Journal of Network Intelligence ©2026 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **11**, Number **1**, February **2026**

Multimodal Emotion Analysis Based BERT and Self-attention Feature Fusion of Improved Inception

Qingfang Li[1]

[1]Institute of Higher Education,
Chengdu Technological University, Chengdu 611730, China
summersanmer@163.com

Jun Li[2]

[2]Teaching Quality Monitoring and Evaluation Center/Teacher Development Center,
Chengdu Technological University, Chengdu 611730, China
ljun1@cdtu.edu.cn

Xu Li[3,*]

[3]School of Intelligent Manufacturing,
Chengdu Technological University, Chengdu 611730, China
faresless000@126.com

Mingrong Li[4]

[4]School of Architecture,
Southwest Minzu University, Chengdu 610225, China
21900039@swun.edu.cn

Ruihui Wu[5]

[5]Faculty of Entrepreneurship and Business,
Universiti Malaysia Kelantan, Kuantan 16100, Malaysia
a20e0276f@siswa.umk.edu.my

*Corresponding author: Xu Li

Received September 21, 2024, revised January 22, 2025, accepted June 20, 2025.

Abstract. *Multimodal sentiment analysis judges users' emotions through videos and some text information uploaded by users on social platforms. Current research on multimodal sentiment analysis mainly designs complex multimodal fusion networks to learn the consistency information between modalities, while ignoring the complementary role of differential information. In addition, the learning rate of the modalities is also unbalanced, which may lead to the fact that when one modality converges, the other modalities do not converge, resulting in poor collaborative decision-making. To this end, a multimodal sentiment analysis method based on BERT and improved Inception collaborative feature representation is proposed, in which short-time Fourier transform (STFT) is used to extract audio information, BERT is used to extract text information, and improved Inception is used to extract image information. Feature alignment is performed in three different modalities of text, image and language, thereby promoting deep fusion of cross-modal information, and the self-attention mechanism is used to increase the weight of key features. Experiments are conducted on two public datasets, CMU-MOSEI and CMU-MOSI. The experimental results show that the proposed method is superior to other excellent methods in terms of binary classification accuracy and 7-classification accuracy.*

Keywords: Multimodal sentiment analysis, Learning rate, BERT, Inception, Fusion networks, Self-attention mechanism

1. **Introduction.**

With the rapid development of social media, online comments have grown at an explosive rate. How to mine valuable information from massive comment data has become a focus of discussion in the industry and academia. The rapid development of artificial intelligence is changing thinking and reshaping the industry. The new generation of AI language models such as "GPT-4" has promoted a new wave of natural language processing and computer vision [1]. Sentiment analysis, as one of the tasks in the field of natural language processing, aims to identify and extract emotional tendencies, attitudes and emotions in text. According to the different granularity of information, sentiment analysis is divided into coarse-grained sentiment analysis and fine-grained sentiment analysis. Among them, fine-grained sentiment analysis, also known as aspect-level sentiment analysis, mainly includes aspect word extraction and aspect-level sentiment classification [2, 3], which identifies the emotional tendencies of commentators towards the commented object from different aspects.

However, information technology is constantly updated and iterated, and users can publish comments containing other modal information such as text, images and even videos. Users prefer to use different modal data to express their personal multi-faceted emotional tendencies more specifically, or publish comment data that is opposite to the text and image information to express strong appreciation or irony. Therefore, it is an inevitable trend for aspect-level sentiment analysis to shift from text modality to multimodal information fusion.

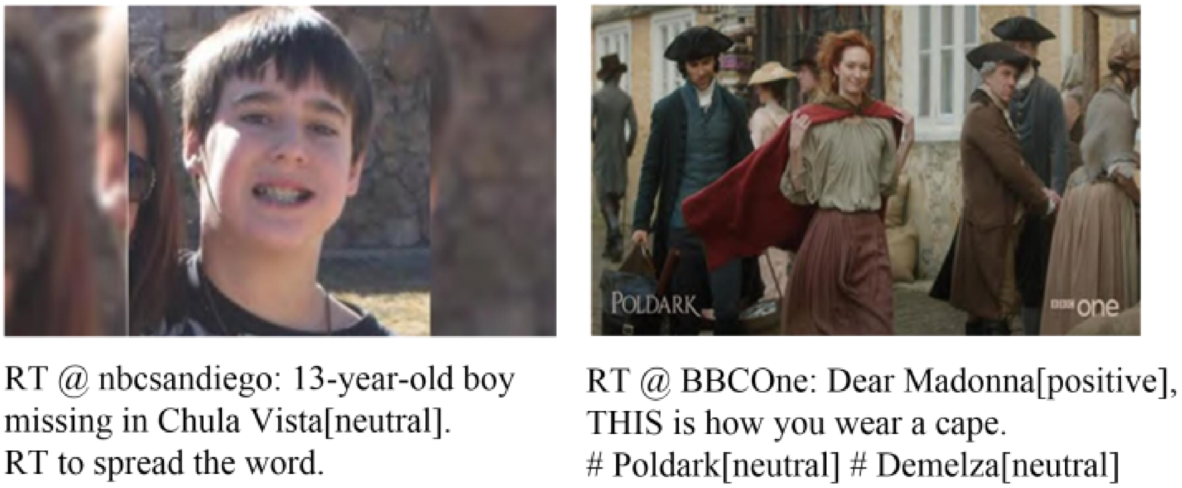
Multimodal sentiment analysis aims to comprehensively analyze and study multiple modalities (such as images, videos, text, sounds, etc.) to achieve more accurate analysis of users' emotions in different aspects. As shown in Figure 1, two data samples fully illustrate the richness of multi-source data and the importance of aspect-level sentiment analysis tasks. First, as shown in the left figure, image information alone may mislead the model to predict that the comment conveys positive emotions, but further combined with the comment text, it is found that the description is a child lost. The model needs to integrate the image and text information to accurately predict that the comment text expresses negative emotions for the entity "Chula Vista". Secondly, from the comment text on the right, it can be seen that the comment contains three entities "Madonna", "Poldark" and "Demelza". The emotional polarity of the commentator cannot be judged based on the text information. Further combined with the image information, it can be found that the entity "Madonna" conveys positive emotions, while other entities show neutral emotions. It can be seen that there is interrelated and complementary semantic information between text sentiment words and local image regions in multimodal data. Only by fully extracting effective features can the performance of sentiment analysis be further improved.

In order to more comprehensively mine text features and image features and improve the effect of sentiment analysis, this paper proposes a multimodal sentiment analysis method based on BERT and improved Inception to achieve feature alignment in different modalities, thereby promoting deep fusion of cross-modal information and improving the accuracy of multimodal aspect-level sentiment analysis.

2. **Related works.**

2.1. **Multimodal feature extraction.**

In the study of multimodal sentiment analysis, feature extraction of different modalities has attracted attention [4]. ELMo, BERT and GPT series [5, 6, 7, 8] have been proposed in the field of natural language processing, and



RT @ nbcsandiego: 13-year-old boy missing in Chula Vista[neutral]. RT to spread the word.

RT @ BBCOne: Dear Madonna[positive], THIS is how you wear a cape. # Poldark[neutral] # Demelza[neutral]

Figure 1. Data examples.

models such as ResNet series, VGG series and ViT have been updated and iterated in the field of image processing.

In the process of text feature extraction, Song et al. [9] designed an attention encoder network (AEN) as a downstream structure to associate aspect words and text features encoded by the BERT model. After the BERT model encodes each word representation, the “CLS” position vector integrates the overall text information and is directly used by the BERT-SPC model for aspect-level sentiment analysis, achieving good results with a simple structure.

Similarly, Gao et al. [10] pooled the aspect word features encoded by BERT and used the pooled features for sentiment analysis to further improve the model effect.

The BERT model [5] has many advantages in feature extraction tasks, such as ease of use and strong stability. However, some scholars [11] pointed out that there is anisotropy in the text encoding process of the BERT model, and contrastive learning can reduce the deviation of the real semantic understanding, avoid model collapse, and improve the model’s feature representation ability.

In the process of image feature extraction, early research focused on the contour and color of the image, but the rapid development of deep learning has promoted revolutionary changes in the field of computer vision. The use of convolutional neural networks to extract deep features of images has become very popular. For example, ESAFN [12], ModalNet-Bert [13] and other models all use ResNet networks as image feature extraction tools. The difference lies in the different ways in which the downstream structures used by each model fuse multi-source features.

However, with the widespread application and success of Transformer in the field of natural language processing, people have gradually applied it to computer vision tasks. The emergence of ViT [14] breaks the spatial local limitations of traditional convolutional neural networks. By introducing the self-attention mechanism, it realizes the modeling of local features of images and captures richer image information, thereby showing excellent performance in multiple tasks.

2.2. **Multimodal sentiment analysis.** Sentiment analysis is widely defined as the computational study of subjective factors, including people’s opinions, attitudes and emotions [15]. Multimodal sentiment analysis is a sentiment analysis method that uses multiple forms of subjective expressions to analyze and judge sentiment. In theory, existing multimodal sentiment analysis models receive more diverse sentiment information and should be better than the analysis results of unimodal models. However, recent studies have pointed out [16] that the analysis results of unimodal models are better than multimodal models in some scenarios, and we have found that this phenomenon also occurs in multimodal sentiment analysis models.

The reason for the above phenomenon is that different modalities are fitted and generalized at different rates. Using a single optimization strategy to train all modalities will result in a situation where one modality is already fitted while the other modalities are still underfitted, resulting in the inability of the unimodal learning network to be fully learned, which in turn affects the fusion effect [16, 17, 18]. In summary, current multimodal sentiment analysis methods mainly focus on multimodal fusion and multimodal representation.

In terms of multimodal fusion, Zadeh et al. [19] proposed a multimodal tensor fusion method in 2017 to fuse the three modalities by outer product. The fusion results were used for final sentiment analysis through a fully connected deep network model. This method can not only preserve the internal information of a single modality, but also learn the complementary information between modalities. Tsai et al. [20] proposed a multimodal converter structure that uses cross-modal attention to cross-fuse the three modalities of text, audio, and image, so that a single modality can obtain information from other modalities.

In terms of multimodal representation, Li et al. proposed a method for representing modal commonalities and characteristics [21]. The loss function is used to learn the commonalities and characteristics between single-modal representations, and the commonalities and characteristics between these modalities are fused to reduce the information redundancy between modalities. Rahman et al. [22] proposed a multimodal adaptation gate structure for fine-tuning multimodal models. Yu et al. [23] designed a label generation module with a self-supervised learning strategy, and learned the consistency representation and difference representation between modalities in a multi-task learning manner. Wu et al. [24] used the shared semantics and private semantics of video and audio modalities to enhance and complement the text modality, and proposed a text-centric shared private framework for sentiment analysis. The CLIP model proposed by Radford et al. [25] uses a large-scale dataset to learn the relationship between images and texts, achieve image-text matching, and perform image classification tasks. Based on this, Yang et al. [26] simultaneously used the information of images, texts, and labels to complement each other and construct a unified contrastive learning framework.

The above multimodal sentiment analysis models all use complex fusion methods to form multimodal feature representations for decision-making, without considering the impact of imbalanced learning rates between modalities on sentiment analysis tasks. In addition, there are differences in the intensity of sentiment semantic expression of different modalities, and thus their contribution to the task is also different.

3. **Proposed method.** The main structure of the proposed method is shown in Figure 2. First, there is the encoding stage: the BERT model is used to extract text features, and the improved 2D-Inception model is used to extract image features; then there is the multi-source feature alignment stage: by constructing positive and negative samples on the text side, image side, and audio side, similar feature representations are mapped to the same space to achieve feature alignment in different modalities; finally, there is the joint learning stage: multi-modal aspect-level sentiment analysis and contrastive learning are collaboratively trained and fine-tuned to improve the sentiment classification effect.

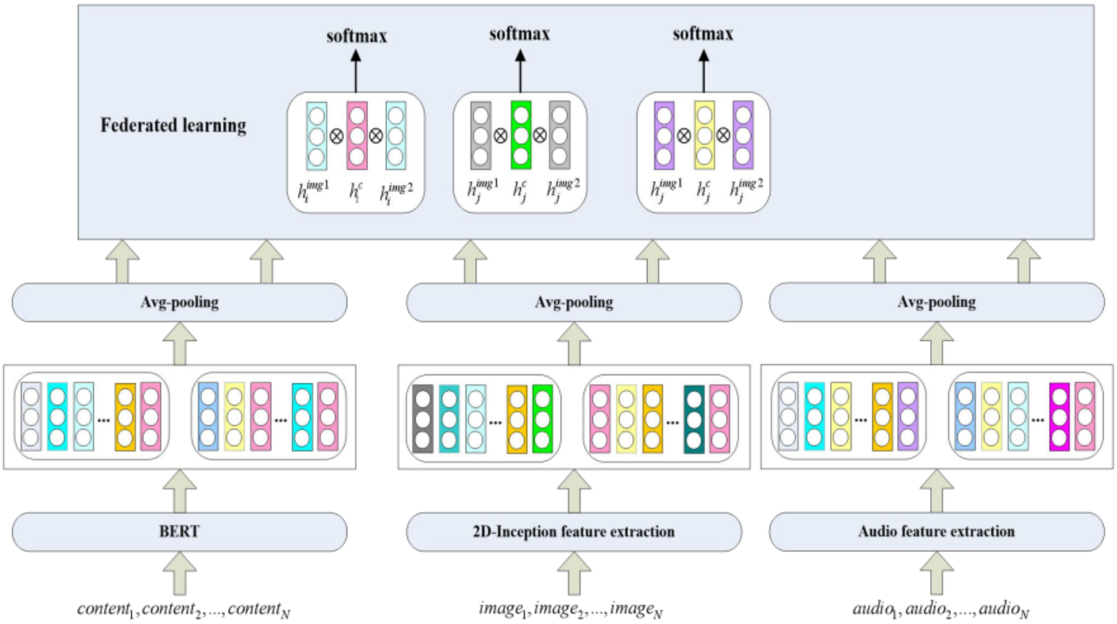

softmax
softmax
softmax
Federated learning
h_i^{img1} h_i^c h_i^{img2}
h_j^{img1} h_j^c h_j^{img2}
h_j^{img1} h_j^c h_j^{img2}
Avg-pooling
Avg-pooling
Avg-pooling
BERT
2D-Inception feature extraction
Audio feature extraction
$content_1, content_2, ..., content_N$
$image_1, image_2, ..., image_N$
$audio_1, audio_2, ..., audio_N$

Figure 2. The structure of the proposed method.

3.1. **Text encoding representation module.** Traditional "Word2Vec" and "GloVe" use the co-occurrence of target words and context words to generate text representations, so the generated tags only provide a representation independent of the context. The order of words in the sentence is not considered, and the word embeddings trained by "Word2Vec" and "GloVe" are static, which will lead to the problem of polysemy. Although the introduction of language model embedding ELMo (embedding from language models) has solved the problem of polysemy to a certain extent, ELMo only captures context from two directions (i.e., bidirectional RNN). BERT [5] uses a Transformer composed of an attention network as an encoder, which can capture context from all possible directions (fully connected). In contrast, this paper selects BERT as the embedding layer, models the input sentence, and inputs the context representation generated by BERT into the self-attention network layer for feature extraction to generate the context representation.

BERT has two input methods: single sentence input $< [CLS]segment[SEP] >$ and sentence pair input $< [CLS]segment_1[SEP]segment_2[SEP] >$, where $[CLS]$ indicates the beginning of the data and $[SEP]$ indicates the end of the segment or data.

In order to construct the positive and negative samples required for text-side contrastive learning, this paper uses the similarity between aspect words and the encoding of aspect words in sentences, uses the $[SEP]$ delimiter to separate the review text and aspect words, and concatenates the review text fragment 1 containing aspect words and the aspect word fragment 2 into the BERT model. In a batch of data, the comment text in the i-th data is $s_i = w_1^i, w_2^i, \ldots, \alpha_{\lambda+1}^i, \ldots, \alpha_{\lambda+n}^i, \ldots, w_m^i$, the aspect word is $\alpha_i = \alpha_{\lambda+1}^i, \ldots, \alpha_{\lambda+n}^i$, and the i-th input text is obtained by concatenating s_i and α_i. Then the N input texts in this batch of data are $content = content_1, content_2, \ldots, content_N$.

After the data is input into the BERT model, the encoding operation is shown in Formula (1), that is, the features of each word H^{c+a} in the output text are:

$$H^{c+a} = BERT(content) \tag{1}$$

Based on the BERT encoding to generate text and aspect word feature representations, the model uses average pooling to further extract corresponding features according to the downstream task feature requirements. Specifically, as shown in Formulas (2)-(4), the model uses average pooling to generate the fused overall context semantic feature h^c, the aspect word feature h^{seg1} in segment 1, and the aspect word feature h^{seg2} in segment 2, according to the location of each word feature. [:] represents feature sequence slicing, and $left\_index$ represents the number of words to the left of the aspect word in segment 1.

$$h^c = avg\_pooling(H^{text}[1 : n + m + 3]) \tag{2}$$

$$h^{seg1} = avg\_pooling(H^{text}[left\_index + 1 : left\_index + n + 1]) \tag{3}$$

$$h^{seg2} = avg\_pooling(H^{text}[m + 2 : m + 2 + n]) \tag{4}$$

3.2. Image feature learning module.

3.2.1. *Proposed 2-D inception structure.* In the process of sentiment analysis, it is effective to use CNN-based learning methods for image features [27]. However, the problem with CNN is that the size of the convolution kernel is fixed within a layer of convolution, so the field of view of a single layer of convolution is also fixed. If you want to expand the field of view of the convolution, you need to stack multiple layers of convolution, which increases the parameter scale and training cost, and is also prone to overfitting. The Inception structure is an improvement on CNN. The Inception structure uses convolution kernels of multiple scales within a layer of convolution to provide various fields of view. A single layer of convolution can obtain features with rich information and has a smaller parameter scale. Therefore, this paper improves the original Inception structure and proposes a 2D-Inception structure with multi-scale convolution, as shown in the box in Figure 3. The convolution operation of 2D-Inception contains three branches: the first branch first uses a convolution kernel with a width of 1 to increase the dimension of the original data, and then uses a convolution kernel with a width of d for convolution; the second branch uses a convolution kernel with a width of $2d$, which doubles the field of view compared to the first branch; the third branch performs pooling sampling and dimension increase on the original data, with a sampling width of $1.5d$. This branch retains the original data information and performs channel mapping. Finally, the convolution pooling results of the three branches are connected in the channel dimension. The calculation results contain both the convolution results of two scales and the original data features, so features with richer information can be obtained compared to general convolution.

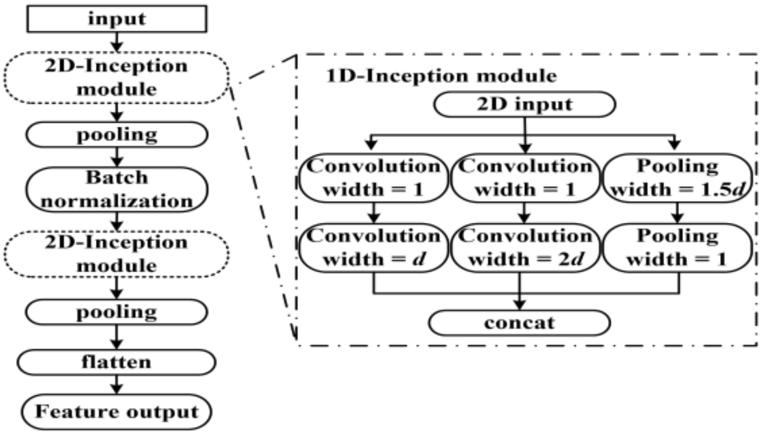


Figure 3. Proposed module of image feature learning.

In order to further reduce the parameter scale and training cost, this paper uses pooling layers and batch normalization to process the convolution results of 2D-Inception. After using average pooling to sample the calculation results, batch normalization is used to adjust the features within the batch to the standard normal distribution, making the loss function flatter and accelerating the learning process [28]. Since the parameter scale and training cost of a single layer are not high, this paper stacks two 2D-Inception blocks to increase the module learning ability, and uses pooling sampling and batch normalization to connect them in the middle to form the 2D-Inception feature learning module used in this paper, as shown in the dotted box in Figure 3.

3.2.2. *Self-attention module.* Google Machine Translation team proposed the self-attention mechanism in June 2017, and it has been widely used in many fields in recent years. However, there are not many applications of the self-attention mechanism combined with deep learning models in the field of sentiment analysis. Unlike standard ordinary attention, self-attention focuses on the joint learning and self-matching of two sequences, where the attention weight of one sequence depends on the other sequence, and vice versa. Usually, the self-attention mechanism does not use other additional information, but it can use self-attention to focus on itself and extract more relevant information from the input, learn the contextual information of the input, and capture their mutual relationship. Its basic structure is shown in Figure 4.

In sentiment analysis, the self-attention mechanism can be employed to extract nuanced sentiment signals from text data. By allowing the model to focus on the most relevant parts of the input, it can learn complex interactions between words and phrases, identifying sentiment shifts, polarity inversions, and sarcasm, all of which are crucial for accurate sentiment classification. The self-attention mechanism, with its unique ability to capture long-range dependencies, focus on relevant information, and scale efficiently, holds immense potential for advancing the state-of-the-art in sentiment analysis. As research in this area progresses, we can expect to see more innovative applications of self-attention, integrated with deep learning models, to unlock new insights and enhance the performance of sentiment analysis systems.

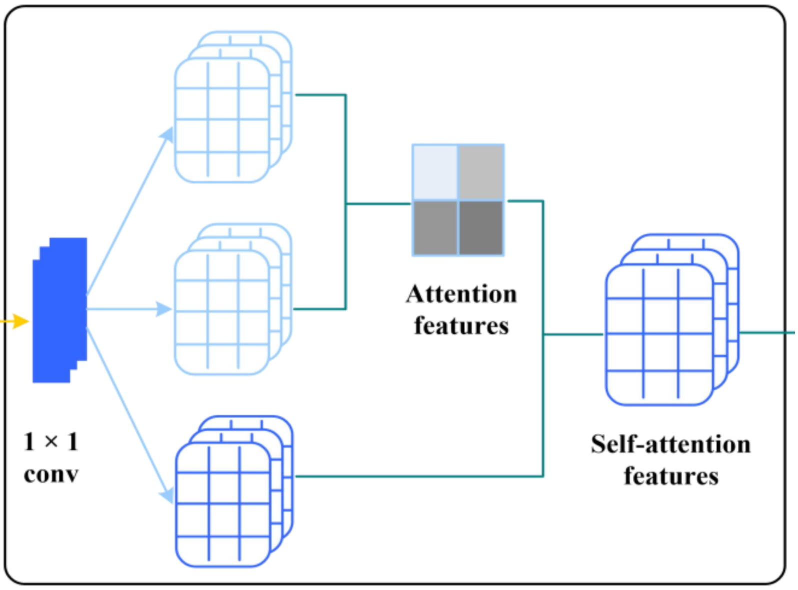


Figure 4. The basic architecture of self-attention module.

The self-attention module is located in the middle of the convolutional layer. The input is the feature map output by the previous convolutional layer, and the output is added to the original feature map to obtain the input of the next layer. The feature map $x \in R^{C\times N}$

from the previous hidden layer is first converted to two specific feature spaces f and g to calculate the attention value, where $f(x) = W_f \cdot x, g(x) = W_g \cdot x$.

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, \text{ and } s_{ij} = f(x_i)^\top g(x_j) \tag{5}$$

Its value represents the degree of attention paid by the model to the i-th region when the j-th region is generated. The output of the attention module is $o = (o_1, o_2, \dots, o_N) \in R^{C\times N}$. Finally, the weighted output of the attention module is added to the original feature map to obtain the final result, which is used as the input of the next hidden layer.

3.3. **Joint Training and Loss Function.** After the BERT model and image feature learning module encode text and image features respectively, how to align the features of different modal data is a key task in the multimodal field. Contrastive learning maps the original sample and the transformed sample as positive sample pairs to similar representation spaces, which not only enables the BERT model and image module to learn more informative feature representations, but also increases the diversity and richness of data, and improves the robustness and generalization ability of the pre-trained model. By designing positive and negative samples on the text side and the image side respectively, similar entity feature representations are mapped to the same coordinate space, thereby achieving alignment of multi-source features.

Inspired by InfoNCE [29], by designing a loss function, the model can be encouraged to learn feature representations that can distinguish samples. Formula (6) gives the loss function on the text side, where $sim(h_i, h_j)$ represents the similarity value between features h_i and h_j. τ is called the temperature coefficient, which is used to scale the similarity value between features. N samples are used in each batch of data to obtain N pairs of word representations. To minimize the loss, the model is required to minimize the similarity distance of aspect word representations in the same sample and maximize the similarity distance between the aspect word representations h_i^{seg1} and h_i^{seg2} of two fragments in any sample of a batch of data and the other $2 * (N - 1)$ aspect word representations.

$$Loss_{\text{text}} = -\log\left(\frac{\exp(sim(h_i^{seg1}, h_i^{seg2})/\tau)}{\sum_{i=1}^{N} \exp(sim(h_i^{seg}, h_j^{seg})/\tau)}\right) \tag{6}$$

On the image side, extracting the $[CLS]$ vector as the feature representation of each image can obtain positive sample pairs of images $< h_i^{img1}, h_i^{img2} >$ with similar features. The two image feature representations of the current sample and the two image feature representations of other samples in the same batch of data are used as negative sample pairs $< h_i^{imgw}, h_j^{imgq} >$, where $i \neq j; w = 1, 2; q = 1, 2$.

Similar to the text aspect, Formula (7) gives the corresponding loss function for the image side. N samples are used in each batch of data to obtain N pairs of image representations. To minimize the loss, the model is required to minimize the distance between image feature representations in the same sample and maximize the similarity distance between two image feature representations h_i^{img1}, h_i^{img2} in any sample of a batch of data and other $2 * (N - 1)$ image feature representations.

$$Loss_{\text{img}} = -\log\left(\frac{\exp(sim(h_i^{img1}, h_i^{img2})/\tau)}{\sum_{i=1}^{N} \exp(sim(h_i^{img}, h_j^{img})/\tau)}\right) \tag{7}$$

Similar to the text aspect, Formula (8) gives the loss function for audio. This paper uses short-time Fourier transform (STFT) to convert the audio signal from the time domain to the frequency domain to capture time and frequency information. The analysis is mainly carried out through the time-frequency graph generated by STFT. N samples are used in each batch of data to obtain N pairs of audio representations. To minimize the loss, the model is required to minimize the distance between the audio feature representations in the same sample and maximize the similarity distance between the two audio feature representations h_i^{audio1} and h_i^{audio2} in any sample of a batch of data and the other $2*(N-1)$ image feature representations.

$$Loss_{\text{audio}} = -\log\left(\frac{\exp(sim(h_i^{audio1}, h_i^{audio2})/\tau)}{\sum_{i=1}^{N}\exp(sim(h_i^{audio}, h_j^{audio})/\tau)}\right) \tag{8}$$

Based on contrastive learning to align multi-source features, this paper makes full use of the overall features of text and the local features of images. In terms of text, the pooled text feature representation h^c fully integrates the overall contextual semantics. In terms of image, after encoding the randomly cropped image, the generated image features h^{img1} and h^{img2} represent part of the image features respectively. The text-side h^c is concatenated with the image-side features h^{img1} and h^{img2}, as shown in Formula (9), to generate the final features for multimodal aspect-level sentiment analysis.

$$h^{\text{text+img+audio}} = concat(h_i^{img1}; h^c; h_i^{img2}) \tag{9}$$

According to the predefined emotion types in the target data corpus, the feature vector $h^{\text{text+img+audio}}$ is input into the fully connected layer and the softmax classification layer in sequence to calculate the probability distribution of the emotion category, as shown in Formula (10). The emotion corresponding to the maximum probability is the result predicted by the model.

$$\hat{y} = softmax(W_o h^{\text{text+img+audio}} + b_o) \tag{10}$$

In order to fine-tune the overall model through the multimodal aspect-level sentiment analysis task, minimizing the cross entropy loss function is adopted as the objective function of this task, as shown in Formula (11).

$$Loss_{\text{Ft}} = -\sum_{i=1}^{N} y_i \log \hat{y}_i + \lambda||\theta||_2 \tag{11}$$

The multimodal aspect-level sentiment analysis task and the contrastive learning task share the feature representation generated by the text and image modules, and the two tasks are combined to fine-tune the parameters of each encoder model. The specific fine-tuning process is manifested in the training process of the two tasks, during which the overall model is back-propagated to minimize the sum of the training losses to achieve the effect of parameter update. The joint objective loss function used by the proposed model includes two parts: the cross entropy loss $Loss_{\text{Ft}}$ used to complete the multimodal aspect-level sentiment analysis and the contrast loss composed of $Loss_{\text{text}}$, $Loss_{\text{img}}$ and $Loss_{\text{audio}}$, as shown in Formula (12).

$$Loss_{\text{total}} = Loss_{\text{Ft}} + Loss_{\text{img}} + Loss_{\text{text}} + Loss_{\text{audio}} \tag{12}$$

4. **Experiments and analysis.**

4.1. **Datasets.** The public datasets used in this paper are CMU-MOSI [18, 19] and CMU-MOSEI [30].

The CMU-MOSI dataset is the first sentiment analysis dataset annotated by opinion proposed by Zadeh et al. [18]. It includes opinion data in various forms such as monologues, speeches, and movies, with a total of 93 videos and 2,198 video clips. These video clips are manually annotated with sentiment scores between $[-3, 3]$, where -3 "indicates" very negative sentiment and 3 "indicates" very positive sentiment. Among them, the training set has 1,281 utterances, the validation set has 229 utterances, and the test set has 685 utterances.

The CMU-MOSEI dataset is a sentiment analysis dataset proposed by Zadeh et al. [30] in 2018 to improve CMU-MOSI. CMU-MOSEI has a larger number of samples, and the expressers and topics are more diverse. The dataset contains 23,453 video clips from 5,000 different videos. The training set has 16,265 utterances, the validation set has 1,869 utterances, and the test set has 4,643 utterances.

4.2. **Experimental indicators and hyper-parameters.** Generally, multimodal sentiment analysis tasks can be considered as regression tasks. In order to comprehensively evaluate the proposed method, this paper uses various standard regression task evaluation methods such as mean absolute error (MAE) and Pearson correlation coefficient (Corr). The specific calculation formulas of MAE and Corr are as follows:

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|y_i - \hat{y}_i| \tag{13}$$

$$Corr = \frac{E[(y-\mu_y)(\hat{y}-\mu_{\hat{y}})]}{\sigma_y \sigma_{\hat{y}}} \tag{14}$$

Where N represents the total number of samples, y represents the true label value, $\hat{y}$ represents the predicted value, σ is the standard deviation, and μ is the expectation.

In addition, to be consistent with other tasks, the accuracy of two categories (Acc-2), seven categories (Acc-7) and F1 score are also used to evaluate the performance of the model. Among the above evaluation indicators, the lower the MAE value, the better the performance of the model, and the higher the other indicators, the better the performance of the model.

This method uses Python programming, and the deep learning framework uses PyTorch. It is trained and tested on an NVIDIA 4070 GPU with a video memory of 12GB. The optimizer uses Adam, and the learning rate of the optimizer is 1e-5. The batch size is 48, the learning rate is 1e-5 when BERT is fine-tuned, and the number of transform layers is set to 5.

4.3. **Verification of the effectiveness of the 2D-Inception module.** In order to prove the effectiveness of the 2D-Inception feature learning module compared with traditional feature extraction methods and traditional CNN, this paper conducted an effectiveness verification experiment. First, SVM was used as the classifier and Gaussian kernel was used as the kernel function. Then a simple 3-layer CNN was built for direct classification of images. The 2D-Inception feature learning module was set up separately, and classification was performed directly after feature learning to verify the classification capabilities of the three. The experimental results are shown in Table 1. ACC (Accuracy) and STD (STandard Deviation) represent the average classification accuracy and standard deviation of accuracy, respectively. The 2D-Inception module achieved the highest classification accuracy, and the parameter scale of the 2D-Inception module is smaller than that of the 3-layer CNN, which shows that the module in this paper has a smaller parameter cost and higher feature learning performance, the learned features are more classifiable, and the generalization ability between different individuals is stronger. This proves that the 2D-Inception feature learning module proposed in this paper is more suitable for image feature learning.

Table 1. The accuracy of the 2D-Inception module and other extraction methods

Extraction method	ACC (%)	STD (%)
SVM	53.81	13.22
CNN	66.27	10.18
2D-Inception	80.91	9.01

4.4. **Experimental results on public datasets.** Table 2 and Table 3 show the comparative experimental results of the proposed method and the baseline model on the CMU-MOSEI and CMU-MOSI data sets respectively, where "—" means that the data was not reported in the original paper, and the bold content is Indicates the model with the best performance for this indicator. We can see that the proposed method performs very well on both public datasets and is higher than or on par with the baseline model on all evaluation metrics.

Table 2. Results of the proposed method and other methods on CMU-MOSEI dataset

Model	MAE	Corr	Acc-2	F1 score	Acc-7
[18]	—	—	76.2	76.5	—
[20]	0.619	0.663	79.1	79.2	50.1
[21]	0.594	0.710	82.2	82.4	51.7
[22]	0.571	0.719	84.2	84.2	51.3
[23]	0.563	0.712	84.1	84.3	51.2
[24]	0.558	0.761	85.5	85.3	52.9
[25]	0.541	0.769	85.3	85.2	—
Proposed	**0.529**	**0.773**	**87.1**	**86.8**	**53.9**

Many early research works use traditional fusion methods to establish inter-modal interactions, while the proposed method uses the Transformer architecture to learn consistent information between modalities, so its performance far exceeds these models. [21] used a method based on the cross-modal attention mechanism to achieve inter-modal interaction. However, this method uses a directional paired cross-modal Transformer, which generates a lot of redundant information and fails to take into account the difference information between different modalities [24]. Based on the concept of modality-invariant and modality-specific representation learning, the regularization terms learned from different sub-tasks are added to the loss function of the final prediction task to obtain high-quality multimodal representations.

Compared with all baseline models, the proposed method performs very well in sentiment prediction. Experimental results show that the proposed method achieves excellent results on both the large dataset CMU-MOSEI and the smaller dataset CMU-MOSI,

Table 3. Results of the proposed method and other methods on CMU-MOSI dataset

Model	MAE	Corr	Acc-2	F1 score	Acc-7
[18]	0.960	0.631	77.3	77.2	34.2
[20]	0.917	0.696	78.1	76.4	33.3
[21]	0.872	0.692	83.1	82.9	40.2
[22]	0.809	0.711	83.2	83.1	39.2
[23]	0.789	0.784	84.9	84.9	—
[24]	0.787	0.762	85.1	83.5	45.8
[25]	0.712	**0.796**	86.0	86.1	—
Proposed	**0.689**	**0.796**	**86.7**	**86.8**	**49.1**

which shows that the proposed method is effective in combining image feature extraction based on hierarchical attention and target modality encoding, and is applicable to different data scenarios. In addition, the designed improved 2D-Inception also achieved the expected results, combining multimodal information with unimodal specific semantics. In summary, these network structures have made certain contributions to multimodal sentiment analysis tasks.

4.5. **Ablation experiments.** In order to study the effect of learning based on different optimizers, this paper uses the stochastic gradient descent optimizer (SGD) and the adaptive gradient optimizer (Adam) to learn and determine whether the method can achieve good results based on different optimizers. The experimental results are shown in Table 4. As can be seen from Table 4, the method in this paper can achieve good improvement effects for different optimizers. In the experiment with stochastic gradient descent as the optimizer, the method in this paper can achieve the best improvement effect, while the improvement of the experimental results using the adaptive gradient optimizer is small.

Table 4. The results of different optimizers

Optimizer	CMU-MOSI		CMU-MOSEI	
	Acc-2	F1 score	Acc-2	F1 score
SGD	82.1	86.2	84.3	84.9
After fusion	83.0	86.8	84.9	85.1
Adam	83.2	83.9	85.1	85.5
After fusion	83.9	84.1	86.0	86.7

In addition, based on the MOSI dataset, this paper compares the changes in the learning gradient amplitude of the three modalities before and after the fusion method. The comparison results are shown in Figure 5. The ordinate represents the learning gradient amplitude of the modality, and the abscissa represents the training round of the model. It can be seen that after 10 epochs, the gradient changes of the three modalities of Text, Audio, and Video tend to be flat, and basically reach a convergence state.

To further reflect the impact of the proposed method on the results of unimodal and multimodal tasks, the CMU-MOSI dataset is used as the object to compare the results

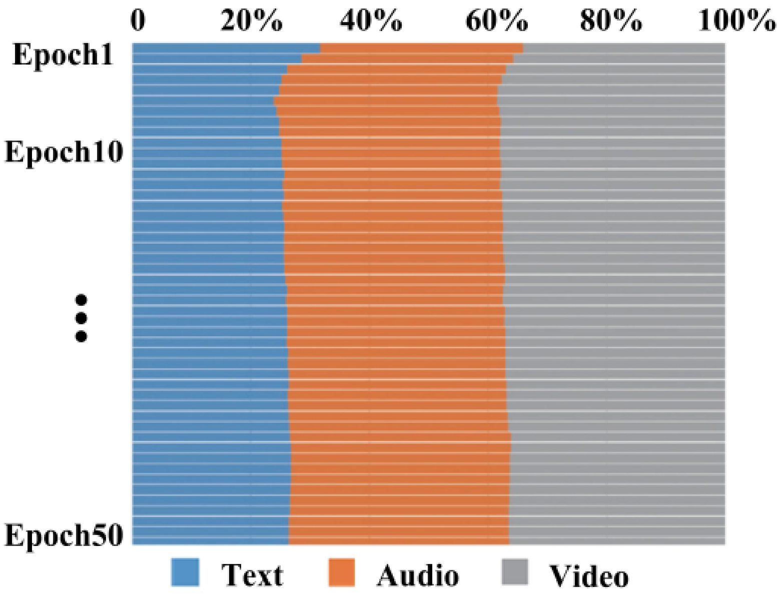


Figure 5. Gradient transformation trend chart after fusion.

of unimodal and multimodal tasks before and after feature fusion. The results are shown in Table 5. Among them, the experimental results of unimodal are obtained by using the output of the unimodal feature extraction network for sentiment analysis. From the results, it can be seen that when fusion is not used, the results of multimodal tasks are dominated by text modality, and the task accuracy of audio and video modalities is low. After fusion, the accuracy of unimodal task analysis is increased, and the results of multimodal tasks are also improved.

Table 5. The comparison of single modal and multimodal on CMU-MOSI dataset

Modal	No fusion	Proposed method
Text	83.2	84.9
Audio	65.1	72.1
Video	66.9	73.2
Multimodal	84.5	85.6

5. **Conclusion.** This paper proposes a multimodal sentiment analysis method based on collaborative feature representation of BERT and improved Inception. Among them, short-time Fourier transform is used to extract audio information, BERT is used to extract text information, and improved Inception is used to extract image information. Feature alignment is performed in three different modalities: text, image, and language, so as to promote deep fusion of cross-modal information and improve the accuracy of multimodal aspect-level sentiment analysis. Experiments are conducted on two public datasets, and the results show that the proposed method is superior to other excellent comparative methods in multiple indicators. The experimental results also prove that the combination of hierarchical attention image feature extraction and target modality encoding is effective and can be applied to different data scenarios.

In the future, we will use better strategies to combine modality-invariant and modality-specific representations, and we can also introduce mathematical knowledge to design loss functions to aggregate the correlation between two features. At the same time, we consider applying contrastive learning to two features of the same sequence or the same type to explore the potential relationship between them. What's more, contrastive learning has emerged as a powerful framework for learning representations that capture the intrinsic structure of the data. For sequential data, contrastive learning can be extended to compare representations of the same sequence or type but under different conditions or augmentations. This can help the model learn temporal dynamics and capture subtle differences that may indicate important relationships or sentiment shifts.

REFERENCES

[1] H. Zhao, M. Yang, X. Bai, and H. Liu, "A Survey on Multimodal Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 12, pp. 12039-12052, 2024.

[2] X. Ju, D. Zhang, R. Xiao, J. Li, S. Li, M. Zhang, and G. Zhou, "Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4395-4405, 2021.

[3] Z. Chen, T. Qian, "Enhancing aspect term extraction with soft prototypes," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2107-2117, 2020.

[4] Y.-Z. Zhang, R. Lu, D.-W. Song, and P. Zhang, "A Survey on Multimodal Sentiment Analysis," *Pattern Recognition and Artificial Intelligence*, vol. 33, pp. 426-438, 2020.

[5] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, pp. 2018.

[6] N. Xu, W. Mao, G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 371-378, 2019.

[7] T. B. Brown, "Language models are few-shot learners," *arXiv preprint ArXiv:2005.14165*, pp. 2020.

[8] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, and L. He, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *arXiv preprint arXiv:2302.09419*, pp. 2023.

[9] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Targeted sentiment classification with attentional encoder network," *Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV 28*, pp. 93-103, 2019.

[10] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290-154299, 2019.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *International Conference on Machine Learning*, pp. 1597-1607, 2020.

[12] J. Yu, J. Jiang, R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429-439, 2019.

[13] Z. Zhang, Z. Wang, X. Li, N. Liu, B. Guo, and Z. Yu, "ModalNet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network," *World Wide Web*, vol. 24, pp. 1957-1974, 2021.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, pp. 2020.

[15] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3454-3466, 2018.

[16] H. Wang, C. Ma, Y. Liu, Y. Chen, Y. Tian, J. Avery, L. Hull, and G. Carneiro, "Enhancing Multimodal Learning: Meta-learned Cross-modal Knowledge Distillation for Handling Missing Modalities," *arXiv preprint arXiv:2405.07155*, pp. 2024.

[17] Y. Sun, S. Mai, H. Hu, "Learning to balance the learning rates between various modalities via adaptive tracking factor," *IEEE Signal Processing Letters*, vol. 28, pp. 1650-1654, 2021.

[18] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, “Balanced multimodal learning via on-the-fly gradient modulation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238-8247, 2022.

[19] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor Fusion Network for Multimodal Sentiment Analysis,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017.

[20] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” *Proceedings of the Conference Association for Computational Linguistics. Meeting*, pp. 6558, 2019.

[21] J. Li, C. Wang, Z. Luo, Y. Wu, and X. Jiang, “Modality-Dependent Sentiments Exploring for Multi-Modal Sentiment Classification,” *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7930-7934, 2024.

[22] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, “Integrating multimodal information in large pretrained transformers,” *Proceedings of the Conference Association for Computational Linguistics. Meeting*, pp. 2359, 2020.

[23] W. Yu, H. Xu, Z. Yuan, and J. Wu, “Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis,” *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10790-10797, 2021.

[24] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, “A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis,” *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4730-4738, 2021.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, “Learning transferable visual models from natural language supervision,” *International Conference on Machine Learning*, pp. 8748-8763, 2021.

[26] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, “Unified contrastive learning in image-text-label space,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19163-19173, 2022.

[27] J. Chen, P. Zhang, Z. Mao, Y. Huang, D. Jiang, and Y. Zhang, “Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks,” *IEEE Access*, vol. 7, pp. 44317-44328, 2019.

[28] S. Ioffe, C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *International Conference on Machine Learning*, pp. 448-456, 2015.

[29] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729-9738, 2020.

[30] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236-2246, 2018.