Cross-Modal Sentiment Analysis Based on CLIP Image-Text Attention Interaction

Xintao Lu[1], Yonglong Ni[2], Zuohua Ding[3]
Faculty of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China[1,3]
Zhejiang Petroleum Comprehensive Energy Sales Co., Ltd., Hangzhou, Zhejiang, China[2]

***Abstract*—Multimodal sentiment analysis is a traditional text-based sentiment analysis technique. However, the field of multi-modal sentiment analysis still faces challenges such as inconsistent cross-modal feature information, poor interaction capabilities, and insufficient feature fusion. To address these issues, this paper proposes a cross-modal sentiment model based on CLIP image-text attention interaction. The model utilizes pre-trained ResNet50 and RoBERTa to extract primary image-text features. After contrastive learning with the CLIP model, it employs a multi-head attention mechanism for cross-modal feature interaction to enhance information exchange between different modalities. Subsequently, a cross-modal gating module is used to fuse feature networks, combining features at different levels while controlling feature weights. The final output is fed into a fully connected layer for sentiment recognition. Comparative experiments are conducted on the publicly available datasets MSVA-Single and MSVA-Multiple. The experimental results demonstrate that our model achieved accuracy rates of 75.38% and 73.95% , and F1-scores of 75.21% and 73.83% on the mentioned datasets, respectively. This indicates that the proposed approach exhibits higher generalization and robustness compared to existing sentiment analysis models.**

***Keywords*—Multi-modal; image-text interaction; multi-head attention mechanism; sentiment analysis; cross-modal fusion**

I. Introduction

2023 Global Digital Report [1] indicates that there are currently 5.16 billion internet users and 4.76 billion social media users worldwide, accounting for 59.4% of the global population. The global social media user base has grown by 3.0% year-on-year, equivalent to 137 million people. With the continuous flourishing development of social networks and internet-enabled mobile devices, there is a growing diversity of expressions for emotions or opinions on various topics posted on social media and various website platforms. This evolution has transitioned from initial text information to gradually include multimodal information such as images, audio, and videos. Consequently, utilizing multimodal feature information for sentiment analysis has become one of the research hotspots in recent years [2], and it has been successfully applied in various potential applications, including decision-making [3], personalized advertising [4], emotion retrieval [5][6], and other domains.

Early sentiment analysis (SA) models mainly focused on text, and manual features were usually designed using limited human knowledge. Text features can quickly summarize subjective emotions, but cannot fully describe the highly abstract nature of emotions. For single-modal approaches, extreme cases such as irony often posed challenges in meeting sentiment analysis needs. In recent years, the rise of deep learning has provided powerful tools for Multimodal Sentiment Analysis (MSA). MSA leverages massive multimodal data generated on social media for integrated analysis, combining various multimodal features. This approach not only enables a more comprehensive understanding of user emotional expressions but also effectively addresses limitations of single-modal methods in handling complex emotions, ambiguity, or extreme cases like irony [7].

However, there are still some shortcomings in multimodal feature fusion methods. In early multimodal fusion methods, they either simply concatenate the extracted multimodal features [8] or roughly integrate relationships between images and text on a horizontal feature level [9] to obtain concatenated or linearly fused feature representations. These methods lack in-depth exploration of the complex relationships between multiple modal features. On the other hand, information loss, redundancy, and noise among different modal features can affect sentiment judgments. Effectively utilizing the complex correlations between high-level abstract features and low-level abstract features across modalities and improving method fusion effectiveness pose significant challenges in the field of multimodal analysis [10].

In order to solve the problem of multi-modal model fusion, this paper proposes a Multimodal Sentiment Analysis model, named CLIP-CA-CG, based on Contrastive Language-Image Pretraining (CLIP) [11], cross-attention, and cross-modal gating.The CLIP model maps text and images into a shared embedding space, making related text descriptions and image representations in this space closer, modeling at fine-grained features, and the model uses contrastive learning and pre-training. This method can learn feature representations with good generalization capabilities and reduce computational pressure and speed. Additionally, considering the complementary role of the contextual information of image text in sentiment analysis, where the same word may cause different emotions in different contexts, so this model integrates the original cross-modal feature information through the self-attention mechanism. This can extract high-level abstract features while maximizing the fusion of environmental features, which helps the model learn the correlation between different modalities to more comprehensively explore multi-modal emotions.

The remainder of the paper is structured as follows: Section II will review related research on sentiment analysis of unimodal and multimodal models, Section III will provide the methods proposed in this study, Section IV will introduce the experimental analysis and discussion, and finally, conclusion and future works are provided on Section V.

II. Related Work

A. Single-modal Sentiment Analysis

1) Text sentiment analysis: In the past, conventional techniques for text sentiment analysis primarily utilized dictionary methods [12]. In this approach, the sentiment scores of individual words within the text are combined based on predefined values. Text sentiment classification methods can be roughly divided into two categories, namely dictionary based models and machine learning models. Hu et al. [13] predicted the semantic analysis orientation of opinion sentences by using adjectives as prior positive or negative polarity. Taboada et al. [14] introduce a dictionary-based method called the Semantic Orientation Calculator (SO-CAL), which not only utilizes word dictionaries annotated with semantic orientations but also incorporates reinforcement and negation factors. Barbosa et al. [15] proposed a two-step sentiment classification method for Twitter messages using online tags as training data.

With the evolution of machine learning, Pang et al. [16] are the first to introduce machine learning methods into text sentiment classification, including Naive Bayes (NB) [17], Support Vector Machine (SVM) [18], and Maximum Entropy Classifier. However, machine learning performance heavily relies on the quality and quantity of the training set. Inspired by the success in the field of Natural Language Processing (NLP), Kim et al. [19] first apply Convolutional Neural Networks (CNN) in text sentiment classification. Tai et al. [20] considered the complex structure of text features and introduced Tree LSTM for sentence sentiment classification. Tang et al. [17] first combined CNN and LSTM to obtain text sentence representations, and then used recursive neural networks to encode their intrinsic connections [21]. Researchers also adopt various neural network models for sentiment analysis, such as hierarchical attention network model (HAN) to select important feature information [22] and facial expression recognition network based on enhanced attention [23].

As large models gain prominence, word embeddings and pre-trained models have seen significant success in sentiment analysis. Word2Vec [24] maps semantically similar words to similar vector spaces, while GloVe [25] derives semantic relationships between words based on global co-occurrence. ELMo [26] introduces context-aware embeddings, allowing word representations to vary based on their specific contexts within sentences. The emergence of BERT [27] further propels the development of sentiment analysis. Built on self-attention mechanisms, BERT captures long-range dependencies and contextual information more effectively. This context-sensitive representation enables BERT to achieve outstanding performance in sentiment analysis tasks, particularly excelling in handling complex sentence structures and context-dependent sentiment expressions.

2) Visual sentiment analysis: Visual sentiment analysis has undergone significant development. In the early stages, image sentiment analysis involved inferring emotions from low-level features. For example, Machajdik et al. [28] predict emotions by extracting features such as texture and color. Borth et al. [29] used the SentiBank model to identify adjective noun pairs (ANP) and extract visual semantic information.Yuan et al. [30] proposed an image sentiment method that utilizes 102 intermediate visual attributes to make the classification results more interpretable.

In recent years, with the continuous advancement of deep learning, researchers have explored the coordination of image color and content in relation to emotional expression. Yang et al. [31] developed a multi task framework to optimize visual emotion models by considering mixed images of multiple emotions. Ruan et al. [32], for instance, employ CNN networks to extract both content and color features from images. By introducing attention mechanisms and sequence convolution, they adeptly model the correlations between content and color features. To delve deeper into the semantic associations among visual emotion regions, Zhang et al. [33] utilize a fully convolutional neural network for image saliency detection. The CNN selection strategy is employed for filtering, and ultimately, Transformer encoders [34] are used to analyze the correlations between different emotion regions, thereby obtaining a comprehensive emotional output.

B. Multimodal Sentiment Analysis

In the field of multimodal research, psychologists have confirmed that emotions are primarily influenced by the joint effects of multimodal data, with visual-text emotional features being particularly prominent. The same piece of text pairs with different images may elicit completely opposite emotions. In early multimodal sentiment analysis, researchers concatenate, added, or weighted shallow features. Cao et al. [35], for example, analyze cross-media sentiment analysis through visual and textual methods. Yu et al. [36] use a pre-trained CNN model to extract feature representations and ultimately fused textual features for sentiment classification.Zhao et al. [37] proposed an image text consistency driven method that utilizes text features, social features, low-level and intermediate visual features, and image text similarity.

As deep learning continues to evolve, mid-term model fusion and late-stage decision fusion methods are showing remarkable success. Yang et al. [38] achieve good results by stacking and gradually pairing different feature vectors on datasets like CMU-MOSI. Poria et al. [39] detail an approach using Long Short-Term Memory (LSTM) networks to capture interdependencies and relationships between utterances in multimodal sentiment prediction. Huang et al. [40] proposed a Deep Multimodal Attention Fusion (DMAF) method, which utilizes both intermediate and post fusion, combining unimodal features and internal cross modal correlations to improve accuracy. Liu et al. [41] introduce a shared memory attention mechanism, capturing interactions between two modalities and their impact on sentiment using similar features.

In recent years, multimodal tasks have made significant progress, benefiting from the latest developments in visual language models. Cheema et al. [42] apply CLIP in multimodal sentiment analysis, demonstrating its potential as a powerful baseline for emotion prediction tasks in tweets. Arevalo et al. [43] propose the Gated Multimodal Unit (GMU) model, which controls the influence of input modalities on unit activation levels for data fusion. Gupta et al. [44] introduce a Collaborative Attention Model based on RoBERTa and FiLMed ResNet, addressing the issue of visual-text inconsistency through joint attention mechanisms.

Although multi-modality has made certain progress in emotional tasks, there is still much room for improvement in image-text feature interaction. Most existing methods simply connect features extracted from different modalities, or simply learn The relationship between images and text leads to bias in complex tasks. Considering the complex relationship between the two modalities and the efficiency of the model, we use a pre-trained model to extract feature networks while capturing the potential alignment between image regions and text words, and finally consider the complementary role of individual modalities in emotion prediction. , situational features are also integrated into our network.

III. Methods

This paper proposes a cross-modal sentiment model, CLIP-CA-CG, based on CLIP image-text attention interaction, as illustrated in Fig. 1. The model architecture consists mainly of a feature extraction layer, an interaction attention layer, a gating fusion layer, and a regression layer. The feature extraction module utilizes existing methods for extracting features from images and text, producing feature vectors for each and a fused feature vector. The interaction attention module enhances the feature representation of images and text based on a multi-head attention mechanism, further exploring consistent emotional features in the image-text pairs. The gating fusion module aligns high-level abstract image-text features, fuses global concrete features, and introduces an adaptive cross-selective block to determine how much interaction information each component should transmit. Finally, sentiment is comprehensively predicted through a multi-layer perceptron and a Softmax regression layer.

A. Image-Text Feature Extraction

The original input of the model consists of two modalities: text and image. For the raw textual data, a set of textual data T can be represented as n words forming $T = [T_1, T_2, ..., T_n]$, where n represents the maximum length of the sequence. Considering the need for a more comprehensive understanding of context and capturing bidirectional language relationships, this paper utilizes the pre-trained RoBERTa (Robustly optimized BERT approach) model to encode the text sequence T. The advantage of the RoBERTa model lies in further optimizing the BERT model by adjusting training tasks, datasets, learning rates, etc. Additionally, unlike BERT, RoBERTa does not add special token embeddings at the beginning and end of the input text, enhancing the generalization of text feature extraction. The textual data is embedded into vectors $F_{T-RoBERTa}$ by the RoBERTa model, where each word is represented in the vector space.

$$F_{T-RoBERTa} = [f_1, f_2, ..., f_x, ..., f_N] \subseteq R^{d \times N} \tag{1}$$

In the equation: f_x represents the contextual semantic feature of the x-th word, d is the output dimension of the RoBERTa model (768), and N is the maximum length of the RoBERTa model's word encoding.

Then, to summarize the contextual information in the sentence, a Bidirectional Gated Recurrent Unit (Bi-GRU) [45] is employed. The combination of RoBERTa and Bi-GRU ensures the learning of text semantics while preserving multi-granular, multi-level information extraction from the text. The vector $F_{T-RoBERTa}$ is passed through the Bi-GRU gated units to further extract and generate the feature h_x.

$$h_x = [\overrightarrow{GRU}(f_x) \oplus \overleftarrow{GRU}(f_x)] \subseteq R^d \tag{2}$$

In the equation: h_x is the feature extracted from f_x through Bi-GRU, $\overrightarrow{GRU}(f_x)$ denotes obtaining the forward hidden state information, and $\overleftarrow{GRU}(f_x)$ represents acquiring the backward hidden state information. Finally, the average of the bidirectional hidden state information, h_x, is obtained, yielding the ultimate textual semantic feature F_T.

$$F_T = [h_1, h_2, ..., h_N] \in R^{d \times N} \tag{3}$$

For image features, ResNet introduces a residual network structure, addressing the issue of gradient vanishing that arises with increasing network depth. Moreover, deeper network structures can handle images under different sizes, angles, and lighting conditions. In this paper, pre-trained ResNet50 is used for feature extraction. Simultaneously, each original image is cropped to 224×224×3 as input for ResNet50. After convolution and pooling, the image feature F_{In} is obtained. Finally, aligning visual feature F_{In} and textual feature FT through a perceptron results in the ultimate image feature F_I.

$$F_I = Linear(F_{In}) \tag{4}$$

After obtaining the original visual and textual features, this paper further utilizes Contrastive Language-Image Pre-training (CLIP) to integrate image and text features, thereby establishing a close connection between them. The core idea of CLIP involves using contrastive learning to represent images and text in a shared embedding space. It maximizes the cosine similarity of paired image and text embeddings while minimizing the cosine similarity of unpaired image and text embeddings. This ultimately brings related images and text closer in this shared space. The original visual-text features, after passing through the CLIP model, result in the fused feature F_{IT}.

$$F_{IT} = CLIP(F_I) \odot CLIP(F_T) \tag{5}$$

B. Multi-Head Attention Mechanism

Multi-Head Attention (MHA) is an extended form of the self-attention mechanism initially introduced in the Transformer model. The core idea is to use multiple distinct attention heads, allowing the model to learn various attention patterns in parallel, with each head focusing on different parts of the sequence. Subsequently, by concatenating the outputs of these heads and projecting them through a linear layer, the final output of multi-head attention is generated. The input to the self-attention mechanism consists of key vectors, query vectors, and value vectors. The mechanism calculates the similarity between query and key vectors, applies a Softmax operation to obtain attention weights for weighted summation, resulting in the final self-attention output as expressed in Formula (6).

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

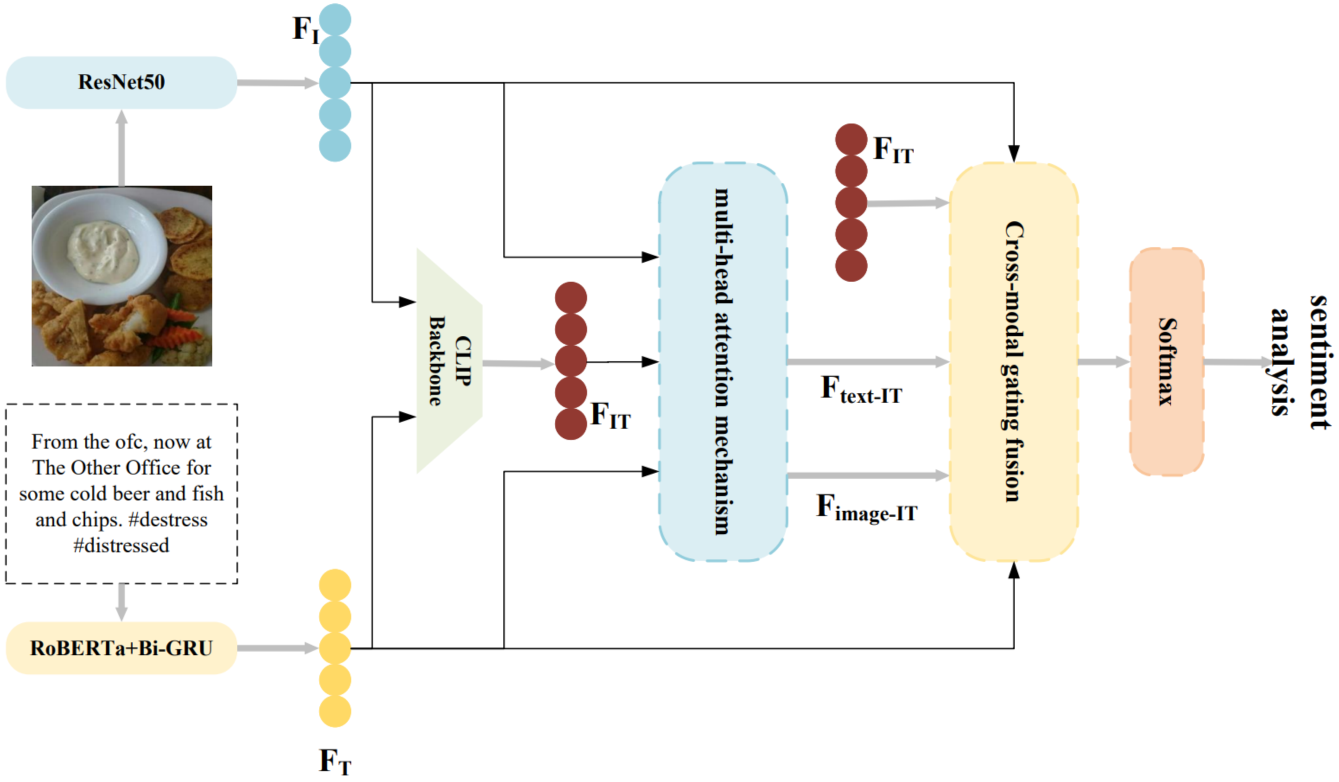
F_I
ResNet50
From the ofc, now at The Other Office for some cold beer and fish and chips. #destress #distressed
RoBERTa+Bi-GRU
F_T
CLIP Backbone
F_IT
multi-head attention mechanism
F_IT
F_text-IT
F_image-IT
Cross-modal gating fusion
Softmax
sentiment analysis

Fig. 1. Overall architecture diagram of the model.

In the equation: Q represents the query matrix, K represents the key matrix, V represents the value matrix, and dk is the dimensionality of the query vectors. Multi-Head Attention is an operation that stacks multiple self-attention mechanisms to focus on different representations of information at different positions.

$$MHA(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{7}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

In the equation: $head_i$ represents the calculation of the i-th attention head, W_i^Q, W_i^K,W_i^V are the weight matrices for linear mappings, and W^O is the weight matrix for the linear mapping of the output. Multi-Head Attention typically includes h attention heads, each with independent weights. The schematic diagram of the multi-head attention mechanism used in this paper is shown in Fig. 2.

By utilizing the image-text features from CLIP, we can obtain more comprehensive global information. In this study, we choose the fusion feature F_{IT} as the main modality for multi-head attention, while visual feature F_I and text feature F_T serve as secondary modality inputs. The main modality learns the sequential information of the secondary modality and ultimately improves the convergence speed and expressive capability of the model through forward propagation. The final output yields feature vectors $F_{image-IT}$ and $F_{text-IT}$.

$$F_{image-IT} = LayerNorm(F_I + MHA(Q_I, K, V)) \tag{9}$$

$$F_{text-IT} = LayerNorm(F_T + MHA(Q_T, K, V)) \tag{10}$$

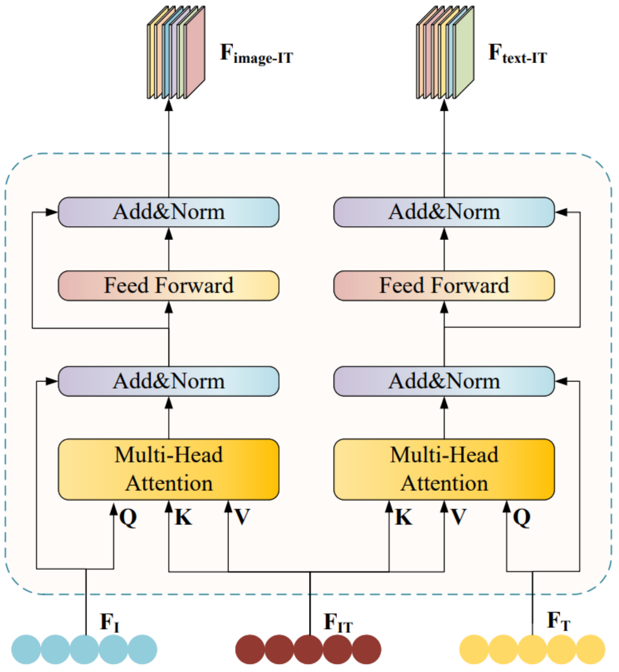

Fig. 2. Principle diagram of multi-head attention mechanism.

C. Cross-modal Gating Fusion

The cross-modal joint feature vector is generated through the interaction and fusion of features from two modalities. It allows vector features to pass fragmentary messages across both modalities for cross-modal interaction. However, in practice, there are still issues such as information redundancy, loss, noise, and region misalignment. To overcome these drawbacks and fully utilize the complementary information of modality correlations contained in the joint features, this paper further proposes a cross-modal gating fusion module. This module adaptively controls the fusion strength through model training to obtain multi-modal fusion features by concatenating them. Considering the significant role of environmental information in sentiment analysis, where the same object may evoke different emotions in various text or visual contexts, it is essential to supplement the fusion with the original image and text features. The structure of the cross-modal gating fusion is illustrated in Fig. 3.

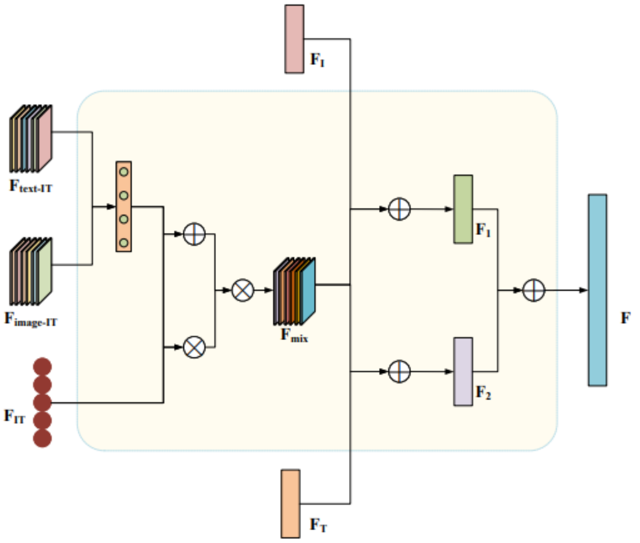

Fig. 3. Cross-modal gated fusion structure diagram.

Firstly, the feature vectors $F_{image-IT}$ and $F_{text-IT}$ obtained through multi-head attention are concatenated to achieve the weight adjustment of joint features. This preserves effectively correlated information in the features, ultimately obtaining the complementary information feature FVT from both, as shown in Formula (11).

$$F_{VT} = [F_{image-IT}, F_{text-IT}] \tag{11}$$

To balance obtaining superior global feature information and the output from higher layers, the output obtained from the features F_{VT} and the globally extracted features F_{IT} by CLIP are used as inputs for fusion. After concatenation and non-linear transformation, the final mixed visual-textual feature information F_{mix} is obtained.

$$F_{mix} = \sigma(Concat(F_{IT}, F_{VT}), Norm(F_{IT}, F_{VT})) \tag{12}$$

In the equation, the σ function represents the fusion of concatenated visual-textual features and non-linearly transformed visual-textual features through trainable parameters. Finally, considering the complementary role of contextual information, we combine the single-modal contextual information feature with the mixed feature F_{mix}, and ultimately generate the final feature F through an MLP.

$$F_1 = MLP(F_I \oplus F_{mix}) \tag{13}$$

$$F_2 = MLP(F_T \oplus F_{mix}) \tag{14}$$

$$F = \lambda F_1 + (1 - \lambda) F_2 \tag{15}$$

In the equation, λ represents the concatenation operation, which is used to control the balance between aggregating visual and textual features.

D. Multimodal Sentiment Classification

The ultimate goal of sentiment analysis is to accurately classify the emotions expressed in multimodal data, such as Positive, Neutral, Negative, etc. To achieve this, the multimodal fusion feature F obtained through the multi-head attention and fusion module is fed into a fully connected layer and a Softmax layer, ultimately producing a probability distribution y for possible sentiment labels.

$$y = Softmax(Linear(F)) \tag{16}$$

In the equation: The Linear network represents the fully connected layer, and the classification results are obtained through Softmax. For model training, this paper utilizes the Adam optimizer to train the model, minimizing the cross-entropy loss.

IV. Experimental Analysis

A. Datasets

In this study, to validate the sentiment analysis performance of the CLIP-CA-CG model, we utilize two publicly available datasets, MVSA-Single and MVSA-Multi, established by Niu et al. [46]. These datasets are collected from the popular social media platform Twitter. The MVSA-Single dataset comprises 5129 pairs of images and text, while the MVSA-Multi dataset includes 196,000 pairs of images and text. The MVSA project provides standardized benchmarks, representing a significant development in the multimodal domain. The data is labeled with sentiment polarity, including positive, neutral, and negative emotions.

For a fair comparative study, we conduct preprocessing on both datasets. During this process, we remove cases where there is emotional inconsistency between the image and text labels, such as one label being positive (or negative) while the other is neutral. Such cases are considered as having a positive (or negative) sentiment label. The resulting new datasets are denoted as the revised MVSA-Single dataset and revised MVSA-Multi dataset, as shown in Table I.

TABLE I. MVSA-Single and MVSA-Multi Datasets

Dataset	Positive	Neutral	Negative	Total
MSVA-Single	2683	470	1358	4511
MSVA-Multiple	11318	4408	1298	17024

B. Implementation Details

In the experiments, we randomly divide the new datasets into training, validation, and test sets, with a data split ratio of 8:1:1. Regarding the experimental environment and parameters, the proposed model is implemented using Python 3.7, developed in the PyTorch 1.9.0 framework, and executed on CUDA 12.0. To eliminate external influences, all experiments are conducted on a server with 64GB of memory and an NVIDIA GeForce RTX 4090 GPU.

In terms of hyperparameter configuration for the model, this experiment employs the cross-entropy loss function and mean squared error loss function for computing the loss of classification and regression tasks, respectively. Adam is utilized as the optimizer for the CLIP-CA-CG model, initialized with a learning rate of 0.0001, executed over 100 epochs, with a 10-fold reduction in learning rate every 10 epochs, and a weight decay of 1e-5. For visual encoding,we utilize the pre-trained ResNet50 to extract image features, taking as input pre-processed image information in the form of a 224×224×3 matrix. In text encoding, we employ pre-trained RoBERTa for extracting text features, where the dimensionality of the extracted word vectors is 768, and subsequently align them for input into the model network. Given the disparate sample sizes in the two datasets, the batch size is set to 64 for the MVSA-Single dataset and 128 for the MVSA-Multi dataset. The initial hyperparameter settings are configured as shown in Table II.

TABLE II. Experimental Parameter Environment

Parameter	Value
Batch_size	64 / 128
Learning_rate	0.0001
Optimizer	Adam
Dropout	0.3
Epochs	100
Text_dimension	768

Finally, to validate the model's effectiveness, comparative experiments are conducted, wherein the proposed model is compared with other mainstream single-modal and multi-modal fusion experiments. Performance evaluation metrics include accuracy and $F_{1-score}$ (F1), calculated as follows.

$$P = \frac{T_P}{T_P + F_P} \tag{17}$$

$$R = \frac{T_P}{T_P + F_N} \tag{18}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{19}$$

$$Acc = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{20}$$

In the equation: T_P represents true positive, T_N represents true negative, F_P represents false positive, F_N represents false negative, P represents precision, and R represents recall.

C. Model Comparison Experiment

In this study, we compare the proposed model with the following benchmark models in terms of accuracy and F1 score. SentiBank and SentiStrength [47] models rely solely on traditional statistical features and are unable to effectively extract key intrinsic features from both images and text, resulting in lower accuracy. Compared to other models, CNN-Multi [48] extracts text and image features separately using two individual CNNs. Benefiting from the powerful feature extraction capability of deep neural networks, this enhances the expressiveness of emotions, and the final prediction is made by connecting these features. The DNN-LR model [49] adopts transfer learning by using pretrained models and utilizes logistic regression for decision analysis. The Co-Memory model [9], introducing a fusion module in sentiment analysis, promotes feature connections between modalities. The MVAN model [38] enhances the semantic image-text features by employing a memory network module on the basis of a multi-view attention network, further improving dataset accuracy. The CLMLF model [50] utilizes contrastive learning to enhance the representation capability of image-text features, fostering relationships between images and text, thus improving model accuracy. The ITIN model [51] introduces cross-modal alignment operations and an adaptive cross-modal gate fusion module, significantly improving accuracy in sentiment analysis tasks.

TABLE III. Comparative Experiments of Several Models

Model	MVSA-S		MVSA-M	
	Accuracy (%)	F1 (%)	Accuracy (%)	F1 (%)
SentiBank&SentiStrength	52.05	50.08	65.62	55.36
CNN-Multi	61.20	58.35	66.39	64.19
DNN-LR	61.42	61.03	67.86	66.33
Co-Memory	70.51	70.01	69.92	69.83
MVAN	72.98	72.98	72.36	72.30
CLMLF	75.33	73.46	72.00	69.83
ITIN	75.19	74.97	73.52	73.49
CLIP-CA-CG (Ours)	**75.38**	**75.21**	**73.95**	**73.83**

As indicated in Table III, the proposed CLIP-CA-CG model achieves the best performance compared to other benchmark models on both the MVSA-S and MVSA-M datasets. This suggests that our model can effectively exploit the correlations between different modalities. Additionally, by preprocessing the model, we can effectively reduce the difficulty of model training. Finally, the model considers the adjustment of weights based on joint features between modalities and environmental features, thus achieving more accurate sentiment classification.

Compared to the SentiBank and SentiStrength models, both SentiBank and SentiStrength models exhibit inferior overall performance. This is attributed to the conventional feature statistics often failing to comprehensively encapsulate the intrinsic features of multimodal information, leading to missing or erroneous feature information inputted into the model, consequently resulting in inaccurate model predictions.

The CNN-Multi, DNN-LR, and Co-Memory models all utilize deep learning for feature extraction, which facilitates the extraction of data features. It is noteworthy that the Co-Memory model introduces a fusion module into sentiment analysis, resulting in a significant improvement in model accuracy. This suggests that effectively integrating image and text features is a viable approach for enhancing the accuracy of multimodal sentiment analysis. Although this approach can learn invariant or specific representations across multiple modalities, it also brings about issues such as excessively redundant feature representations, thereby affecting the effectiveness of fused features.

The MVAN, CLMLF, and ITIN models all incorporate attention mechanisms, which, as observed from the results, further enhance model performance. This indicates that attention mechanisms can focus on more valuable and contributory features. Additionally, considering issues such as feature fusion across modalities and feature redundancy, methods such as contrastive learning and adaptive cross-modal gating fusion have also, to some extent, improved model performance.

Building upon the strengths and weaknesses of baseline models, the proposed CLIP-CA-CG model first enhances the representation capability of image-text data by leveraging pre-trained vision and language models along with contrastive learning techniques. Concurrently, it incorporates a multi-head attention mechanism to capture and express image-text features at a finer granularity. Finally, by exploiting the interaction between images and text, the model utilizes a fusion interaction module to extract both global and focal features of image-text features. These features are complemented with environmental features for more accurate sentiment prediction. Experimental results demonstrate superior performance across public datasets.

D. Ablation Experiment

To validate the performance improvement of each module in multimodal sentiment analysis, we conduct a series of experiments focusing on image and text feature extraction methods, feature fusion methods, etc., to verify the effectiveness of the CLIP-CA-CG model. The details of the model ablation experiments are explained below.

- V_{only} and T_{only}: Represent the evaluation of sentiment analysis using only the visual modality and only the text modality, respectively.
- CLIP-CA-CG w/o Clip: Remove the Clip image-text contrastive model from the complete model, eliminate further feature extraction and fusion, and directly input the preliminary extracted image features and text features into the multi-head attention module.
- CLIP-CA-CG w/o CA: Remove the multi-head attention mechanism from the complete model, and directly input the obtained joint features along with the image and text features into the fusion module.
- CLIP-CA-CG w/o CG: Remove the cross-modal interaction fusion module from the complete model. Instead, use a simple concatenation method to combine multimodal data and process the fused features with an encoder.

TABLE IV. Ablation Experiments on MSVA-Single Dataset

Model	Accuracy (%)	F1 (%)
V_only	63.04	62.76
T_only	71.87	71.19
CLIP-CA-CG w/o Clip	73.65	73.36
CLIP-CA-CG w/o CA	72.15	71.56
CLIP-CA-CG w/o CG	72.41	71.98
CLIP-CA-CG (Ours)	**75.38**	**75.21**

According to the experimental settings, we conduct ablation experiments on the MSVA-Single dataset. As shown in Table IV, proposed CLIP-CA-CG model performs the best, and the absence of any modality or module results in a decrease in model performance. The V_{only} and T_{only} models, which extract features and make sentiment judgments using only a single modality, have the lowest accuracy compared to other experiments. The accuracy of the text model is 71.87, while the accuracy of the image model is only 63.04. This indicates that in the field of sentiment analysis, text has a stronger expressive capability than images. Additionally, incorporating multimodal features can complement information, improving the performance of sentiment analysis models. This provides a solid foundation for subsequent multimodal fusion experiments.

CLIP-CA-CG w/o Clip, CLIP-CA-CG w/o CA, and CLIP-CA-CG w/o CG models respectively remove the Clip contrastive learning module, the multi-head interaction attention module, and the gate fusion module. The experimental results show that the removal of these three modules led to varying degrees of performance degradation in all evaluation metrics. This indicates that these three modules have a promoting effect on the proposed CLIP-CA-CG model.

Specifically, the CLIP-CA-CG w/o Clip model, lacking the utilization of the CLIP pre-trained model, suffers from partial information interaction loss in the early feature extraction, affecting the model's feature fusion to some extent. The CLIP-CA-CG w/o CA model, due to the removal of the attention mechanism, hinders the effective capture of complex relationships between images and text. It fails to extract information components between modalities, making it challenging to ensure the model's robustness at a fine-grained level. The CLIP-CA-CG w/o CG model, obtaining fusion features through direct concatenation, often experiences information loss, redundancy, and noise, leading to a reduction in model accuracy.

V. Conclusion

Addressing the challenges of insufficient inter-modal information, information redundancy, and low effectiveness of fused features in existing multi-modal sentiment analysis, this paper proposes a cross-modal sentiment model, CLIP-CA-CG. The paper first elaborates on the overall architecture of the CLIP-CA-CG model. This model utilizes pre-trained RoBERTa and ResNet50 models to extract textual and visual features. Subsequently, the obtained features are further processed through CLIP contrastive learning to acquire deeper-level features. The model then employs multi-head attention mechanisms and cross-modal fusion modules for global feature, fine-grained feature, and contextual feature extraction, ultimately the control feature weights are input to the fully connected layer for sentiment analysis. In the experimental setup, this paper conducts comparative experiments and ablation experiments with several commonly used multi-modal sentiment analysis models on the public datasets MSVA-Single and MSVA-Multiple. The experimental results show that the accuracy of the CLIP-CA-CG model reaches 75.38% and 73.95%, and the F1 score reaches 75.21% and 73.83%, respectively, validating the generalization and robustness of the CLIP-CA-CG model.

The paper also has some limitations. Due to constraints on data resources, we did not further validate the robustness of the model using other publicly available datasets. Additionally, only two modalities, namely image features and text features, were utilized for experimentation, which might lead to misjudgment in complex scenarios. In future research, we intend to incorporate more modalities to form a more sophisticated multi-modal sentiment analysis model, aiming to further improve the accuracy and generalization of sentiment analysis.

Acknowledgment

This paper is supported by the School of Computer Science, Zhejiang Sci-Tech University. Finally, the authors sincerely thank and respect those who helped complete the work of this article and sincerely hope that many other great studies or projects will result in the future.

References

[1] "Digital 2023: Global overview report," https://datareportal.com/reports/digital-2023-global-overview-report, 2023.

[2] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[3] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of The Web Conference 2020*, 2020, pp. 2514–2520.

[4] Y. Gao, Y. Zhen, H. Li, and T.-S. Chua, "Filtering of brand-related microblogs using social-smooth multiview embedding," *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 2115–2126, 2016.

[5] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[6] Q. T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, and A. Rehman, "Sentiment analysis using deep learning techniques: a review," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017.

[7] L. Alhaidari, K. Alyoubi, and F. Alotaibi, "Detecting irony in arabic microblogs using deep convolutional neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[8] N. Xu, "Analyzing multimodal public sentiment based on hierarchical semantic attentional network," in *2017 IEEE international conference on intelligence and security informatics (ISI)*. IEEE, 2017, pp. 152–154.

[9] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 929–932.

[10] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1424–1437, 2021.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[12] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[13] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[14] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[15] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Coling 2010: Posters*, 2010, pp. 36–44.

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*, 2002.

[17] J. Song, K. T. Kim, B. Lee, S. Kim, and H. Y. Youn, "A novel classification approach based on naïve bayes for twitter sentiment analysis," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 11, no. 6, pp. 2996–3011, 2017.

[18] S. Naz, A. Sharan, and N. Malik, "Sentiment classification on twitter data using support vector machine," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018, pp. 676–679.

[19] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[20] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[21] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1422–1432.

[22] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[23] R. Ni, X. Liu, Y. Chen, X. Zhou, H. Cai, and L. C. Kiong, "Negative emotions sensitive humanoid robot with attention-enhanced facial expression recognition network," *INTELLIGENT AUTOMATION AND SOFT COMPUTING*, vol. 34, no. 1, pp. 149–164, 2022.

[24] H.-J. Yang, G.-S. Lee, S.-H. Kim *et al.*, "End-to-end learning for multimodal emotion recognition in video with adaptive loss," *IEEE MultiMedia*, vol. 28, no. 2, pp. 59–66, 2021.

[25] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[26] B. Büyüköz, A. Hürriyetoğlu, and A. Özgür, "Analyzing elmo and distilbert on socio-political news classification," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, 2020, pp. 9–18.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[28] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83–92.

[29] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 223–232.

[30] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *Proceedings of the second international workshop on issues of sentiment discovery and opinion mining*, 2013, pp. 1–8.

[31] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network." in *IJCAI*, 2017, pp. 3266–3272.

[32] S. Ruan, K. Zhang, L. Wu, T. Xu, Q. Liu, and E. Chen, "Color enhanced cross correlation net for image sentiment analysis," *IEEE Transactions on Multimedia*, 2021.

[33] J. Zhang, X. Liu, M. Chen, Q. Ye, and Z. Wang, "Image sentiment classification via multi-level sentiment region correlation analysis," *Neurocomputing*, vol. 469, pp. 221–233, 2022.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] D. Cao, R. Ji, D. Lin, and S. Li, "A cross-media public sentiment analysis system for microblog," *Multimedia Systems*, vol. 22, pp. 479–486, 2016.

[36] Y. Yu, H. Lin, J. Meng, and Z. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 2, p. 41, 2016.

[37] Z. Zhao, H. Zhu, Z. Xue, Z. Liu, J. Tian, M. C. H. Chua, and M. Liu, "An image-text consistency driven multimodal sentiment analysis approach for social media," *Information Processing & Management*, vol. 56, no. 6, p. 102097, 2019.

[38] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2020.

[39] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[40] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," *Knowledge-Based Systems*, vol. 167, pp. 26–37, 2019.

[41] Y. Liu, X. Zhang, Q. Zhang, C. Li, F. Huang, X. Tang, and Z. Li, "Dual self-attention with co-attention networks for visual question answering," *Pattern Recognition*, vol. 117, p. 107956, 2021.

[42] G. S. Cheema, S. Hakimov, E. Müller-Budack, and R. Ewerth, "A fair and comprehensive comparison of multimodal tweet sentiment analysis methods," in *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*, 2021, pp. 37–45.

[43] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[44] S. Gupta, A. Shah, M. Shah, L. Syiemlieh, and C. Maurya, "Filming multimodal sarcasm detection with attention," in *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28*. Springer, 2021, pp. 178–186.

[45] Z. Zhang, Z. Dong, H. Lin, Z. He, M. Wang, Y. He, X. Gao, and M. Gao, "An improved bidirectional gated recurrent unit method for accurate state-of-charge estimation," *IEEE Access*, vol. 9, pp. 11 252–11 263, 2021.

[46] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*. Springer, 2016, pp. 15–27.

[47] D. Borth, T. Chen, R. Ji, and S.-F. Chang, "Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 459–460.

[48] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis," in *Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings 4*. Springer, 2015, pp. 159–167.

[49] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[50] Z. Li, B. Xu, C. Zhu, and T. Zhao, "Clmlf: a contrastive learning and multi-layer fusion method for multimodal sentiment detection," *arXiv preprint arXiv:2204.05515*, 2022.

[51] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, 2022.