Intelligent Systems with Applications 23 (2024) 200410

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

AdaFN-AG: Enhancing multimodal interaction with Adaptive Feature Normalization for multimodal sentiment analysis

Weilong Liu [a,b], Hua Xu [b], Yu Hua [a], Yunxian Chi [a], Kai Gao [a,*]

[a] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, 050018, China
[b] Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Cross-modal interaction
Adaptive Feature Normalization
Attention Gating

ABSTRACT

In multimodal sentiment analysis, achieving effective fusion among text, acoustic, and visual modalities for enhanced sentiment prediction is a crucial research topic. Recent studies typically employ tensor-based or attention-based mechanisms for multimodal fusion. However, the former fails to achieve satisfactory prediction performance, and the latter complicates the computation of fusion between non-textual modalities. Therefore, this paper proposes the multimodal sentiment analysis model based on Adaptive Feature Normalization and Attention Gating mechanism (AdaFN-AG). Firstly, facing highly synchronized non-textual modalities, we design the Adaptive Feature Normalization (AdaFN) method, which focuses more on sentiment features interaction rather than timing features association. In AdaFN, acoustic and visual modality features achieve cross-modal interaction through normalization, inverse normalization, and mix-up operations, with weights utilized for adaptive strength regulation of the cross-modal interaction. Meanwhile, we design the Attention Gating mechanism that facilitates cross-modal interactions between textual and non-textual modalities through cross-attention and captures timing associations, while the gating module concurrently regulates the intensity of these interactions. Additionally, we employ self-attention to capture the intrinsic correlations within single-modal features. Subsequently, we conduct experiments on three benchmark datasets for multimodal sentiment analysis, with the results indicating that AdaFN-AG outperforms the baselines across the majority of evaluation metrics. Through research and experiments, we validate that AdaFN-AG not only enhances performance by adopting appropriate methods for different types of cross-modal interactions while conserving computational resources but also verifies the generalization capability of the AdaFN method.

1. Introduction

As the internet technology advances swiftly, multimodal data has emerged as a crucial resource for understanding and analyzing the world. Researchers can mine and recognize information and knowledge from multimodal data, enhancing the performance and reliability of data analysis and applications (Chen, 2021). However, the essence of the research resides in exploring the significance of rich multimodal data for comprehensive analysis and application. As a technology capable of identifying sentiment information within data, sentiment analysis holds significant research and exploitation value, such as enhancing the precision of personalized recommendations through the analysis of sentiment in online data (Wang et al., 2022). Furthermore, Multimodal Sentiment Analysis (MSA), as a pivotal branch of sentiment analysis that integrates data from text, audio, and video, enabling more comprehensive and accurate sentiment analysis, has wide applications in product services, medical health, affective computing, and other fields (Chandrasekaran et al., 2021; Stappen et al., 2021; Wang et al., 2021).

The core process of MSA encompasses data collection, post-processing, annotation, unimodal feature extraction, multimodal feature fusion, and sentiment discrimination, with multimodal feature fusion being the pivotal step in multimodal sentiment analysis (Gandhi et al., 2023). Multimodal feature fusion integrates diverse aspects of sentiment across different modalities, thereby enhancing the precision and stability of sentiment analysis outcomes (Liu et al., 2024; Xue et al., 2022). However, how to effectively integrate cross-modal information to improve the accuracy of sentiment analysis, represents major challenges in multimodal fusion (Zhu et al., 2023). The attention mechanism is a commonly used fusion method in MSA tasks, capable of handling incongruities between modalities and enhancing the accuracy of sentiment recognition (Vaswani et al., 2017; Zhu et al., 2023). However, Audio–video data are highly heterogeneous

* Corresponding author.
E-mail address: 1552556972@qq.com (K. Gao).

https://doi.org/10.1016/j.iswa.2024.200410
Received 30 April 2024; Received in revised form 28 May 2024; Accepted 17 June 2024
Available online 22 June 2024

cross-modal inputs, but they also exhibit temporal synchronization; therefore, the computational processing becomes more complex when attention mechanism are used (Frintrop et al., 2010; Yang et al., 2023; Zhu et al., 2023). Additionally, it is clear that conventional algorithms like early fusion do not require extensive computational resources, yet the performance are not strong (Zadeh et al., 2017). Thus, the fusion methods between different modalities for MSA tasks still need improvement.

In this paper, we focus on a fusion method that pursues modal interaction while ensuring computational efficiency. Unlike (Tsai et al., 2019), we do not employ attention mechanism for fusion across all modalities; instead, we adopt appropriate interaction methods tailored to different cross-modal interactions. Our cross-modal fusion method is more akin to an interactive process between different modalities, hence we refer to it as cross-modal interaction. Inspired by the normalization-based fusion (Huang & Belongie, 2017) and mix-up (Liu et al., 2022), we introduce an MSA model based on Adaptive Feature Normalization and Attention Gating mechanism (AdaFN-AG).

Firstly, we designed the Adaptive Feature Normalization (AdaFN) method to facilitate the fusion between acoustic and visual modalities. AdaFN simplifies the interaction process between acoustic and visual modalities through operations such as normalization, weighted mix-up, and inverse normalization per sample, and can adaptively adjust the intensity of interaction. Subsequently, since textual and non-textual modalities are not only heterogeneous but also lack temporal synchronization, we employed the attention mechanism to model their cross-modal interactions. Additionally, we incorporated a gated mechanism to control the interaction intensity of the attention mechanism.

The contributions provided by our research can be encapsulated as detailed below:

- AdaFN-AG employs suitable methods for distinct types of cross-modal interactions, where AdaFN achieves interaction between acoustic and visual modalities while saving computational complexity and enhancing performance.
- The Attention Gating mechanism not only facilitates the fusion of textual and non-textual modalities but also captures the intrinsic relationships within individual modal features.
- Extensive experimental results on MSA benchmark datasets demonstrate that AdaFN-AG surpasses most current baselines, thereby validating the stability and reliability of our proposed model.

2. Related works

We emphasize the construction and innovation of our model by discussing related work on Multimodal Sentiment Analysis and Normalization Fusion.

2.1. Multimodal sentiment analysis

In MSA tasks involving textual, acoustics, and visual modalities, related work has employed various fusion strategies, including early fusion, late fusion, tensor-based fusion, attention-based fusion, and so on, to achieve more accurate analysis of sentiment polarity (Gandhi et al., 2023; Zhu et al., 2023). Zadeh et al. (2017) propose a Tensor Fusion Network approach that obtains tensor representations by computing the outer product between single-modal representations. Liu et al. (2018) propose the Low-rank Multimodal Fusion method, which utilizes low-rank tensors to reduce computational complexity and the number of parameters introduced by the input transformation. Subsequently, Zadeh, Liang, Mazumder, et al. (2018) further proposes the Memory Fusion Network, which continuously models both view-specific and cross-view interactions in multi-view sequences, effectively capturing the dynamic relationships between the visual and acoustic modalities. The Multimodal Transformer model proposed by Tsai et al. (2019), which leverages cross-modal attention mechanism, achieves improvements in handling multimodal human language time-series data without explicit alignment, capturing long-range dependencies. Hazarika et al. (2020) proposed the Modality-Invariant and -Specific Representations, which utilizes cross-modal attention mechanism to learn effective modality representations for multimodal sentiment analysis, surpassing prior models on various benchmarks. Han et al. (2021) proposed the MultiModal InfoMax model, which maintains task-related information through multimodal fusion by hierarchically maximizing mutual information between cross-modal and between the multimodal fusion result and unimodal inputs. Yu et al. (2021) proposed the Self-MM model, which employs a hard sharing strategy for unimodal sub-tasks and designs a weight adjustment method to balance the learning of modality-specific representations, thereby better revealing the association between unimodal and multimodal sentiment information through multi-task learning.

In this paper, we employ above multi-task learning strategy to construct the MSA model and utilize the attention-based mechanism to facilitate interaction between textual and non-textual modalities. However, the interaction method between the acoustic and visual modalities is discussed subsequently.

2.2. Normalization fusion

In the field of image style transfer, methods based on normalization such as batch normalization and batch instance normalization have been widely applied to achieve style transfer (Awais et al., 2020; Choi et al., 2021). In this domain, Adaptive Instance Normalization (AdaIN) can automatically adjust features based on the statistical information of both the source and target images to achieve style transfer, and it accomplishes this at a near real-time processing speed (Huang & Belongie, 2017). In the field of voice conversion, an innovative Voice Conversion system based on AdaIN has been proposed, which effectively balances the trade-off between synthesis quality and speaker similarity (Chen et al., 2021). Wang et al. (2023) introduced an adaptive multimodal learning approach for sentiment analysis, which incorporates a self-adaptive network to facilitate dynamic fusion and adaptive weight assignment among different modalities, but the fusion strategy is implemented based on transformers. The remarkable performance of AdaIN in both visual and acoustic modalities has triggered our investigation into the possibility of cross-modal interaction between these two modalities.

In this paper, inspired by the fast processing speed of AdaIN and its applicability to both visual and acoustic modalities, we implement cross-modal normalization interaction on the feature representations of visual and acoustic modalities for each sample.

3. Model

In this study, the MSA task is defined as a regression problem, and binary classification tasks are achieved based on the polarity of the regression values. Our MSA model based on Adaptive Feature Normalization and Attention Gating (AdaFN-AG) is depicted in Fig. 1. AdaFN-AG takes textual, acoustic, and visual modalities as inputs, and after training, it ultimately outputs the multimodal sentiment prediction result $\hat{y}_{\mathrm{m}}$. During training, the single-modal tasks within the multi-task module generate single-modal sentiment prediction results $\hat{y}_s$. Here and subsequently, $s \in \{\mathrm{t}, \mathrm{a}, \mathrm{v}\}$ respectively represent textual, acoustic and visual modalities. This study only focuses on the multimodal prediction results. The detailed architecture and specific implementation of AdaFN-AG are discussed in subsequent sections.

3.1. Feature extraction

Initially, we extract features from the raw data of the three modalities: text X_{t}, audio X_{a} and video X_{v}, to acquire sentiment features

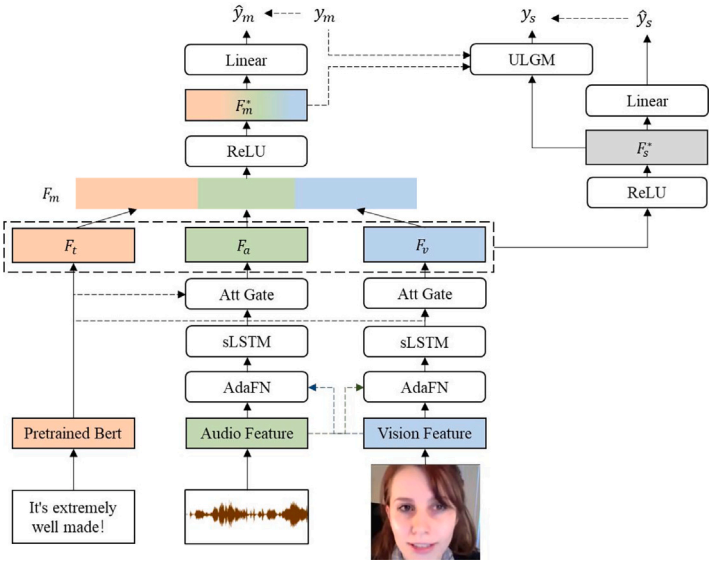


Fig. 1. The overall architecture of AdaFN-AG.

that are optimally aligned for analysis. For the textual modality, we utilize a pre-trained 12-layer BERT (Devlin et al., 2019) model as the text encoder, with a hidden layer size of 768, and select the output from the final layer as the textual feature representation. For the acoustic and visual modalities, this section utilizes openSMILE and OpenFace tools to extract frame-level acoustic and facial features, respectively (Vogel & Ahmad, 2023). Subsequently, Long Short-Term Memory (sLSTM) networks are applied to each modality to capture temporal characteristics, with the terminal state vectors used as the modality feature representations. The process of extracting feature representations for each modality is shown in formulas (1), (2) and (3).

$$\boldsymbol{F}_{\mathrm{t}}^{\mathrm{o}} = \mathrm{BERT}\left(X_{\mathrm{t}}, \theta_{\mathrm{t}}\right) \tag{1}$$

$$\boldsymbol{F}_{\mathrm{a}}^{\mathrm{o}} = \mathrm{sLSTM}\left(X_{\mathrm{a}}, \theta_{\mathrm{a}}\right) \tag{2}$$

$$\boldsymbol{F}_{\mathrm{v}}^{\mathrm{o}} = \mathrm{sLSTM}\left(X_{\mathrm{v}}, \theta_{\mathrm{v}}\right) \tag{3}$$

Where $\boldsymbol{F}_s^{\mathrm{o}} \in \mathbb{R}^{d_s \times l_s}$. $\boldsymbol{F}_s^{\mathrm{o}}$ represents the initial features extracted from each modality, d_s denotes the feature dimension of the modality, l_s represents the feature sequence length of the modality, and θ_s represents the configuration parameters of each modality encoder. Currently, the feature representations of each modality are independent of one another.

3.2. Adaptive feature normalization

This section proposes the Adaptive Feature Normalization (AdaFN) method to facilitate cross-modal interaction between acoustic and visual modalities. Additionally, non-textual modalities are highly synchronized in the temporal sequence, thus theoretically employing the AdaFN method can save computational complexity compared to using attention mechanism to establish cross-modal interactions, and we validate this in our experiments.

As shown in the AdaFN module in Fig. 1, the original modality gains interactive sentiment information from the auxiliary modality through the AdaFN, thereby obtaining a new feature representation. Firstly, the mean and variance need to be calculated along the sequence length for the feature representations of the original and auxiliary modalities of each sample. If the acoustic modality is considered the original modality, then the visual modality acts as the auxiliary modality; Conversely, the same applies. The calculation process of the mean $\boldsymbol{\mu}_x$ and variance $\boldsymbol{\sigma}_x^2$ for the original modality of each sample is shown in formulas (4) and (5).

$$\boldsymbol{\mu}_x = \frac{\sum_{i=1}^{l_x} \boldsymbol{n}_x^{:,i}}{l_x} \tag{4}$$

$$\boldsymbol{\sigma}_x^2 = \frac{\sum_{i=1}^{l_x} \left(\boldsymbol{n}_x^i - \boldsymbol{\mu}_x\right)^2}{l_x} \tag{5}$$

Where the subscript $x \in \{\mathrm{a}, \mathrm{v}\}$ denotes the original modality, $\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2 \in \mathbb{R}^{d_x \times 1}$, $\boldsymbol{n}_x^{:,i}$ represents the slicing operation on a specific sample of the original modality. Subsequently, before calculating the mean and variance of the auxiliary modality, it is necessary to align the dimensions of the auxiliary modality features with the original modality features. Specifically, after separately calculating the mean and variance of the auxiliary modality features along the sequence length, we need to continue the calculation along the feature dimension to obtain a tensor with a shape of $\mathbb{R}^{1\times 1}$. The tensor is replicated along the feature dimension, ultimately resulting in $\mathbb{R}^{d_x \times 1}$ shaped mean and variance values. The calculation process of the mean and variance for the auxiliary modality is shown in formulas (6) and (7).

$$\boldsymbol{\mu}_y = \left(\frac{\sum_{i=1}^{d_y} \boldsymbol{n}_y^{i,:}}{d_y}, \frac{\sum_{i=1}^{l_y} \boldsymbol{n}_y^{:,i}}{l_y}\right) \otimes \boldsymbol{\delta}_{(d_x,1)} \tag{6}$$

$$\boldsymbol{\sigma}_y^2 = \left(\frac{\sum_{i=1}^{d_y} \left(\boldsymbol{n}_y^{i,:} - \boldsymbol{\mu}_y\right)^2}{d_y}, \frac{\sum_{i=1}^{l_y} \left(\boldsymbol{n}_y^{:,i} - \boldsymbol{\mu}_y\right)^2}{l_y}\right) \otimes \boldsymbol{\delta}_{(d_x,1)} \tag{7}$$

Where the subscript $y \in \{\mathrm{a}, \mathrm{v}\}$ denotes the auxiliary modality, $\boldsymbol{\delta}_{(d_x,1)} \in \mathbb{R}^{d_x \times 1}$ represents the identity matrix, and the other symbols is analogous to that used previously. Next, a random number $\alpha \in \mathbb{R}^{1\times 1}$ is generated as the weight parameter to control the intensity of interaction between the auxiliary modality. Subsequently, the new interaction mean $\boldsymbol{\mu}_{\mathrm{mix}}$ and interaction variance $\boldsymbol{\sigma}_{mix}^2$ are calculated through a weighted fusion, as shown in formulas (8) and (9).

$$\boldsymbol{\mu}_{mix} = (1-\alpha)\boldsymbol{\mu}_x + \alpha\boldsymbol{\mu}_y \tag{8}$$

$$\boldsymbol{\sigma}_{mix}^2 = (1-\alpha)\boldsymbol{\sigma}_x^2 + \alpha\boldsymbol{\sigma}_y^2 \tag{9}$$

Where the subscript $mix \in \{\mathrm{a{-}v}, \mathrm{v{-}a}\}$ denotes two interactive states. Upon the completion of data and parameter preparation, the feature representations of the original modality in the samples undergo normalization. Subsequently, inverse normalization is performed using the interaction mean and interaction variance. Eventually, the novel

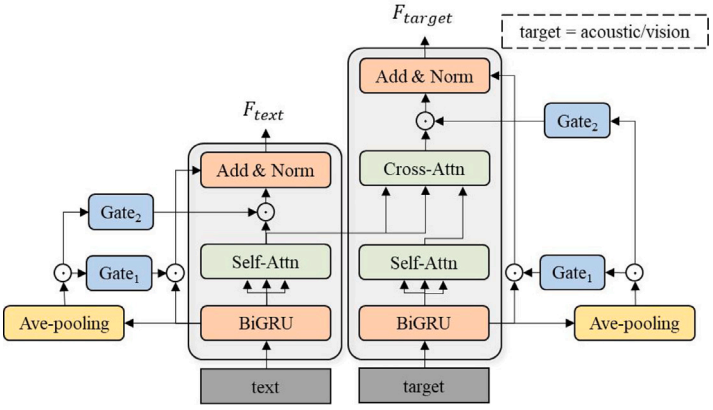


Fig. 2. The architecture of attention gating mechanism.

feature representation of the original modality incorporates the interaction information from the auxiliary modality. The computation of the AdaFN is depicted in formula (10).

$$\text{AdaFN}\left(\boldsymbol{n}_x, \boldsymbol{n}_y\right) = \frac{\sqrt{\boldsymbol{\sigma}_{mix}^2}\left(\boldsymbol{n}_x - \boldsymbol{\mu}_x\right)}{\sqrt{\boldsymbol{\sigma}_x^2 + \varepsilon}} + \boldsymbol{\mu}_{mix} \tag{10}$$

Where ε represents a minuscule value. The final output aligns with the dimensions of the initial feature representation. The AdaFN method is tailored to individual samples, enabling adaptive interaction adjustment without the need for learning new parameters, thus boasting higher computational efficiency than Choi et al. (2021). Acoustic and visual modalities mutually utilize the AdaFN method to achieve cross-modal interaction between non-textual modalities.

3.3. Attention gating mechanism

To address the interaction between textual and non-textual modalities, this section employs an attention module to facilitate cross-modal interaction between the textual and non-textual modalities. Furthermore, this section controls the weights of cross-modal interaction through the gating module.

Fig. 1 delineates the Attention Gating mechanism within AdaFN-AG; subsequent Fig. 2 and the text that follows introduce the details of the Attention Gating mechanism.

Firstly, the initial feature representation of the textual modality is processed through a Bidirectional Gated Recurrent Unit and a Self-Attention module (BiGRU&Self-Attention) to capture the intricate relationships within the textual modality. Here, the gating module generates two weight scores $\boldsymbol{Gate}_{\mathrm{t}}^{\mathrm{Attn}}$, $\boldsymbol{Gate}_{\mathrm{t}}^{\mathrm{BiGRU}} \in \mathbb{R}^{1\times l_{\mathrm{t}}}$ through an activation function (sigmoid) to adjust the balance between the BiGRU encoding and the self-attention layer. The implementations of the module are presented in formulas (11), (12), and (13).

$$\boldsymbol{F}_{\mathrm{t}}^{\mathrm{BiGRU}} = \text{BiGRU}\left(\boldsymbol{F}_{\mathrm{t}}^{\mathrm{o}}\right) \tag{11}$$

$$\boldsymbol{F}_{\mathrm{t}}^{\mathrm{Attn}} = \text{Attention}\left(\boldsymbol{F}_{\mathrm{t}}^{\mathrm{BiGRU}}, \boldsymbol{F}_{\mathrm{t}}^{\mathrm{BiGRU}}, \boldsymbol{F}_{\mathrm{t}}^{\mathrm{BiGRU}}\right) \tag{12}$$

$$\boldsymbol{F}_{\mathrm{t}} = \text{Norm}\left(\boldsymbol{Gate}_{\mathrm{t}}^{\mathrm{BiGRU}} \times \boldsymbol{F}_{\mathrm{t}}^{\mathrm{BiGRU}} + \boldsymbol{Gate}_{\mathrm{t}}^{\mathrm{Attn}} \times \boldsymbol{F}_{\mathrm{t}}^{\mathrm{Attn}}\right) \tag{13}$$

Where $\boldsymbol{F}_{\mathrm{t}}^{\mathrm{BiGRU}}$ and $\boldsymbol{F}_{\mathrm{t}}^{\mathrm{Attn}}$ represent the processing of the textual modality feature representation by the BiGRU layer and the attention layer, respectively. Attention(·) denotes the attention layer, which at this point functions as a self-attention mechanism. Norm(·) is the normalization layer, and the output $\boldsymbol{F}_{\mathrm{t}} \in \mathbb{R}^{\mathrm{d_t}}$ is the final feature representation of the textual modality.

The interaction between the textual modality and the non-textual modalities is realized through a module combining Bidirectional Gated Recurrent Units, self-attention layers, and cross-attention layers (BiGRU&Integrated-Attention). Similarly, the gating module generates two weight scores $\boldsymbol{Gate}_{tar}^{\mathrm{Att}}$, $\boldsymbol{Gate}_{tar}^{\mathrm{BiGRU}} \in \mathbb{R}^{1\times l_{tar}}$ to regulate the balance between the BiGRU encoding and the Integrated-Attention layer, where $tar \in \{\mathrm{a}, \mathrm{v}\}$. The implementations are outlined in formulas (14), (15), and (16).

$$\boldsymbol{F}_{tar}^{\mathrm{BiGRU}} = \text{BiGRU}\left(\boldsymbol{F}_{tar}^{\mathrm{o}}\right) \tag{14}$$

$$\boldsymbol{F}_{tar}^{\mathrm{Attn}} = \text{Cross}-\text{Attn}\left(\boldsymbol{F}_{tar}^{\mathrm{BiGRU}}, \boldsymbol{F}_{t}^{\mathrm{Attn}}, \boldsymbol{F}_{t}^{\mathrm{Attn}}\right) \tag{15}$$

$$\boldsymbol{F}_{tar} = \text{Norm}\left(\boldsymbol{Gate}_{tar}^{\mathrm{BiGRU}} \times \boldsymbol{F}_{tar}^{\mathrm{BiGRU}} + \boldsymbol{Gate}_{tar}^{\mathrm{Attn}} \times \boldsymbol{F}_{tar}^{\mathrm{Attn}}\right) \tag{16}$$

Where Cross-Attn(·) represents the cross-attention layer, and the output $\boldsymbol{F}_{tar} \in \mathbb{R}^{d_{tar}}$ is the final feature representation of the non-textual modalities.

3.4. Sentiment prediction

The final feature representations of the three modalities are denoted as $\boldsymbol{F}_{\mathrm{t}}, \boldsymbol{F}_{\mathrm{a}}$ and $\boldsymbol{F}_{\mathrm{v}}$. Subsequently, the concatenated feature representations of each modality are mapped into a low-dimensional tensor $\boldsymbol{F}_m^* \in \mathbb{R}^{d_m}$.

$$\boldsymbol{F}_m^* = \text{ReLU}\left(\boldsymbol{W}_m^1\left[\boldsymbol{F}_{\mathrm{t}}; \boldsymbol{F}_{\mathrm{a}}; \boldsymbol{F}_{\mathrm{v}}\right] + \boldsymbol{b}_m^1\right) \tag{17}$$

Where ReLU(·) refers to the linear activation function, and $\boldsymbol{W}_m^1, \boldsymbol{b}_m^1 \in \mathbb{R}^{(d_{\mathrm{t}}+d_{\mathrm{a}}+d_{\mathrm{v}})\times d_m}$, denoting the weight matrix and bias vector, respectively. $\boldsymbol{F}_m^*$ is used for multimodal sentiment prediction, and its prediction formula is as follows.

$$\hat{y}_m = \boldsymbol{W}_m^2 \boldsymbol{F}_m^* + \boldsymbol{b}_m^2 \tag{18}$$

Where $\boldsymbol{W}_m^2, \boldsymbol{b}_m^2 \in \mathbb{R}^{d_m\times 1}$.

The single-modal prediction tasks and optimization objectives are referenced from Yu et al. (2021), ultimately completing the construction of the model AdaFN-AG.

4. Experiment setup

This section delineates the experimental setup, encompassing the selection of datasets and baselines, the configuration of parameter settings, and the specification of evaluation metrics.

4.1. Datasets

Table 1 illustrates the distributional information of each benchmark dataset, including the range of sentiment labels for the corresponding dataset samples.

Table 1
Data Statistics for MOSI, MOSEI, and SIMS Datasets. Labels denote the range of true multimodal sentiment labels for each video clip.

Datasets	Training	Validation	Test	Sum	Labels
MOSI	1284	229	686	2199	$[-3, 3]$
MOSEI	16 326	1871	4659	22 856	$[-3, 3]$
SIMS	1368	456	457	2281	$[-1, 1]$

MOSI. The CMU-MOSI (MOSI) (Zadeh et al., 2016) dataset consists of 2199 opinion video clips, each annotated with a sentiment label ranging from −3 to 3.

MOSEI. The CMU-MOSEI (MOSEI) (Zadeh, Liang, Poria, et al., 2018) dataset, an extension of the original MOSI dataset, includes a broader vocabulary, more diverse samples, and a wider range of speakers and topics.

SIMS. The CH-SIMS (SIMS) (Yu et al., 2020) dataset is a Chinese multimodal sentiment analysis dataset with both multimodal and independent unimodal annotations, comprising 2281 video clips from television shows and variety programs.

4.2. Baselines

We compare AdaFN-AG to the model with outstanding performance in the field of multimodal sentiment analysis, thereby fully validating its performance.

TFN. The Tensor Fusion Network (TFN) (Zadeh et al., 2017) obtains tensor representations by concatenating the outer products of single-modal representations.

LMF. The Low-rank Multimodal Fusion (LMF) (Liu et al., 2018) utilizes low-rank tensors to reduce the computational complexity and parameter amount of input transformation.

MFN. The Memory Fusion Network (MFN) (Zadeh, Liang, Mazumder, et al., 2018) enables interaction between visual and acoustic modalities through continuous modeling of cross-view interactions and temporal integration.

MulT. The Multimodal Transformer (MulT) (Tsai et al., 2019) utilizes a bidirectional cross-modal Transformer structure to enable interaction between unaligned multimodal sequences.

MISA. The Modality Invariant and Specific Representations (MISA) (Hazarika et al., 2020) integrates various comprehensive losses to learn modality invariant and modality-specific representations.

MMIM. The MultiModal InfoMax(MMIM) (Han et al., 2021) guides the model to learn shared representations from multimodal data through hierarchical interaction information maximization.

SELF-MM. The Self-supervised Multi-task Multimodal analysis model (SELF-MM) (Yu et al., 2021) features a label generation module based on self-supervised learning strategies, acquiring independent single-modal supervision. It learns consistency and difference through joint training of multi-modal and single-modal tasks.

CENet. The Cross-modal Enhancement Network (CENet) (Di Wang, Liu, Wang, et al., 2023) leveraging a transformer to integrate visual and acoustic data with text for sentiment analysis, utilizes an enhancement module to align nonverbal and verbal cues, enhancing recognition accuracy.

TETFN. The Text Enhanced Transformer Fusion Network (TETFN) (Di Wang, Guo, Tian, et al., 2023) creates unified multimodal representations by learning text-oriented cross-modal mappings, integrating text with nontextual representations through attention and prediction mechanisms.

4.3. Parameter settings

The proposed AdaFN-AG utilizes the Adam optimizer, initializing the learning rate for Bert at $4.5e-5$ and $1e-3$ for other parameters. Theoretically, the training parameter α in AdaFN should be within the range of $(0.0, 1.0)$. However, experimental results present that the model trains most effectively within the range of $(0.0, 0.3)$. The default configuration of Attention Gating mechanism includes two heads and a dropout rate of 0.1.

4.4. Metrics

To fairly and comprehensively evaluate the models' performance, we integrate AdaFN-AG and baselines on the M-SENA (Mao et al., 2022) platform, which provides a unified environment for integrating diverse MSA models and a standardized set of evaluation metrics. In regression tasks, the mean absolute error (MAE) and the Pearson correlation coefficient (Corr) serve as the evaluation metrics. For binary classification tasks that classify sentiment values into positive/negative (excluding zero), the metrics include binary classification accuracy (Acc2) and the F1 score (F1). The MAE is prioritized as the primary metric, with Acc2, F1, and Corr following. Higher values are considered to indicate better model performance, except for MAE.

5. Results and analysis

5.1. Comparison study

We evaluate the performance of our AdaFN-AG by comparing it with advanced baselines, each employing distinct mechanisms for multimodal fusion. Table 2 presents the comparative results of each model in the multimodal sentiment analysis task across three benchmark datasets, with AdaFN-AG achieving superior performance than all baseline models on the majority of evaluation metrics. Specifically, it outperforms the best performing SELF-MM by 0.005 MAE, 0.61% Acc2, 0.59% F1, and 0.008 Corr on MOSI. Additionally, the quantitative results on MOSEI and SIMS are also improved compared to the baselines.

Quantitative analysis indicates that AdaFN-AG demonstrates state-of-the-art performance across three datasets, concurrently validating its robust generalization. However, AdaFN-AG is slightly less effective than CENet in correlation on MOSEI, which can be attributed to its enhanced long-term emotional cue capture and the complexity of the MOSEI dataset. Additionally, CENet's limited generalization leads to poor performance on SIMS. Including AdaFN-AG, SELF-MM and TETFN exhibit outstanding performance across different datasets, reflecting the reliable generalization capabilities of employing the multitask backbone network. Furthermore, attention-based baselines, such as MulT and TETFN, are generally more effective than tensor-based baselines like TFN and LMF, but the use of attention mechanisms comes with increased computational complexity.

5.2. Ablation study

In this section, in order to delve into the roles of each module in AdaFG-AG, we conducted a comprehensive ablation study of AdaFG-AG on MOSI. We conduct ablation studies focused on the AdaFN approach and the Attention Gating (AG) mechanism within AdaFN-AG to assess the contributions of each module to the model's performance across different cross-modal interactions. The specific ablation variants of AdaFN-AG are shown as follows, and Table 3 presents the results of the ablation study.

- N. This means that AdaFN or AG module is completely ablated.
- $V \rightarrow A$. Cross-modal interaction module in AdaFN, where the visual modality serves as an auxiliary modality, and the acoustic modality is the original modality.

Table 2
Comparison study results on MOSI, MOSEI and SIMS. The AdaFN-AG is our model, while the other models are baselines.

Datasets	Models	MAE	Acc 2 (%)	F1 (%)	Corr
MOSI	TFN	0.931	79.12	79.08	0.657
	LMF	0.928	79.57	79.42	0.656
	MFN	0.954	78.96	78.92	0.651
	MulT	0.881	81.40	81.45	0.705
	MISA	0.782	83.08	83.17	0.786
	MMIM	0.751	83.99	84.08	0.782
	SELF-MM	0.711	84.45	84.48	0.795
	CENet	0.736	84.91	84.93	0.796
	TETFN	0.723	84.15	84.18	0.798
	ADAFN-AG	**0.706**	**85.06**	**85.07**	**0.803**
MOSEI	TFN	0.573	83.16	83.09	0.720
	LMF	0.576	83.48	83.36	0.717
	MFN	0.566	83.02	83.13	0.729
	MulT	0.559	84.37	84.30	0.732
	MISA	0.554	85.14	85.14	0.755
	MMIM	0.565	83.57	83.28	0.742
	SELF-MM	0.530	84.87	84.92	0.757
	CENet	0.532	84.59	84.70	**0.774**
	TETFN	0.542	85.61	85.61	0.764
	ADAFN-AG	**0.529**	**85.80**	**85.70**	0.768
SIMS	TFN	0.428	78.99	79.32	0.601
	LMF	0.442	76.37	79.37	0.563
	MFN	0.438	78.99	78.62	0.571
	MulT	0.441	78.34	77.86	0.589
	MISA	0.447	76.37	75.82	0.542
	MMIM	0.449	76.81	76.66	0.520
	SELF-MM	0.424	79.43	79.30	0.572
	CENet	0.446	74.62	74.97	0.548
	TETFN	0.417	78.56	78.64	0.588
	ADAFN-AG	**0.408**	**79.43**	**79.30**	**0.607**

Table 3
Ablation study results of AdaFN-AG on MOSI.

AdaFN	AG	MAE	Acc2(%)	F1(%)	Corr
N	N	0.756	83.08	83.16	0.795
$V \rightarrow A$	N	0.711	84.45	84.43	0.796
$A \rightarrow V$	N	0.711	83.99	84.00	0.799
$A \leftrightarrow V$	N	0.710	83.99	83.93	0.792
N	$T \leftrightarrow A$	0.714	84.76	84.68	0.792
N	$T \leftrightarrow V$	0.722	83.99	83.97	0.788
N	$T \leftrightarrow A+V$	0.708	84.45	84.43	0.795
$A \leftrightarrow V$	$T \leftrightarrow A+V$	**0.706**	**85.06**	**85.07**	**0.803**

- $A \rightarrow V$. Cross-modal interaction module in AdaFN, where the acoustic modality serves as an auxiliary modality, and the visual modality is the original modality.
- $T \leftrightarrow A$. Cross-modal interaction module between textual modality and acoustic modality in AG.
- $T \leftrightarrow V$. Cross-modal interaction module between textual modality and visual modality in AG.
- $A \leftrightarrow V$. Cross-modal interaction module in AdaFN, including both $V \rightarrow A$ and $A \rightarrow V$.
- $T \leftrightarrow A+V$. Cross-modal interaction module in AG, including both $T \leftrightarrow A$ and $T \leftrightarrow V$.

The outcomes of our ablation study underscore a progressive enhancement in performance with the incrementation of cross-modal interaction modules, with the non-ablated AdaFN-AG yielding the most favorable results. This validates the soundness and efficacy of the design of each individual module. Additionally, the results of only retaining the AdaFN module in AdaFN-AG are superior to the attention-based baselines like MulT, indicating that AdaFN is more effective in interactions with non-textual modalities. However, the slight decrease in performance when only retaining AdaFN compared to only retaining the AG module is due to the fact that AG is established based on the textual modality, while the dominance of textual modalities is an inherent challenge in MSA.

Table 4
Portability study results of AdaFN on MOSI. N and Y denote the original and AdaFN-ported models, respectively.

Models	AdaFN	MAE	Acc2 (%)	F1 (%)	Corr
MFN	N	0.954	78.96	78.92	0.651
	Y	**0.928**	**80.64**	**80.64**	**0.687**
MulT	N	**0.881**	81.40	81.45	**0.705**
	Y	0.893	**82.47**	**82.27**	0.694
MISA	N	0.782	83.08	83.17	0.786
	Y	**0.721**	**84.30**	**84.27**	**0.799**
MMIM	N	0.751	83.99	**84.05**	0.782
	Y	**0.723**	**83.99**	84.05	**0.791**
SELF-MM	N	0.711	84.45	84.48	**0.795**
	Y	**0.711**	**84.60**	**84.61**	0.794

5.3. Portability study

The AdaFN module constitutes our core innovation, while theoretically possessing portability. To validate this portability and assess the efficacy of its integration with different fusion mechanisms, we incorporate the AdaFN into select representative baselines. Table 4 presents the comparative results of these baselines when augmented with the AdaFN module.

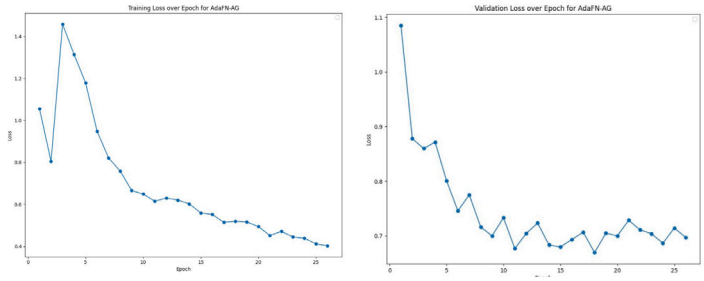
Compared to the original baselines, the performance of each baseline integrated with AdaFN has been enhanced to varying degrees, particularly in the cases of MISA and MMIM, which improved by 0.061 and 0.028 MAE respectively, showcasing the portability and adaptability of AdaFN. However, since MulT already utilizes an attention mechanism to facilitate interaction between non-text modalities, the addition of AdaFN does not yield substantial improvements. Therefore, it can be confirmed that in future MSA research, AdaFN can serve not only as a suitable and effective method to enhance interactions among non-text modalities but also be combined with other fusion mechanisms to further facilitate interaction between non-text modalities.

5.4. Case study

To further demonstrate the reliability and rationality of AdaFN-AG, we select three multimodal cases from the MOSI dataset, as shown in Table 5. From cases 1 and 2, it can be observed that when the expression of sentiment each modality is more prominent, the prediction values of the AdaFN-AG model are very close to the true label values. However, we find case 3 where the prediction values deviate from the true labels. This is attributed to the extensive content in the textual modality and the less pronounced sentiment expression in each modality. Nonetheless, the AdaFN-AG's prediction of sentiment polarity is accurate. Overall, the variant without the AdaFN module performs worse than the original model, further validating the reliability of the AdaFN module.

5.5. Training process study

To comprehensively analyze the AdaFN-AG training process, we compare the detailed information of the AdaFN-AG and baseline training processes and visualize the changes in loss for the training and validation sets during the AdaFN-AG training, as shown in Fig. 3. From the figure, we can observe that the loss value for the training set consistently decreases, and the loss for the validation set also trends downwards overall, indicating that the AdaFN-AG is being effectively trained.



(a) Training loss (b) Validation loss

Fig. 3. Training and validation loss curve for the AdaFN-AG on MOSI. The horizontal axis represents the epoch, and the vertical axis shows the loss value calculated for each epoch.

Table 5
Case study for the AdaFN-AG on MOSI. Example illustrates the content of each multimodal case, where T represents text, A stands for acoustics, and V signifies visual modalities. Label denotes the manually annotated truth labels, while AdaFN-AG and AG represent the automatic prediction values of the original model and the variant model without the AdaFN module, respectively.

	Example	Label	AdaFN-AG	AG
T	But it was a fun experience going.			
A	Pauses and emphasis.	2.25	2.07	1.90
V	Nod and smile.			
T	He's very boring as well.			
A	Gradually increasing tone.	−1.75	−1.57	−1.52
V	Frown and pout.			
T	I don't think he got mad when hah…			
A	Slight Sighs.	−0.25	−1.04	−0.91
V	Upward gaze.			

Previous work in MSA has concluded that attention-based models exhibit good performance but are computationally complex, especially models like MulT that rely solely on attention-based interactions. Meanwhile, tensor-based models are computationally less complex, but their performance is generally lower, as exemplified by TFN. Our model, AdaFN-AG, as well as certain baselines like TETFN, are still considered attention-based models but are not entirely dependent on attention mechanisms. AdaFN-AG, for instance, reduces its reliance on attention mechanisms for interactions between non-textual modalities through AdaFN method. Therefore, we choose to compare the training times of AdaFN-AG, MulT, TFN and TETFN, aiming to investigate the trade-off between performance and computational complexity. Specifically, under identical conditions, we repeat the experiments with AdaFN-AG and the baselines multiple times, automatically logging the training duration for each trial. After discarding the longest and shortest training duration, we obtain the average training duration for each model.

Fig. 4 offers an intuitive visualization of the training times of AdaFN-AG and baselines across different datasets. The figure reveals that TFN's shortest training duration underscores its exceptional computational efficiency, but its lower predictive performance can not be ignored. It is evident that MulT requires the longest training time, which signifies that the utilization of attention mechanisms has increased computational complexity. Simultaneously, the effective capture of complementary information is a significant contribution of the attention mechanism. The most significant point is that AdaFN-AG and TETFN achieve relatively favorable training durations, with training time on the large-scale dataset MOSEI being less than one-third of MulT's. AdaFN-AG exhibits improved training times on MOSI and SIMS compared to TETFN, while it is slightly less efficient on MOSEI, with the differences being marginal. When combined with the comparison study and computational efficiency analysis, AdaFN-AG exhibits superior MSA performance. It can be attributed to the appropriate use of attention mechanisms and AdaFN method in cross-modal interactions.

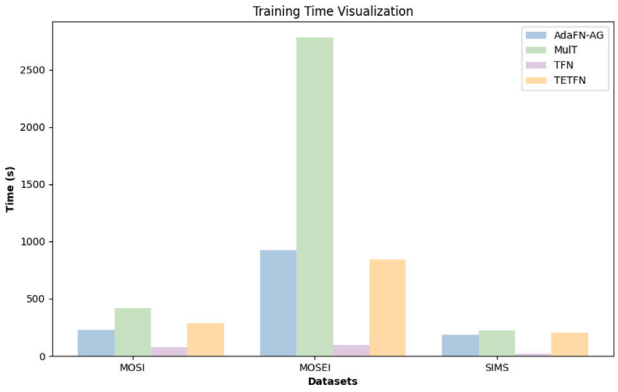


Fig. 4. Visualization of training time for AdaFN-AG and representative baselines on MOSI, MOSEI, and SIMS.

6. Conclusions and further works

We introduce a multimodal sentiment analysis model based on Adaptive Feature Normalization and Attention Gating, employing suitable interaction methods for diverse cross-modal interactions and achieving significant performance improvements. Specifically, the AdaFN method utilizes more efficient design to achieve interaction between non-textual modalities and improve task performance. Concurrently, the Attention Gating mechanism facilitates interaction between textual and non-textual modalities, capturing temporal relationships across modalities.

In the event that large multimodal models proficient in analyzing video data gain wider adoption in the forthcoming years, we intend to delve into the underlying principles of these models and explore their optimization strategies.

CRediT authorship contribution statement

Weilong Liu: Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, Writing – original draft. **Hua Xu:** Project administration, Data curation, Supervision, Resources, Writing – review & editing. **Yu Hua:** Resources, Writing – review & editing. **Yunxian Chi:** Funding acquisition, Supervision, Writing – review & editing. **Kai Gao:** Project administration, Resources, Funding acquisition, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the Hebei Natural Science Foundation, China [grant numbers F2022208006, F2023207003] and the Science Research Project of the Hebei Education Department, China [grant number QN2024196].

References

Awais, M., Iqbal, M. T. B., & Bae, S. H. (2020). Revisiting internal covariate shift for batch normalization. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 5082–5092.

Chandrasekaran, G., Nguyen, T. N., & D., J. H. (2021). Multimodal sentimental analysis for social media applications: A comprehensive review. *WIREs Data Mining Knowledge Discovery*, *11*(5), 1415.

Chen, S. (2021). Embracing multimodal data in multimedia data analysis. *IEEE Multimedia*, *28*(3), 5–7.

Chen, Y. H., Wu, D. Y., Wu, T. H., & Lee, H. (2021). Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 5954–5958). IEEE.

Choi, S., Kim, T., Jeong, M., Park, H., & Kim, C. (2021). Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3425–3435).

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).

Di Wang, Guo, X., Tian, Y., Liu, J., He, L., & Luo, X. (2023). TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, *136*, Article 109259.

Di Wang, Liu, S., Wang, Q., Tian, Y., He, L., & Gao, X. (2023). Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, *25*, 4909–4921.

Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans.actions on Applied Perception*, *7*(1), 1–39.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444.

Han, W., Chen, H., & Poria, S. (2021). Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 9180–9192).

Hazarika, D., Zimmermann, R., & Poria, S. (2020). Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1122–1131).

Huang, X., & Belongie, S. J. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 1501–1510).

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2247–2256).

Liu, Y., Yuan, Z., Mao, H., Liang, Z., Yang, W., Qiu, Y., Cheng, T., Li, X., Xu, H., & Gao, K. (2022). Make acoustic and visual cues matter: CH-SIMS v2. 0 dataset and AV-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction* (pp. 247–258).

Liu, Z., Zhou, B., Chu, D., Sun, Y., & Meng, L. (2024). Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, *101*, Article 101973.

Mao, H., Yuan, Z., Xu, H., Yu, W., Liu, Y., & Gao, K. (2022). M-SENA: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th annual meeting of the association for computational linguistics: system demonstrations* (pp. 204–213).

Stappen, L., Baird, A., Schumann, L., & Schuller, B. W. (2021). The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing*, *14*(2), 1334–1350.

Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th conference of the association for computational linguistics* (pp. 6558–6569).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems 30: annual conference on neural information processing systems 2017* (pp. 5998–6008).

Vogel, C., & Ahmad, K. (2023). Agreement and disagreement between major emotion recognition systems. *Knowledge-Based Systems*, *276*, Article 110759.

Wang, Z., Gao, P., & Chu, X. (2022). Sentiment analysis from Customer-generated online videos on product review using topic modeling and Multi-attention BLSTM. *Advanced Engineering Informatics*, *52*, Article 101588.

Wang, J., Mou, L., Ma, L., Huang, T., & Gao, W. (2023). AMSA: adaptive multimodal learning for sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(3s), 1–21.

Wang, Z., Yin, Z., & Argyris, Y. A. (2021). Detecting medical misinformation on social media using multimodal deep learning. *IEEE Journal of Biomedical Health Informatics*, *25*(6), 2193–2203.

Xue, X., Zhang, C., Niu, Z., & Wu, X. (2022). Multi-level attention map network for multimodal sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 5105–5118.

Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., Cao, X., & Ren, K. (2023). Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, *18*, 2015–2029.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., & Yang, K. (2020). Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3718–3727).

Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 12* (pp. 10790–10797).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1103–1114).

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. P. (2018). Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1* (pp. 5634–5641).

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). 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* (pp. 2236–2246).

Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, *31*(6), 82–88.

Zhu, L., Zhu, Z., Zhang, C., Xu, Y., & Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, *95*, 306–325.