal emotion dataset. It is collected through recordings of emotional dialogues and interactions between actors. The dataset consists of five sessions, each containing approximately 30 videos, and each video contains at least 24 utterances. The annotated labels in IEMOCAP are: neutral, frustrated, angry, sad, happy, excited, surprised, fearful, disappointed, and others.

Following previous research, we conducted three classification experiments on the CMU-MOSI dataset and two on the IEMOCAP dataset. Therefore, in our experiment, for the CMU-MOSI dataset, we transformed the sentiment scores into negative, neutral, and positive labels (i.e., negative: [−3, 0), neutral: [0], and positive: (0, 3]). For the

Table 2
Parameter settings of MTMSA.

Description	Symbol	Value
Batch size	b	32
Epoch number	e	20
Dropout rate	p	0.8
Hidden size	d	300
Learning rate	lr	0.001
Missing rate	η	[0.1–0.5]
Maximum text length	n_t	25
Maximum audio length	n_a	150
Maximum video length	n_v	100
Loss weights	λ_1, λ_2, λ_3, λ_4	0.1

IEMOCAP dataset, we transformed the sentiment scores into negative and positive labels (i.e., negative: [frustrated, angry, sad, fearful, disappointed], positive: [happy, excited]).

4.2. Data preprocessing

The feature extraction process for the two datasets is as follows [21]. First, the visual features in the CMU-MOSI and IEMOCAP datasets consist mainly of human faces. Facial features are extracted by the OpenFace2.0 toolkit [42], which includes face, head, and eye movements. The dimensions of a visual representation are 709. Secondly, for the textual representations, the pre-trained bidirectional encoder representations from transformers (BERT) method [43](including 12-layer, 768-hidden, 12-heads) is used to extract textual features. The dimension of the textual feature is 768. Third, the audio features are extracted using Librosa [44]. Each audio sample is mixed into a mono signal and resampled at 16,000 Hz. In addition, each frame is separated into 512 samples, and the Zero-Crossing Rate, Mel-Frequency Cepstral Coefficients (MFCCs), and Constant-Q Transform (CQT) features are selected to represent the audio segments. Finally, these three features are concatenated to produce 33-dimensional audio features. In our experiment, we used the preprocessed data provided by [21] to conduct the experiments.

4.3. Experimental setting

Our experimental platform was a personal computer with the following configuration: OS: Windows 10, CPU: Intel(R) Core(TM) i9-10900K CPU, GPU: Nvidia 3090, and RAM: 96G. We implemented the proposed model on TensorFlow 1.14.0 using Python 3.6. For our proposed model MTMSA, we set the learning rate lr to 0.001, batch size b to 32, and hidden size d to 300. We adopted the Adam optimizer [45] to minimize the total loss $\mathcal{L}$. The epoch number was set to 20 and the weight loss was set to 0.1. The parameter setting of our proposed model is presented in Table 2.

In our experiment, we adopted the accuracy (Acc) metric and macro-F1 score (M-F1) as the evaluation metrics for the performance comparison between our proposed model and baseline models. Acc and M-F1 are defined in Eqs. (38) and (39) as follows:

$$\text{Acc} = \frac{N_{true}}{N} \tag{38}$$

$$\text{M-F1} = \frac{2PR}{P+R}. \tag{39}$$

where N_{true} is the number of correctly predicted samples, N is the total number of samples, P is the positive predictive value, and R is the recall value.

4.4. Baseline models

To verify the performance of MTMSA, we selected eight state-of-the-art models as baseline models, which are introduced as follows:

- AE [46]: This model makes the target value in the neural network equal to the input value and uses the back-propagation algorithm to learn the inherent structure of the data.
- CRA [18]: This is a reconstruction framework for missing modalities based on a cascaded residual autoencoder that employs a residual connection mechanism to approximate the differences between input data.
- MCTN [36]: This is a method to learn robust joint representations by the cyclic transformation between source and target modalities.
- TransM [28]: This is a multimodal fusion method based on end-to-end translation that utilizes the transformer method to perform cyclic translation between source and target modalities to improve translation performance.
- MMIN [20]: This is a unified multimodal emotion recognition model that uses a cascade residual auto-encoder and cyclic consistency learning method to predict missing modes with the available modalities.
- ICDN [12]: This model integrates consistency and difference networks to address the missing modality problem. Additionally, it maps other modalities to the target modality through a cross-modal transformer to solve for the missing modality.
- MRAN [17]: This is a multimodal reconstruction and aligning net that tackles the missing modality problem through guiding the reconstruction of missing modality features by introducing multimodal embeddings and missing index embeddings.
- TATE [21]: This is a TATE network that adopted a tag encoding technique to cover all uncertain missing cases and supervise joint representation learning.

4.5. Performances comparison

In this experiment, we tested the performance of our proposed model MTMSA by conducting three classifications on the CMU-MOSI dataset and two classifications on the IEMOCAP dataset. Our experiment consisted of two parts: the first part considered the case of a single missing modality, and the second part considered the case of multiple missing modalities. The performances of the baseline models were obtained from a previous study [21]. The experimental results are presented in Tables 3 and 4; the best results are bolded. The experimental results for the MTMSA, ICDN, and MRAN models were obtained by employing the trained model on our experimental platform, while the experimental results for the other six models were selected from [21].

Experiment on a single missing modality: In this experiment, the missing modality rate was set from 0 to 0.5. The experimental results are presented in Table 3. From Table 3, for the CMU-MOSI dataset, our proposed model MTMSA outperforms other baseline models on both evaluation metrics (ACC and M-F1) when the missing modality rate is set to 0.2, 0.3, 0.4, and 0.5. However, when the missing modality rate is set to zero, the M-F1 score for MTMSA is 2.29% lower than that of the MMIN model, and the ACC value for MTMSA is 0.01% lower than that of the TATE model. When the missing modality rate is set to 0.1, the M-F1 value of MTMSA is 0.78% lower than that of the TATE model. Moreover, for the IEMOCAP dataset, MTMSA outperforms other baseline models on both evaluation metrics (ACC and M-F1) when the missing modality rate is set to 0, 0.1, 0.2, 0.3, 0.4, and 0.5. Therefore, based on the results in Table 3, we can conclude that the overall performance of our proposed model is better than that of the other baseline models on the CMU-MOSI and IEMOCAP datasets.

Experiment on missing multiple modalities: In this experiment, the missing modality rate was set from 0 to 0.5. The experimental

Table 3
Experimental results for all the models with a single missing modality.

Datasets	Models	0		0.1		0.2		0.3		0.4		0.5	
		M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
CMU-MOSI	AE	56.78	79.69	54.07	79.17	53.40	78.13	51.28	72.53	50.75	73.48	44.99	69.32
	CRA	56.85	79.73	54.37	79.38	53.57	78.24	51.67	72.84	51.02	73.79	45.38	69.45
	MCTN	57.32	79.75	55.48	79.87	53.99	77.49	52.31	71.59	51.64	73.81	45.76	68.11
	TransM	57.84	80.21	57.53	79.69	55.21	78.42	52.87	72.92	52.49	72.40	45.86	68.23
	ICDN	55.71	82.30	54.85	81.25	54.54	80.73	53.51	79.69	46.09	68.23	41.00	60.42
	MRAN	56.79	83.85	56.21	82.81	55.20	82.29	54.52	81.25	54.15	80.73	53.99	78.65
	MMIN	**60.41**	82.29	57.75	81.86	55.38	80.20	53.65	79.24	52.55	76.33	48.95	70.76
	TATE	58.32	**84.90**	**58.21**	84.46	55.46	81.25	55.11	80.73	54.11	80.21	51.71	74.04
	Ours	58.12	84.89	57.43	**84.89**	**57.48**	**83.85**	**55.91**	**81.77**	**54.87**	**81.25**	**54.31**	**79.16**
IEMOCAP	AE	76.15	82.09	75.24	80.26	75.02	78.01	73.92	77.43	70.19	76.01	67.27	76.43
	CRA	77.05	82.13	75.95	80.97	75.13	78.09	74.02	78.11	70.69	76.12	67.75	76.49
	MCTN	78.57	82.27	77.74	81.02	75.37	78.27	74.69	78.52	71.75	76.29	68.17	76.63
	TransM	79.57	82.64	78.03	81.86	76.33	80.43	75.83	78.64	72.01	77.27	68.57	76.65
	ICDN	77.37	82.81	76.46	81.34	74.13	80.56	65.00	78.04	73.26	75.17	60.50	73.35
	MRAN	81.21	85.98	81.06	84.88	80.61	84.38	79.99	83.51	78.63	82.90	75.82	81.33
	MMIN	80.83	83.43	78.85	82.58	77.09	81.27	76.63	80.43	72.81	78.43	70.58	77.45
	TATE	81.15	85.39	79.99	85.09	79.10	84.07	78.45	83.25	76.74	82.75	74.43	82.43
	Ours	**81.36**	**86.14**	**81.81**	**85.24**	**81.47**	**84.46**	**80.20**	**84.28**	**79.53**	**82.94**	**75.84**	**82.55**

Table 4
Experimental results for all the models with multiple missing modalities.

Datasets	Models	0		0.1		0.2		0.3		0.4		0.5	
		M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
CMU-MOSI	AE	56.78	79.69	52.80	75.65	50.84	74.18	46.23	69.18	44.40	69.05	40.29	66.01
	CRA	56.85	79.73	52.85	75.68	51.02	74.73	46.87	69.23	45.17	69.48	41.77	66.82
	MCTN	57.32	79.75	52.97	75.89	51.75	74.16	46.98	69.29	45.73	69.55	42.98	67.02
	TransM	57.84	80.21	53.49	77.08	51.97	74.24	48.23	70.51	47.02	70.38	43.28	67.74
	ICDN	55.71	82.30	54.55	80.21	53.34	78.13	47.04	69.27	42.41	62.50	37.81	57.29
	MRAN	56.79	83.85	55.74	83.33	54.09	80.73	51.60	77.08	49.53	73.96	49.17	72.40
	MMIN	**60.41**	82.29	55.49	80.12	52.79	76.26	48.97	73.27	47.39	74.28	44.63	68.92
	TATE	58.32	**84.90**	56.38	81.77	54.87	81.07	52.12	77.60	51.19	**76.56**	51.15	73.23
	Ours	58.12	84.89	**57.43**	**84.37**	**55.71**	**81.25**	**52.82**	**78.12**	**51.32**	76.04	**51.72**	**73.43**
IEMOCAP	AE	76.15	82.09	75.07	79.84	74.20	76.91	71.55	76.07	69.73	75.16	67.15	75.22
	CRA	77.05	82.13	75.21	79.95	74.22	77.03	71.86	76.41	70.13	75.29	67.31	75.42
	MCTN	78.57	82.27	76.83	80.56	74.77	77.89	72.27	77.03	71.02	75.84	67.51	75.88
	TransM	79.57	82.64	77.21	81.13	75.87	79.01	72.36	78.15	71.38	76.88	68.02	76.04
	ICDN	77.37	82.81	72.56	79.25	71.73	78.99	69.94	77.17	69.59	74.65	68.98	73.26
	MRAN	81.21	85.98	80.22	85.07	79.86	83.60	79.14	82.89	75.80	81.25	68.61	78.30
	MMIN	80.83	83.43	78.02	82.23	76.38	79.53	73.05	79.02	71.22	77.27	69.39	77.01
	TATE	81.15	85.39	78.37	83.63	77.55	82.33	76.14	82.21	74.09	81.94	72.49	80.57
	Ours	**81.36**	**86.14**	**80.28**	**85.17**	**80.39**	**84.12**	**79.30**	**83.85**	**76.07**	**83.07**	**74.80**	**82.03**

results are presented in Table 4. From Table 4, for the dataset CMU-MOSI, our proposed model MTMSA outperforms other baseline models on both evaluation metrics (ACC and M-F1) when the missing modality rate is set to 0.1, 0.2, 0.3, and 0.5. However, when the missing modality rate is set to zero, the M-F1 score for MTMSA is 2.29% lower than that of the model MMIN, and the ACC value for MTMSA is 0.01% lower than that of the TATE model. When the missing rate is set to 0.4, the ACC value for MTMSA is 0.52% lower than that of the TATE model.

MTMSA outperformed other baseline models on the IEMOCAP dataset in terms of both evaluation metrics (ACC and M-F1) when the modality missing rate was set to 0, 0.1, 0.2, 0.3, 0.4, and 0.5. Moreover, compared with other baseline models, our proposed model enhanced the value of the M-F1 score from 0.21% to 5.21% and the value of ACC from 0.75% to 4.05% on the IEMOCAP dataset. Therefore, based on the above results, we can conclude that the proposed model MTMSA outperforms the other baseline models on the CMU-MOSI and IEMOCAP datasets.

Theoretical Analysis: From Tables 3 and 4, we find that MCTN and TransM perform better than AE and CRA because of the cyclic translation operation used in the MCTN and TransM models. Compared with the autoencoder operation in the AE and CRA models, the cyclic translation operation can extract and integrate information from different modalities. By comparing our proposed model MTMSA with MCTN and TransM, we can find that MTMSA achieves better results because it considers the quality differences among the different modalities. Through modality translation operations, it translates lower-quality modalities (auditory and visual) into higher-quality modalities (textual) to improve the sentiment analysis performance.

Comparing ICDN with other models, it can be observed that when the missing modality rate is 0.4, the ACC and F1 values for ICDN on the CMU-MOSI and IEMOCAP datasets drop sharply. This is because ICDN addresses the missing modality by mapping between modalities; however, when there are too many missing modalities, effectively mapping different modalities becomes a significant challenge. Therefore, as the missing rate increases, the performance of ICDN decreases significantly. Moreover, the ACC and F1 values for MRAN on the CMU-MOSI and IEMOCAP datasets drop sharply when the modality missing rate is 0.5. This is because the visual and auditory features of the MRAN model are projected onto the text feature space, and the features of all three modalities are learned to be close to their corresponding emotional word embeddings such that the visual and auditory features are consistent with the textual features. Meanwhile, the intermodal feature projection is limited when the modal missing rate is severe.

When all modalities are available, MTMSA still differs slightly from the aforementioned models. However, when the modalities are missing, our model generally outperforms the MMIN and TATE models. This is because our proposed modality translation operation can compensate for missing modalities in cases of uncertainty. Furthermore, compared

Table 5
Results of modality ablation experiment.

Modules	0		0.1		0.2		0.3		0.4		0.5	
	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
V	35.02	57.29	–	–	–	–	–	–	–	–	–	–
A	40.50	60.93	–	–	–	–	–	–	–	–	–	–
T	53.46	78.64	–	–	–	–	–	–	–	–	–	–
V+A	41.12	61.97	40.20	60.41	37.78	58.33	33.94	55.20	33.83	54.68	33.77	53.12
V+T	54.73	80.72	53.83	78.64	53.42	78.12	52.97	77.60	51.53	76.56	52.15	76.04
A+T	55.54	82.29	55.53	81.25	55.17	80.72	54.47	79.68	53.10	78.64	52.91	77.08
V+A+T	**58.12**	**84.89**	**57.43**	**84.89**	**57.48**	**83.85**	**55.91**	**81.77**	**54.87**	**81.25**	**54.31**	**79.16**

Table 6
Results of module ablation experiment.

Modules	0		0.1		0.2		0.3		0.4		0.5	
	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
MTMSA-MT	56.84	83.85	56.67	82.82	55.61	81.25	53.83	78.64	53.46	78.13	52.55	76.56
MTMSA-CSP	56.69	82.81	54.76	81.25	54.15	79.16	53.34	78.13	52.02	76.04	51.27	75.52
MTMSA-preTMTN	55.71	82.29	55.58	81.20	54.52	79.68	53.07	77.60	50.96	74.47	49.81	72.91
MTMSA	**58.12**	**84.89**	**57.43**	**84.89**	**57.48**	**83.85**	**55.91**	**81.77**	**54.87**	**81.25**	**54.31**	**79.16**

with the above-mentioned two models, our model uses complete joint modalities during pre-training to supervise missing joint modalities. Therefore, it does not need to consider specific scenarios of missing modalities; it only needs to approximate the missing joint modalities to the complete joint modalities, thus reducing the complexity of the model.

4.6. Ablation experiment

To verify the performance of MTMSA on different modalities and the effectiveness of different modules of MTMSA, modality and module ablation experiments were conducted based on the dataset CMU-MOSI. Here, we use "T", "A", and "V" to represent the text, audio, and video modalities, respectively. The experimental settings and results of these two experiments are described in the following paragraphs.

Modality ablation experiment: In this experiment, the three following scenarios were considered: A. Only one modality was used to analyze the sentiment. In this scenario, sentiment analysis results were obtained by directly extracting features from a single modality with a transformer encoder and subsequently performing sentiment classification. Since only one modality was used, there were no cases where the modality was missing. Therefore, the missing modality rate in this situation was set to 0; B. Any two modalities were used for sentiment analysis (e.g., T + V, T + A, and V + A). In this scenario, we set the missing modality ratios to 0, 0.1, 0.2, 0.3, 0.4, and 0.5. For the combination of video and audio (V + A), since the text modality was not involved, the video and audio modalities were encoded by the transformer encoder and then input into the common space for concatenation without a modality translation operation; C. The three modalities (T + A + V) were used simultaneously for sentiment analysis. In this scenario, the missing modality ratios were set to 0, 0.1, 0.2, 0.3, 0.4, and 0.5.

The experimental results of the modality ablation are presented in Table 5; the best results are bolded. From Table 5, in situation A, the best results are achieved with the text modality, where the ACC values of MTMSA are 21.35% and 17.71% higher than those of MTMSA when using video or audio modalities, respectively. These experimental results verify the dominance of text modality in multimodal sentiment analysis. In Scenario B, bimodal combinations that include the text modality achieve better results than bimodal combinations that do not. In the bimodal combination, the ACC value of the bimodal combination without the text modality decreases by 20% compared to the bimodal combination without video or audio modality. In addition, by comparing the experimental results using only one modality with the results using a bimodal modality, we find that the results based on the two modalities are better than those based on a single modality. In Scenario C, the best results are obtained when all three modalities are used simultaneously. Moreover, the experimental results also verify that complementary features can be learned from multiple modalities.

Module ablation experiment: In this experiment, some model variants were generated by removing different modules from MTMSA, and the effectiveness of different modules of MTMSA was verified by testing the performance of model variants. The model variants were generated as follows: (1) The model variant MTMSA-MT was generated by removing the modality translation module from MTMSA. (2) The Model variant MTMSA-preTMTN was generated by removing the pre-trained module from MTMSA. (3) The model variant MTMSA-CSP was produced by removing the common-space projection module from MTMSA.

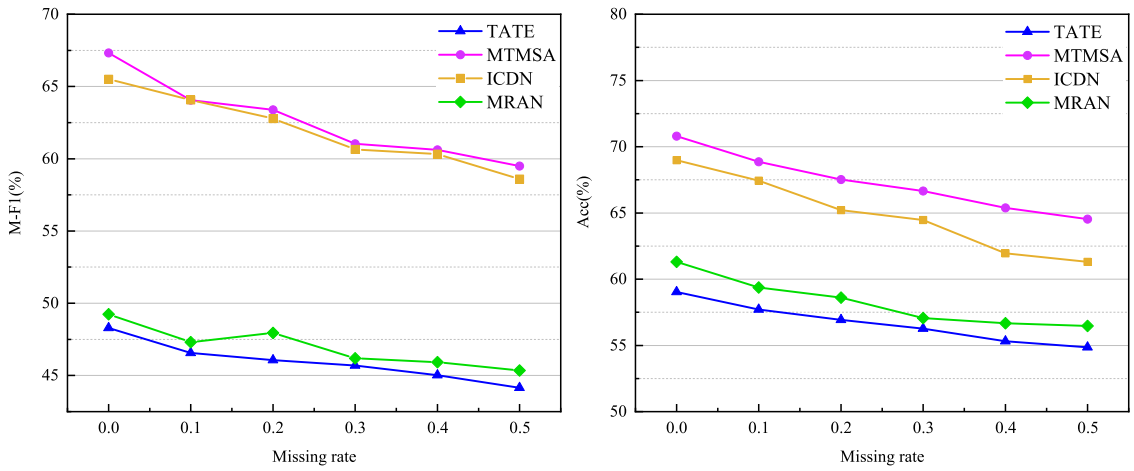
The experimental results of module ablation are presented in Table 6. From Table 6, the MTMSA-MT model decreases by 1.28% in M-F1 and by 1.04% in ACC compared with the MTMSA model when the missing rate is set to 0. The performance of MTMSA-MT decreases by a 2.08% in M-F1 and by 3.13% in ACC when the missing rate is set as 0.3. The above experimental results demonstrate that the modality translation module in the MTMSA model is effective.

For the MTMSA-CSP model, we used a concatenation operation to replace the common-space projection module. Compared with MTMSA, the performance of MTMSA-CSP decreases by approximately 1.43% in M-F1 and by approximately 2.08% in ACC. When the missing rate is set to 0.2, the M-F1 of MTMSA-CSP decreases the most, that is, by 3.33%. When the missing rate is set to 0.4, the ACC value of MTMSA-CSP decreases the most, by 5.21%. These results verify that the common-space projection module can enhance the performance of MTMSA.

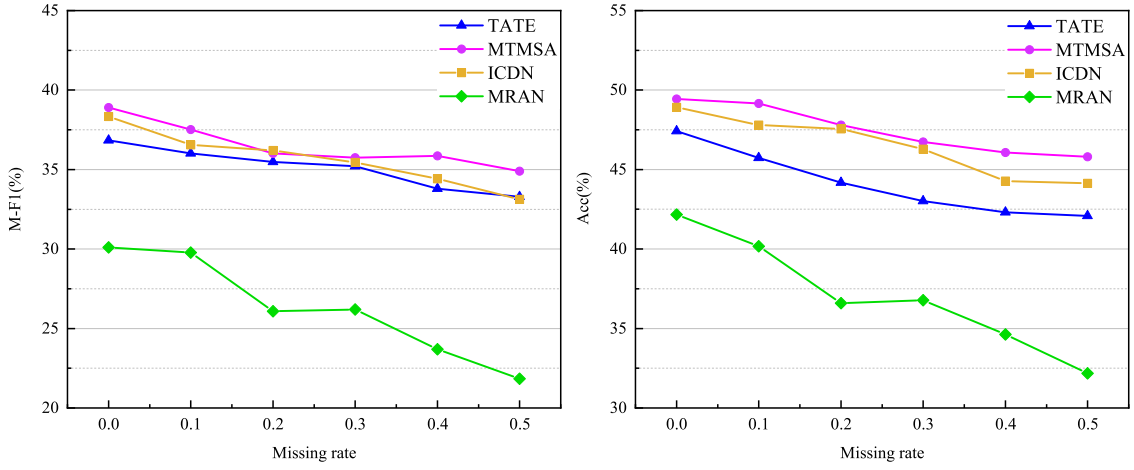
Compared with MTMSA, MTMSA-preTMTN decreases by 2.41% in M-F1 and by 2.6% in ACC when the missing rate is set to 0. When the missing rate is set to 0.5, MTMSA-preTMTN model decreases by 4.5% in M-F1 value. Moreover, MTMSA-preTMTN achieves the largest decrease (6.78%) in ACC when the missing rate is set to 0.4. These results prove that the pre-trained module significantly contributes to the performance of MTMSA.

4.7. Multi-classification verification

To verify MTMSA's performance on the multi-classification of sentiment based on the IEMOCAP dataset, we conducted experiments on four (happy, angry, sad, and neutral) and seven classes (happy, angry, sad, neutral, frustrated, excited, and surprised). The distribution of the multi-classification labels in IEMOCAP is shown in Table 7. In



(a) Four-classification M-F1 value (b) Four-classification Acc value



(c) Seven-classification M-F1 value (d) Seven-classification Acc value

Fig. 4. Experimental results of four models on four-classification and on seven-classification.

Table 7
Distributions of labels in the IEMOCAP dataset.

Category		Hap.	Ang.	Sad.	Neu.	Fru.	Exc.	Sur.
Four-classes	Train	367	655	661	1016	–	–	–
	Test	228	448	423	692	–	–	–
Seven-classes	Train	354	670	636	1013	1109	608	57
	Test	241	433	448	695	740	433	50

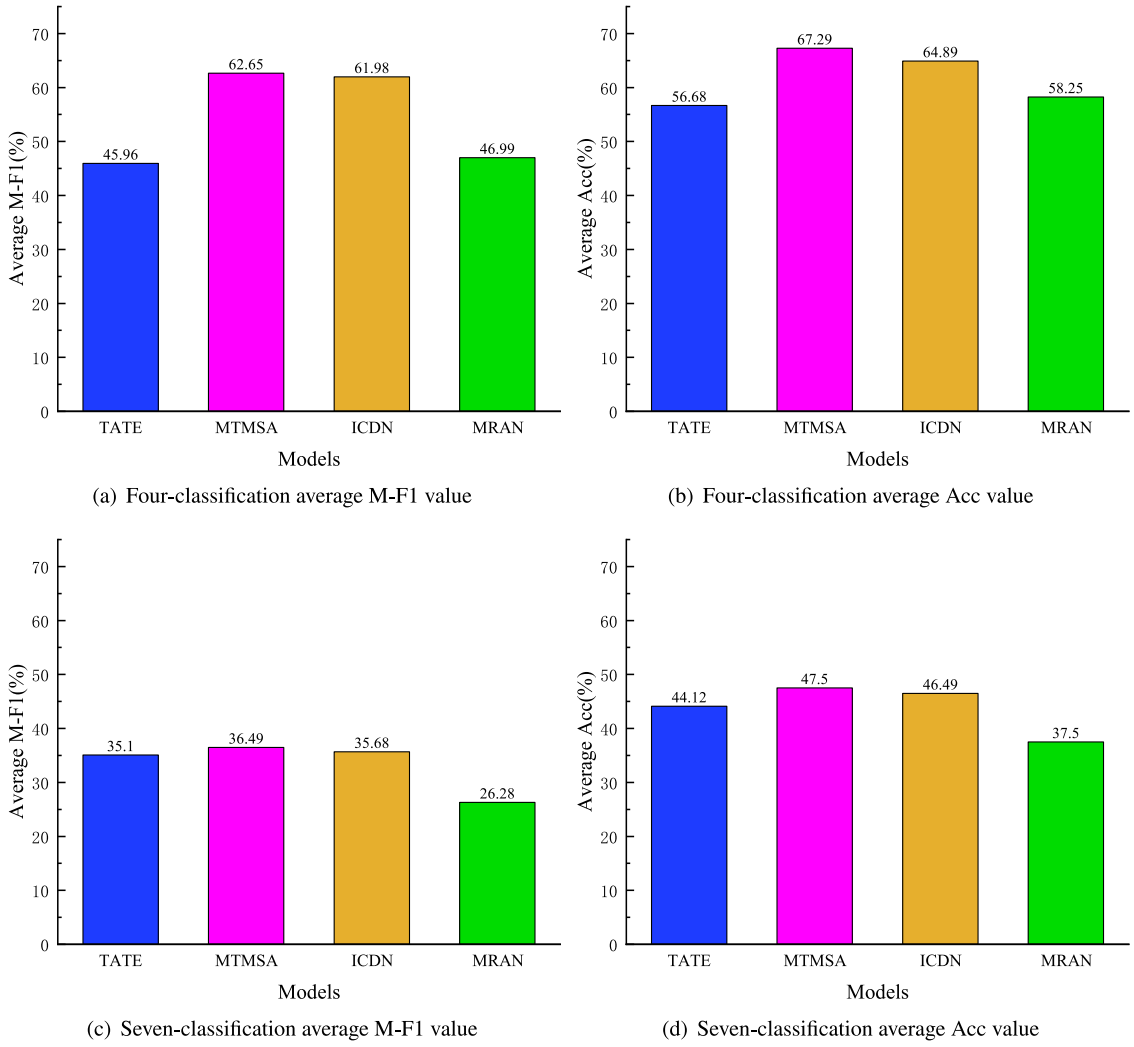
this experiment, we selected TATE, ICDN, and MRAN as the baseline models, and the average results obtained by the four models were recorded. The experimental results are presented in Figs. 4 and 5, where the experimental results for MTMSA, ICDN and MRAN are obtained through employing the trained model on our experimental platform, while those of TATE are obtained from work [21].

For Fig. 4, the vertical axis represents the evaluation metrics (M-F1 or ACC), and the horizontal axis represents the missing rate of modalities. From Fig. 4, for the four- and seven-class classifications, the performance of the four models continue to decline as the rate of missing modalities increases. Moreover, From Fig. 4, our proposed model MTMSA has the best performance among the four models in terms of both four- and seven-class classifications. Moreover, ICDN achieves suboptimal performance. The Experimental results in Fig. 4 verify that MTMSA is effective in multi-class sentiment classification.

For Fig. 5, the vertical axis represents the average values of evaluation metrics (M-F1 or ACC), and the horizontal axis indicates the four models. From Fig. 5(a) and (b), for the four-class classification, the M-F1 value of MTMSA is higher than that of TATE by 16.69%, and the Acc value of MTMSA is larger than that of TATE by 10.61%. Compared with the ICDN model, MTMSA obtains larger M-F1 and ACC values than those of ICDN by 0.67% and 2.40%, respectively. Compared with MARN, MTMSA achieves a 15.67% improvement in M-F1 and a 9.04% improvement in Acc.

From Fig. 5(c) and (d), for the seven-class classification, compared with TATE, MTMSA achieves 1.39% increase in M-F1 and obtains a 3.38% improvement in Acc. Compared with ICDN, MTMSA achieves 0.81% improvement in M-F1 and a 1.01% increase in Acc. Compared with MRAN, MTMSA obtains a 10.21% increase in M-F1 and a 10.00% improvement in Acc. Based on the above experimental results, we can conclude that MTMSA has a better performance in multi-class sentiment classification.

Moreover, from Fig. 4(b) and (d), the four- and seven-class scenarios, the accuracy of ICDN also experience a sharp decline when the missing rates are set to 0.3 and 0.4, respectively. This is because an excessive number of missing modalities prevents the model from effectively capturing the interactions between modalities and utilizing them to fill in the missing modalities. Meanwhile, it can be seen that the accuracy of MRAN in the case of seven-class classification experience a sharp decrease in all cases with different missing rates (except for the missing rate of 0.3 when the correct rate remained basically the same as before). This is because when the modality is severely missing, the intermodal feature projection of the MRAN model is restricted, thereby



(a) Four-classification average M-F1 value

(b) Four-classification average Acc value

(c) Seven-classification average M-F1 value

(d) Seven-classification average Acc value

Fig. 5. Average results of four models on four-classification and seven-classification.

affecting the projection of visual and auditory features onto the text feature space.

5. Conclusion

In this study, we proposed the MTMSA model to address the problem of MSA uncertain missing modalities. Owing to the modality translation technique, the quality of the visual and auditory modalities was improved by translating them into a textual modality, thus enabling MTMSA to fill in the gaps of missing modalities through translation between modalities. Furthermore, MTMSA utilizes a pre-trained model to guide the generation of the most similar joint features of the missing modalities to those of complete modalities, thus resolving the missing modality problem. Owing to the modality translation technique and the joint feature generation method, MTMSA can not only solve all uncertain missing cases, but also does not need to consider which modalities are missing. Moreover, classification, pre-training, encoder, and modality translation losses were proposed to supervise the learning process. All the experiments and further analyses were conducted on two popular benchmark datasets (CMU-MOSI and IEMOCAP), and the experimental results verified the effectiveness of the proposed model. In future works, we will explore cases in which there are no complete modalities for training a pre-trained model. Hence, our future research will be more suitable for practical and real-world applications.

CRediT authorship contribution statement

Zhizhong Liu: Conceptualization, Methodology, Writing – original draft, Investigation, Resources. **Bin Zhou:** Methodology, Experiments validation, Writing – original draft, Visualization. **Dianhui Chu:** Conceptualization, Review, Editing, Investigation. **Yuhang Sun:** Formal analysis, Validation, Visualization. **Lingqiang Meng:** Experiments validation, Editing.

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 nos. 62273290, 61872126), Key projects of Shandong Natural Science Foundation (Grant no. ZR2020KF019), and the Special Funding Program of Shandong Taishan Scholars Project.

References

[1] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, Xiangjie Kong, Multimodal sentiment analysis based on fusion methods: A survey, Inf. Fusion 95 (2023) 306–325.

[2] P.D. Mahendhiran, SJIJoIT Kannimuthu, Deep learning techniques for polarity classification in multimodal sentiment analysis, Int. J. Inf. Technol. Decis. Mak. 17 (03) (2018) 883–910.

[3] P.D. Mahendhiran, Kannimuthu Subramanian, CLSA-CapsNet: Dependency based concept level sentiment analysis for text, J. Intell. Fuzzy Systems (Preprint) (2022) 1–17.

[4] José Ramón Trillo, Enrique Herrera-Viedma, Juan Antonio Morente-Molinera, Francisco Javier Cabrerizo, A large scale group decision making system based on sentiment analysis cluster, Inf. Fusion 91 (2023) 633–643.

[5] Chaima Messaoudi, Zahia Guessoum, Lotfi Ben Romdhane, Opinion mining in online social media: a survey, Soc. Netw. Anal. Min. 12 (1) (2022) 25.

[6] Nitidetch Koohathongsumrit, Warapoj Meethom, A fuzzy decision-making framework for route selection in multimodal transportation networks, Eng. Manag. J. (2022) 1–16.

[7] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, Amir Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion (2022).

[8] Bo Yang, Bo Shao, Lijun Wu, Xiaola Lin, Multimodal sentiment analysis with unidirectional modality translation, Neurocomputing 467 (2022) 130–137.

[9] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al., A systematic review on affective computing: Emotion models, databases, and recent advances, Inf. Fusion (2022).

[10] Sarah A. Abdu, Ahmed H. Yousef, Ashraf Salem, Multimodal video sentiment analysis using deep learning approaches, a survey, Inf. Fusion 76 (2021) 204–226.

[11] Zhibang Quan, Tao Sun, Mengli Su, Jishu Wei, Multimodal sentiment analysis based on cross-modal attention and gated cyclic hierarchical fusion networks, Comput. Intell. Neurosci. 2022 (2022).

[12] Qiongan Zhang, Lei Shi, Peiyu Liu, Zhenfang Zhu, Liancheng Xu, ICDN: integrating consistency and difference networks by transformer for multimodal sentiment analysis, Appl. Intell. (2022) 1–14.

[13] Feng Zhang, Xi-Cheng Li, Chee Peng Lim, Qiang Hua, Chun-Ru Dong, Jun-Hai Zhai, Deep emotional arousal network for multimodal sentiment analysis and emotion recognition, Inf. Fusion 88 (2022) 296–304.

[14] Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, Jianwu Dang, Context-and knowledge-aware graph convolutional network for multimodal emotion recognition, IEEE MultiMedia 29 (3) (2022) 91–100.

[15] Yuntao Shou, Tao Meng, Wei Ai, Sihan Yang, Keqin Li, Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis, Neurocomputing 501 (2022) 629–639.

[16] Hao Sun, Jiaqing Liu, Yen-Wei Chen, Lanfen Lin, Modality-invariant temporal representation learning for multimodal sentiment classification, Inf. Fusion 91 (2023) 504–514.

[17] Wei Luo, Mengying Xu, Hanjiang Lai, Multimodal reconstruct and align net for missing modality problem in sentiment analysis, in: MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part II, Springer, 2023, pp. 411–422.

[18] Luan Tran, Xiaoming Liu, Jiayu Zhou, Rong Jin, Missing modalities imputation via cascaded residual autoencoder, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1405–1414.

[19] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, Shuiwang Ji, Deep adversarial learning for multi-modality missing data completion, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1158–1166.

[20] Jinming Zhao, Ruichen Li, Qin Jin, Missing modality imagination network for emotion recognition with uncertain missing modalities, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 2608–2618.

[21] Jiandian Zeng, Tianyi Liu, Jiantao Zhou, Tag-assisted multimodal sentiment analysis under uncertain missing modalities, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1545–1554.

[22] P.M. Arunkumar, Soundararajan Chandramathi, S. Kannimuthu, Sentiment analysis-based framework for assessing internet telemedicine videos, Int. J. Data Anal. Tech. Strateg. 11 (4) (2019) 328–336.

[23] Sijie Mai, Songlong Xing, Haifeng Hu, Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network, IEEE/ACM Trans. Audio, Speech, Lang. Process. 29 (2021) 1424–1437.

[24] Bowen Zhang, Xutao Li, Xiaofei Xu, Ka-Cheong Leung, Zhiyao Chen, Yunming Ye, Knowledge guided capsule attention network for aspect-based sentiment analysis, IEEE/ACM Trans. Audio, Speech, Lang. Process. 28 (2020) 2538–2551.

[25] Chunjun Zheng, Chunli Wang, Ning Jia, Emotion recognition model based on multimodal decision fusion, in: Journal of Physics: Conference Series, Vol. 1873, IOP Publishing, 2021, 012092.

[26] Sijie Mai, Ying Zeng, Shuangjia Zheng, Haifeng Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022).

[27] Sijie Mai, Haifeng Hu, Songlong Xing, Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 164–172.

[28] Zilong Wang, Zhaohong Wan, Xiaojun Wan, Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis, in: Proceedings of the Web Conference 2020, 2020, pp. 2514–2520.

[29] Diederik P. Kingma, Max Welling, Auto-encoding variational bayes, 2013, arXiv preprint arXiv:1312.6114.

[30] Chao Shang, Aaron Palmer, Jiangwen Sun, Ko-Shin Chen, Jin Lu, Jinbo Bi, VIGAN: Missing view imputation with generative adversarial networks, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 766–775.

[31] Tongxue Zhou, Stéphane Canu, Pierre Vera, Su Ruan, Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing MR modalities, Neurocomputing 466 (2021) 102–112.

[32] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, Qinghua Hu, Deep partial multi-view learning, IEEE Trans. Pattern Anal. Mach. Intell. (2020).

[33] Srinivas Parthasarathy, Shiva Sundaram, Training strategies to handle missing modalities for audio-visual expression recognition, in: Companion Publication of the 2020 International Conference on Multimodal Interaction, 2020, pp. 400–404.

[34] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, Boqing Gong, Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, Adv. Neural Inf. Process. Syst. 34 (2021) 24206–24221.

[35] Jing Han, Zixing Zhang, Zhao Ren, Björn Schuller, Implicit fusion by joint audiovisual training for emotion recognition in mono modality, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5861–5865.

[36] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, Barnabás 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.

[37] Ziqi Yuan, Wei Li, Hua Xu, Wenmeng Yu, Transformer-based feature reconstruction network for robust multimodal sentiment analysis, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4400–4407.

[38] Wei Peng, Xiaopeng Hong, Guoying Zhao, Adaptive modality distillation for separable multimodal sentiment analysis, IEEE Intell. Syst. 36 (3) (2021) 82–89.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[40] Amir Zadeh, Rowan Zellers, Eli Pincus, Louis-Philippe Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88.

[41] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, Shrikanth S Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359.

[42] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, Louis-Philippe Morency, Openface 2.0: Facial behavior analysis toolkit, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 59–66.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[44] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, Librosa: Audio and music signal analysis in python, in: Proceedings of the 14th Python in Science Conference, Vol. 8, 2015, pp. 18–25.

[45] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[46] Pierre Baldi, Autoencoders, unsupervised learning, and deep architectures, in: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, JMLR Workshop and Conference Proceedings, 2012, pp. 37–49.