Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Extracting method for fine-grained emotional features in videos

Cangzhi Zheng [a], Junjie Peng [a,b,*], Zesu Cai [c]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[b] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China
[c] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Fine-grained feature extraction
Non-verbal modalities

ABSTRACT

Multimodal Sentiment Analysis (MSA) has significant applications in social media analysis and healthcare. It utilizes features from multiple modalities e.g., video, general text, acoustics, and vision, to obtain more credible sentiment analysis results. However, previous studies have ignored substantial variations in sub-features emerging within both the acoustic and visual modalities during feature extraction. Instead, they primarily rely on a simple concatenation method to derive representations. Consequently, these approaches prevent feature extractors from effectively leveraging non-verbal (acoustic and visual) features, thereby, constraining the model's overall performance. To solve this problem, this study proposes a method for extracting fine-grained emotional features from videos. By segregating the initial features of non-verbal modalities into distinct domains, then separately modeling and uniformly re-integrating them, our method effectively exploits the modality-specific information in these original features. Through extensive experiments, the performance of all models significantly improves using the features extracted following our method compared with original ones. This substantiates that our proposed approach more effectively utilizes features from non-verbal modalities compared with conventional approaches. This also underscores that processing non-verbal sub-features separately before integration represents a viable solution for enhancing the performance of the MSA model.

1. Introduction

Sentiment analysis is a technology that exploits computer programs to automatically identify the sentiments, attitudes, and emotions in information expressed by humans [1]. It can help people better analyze and understand large amounts of data in the service industry and help businesses and organizations better communicate and interact with customers [2]. Early methods primarily relied on unimodal data, such as text, speech, and facial expressions. Using only a single modality for analysis can result in deviations in sentiment prediction. Therefore, these methods often require additional modalities to leverage the comprehensive contextual information and obtain accurate results [3].

Multimodal Sentiment Analysis (MSA) is a rapidly growing research field that aims to improve analytical performance by leveraging the features of multiple modalities in multimedia data [4,5]. Recently, researchers explored fusion methods for features extracted from different modalities to obtain more accurate sentiment predictions [6–11]. By integrating multimodal information, models can capture broader contexts, thereby improving their performance in predicting sentiment.

Models designed for MSA typically comprise three key stages: feature extraction, modality fusion, and label prediction. Feature extraction derives effective features from each mode, modality fusion facilitates meaningful interactions between them, and label prediction transforms abstract features into task-specific labeled data. Feature extraction is as important as modality fusion and label prediction because together they collectively influence the effective utilization of features and the model's overall performance.

Research [4,5,12–19] predominantly emphasizes the fusion and prediction stages, often overlooking further utilization of the original modality-specific information during feature extraction, particularly among the non-verbal modalities. This gap has resulted in performance bottlenecks in the model. Distinctly from these, our research focuses on optimizing feature extraction to utilize modality-specific information more effectively.

In MSA, the verbal modality refers to text features trained by large-scale Pre-trained Language Models (PLMs) [20,21], whereas non-verbal modalities refer to acoustic and visual features extracted by feature extraction tools [22–25].

* Corresponding author at: School of Computer Engineering and Science, Shanghai University, Shanghai, China.
E-mail addresses: cangzhizheng@shu.edu.cn (C. Zheng), jjie.peng@shu.edu.cn (J. Peng), caizesu@hit.edu.cn (Z. Cai).

https://doi.org/10.1016/j.knosys.2024.112382
Received 1 May 2024; Received in revised form 10 August 2024; Accepted 11 August 2024
Available online 14 August 2024

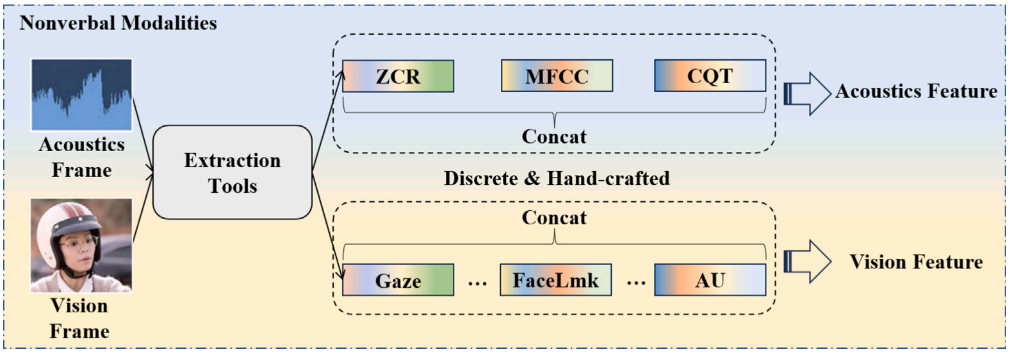


Fig. 1. Schematic diagram of feature extraction from non-verbal modalities through extraction tools used in studies on CH-SIMS [34]. In the non-verbal modalities, discrete and hand-crafted features are internally spliced to obtain the original features. The information about features (e.g., **Gaze** in the figure) is detailed in Section 3.3.

In textual feature extraction, considering the superior performance of large pre-trained models [26], most studies leverage PLMs, such as Bidirectional Encoder Representations from Transformers (BERT) [20], whereas for acoustic and visual modalities, tools such as COVAREP [23] and OpenFace [25] are commonly employed to extract handcrafted features related to sentiments. For instance, in the visual domain, facial-expression-related features are crucial in the expression of various emotions, with facial expressions and body postures providing valuable insights into speakers' emotional states. In audio data, intonation and pitch convey important emotional cues [27,28].

It is noteworthy that non-verbal features are relatively low compared with verbal features learned through PLMs. Raw acoustic and visual features possess distinct low-level characteristics that are inherently discriminative and significant in their own right. Existing methodologies typically integrate low-level features to represent coarse-grained unimodal features, as illustrated in Fig. 1. Various models, such as the Feedforward Neural Networks (FNN) [29], Multi-Layer Perceptron (MLP) [30], recurrent neural networks (RNN), including Long Short-Term Memory (LSTM) [31] and Gated Recurrent Unit (GRU) [32], as well as Convolutional Neural Networks (CNNs) [33], are utilized to learn utterance-level representations from these handcrafted features. However, these fine-grained low-level features possess distinct characteristics that may be independent of each other. Recent connection methods potentially constrain shallow extraction networks from adequately capturing intricate interactions among low-level features to construct coarse-grained representations. Consequently, the unique attributes of non-verbal modalities may not have been thoroughly explored and exploited.

To further clarify our motivation, we analyze how the non-verbal sub-features are distributed, as shown in Fig. 2. Traditional MSA research typically concatenates each sub-modality to form a representative modality feature (Fig. 2(a)). These concatenated features are then fed into the feature extraction network to obtain the initial features. However, as shown in Fig. 2(b), these sub-features differ significantly. The concatenation operation causes interference, which significantly dilutes the emotional information that they contain. This increases the feature extraction network's learning burden and affects how efficiently the original emotional information of the sub-features is utilized. Furthermore, it potentially limits performance during the subsequent fusion stages.

In essence, inadequate utilization of feature information may hinder overall performance enhancements in sentiment prediction.

To utilize the sentiment information in original non-verbal features more effectively, a method is proposed for extracting fine-grained sentiment features from non-verbal modalities. It internally separates the raw features of different modalities and independently models the features from distinct domains. This allows the feature extractor to explore various sub-features within each non-verbal modality and preserve modality-specific information. Moreover, a unified architecture integrates these sub-features, thereby obtaining enhanced coarse-grained modal features for improved interaction.

To validate the effectiveness of the proposed method, we conduct experiments with MSA models in which emotional features are extracted by applying our approach instead of using conventional methods. Extensive experiments demonstrate that applying the features extracted using our method significantly improves the performance of MSA across existing models. This demonstrates that our method effectively harnesses features from non-verbal modalities more efficiently than traditional approaches.

In summary, the main contributions of this study are as follows.

- We present a method for extracting fine-grained emotional features within non-verbal modalities. It promotes the model's in-depth exploration of the interactive relationships between different aspects of sub-features found within the non-verbal modalities.
- Without altering the original fusion strategy, our method internally segregates the original features of different modalities and models the fine-grained features separately. It reduces the learning complexity of the shallow feature extractors and enhances their capability for effectively utilizing the original features.
- Extensive experiments show the performance improvement of classic MSA models after replacing the features obtained by their methods with that extracted by ours, proving that the proposed method can extract emotional features more effectively.

2. Related work

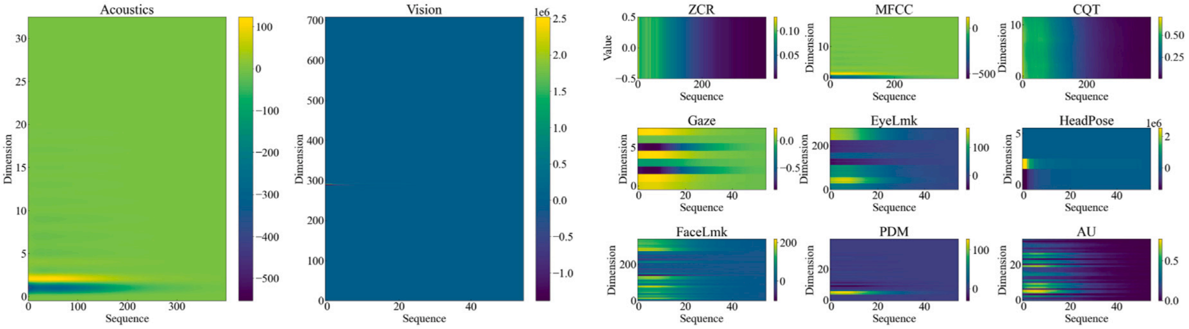
2.1. Multimodal sentiment analysis

Numerous fusion strategies have emerged in multimodal sentiment analyses. We review several studies on classic fusion strategies.

Williams et al. [35] proposed an approach based on input-level feature fusion with sequence learning using bidirectional long short-term memory (BLSTM) deep neural networks (DNNs).

Zadeh et al. [12] proposed a Tensor Fusion Network (TFN) to learn intra-modal dynamics through a modality embedding sub-network and obtain intermodal interactions by computing the outer product. Liu et al. [13] reduced the computational cost of TFN using low-rank tensors in a method named Low-rank Multimodal Fusion (LMF).

The Graph-based Memory Fusion Network (Graph-MFN) [4] is another version of the Memory Fusion Network (MFN) [14] that explores the interactions between unimodal, bimodal, and trimodal features through dynamic graph fusion. Heterogeneity fusion networks based on Graph CNNs (HFNGC) [9] use a shared convolutional aggregation mechanism to overcome the semantic gap between modalities and reduce the noise caused by modality heterogeneity.



(a) Concatenated original sub-features in a general MSA model.

(b) Each fine-grained sub-feature in our method.

Fig. 2. Feature distribution in the acoustic and visual modalities on CH-SIMS [34]. The splicing operation reduces the original sub-feature's emotional information; the sub-features also interfere with each other, as shown in Fig. 2(a). There are obvious differences between the sub-features, and each contains its own emotional information, as shown in Fig. 2(b).

Rahman et al. [36] introduced a well-embedded Multimodal Adaptation Gate network (MAG-BERT) that enables BERT to accept representations from non-verbal modalities aligned with the text. Tsai et al. [15] proposed the Multimodal Transformer (MulT) to apply cross-modal attention to transform one modality into another and vice versa, thus constructing an interaction between different pairs of modalities. An Efficient Multimodal Transformer (EMT) [37] employs utterance-level representations from each modality as a global multimodal context to interact with local unimodal features and mutually promote each other.

Hazarika et al. [5] proposed a framework of Modality-Invariant and Modality-Specific representations for sentiment analysis (MISA) to learn invariant and specific representations for each modality through four different loss functions, and fused different representations to predict sentiments. Yu et al. [16] proposed a Self-supervised Multi-Modal learning strategy (Self-MM) to obtain unimodal labels and learn intermodal consistency and intra-modal specificity through a multi-task framework based on multimodal and unimodal labels. A Fine-grained modal label-based Multi-Stage learning Network (FmlMSN) [8] aids in the final sentiment prediction with six unimodal tasks and three bimodal tasks to improve the robustness of the model. Cross-Modal Hierarchical Fusion Model (CMHFM) [11] combines unimodal, bimodal, and trimodal tasks to enhance multimodal feature representation for the final sentiment prediction.

The Interaction Canonical Correlation Network (ICCN) [38] uses pre-trained BERT in a shared semantic space for vision-to-text and audio-to-text translation. The Bi-Bimodal Fusion Network (BBFN) [6] separates and fuses the representations of each modality to predict sentiments through extra task loss. The Bimodal Information-augmented Multi-Head Attention (BIMHA) [7] system applies the attention mechanism to distill the relative relationships and importance between two pair-wise modalities.

The token disentangling mutual transformer (TMT) [19] utilizes a token disentanglement module to disentangle inter-modality consistency features and intra-modality heterogeneity features. Mai et al. developed a novel multimodal Hybrid Contrastive learning framework (HyCon-BERT) [39] based on BERT, designed with three types of loss for comprehensive intermodal and intra-modal dynamics for both supervised and unsupervised learning.

Related research has combined different learning methods and downstream tasks. The Polar vector and Strength vector Mixer (PS-Mixer) [40] blends the polarity vector and intensity vector models based on MLP to achieve better communication between different modality data for multimodal sentiment analysis. Sun et al. [41] introduced a meta-learning-based method, named Adaptive Multimodal Meta-Learning (AMML) to learn optimized unimodal representations for multimodal fusion. Network-expanding vision-language pre-training (VLP) models for Multimodal Sentiment Analysis (VLP2MSA) achieved improved performance [42]. The Co-space Representation Interaction Network (CRNet) [43] leverages different acoustic and visual representation sub-spaces to interact with language modalities.

However, previous studies mainly focused on the design of modality interaction and label prediction, and virtually overlooked the potential negative impacts of the premature integration of sub-features from non-verbal modalities during feature extraction. The primary focus of our study is the further feature extraction from the non-verbal modalities. For this we apply our method based on the opensource architectures from previous work to validate its effectiveness. For more details, please refer to our experiments.

2.2. Feature extraction

In MSA, most studies have primarily utilized the text, acoustic, and visual modalities for sentiment prediction. Each modality employs a specific method for feature extraction.

For text data, feature extraction methods, such as Word2Vec [44] and GloVe [45], are used for text sentiment analysis. BERT [20] is a PLM based on Transformer [46] that combines auto-regressive and auto-encoding training methods for pre-training bidirectional language models. It has achieved significant results in multiple Natural Language Processing (NLP) tasks. Since the advent of BERT and its variants, text features in MSA have mainly been extracted from them.

Researchers have developed various feature-extraction tools to extract valuable emotional information from non-verbal data. Acoustic features such as intonation, pitch, and energy are often leveraged in speech sentiment analysis. LibROSA [24], OpenSMILE [22], and COVAREP [23] can be employed as acoustic modality extraction tools. Visual features such as facial expressions, body posture, and scenes are widely used for image and video data. Commonly used visual modality extraction tools include FaceNet [47] and OpenFace 2.0 [48]. In addition, considering that the original modality features are unaligned in the time dimension, several studies have used alignment tools such as the Multi-Task Convolutional Neural network (MTCNN) [49] to align the acoustic and visual modalities to the text for subsequent integration research.

However, most research has overlooked the unique characteristics of the original data structure regarding non-verbal modal features, particularly their spatial dimensions. These sub-features, which relate to different aspects of emotions, exhibit distinct features, and may have weak correlations. These weakly correlated features are prematurely concatenated and fed into subsequent extraction methods, e.g., MLP, LSTM, and CNN. Internal disparities among non-verbal features lead to interference among sub-features when the feature extractor learns complex interactions. This blurs specific sub-feature information and diminishes the model's capacity to utilize emotional information effectively during fusion, thereby constraining the model's overall performance.

Our method categorizes non-verbal modalities into distinct domains and models each domain feature as a low-level sub-feature within the modality. This approach aims to preserve specific sentiment information inherent in the original features. By doing so, it reduces the learning complexity of the non-verbal feature extraction network to achieve improved features and enhances the potential for effectively utilizing the sentiment information in modal interactions. Consequently, the proposed method enhances the model's overall performance.

3. Methodology

3.1. Task setting

Our study aims to obtain better initial features from non-verbal modalities to improve MSA and the proposed method is applied to MSA models for verification.

The goal of MSA is to predict the sentiment intensity variable $\hat{y} \in \mathbb{R}$ in a video clip, $X_m = \left\{X_t, X_a, X_v\right\}$ using multimodal signals. Specifically, text (t), acoustic (a), and visual (v) sequences are denoted by X_t, X_a, and $X_v \in \mathbb{R}^{S_m}$, where $m \in \{t, a, v\}$ and S_m represent the sequence length of signal m.

According to previous studies [4,34], MSA prediction is divided into regression and classification tasks. For the regression task, the sentiment intensity variable $\hat{y}$ is continuous, $\hat{y} > 0$, $\hat{y} < 0$, and $\hat{y} = 0$ respectively represent positive, negative, and neutral sentiments. For the classification task, $\hat{y}$ represents the possibility of a sentiment polarity category.

Several studies have focused on either regression or classification tasks, whereas others achieved experimental results on both tasks; we adopt the task settings of previous studies and assess the effectiveness of our approach.

3.2. Original features extraction

For the text sequence X_t, researchers have generally used the PLM to obtain the original feature $E_t \in \mathbb{R}^{S_t \times d_t}$ with dimension d_t. For acoustic and visual sequences, most researchers have extracted the original features using extraction tools. The original non-verbal feature in a video clip is represented by $I_m \in \mathbb{R}^{S_m \times d_m}$, where d_m is the feature dimension and $m \in \{a, v\}$. The calculation is:

$$E_t = \mathrm{PLM}\left(X_t\right), \tag{1}$$

$$I_m = \mathrm{Ext}_m\left(X_m\right), \tag{2}$$

where $\mathrm{Ext}_m(\cdot)$ refers to the original feature extraction of modality m using the extraction tools described in Section 2.2.

3.3. Extraction of fine-grained features

Fig. 3 shows our method applied to the verification model for extracting features from non-verbal modalities. Specifically, it divides the features of non-verbal modalities into fields. The calculation process is as follows.

$$I_m \Rightarrow I_m^s = \left\{I_m^{s_1}, \ldots, I_m^{s_n}\right\}, \tag{3}$$

where I_m^s refers to the set of fine-grained features from non-verbal modality m. $\Rightarrow$ represents the operation for dividing the original features. $I_m^{s_n} \in \mathbb{R}^{S_m \times d_n}$ refers to the nth original non-verbal sub-feature with dimension d_n. n denotes the number of sub-features in modality m.

To further describe our method, we consider the processing of CH-SIMS [34], a widely used dataset in MSA, as an example.

Extract the description in study [34], authors used LibROSA and OpenFace 2.0 to extract 33-dimensional acoustic features $I_a \in \mathbb{R}^{S_a \times 33}$ and 709-dimensional visual features $I_v \in \mathbb{R}^{S_v \times 709}$ respectively from a video clip.

The acoustic features include Zero-Crossing Rate (ZCR) $I_a^{\mathrm{ZCR}} \in \mathbb{R}^{S_a \times 1}$, Mel Frequency Cepstrum Coefficient (MFCC) $I_a^{\mathrm{MFCC}} \in \mathbb{R}^{S_a \times 20}$, and Constant-Q Transform chromatogram (CQT) $I_a^{\mathrm{CQT}} \in \mathbb{R}^{S_a \times 12}$.

The visual modality includes Eye-Gaze coordinates (Gaze) $I_v^{\mathrm{Gaze}} \in \mathbb{R}^{S_v \times 8}$, Eye Landmarks (EyeLmk) $I_v^{\mathrm{EyeLmk}} \in \mathbb{R}^{S_v \times 280}$, Head Pose (HeadPose) $I_v^{\mathrm{HeadPose}} \in \mathbb{R}^{S_v \times 6}$, Facial Landmarks (FaceLmk) $I_v^{\mathrm{FaceLmk}} \in \mathbb{R}^{S_v \times 340}$, Point Distribution Model parameters (PDM) $I_v^{\mathrm{PDM}} \in \mathbb{R}^{S_v \times 40}$, and facial Action Units (AU) $I_v^{\mathrm{AU}} \in \mathbb{R}^{S_v \times 35}$.

We divide the original features according to the above fields and obtain the acoustic and visual modalities' fine-grained features:

$$I_a^s = \left\{I_a^{\mathrm{ZCR}}, I_a^{\mathrm{MFCC}}, I_a^{\mathrm{CQT}}\right\}, \tag{4}$$

$$I_v^s = \left\{I_v^{\mathrm{Gaze}}, I_v^{\mathrm{EyeLmk}}, \ldots, I_v^{\mathrm{AU}}\right\}. \tag{5}$$

As shown in Fig. 2, the general MSA model considers the concatenated original sub-features as input, whereas our approach focuses on the differences between the sub-features and models each sub-feature separately.

In Fig. 3(b), after obtaining the set of these original features, we do not simply internally splice them as representative features of the modality as in the reference work but regard them as sub-features for separate feature extraction. The calculation process is:

$$F_m^{s_n} = \mathrm{SubNet}_m^{s_n}\left(I_m^{s_n}\right), \tag{6}$$

where $F_m^{s_n} \in \mathbb{R}^{S_m \times h_m}$ refers to the nth initial sub-feature of modality m. $\mathrm{SubNet}_m^{s_n}(\cdot)$ represents the nth extraction sub-network for nth sub-feature, as detailed in Section 4.1.4. h_m is the feature dimension of the hidden layers in the nth sub-network.

The fine-grained non-verbal features $\{F_m^{s_1}, \ldots, F_m^{s_n}\}$ are processed to obtain a unified non-verbal representation $F_m^s \in \mathbb{R}^{S_m \times h_m}$ using a Multi-Head Self-Attention (MHSA) [46] mechanism. We consider these sub-features as different heads $\left[Head_m^{s_1}; \ldots; Head_m^{s_i}\right]$ in MHSA to learn domain-specific information, that is, each head learns specific sub-features separately. The difference between the proposed method and the original method is presented in Fig. 4. The calculation process is:

$$Head_m^{s_i} = \mathrm{Softmax}\left(\frac{Q_m^{s_i} K_m^{s_i T}}{\sqrt{h_m}}\right) V_m^{s_i}, \tag{7}$$

$$\mathrm{MHSA}\left(F_m^{s_1}, \ldots, F_m^{s_i}\right) = \left[Head_m^{s_1}; \ldots; Head_m^{s_i}\right] W_O^s, \tag{8}$$

$$F_m^s = \mathrm{MHSA}\left(F_m^{s_1}, \ldots, F_m^{s_i}\right), \tag{9}$$

where $i \in [1, n]$, and n is the number of sub-features in modality m. $Q_m^{s_i}$, $K_m^{s_i}$, and $V_m^{s_i} \in \mathbb{R}^{S_m \times h_m}$ are transformed by ith fine-grained feature $F_m^{s_i}$ with linear matrices $W_Q^{s_i}$, $W_K^{s_i}$ and $W_V^{s_i} \in \mathbb{R}^{h_m \times h_m}$, respectively. $W_O^s \in \mathbb{R}^{(n \times h_m) \times h_m}$ is the output matrix. $[;]$ means concatenation operation. $\mathrm{MHSA}(\cdot)$ denotes the multi-head attention mechanism.

The initial features of the text modality are abstract features pretrained by PLM, which are closely related and cannot be divided into finer-grained parts. Thus, for the verbal modality E_t, we directly obtain the features of the text modality $F_t \in \mathbb{R}^{S_t \times h_t}$ through the original model's extraction network, where h_t is the dimension of the textual initial features.

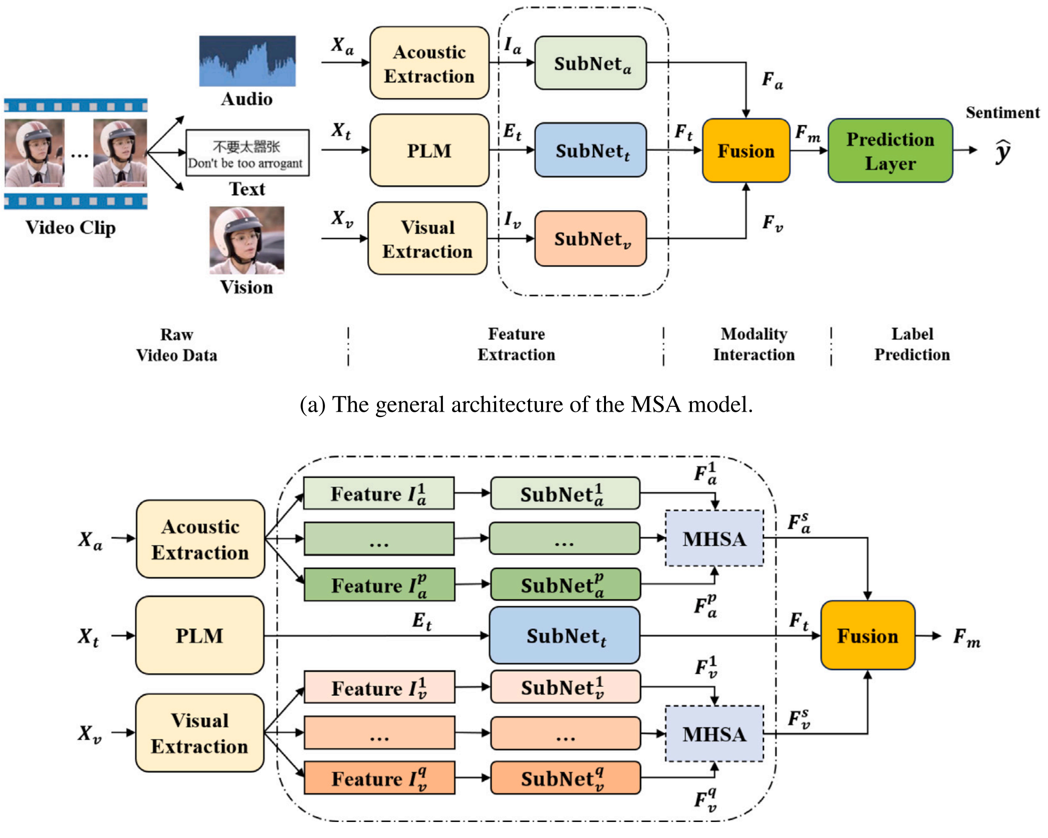
$$F_t = \mathrm{SubNet}_t\left(E_t\right), \tag{10}$$

where $\mathrm{SubNet}_t(\cdot)$ refers to the verbal feature extraction network in the model.

3.4. Fusion and prediction

The initial features F_a^s and F_v^s are extracted using our method, and F_t is delivered to the fusion module of the verification model to predict the sentiment value $\hat{y}$. The calculation process is:

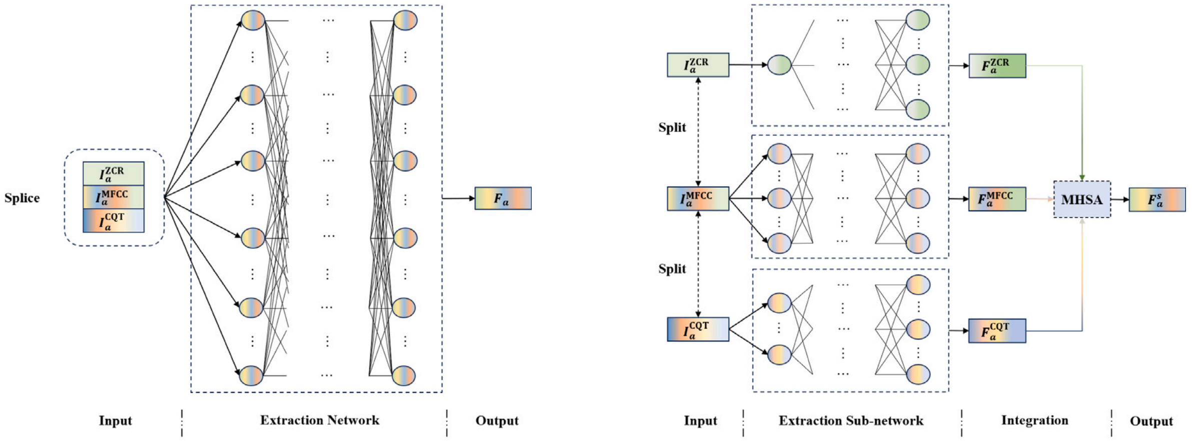
$$\hat{y} = \mathrm{Fusion}\left(F_t, F_a^s, F_v^s\right), \tag{11}$$



(a) The general architecture of the MSA model.

(b) Fine-grained feature extraction (our method).

Fig. 3. Illustration of applying fine-grained feature extraction method within non-verbal modalities on the verification model. (a) represents the unified architecture of the verification model, and our method focuses on extracting features from the non-verbal modalities. (b) symbolizes the modifications of the proposed method. F_a, F_v, and F_m respectively represent acoustic, visual, and fusion features in the original model. p and q represent the number of acoustic and visual sub-features.



(a) The general feature extraction in the MSA model.

(b) The fine-grained feature extraction (our method).

Fig. 4. Acoustic modality: Diagram of the original method and our method on CH-SIMS [34]. (a) represents the initial features of the acoustic modality obtained by the original method. (b) symbolizes the procedure of extracting the initial features of the acoustic modality using our method. In (b), the gradient lines indicate that the features are treated as different heads in MHSA. The dimensions of initial features F_a and F_a^s remain consistent.

Table 1
Data distribution of the datasets.

Dataset	Train	Valid	Test	Total
MOSEI	16,326	1,871	4,659	22,856
CH-SIMS	1,368	456	457	2,281

Table 2
Fine-grained features of non-verbal modalities on the datasets.

Dataset	Modality	Fine-grained features	Dimension
MOSEI	Audio	F0, VUV, NAQ, QOQ, H1H2, PSP, MDQ, Peak Slope, Rd, Creak, MECP, HMPDM, HMPDD	74
	Vision	–	35
CH-SIMS	Audio	ZCR, MFCC, CQT	33
	Vision	Gaze, EyeLmk, HeadPose, FaceLmk, PDM, AU	709

where $\text{Fusion}(\cdot)$ denotes the models' fusion and prediction networks.

We follow previous studies [4,34] and use L1 loss for regression and cross-entropy loss for classification.

4. Experiments

4.1. Experimental settings

We substitute the emotional features in the original models with those extracted using our method and perform several experiments combined with the original fusion mechanisms. These classic models provide reliable reproduction results on the two datasets commonly used to compare performance improvements. In addition, we elaborated on the details of the fine-grained features extracted from the original model and its feature extraction network. Finally, our experiment demonstrate the model parameter settings and metrics.

4.1.1. Validation models

To verify the effectiveness of the proposed method, we conduct experiments on classical MSA models using fine-grained features extracted using the proposed method instead of their original features. These models are described as follows:

- **LF-DNN**: The Late Fusion-Deep Neural Network (LF-DNN) uses a late fusion method to use Multi-Layer Perceptrons (MLP) for sentiment prediction [50].
- **TFN**: The Tensor Fusion Network (TFN) utilizes the Cartesian product of tensors to aggregate unimodal, bimodal, and trimodal interactions [12].
- **LMF**: The Low-rank Multimodal Fusion network (LMF) applies low-rank decomposition based on TFN to reduce the amount of calculation and improve the efficiency of the model [13].
- **MFN**: The Multi-Modal Fusion Network (MFN) stores the modalities' internal information and the interactive information between modalities using gated memory units and adds dynamic fusion maps to reflect emotional information effectively [14].
- **Graph-MFN**: The Graph-MFN extends MFN by using a dynamic fusion graph [4].
- **MulT**: The Multimodal Transformer (MulT) develops the standard Transformer model to focus on cross-modal interactions of the entire utterance, learning representations from unaligned multimodal data [15].
- **MISA**: The framework of Modality-Invariant and -Specific representations for sentiment Analysis (MISA) projects each modality into two subspaces to learn modality-invariant and modality-specific representations and fuses the two representations to predict sentiments [5].

Table 3
Sub-networks and task settings for the verification models. L-C indicates that the model uses LSTM and CNN for modality interaction.

Model	Text	Acoustics	Vision	Regression	Classification
LF-DNN	LSTM	MLP	MLP	✓	✓
TFN	LSTM	MLP	MLP	✓	✓
LMF	LSTM	MLP	MLP	✓	✓
MulT	CNN	CNN	CNN	✓	✓
MFN	LSTM	LSTM	LSTM	✓	✓
Graph-MFN	LSTM	LSTM	LSTM	✓	✓
MISA	BERT	LSTM	LSTM	✓	✓
MLF-DNN	LSTM	MLP	MLP	–	✓
MTFN	LSTM	MLP	MLP	–	✓
MLMF	LSTM	MLP	MLP	–	✓
Self-MM	BERT	LSTM	LSTM	✓	–
TETFN	BERT	L-C	L-C	✓	–
ALMT	BERT	FNN	FNN	✓	–
TMT	BERT	MLP	MLP	✓	–

- **MLF-DNN**: This system utilizes multi-task learning method on LF-DNN to improve model performance [16].
- **MTFN**: This system predicts sentiments using multi-task learning method based on TFN [16].
- **MLMF**: This system uses a multi-task learning method on LMF to improve the model's performance in identifying sentiments [16].
- **Self-MM**: The Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) builds a unimodal label generator based on multi-task learning. It can learn the specific representation of each modality through self-supervised way, enhancing the representation of each modality [16].
- **TETFN**: The Text Enhanced Transformer Fusion Network (TETFN) learns text-oriented pairwise cross-modal mappings for obtaining effective unified multimodal representations [17].
- **ALMT**: The Adaptive Language-guided Multimodal Transformer (ALMT) incorporates an Adaptive Hyper-modality Learning (AHL) module to learn an unrelated or conflict-suppressing representation from visual and audio features under the guidance of language features at different scales [18].
- **TMT**: The Token-disentangling Mutual Transformer (TMT) utilizes a simple and easy-to-implement multimodal emotion token disentanglement module to disentangle the inter-modality consistency features and intra-modality heterogeneity features [19].

4.1.2. Datasets

To verify the effectiveness of the method, experiments are conducted on two public datasets, MOSEI [4] and CH-SIMS [34], both commonly used for MSA. Because all models provide the results on these datasets using their own feature extraction methods, they are fairer and more realistic for validation experiments. Basic information on the datasets is provided in Table 1.

MOSEI: MOSEI [4] is a MSA dataset released in 2018. It contains 3,228 videos divided into 22,856 utterance-level video clips. The dataset is annotated using multi-modal sentiment labels from −3 (strong negative) to three (strongly positive). The multimodal sentiment labels are categorized into five classifications: {−2, −1, 0, 1, 2}. The multimodal sentiment labels are classified into seven categories: {−3, −2, −1, 0, 1, 2, −3}. The partitioning of the dataset is speaker-independent, with 16,326 video clips in the training set, 1,871 in the validation set, and 4,659 in the test set. It contains both aligned and unaligned data.

CH-SIMS: CH-SIMS [34] is a Chinese multimodal sentiment analysis dataset released in 2020 and contains a total of 60 videos, which are split into 2,281 utterance-level video clips. It contains a multimodal sentiment label and three unimodal sentiment labels for each video clip. Each clip is labeled by sentiment intensity ranging between [−1, 1]. Unlike MOSEI, the multimodal sentiment labels are categorized into three intervals, where [−1.0, −0.1] represents negative, (−0.1, 0.1] represents neutral, (0.1, 1.0] represents positive. The multimodal

Table 4
Results of applying the proposed method to the models with the regression task on MOSEI.

Model	Acc-2 (%)↑	F1 (%)↑	Acc-5 (%)↑	Acc-7 (%)↑	MAE↓	Corr↑	Params	FLOPs	Data Setting
LF-DNN*	80.60/82.74	80.85/82.52	51.97	50.83	0.580	0.709	–	–	Unaligned
#LF-DNN	79.10/82.34	79.40/82.11	53.96	52.66	0.559	0.732	5.72×10^5	1.48×10^9	Unaligned
#LF-DNN-fg	80.43/81.86	80.37/81.38	53.93	52.55	0.559	0.730	6.14×10^5	1.49×10^9	Unaligned
Δ_{fg}	+**1.33**/−0.48	+**0.97**/−0.73	−0.03	−0.11	+0.000	−0.002	+7.34%	+0.68%	–
TFN*	78.50/81.89	78.96/81.74	53.10	51.60	0.573	0.714	–	–	Unaligned
#TFN	80.92/82.60	81.06/82.30	53.27	51.98	0.570	0.720	5.04×10^6	3.54×10^9	Unaligned
#TFN-fg	81.65/82.77	81.69/82.42	53.04	51.65	0.572	0.720	5.05×10^6	3.54×10^9	Unaligned
Δ_{fg}	+**0.73**/+**0.17**	+**0.63**/+**0.12**	−0.23	−0.33	−0.002	+0.000	+0.20%	+0.00%	–
LMF*	80.54/83.48	80.94/83.36	52.99	51.59	0.576	0.717	–	–	Unaligned
#LMF	79.22/83.59	79.85/83.58	53.87	52.42	0.564	0.735	5.08×10^5	7.39×10^8	Unaligned
#LMF-fg	81.02/84.14	81.43/84.04	53.57	52.14	0.562	0.736	5.18×10^5	7.39×10^8	Unaligned
Δ_{fg}	+**1.80**/+**0.55**	+**1.58**/+**0.46**	−0.30	−0.28	+**0.002**	+**0.001**	+1.97%	+0.00%	–
MFN*	78.94/82.86	79.55/82.85	52.76	51.34	0.573	0.718	–	–	Aligned
#MFN	81.27/83.93	81.60/83.79	52.54	51.08	0.573	0.723	1.28×10^6	3.92×10^9	Aligned
#MFN-fg	81.35/83.58	81.66/83.43	52.78	51.33	0.573	0.722	1.35×10^6	4.44×10^9	Aligned
Δ_{fg}	+**0.08**/−0.35	+**0.06**/−0.36	+**0.24**	+**0.25**	+0.000	−0.001	+5.47%	13.27%	–
Graph-MFN*	81.28/83.48	81.48/83.23	52.69	51.37	0.575	0.713	–	–	Aligned
#Graph-MFN	82.84/83.99	82.85/83.66	53.07	51.84	0.568	0.724	6.76×10^5	1.98×10^9	Aligned
#Graph-MFN-fg	82.26/84.11	82.44/83.89	52.74	52.92	0.564	0.728	6.94×10^5	2.11×10^9	Aligned
Δ_{fg}	−0.58/+**0.12**	−0.41/+**0.23**	−0.33	+**1.08**	+**0.004**	+**0.004**	+2.66%	+6.57%	–
MulT*	81.15/84.63	81.56/84.52	54.18	52.84	0.559	0.733	–	–	Unaligned
#MulT	78.87/83.84	79.61/83.89	54.66	53.15	0.555	0.739	7.84×10^5	8.70×10^8	Unaligned
#MulT-fg	81.45/84.57	81.82/84.45	54.80	53.34	0.556	0.736	7.99×10^5	9.64×10^8	Unaligned
Δ_{fg}	+**2.58**/+**0.73**	+**2.21**/+**0.56**	+**0.14**	+**0.19**	−0.001	−0.003	+1.91%	+10.80%	–
MISA*	80.67/84.67	81.12/84.66	53.63	52.05	0.558	0.752	–	–	Unaligned
#MISA	81.64/84.59	82.01/84.50	53.51	51.82	0.550	0.760	8.70×10^7	2.73×10^{11}	Unaligned
#MISA-fg	81.64/84.83	82.03/84.74	53.32	51.77	0.552	0.759	8.79×10^7	2.79×10^{11}	Unaligned
Δ_{fg}	+0.00/+**0.24**	+**0.02**/+**0.24**	−0.19	−0.05	−0.002	−0.001	+1.03%	+2.20%	–
Self-MM*	83.76/85.15	83.82/84.90	55.53	53.87	0.530	0.765	–	–	Unaligned
#Self-MM	80.09/84.38	80.69/84.40	55.20	53.38	0.535	0.766	8.58×10^7	1.36×10^{11}	Unaligned
#Self-MM-fg	82.77/84.97	83.04/84.84	55.06	53.30	0.534	0.763	8.58×10^7	1.36×10^{11}	Unaligned
Δ_{fg}	+**2.68**/+**0.59**	+**2.35**/+**0.44**	−0.14	−0.08	+**0.001**	−0.003	+0.00%	+0.00%	–
TETFN*	84.12/86.21	84.35/86.11	55.78	53.90	0.537	0.770	–	–	Aligned
#TETFN	79.47/84.50	80.12/84.51	56.00	54.09	0.539	0.761	8.66×10^7	1.38×10^{11}	Aligned
#TETFN-fg	81.20/85.11	81.72/85.10	55.73	53.81	0.542	0.762	8.67×10^7	1.38×10^{11}	Aligned
Δ_{fg}	+**1.73**/+**0.61**	+**1.60**/+**0.59**	−0.27	−0.28	−0.003	+**0.001**	+0.12%	+0.00%	–
ALMT	84.78/86.79	85.19/86.86	55.96	54.28	0.526	0.779	–	–	Aligned
#ALMT	81.54/85.44	81.98/85.36	54.15	52.54	0.547	0.763	8.87×10^7	2.77×10^{11}	Aligned
#ALMT-fg	82.87/84.89	83.07/84.71	54.20	52.54	0.547	0.765	8.90×10^7	2.80×10^{11}	Aligned
Δ_{fg}	+**1.33**/−0.55	+**1.09**/−0.65	+**0.05**	+0.00	+0.000	+**0.002**	+0.39%	+1.08%	–
TMT	–/86.50	–/86.50	–	53.70	0.542	0.775	–	–	Aligned
#TMT	–/83.68	–/83.69	47.76	46.77	0.620	0.711	4.47×10^6	4.55×10^9	Aligned
#TMT-fg	–/83.21	–/83.14	48.77	47.67	0.614	0.705	5.52×10^6	7.28×10^9	Aligned
Δ_{fg}	–/−0.47	–/−0.55	+**1.01**	+**0.90**	+**0.006**	−0.006	+23.49%	+60.00%	–

sentiment labels are also categorized into five intervals, where [−1.0, −0.7] represents negative, (−0.7, −0.1] represents weakly negative, (−0.1, 0.1] represents neutral, (0.1, 0.7] represents weakly positive, and (0.7, 1.0] represents positive. It contains only the unaligned data.

4.1.3. Fine-grained features

Table 2 shows the fine-grained features information in the datasets. The fine-grained features of the audio modality in MOSEI are derived from COVAREP, and the visual modality is obtained using the FACET tool. As this work [4] does not disclose the details of the selected features, we are unable to conduct fine-grained extraction from the visual modality. The fine-grained features of the audio modality on CH-SIMS are from LibROSA, and the fine-grained features of the visual modality are from OpenFace 2.0. The details of the fine-grained features can be obtained from the extraction tools' official websites.

4.1.4. Sub-networks and task settings of models

Table 3 illustrates the details of the models' sub-networks. From Table 3, we find that simple networks initially extract features from the non-verbal modalities. Notably, our approach focuses on partitioning fine-grained features and modeling individual subnetworks based on the original models, which minimizes the introduction of additional parameters and uncertainties.

To verify the effectiveness of the proposed method, we adopt evaluation metrics widely utilized in MSA [4,34]. The task settings for the verification model are provided in Table 3.

4.1.5. Experimental parameters and metrics

Considering that the experiment is fair and the results are reproducible, we experiment with opensource project, which contains model implementations and has attracted the attention of many researchers. It is noteworthy that our method only changes the feature extraction process of the original data and does not change the fusion parts or the models' hyperparameters.

For regression, we report on the 2-class Accuracy (Acc-2) and weighted F1-Score (F1), 5-class Accuracy (Acc-5), and 7-class Accuracy (Acc-7) for MOSEI. There are two methods for calculating Acc-2 and F1: negative or non-negative (neutral is included) and negative or positive

Table 5
Results of applying the proposed method to the models with the regression task on CH-SIMS.

Model	**Acc-2 (%)↑**	**F1 (%)↑**	**Acc-3 (%)↑**	**Acc-5 (%)↑**	**MAE↓**	**Corr↑**	**Params**	**FLOPs**
LF-DNN*	77.02	77.27	64.33	39.74	0.446	0.555	–	–
#LF-DNN	77.59	76.92	63.94	38.38	0.451	0.553	6.36×10^5	2.32×10^9
#LF-DNN-fg	78.03	77.75	66.13	41.75	0.443	0.568	9.60×10^5	2.40×10^9
Δ_{fg}	**+0.44**	**+0.83**	**+2.19**	**+3.37**	**+0.008**	**+0.015**	+50.94%	+3.45%
TFN*	78.38	78.62	65.12	39.30	0.432	0.591	–	–
#TFN	76.15	76.49	63.98	36.50	0.440	0.574	3.56×10^7	1.70×10^9
#TFN-fg	79.87	79.69	66.43	40.17	0.414	0.636	3.60×10^7	1.72×10^9
Δ_{fg}	**+3.72**	**+3.20**	**+2.45**	**+3.67**	**+0.026**	**+0.062**	+1.12%	+1.18%
LMF*	77.77	77.88	64.68	40.53	0.441	0.576	–	–
#LMF	77.46	77.47	65.56	37.24	0.446	0.572	2.72×10^5	5.38×10^8
#LMF-fg	78.29	77.66	64.99	40.35	0.443	0.572	3.53×10^5	5.47×10^8
Δ_{fg}	**+0.83**	**+0.19**	−0.57	**+3.11**	**+0.003**	+0.000	+29.78%	+1.67%
MFN*	77.90	77.88	65.73	39.47	0.435	0.582	–	–
#MFN	76.85	76.69	65.21	38.60	0.441	0.566	6.02×10^5	7.35×10^8
#MFN-fg	79.69	79.79	66.87	40.26	0.411	0.634	7.27×10^5	1.10×10^9
Δ_{fg}	**+2.84**	**+3.10**	**+1.66**	**+1.66**	**+0.030**	**+0.068**	+20.76%	+49.66%
Graph-MFN*	78.77	78.21	65.65	39.82	0.445	0.578	–	–
#Graph-MFN	75.84	76.04	65.86	41.01	0.435	0.586	2.72×10^6	6.43×10^9
#Graph-MFN-fg	77.55	77.50	65.74	41.92	0.433	0.582	4.65×10^6	1.77×10^{10}
Δ_{fg}	**+1.71**	**+1.46**	−0.10	**+0.91**	**+0.002**	−0.004	+70.96%	+175.27%
MulT*	78.56	79.66	64.77	37.94	0.453	0.564	–	–
#MulT	78.16	77.85	65.56	38.82	0.439	0.587	1.51×10^6	3.02×10^9
#MulT-fg	81.01	81.16	65.45	39.17	0.406	0.653	1.54×10^6	3.27×10^9
Δ_{fg}	**+2.85**	**+3.31**	−0.11	**+0.35**	**+0.033**	**+0.066**	+1.99%	+8.28%
Self–MM*	80.04	80.44	65.47	41.53	0.425	0.595	–	-
#Self-MM	77.15	77.35	64.25	41.53	0.431	0.576	8.60×10^7	1.07×10^{11}
#Self-MM-fg	78.43	78.48	64.95	42.32	0.411	0.600	8.61×10^7	1.07×10^{11}
Δ_{fg}	**+1.28**	**+1.13**	**+0.70**	**+0.79**	**+0.020**	**+0.024**	+0.12%	+0.00%
TETFN*	81.18	80.24	63.24	41.79	0.420	0.577	–	–
#TETFN	77.59	77.83	64.51	42.23	0.428	0.579	8.69×10^7	2.15×10^{11}
#TETFN-fg	79.21	79.20	64.55	43.28	0.419	0.590	8.71×10^7	2.16×10^{11}
Δ_{fg}	**+1.62**	**+1.37**	**+0.04**	**+1.05**	**+0.009**	**+0.011**	+0.23%	+0.47%
ALMT	81.19	81.57	68.93	45.73	0.404	0.619	–	–
#ALMT	77.24	77.41	65.43	42.23	0.425	0.562	8.88×10^7	2.24×10^{11}
#ALMT-fg	79.21	79.39	65.86	42.89	0.417	0.578	8.91×10^7	2.31×10^{11}
Δ_{fg}	**+1.97**	**1.98**	**+0.43**	**+0.66**	**+0.008**	**+0.016**	+0.34%	+3.13%
TMT	80.53	81.11	68.71	48.14	–	–	–	–
#TMT	78.34	78.52	65.65	39.39	0.451	0.592	4.63×10^6	1.36×10^{10}
#TMT-fg	80.74	80.45	68.05	38.73	0.436	0.628	5.62×10^6	2.00×10^{10}
Δ_{fg}	**+2.40**	**+1.93**	**2.40**	−0.66	**0.015**	**0.036**	+21.38%	+47.06%

(neutral is excluded). We report on the Acc-2 and F1, 3-class Accuracy (Acc-3), and 5-class Accuracy (Acc-5) for CH-SIMS.

According to previous studies [4,34], in the indicators (Accuracy and F1-score) of the regression task, it is in the correctly predicted samples that the predicted labels match the true labels (for MOSEI), or the predicted and true labels fall within the same range (for CH-SIMS). The calculated accuracy is consistent with that of the original model.

We also report the Mean Absolute Error (MAE) between the predicted and true labels and the Pearson Correlation coefficient (Corr) related to the degree of linear correlation between the prediction and ground truth.

We report on the 2-class Accuracy (Acc-2), 2-class weighted F1-Score (F1), 3-class Accuracy (Acc-3), and 3-class weighted F1-Score (F1-3) on MOSEI and CH-SIMS for classification.

Excepting for MAE, the higher the values of the other metrics, the better the performance.

To consider the computational efficiency, we also report on the trainable parameters (Params) and floating points of operations (FLOPs).

Experiments on the regression and classification tasks of the same model share the same set of hyperparameters. The difference in the models' computational efficiency in executing the tasks is minimal compared with the parameter calculation of the entire model. Thus, for brevity, we only show our method's impact on the computational efficiency in executing the regression task.

4.2. Results analysis

The results of applying our method to different MOSEI (Tables 4 and 6) and CH-SIMS (Tables 5 and 7) models typically indicate a noticeable performance enhancement.

In these tables, * indicates that the result is from opensource project, # denotes the result of the model reproduced under the same conditions, and -fg represents the result using the proposed method based on reproduction. Δ_{fg} represents the degree of impact on the original performance of the model after applying this method. ↑ indicates that higher is better and ↓ indicates that lower is better. In all the results, the improvement value of the performance is in bold. For Acc-2 and F1, the number on the left of "/" means "negative or non-negative", and the number on the right of "/" is "negative or positive".

The results for these datasets are as follows:

The results in Tables 4 and 6, confirm significant performance improvements in the models using the features extracted by the proposed method, which indicates that our method has definite advantages over the original methods. For the regression task, the models with the most remarkable improvements in classification accuracy are Self-MM,

Table 6
Results of applying the proposed method to the models with the classification task on MOSEI.

Model	Acc-2 (%)↑	F1 (%)↑	Acc-3 (%)↑	F1-3 (%)↑
LF-DNN*	71.18/27.19	70.14/28.67	67.35	64.65
#LF-DNN	70.99/27.95	70.03/28.82	66.97	65.16
#LF-DNN-fg	72.21/28.86	71.54/29.21	67.58	65.61
Δ_{fg}	**+1.22**/**+0.91**	**+1.51**/**+0.39**	**+0.61**	**+0.45**
TFN*	71.54/28.66	70.91/28.75	66.63	63.93
#TFN	70.94/27.63	69.99/28.68	67.34	65.41
#TFN-fg	72.01/28.47	71.34/28.92	66.52	64.71
Δ_{fg}	**+1.07**/**+0.84**	**+1.35**/**+0.24**	−0.82	−0.70
LMF*	71.90/28.86	71.26/28.92	66.59	64.86
#LMF	71.18/28.05	70.30/28.91	67.29	65.15
#LMF-fg	71.53/27.88	70.65/28.90	67.27	65.11
Δ_{fg}	**+0.35**/−0.17	**+0.35**/−0.01	−0.02	−0.04
MFN*	71.49/28.61	70.80/28.70	66.59	64.31
#MFN	71.00/27.44	70.01/28.73	67.27	64.83
#MFN-fg	71.47/28.48	70.71/28.70	66.79	64.89
Δ_{fg}	**+0.47**/**+1.04**	**+0.70**/−0.03	−0.48	**+0.06**
Graph-MFN*	71.25/28.47	70.51/28.77	66.39	64.00
#Graph-MFN	70.89/27.83	70.00/28.70	66.76	64.63
#Graph-MFN-fg	71.29/27.99	70.43/28.86	66.96	64.65
Δ_{fg}	**+0.40**/**+0.16**	**+0.43**/**+0.16**	**+0.20**	**+0.02**
MulT*	71.23/27.38	70.23/28.67	67.04	65.01
#MulT	71.49/27.43	70.49/28.82	67.43	66.36
#MulT-fg	71.86/28.17	70.95/29.03	68.09	66.69
Δ_{fg}	**+0.37**/**+0.74**	**+0.46**/**+0.21**	**+0.66**	**+0.33**
MISA*	71.44/28.40	70.56/29.03	67.63	65.39
#MISA	70.42/27.02	69.16/28.78	67.25	62.57
#MISA-fg	71.05/28.05	69.96/29.01	67.69	65.06
Δ_{fg}	**+0.63**/**+1.03**	**+0.80**/**+0.23**	**+0.44**	**+2.49**

MulT, and LMF, with improvements of 2.68%, 2.58%, and 1.80%, respectively. In the classification task, the models' performance also improves significantly for several indicators, which indicates that our method improves the models' performance compared with the original methods. However, because MOSEI only uses the acoustic modality's fine-grained features, improvements in the fine-grained metrics of all models remain limited.

The feature-extraction methods in MFN, Graph-MFN, ALMT, and TMT are designed for the aligned data on MOSEI. Word-level alignment of the original complete information on the acoustic and visual modalities with text destroys the time-series characteristics of the original data. It loses adequate information in the time dimension and reduces the models' ability to capture the interaction of the fundamental modality's internal features. However, unaligned data is used in CH-SIMS, which preserves the temporal interactions within the original modalities, whereas our method exploits the emotional information of the temporal dimension more effectively than do the original methods.

Tables 5 and 7 show that on CH-SIMS, with the features extracted using our method, all models' performance significantly improves. Compared with MOSEI, the degree of improvement after using our method is more apparent. Because the non-verbal features extracted from the CH-SIMS dataset are richer, our method promotes the models' efficiency in utilizing the emotional information in the original features. LF-DNN uses a simple fusion mechanism and exhibits limited improvement in 2-class accuracy, while the improvement in the fine-grained metrics is obvious. It improves the 3-class and 5-class accuracies by 2.19% and 3.37% respectively. For the regression task, our method achieves significant improvements in 2-class accuracy and F1. The 2-class accuracy of TFN, MFN, Graph-MFN, MulT, and Self-MM improves by 3.72%, 2.84%, 1.71%, 2.85%, and 1.28%, respectively. For F1, they increase by 3.20%, 3.10%, 1.46%, 3.31% and 1.13% respectively. These results further indicate that our method extracts compelling features from the original data more effectively than the original methods and promotes the effective integration of emotional information using the MHSA mechanism.

Table 7
Results of applying the proposed method to the models with the classification task on CH-SIMS.

Model	Acc-2 (%)↑	F1 (%)↑	Acc-3 (%)↑	F1-3 (%)↑
LF-DNN*	79.87/56.70	79.97/55.27	70.20	65.29
#LF-DNN	77.94/57.89	77.66/54.31	67.53	61.45
#LF-DNN-fg	78.08/57.68	77.81/54.25	67.40	61.40
Δ_{fg}	**+0.14**/−0.21	**+0.15**/−0.06	−0.13	−0.05
TFN*	75.32/53.56	75.66/52.79	65.95	62.04
#TFN	76.02/53.61	76.49/53.45	67.40	63.20
#TFN-fg	79.39/56.81	79.40/54.69	68.40	65.21
Δ_{fg}	**+3.37**/**+3.20**	**+2.91**/**+1.24**	**+1.00**	**+2.01**
LMF*	77.99/57.06	77.59/53.83	66.87	62.46
#LMF	76.15/53.87	76.53/53.47	67.66	63.59
#LMF-fg	78.03/57.42	77.82/54.31	67.57	62.34
Δ_{fg}	**+1.88**/**+3.55**	**+1.29**/**+0.84**	−0.09	−1.25
MFN*	78.25/56.96	78.08/54.14	67.57	-
#MFN	77.55/56.08	77.63/54.07	67.57	62.90
#MFN-fg	80.74/55.16	81.12/55.46	70.55	66.34
Δ_{fg}	**+3.19**/−0.92	**+3.49**/**+1.39**	**+2.98**	**+3.44**
Graph-MFN*	79.21/57.99	78.92/54.66	68.44	63.44
#Graph-MFN	77.68/55.77	77.79/54.22	68.67	64.75
#Graph-MFN-fg	79.83/57.99	79.66/55.01	69.41	64.96
Δ_{fg}	**+2.15**/**+2.22**	**+1.87**/**+0.79**	**+0.74**	**+0.21**
MulT*	78.07/56.34	78.07/54.26	68.27	64.23
#MulT	76.89/56.55	76.54/53.65	66.83	61.75
#MulT-fg	82.97/57.27	83.10/56.27	71.73	68.55
Δ_{fg}	**+6.08**/**+0.72**	**+6.56**/**+2.62**	**+4.90**	**+6.80**
MLF-DNN*	80.79/58.19	80.59/55.55	70.37	65.94
#MLF-DNN	76.72/54.38	77.24/54.19	68.93	64.45
#MLF-DNN-fg	79.35/56.03	79.76/55.65	71.56	67.06
Δ_{fg}	**+2.63**/**+1.65**	**+2.52**/**+1.46**	**+2.63**	**+2.61**
MTFN*	81.23/56.91	81.24/55.29	70.28	66.44
#MTFN	76.45/55.57	76.64/53.75	67.09	61.64
#MTFN-fg	78.91/55.36	79.28/54.93	68.93	64.13
Δ_{fg}	**+2.46**/−0.21	**+2.64**/**+1.18**	**+1.84**	**+2.49**
MLMF*	81.45/56.60	81.62/55.66	71.60	70.45
#MLMF	79.43/55.98	79.60/54.76	69.54	65.34
#MLMF-fg	80.66/56.34	80.89/55.63	70.46	67.05
Δ_{fg}	**+1.23**/**+0.36**	**+1.29**/**+0.87**	**+0.92**	**+1.71**

Regarding the fine-grained evaluation metrics, on the different models, there is a significant gap in the improvement effected by our method. The improvement in 5-class accuracy of LF-DNN, TFN, LMF, and MFN is pronounced, with the highest and lowest performance improvements being 3.67% and 1.66%, respectively. Compared with the aforementioned models, the others exhibit weaker improvement effects under the same metrics. Other models employ more intricate fusion mechanisms, thereby capturing richer emotional information when ample data are available, as evidenced by enhancements in the fine-grained metrics. However, compared with MOSEI, the amount of CH-SIMS could be higher. The size of the training set limits the extraction methods for these models. This could potentially cause overfitting, including of the sub-feature extraction networks, which may weaken the performance improvement in the fine-grained metrics introduced by our method.

In addition, the other models, excluding LF-DNN, show significant performance improvements in the classification task. MulT records the highest performance improvement, which increases by 6.08% and 6.56% in 2-class accuracy and F1, respectively. It also improves by 4.90% and 6.80% in 3-class accuracy and F1-3, respectively. MulT uses unaligned acoustic and visual data to fully retain the emotional information in the original features. This indicates that our method improves the efficiency of utilizing the model's extraction features, confirming that our method is more effective than the original methods.

Because our method introduces MHSA, the number of parameters inevitably increases. The parameters and computational overheads depend entirely on the feature dimensions of the original parameter

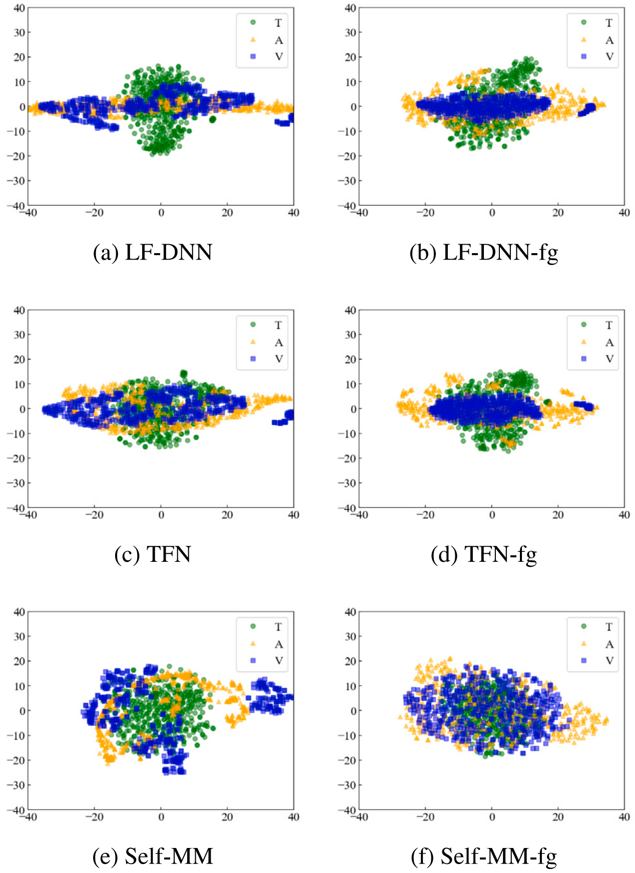
Table 8
Results of ablation study on CH-SIMS. -a and -v represent the reproduction results after using only the acoustic and visual modality fine-grained feature extraction methods, respectively. In all results, bold represents the best result in the results column, and the underlined is the sub-optimal result.

Model	Acc-2 (%)↑	F1 (%)↑	MAE↓	Corr↑
#LF-DNN	77.59	76.92	0.451	0.553
#LF-DNN-fg	78.03	77.75	0.443	0.568
#LF-DNN-a	77.64	77.21	0.448	0.562
#LF-DNN-v	**78.16**	**77.78**	**0.439**	**0.584**
#TFN	76.15	76.49	0.440	0.574
#TFN-fg	**79.87**	**79.69**	0.414	**0.636**
#TFN-a	77.55	77.76	0.434	0.581
#TFN-v	78.99	79.21	**0.413**	0.626
#LMF	77.46	77.47	0.446	0.572
#LMF-fg	**78.29**	**77.66**	0.443	**0.572**
#LMF-a	76.80	76.84	**0.442**	0.570
#LMF-v	77.24	76.91	**0.442**	0.564
#MFN	76.85	76.69	0.441	0.566
#MFN-fg	**79.69**	**79.79**	0.411	0.634
#MFN-a	77.16	76.93	0.439	0.569
#MFN-v	78.25	78.49	**0.406**	**0.642**
#Graph-MFN	75.84	76.04	0.435	0.586
#Graph-MFN-fg	77.55	**77.50**	**0.433**	**0.582**
#Graph-MFN-a	76.98	76.98	0.439	0.575
#Graph-MFN-v	**77.99**	77.46	0.444	0.565
#MulT	78.16	77.85	0.439	0.587
#MulT-fg	81.01	81.16	0.406	0.653
#MulT-a	77.81	77.41	0.450	0.583
#MulT-v	**81.40**	**81.67**	**0.385**	**0.680**
#Self-MM	77.15	77.35	0.431	0.576
#Self-MM-fg	78.43	78.48	**0.411**	**0.600**
#Self-MM-a	77.42	77.61	0.424	0.581
#Self-MM-v	**78.60**	**78.52**	0.430	0.581
#TETFN	77.59	77.83	0.428	0.579
#TETFN-fg	**79.21**	**79.20**	**0.419**	**0.590**
#TETFN-a	78.55	78.26	0.425	0.581
#TETFN-v	78.33	78.24	0.421	0.584
#ALMT	77.24	77.41	0.425	0.562
#ALMT-fg	**79.21**	**79.39**	**0.417**	0.578
#ALMT-a	77.46	77.44	0.433	0.567
#ALMT-v	78.34	78.36	0.421	**0.590**
#TMT	78.34	78.52	0.451	0.592
#TMT-fg	**80.74**	80.45	0.436	0.628
#TMT-a	78.56	78.37	0.447	0.594
#TMT-v	**80.74**	**80.58**	**0.426**	**0.651**

configuration. MFN and Graph-MFN have similar structures. They use a serial memory mechanism to store and retrieve historical and process information at the current time step. This increases the proposed method's computational overhead in the time dimension. Therefore, applying our method to models with serial feature extraction does not improve computational efficiency. In addition to MFN and Graph-MFN, the performance improvement occasioned by applying the features extracted using our method has apparent advantages regarding parameter quantity and computational efficiency.

4.3. Ablation study

Table 8 shows the proposed method's results for different modalities. From Table 8, it is evident that most models exhibit apparent patterns. On the acoustic or visual modalities alone, the model performance increases to varying degrees using the features extracted with our method. The performance improvement achieved using this method in the visual modality is much greater than that in the acoustic modality. This is because the acoustic modality has more initial feature dimensions and emotion-related domain categories than the acoustic modality. The visual modality contains richer emotional information. Therefore, the proposed method further improves the efficiency of utilizing the feature extraction network.



(a) LF-DNN

(b) LF-DNN-fg

(c) TFN

(d) TFN-fg

(e) Self-MM

(f) Self-MM-fg

Fig. 5. Visualization in unimodal representation. The first column displays the representations learned from the original model. The second column displays the representations learned from fine-grained feature extraction methods. Compare two subgraphs on the same line.

When applying the proposed method to both the acoustic and visual modalities, we find that the models' performance is lower than when using the method for the visual modality only. This does not seem to follow the general rules. Below, we analyze the reasons.

Information Imbalance Between Modalities: Visual features have much higher dimensions than acoustic features (709-d vs. 33-d), which means that there is a large information difference between the two modalities. The model using our method may tend to rely more on visual information than on acoustic information, causing the acoustic modality to exhibit degraded performance.

Data Quality and Quantity: There are noisy data in CH-SIMS that require more data coverage, and the model's acoustic feature extraction network may cause overfitting, thus exacerbating the adverse effects mentioned above.

Hyperparameter Settings: We improve the original feature extraction methods of the models but, for a fair experiment, do not fine-tune the hyperparameters. This could have caused performance deviations in the model when the proposed method is used.

4.4. Visualization and analysis

Another motivation for proposing our method is that the differences in unimodal representations can be more pronounced when modeling the independent sub-features of non-verbal modalities. These discrepancies may reflect modality-specific information that is crucial for modality interactions and sentiment prediction.

Table 9
The table shows several examples of sentiment labels predicted by MulT [15] retaining the original method **L(Ori)**, applying our method **L(Fg)**, and Ground Truth (**GT**). Δ denotes the absolute value of the difference between the predicted value and the ground truth, with a smaller absolute value indicating better performance. Bold indicates the best result for each example.

Modality	Example	L(Ori)	L(Fg)	GT
T	不太理想，话题太高端。 (Not really. This subject is serious.)	0.02 Δ0.62	−0.55 **Δ0.05**	−0.60
A	Slow			
V	Smile			
T	03年我一年见了三百多个剧组。 (In 2003, I met with more than 300 film crews in one year.)	0.82 Δ0.82	−0.03 **Δ0.03**	0.00
A	Fast			
V	Normal			
T	是吗？挺好，为民除害。 (Really? That is good, it is a relief for the people.)	0.84 Δ0.44	0.31 **Δ0.09**	0.40
A	Low voice			
V	Serious			

We utilize t-SNE [51] to visualize the intra-modal representations learned in the original models (LF-DNN, TFN, and self-MM) and new models (LF-DNN-fg, TFN-fg, and Self-MM-fg), as depicted in Fig. 5. The models that applied our method learn more distinct unimodal representations of the non-verbal modalities than those that retain the original methods. Therefore, our method enhances the models' ability to acquire more diverse information and improves inter-modality complementarity.

4.5. Case study

Further to elucidate the enhancements brought about by our approach, we conduct a case analysis of CH-SIMS [34] using MulT [15], which produce the best overall performance, as depicted in Table 9.

Table 9 shows that in the first example, our method aids the model in acquiring additional non-verbal modality-specific information to determine sentiment polarity. In the second and third examples, our method assists the model in gaining more precise sentiment intensity from the non-verbal modalities. These observations confirm that our method more effectively utilizes information from both the acoustic and visual modalities than the original methods and achieves more accurate sentiment predictions.

5. Conclusion

To utilize the information in non-verbal features effectively, we propose a method for extracting fine-grained emotional features from videos. The performance of several MSA models is significantly improved by using the fine-grained features extracted by our method, indicating that it effectively mines emotional feature information. In the future, we will focus on designing a method that more effectively extracts and utilizes the internal features of modalities to achieve a higher MSA performance.

CRediT authorship contribution statement

Cangzhi Zheng: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Junjie Peng:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Zesu Cai:** Writing – review & editing, Conceptualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the 2022 Shanghai Service Industry Development Fund, China (No. 06162021592), the Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University, China.

References

[1] J. Peng, Emotion analysis for machine intelligence, Chin. J. Nat. 46 (02) (2024) 150–156.
[2] G. Vinodhini, R. Chandrasekaran, Sentiment analysis and opinion mining: A survey, Int. J. 2 (6) (2012) 282–292.
[3] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.
[4] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.
[5] D. Hazarika, R. Zimmermann, S. Poria, MISA: Modality-invariant and -specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.
[6] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, S. Poria, Bi-bimodal modality fusion for correlation controlled multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 6–15.
[7] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, Y. Huang, Video sentiment analysis with bimodal information-augmented multi-head attention, Knowl.-Based Syst. 235 (2022) 107676.
[8] J. Peng, T. Wu, W. Zhang, F. Cheng, S. Tan, F. Yi, Y. Huang, A fine-grained modal label-based multi-stage network for multimodal sentiment analysis, Expert Syst. Appl. 221 (2023) 119721.
[9] T. Zhao, J. Peng, Y. Huang, L. Wang, H. Zhang, Z. Cai, A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis, Appl. Intell. 53 (24) (2023) 30455–30468.
[10] H. Lin, P. Zhang, J. Ling, Z. Yang, L.K. Lee, W. Liu, PS-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis, Inf. Process. Manage. 60 (2) (2023) 103229.
[11] L. Wang, J. Peng, C. Zheng, T. Zhao, et al., A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning, Inf. Process. Manage. 61 (3) (2024) 103675.
[12] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.
[13] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 2247–2256.
[14] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 5634–5641.

[15] Y.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019, pp. 6558–6569.

[16] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, 2021, pp. 10790–10797.

[17] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognit. 136 (2023) 109259.

[18] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, T. Yu, Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 756–767.

[19] G. Yin, Y. Liu, T. Liu, H. Zhang, F. Fang, C. Tang, L. Jiang, Token-disentangling mutual transformer for multimodal emotion recognition, Eng. Appl. Artif. Intell. 133 (2024) 108348.

[20] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[21] Z. Yang, Z. Dai, Y. Yang, J.G. Carbonell, R. Salakhutdinov, Q.V. Le, XLNet: Generalized autoregressive pretraining for language understanding, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, 2019, pp. 5754–5764.

[22] F. Eyben, M. Wöllmer, B.W. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th International Conference on Multimedia 2010, 2010, pp. 1459–1462.

[23] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP - a collaborative voice analysis repository for speech technologies, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, pp. 960–964.

[24] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, O. Nieto, Librosa: Audio and music signal analysis in python, in: Proceedings of the 14th Python in Science Conference 2015, 2015, pp. 18–24.

[25] T. Baltrusaitis, P. Robinson, L. Morency, OpenFace: An open source facial behavior analysis toolkit, in: 2016 IEEE Winter Conference on Applications of Computer Vision, 2016, pp. 1–10.

[26] Q. Xu, J. Peng, C. Zheng, S. Tan, F. Yi, F. Cheng, Short text classification of Chinese with label information assisting, ACM Trans. Asian Low-Resource Lang. Inf. Process. 22 (4) (2023) 1–19.

[27] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359.

[28] R. Li, Z. Wu, J. Jia, J. Li, W. Chen, H. Meng, Inferring user emotive state changes in realistic human-computer conversational dialogs, in: 2018 ACM Multimedia Conference on Multimedia Conference, 2018, pp. 136–144.

[29] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133.

[30] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back propagating errors, Nature 323 (6088) (1986) 533–536.

[31] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[32] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1724–1734.

[33] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[34] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotations of modality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3718–3727.

[35] J. Williams, S. Kleinegesse, R. Comanescu, O. Radu, Recognizing emotions in video using multimodal DNN feature fusion, in: Proceedings of Grand Challenge and Workshop on Human Multimodal Language, 2018, pp. 11–19.

[36] W. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, 2020, pp. 2359–2369.

[37] L. Sun, Z. Lian, B. Liu, J. Tao, Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis, IEEE Trans. Affect. Comput. 15 (1) (2023) 309–325.

[38] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8992–8999.

[39] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. 14 (3) (2022) 2276–2289.

[40] H. Lin, P. Zhang, J. Ling, Z. Yang, L.K. Lee, W. Liu, PS-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis, Inf. Process. Manage. 60 (2) (2023) 103229.

[41] Y. Sun, S. Mai, H. Hu, Learning to learn better unimodal representations via adaptive multimodal meta-learning, IEEE Trans. Affect. Comput. 14 (2023) 2209–2223.

[42] G. Yi, C. Fan, K. Zhu, Z. Lv, S. Liang, Z. Wen, G. Pei, T. Li, J. Tao, VLP2MSA: Expanding vision-language pre-training to multimodal sentiment analysis, Knowl.-Based Syst. 283 (2024) 111136.

[43] H. Shi, Y. Pu, Z. Zhao, J. Huang, D. Zhou, D. Xu, J. Cao, Co-space representation interaction network for multimodal sentiment analysis, Knowl.-Based Syst. 283 (2024) 111149.

[44] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.

[45] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 2017, pp. 5998–6008.

[47] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

[48] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L. Morency, OpenFace 2.0: Facial behavior analysis toolkit, in: 13th IEEE International Conference on Automatic Face & Gesture Recognition, 2018, pp. 59–66.

[49] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.

[50] E. Cambria, D. Hazarika, S. Poria, A. Hussain, R.B.V. Subramanyam, Benchmarking multimodal sentiment analysis, in: Computational Linguistics and Intelligent Text Processing - 18th International Conference, 10762, 2017, pp. 166–179.

[51] L. Van Der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2625.