Information Sciences 675 (2024) 120682

Contents lists available at ScienceDirect

Information Sciences

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

An autoencoder-based self-supervised learning for multimodal sentiment analysis

Wenjun Feng, Xin Wang, Donglin Cao *, Dazhen Lin *

Department of Artificial Intelligence, Xiamen University, Xiamen, Fujian, China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Self-supervised learning
Autoencoder
Contrastive learning

ABSTRACT

Representation learning is a crucial and challenging task within multimodal sentiment analysis. Effective multimodal sentiment representations contain two key aspects: consistency and difference. However, the state-of-the-art multimodal sentiment analysis approaches failed to capture the difference and consistency of sentiment information across diverse modalities. To address the multimodal sentiment representation problem, we propose an autoencoder-based self-supervised learning framework. In the pre-training stage, an autoencoder is designed for each modality, leveraging unlabeled data to learn richer sentiment representations for each modality through sample reconstruction and modality consistency detection tasks. In the fine-tuning stage, the pre-trained autoencoder is injected into MulT (AE-MT) and enhance the model's ability to extract deep sentiment information by incorporating a contrastive learning auxiliary task. Our experiments on the popular Chinese sentiment analysis benchmark (CH-SIMS v2.0) and English sentiment analysis benchmark (MOSEI) demonstrate significant gains over baseline models.

1. Introduction

Multimodal sentiment analysis has attracted more and more attention in recent years [1,2]. With the thriving development of the social media world, multimodal sentiment analysis has found widespread applications in areas such as public opinion management [3] and video comprehension. Multimodal models, when compared to traditional text-based unimodal sentiment analysis [4], demonstrate enhanced robustness and the ability to seamlessly adapt to social media data. Many advanced approaches have focused on extracting rich multimodal features containing abundant sentiment information from heterogeneous multimodal data [5,6].

Though previous works have made impressive improvements on benchmark datasets [7,8], these state-of-the-art methods have overlooked the consistency and difference multimodal representation. Therefore, they cannot learn a complementary and unified multimodal sentiment representation which is important for multimodal sentiment analysis. As for consistent sentiment information, it manifests as the inter-modal correlation between sentiment and semantics. As for differentiated sentiment information, it reflects the uniqueness and complementarity of heterogeneous modalities in sentiment expression. To achieve accurate multimodal sentiment analysis, it is essential to acquire improved multimodal representations. Specifically, the challenge lies in obtaining better consistent representations without losing the modality-specific differentiation. For example, concerning the text and visual modalities of the same video sample, the text modality is more proficient in conveying comprehensive semantic information, while the visual modality

* Corresponding authors.
E-mail addresses: 1257996440@qq.com (W. Feng), 845003771@qq.com (X. Wang), another@xmu.edu.cn (D. Cao), dzlin@xmu.edu.cn (D. Lin).

https://doi.org/10.1016/j.ins.2024.120682
Received 8 January 2024; Received in revised form 25 February 2024; Accepted 29 April 2024

Available online 22 May 2024

excels at capturing subtle emotional changes over time. Through the combination of these two modalities, the sentiment can be expressed more comprehensively.

To address the issue of multimodal sentiment representation mentioned above, we propose a self-supervised learning framework based on autoencoders. During the pre-training stage, an independent autoencoder is designed for each modality and trained by using unlabeled data with two pre-training tasks: sample reconstruction and modality-consistency detection. The sample reconstruction task involves reconstructing the representation of each modality, allowing the model to learn modality-specific differentiated sentiment information within each modality. The modality-consistency detection task focuses on learning the ability to discern tampered samples by jointly considering the consistency of sentiment information across all modalities. The autoencoders, after pre-training, possess the capability to extract both modality-specific differentiated sentiment information and modality-consistent sentiment information. The combination of these two types of information guides the model to make more accurate decisions. During the fine-tuning stage, the pre-trained autoencoders and MulT [9] (AE-MT) are integrated. Furthermore, the cross-modal attention mechanism is used to facilitate inter-modal interactions. Additionally, to address the issue of sentiment diffusion in the high-level semantic space, a contrastive learning [10] auxiliary task is introduced in the second stage to enhance the model's comprehension of sentiment expressions.

The novel contributions of this paper can be summarized as:

- A self-supervised learning framework based on autoencoders is proposed to improve the model's ability to extract both modality-specific differentiated and modality-consistent sentiment information and give a comprehensive and disentangled view of the multimodal data.
- To solve the sentiment diffusion problem in high-level semantic spaces, we introduce a contrastive learning auxiliary task which improves the model's understanding of sentiment expression by maximizing the similarity between the sample and positive sample, minimizing the similarity between the sample and negative sample.
- Extensive experiments on CH-SIMS v2.0 [11] Chinese multimodal dataset and MOSEI [8] dataset validate the effectiveness of our framework.

The remainder of this paper is organized as follows. Section 2 briefly describes the background and related work in multimodal sentiment analysis. Section 3 presents the self-supervised learning framework based on autoencoders and analyzes the time complexity of the proposed model. Experimental results are given in Section 4. Section 5 draws conclusions and gives directions for future work.

2. Related works

Multimodal sentiment analysis has become a significant research topic that integrates both language and non-language information such as text, vision, and audio data. Learning from heterogeneous multimodal data primarily relies on two key techniques: multimodal representation and multimodal fusion.

2.1. Multimodal representation learning

Multimodal representation learning [12–14] can be categorized into joint representation and coordinated representation based on whether the representations share a common feature space. Joint representation, also known as a single-stream network, integrates information from multiple unimodal sources and maps them into a unified feature space, with a focus on capturing the complementarity between modalities and eliminating redundant information in multimodal data. References [15–17] fall under the category of joint representation learning. They concatenate multimodal inputs and feed them into Transformer-based encoders, using attention mechanisms to learn implicit alignment between modalities, thereby achieving a unified representation of multimodal information. Coordinated representation, also known as dual-stream networks, involves using multiple networks to separately process information from each modality and project them to their respective feature spaces. Similarity constraints are imposed between modalities to strengthen their correlations. Lu et al. [18] independently encoded different modalities and utilized cross-modal attention mechanisms to learn the interaction of representations from each modality.

2.2. Multimodal fusion learning

Heterogeneous multimodal information resides in different feature spaces, containing sentiment information from various perspectives and levels. Therefore, it's crucial to choose effective multimodal fusion [19] strategies to select important sentiment information from different modalities. For multimodal fusion, according to the fusion strategy, previous works can be classified into three categories: simple fusion, tensor-based fusion, and attention-based fusion. Simple fusion involves merging information from different modalities through operations such as concatenation or averaging of features. References [20,21] utilized multiple networks to extract representations separately from each modality and then directly concatenated them to obtain multimodal representations. However, this straightforward fusion approach difficult to fully leverage the interaction information between different modalities, thereby affecting the effectiveness of multimodal fusion. Tensor-based fusion methods involve performing tensor outer product operations on features from different modalities, resulting in higher-dimensional multimodal tensors. Zadeh et al. [12] employed a tensor fusion network to acquire tensor representations through the computation of outer products between unimodal representations.

Table 1
Method comparison.

Methods	Transformer	Unlabeled data	Consistency	Difference
MulT (2019)	used	unused	no	no
BERT_MAG (2020)	used	unused	no	no
MISA (2020)	used	unused	similarity	orthogonality constraint
MMIM (2021)	unused	unused	no	no
SELF-MM (2021)	unused	unused	no	label generation
AV-MC (2022)*	unused	unused	no	no
AV-MC (semi) (2022)*	unused	used	no	no

Furthermore, Liu et al. [13] introduced a low-rank multimodal fusion approach to reduce the computational complexity associated with tensor-based methods. Attention-based fusion methods leverage attention mechanisms to dynamically and selectively focus on features from different modalities, allowing for the full utilization of the interrelationships between different modalities. MulT [9], an attention-based method, proposed cross-modal transformers, which learn the cross-modal attention to reinforce a target modality. Sunny et al. [22] proposed a multimodal fusion framework, named DeepCU, which extracts the common information from the multimodal representations and integrate two aspects of multimodal information via a fusion layer.

2.3. Multimodal sentiment analysis

By using the multimodal representation and multimodal fusion, some state-of-the-art approaches have been proposed in recent years. Their characteristics can be summarized in Table 1. It shows that most of the model have overlooked consistency and difference multimodal sentiment representation. Although MISA has noticed this problem, how to measure the consistency and difference in representation learning is not further studied. To learn an effective multimodal sentiment representation, in this paper, a self-supervised learning framework based on autoencoders is proposed to improve the model's ability to extract both modality-specific differentiated and modality-consistent sentiment information and give a comprehensive and disentangled view of the multimodal data. Furthermore, our approach effectively uses the unlabeled data to learn a multimodal sentiment representation which greatly reduces the manual labeling work.

The two-stage approach proposed in this paper utilizes both single-stream and dual-stream networks to extract multimodal features in the respective stages. Ultimately, the model leverages MulT-based cross-modal Transformers for modal fusion.

3. Methodology

In this section, we provide a detailed explanation of our self-supervised learning framework based on autoencoders, which consists of two stages: pre-training and fine-tuning. In the pre-training stage, unlabeled data and two pre-training tasks are utilized to equip the autoencoders with the capability to extract differentiated and consistent sentiment information from heterogeneous modalities. In the fine-tuning stage, a contrastive learning auxiliary task based on sentiment intensity is designed on labeled data. This stage combines the multimodal representation extraction ability of the pre-trained autoencoders and the multimodal fusion capability of MulT to capture richer and deeper sentiment information.

3.1. Task setup

Multimodal sentiment analysis assesses sentiment intensity based on three heterogeneous sources of information: text (t), vision (v), and audio (a). Typically, multimodal sentiment analysis can be modeled as either a classification task or a regression task. In the context of a classification task, the model is tasked with determining which sentiment category a sample belongs to. Conversely, in a regression task, the model is expected to directly provide the sentiment intensity of a sample as a specific numerical value. In this paper, our approach is evaluated from both of these perspectives.

3.2. Modality representation learning

To provide AE-MT with a better initialization that can capture differentiated and consistent sentiment information across heterogeneous modalities, we employ two pre-training tasks to train the autoencoders in the pre-training stage.

3.2.1. Feature extraction

As illustrated in Fig. 1, for a given sample $X = (X_t, X_v, X_a)$, where X_t, X_v, and X_a represent text, vision, and audio time-series modalities, respectively. For the text modality X_t, we use pre-trained BERT [23] to extract representations and take the vector of the first word as the whole sentence embedding. For the vision modality X_v, we employ TalkNet [24] to locate the speaker and then use the OpenFace2.0 [25] toolkit to extract facial features. For the audio modality X_a, we utilize the openSMILE [26] open-source tool to extract the frequency spectrum features.

To handle the temporal features of the three modalities, a 1D temporal convolutional layer is used to capture contextual information and uniformly transform the dimensions to d for subsequent computations:

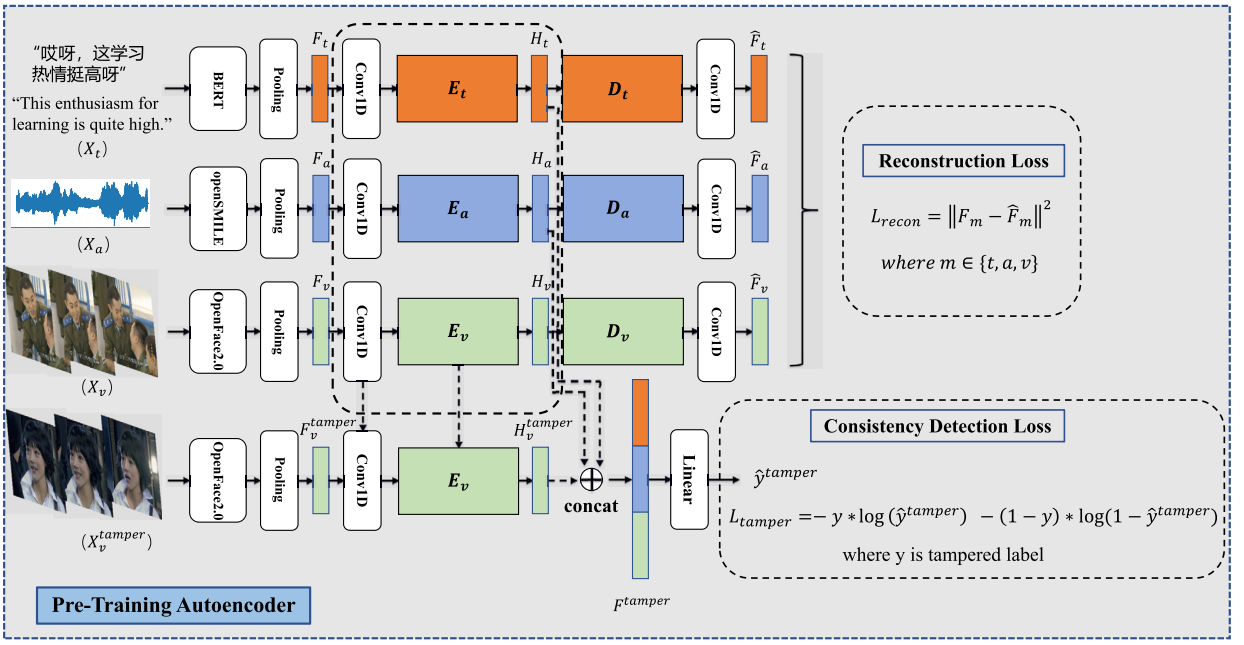

Fig. 1. Pipeline of the pre-training stage. In this stage, the autoencoders are pre-trained by using the sample reconstruction task and the modality-consistency detection task. These two pre-training tasks provide the model with the capability to extract differential sentiment information and consistent sentiment information across heterogeneous modalities. Finally, the model is optimized through L_{recon} and L_{tamper}.

$$F_m = Conv1D(X_m, k_m) \tag{1}$$

where $m \in \{t, a, v\}$, k_m denotes the size of the convolutional kernel for each modality. Additionally, the key component, the autoencoders, consists of three sub-networks corresponding to the text, audio, and vision modalities. Each sub-network adopts the Transformer architecture [27], which includes an encoder E_m and decoder D_m. The self-attention mechanism within the Transformer captures the dependencies between different positions in the input sequence, making it particularly suited for handling variable-length multimodal sequence data. Then, the encoders E_m compute the hidden feature H_m:

$$H_m = E_m(F_m) = Encoder_m(F_m, \theta_{E_m}) \tag{2}$$

where the encoder E_m consists of multiple layers of Transformer encoder, and θ_{E_m} represents the parameters of the encoder for modality m.

3.2.2. Differentiated sentiment representation

To enable the autoencoders to capture differentiated sentiment information for each modality, we introduce the sample reconstruction task during the pre-training stage. Specifically, the decoder D_m reconstructs the input data F_m based on the hidden features H_m,

$$F'_m = D_m(H_m) = Decoder_m(H_m, \theta_{D_m}) \tag{3}$$

Where the decoder D_m consists of multiple layers of Transformer decoder, and θ_{D_m} represents the parameters of the decoder for modality m. To facilitate the computation of sample reconstruction loss, the output F'_m from the decoder is passed through a 1D temporal convolutional layer to obtain the reconstructed feature $\hat{F}_m$ with the same dimension as the original sample F_m,

$$\hat{F}_m = Conv1D(F'_m, k_m) \tag{4}$$

where k_m denotes the kernel size for each modality. We introduce a reconstruction loss, denotes L_{recon}, to ensure that the hidden representations capture the specific details of each modality. In this paper, the Mean Squared Error (MSE) loss is used for calculating the reconstruction loss.

$$L_{recon} = ||\hat{F}_m - F_m||^2 \tag{5}$$

3.2.3. Consistent sentiment representation

To capture the inter-modal consistency of sentiment information, the modality-consistency detection auxiliary task is designed for the autoencoder. During the construction of training data, one of the modalities of the sample is replaced with a certain probability p^{tamper} and assigned a label y^{tamper} to the sample after replacement. Specifically, for the samples that have undergone replacement,

their modality-consistency detection label is set to 1, while those that have not been replaced are labeled as 0. Taking the case where the vision modality has been tampered with as an example, suppose the original sample is $X = X_t, X_v, X_a$, and the sample with tampered vision information becomes X^{tamper},

$$X = \{X_t, X_v^{tamper}, X_a, y^{tamper}\} \tag{6}$$

where X_m^{tamper} represents samples in which one modality has been replaced and $y^{tamper} \in \{0, 1\}$ denotes whether the sample has been tampered. Finally, the optimized objective of the modality-consistency detection task is to train the autoencoder to detect which samples have been tampered with, thereby learning the inter-modal consistency of sentiment information.

In this stage, the modality-consistency detection task is considered as a classification task. We define a binary classifier C_m to determine whether a sample has been tampered with, indicating whether the modal information is consistent. The output feature H_m of the autoencoder E_m is concatenated and used as input to the classifier C_m,

$$F^{tamper} = [H_t; H_v^{tamper}; H_a] \tag{7}$$

$$\hat{y}^{tamper} = C_m(F^{tamper}) \tag{8}$$

where $\hat{y}^{tamper}$ represents the predicted probability of the sample being tampered with given the concatenated feature representation F^{tamper} from the hidden layers. With the assistance of the modality-consistency detection proxy task, the autoencoders can effectively align the consistent sentiment information across different modalities. In this work, binary cross-entropy loss is employed as the loss function for the consistency detection task.

$$L_{consistency} = -y * log(\hat{y}^{tamper}) - (1 - y) * log(1 - \hat{y}^{tamper}) \tag{9}$$

3.2.4. Optimization objectives

During the pre-training stage, two pre-training tasks, namely sample reconstruction and modality-consistency detection, are designed to enable the autoencoder to learn differentiated and consistent sentiment information across modalities. Consequently, by combining the reconstruction loss and the consistency detection loss, we optimize the autoencoder, and the final self-supervised pre-training loss of the autoencoder is denoted as L_{ae}:

$$L_{ae} = L_{recon} + \lambda_1 * L_{consistency} \tag{10}$$

where λ_1 is a weighting coefficient used to control the relative importance of the sample reconstruction and consistency detection tasks. After the pre-training with two proxy tasks, our autoencoders are capable of effectively capturing differentiated and consistent sentiment information across heterogeneous modalities, which is subsequently utilized for the fine-tuning stage.

3.3. Fine-tuning

During the fine-tuning stage, we construct a multimodal sentiment analysis model based on the concept of MulT's [9] cross-modal attention mechanism and further enhance the model's performance using labeled data. A key distinction from MulT is that our encoder component is derived from the encoder E_m of the pre-trained autoencoders, resulting in AE-MT. Encoder E_m has already learned to capture differentiated and consistent sentiment information across modalities from unlabeled data.

3.3.1. Crossmodal learning based MulT

As shown in Fig. 2, for each modality, the feature extraction process remains consistent with the pre-training stage. Subsequently, the representations F_m from each modality are passed to their respective encoders E_m, resulting in encoded hidden representations H_m. The hidden representations H_m contain rich sentiment information, and we utilize the cross-modal Transformer proposed by MulT to compute interaction representations between each pair of modalities. For modality i and modality j (where $i \neq j$), we compute cross-modal representations $Z_{i\to j}$, with modality i as the query and modality j as the key and value. This is achieved through multiple layers of cross-modal Transformer modules, allowing us to obtain interaction representations between each modality and the other modalities.

$$Z_{i\to j} = CrossmodalTransformer(H_i, H_j, H_j) \tag{11}$$

As a result, 6 different combinations of cross-modal representations are obtained, denoted as $Z_{a\to t}$, $Z_{v\to t}$, $Z_{t\to a}$, $Z_{v\to a}$, $Z_{t\to v}$, and $Z_{a\to v}$. For each modality, we can obtain cross-modal interaction representations that fuse information from the other two modalities through concatenation, resulting in Z_t, Z_a, and Z_v.

$$Z_t = [Z_{a\to t}; Z_{v\to t}] \tag{12}$$

$$Z_a = [Z_{t\to a}; Z_{v\to a}] \tag{13}$$

$$Z_v = [Z_{a\to v}; Z_{t\to v}] \tag{14}$$

Then, each representation is separately fed into a Transformer block to capture contextual information within each modality. Finally, the last layer embeddings for each modality within the network are extracted, concatenated, and then fed into a fully connected

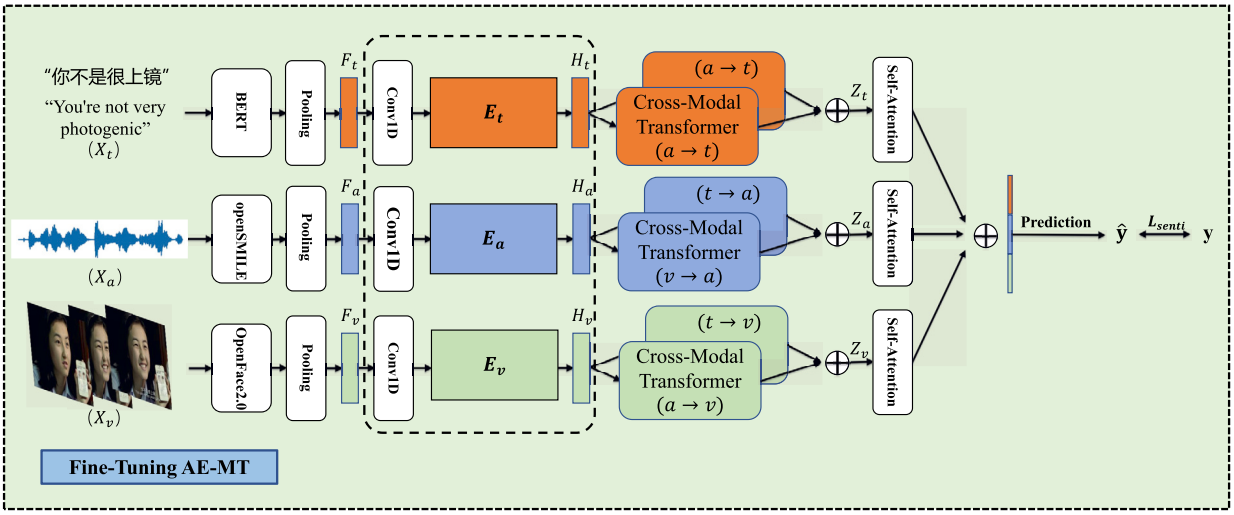

Fig. 2. Pipeline of fine-tuning stage. In this stage, the encoder of MulT is initialled by the pre-trained autoencoders in the pre-training stage resulting in AE-MT. Then, AE-MT utilizes MulT's cross-modal attention mechanism to facilitate the interaction and fusion of the sentiment information of heterogeneous modality, ultimately making sentiment predictions.

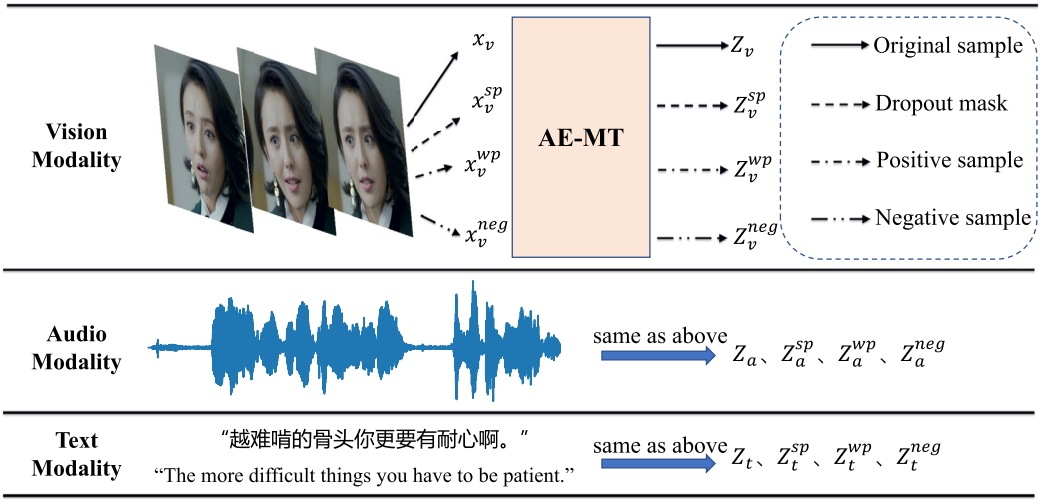

Fig. 3. Pipeline of contrastive learning. For each sample, the strong positive sample, weak positive sample, and negative sample are constructed. Then, AE-MT is used to extract features from samples and optimize the model's ability to distinguish different sentiment polarities by minimizing the distance between the original sample and positive samples and maximizing the distance between the original sample and negative samples.

layer to obtain the final multimodal sentiment intensity prediction, denoted as $\hat{y}$. Therefore, for the multimodal sentiment analysis task, the loss function is denoted as L_{senti}.

$$L_{senti} = \frac{1}{N}\sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2^2 \tag{15}$$

where N denotes the number of samples, y_i and $\hat{y}_i$ represent the ground truth sentiment score and the model's predicted sentiment score, respectively.

3.3.2. *Contrastive learning auxiliary task*

To enhance the model's perception of different sentiment intensities and improve its ability to address sentiment diffusion in high-level semantic spaces during the fine-tuning stage, a contrastive learning auxiliary task is introduced. We construct strong positive samples, weak positive samples, and negative samples based on the sentiment polarity of samples from the same batch. As shown in Fig. 3, for a given sample $X_i = \{X_t, X_v, X_a\}$, we create strong positive samples X_i^{sp} by applying a dropout mask to the original sample, where a portion of the original sample's information is set to 0 with a certain probability.

$$X_i^{sp} = \text{dropout}(X_i) \tag{16}$$

Additionally, samples with the same unimodal labels within the same batch are considered weak positive samples X_i^{wp}, while samples with opposite sentiment polarities are considered negative samples X_i^{neg}. The objective of the contrastive learning auxiliary task is to minimize the distance between sample X_i and positive samples X_i^{sp} and X_i^{wp} while maximizing the distance between sample X_i and negative samples X_i^{neg}. To measure the distance between samples, cosine similarity is used.

$$s(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2} \tag{17}$$

The multimodal representations obtained by passing the above-mentioned samples to the model are denoted as Z_i, Z_i^{sp}, Z_i^{wp}, and Z_i^{neg}, respectively. In this paper, the contrastive learning task loss function is denoted as L_{con},

$$L_{con} = \sum_{i=1}^{N} -\log \frac{exp(s(Z_i, Z_i^{sp})/\tau) + \sum_{k=1}^{K} exp(s(Z_i, Z_k^{wp})/\tau)}{\sum_{j=1}^{J} exp(s(Z_i, Z_j^{neg})/\tau)} \tag{18}$$

where K represents the number of weak positive samples, J denotes the number of negative samples, and τ is the temperature parameter used for smoothing probability distributions. During the training process, we optimize the model parameters by minimizing L_{con} to learn robust and discriminative representations, thereby enhancing the model's perception of different levels of sentiment intensity.

3.3.3. Optimization objectives

During the fine-tuning stage, the multimodal sentiment analysis main task and the contrastive learning auxiliary task are combined to obtain the final optimization objective L_{ft},

$$L_{ft} = L_{senti} + \lambda_2 * L_{con} \tag{19}$$

where λ_2 is used to balance the relative importance of the main task and the auxiliary task.

3.3.4. Complexity analysis

Our approach contains two stages: the pre-training stage and the fine-tuning stage.

In the pre-training stage, the model mainly contains Conv1D layer, Transformer Encoder layer and Transformer Decoder layer. Since the time complexity of a Transformer block is $O(n^2d + nd^2)$, the time complexity of Transformer Encoder layer and Transformer Decoder layer are $O(h_1(n^2d + nd^2))$ and $O(h_2(n^2d + nd^2))$ respectively, where n is the sequence length, d is the representation dimension, h_1 and h_2 are the number of block of Transformer Encoder layer and Transformer Decoder layer respectively. The time complexity of Conv1D is $O(knd^2)$, where k is the kernel size of convolution. Therefore, the time complexity of pre-training stage is $O(n_1(knd^2 + h_1(n^2d + nd^2) + h_2(n^2d + nd^2))$, where n_1 is the number of training data.

In the fine-tuning stage, the model mainly contains Conv1D layer, Transformer Encoder layer, Cross Modal Transformer layer and self-attention layer. Since the time complexity of self-attention layer is $O(n^2d)$. Therefore, the time complexity of fine-tuning stage is $O(n_2(knd^2 + h_1(n^2d + nd^2) + h_3(n^2d + nd^2) + n^2d))$, where h_3 is the number of block of Cross Modal Transformer layer, where n_2 is the number of training data.

Finally, the total time complexity of our model is $O(n_1(knd^2 + h_1(n^2d + nd^2) + h_2(n^2d + nd^2)) + n_2(knd^2 + h_1(n^2d + nd^2) + h_3(n^2d + nd^2) + n^2d))$.

4. Experiments

In this section, we first introduce the dataset that was used to evaluate our framework and empirically compare our AE-MT with previous strong baselines.

4.1. Datasets

In this work, experiments are conducted based on the CH-SIMS (Chinese) v2.0 [11] dataset and MOSEI (English) [8] dataset.

The CH-SIMS v2.0 dataset is a distinctive Chinese multimodal sentiment analysis benchmark with fine-grained annotations of each modality. It includes both labeled and unlabeled data. In the pre-training stage, the unlabeled data is divided into a training set and a testing set, totaling 10,161 samples. In the fine-tuning stage, the model is optimized by using the labeled data for downstream tasks, and the dataset is further divided into three subsets: training (2,722 samples), validation (647 samples), and testing (1,034 samples) sets, consisting of a total of 4,403 samples.

The MOSEI dataset contains 22,856 annotated video segments over 250 different topics. Each clip has a sentiment score between -3 (strongly negative) to +3 (strongly positive). In our experiment, it is divided into three subsets: training (16,326 samples), validation (1,871 samples), and testing (4,659 samples) sets. Since The MOSEI dataset only has multi-modal sentiment annotation, and there is no single modal sentiment annotation, it is hard to train our AE-MT model in the pre-training stage. To solve that problem, we assume that all single modal annotation is the same as its multi-modal annotation.

Table 2
The detail of experiment parameters.

Parameters	Value
activation function	ReLU
number of epochs in the pre-training stage	500
number of epochs in the fine-tuning stage	50
optimizer	Adam
learning ratio scheduler	ReduceLROnPlateau
number of Transformer Encoder layer	3
number of Transformer Decoder layer	3
number of Cross Modal Transformer layer	4
λ_1	0.4
λ_2	1.5

Those two datasets are collected from real-world. Although there are some simplifications in the datasets, such are simple background image, clean speech and simple sentence, they reflect some of the current simple multimodal sentiment analysis scenarios. For example, short video platform.

4.2. Basic settings

4.2.1. Standardize sequence lengths

The audio and vision modality features from the CH-SIMS v2.0 dataset are preprocessed. For audio and vision temporal data, their sequence lengths represent the continuity of information along the temporal axis. In CH-SIMS v2.0, the lengths of the audio and vision modalities are significantly longer at 925 and 232, respectively, compared to the text modality with a length of 50. This discrepancy results in unbalanced and unaligned sequence lengths. To meet the computational resource requirements of our work, average pooling is applied to the three modalities' data, transforming them into a unified length. This process helps reduce redundant information in the multimodal data and minimizes the model size and parameter count for our proposed method.

4.2.2. Experimental details

During training, Adam is used as the optimizer and implement learning rate decay and early stopping to control the model's convergence. The proposed method is implemented using the PyTorch framework and experiments are conducted on one NVIDIA 2080Ti GPU which contains 11 GB video memory. The total amount of video memory used is 2.5 GB. The details of our experiments are shown in Table 2. In the pre-training stage, the runtime of our approach is 10 seconds per epoch. In the fine-tuning stage, the runtime of our approach is 33 seconds per epoch. For a fair comparison, we run five times and report the average performance for our AE-MT and all baselines.

4.2.3. Evaluation metrics

Following the previous works [6], our experimental results are reported in two forms: classification and regression. For classification, the experiment shows the Weighted F1 score (F1-2) and classification accuracy at different numbers of classes (Acc-2, Acc-3, and Acc-5). For regression, the experiment shows the Mean Absolute Error (MAE) and Pearson correlation (Corr). Except for MAE, higher values represent better performance for all metrics.

4.3. Baselines

Comprehensive comparative study is performed against the proposed **AE-MT** by considering various baselines as detailed below.

MulT [9] (2019) is a classic model in the field of multimodal sentiment analysis and was the first to introduce Transformer models to this task. By adopting the powerful Transformer model, MulT learns cross-modal attention mechanisms between three modalities, enhancing interactions between different modalities and acquiring richer multimodal feature representations. As a result, it achieved outstanding performance in multimodal sentiment analysis.

BERT_MAG [6] (2020) (multimodal adaptation gate for BERT) is the application of multimodal adaptation gates at different layers of the BERT backbone in RAVEN to handle aligned data.

MISA [5] (2020) is an approach that addresses the characteristics of multimodal data, where there are similarities between modalities but also differences within each modality. It employs an encoder to map the original features into both a modality-invariant feature space and a modality-specific feature space. This enables the learning of modality representations in their respective subspaces.

MMIM [21] (2021) is the first time introduces the maximization of mutual information into the multimodal sentiment analysis task. It hierarchically maximizes mutual information to preserve task-relevant information through multimodal fusion.

Self-MM [20] (2021) proposes a method that generates pseudo-labels for each modality based on the self-supervised information between multimodal labels and samples. It then utilizes these pseudo unimodal labels for multi-task learning.

AV-MC [11] (2022) and CH-SIMS v2.0 are both from the same work and represent the first attempt on the dataset as a strong baseline. AV-MC introduces an audio-visual consistency framework, enabling the model to learn from different non-textual contexts for sentiment prediction, achieving impressive results.

Table 3
Results on CH-SIMS v2.0 dataset. All baseline models were reproduced and evaluated under the same experimental environment. And the results with ∗ denote the performance reported in the original papers.

Methods	Acc-2	Acc-3	Acc-5	F1-2	MAE	Corr
MulT (2019)	81.24	74.18	57.45	81.3	27.78	74.53
BERT_MAG (2020)	75.16	68.82	47.72	75.23	36.06	60.61
MLMF (2020)	72.28	60.87	43.13	71.87	39.64	51.85
MLF-DNN (2020)	72.01	62.38	42.13	71.73	40.29	48.55
MTFN (2020)	72.40	62.32	43.54	71.97	39.97	48.56
MISA (2020)	77.58	69.56	47.12	77.66	34.41	66.73
MMIM (2021)	75.15	68.96	48.30	75.22	35.78	60.53
SELF-MM (2021)	79.98	72.44	52.22	79.97	35.04	64.93
AV-MC (2022)*	81.72	74.08	53.09	81.78	29.89	73.06
AV-MC (semi) (2022)*	83.46	-	-	83.52	28.60	**76.04**
ConFEDE (2023)	82.23	70.15	46.30	82.08	39.20	63.70
AE-MT (ours)	**84.33**	**77.85**	**58.03**	**84.52**	**27.69**	75.78

Table 4
Results on MOSEI dataset (The Acc-2 and F1-2 correspond to "negative/non-negative" sentiment classification).

Methods	Acc-2	F1-2	MAE	Corr
MulT (2019)	81.15	81.56	55.90	73.30
BERT_MAG (2020)	79.86	80.47	58.30	74.10
MISA (2020)	82.59	82.67	56.80	72.40
MMIM (2021)	82.68	82.95	52.60	77.20
SELF-MM (2021)	82.81	82.53	53.00	76.50
ConFEDE (2023)	81.65	82.17	**52.20**	**78.00**
AE-MT (ours)	**83.17**	**83.29**	55.10	74.30

AV-MC [11] (Semi-supervised) builds upon AV-MC and incorporates the unsupervised data from CH-SIMS v2.0. It introduces a modality-mixing module as an enhancement, which combines audio and visual modalities from different videos. By integrating unobserved multimodal context with text, the model can learn to perceive various non-linguistic contexts for sentiment prediction.

ConFEDE [28] (2023) constructs a unified learning framework that jointly performs contrastive representation learning and contrastive feature decomposition to enhance representation of multimodal information.

4.4. Quantitative results

The experiment results are shown in Table 3 and 4. According to our results, AE-MT is generalized and scalable on language.

Table 3 shows the comparative results on the CH-SIMS v2.0 dataset. Our AE-MT shows significant performance improvements compared to all the baseline models without pre-training across various evaluation metrics. This demonstrates that pre-training with self-supervised learning using unlabeled data provides a better initialization for the model, significantly improving its performance on downstream tasks. Despite AV-MC (semi) pre-trained by unlabeled data, AE-MT still outperforms it on all metrics except Corr. This indicates that our pre-training method enables our autoencoders to effectively extract both differential and consistent sentiment information between modalities, providing greater assistance in multimodal sentiment analysis.

Table 4 shows the comparative results on the MOSEI dataset. It shows that AE-MT outperforms all baseline models in Acc-2 and F1-2 evaluation metrics. However, AE-MT is not the best in MAE and Corr metrics. That is because the MOSEI dataset only has multi-modal sentiment annotation, and there is no single modal sentiment annotation. To perform the pre-training stage, we assume that all single modal annotation is the same as its multi-modal annotation. It is not true for all sample and leads to wrong annotations. Most of models achieve better performance on MOSEI than on CH-SIMS. That lies in two reasons. First, Chinese is more complicated than English. Second, the MOSEI dataset is bigger than CH-SIMS.

4.5. Ablation study

To further study the influence of the individual pre-training tasks in AE-MT, we perform a comprehensive ablation analysis. Additionally, to validate the effectiveness of the Transformer architecture [27] in the autoencoder, the autoencoder structure is replaced with an LSTM architecture [29] to capture temporal context information in the data.

From Table 5, it shows that as contrastive learning and modality consistency detection tasks, and the Transformer architecture inside the autoencoder are removed, the model's performance significantly decreases. This indicates that the contrastive learning auxiliary task helps improve the model's ability to capture higher-level sentiment semantics in deeper network layers. Additionally, the modality consistency detection task allows the model to learn superior feature spaces from a large amount of unlabeled data, effectively capturing consistent sentiment information between multimodal data. Furthermore, the Transformer architecture based on self-attention mechanisms is more effective in handling long-range dependencies in time series, thus improving model performance.

Table 5
Ablation study. We conducted a structural analysis of the AE-MT model, where w/o CL represents the removal of the contrastive learning auxiliary task, w/o MCD represents the removal of the modality consistency detection task, and w/o TR represents the substitution of the Transformer architecture with the LSTM architecture.

Methods	Acc-2	Acc-3	Acc-5	F1-2	MAE	Corr
AE-MT w/o CL	83.17	77.08	57.54	84.02	27.84	75.39
AE-MT w/o MCD	82.69	76.11	55.71	83.24	28.88	74.92
AE-MT w/o TR	80.17	73.4	52.71	80.29	30.18	72.83
AE-MT	**84.33**	**77.85**	**58.03**	**84.52**	**27.3**	**75.98**

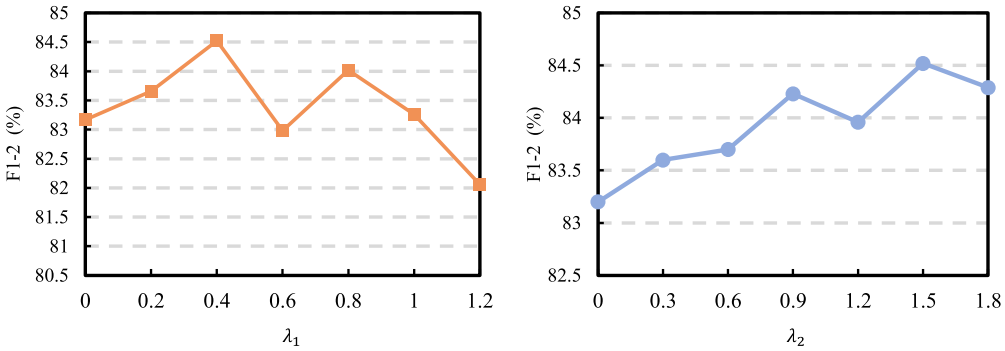

Fig. 4. The sensitivity analysis of hyperparameters. The left plot shows the sensitivity analysis of hyperparameter λ_1, which controls the strength of the modality consistency detection auxiliary task. The right plot illustrates the sensitivity analysis of hyperparameter λ_2, which controls the strength of the sentiment-guided contrastive learning auxiliary task.

4.6. Parameter sensitivity experiments

This section aims to investigate the mechanisms of the modality consistency detection task in the pre-training stage and the contrastive learning auxiliary task in the fine-tuning stage. Through a sensitivity analysis of hyperparameters λ_1 and λ_2, this study adjusts their values to control the strength of the two auxiliary tasks and analyze the generalization performance of the trained model under different task strengths.

As shown in Fig. 4, during the pre-training stage of the autoencoders, introducing a certain degree of modality consistency detection task helps the model extract consistent sentiment information and enhances the model's expressive capability. However, there is no specific golden standard for the exact value of λ_1. The change of F1 score with λ_1 has three stages. First, when λ_1 is less than 0.4, the F1 score of AE-MT continues to increase. Second, when λ_1 is great than 0.4 and less than 0.8, the F1 score of AE-MT fluctuates. Third, when λ_1 is great than 0.8, the F1 score of AE-MT continues to decrease. According to these three stages, higher λ_1 will decrease the performance of AE-MT. It indicates that the reconstruction loss and consistency detection loss is complementary. Furthermore, since the F1 score obtained when λ_1 is 0 is higher than when λ_1 is 1.2, it shows that the intra-modal reconstruction loss plays more important role than the consistency detection loss.

During the fine-tuning stage of the multimodal sentiment analysis task, a larger degree of influence from the contrastive learning task helps the model better capture information about different sentiment polarities when obtaining the final multimodal representation. Although there are some fluctuations in the value of F1, the general trend is to increase as λ_2 increases. It indicates that the contrastive learning loss is important for improving the performance of sentiment classification. This further confirms the importance of the contrastive learning auxiliary task in helping the model delve deeper into sentiment semantics, thus improving the accuracy of sentiment analysis.

In conclusion, the appropriate degree of influence from the modality consistency detection auxiliary task and contrastive learning auxiliary task is crucial for improving the model's generalization performance on multimodal sentiment analysis tasks.

4.7. Case study

In this section, we present the performance of our proposed AE-MT method and baseline models on specific samples from the CH-SIMS v2.0 test data. For each sample, the top row shows the predicted sentiment scores and multimodal sentiment labels for each model, representing the overall sentiment score of the sample. Values closer to -1 indicate a more negative sentiment, while values closer to +1 indicate a more positive sentiment. The subsequent rows represent the specific content of the visual, audio, and text modalities, along with their respective individual labels.

In the first case of Fig. 5, the video clip represents a case with a negative sentiment score. In the visual modality, the characters display obvious negative facial expressions including frowning and mouth movements. In the audio modality, the speech is intense and the voice sounds agitated, conveying a strong sense of anger. These two modalities express extremely negative sentiments. While

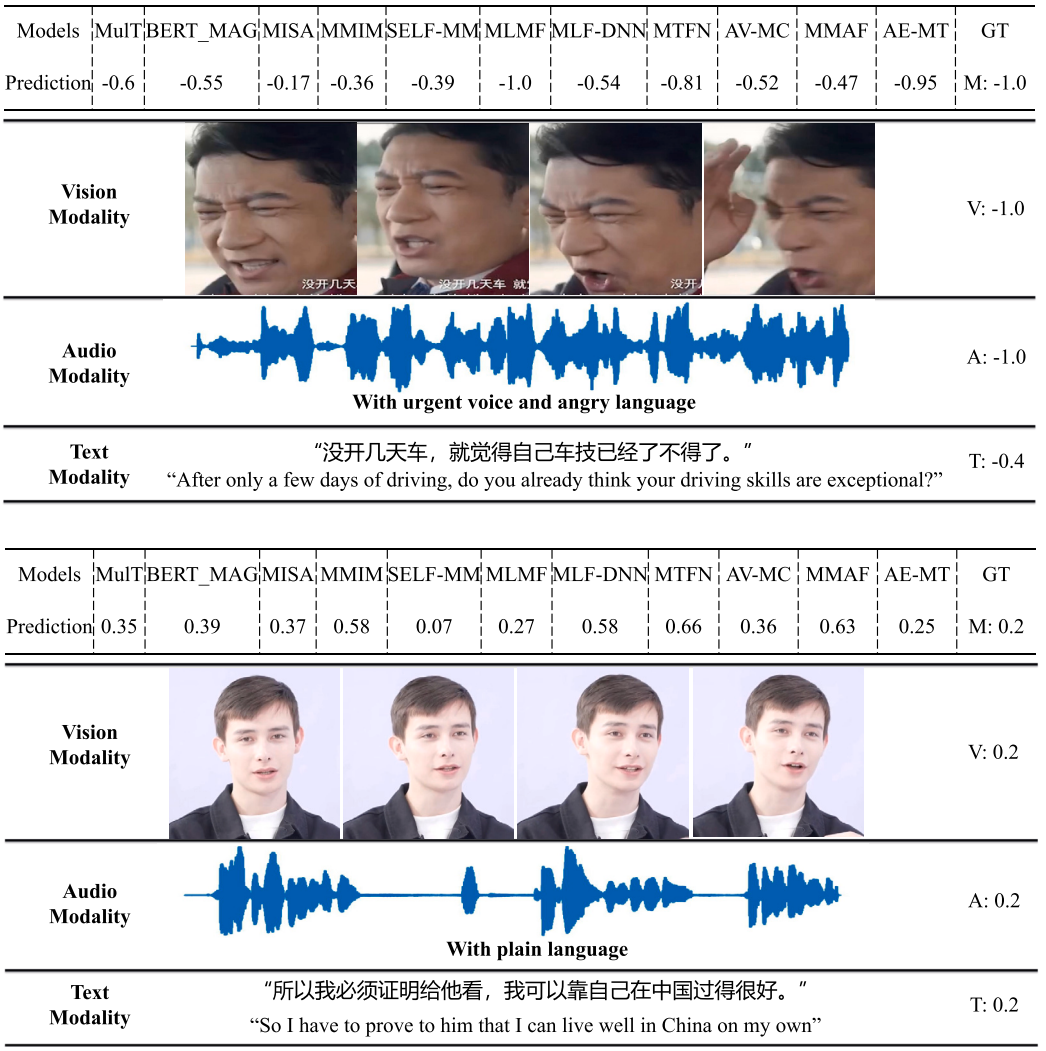

Fig. 5. Case study on CH-SIMS v2.0 dataset. The first case contains negative sentiment information, while the second case contains positive sentiment information.

the text modality contains a subtle sarcastic tone, indicating a relatively weak negative sentiment. All three modalities show strong consistent negative sentiment information. Although all the compared models in the figure predict that the sample has negative sentiment, only MLMF, MTFN, and the proposed AE-MT are more accurate in predicting the intensity of its negative sentiment, tending to have scores closer to -1.

In the second case of Fig. 5, the video clip represents a case with weak positive sentiment. In the visual modality, the speaker has a slight smile, and the voice tone is relatively calm and neutral. The text modality also contains content with a weak positive sentiment. All three modalities show consistent weak positive sentiment information. For this type of sample with weak positive sentiment, AE-MT can also accurately predict its sentiment intensity.

5. Conclusion

In this paper, we propose an autoencoder-based self-supervised learning framework, which aims to improve the model's ability to extract modality-specific differentiated and modality-consistent sentiment information. In the pre-training phase, we design an autoencoder for each modality and use unlabeled data to learn richer sentiment representations for each modality through sample reconstruction and modality-consistency detection tasks. In the fine-tuning phase, we inject the pre-trained autoencoders into MulT, resulting in AE-MT, and enhance the model's ability to extract deep sentiment information by combining contrastive learning auxiliary tasks. AE-MT efficiently integrates differential and consistent sentiment information, enhancing the model's capability from a multimodal representation perspective. Additionally, it utilizes MulT's cross-modal attention mechanism to facilitate interaction and fusion of information from various modalities, leading to accurate sentiment predictions.

Although AE-MT shows some advantages compared with the state-of-the-art approaches, there is a problem when data is missing for a certain modal. That is because we assume that data exists for all modalities. In our future work, we will construct an adaptive learning strategy which allows the model to better handle missing modality problem and achieve effective fusion of multimodal information.

CRediT authorship contribution statement

Wenjun Feng: Writing – original draft, Validation, Methodology. **Xin Wang:** Software, Methodology, Investigation. **Donglin Cao:** Writing – review & editing, Methodology. **Dazhen Lin:** Writing – review & editing, Methodology, Funding acquisition, Data curation, 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.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62076210) and the Natural Science Foundation of Xiamen, China (No. 3502Z20227188).

References

[1] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, M. Pantic, A survey of multimodal sentiment analysis, Image Vis. Comput. 65 (2017) 3–14.
[2] R. Kaur, S. Kautish, Multimodal sentiment analysis: a survey and comparison, in: Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines, 2022, pp. 1846–1870.
[3] R. Creemers, Cyber China: upgrading propaganda, public opinion work and social management for the twenty-first century, J. Contemp. China 26 (103) (2017) 85–100.
[4] R. Das, T.D. Singh, Multimodal sentiment analysis: a survey of methods, trends and challenges, ACM Comput. Surv. (2023).
[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. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2020, NIH Public Access, 2020, p. 2359.
[7] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, arXiv preprint, arXiv:1606.06259, 2016.
[8] 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.
[9] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2019, NIH Public Access, 2019, p. 6558.
[10] P.H. Le-Khac, G. Healy, A.F. Smeaton, Contrastive representation learning: a framework and review, IEEE Access 8 (2020) 193907–193934.
[11] Y. Liu, Z. Yuan, H. Mao, Z. Liang, W. Yang, Y. Qiu, T. Cheng, X. Li, H. Xu, K. Gao, 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, 2022, pp. 247–258.
[12] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, arXiv preprint, arXiv:1707.07250, 2017.
[13] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, arXiv preprint, arXiv:1806.00064, 2018.
[14] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.
[15] L.H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: a simple and performant baseline for vision and language, arXiv preprint, arXiv:1908.03557, 2019.
[16] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, Videobert: a joint model for video and language representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7464–7473.
[17] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, A. Sacheti, Imagebert: cross-modal pre-training with large-scale weak-supervised image-text data, arXiv preprint, arXiv:2001.07966, 2020.
[18] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. 32 (2019).
[19] 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.
[20] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 10790–10797.
[21] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, arXiv preprint, arXiv:2109.00412, 2021.
[22] S. Verma, C. Wang, L. Zhu, W. Liu, Deepcu: integrating both common and unique latent information for multimodal sentiment analysis, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 3627–3634.
[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, 2018.

[24] S. Beliaev, Y. Rebryk, B. Ginsburg, Talknet: fully-convolutional non-autoregressive speech synthesis model, arXiv preprint, arXiv:2005.05514, 2020.
[25] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.-P. Morency, Openface 2.0: facial behavior analysis toolkit, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 59–66.
[26] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the Munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, 2010, pp. 1459–1462.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[28] J. Yang, Y. Yu, D. Niu, W. Guo, Y. Xu, ConFEDE: contrastive feature decomposition for multimodal sentiment analysis, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 7617–7630, https://aclanthology.org/2023.acl-long.421.
[29] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, Neural Comput. 31 (7) (2019) 1235–1270.