



Full Length Article

Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities

Yuhang Sun^a, Zhizhong Liu^{a,*}, Quan Z. Sheng^b, Dianhui Chu^c, Jian Yu^d, Hongxiang Sun^a

^a The School of Computer and Control Engineering, Yantai University, Yantai, 264005, China

^b School of Computing, Macquarie University, Sydney, NSW, 2109, Australia

^c College of Computer Science and Technology, Harbin Institute of Technology, Weihai, 264209, China

^d Department of Computer Science, Auckland University of Technology, Auckland, 1142, New Zealand

ARTICLE INFO

Keywords:

Multimodal sentiment analysis
Uncertain missing modalities
Similar modality completion
Transformer

ABSTRACT

Recently, uncertain missing modalities in multimodal sentiment analysis (MSA) brings a new challenge for sentiment analysis. However, existing research cannot accurately complete the missing modalities, and fail to explore the advantages of the text modality in MSA. For the above problems, this work develops a Similar Modality Completion based-MSA model under uncertain missing modalities (termed as SMCMSA). Firstly, we construct the full modalities samples database (FMSD) by screening out the full modality samples from the whole multimodal dataset, and then predicting and marking the sentiment labels of each modality of the samples with three pre-trained unimodal sentiment analysis model (PTUSA). Next, for completing the uncertain missing modalities, we propose a set of missing modalities completion strategies based on the similar modalities selected from FMSD. For the completed multimodal data, we first encode the text, video and audio modality using the encoder of transformer, then we fuse the representation of text into the representations of video and audio under the guidance of a pre-trained model, thereby improving the quality of video and audio. Finally, we conduct sentiment classification based on the representations of text, video and audio with the softmax function respectively, and get the final decision with the decision-level fusion method. Based on benchmark datasets CMU-MOSI and IEMOCAP, extensive experiments have been conducted to verify that our proposed model SMCMSA has better performance than that of the state-of-the-art baseline models. The codes of our model are available at <https://github.com/Astro2Sun/SMCMSA>.

1. Introduction

Recently, along with the popular of social network platforms (e.g., YouTube, Twitch and Tiktok) and the rapid development of social media (e.g., Meta, X and Weibo), a growing quantity of individuals are inclined to convey their personal sentiments and viewpoints by posting videos, graphics, etc., which produces a large amount of video, audio, and text information. To effectively identify and understand the sentiments in these messages, multimodal sentiment analysis (MSA) has become an important research direction in the sentiments analysis field. The aim of MSA is to achieve a deeper understanding and detection of the user's sentiment by leveraging multimodal information including text, audio and video in a monologue video [1]. Simultaneously, automatic and accurate sentiment analysis is playing a pivotal role in various domains, such as Human-Computer Interaction Systems [2], Decision Support Systems [3], Intelligent Service Systems [4], Evaluating Systems [5] and Emotional Health Management [6].

Compared to unimodal data, multimodal data contains complementary information for sentiment analysis. Therefore, the accuracy of sentiment analysis has been significantly enhanced by complementary learning from multimodal features [7,8]. Over the years, some effective MSA models have been proposed based on new deep learning models, such as recurrent neural networks [9], transformers [10] and graph convolutional neural networks [11,12]. Existing research work has achieved good results and facilitated the successful application of MSA technology in fields such as education, healthcare, and elderly care.

Unfortunately, most exiting works deem that the three modalities (text, video and audio) are always usable [13]. In fact, uncertain modalities missing often occurs in real-world applications because of some uncontrollable factors [14]. For instance, users' images cannot be captured when the camera device is blocked; Words spoken by users cannot be obtained when there is some sudden noise that cannot

* Corresponding author.

E-mail addresses: 202200358047@s.ytu.edu.cn (Y. Sun), zhizhongliu@ytu.edu.cn (Z. Liu), michael.sheng@mq.edu.au (Q.Z. Sheng), cdh@hitwh.edu.cn (D. Chu), jian.yu@aut.ac.nz (J. Yu), 2363453828@s.ytu.edu.cn (H. Sun).

<https://doi.org/10.1016/j.infus.2024.102454>

Received 17 February 2024; Received in revised form 14 April 2024; Accepted 2 May 2024

Available online 7 May 2024

1566-2535/© 2024 Elsevier B.V. All rights reserved.



Modality	Demonstration	Possible Reasons
Visual		Camera equipment is obscured by obstacles
Audio		Excessive ambient noise causes the audio signal to be unavailable
Text	Oh hey, do not thank me, thank yourself. You are the one who faced her fears and ultimately overcame them.	Inability to capture text information due to privacy issues

Fig. 1. Instances of Uncertain missing modalities in MSA.

be removed, and so on. So that, in many real-world scenarios, the phenomenon of uncertain modal missing will often occur (illustrated as Fig. 1), which will results in failures of most existing MSA models. Therefore, how to deal with uncertain missing in MSA has become a critical problem to be tackled.

Recently, some research has been carried out to tackle the challenging issue of MSA under uncertain missing modalities, a number of wonderful methods have been proposed, which can be divided into two categories, namely *Generative Methods* and *Joint Learning Methods*. The first kind of methods treat the uncertain missing modalities by generating new data that would match the observed distributions of existing modalities. Such as work [15], which developed a cascaded residual autoencoder (CRA) model to compensate for the missing modalities through stacking residual autoencoders and exploiting the correlation between different modalities. Work [16] conducted semantic learning from incomplete data through low-level feature reconstruction, and then applied the siamese representation learning to align high-level representations of the complete data and incomplete data.

The second kinds of methods aim to learn the latent representations of the missing modalities from the available modalities. Such as work [17], it devised a framework for handling signal missing in MSA tasks and other multimodal contexts, including a crossmodal interaction module, a feature refinement module, and a knowledge-integrated self-distillation module for precise missing semantics reconstruction. Work [18] developed MSA model considering uncertain missing modalities, this model applied a tag-assisted transformer encoder network to guide the learning of joint distributions with tags, and adopted a pre-trained model to treat the situations of uncertain missing modalities. Despite several wonderful models have been developed to solve the problem of MSA under modalities missing, there are still some shortcomings to be addressed, which are presented as follows:

- For the Generative Methods, although this kind of methods can generate data for the missing modalities though learning the distribution of available modalities, the quality of the data generated by models is usually poor compared with the real data, which will decrease the performance of MSA models
- The Joint Learning Methods always feed the multimodal data with uncertain missing modalities into the MSA model directly for representations learning. However, the uncertain missing modalities will result in poor normal common space projection, thus affects the performance of the MSA model.
- Actually, the text modality always contains more useful emotional information than the other two modalities (video and audio). Some works [8,14] have proved that the accuracy of sentiment analysis based on the text modality is better than that of sentiment analysis based on the other two modalities. However, existing research fails to take the advantage of the text modality to improve the quality of the other two modalities.

To attack the above problems, we develop a similar modality completion-based multimodal sentiment analysis model for uncertain missing modalities (named SMCMSA). Our proposed model is composed of three modules, they are *missing modality completion module*, *multimodal features fusion module* and *sentiment prediction module*. For the first module, we construct a full modalities samples database (FMSD) as follows, we first screen out the full modalities samples from the multimodalities dataset; then, we predict and mark the sentiment labels of each modality on the samples with three pre-trained unimodal sentiment analysis models (PTUSA). In the first module of SMCMSA, considering different modality missing situations, we propose a set of similar modality completion strategies to complete the missing modalities with similar modalities selected from FMSD. In the second module of SMCMSA, we first encode the text, video and audio with the transformer encoder; then, we fuse the representation of text into the representations of video and audio, thus to improve the quality of the video and audio modalities. Next, the encoded text, video and audio modalities are fused with the guidance of a pre-trained model. In the sentiment prediction module, we first apply the softmax function to conduct sentiment analysis based on the encoded text, fused video and fused audio, respectively. Then, the final sentiment classification is obtained the above three sentiment analysis results with the decision level fusion strategy. The contributions of our work are presented as follows:

- For solving the problem of MSA under uncertain missing modalities, we firstly bring forward the idea of completing missing modalities with the similar modalities, which are selected according to the similarities and the predicted sentiment labels. For different modalities missing scenarios, we propose a set of strategies for completing the missing modalities.
- Inspired by the advantage of the text modality in MSA, we propose to fuse the text's representation into the video and audio with the guidance of a pre-trained model, thus to enhance the quality of the video and audio. The pre-trained model is trained with the full modalities, which not only can guide the fusion between modalities, but also can help to boost fusion result of the incomplete multimodal data closer to that of the full modalities.
- Base on two public and popular datasets (CMU-MOSI and IEMO-CAP), we carried out extensive experiments to prove the superiority of our proposed model SMCMSA. Experimental results have proved that our proposed model has better performance than that of the Ten baseline models.

The structure of this work is described as follows. Section 2 overviews existing research work. Section 3 introduces our proposed model SMCMSA. Section 4 presents our experiments and results analysis. At last, Section 5 summarized our work and discusses some future research activities.

2. Related work

In this section, we first review existing research on multimodal sentiment analysis (MSA). Then, we review the representative works on the problem of MSA with missing modalities.

2.1. Research on multimodal sentiment analysis

Actually, compared to the unimodal data, multimodal data integrates information from multiple aspects (such as video, audio and text), which can bring more comprehensive and robust sentiment analysis results [14]. Recently, the topic of MSA has attracted much attention [14,18]. In the early stages of MSA research, some traditional machine learning methods have been adopted. Rozgić et al. [19] proposed an automatically generated tree ensemble, which is constructed with binary support vector machine classifiers. Cummins et al. [20] improved

the performance of the sentiment detection system with multiple Bag-of-Words paradigms and additional data, especially when considering information from the test domain and out-of-domain datasets. Arunkumar et al. [21] proposed several machine learning classifier frameworks for opinion mining, which demonstrated that Support Vector Machine with particle swarm optimization performs best in evaluating video content reviews.

With the popularity of deep learning models, some MSA methods based on deep learning have been proposed and demonstrated excellent performance [2,7]. For this kind of methods, the integration of multimodal features has important influences on the performance of MSA [22,23]. Currently, there are four commonly used multimodal fusion strategies to deeply investigate the interactions between different modalities, which are as follows: (1) *Early fusion*, which combines the features of different modalities into a fused feature, then inputs the fused feature into the sentiment prediction model. Work [13] developed a gated inter-modality attention mechanism to enhance modality interactions and employed parallel structures to acquire comprehensive sentimental information in pairs.

(2) *Late Fusion*, which processes and classifies the features of each modality in a parallel structure, and fuses all classification results into a single decision vector for sentiment prediction. Zheng et al. [24] designed different feature extraction schemes for speech, text and motion modalities respectively and finally used decision fusion to obtain emotion recognition results. (3) *Hybrid Fusion*, which combines the early fusion strategy and the late fusion strategy. Work [25] presented a hybrid contrast learning framework to facilitate cross-modal interactions, maintain inter-class relationships and narrow the gaps between different modalities. (4) *Translation-based Fusion*, which is proposed with the inspire of machine translation. This kind of strategy can capture more meaningful cross-modality relationships by translating one modality and another modality. Liu et al. [14] proposed to translate the visual and auditory modalities into the textual modality with a modality translation module.

Recently, some research has introduced the attention mechanism into MSA model. Wang et al. [26] adopted the text-based multi-head attention to incorporate the information contained in text into the representations of video and audio. Kim et al. [27] introduced a single-stream transformer, which is pretrained on the Multimodal Masked Language Modeling and Alignment Prediction tasks to determine the dependencies between modalities. Ashima et al. [28] developed a multimodal learning model (DMLANet), which obtains emotionally rich features for classification by generating bi-attentive visual maps and modeling the relationship between images and text. However, most outstanding MSA models are proposed with the assumption that all modalities are consistently available, which will invalidate them in scenarios where modalities missing occurs.

2.2. MSA with missing modalities

Recently, multimodal machine learning and MSA with modalities missing have become a challenging problem, some research has been conducted to solve this problem and has achieved promising results. Existing methods for handling the missing modalities can be classified into two categories, which are generative methods [29–34] and joint learning methods [8,10,35–41]. We will review the relevant works in the following.

Generative Methods. This kind of methods usually generate new data that owns similar distributions to the available data through analyzing the available data. Work [29] introduced a variational auto-encoder (VAE) which enables efficient learning in directed probability models. Shang et al. [30] proposed to learn domain mappings through Generative adversarial networks and employ a multi-modal denoising autoencoder for reconstruction. Work [31] utilized an encoder–decoder network to generate data for the missing modalities with an auxiliary

adversarial loss to obtain high quality output. Zhao et al. [39] addressed uncertain missing modality by guiding the generation of joint multimodal representations through forward and backward imagination module that can predict missing modalities under various conditions. Zhou et al. [32] proposed to generate relevant feature-enhanced modalities for the missing modalities with a data generator, which is developed based on an end-to-end feature enhanced generation and a multi-source correlation deep neural network. Zhang et al. [33] converted the learning of latent representation of multi-view into a degenerate process that achieves unification of consistency and complementarity between different views.

Joint Learning Methods. This kind of methods aim to learn the joint representations of multimodalities by exploiting the interaction between different modalities [35]. Work [36] presented a novel joint training model, which incorporates auxiliary modalities during training to map audio and visual features for sentiment prediction. Zhang et al. [10] proposed to model modality interaction using a cross-modal Transformer and self-supervised unimodal sentiment labels to guide sentiment analysis. Work [8] adopted the missing index embeddings to guide the reconstruction of missing modalities' features. Recently, Yuan et al. [38] proposed a transformer-based feature reconstruction network to capture the robust intra- and inter- modality representations and generate the missing modality features. Wei et al. [40] devised a separable tensor fusion network to capture interactions between different modalities, and further improved the computational efficiency with a Tucker decomposition operation. Chi et al. [41] provided an effective add-on training component based on meta sampling.

Moreover, based on our previous work [14], to provide a clearer presentation of current research works on uncertain modality missing, we compare and summarize the related works, which is shown in Table 1. Although existing works mentioned above have achieved excellent results, they cannot overcome the adverse effects of low quality modalities on the performance of MSA. Furthermore, existing works primarily generate data for missing modality by learning from the available modalities, but the generated data often deviates from the real modalities greatly, which will effect the performances of MSA models.

3. Methodology

In the following sections, first of all, we introduce the problem studied in this work; Then, we describe the structure of our proposed model. Finally, we introduce the main function of each module in detail.

3.1. Problem

Assume that there is a set of multimodal data including three modalities: $P = [X_v, X_a, X_t]$, where X_v , X_a and X_t denote the video, audio and text modalities, respectively. Without loss of generality, we use X_M^m to represent the missing modality, where $M \in \{v, a, t\}$. There are some possible scenarios of uncertain missing modalities, which are summarized in Table 2. The problem studied in this work is how to perform robust sentiment analysis based on data set P with uncertain missing modalities. For convenience of presentation, in this work we use $\{X_v^m, X_a, X_t\}$ to denote the multimodal data with uncertain missing modalities.

3.2. Overview of our proposed model

To address the challenging issue of MSA under uncertain missing modalities, we develop a similar modality completion-based MSA model (termed as SMCMSA), which is illustrated as Fig. 2. SMCMSA is composed of three modules, which are missing modality completion module, multimodal feature fusion module and sentiment prediction module. Next, we introduce the operation process of each module as follows:

Table 1
Summary of related works about missing modalities.

Category	Model	Technique	Problem	Modality missing	Advantages	Disadvantages
Generative	MFNet [32]	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 [33]	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 [31]	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 [15]	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	MMIN [39]	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
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
Generative	EMT-DLFR [16]	Transformer	Multimodal Sentiment Analysis	Random Modality Feature Missing	Enhances semantic learning from incomplete data and promotes high-level representation alignment between complete and incomplete data	Lacks of utilizing the superiority of the text modality
Joint learning	ICDN [10]	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 [8]	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	TATE [18]	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 [37]	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 [38]	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 [40]	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
Joint learning	RMSF [17]	Transformer and Knowledge Distillation	Multimodal Sentiment Analysis	Uncertain Signal Missing	Recovers missing semantics through knowledge integration and transfer from diverse missing-modality samples	Lacks of utilizing the superiority of the text modality
Joint learning	M3S [41]	MMIM , MISA, Self-MM, MMIN, and Meta Learning	Multimodal Sentiment Analysis	A Mixture Of Missing Across Multiple Modalities	Provide an effective add-on training component to existing multimodal models that can significantly improve the model performance on mixed missing modalities	Lacks of utilizing the superiority of the text modality

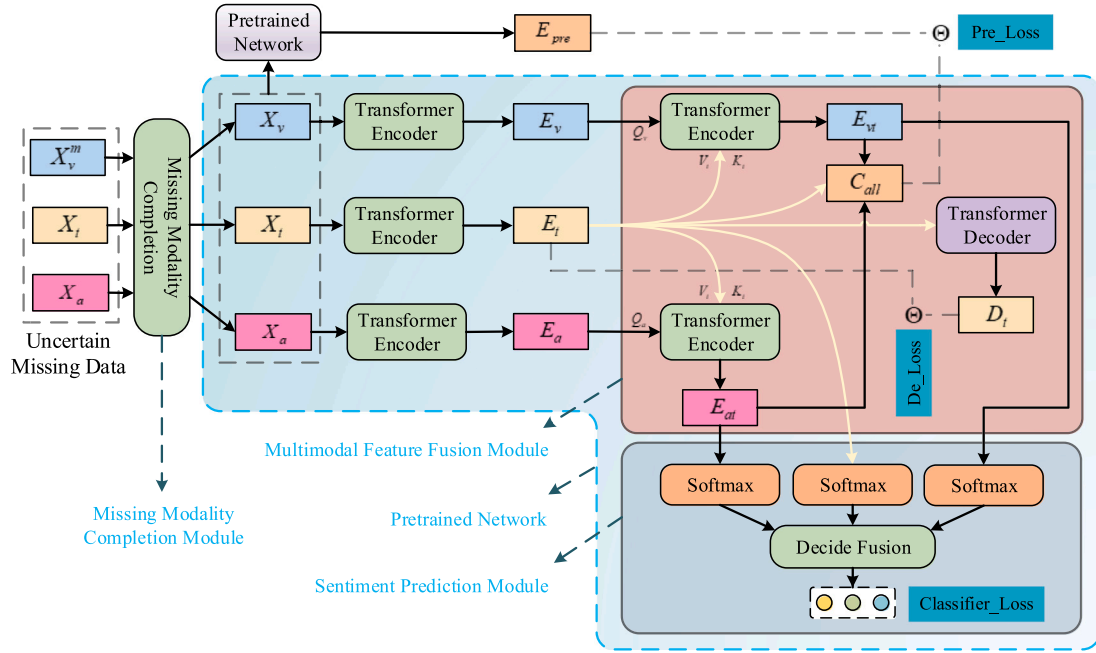


Fig. 2. The structure of SMCMSA.

Table 2

The seven possible scenarios of uncertain missing modalities.

Missing scenario	Missing modality	Multimodal data
No missing	/	$P = [X_v, X_a, X_t]$
Single missing	Visual	$P = [X_v^m, X_a, X_t]$
Single missing	Audio	$P = [X_v, X_a^m, X_t]$
Single missing	Text	$P = [X_v, X_a, X_t^m]$
Multiple missing	Visual & Audio	$P = [X_v^m, X_a^m, X_t]$
Multiple missing	Visual & Text	$P = [X_v^m, X_a, X_t^m]$
Multiple missing	Audio & Text	$P = [X_v, X_a^m, X_t^m]$

- Firstly, for completing the missing modalities with similar modalities, we construct the full modalities samples database (FMSD) through screening out the full modalities samples from the multimodalities dataset, which are collected in daily life, and then predicting and marking the sentiment labels of each modality on the samples with three pre-trained unimodal sentiment analysis models.
- For the missing modality completion module, the multimodal data $\{X_v^m, X_a, X_t\}$ with uncertain missing modalities are completed according to our proposed modality completion strategies (which will be introduced in Section 3.4).
- For the multimodal feature fusion module, the video, audio and text are first encoded by the transformer encoder. Then, the encoded text is fused into the video and audio to enhance the quality of the two modalities. Next, the encoded text, fused video and fused audio are concatenated, and a pre-trained model is used to guide the multimodal feature fusion.
- For the sentiment prediction module, the softmax function is adopted to perform sentiment classification based on the encoded text, fused video and audio, respectively. Then, the final sentiment classification is obtained based on the above sentiment classifications with the decision level fusion strategy.

In subsequent sections, we first present the basics concept and formulas of Transformer, then introduce the three main modules of our proposed model SMCMSA.

3.3. Transformer

Since Transformer has been proposed [42], it demonstrates great advantage is solving different problems and has been applied in many fields. The main formulas and mechanisms of Transformer are presented as follows. Assume that X is the input, 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 $W_V \in \mathbb{R}^{d \times d}$ mean the weight matrices. In Transformer, the multi-head dot-product attention is an important operator, and its calculation process is as Eq. (1):

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

For Transformer, its multi-head attention mechanism includes several multiple attention heads. Therefore, Transformer can learning useful information from multiple perspectives. In this work, we adopt the multi-head attention mechanism to learn important information from different semantic spaces of each modality. The calculation process of multi-head attention mechanism is described as Eq. (2):

$$\begin{aligned} E_M &= \text{MultiHead}(Q, K, V) \\ &= \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) W_O. \end{aligned} \quad (2)$$

where $W_O \in \mathbb{R}^{d \times d}$ indicates the weight matrix, h indicates the number of heads. The i th head_i is calculated as Eq. (3):

$$\text{head}_i = \text{Attention} \left(XW_Q^i, XW_K^i, XW_V^i \right). \quad (3)$$

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 $W_V^i \in \mathbb{R}^{\frac{d}{h} \times \frac{d}{h}}$ denote the i th weight matrices of the Query, Key and Value.

3.4. Missing modalities completion module

For multimodal data with uncertain missing modalities, existing research always fills in the missing modalities with simulated data though learning the available modalities. However, the quality of the simulation results is often inferior to the quality of the real data. For this problem, we put forward the idea that to complete the uncertain

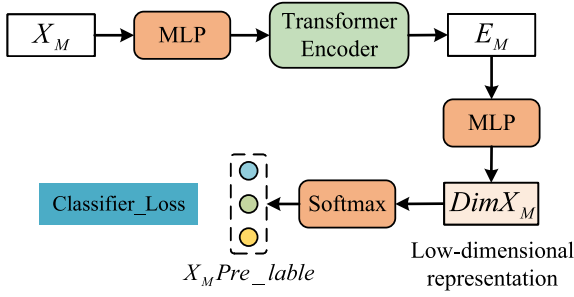


Fig. 3. Pre-trained unimodal sentiment analysis model.

missing modalities with the similar data, which is selected from the full modalities samples database (FMSD). We argue that the quality of the similar data is superior to the quality of data generated by MSA models, thus, completing the missing modalities with real modalities will obtain better performance in MSA.

Moreover, for uncertain missing modalities, there usually exist various situations of modalities missing. For this problem, we categorize uncertain modality missing into two kinds and propose a set of modalities completion strategies for different modality missing situations. Next, we first introduce how to construct the full modalities samples database and then present missing modalities completion strategies.

(1) Full Modality Samples Database Construction

For our proposed approach, a diverse and comprehensive full modalities database is the key foundation for ensuring the performance of our model. Therefore, to complete the missing modalities well, it is necessary to build a sound full modalities database. Actually, as shown in Fig. 1 and Table 2, in users' daily life, although there exist uncertain missing modalities, there are still cases that three modalities are complete. Therefore, we first collect the sentimental data from a large amount of users, and then, we select the full modalities data from the database.

Moreover, for multimodal sentiment data, although the data of a certain modality is similar, the emotions expressed by the data may be different. Therefore, to find the semantically consistent data to complete the missing modalities, we propose to apply the predicted emotional labels as the auxiliary criteria for similar data selection. Thus, after we select the full modalities from all users' multimodal sentiment data, we predict the sentiment labels for each modal data with the pre-trained unimodal sentiment analysis model (PTUSA), and mark the labels on the each modal data. Thus, the full modalities database is constructed. In this work, the pre-trained unimodal sentiment analysis model (PTUSA) is proposed based on the Transformer encoder, the structure of the model is illustrated as Fig. 3.

Here, we take X_M to indicate the unimodal data. First, data X_M is input into a fully connected layer for dimensional transformation, and X_M is converted to $X'_M \in \mathbb{R}^{l \times d}$. (In the remainder of this work, we adopt $l(\cdot)$ and $d(\cdot)$ to denote the sequence length and feature dimension, respectively.) Then, a transformer encoder is used to extract the contextual features of the unimodal data. The representation learning process of the unimodal data can be formulated as follows:

$$X'_M = \text{MLP}(X_M) \quad (4)$$

$$E_M = \text{MultiHead}(X'_M, X'_M, X'_M). \quad (5)$$

where $E_M \in \mathbb{R}^{l \times d}$, and $M \in \{v, a, t\}$. Next, we encode E_M with the fully-connected layers to obtain the low-dimensional representation $DimX_M$. The calculation process is illustrated in Eq. (6):

$$DimX_M = \text{MLP}(E_M). \quad (6)$$

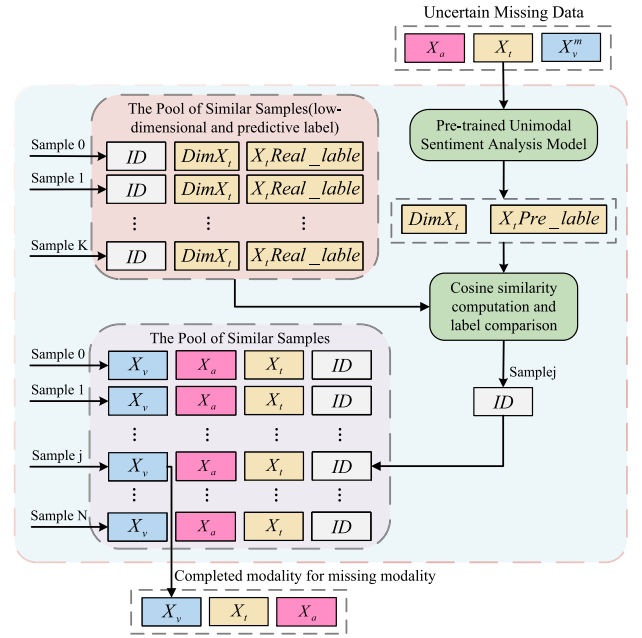


Fig. 4. Text-based missing modalities completion.

where $DimX_M \in \mathbb{R}^{l \times d}$. Finally, we predict the sentiment label \hat{y}_{DimX_M} based on $DimX_M$ according to Eq. (7).

$$\hat{y}_{DimX_M} = \text{softmax}(W_M DimX_M + b_M). \quad (7)$$

where W_M and b_M denote the weights and biases. In this work, the cross-entropy loss function is adopted for sentiment classification, which is described as Eq. (8):

$$\mathcal{L}_{clsPre} = -\frac{1}{N} \sum_{n=1}^N y_{Mn} \log \hat{y}_{DimX_Mn}. \quad (8)$$

where N denotes the scale of samples, y_{Mn} indicate the true label of the n th sample, and \hat{y}_{DimX_Mn} is the sentiment prediction label, which is denoted as $X_M Pre_label$.

(2) Missing Modalities Completion Strategies

For multimodal sentiment analysis, there are seven possible scenarios of uncertain missing modalities, which is illustrated in Table 2. For MSA, some works have presented that the accuracy of sentiment analysis based on the text is about 70%–80%, while it is about 60%–70% when based on the video or audio modality [14]. That is, the text modality contains more useful information for sentiment analysis. Inspired by this phenomenon, we categorize different modality missing situations into two categories, which are situation that the text modality is not missing and the situation that the text modality is missing, then, we develop two strategies of modalities missing completion for the above two situations.

Strategy I: Text-based missing modalities completion. For uncertain missing multimodal data, when the text modality is available, the visual modal is missing, or the audio modal is missing, or the other two modalities are missing simultaneously, we take the text modality as the basis to find the similar full modalities data. The process of text-based missing modalities completion strategy is described as follows. We first compute the low-dimensional representation $DimX_t$ of the text modality, and then predict the sentimental label $X_t Pre_label$ of the text modality with PTUSA. Next, we compute the similarities between $DimX_t$ and the low-dimensional representations of all text modalities in FMSD with the Cosine similarity function, which is defined as Eq. (9). Finally, we select three samples which have better similarities, select the sample data that has the same sentimental label with $X_t Pre_label$

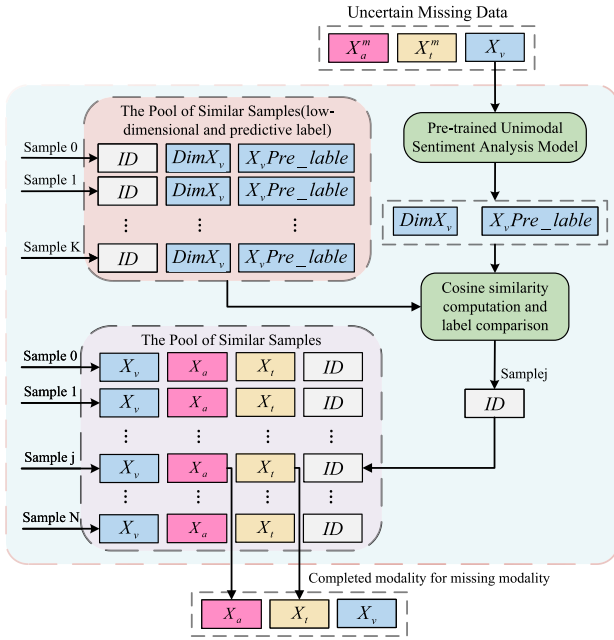


Fig. 5. Video-based missing modality completion.

from the three samples, and take the visual modality and the audio modality of the sample data to complete the missing visual modality or the missing audio modality. Otherwise, we complete the missing modalities with zeros. The process of this strategy is illustrated as Fig. 4.

$$\begin{aligned}
 \text{sim}(a, b) &= \cos(\theta) \\
 &= \frac{\langle a, b \rangle}{\|a\| \times \|b\|} \\
 &= \frac{\sum_{i=1}^D (a_i \cdot b_i)}{\sqrt{\sum_{i=1}^D a_i^2} \times \sqrt{\sum_{i=1}^D b_i^2}}.
 \end{aligned} \quad (9)$$

where θ denotes the angle between two vectors, and D is the dimension of the problem.

Strategy II: Visual or audio-based missing modalities completion. Recently, existing research has proved that when the text modality is missing, the accuracy of sentiment analysis based on video or audio is around 55%–70% [7]. Therefore, the predicted labels of the audio or the video modalities of X_M^m will have large differences with the real labels of the similar samples in the database FMSD, this will bring difficulty for finding the suitable similar samples. For this problem, we develop the visual/audio based missing modality completion strategy. The main idea of this strategy is as follows:

When the audio and video modality exist, and the text modality is missing, we take the audio modality as the basis to find the similar samples from the database FMSD, and then use the similar samples to complete the missing text modality. When the audio modality exists, and the video and text modality are missing, we take the audio modality as the basis to find the similar samples from the database FMSD, and apply the similar samples to complete the missing text and video modality. When the video modality exists, the text and audio modality are missing, we take the video modality as the basis to find the similar samples from the database FMSD, and apply the similar samples to complete the missing text modality and audio modality.

The process of Strategy II is presented as follows: we first compute the low-dimensional representation $\text{Dim}X_v$ and the sentiment prediction label $X_v\text{Pre_label}$ of the video modality through PTUSA. Next, we compute the cosine similarity between $\text{Dim}X_v$ and the low-dimensional representations of all video modalities in FMSD. Finally,

we select three samples which have better similarities; then, we select the sample data that has the same predicted sentimental label with the $X_v\text{Pre_label}$ from the three samples, and take the text modality and the audio modality of the sample data to complete the missing text modality or the missing audio modality. Otherwise, we complete the missing modalities with zeros. Strategy II is illustrated in Fig. 5. The audio-based missing modality completion module is similar to video-based. Our proposed missing modalities completion strategies are fully described in Algorithm 1.

3.5. Multimodal features' fusion

In MSA, it can get the best sentiment analysis result based on the text modality [14], therefore, to enhance the quality of the other two modalities, we propose to fuse the feature of text into the video and audio modalities with the Transformer encoder and the pre-trained model. The process of multimodal features' fusion is described as follows.

After completing the multimodality data with missing modalities with our proposed strategies, we obtain the completed visual, audio, and text modality (denoted as X_v , X_a and X_t). Then, we input each modality data into a fully connected layer for dimensional transformation, and each modality is converted into $X_v \in \mathbb{R}^{l_v \times d}$, $X_a \in \mathbb{R}^{l_a \times d}$, $X_t \in \mathbb{R}^{l_t \times d}$. Next, the representation of each modality are learned with the transformer encoder, which is illustrated as Eqs. (10)–(12):

$$E_{vm} = \text{MultiHead}(X_v, X_v, X_v) \quad (10)$$

$$E_{am} = \text{MultiHead}(X_a, X_a, X_a) \quad (11)$$

$$E_{tm} = \text{MultiHead}(X_t, X_t, X_t). \quad (12)$$

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_{vm} , X_{am} and X_{tm} denote the visual, audio and text modalities, respectively.

Subsequently, the residual connection is used to extract the feature of each modality and the layernorm layer is applied to normalize the data. The calculation process is illustrated as Eqs. (13)–(15):

$$E_v = \text{Layernorm}(X_v + E_{vm}) \quad (13)$$

$$E_a = \text{Layernorm}(X_a + E_{am}) \quad (14)$$

$$E_t = \text{Layernorm}(X_t + E_{tm}). \quad (15)$$

Next, the normalized unimodal features are linearly transformed in the position feed-forward sublayer, and the representations of each modalities is obtained. This process can be illustrated as Eqs. (16)–(18):

$$E_v = \text{Relu}(E_v W_{vl}^1 + b_{vl}^1) W_{vl}^2 + b_{vl}^2 \quad (16)$$

$$E_a = \text{Relu}(E_a W_{al}^1 + b_{al}^1) W_{al}^2 + b_{al}^2 \quad (17)$$

$$E_t = \text{Relu}(E_t W_{tl}^1 + b_{tl}^1) W_{tl}^2 + b_{tl}^2. \quad (18)$$

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

Then, we fuse the text into the visual and audio modality. In this process, we take the encoded textual modality as the Key and Value in the multi-head attention mechanism, and take the features of encoded visual and audio as the Query of the multi-head attention mechanism. The fusion process is illustrated as Eqs. (19) and (20):

$$E_{vfm} = \text{MultiHead}(E_v, E_t, E_t) \quad (19)$$

$$E_{afm} = \text{MultiHead}(E_a, E_t, E_t). \quad (20)$$

Algorithm 1 : Missing Modality Completion Strategies**Input:** Multimodal data with uncertain missing modalities X_M^m .**Output:** The completed multimodal data: $[X_v, X_a, X_t]$.

```

1: Screening the full modality samples from the dataset to obtain the pool of
   similar samples
2: Produce the similar pool with  $K$  groups of low-dimensional representations
   and sentiment prediction labels according to the PTUSA model
3:   if text modality is available then
4:     Obtain the low-dimensional representation  $DimX_{tt}$  and the
     predictive label  $X_{tt}Pre\_label$  of  $X_t$  by Eqs. (4)-(7)
5:      $DimX_{tt}, X_{tt}Pre\_label \leftarrow PTUSA(X_t)$ 
6:     Calculate three samples in the pool that most similar to  $DimX_{tt}$ 
     by Eq. (9)
7:      $Top_3 \leftarrow \text{Cosine Similarity}(DimX_{tt}, KDimX_t)$ 
8:     Compare the  $X_{tt}Pre\_label$  and every  $X_iReal\_label$  in  $Top_3$ 
9:      $j, ID \leftarrow \text{Compare}(X_{tt}Pre\_label, X_iReal\_label)$ 
10:    if  $j$  is not null then
11:      Completing missing modalities according to index  $j$ 
12:       $X_v \leftarrow \text{Pool}_V(j)$ , when  $X_v$  is missing
13:       $X_a \leftarrow \text{Pool}_A(j)$ , when  $X_a$  is missing
14:    else
15:       $X_v \leftarrow [0, 0, \dots, 0]$ , when  $X_v$  is missing
16:       $X_a \leftarrow [0, 0, \dots, 0]$ , when  $X_a$  is missing
17:    elif audio modality is available then
18:      Obtain the low-dimensional representation  $DimX_{aa}$  and the
      predictive label  $X_{aa}Pre\_label$  of  $X_a$  by Eqs. (4)-(7)
19:       $DimX_{aa}, X_{aa}Pre\_label \leftarrow PTUSA(X_a)$ 
20:      Calculate three samples in the pool that most similar to  $DimX_{aa}$ 
      by Eq. (9)
21:       $Top_3 \leftarrow \text{Cosine Similarity}(DimX_{aa}, KDimX_a)$ 
22:      Compare the  $X_{aa}Pre\_label$  and every  $X_iPre\_label$  in  $Top_3$ 
23:       $j, ID \leftarrow \text{Compare}(X_{aa}Pre\_label, X_iPre\_label)$ 
24:      if  $j$  is not null then
25:        Completing missing modalities according to index  $j$ 
26:         $X_v \leftarrow \text{Pool}_V(j)$ , when  $X_v$  is missing
27:         $X_t \leftarrow \text{Pool}_T(j)$ , when  $X_t$  is missing
28:      else
29:         $X_v \leftarrow [0, 0, \dots, 0]$ , when  $X_v$  is missing
30:         $X_t \leftarrow [0, 0, \dots, 0]$ , when  $X_t$  is missing
31:      else
32:        Obtain the low-dimensional representation  $DimX_{vv}$  and the
        predictive label  $X_{vv}Pre\_label$  of  $X_v$  by Eqs. (4)-(7)
33:         $DimX_{vv}, X_{vv}Pre\_label \leftarrow PTUSA(X_v)$ 
34:        Calculate three samples in the pool that most similar to  $DimX_{vv}$ 
        by Eq. (9)
35:         $Top_3 \leftarrow \text{Cosine Similarity}(DimX_{vv}, KDimX_v)$ 
36:        Compare the  $X_{vv}Pre\_label$  and every  $X_iPre\_label$  in  $Top_3$ 
37:         $j, ID \leftarrow \text{Compare}(X_{vv}Pre\_label, X_iPre\_label)$ 
38:        if  $j$  is not null then
39:          Completing missing modalities according to index  $j$ 
40:           $X_a \leftarrow \text{Pool}_A(j)$ , when  $X_a$  is missing
41:           $X_t \leftarrow \text{Pool}_T(j)$ , when  $X_t$  is missing
42:        else
43:           $X_a \leftarrow [0, 0, \dots, 0]$ , when  $X_a$  is missing
44:           $X_t \leftarrow [0, 0, \dots, 0]$ , when  $X_t$  is missing
45:        end if
46:      Return  $[X_v, X_a, X_t]$ 
47: End

```

Next, the representations of the three modalities are updated according to Eqs. (21)–(24):

$$E_{vt} = \text{Layernorm}(E_v + E_{vtm}) \quad (21)$$

$$E_{at} = \text{Layernorm}(E_a + E_{atm}) \quad (22)$$

$$E_{vt} = \text{Relu}(E_{vt}W_{vtl}^1 + b_{vtl}^1)W_{vtl}^2 + b_{vtl}^2 \quad (23)$$

$$E_{at} = \text{Relu}(E_{at}W_{atl}^1 + b_{atl}^1)W_{atl}^2 + b_{atl}^2. \quad (24)$$

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

To maintain the original features of the textual modality during multimodal feature fusion, we take the encoded textual modality as the Query, Key and Value of the multi-head attention mechanism. The calculation process is described as Eq. (25) as follows:

$$D_{tm} = \text{MultiHead}(E_t, E_t, E_t). \quad (25)$$

Similar to the encoder, modality representations can be updated with Eqs. (26)–(27):

$$D_t = \text{Layernorm}(E_t + D_{tm}) \quad (26)$$

$$D_t = \text{Relu}(D_tW_{tl}^1 + b_{tl}^1)W_{tl}^2 + b_{tl}^2. \quad (27)$$

where W_{tl} is the weight matrices, and b_{tl} is the bias.

Then, we concatenate the fused visual, fused audio and encoded text modality as an intermediary to approximate the common joint representation to the joint features of the complete modalities. The process can be illustrated as Eq. (28):

$$C_{all} = [E_{vt} \| E_{at} \| E_t]. \quad (28)$$

where “ $\|$ ” denotes the concatenation operation.

3.6. Sentiment prediction

To obtain the accurate sentiment prediction, considering the advantage of the textual modality, with the decision level fusion strategy, we compute the score of the sentiment label \hat{y} by computing the weighted sum of predicted results of each modality. Specifically, we feed E_{vt} , E_{at} , and E_t into a fully connected network with the softmax function to compute the prediction score \hat{y}_{vt} , \hat{y}_{at} , and \hat{y}_t , respectively, which is illustrated as Eqs. (29)–(31):

$$\hat{y}_{vt} = \text{softmax}(W_{vt}E_{vt} + b_{vt}) \quad (29)$$

$$\hat{y}_{at} = \text{softmax}(W_{at}E_{at} + b_{at}) \quad (30)$$

$$\hat{y}_t = \text{softmax}(W_tE_t + b_t). \quad (31)$$

where W_{vt} , W_{at} , and W_t are the weight matrices, and b_{vt} , b_{at} , and b_t are the biases. Subsequently, we assign the modalities with different learnable weights to fully utilize the textual modality. The final prediction score can be calculated according to Eq. (32):

$$\hat{y} = \text{softmax}(W_{dvt}\hat{y}_{vt} + W_{dat}\hat{y}_{at} + W_{dt}\hat{y}_t). \quad (32)$$

where W_{dvt} , W_{dat} , and W_{dt} are the weight matrices assigned to each modality. To obtain more effective weights, we employ model self-learning to obtain the weights of the three modalities, which can better leverage the advantages of video and audio modalities as auxiliary models and text modalities as the main modality.

3.7. Training objective

To train our proposed model well, we define the whole loss of our proposed model based on the loss of classification module, the loss of the pre-trained model and the loss of the decoder module. Thus, the loss function of SMCMSA is defined as Eq. (33):

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{pretrain} + \lambda_2 \mathcal{L}_{de}. \quad (33)$$

where \mathcal{L}_{cls} indicates the loss of classification module, $\mathcal{L}_{pretrain}$ denotes the loss of the pre-trained model, \mathcal{L}_{de} means the loss of decoder module, λ_1 and λ_2 denote weights of the pre-trained loss and the decoder loss.

In this work, the Jensen–Shannon (JS) divergence is adopted to compute the loss. JS consists of two asymmetric Kullback–Leibler (KL) divergences. The KL divergence and JS divergence are defined as Eqs. (34) and (35):

$$D_{KL}(p||q) = \sum_{i=1}^N p(x_i) \cdot \frac{p(x_i)}{q(x_i)} \quad (34)$$

$$JS(p||q) = \frac{1}{2} D_{KL}(p||q) + \frac{1}{2} D_{KL}(q||p) \quad (35)$$

where p and q are two probability distributions. The pre-trained loss, the decoder loss and the classification loss are introduced as follows.

(1) Loss of the Pre-trained model ($\mathcal{L}_{pretrain}$): which is used to approximate the intermediate feature (C_{all}) to the joint features (E_{pre}) of the full modalities produced by the pre-trained model. The pre-trained model (illustrated in Fig. 2) is trained based on the full modalities. The loss of the pre-trained model is defined as Eq. (36):

$$\begin{aligned} \mathcal{L}_{pretrain} &= JS(C_{all}||E_{pre}) \\ &= \frac{1}{2} D_{KL}(C_{all}||E_{pre}) + \frac{1}{2} D_{KL}(E_{pre}||C_{all}). \end{aligned} \quad (36)$$

(2) Loss of the Decoder module (\mathcal{L}_{de}): which is used to maintain the original features of the textual modality during multimodal feature fusion. The loss function is defined as Eq. (37):

$$\begin{aligned} \mathcal{L}_{de} &= JS(D_{out}||C_{all}) \\ &= \frac{1}{2} D_{KL}(D_{out}||C_{all}) + \frac{1}{2} D_{KL}(C_{all}||D_{out}). \end{aligned} \quad (37)$$

(3) Loss of the Classification module (\mathcal{L}_{cls}): Here, we employ the standard cross-entropy loss function for computing the classification loss. The loss function is defined as Eq. (38):

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{n=1}^N y_n \log \hat{y}_n \quad (38)$$

where N denotes the number of samples, y_n denotes the label of the n th sample, and \hat{y}_n indicates the predicted label. Based on the above description, the whole process of our proposed model SMCMSA is described as Algorithm 2.

4. Experiments

To verify the performance of our proposed model SMCMSA, we carried out extensive experiments on two public benchmark datasets (CMU-MOSI [43] and IEMOCAP [44]). In the following sections, we first introduce the two public datasets and the process of data pre-processing. Then, we introduce the experimental settings and baseline models. Finally, we present and analyze the experimental results.

4.1. Benchmark datasets

As is known to all, CMU-MOSI and IEMOCAP are two public datasets that used for testing MSA models, so, we also adopted these two datasets as the benchmark datasets. Next, we will introduce the two datasets and features extraction process in detail as follows.

CMU-MOSI: Dataset CMU-MOSI (Carnegie Mellon University Multimodal Opinion Sentiment and Intensity) contains 2199 short monologue video clips, which are extracted from 93 movie videos in YouTube. Each sample in CMU-MOSI is marked with an emotional score that ranges in $[-3, 3]$.

IEMOCAP: Dataset IEMOCAP (Interactive Emotional Dyadic Motion Capture) is constructed based on emotional dialogues and interactions between actors. IEMOCAP includes 5 sessions, each session contains about 30 videos, and each video comprises at least 24 utterances. The annotated labels in IEMOCAP are: neutral, frustration, anger, sad, happy, excited, surprise, fear, disappointing, and others.

In this paper, we conducted three-classification experiment on the CMU-MOSI dataset, and map the sentiment values to labels as negative,

Algorithm 2 :Similar Modalities Completion-based MSA

Input: Multimodal data with uncertain missing modalities in Table 2.

Output: The predicted sentiment category \hat{y} .

- 1: Screening the full modality samples from the dataset to obtain the pool of similar samples
- 2: Produce the similar pool with M low-dimensional representations and sentiment prediction labels according to Eqs. (5)-(7)
- 3: **Phase I. In the missing modality completion module**
- 4: Return $[X_v, X_a, X_t]$
- 5: $X_m \leftarrow \text{Dense}(X_m), m \in \{v, a, t\}$
- 6: **Phase II. In the multimodal feature fusion module**
- 7: Encoder: Produce the fused representation E_{vt}, E_{at}, E_t of each modality according to Eqs. (10)-(24)
- 8: $E_{vt} \leftarrow \text{encoder}(X_v, X_t, X_t)$
- 9: $E_{at} \leftarrow \text{encoder}(X_a, X_t, X_t)$
- 10: $E_t \leftarrow \text{encoder}(X_t, X_t, X_t)$
- 11: Decoder: Produce the decoded representation D_t according to Eqs. (25)-(27)
- 12: $D_t \leftarrow \text{decoder}(E_t, E_t, E_t)$
- 13: Calculate the fusion loss using Eq. (37)
- 14: $\mathcal{L}_{de} \leftarrow JS(D_t||E_t)$
- 15: Calculate the missing joint features using Eq. (28)
- 16: $C_{all} \leftarrow [E_{vt}||E_{at}||E_t]$
- 17: Calculate the loss between the output of the pre-trained model (E_{pre}) and the output of the transformer encoder (C_{all}) using Eq. (36)
- 18: $\mathcal{L}_{pretrain} \leftarrow JS(E_{out}||C_{all})$
- 19: **Phase III. In the sentiment prediction module**
- 20: Train the whole model using Eq. (33)
- 21: Calculate the prediction score \hat{y} using Eq. (32)
- 22: $\hat{y} \leftarrow \text{softmax}(W_{dvt}\hat{y}_{vt} + W_{dat}\hat{y}_{at} + W_{dt}\hat{y}_t)$
- 23: Return \hat{y}
- 24: End

Table 3

Distribution of three-class labels in the CMU-MOSI dataset and two-class labels in the IEMOCAP dataset.

Category		Neg.	Neu.	Pos.
three-classes	Train	914	90	996
	Test	108	7	84
two-classes	Train	3219	–	1299
	Test	859	–	337

neutral and positive as follows: map $[-3, 0)$ to negative, map $[0, 3]$ to neutral, and map $(0, 3]$ to positive. Meanwhile, on the IEMOCAP dataset, we conduct two-classification experiment and map the sentiment labels into positive and negative labels as follows: map [frustrated, angry, sad, fearful, disappointed] to negative, and map [happy, excited] to positive. The information of three-class labels in CMU-MOSI and two-class labels in the IEMOCAP are described in Table 3.

4.2. Preprocessing of the two datasets

The feature extraction for each modality in the two benchmark datasets is introduced as follows [18]. For the video modality, a series of features from the human face images are extracted with the OpenFace2.0 toolkit [45]. The features of the video include timestamp, confidence level, recognition success mark, eye movement, head pose and facial movement. The dimension of a visual feature is 709. For the text modality, its features are learned by a pre-trained BERT model [46], and the dimension of a textual feature is 768. The features of the audio modalities are extracted with Librosa [47]. Each audio is mixed into mono and resampled to 16,000 Hz. Moreover, each frame is separated by 512 samples, and zero-crossing rate, Mel-frequency Cepstral Coefficients (MFCC) and Constant Q Transform (CQT) features are selected to represent audio segments. Finally, these three features are concatenated to produce 33-dimensional audio features. In this

Table 4
Parameter settings of SMCMSA.

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	0.1

work, we apply the above pre-processing procedure for all baselines and adopt the preprocessed data provided by [18] to conduct experiments.

4.3. Experimental settings

In this work, we conduct our experiments on a personal computer which is configured as follows. Operating System (OS): Windows 10, CPU: Intel(R) Core(TM) i9-10900K, Graphics Processing Unit (GPU): Nvidia 3090, and a generous 96 GB of RAM. The framework of our proposed model is TensorFlow 1.14.0, and the programming language is Python 3.6. The hyperparameters parameters of our proposed model adopt the settings provided in [18] as described in Table 4. Specifically, the learning rate (lr) is set as 0.001, the batch size (b) is set as 32, and the hidden size (d) is set as 300. To optimize our proposed model, we adopted the Adam optimizer [48] to minimize the overall loss function \mathcal{L} . The epoch size is set as 20, the loss weight is set as 0.1.

To evaluate the performance of our proposed model, the accuracy (Acc) and macro-F1 score (M-F1) are selected as the evaluation metrics. The calculation formulas for Acc and M-F1 are defined as Eqs. (39) and (40):

$$\text{Acc} = \frac{N_{\text{true}}}{N} \quad (39)$$

$$\text{M-F1} = \frac{2PR}{P + R} \quad (40)$$

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

4.4. Baseline models

To demonstrate the effectiveness of our proposed model SMCMSA, ten state-of-the-art models are chosen as the baseline models. The introduction of the ten models are presented as follows:

- AE [49]: A generalized framework for the study of linear and nonlinear self-encoders designed to make the output of a neural network as consistent as possible with the input.
- CRA [15]: A missing modality reconstruction model based on the cascading residual structure of an autoencoder, which approximates the input data by employing a residual connection mechanism.
- MCTN [37]: This method utilizes modality translation to perform inter-modal interactions, which is useful for learning the robust joint relationships.
- TransM [50]: A multimodal features fusion model based on end-to-end translation, which enables inter-modal interaction by performing cyclic translation between modalities.
- MMIN [39]: A feature reconstruction model for dealing with missing modalities, which utilizes a cascaded residual autoencoder as well as forward and backward imagining modules to transform the available and missing modalities into each other.

- ICDN [10]: This method combines consistency and difference networks to enable inter-modality interactions by mapping information from other modalities to the target modality through a cross-modality Transformer.
- MRAN [8]: This model utilizes multimodal and missing index embedding to instruct the features' reconstruction of missing modality, and aligns video and audio features with text features to solve the missing modality problem.
- TATE_C [18]: This model uses a tag-assisted transformer encoder to handle all the cases of uncertain missing modalities, and adopts a pre-trained model to instruct the learning of the joint representation.
- MTMSA [14]: A modality translation network that translate the visual and audio modalities into the textual modality, thus to handle the missing modalities and capture the deep interaction between different modalities. Moreover, this model has utilized the advantage of the text modality.
- TATE_J [7]: This model builds on previous research [18] by adding different weights for different modalities to fully take advantage of each modality.

4.5. Performance verification

Here, we verify model SMCMSA's performance through implementing 3-classification on dataset CMU-MOSI and 2-classifications on dataset IEMOCAP. This experiment includes two segments. In the first segment, the case of single modality missing is considered. While in the second segment, the case of multiple modalities missing is considered. Experimental results are shown in Tables 5 and 6, the best results of this experiment are bolded in the two tables. Moreover, results of models MTMSA, ICDN, and MRAN were selected from work [14], while results of the other seven models come from works [7,18]. Next, we will introduce the two experimental segments in detail.

Experiment on single modality missing. In this experiment, the modality missing rate is set from 0 to 0.5 with the step =0.1. Experimental results are detailed in Table 5. From Table 5 we can find that, on CMU-MOSI dataset, compared with the baseline models, model SMCMSA achieves the best scores in terms of ACC and M-F1 when the missing rate are set as 0.2, 0.3, 0.4, and 0.5. The situations where SMCMSA performs poorly are as follows. When the missing rate is set as 0, SMCMSA's M-F1 score is 0.67% less than model MMIN's M-F1 score, SMCMSA's ACC value is 0.6% less than that of model TATE_J. When missing rate = 0.1, SMCMSA's M-F1 score is 0.41% lower than that of model TATE_J, meanwhile, SMCMSA's ACC value is 0.43% lower than that of model TATE_J.

More importantly, on dataset IEMOCAP, in terms of ACC and M-F1, model SMCMSA gets the best performance than that of other ten baseline models on all the missing rates (0, 0.1, 0.2, 0.3, 0.4, and 0.5). Therefore, according to the above experimental results presented in Table 5, the conclusion that can be drawn is that the overall performance of model SMCMSA is better than that of other ten baseline models on the two datasets CMU-MOSI and IEMOCAP on multiple modality missing rates.

Experiment on multiple modalities missing. In this experiment, we set the modality missing rate from 0 to 0.5 with a step = 0.1. Experimental results are shown in Table 6. According to Table 6, it is easy to find that, on dataset CMU-MOSI, model SMCMSA achieves the best scores in terms of ACC and M-F1 when the values of missing rate are 0.1, 0.2, 0.3 and 0.5. Nevertheless, when the value of missing rate is 0.4, SMCMSA's M-F1 value is 0.27% lower than that of model TATE_J, SMCMSA's ACC value is 0.38% less than that of model TATE_J. Moreover, when the values of missing rate is 0.5, the ACC value of SMCMSA is 1.6% lower than that of model TATE_J.

Further more, on dataset IEMOCAP, SMCMSA outperforms other ten baseline models in terms of ACC and M-F1 when the values of modality missing rate are 0, 0.1, 0.2, and 0.5. However, when the value

Table 5

Experimental results for all the models with a single missing modality (the missing rate increases from 0 to 0.5).

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_C	58.32	84.90	58.21	84.46	55.46	81.25	55.11	80.73	54.11	80.21	51.71	74.04
	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
	TATE_J	60.18	86.54	58.49	85.85	56.52	82.65	56.08	81.77	54.67	80.53	52.66	76.04
	Ours	59.74	85.94	58.08	85.42	57.80	83.85	57.57	82.29	55.89	81.77	54.72	79.68
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_C	81.15	85.39	79.99	85.09	79.10	84.07	78.45	83.25	76.74	82.75	74.43	82.43
	MTMSA	81.36	86.14	81.81	85.24	81.47	84.46	80.20	84.28	79.53	82.94	75.84	82.55
	TATE_J	81.76	86.46	81.25	85.24	79.57	84.80	78.06	83.76	77.84	82.97	75.76	82.51
	Ours	84.17	87.84	82.10	85.64	81.92	85.05	80.21	84.80	79.99	83.78	77.64	83.36

Table 6

Experimental results for all the models with multiple missing modalities (the missing rate increases from 0 to 0.5).

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_C	58.32	84.90	56.38	81.77	54.87	81.07	52.12	77.60	51.19	76.56	51.15	73.23
	MTMSA	58.12	84.89	57.43	84.37	55.71	81.25	52.82	78.12	51.32	76.04	51.72	73.43
	TATE_J	60.18	86.54	57.37	82.81	55.62	81.67	53.65	78.12	53.28	77.98	51.92	76.60
	Ours	59.74	85.94	57.78	84.90	57.41	83.33	54.85	79.69	53.01	77.60	52.60	75.00
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_C	81.15	85.39	78.37	83.63	77.55	82.33	76.14	82.21	74.09	81.94	72.49	80.57
	MTMSA	81.36	86.14	80.28	85.17	80.39	84.12	79.30	83.85	76.07	83.07	74.80	82.03
	TATE_J	81.76	86.46	80.67	85.30	79.12	83.77	78.99	84.64	78.44	82.75	76.97	82.25
	Ours	84.17	87.84	81.55	85.81	80.55	84.38	78.29	82.93	74.47	80.49	77.35	82.85

of missing rate is set as 0.3, the SMCMSA's M-F1 value is 0.7% lower than that of model TATE_J, and SMCMSA's ACC value is 1.71% lower than that of model TATE_J. When the missing rate is set as 0.4, the M-F1 and ACC values of model SMCMSA are 3.97% and 2.26% lower than that of model TATE_J. According to the experimental results presented in Table 6, we can conclude that model SMCMSA outperforms the ten baseline models on datasets CMU-MOSI and IEMOCAP on multiple modality missing rates.

Theoretical Analysis. From Tables 5 and 6 we can find that, the performance of model MCTN and TransM are better than that of model AE and CRA, this is because model MCTN and TransM adopt the cyclic translation operation. Thus, above experimental results prove that the cyclic translation operation is effective in extracting and integrating information from diverse modalities. By comparing the performance

of our proposed model SMCMSA with that of model MCTN, MTMSA and TransM, we can find that SMCMSA outperforms other three models because it fuses the text modality into the video and audio modality, and adopts adaptive learning method to compute weights for different modalities in the decision fusion level, which are useful for enhancing the ability of our model.

Specifically, from Tables 5 and 6 we can find that, SMCMSA is not optimal on the CMU-MOSI dataset as the missing rate is low. When there is no missing data (that is, missing rate = 0), the M-F1 value of SMCMSA is 0.67%, lower than that of MMIN, and the ACC value of SMCMSA is 0.6%, lower than that of TATE_J. Moreover, SMCMSA has a decrease of 0.41% in M-F1 and a decrease of 0.43% in ACC compared with TATE_J when the missing rate is 0.1.

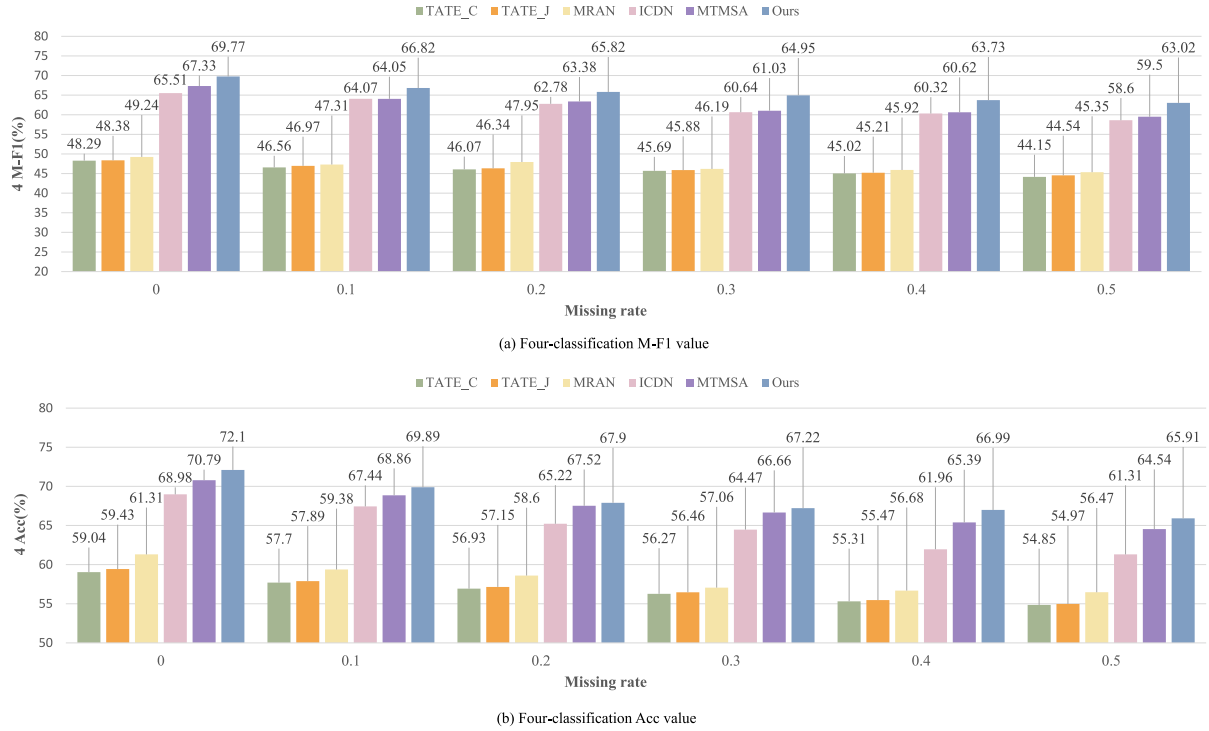


Fig. 6. Experimental results of six models on four-classification.

Table 7

Distribution of multi-class 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

Model MMIN and TATE_J encode multimodal splicing features, which can promote more deeply interactions between different modalities. Therefore, when there is no modality missing or the modality missing rate is low, the encoders of MMIN and TATE_J are able to model inter-modal relationships accurately, thus achieving the better performance. While the Multimodal Features' Fusion Module in SMCMSA mainly focuses on fusing textual information into video and audio modalities, which limits free and deep interaction of information between different modalities. Moreover, it is worth noting that as the rate of missing modalities rises, the baseline model experiences significant challenges in constructing an effective joint multimodal representation due to the influence of a large number of missing modalities. On the other hand, the Multimodal Features' Fusion Module in SMCMSA only incorporates textual modalities into video and audio modalities, thus can limit the information interaction between the missing modalities and the available modalities, which helps to obtain a more effective feature representation and can get better performance when modality missing rates increase.

When comparing the performance of model ICDN with that of other models, we can find that, as the missing rate is 0.4, the ACC and M-F1 of model ICDN exhibit a significant decrease on datasets CMU-MOSI and IEMOCAP. This is because model ICDN addresses the missing modalities through inter-modality mapping. However, when there are too many missing modalities, it becomes a challenge to map across different modalities effectively. So, with the increase of the values of missing rate, the performance of model ICDN is getting worse and worse.

Further more, from Tables 5 and 6 we can find, on datasets CMU-MOSI and IEMOCAP, the ACC and M-F1 values of model MRAN experience a substantial decrease when the missing rate is set as 0.5.

Since model MRAN projects the visual and audio features onto the text feature space, where the three modalities are learned to align closely with their corresponding emotional word embeddings, so as to ensure the visual and auditory features be consistent with the textual features. However, this intermodal feature projection will be constrained as the modality missing rate intensifies.

4.6. Multi-classification verification

To test SMCMSA's ability in multi-classification of sentiment, we take IEMOCAP as the benchmark dataset, and conduct experiments on 4-classification (happy, angry, sad and neutral) and on 7-classification (happy, anger, sad, neutral, frustration, excited, and surprise). The information of multi-class labels in IEMOCAP is described in Table 7. Here, we chose models TATE_C, TATE_J, MRAN, ICDN and MTMSA as the baseline models. Experimental results are presented in Figs. 6 and 7. The results of MTMSA, ICDN and MRAN are selected from work [14], while those of TATE_C and TATE_J are obtained from work [7,18], respectively.

Experimental results of the four-class sentiment classification are shown in Fig. 6. According to Fig. 6(a) and (b) it can be found that, when the missing rate is set as 0, our proposed model SMCMSA gets larger M-F1 and ACC than that of model ICDN by 4.26% and 3.12%. When the missing rate was 0.1, SMCMSA's M-F1 value is 19.85% higher than that of model TATE_J, and SMCMSA's ACC value is 12.00% larger than that of model TATE_J. When the missing rate is set as 0.5, compared with model MTMSA, SMCMSA achieves 3.52% improvement in terms of M-F1 and 1.37% improvement in terms of ACC.

Experimental results of the seven-class sentiment classification are illustrated in Fig. 7. From Fig. 7(a) and (b) it can be found that, as the missing rate = 0, our proposed model SMCMSA achieves 13.18% improvement in terms of M-F1 and gets 7.24% improvement in terms of ACC compared to model ICDN. When the missing rate is set as 0.2, the M-F1 value of SMCMSA is 21.82% higher than that of MRAN, and the ACC value of SMCMSA is 16.24% higher than that of MRAN. Compared with MTMSA, SMCMSA achieves a 12.61% improvement in M-F1 and gets 6.71% improvement in ACC when the missing rate is set as 0. Based

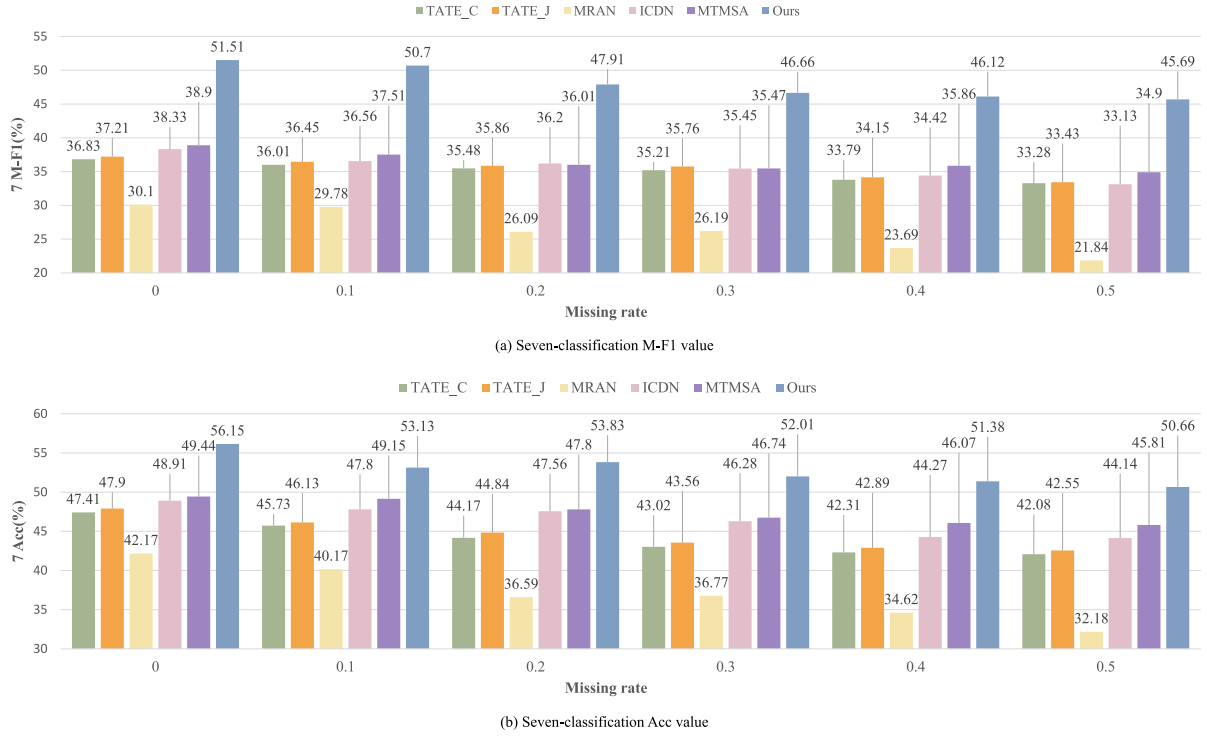


Fig. 7. Experimental results of six models on seven-classification.

on the above experimental results, we can conclude that SMCMSA has better performance in multi-classification of sentiment.

Furthermore, from Figs. 6 and 7 we can find that, the ACC values of model ICDN shows a significant decrease when the missing rate is set as 0.3 and 0.4 for both the four-class and seven-class sentiment classification, respectively. These results prove that when the proportion of modality missing is larger, model ICDN cannot capture the inter-modal interactions to compensate the missing modalities effectively.

Specifically, based on Figs. 6 and 7 we can get that, in the seven-class sentiment classification, the ACC values of model MRAN decrease significantly with the increases of the modality missing rates. This is because the cross-modality feature projection in the MRAN model becomes weak when the modality missing rates increase, thus affecting the alignment of video and audio features with the text features, and making it a challenge to reconstruct the features of the missing modality.

Results in Figs. 6 and 7 demonstrate that, the performance of all the six models become less efficient with the increase of the modality missing rates in the four-class and seven-class sentiment classifications. Importantly, our proposed model SMCMSA outperforms the other five baseline models when the missing rate takes different values. In all, according to the above experimental results we can get that our proposed model SMCMSA has the best performance in multi-class sentiment classification with uncertain missing modalities.

4.7. Ablation study

In this section, we introduce the modality ablation experiment and module ablation experiment on CMU-MOSI dataset. In these two experiments, we take “T” to denote the text modality, take “A” to represent the audio modality, and use “V” to denote the visual modality. Next, we will present the experimental settings and result analysis of these two experiments.

Modality ablation experiment: Here we consider three different scenarios which are as follows: (1) Take one modality to analyze the sentiment. In this situation, the results are achieved by directly using the Transformer encoder to extract features from the single

modality, and then performing sentiment classification subsequently. Because only a single modality is used, there is no scenarios of modality missing. So, the modality missing rate is set as 0; (2) Take any two modalities to conduct sentiment analysis (such as T+V, T+A, V+A). In this experiment, the modality missing rates are set as 0, 0.1, 0.2, 0.3, 0.4 and 0.5, respectively. For the combination of visual and audio (V+A), since the text modality is not involved, the visual and audio modalities are encoded by the Transformer encoder and then is input into the common space for concatenating without modality fusion; (3) Take three modalities (T+A+V) simultaneously for sentiment analysis. In this situation, the modality missing rates are set as 0, 0.1, 0.2, 0.3, 0.4 and 0.5, respectively.

Results of the modality ablation experiments are shown in Table 8, where the optimal outcomes are highlighted in bold. From Table 8 we can find that, in scenario (1), the best results are achieved with the text modality, where the ACC value of SMCMSA are 29.17% and 28.13% higher than that of SMCMSA when adopting the visual modality or audio modality for sentiment analysis. These results prove the dominance of text modality in multimodal sentiment analysis. For scenario (2), experimental results prove that bimodal combinations which include the text modality achieve better results than the bimodal combinations that do not include the text modality. Specifically, the ACC value of the bimodal combination without text modality is decreased by 25.52% compared to the bimodal combination without the visual modality or the audio modality. Based on the experimental results of scenario (3) we can find that, the results are the best when all the three modalities are adopted simultaneously. Furthermore, these results also prove that complementary features can be learned from multi modalities.

Module ablation experiment: To verify the effectiveness of the key modules of SMCMSA, we conducted some module ablation experiment. Here, four variants of SMCMSA were generated for performance comparison by removing different modules from model SMCMSA. The four variants are obtained as follows: (1) SMCMSA-PreL was produced by removing the predicted label from the missing modality completion module. (2) SMCMSA-SMC was obtained through cutting the missing modality completion module from SMCMSA. (3) SMCMSA-MFF was achieved through subtracting the multimodal feature fusion module

Table 8
Results of modality ablation experiment (the missing rate increases from 0 to 0.5).

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	31.81	55.21	–	–	–	–	–	–	–	–	–	–
A	35.01	56.25	–	–	–	–	–	–	–	–	–	–
T	57.12	84.38	–	–	–	–	–	–	–	–	–	–
V+A	38.59	58.33	37.71	57.81	36.52	56.25	35.48	54.69	34.81	54.17	34.13	53.65
V+T	56.34	83.33	55.26	80.73	54.55	79.69	54.36	79.17	53.47	74.48	50.04	72.92
A+T	56.87	83.85	55.61	81.25	55.23	80.73	54.19	79.17	54.63	75.52	52.87	73.44
V+A+T	59.74	85.94	58.08	85.42	57.80	83.85	57.57	82.29	55.89	81.77	54.72	79.68

Table 9
Results of model ablation experiment (the missing rate increases from 0 to 0.5).

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
SMCMSA-PreL	59.74	85.94	57.89	83.33	57.49	82.81	57.49	81.77	54.87	79.17	53.96	78.65
SMCMSA-SMC	59.74	85.94	56.75	82.81	56.70	82.29	56.48	81.77	53.63	78.65	51.92	77.60
SMCMSA-MFF	55.00	81.25	56.31	82.29	55.60	81.25	55.17	80.73	52.84	77.60	50.15	76.56
SMCMSA-Pre	59.74	85.94	57.80	83.55	56.60	81.25	54.89	80.21	51.19	77.08	51.02	74.45
SMCMSA	59.74	85.94	58.08	85.42	57.80	83.85	57.57	82.29	55.89	81.77	54.72	79.68

from SMCMSA. (4) Variant model SMCMSA-Pre was produced by subtracting the pre-trained module from SMCMSA. Experimental results of module ablation are presented in Table 9.

For the SMCMSA-PreL model, instead of matching the predicted sentiment labels in the three similar similarity samples, we take the most similar modality to complete the missing modality. From Table 9 it can be found that, the M-F1 value and the ACC value of SMCMSA-PreL are 1.02% and 2.6% lower than that of SMCMSA when the missing rate is set as 0.4. Moreover, the SMCMSA-PreL model has a decrease of 0.19% in M-F1 and a decrease of 2.09% in ACC compared with SMCMSA when the missing rate is set as 0. Based on experimental results presented in Table 9 we can obtain the conclusion that, using the predicted label for similar modality selecting is effective for improving the performance of our proposed model SMCMSA.

From Table 9 it can also be found that compared with SMCMSA, in terms of M-F1, SMCMSA-SMC gets a decrease about 2.8% when the missing rate is set as 0.5; in terms of ACC, SMCMSA-SMC obtains a decrease about 3.12% when the missing rate is set as 0.4. These experimental results can validate that the missing modality completion module is useful for improving the performance of SMCMSA.

Compared with SMCMSA, SMCMSA-MFF experiences a decrease of 4.74% in terms of M-F1 and a decrease of 4.69% in terms of ACC as the missing rate is 0. Moreover, SMCMSA-MFF gets a decrease of 4.17% in ACC when the missing rate is set to 0.4, SMCMSA-MFF gets the decrease (4.63%) in ACC as the missing rate is 0.5. Results presented in Table 9 can verify that the multimodal feature fusion module can enhance the level of features of multimodal data, and the multimodal feature fusion module has a significant contribution to enhance the performance of SMCMSA.

For SMCMSA-Pre, compared with SMCMSA, it experiences a decrease of 4.7% in terms of M-F1 when the missing rate is 0.4. Additionally, when the missing rate is 0.5, SMCMSA-Pre exhibits the most substantial decline, with a 5.23% decrease in terms of ACC. These experimental results demonstrate that the pre-trained module is useful for enhancing the performance of SMCMSA. In a word, from Table 9 we can find that, SMCMSA outperforms other four variants in all the cases of different modality missing rates. These experimental results can prove that the key modules of SMCMSA are effective, and the SMCMSA model is effective for MSA under uncertain missing modalities.

Verification of missing modalities completion with similar data: This experiment aims to verify that the quality of similar data selected by SMCMSA is superior to the quality of data generated by generative MSA models. In this experiment, we select three generative MAS models (CRA [15], TransM [50] and MMIN [39]) as the validate models. We take CMU-MOSI as the benchmark dataset and set the modality missing

rate from 0.1 to 0.5 with a step = 0.1. In this experiment, the performance of the three models is first tested with the original benchmark data (CMU-MOSI), and then is tested with the completed data that are produced by the Missing Modality Completion Module of model SMCMSA. The experimental results are presented in Table 10, where models tested with data produced by model SMCMSA are denoted as CRA_Sim, TransM_Sim, and MMIN_Sim.

From Table 10 we can find that, CRA_Sim experiences an increase of 2.15% in terms of M-F1 when the missing rate is 0.3 compared with CRA. Additionally, when the missing rate is 0.5, CRA_Sim exhibits the most substantial improvement, with a 7.63% increase in terms of ACC. Compared with TransM, TransM_Sim experiences an improvement of 5.56% in terms of M-F1 and an improvement of 4.69% in terms of ACC as the missing rate is 0.1. TransM_Sim gets an improvement of 10.37% in ACC when the missing rate is 0.5, TransM_Sim gets the improvement (6.77%) in ACC as the missing rate is 0.3 and 0.4. For MMIN, the M-F1 value and the ACC value of MMIN are 4.78% and 8.93% lower than that of MMIN_Sim when the missing rate is set as 0.5. Moreover, the MMIN_Sim model has an improvement of 2.74% in M-F1 and an improvement of 1.99% in ACC compared with MMIN when the missing rate is 0.1.

From Table 10 we can also find that, for all the cases of modality missing rate (from 0.1 to 0.5 with a step = 0.1), when tested with the data produced by model SMCMSA, the performance of the three models is better than that of the three models when tested with the original data. Therefore, based on these experimental results we can conclude that, the quality of the similar data selected by SMCMSA is superior to the quality of data generated by MSA models.

5. Conclusion

To better solve the problem of multimodal sentiment analysis (MSA) with uncertain missing modalities, in this work, we propose a Similar Modality Completion-based Multimodal Sentiment Analysis model (named SMCMSA). SMCMSA propose to complete the missing modalities with the most similar modalities which are found from the full modalities database, and experimental results prove that using the real and similar modalities to fill in the missing modalities is effective. Furthermore, SMCMSA fuses the text modality into the video and audio modalities with our proposed multimodal feature fusion module, this is useful for enhancing the quality of multimodal features, thus to improving the effect of multimodal sentiment analysis. Moreover, SMCMSA adopts the pre-trained model to guide the intermediate feature approximate to the joint features of the full multimodalities, it is also useful for resolving missing modalities that are filled with zeros. Finally, SMCMSA

Table 10

The experimental results for generative models based on similar modality completion data (the missing rate increases from 0.1 to 0.5).

Datasets	Models	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
CMU-MOSI	CRA	54.37	79.38	53.57	78.24	51.67	72.84	51.02	73.79	45.38	69.45
	CRA_Sim	56.41	83.33	55.35	81.77	53.82	79.69	52.77	78.13	52.28	77.08
CMU-MOSI	TransM	57.53	79.69	55.21	78.42	52.87	72.92	52.49	72.40	45.86	68.23
	TransM_Sim	63.09	84.38	56.10	82.81	53.83	79.69	53.27	79.17	53.30	78.65
CMU-MOSI	MMIN	57.75	81.86	55.38	80.20	53.65	79.24	52.55	76.33	48.95	70.76
	MMIN_Sim	60.49	83.85	55.69	82.29	54.90	81.25	54.10	80.21	53.73	79.69

applies the loss of classification, loss of pre-trained model and the loss of encoder to supervise the learning process of model SMCMSA. Based on two public popular benchmark datasets CMU-MOSI and IEMOCAP, we conduct extensive experiments and analysis. Experimental results prove the effectiveness of our proposed model SMCMSA. In our future works, we will develop a more effective MSA model without a pre-trained model, thus to enable our model more suitable for practical application scenarios.

CRedit authorship contribution statement

Yuhang Sun: Writing – original draft, Visualization, Validation, Methodology. **Zhizhong Liu:** Writing – original draft, Resources, Methodology, Investigation, Conceptualization. **Quan Z. Sheng:** Writing – review & editing, Investigation, Conceptualization. **Dianhui Chu:** Visualization, Validation, Formal analysis. **Jian Yu:** Writing – review & editing, Validation, Conceptualization. **Hongxiang Sun:** Writing – review & editing, Validation.

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, 61832004), (Grant no. ZR2020KF019), the Special Funding Program of Shandong Taishan Scholars Project, China, Australian Research Council (ARC) Future Fellowship, Australia FT140101247 and Discovery Project, Australia DP200102298.

References

- [1] Bo Yang, Bo Shao, Lijun Wu, Xiaola Lin, Multimodal sentiment analysis with unidirectional modality translation, *Neurocomputing* 467 (2022) 130–137.
- [2] P.D. Mahendhiran, Kannimuthu Subramanian, CLSA-CapsNet: Dependency based concept level sentiment analysis for text, *J. Intell. Fuzzy Systems* (2022) 1–17, Preprint.
- [3] 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.
- [4] Sanjeev Verma, Sentiment analysis of public services for smart society: Literature review and future research directions, *Gov. Inf. Q.* 39 (3) (2022) 101708.
- [5] Delali Kwasi Dake, Esther Gyimah, Using sentiment analysis to evaluate qualitative students' responses, *Educ. Inf. Technol.* (2023).
- [6] Samir Lamouri Angie Nguyen, Béranger Lekens, Managing demand volatility of pharmaceutical products in times of disruption through news sentiment analysis, *Int. J. Prod. Res.* 61 (9) (2023) 2829–2840.
- [7] Jiandian Zeng, Jiantao Zhou, Tianyi Liu, Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities, *IEEE Trans. Multimed.* (2022).
- [8] 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.
- [9] 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).
- [10] 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.
- [11] Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaying Liu, Jianwu Dang, Context-and knowledge-aware graph convolutional network for multimodal emotion recognition, *IEEE MultiMedia* 29 (3) (2022) 91–100.
- [12] Yuntao Shou, Tao Meng, Wei Ai, Sihang Yang, Keqin Li, Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis, *Neurocomputing* 501 (2022) 629–639.
- [13] Hao Sun, Jiaqing Liu, Yen-Wei Chen, Lanfen Lin, Modality-invariant temporal representation learning for multimodal sentiment classification, *Inf. Fusion* 91 (2023) 504–514.
- [14] Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, Lingqiang Meng, Modality translation-based multimodal sentiment analysis under uncertain missing modalities, *Inf. Fusion* 101 (2024) 101973.
- [15] 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.
- [16] Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis, *IEEE Trans. Affect. Comput.* (2023).
- [17] Mingcheng Li, Dingkang Yang, Lihua Zhang, Towards robust multimodal sentiment analysis under uncertain signal missing, *IEEE Signal Process. Lett.* (2023).
- [18] 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.
- [19] Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Rohit Prasad, Ensemble of SVM trees for multimodal emotion recognition, in: *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1–4.
- [20] Nicholas Cummins, Shahin Amiriparian, Sandra Ottl, Maurice Gerczuk, Maximilian Schmitt, Björn Schuller, Multimodal bag-of-words for cross domains sentiment analysis, in: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2018, pp. 4954–4958.
- [21] P.M. Arunkumar, Soundararajan Chandramathi, S. Kannimuthu, Sentiment analysis-based framework for assessing internet telemedicine videos, *Int. J. Data Anal. Tech. Strategies* 11 (4) (2019) 328–336.
- [22] Sijie Mai, Songlong Xing, Haifeng Hu, Analyzing multimodal sentiment via acoustic-and visual-istm with channel-aware temporal convolution network, *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29 (2021) 1424–1437.
- [23] 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.
- [24] Chunjun Zheng, Chunli Wang, Ning Jia, Emotion recognition model based on multimodal decision fusion, *J. Phys. Conf. Ser.* 1873 (1) (2021) 012092.
- [25] Sijie Mai, Ying Zeng, Shuangjia Zheng, Haifeng Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, *IEEE Trans. Affect. Comput.* (2022).
- [26] Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, Xuemei Luo, TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis, *Pattern Recognit.* 136 (2023) 109259.
- [27] Kyeonghun Kim, Sanghyun Park, AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis, *Inf. Fusion* 92 (2023) 37–45.
- [28] Ashima Yadav, Dinesh Kumar Vishwakarma, A deep multi-level attentive network for multimodal sentiment analysis, *ACM Trans. Multimedia Comput. Commun. Appl.* 19 (1) (2023).
- [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] 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.
- [32] 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.
- [33] 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).
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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, (no. 01) 2019, pp. 6892–6899.
- [38] 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.
- [39] 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.
- [40] Wei Peng, Xiaopeng Hong, Guoying Zhao, Adaptive modality distillation for separable multimodal sentiment analysis, *IEEE Intell. Syst.* 36 (3) (2021) 82–89.
- [41] Haozhe Chi, Minghua Yang, Junhao Zhu, Guan hong Wang, Gaoang Wang, Missing modality meets meta sampling (M3S): An efficient universal approach for multimodal sentiment analysis with missing modality, 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 (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 121–130.
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [43] 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.
- [44] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, Shrikanth S. Narayanan, IEMO-CAP: Interactive emotional dyadic motion capture database, *Lang. Resour. Eval.* 42 (4) (2008) 335–359.
- [45] Tadas Baltrušaitis, 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.
- [46] 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*.
- [47] 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.
- [48] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, *arXiv preprint arXiv:1412.6980*.
- [49] 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.
- [50] 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.