electronics

MDPI

Article

Multi-Modal Sentiment Analysis Based on Image and Text Fusion Based on Cross-Attention Mechanism

Hongchan Li, Yantong Lu and Haodong Zhu *

School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 2011017@zzuli.edu.cn (H.L.); 332107040623@email.zzuli.edu.cn (Y.L.)
* Correspondence: zhdzzuliketizu@163.com

Abstract: Research on uni-modal sentiment analysis has achieved great success, but emotions in real life are mostly multi-modal; there are not only texts but also images, audio, video, and other forms. The various modes play a role in mutual promotion. If the connection between various modalities can be mined, the accuracy of sentiment analysis will be further improved. To this end, this paper introduces a cross-attention-based multi-modal fusion model for images and text, namely, MCAM. First, we use the ALBert pre-training model to extract text features for text; then, we use BiLSTM to extract text context features; then, we use DenseNet121 to extract image features for images; and then, we use CBAM to extract specific areas related to emotion in images. Finally, we utilize multi-modal cross-attention to fuse the extracted features from the text and image, and we classify the output to determine the emotional polarity. In the experimental comparative analysis of MVSA and TumEmo public datasets, the model in this article is better than the baseline model, with accuracy and F1 scores reaching 86.5% and 75.3% and 85.5% and 76.7%, respectively. In addition, we also conducted ablation experiments, which confirmed that sentiment analysis with multi-modal fusion is better than single-modal sentiment analysis.

Keywords: multi-modal sentiment analysis; ALBert; CBAM; DenseNet121; deep learning; feature extraction

Citation: Li, H.; Lu, Y.; Zhu, H. Multi-Modal Sentiment Analysis Based on Image and Text Fusion Based on Cross-Attention Mechanism. *Electronics* **2024**, *13*, 2069. https://doi.org/10.3390/electronics13112069

Academic Editors: Junaid Rashid and Patrick Siarry

Received: 22 April 2024
Revised: 21 May 2024
Accepted: 23 May 2024
Published: 27 May 2024

1. Introduction

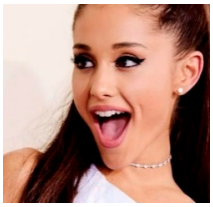
With the rapid development of social media, there are more and more data covering multi-modal social interaction, both domestic, such as via Weibo and Tieba, and foreign, such as via Twitter and Flickr. Sentiment analysis obtains people's views, opinions, attitudes, and emotions by analyzing text [1], audio, video [2], and images [3]. For example, product-based sentiment analysis [4] can judge users' emotional tendencies towards products and brands, and companies can use this information to improve product quality and brand image, so it is favored by many consumers and e-commerce websites. Social media sentiment analysis [5] can assist the government in comprehending public opinions or stances regarding significant events or trending topics.

However, in real life, information is often not uni-modal, and text and images generally appear in pairs. Therefore, integrating multi-modal data, such as images and text, for sentiment analysis is a challenging task and a popular research topic. As shown in Table 1, several groups of social data on Twitter come from the dataset MVSA, including images and their corresponding text descriptions. From the table, we can see that due to the different data forms, the images and their corresponding texts may express the same sentiment or different sentiments.

For example, in the first set of data, we can see that the facial expression of the person in the image is smiling, and their emotion is positive. In addition, the word "Ecstatic" in the text also obviously expresses positive emotion. Therefore, for data where the image and text express the same emotion, they can complement each other, and the most accurate prediction of the emotion can be effectively made by extracting the emotional words and

emotional regions, respectively. In the second group, the smiling faces of the characters in the images expressed positive emotions, but no words with obvious emotional tendencies were found in the text. The images played a leading role in the overall emotions. In the third set of data, a bottle of beer is shown in the image. We cannot mine its emotional tendency from the image, but the meaning of the word "ebullient" in the text expresses positive emotions, so the text plays a leading role in predicting the overall emotion. For this, we need to learn the overall emotional orientation of combining text and image content.

Table 1. Twitter social data.

Image	Text	Image Sentiment	Text Sentiment
	Ecstatic to be holding the first print copy of Wind In Your Sails. #biz #author	Positive	Positive
	RT @FreddyAmazin: When someone insults you & they think you actually care	Positive	Neutral
	This beer goes well with Ghost Bath's 'Moonlover'. Some ebullient stuff going on, just wish the vox were cranked up.	Neutral	Positive

Multi-modal sentiment analysis has garnered increasing attention with the advancement of research in this area. Different fusion methods can be categorized into three groups: early fusion [6,7], intermediate fusion [8,9], and late fusion [10,11]. Early fusion is mainly achieved through feature extraction of multi-modal information and then through splicing, weighting, and other methods of fusion, but the fusion output may include a significant amount of redundant vectors, resulting in information redundancy and information dependence, resulting in poor fusion effect. Intermediate fusion is mainly realized through the neural network, sharing the middle layer of the shared network during the fusion process. Late fusion trains each modality to choose the best fit for it, uses different classifiers to make predictions, and then performs decision fusion. However, late fusion cannot well coordinate the correlations among the various modalities.

This paper presents a multi-modal sentiment analysis model that uses a cross-attention mechanism for image–text fusion. The proposed model leverages the complementarity and relevance between images and texts. The main method is to use the attention mechanism to extract emotional regions and emotional vocabulary from images and texts; then, we use the cross-attention mechanism to perform feature fusion on the extracted emotional features; finally, the fused features are passed through a classifier to output prediction results. This paper's primary contributions are as follows:

1. For text data, initially, the ALBert pre-training model is utilized to convert text into vectors; text context features are then obtained using BiLSTM.

2. We first use the DenseNet121 network to extract features from the image, and then we use the CBAM mechanism to obtain the corresponding emotional regions from two aspects of channel and space.
3. For the acquired emotional vocabulary and emotional region features, the cross-attention mechanism is used for feature fusion, and the resulting output is obtained through a softmax classifier.

The paper is organized as follows: Section 2 reviews relevant prior research; Section 3 outlines the necessary theoretical foundations; Section 4 introduces the proposed model; Section 5 presents experimental results and analysis; Section 6 provides a conclusion, summarizes the findings, and proposes future research directions.

2. Related Work

This section reviews sentiment analysis, focusing on three distinct research objects.

2.1. Sentiment Analysis of Text

Text sentiment analysis has achieved great success and is widely used in public opinion monitoring and product reviews. Text sentiment analysis research primarily concentrates on three aspects: the sentiment dictionary method, the machine learning method, and the deep learning method.

The method based on the sentiment lexicon obtains the sentiment value of the sentiment words in the document and then weights the calculation to determine the overall sentiment tendency of the document. Zargari H et al. [12] proposes a sentiment lexicon method using N-Gram is proposed, which incorporates global intensifiers to expand the emotional phrase dictionary's coverage. The method considers the relationship between multiple intensifiers and emotional words. Xu G et al. [13] constructs a sentiment dictionary that includes basic sentiment, scene sentiment, and polysemous words. The method utilizes machine learning to extract features and output them into a classifier. Goel A et al. [14] employs the naive Bayesian algorithm for sentiment analysis. Rathor A S et al. [15] compares three machine learning algorithms, SVM, NB, and ME, and achieves good classification results. Deep learning has gained popularity as a text sentiment analysis method, similar to its success in the domain of computer vision. Zhou X et al. [16] employs the long short-term memory network (LSTM), which is utilized for sentiment analysis in various languages. Sun C et al. [17] further improves the accuracy of the results by utilizing the pre-trained Bert model. Miao et al. [18] proposes a CNN-BiGRU model that combines the convolutional neural network and the gating mechanism and achieves favorable outcomes. Yenduri et al. [19] proposed a new customized BERT-type sentiment classification method, which consists of two main phases—preprocessing and tokenization, as well as a classification method based on the "Customized Bi-directional Encoder Representation Transformer (BERT)"—and experimentally demonstrated the enhancement effect of the proposed model. Cauteruccio F et al. [20] proposed a social-network-based model and topic analysis technology to study the emotional aspects of e-sports viewing. The research method is universal and can study the audience identity of e-sports from a heterogeneous perspective.

2.2. Sentiment Analysis of Images

Image processing is a prominent area of research in computer vision, and images can bring more visual impact than text and contain richer semantic information. Therefore, sentiment analysis of images has attracted more and more researchers' interest.

Early sentiment analysis on images mainly focused on low-level features, such as the shape and color of the image. This method mainly relied on people's manual annotation, and the effect was not good. As the research progressed, researchers discovered middle-level properties of image sentiment analysis. Yuan J et al. [21] proposes an image sentiment prediction framework that incorporates facial expression detection. D. Borth et al. [22] propose visual entities or attributes as features for image sentiment analysis. However, these middle-level attributes rely on extensive knowledge of psychology or linguistics and

require human intervention to fine-tune the emotional prediction results, leading to less accurate predictions. The advent of deep learning has brought about higher-level semantic features that are widely used in image emotion analysis. He X et al. [23] introduces an attention mechanism that focuses on areas related to emotion.

2.3. Sentiment Analysis of Image and Text Fusion

Multi-modal sentiment analysis has received more and more attention in recent years, and it is also a very challenging research topic. Multi-modal sentiment analysis integrates multiple fields, such as natural language processing, computer vision, and more, whereas single-modal sentiment analysis does not. It is an interdisciplinary research area.

Based on the early fusion method, the feature extraction of various modal information is first performed, and then the fusion is carried out via splicing, weighting, and other methods. Wang M et al. [24] performs feature splicing by fusing text and images into a unified bag of words to output the final representation. Zhang Y et al. [25] extracts text features using binary representation and utilizes the interactive information method to extract the underlying image features. The method performs binary classification on the resulting similarity-based neighborhood classifier. Late fusion trains the data of each modality separately, selects the most appropriate classifier, and outputs the final fused result. Yu Y et al. [26] initially extracts image and text features independently using CNN. The method then employs logistic regression to predict and analyze different emotions. Finally, an average and weighted fusion strategy is utilized to perform the final emotion prediction analysis. Kumar A et al. [27] proposes a hybrid deep learning model that initially performs fine-grained analysis on multi-modal data and then utilizes a decision-level multi-modal combination to classify and output the data. Xu J et al. [28] proposes a new bidirectional multi-modal attention model to analyze the complementarity and correlation between images and text at both levels.

3. Background

3.1. DenseNet Network

DenseNet (densely connected convolutional network) is a deep convolutional neural network model proposed by Gao Huang et al. [29] in 2017. It uses the idea of dense connection to make the network have stronger feature transfer. In addition to its reusability, deep learning can attain high accuracy by training very deep neural networks. In a traditional convolutional neural network, the output of each layer is computed by convolving the previous layer's input using a nonlinear activation function.

Assuming that the input is an image X_0, after an L-layer neural network, the i-th layer's nonlinear transformation is represented as $H_i(*)$, which can comprise several functional operations, including BN, ReLU, Pooling, or Conv. The i-th layer's output feature is denoted as X_i.

In a conventional convolutional feed-forward neural network, the output X_i of the i-th layer serves as the input of the $i+1$ layer and can be represented as $X_i = H_i(X_{i-1})$. However, ResNet introduces a bypass connection, which can be written as the following Formula (1):

$$X_l = H_l(X_{l-1}) + X_{l-1}. \quad (1)$$

A key benefit of ResNet is that it allows the gradient to flow through the identity function to reach the previous layer. However, the use of addition to combine the identity mapping and nonlinear transformation output in the layer stacking process can potentially disrupt the information flow in the network. And DenseNet proposes the concept of dense block. Each dense block contains several convolutional layers and a skip connection so that the features of all the previous layers can be directly transferred to the subsequent layers, thus forming a dense network. A dense block consists of several convolutional layers and a skip connection, as shown in Figure 1 below.

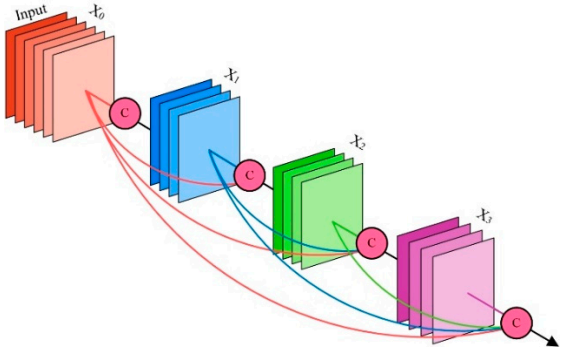


Figure 1. The computational process of the attention mechanism. DenseNet network dense link mechanism (where *C* stands for channel-level connection operation).

In DenseNet, each layer's input is the concatenation of all the preceding layer's feature maps generated within the same dense block. The output of layer $(l-1)$ is denoted as X_{l-1}. As depicted in Figure 1, the input of the i-th layer depends on both the output of the $i-1$ layer and the output of all previous layers. The output of the l-th layer can be expressed as follows using Equation (2):

$$X_l = H_l([X_0, X_1, \cdots, X_{l-1}]). \tag{2}$$

Among them, [] operator represents concatenation, that is, all output feature maps of layers X_0 to X_{l-1} are combined by Channel. The nonlinear transformation function $H_l(\cdot)$ used here is a combination of BN + ReLU + Conv(3×3).

Each convolutional layer's input includes the output from all previous layers, and the output will be directly passed to all subsequent layers. If the input and output dimensions are different, it needs to be transformed by an additional convolutional layer so that the input and output can be concatenated. The advantage of using dense blocks is that it can make the network have stronger feature transfer and reuse capabilities, thereby improving accuracy. In addition, skip connections can also make the network have stronger feature reuse ability and generalization ability, which further improves the accuracy.

3.2. ALBert Pre-Trained Model

ALBert [30] (a little Bert) is a natural language processing model developed by Alibaba DAMO Academy and based on the Bert model. The model is pre-trained for self-supervised learning and can undertake various natural languages processing tasks, such as text classification, named entity recognition, question answering, and text generation.

The structure of ALBert is similar to that of Bert. It is composed of multiple transformer modules. Each transformer module comprises a multi-head self-attention layer and a feed-forward neural network layer. Different from Bert, ALBert uses cross-layer connections and global pathways to enhance the ability of feature transfer and information flow, thereby improving the efficiency and performance of the model.

The main advantages of ALBert over Bert are as follows:

1. More efficient training: ALBert uses two new training strategies. ALBert uses cross-layer parameter sharing as a training strategy, which significantly reduces the number of model parameters and simplifies the model training process. Sentence order prediction enhances the model's generalization ability and augments the training data.
2. Better performance: ALBert achieved better results than Bert in the GLUE (general language understanding evaluation) benchmark evaluation task, indicating that ALBert has better performance in natural language understanding tasks than Bert.
3. Better generalization ability: ALBert uses more sufficient pre-training, which can better learn general language representation and thus has better generalization ability in downstream tasks.

4. More flexible model structure: ALBert provides a variety of different model structures and hyperparameters, which can be adjusted flexibly according to specific application scenarios.

In short, compared with Bert, ALBert has improved in terms of training efficiency, performance, generalization ability, and model flexibility, so this paper chooses ALBert as the pre-training model.

3.3. CBAM Attention Mechanism

CBAM (convolutional block attention module) is an attention mechanism module for image recognition, proposed by Sanghyun Woo et al. [31] of the KAIST Machine Learning Research Center in 2018. The CBAM module enhances the accuracy and robustness of image recognition in convolutional neural networks by introducing spatial and channel attention mechanisms.

Specifically, the CBAM module includes two attention sub-modules: the channel attention module and the spatial attention module. The channel attention module dynamically adjusts the weights of various channels to enhance the model's focus on essential features. The spatial attention module is used to adaptively change the weights of different spatial locations to enhance the model's attention to important locations. Through the combination of these two sub-modules, the CBAM module enables dynamic weight adjustment of both channel and position within the feature map in an adaptive manner, thereby alleviating the dependence on global features and improving the robustness and generalization ability of the model.

CBAM takes an intermediate feature map $F \in R^{C\times H\times W}$ as its input and typically consists of two operational stages. To generate channel attention $M_C \in R^{C\times 1\times 1}$, the input undergoes global maximum and mean pooling by channel. The resulting one-dimensional vectors are then sent to a fully connected layer and added together, and to obtain the adjusted feature map F', we multiply the channel's attention with the input elements. Next, we perform global maximum and mean pooling on F' based on the spatial dimensions. We concatenate the resulting two-dimensional vectors and apply a convolution operation to generate a two-dimensional spatial attention $M_S \in R^{1\times H\times W}$. Finally, we multiply the element-wise product of the spatial attention and F' to obtain the adjusted feature map. The specific process is shown in Figure 2. CBAM generates attention the process as follows:

$$F' = M_C(F) \bigotimes F, \tag{3}$$

$$F'' = M_S(F') \bigotimes F'. \tag{4}$$

Among them, we use $\otimes$ to denote the element-wise multiplication of corresponding elements. Before performing the multiplication operation, we broadcast the channel attention and spatial attention based on the channel and spatial dimensions, respectively.

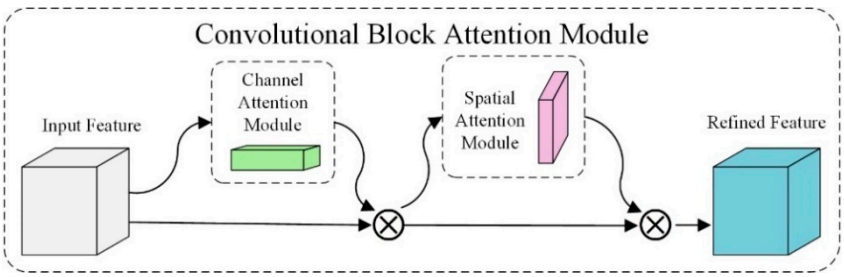


Figure 2. CBAM attention model structure.

4. Multi-Modal Sentiment Analysis Model Based on Cross-Attention Mechanism

Based on the above analysis, this paper proposes the multi-cross-attentive model (MCAM), which is a multi-modal model capable of processing multiple types of data

(such as text, images, audio, etc.). Traditional sentiment analysis models can only use one or several data types for sentiment analysis and cannot make full use of the interaction between different data types. MCAM, on the other hand, can process multiple data types simultaneously, and it employs the cross-attention mechanism to learn the interaction between diverse data types, enhancing sentiment analysis's accuracy and robustness.

Figure 3 illustrates the architecture of the model proposed in this paper. First, we utilize ALBert and DenseNet121 to extract vectorized features from text and images, respectively; then, the text features are processed by BiLSTM to obtain the context features containing emotional words, and the CBAM attention is obtained from the two aspects of space and channel, respectively. In the area of emotional characteristics; the single-modal attention mechanism considers the relationship within the modality, and the cross-attention mechanism can consider the relationship between the two modalities, fully considering the complementarity and relevance between different modalities. Utilizing the cross-attention mechanism, we fuse the emotional features extracted from both text and images. The final sentiment analysis result is then obtained using a softmax classifier.

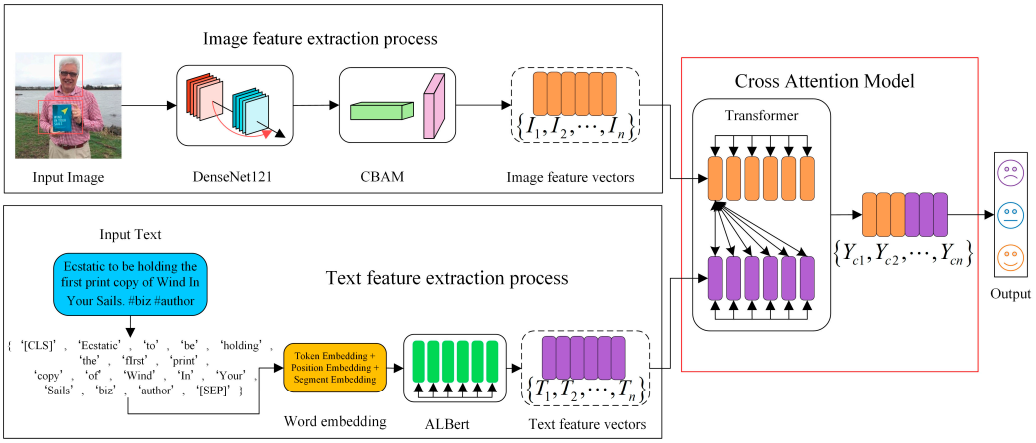


Figure 3. The MCAM model structure proposed in this paper.

4.1. Image Feature Extraction

In an image, usually, certain regions can better reflect the emotional tendency of the whole image. In these regions where the most emotional characteristics can be mined, the results of sentiment analysis will be more accurate. You Quanzeng et al. [32] proposes an attention mechanism to detect emotion-related local areas and creates an emotion classifier based on these regions for image sentiment analysis. This paper utilizes the pre-trained DenseNet121 network model to extract image features. CBAM attention is then employed to extract the most emotional region in the image from both the spatial and channel dimensions, enhancing the expressive capabilities of both overall and local features. Figure 4 illustrates the image features extraction process.

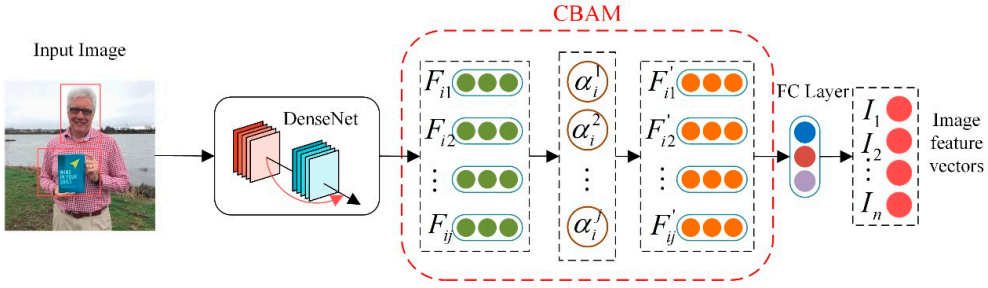


Figure 4. Image feature extraction process.

Let $X = \{X_1, X_2, \cdots, X_n\}$ denote a dataset with n images. For each image X_i, we use the DenseNet network to preprocess the image of the input layer. Next, the image undergoes the CBAM attention mechanism, and its feature vector I_i, is extracted.

The feature map obtained by the i-th image after the j-th layer of convolution is $F_{ij} = \{F_{1j}, F_{2j}, \cdots, F_{nj}\} \in \mathbb{R}^{C \times H \times W}$, where C represents the number of channels, while H and W denote the feature map's length and width, respectively. F'_{ij} is the resulting attention feature map after applying attention weights. The attention weight α_i^j for the j-th feature map of the i-th image can be computed using Formula (5), which is expressed as follows:

$$\alpha_i^j = \{M_c(F_{ij}), M_s(F_{ij})\}. \tag{5}$$

$M_c(F_{ij})$ and $M_s(F_{ij})$, respectively refer, to the channel and spatial attention weights of the j-th feature map of the i-th image.

Channel attention focuses on which feature on which channel is meaningful, specifically in terms of the contribution of each feature of the convolutional feature map to key information. Formulae (6) and (7) compute the channel attention weight $M_c(F_{ij})$ as follows:

$$M_c(F_{ij}) = \sigma(MLP(AvgPool(F_{ij})) + MLP(MaxPool(F_{ij}))), \tag{6}$$

$$M_c(F_{ij}) = \sigma(W_1(W_0(AvgPool(F_{ij}))) + W_1(W_0(MaxPool(F_{ij})))). \tag{7}$$

Here, σ is the sigmoid activation function; the weights W_0 and W_1 of the MLP are shared between the two inputs and are followed by the *ReLu* activation function; $AvgPool(\cdot)$ and $MaxPool(\cdot)$, respectively, calculate the mean and maximum feature values within each feature map in a global manner.

The input feature map is represented by $F_{ij} \in \mathbb{R}^{C \times H \times W}$. First, we obtain two feature maps through global average and maximum pooling, with each result having a shape of $\mathbb{R}^{C \times 1 \times 1}$. Next, each of the two feature maps is fed into a two-layer fully connected neural network that shares parameters. The next step is to add the two feature maps together, utilizing the sigmoid function to obtain a weight coefficient between 0 and 1. Multiplying the weight coefficient with the input feature map produces the final output feature map.

Spatial attention emphasizes the significance of specific features within a given space, particularly the local regions of an image that contribute essential information. This process effectively identifies areas within the image that warrant attention. Calculation of the spatial attention weight, denoted by $M_s(F_{ij})$, is performed using the following Formula (8):

$$M_s(F_{ij}) = \sigma\Big(f^{7\times7}([AvgPool(F_{ij}), MaxPool(F_{ij})])\Big). \tag{8}$$

Among them, the σ activation function represents sigmoid, while the operation of connection is denoted by $[\cdot]$; the convolution operation, represented by $f^{7\times7}(\cdot)$, assesses the impact of critical information from distinct local regions of the feature map. The convolution kernel size is 7×7; the average pooling function is denoted by $AvgPool(\cdot)$, while the maximum pooling function is denoted by $MaxPool(\cdot)$.

The input feature map, denoted by F_{ij} and having a shape of $\mathbb{R}^{C \times H \times W}$, undergoes maximum and average pooling along a single channel dimension, resulting in two feature maps of $\mathbb{R}^{1 \times H \times W}$. The two feature maps are concatenated in the channel dimension, resulting in a feature map of size $\mathbb{R}^{2 \times H \times W}$, which is passed through a convolutional layer to reduce it to 1 channel. The convolutional kernel, which maintains the size of $H \times W$ while using a 7×7 filter, produces an output feature map of $\mathbb{R}^{1 \times H \times W}$. The sigmoid function generates a spatial weight coefficient, which is used to multiply the output feature map. The final feature map is obtained by multiplying the input feature map with the spatial weight coefficient. Formula (9) represents the calculation formula for the CBAM attention feature map, as follows:

$$F'_{ij} = F_{ij} \bigotimes M_c(F_{ij}) \bigotimes M_s(F_{ij}). \tag{9}$$

Among them, $\otimes$ represents the bitwise multiplication of corresponding elements.

Finally, after the convolution operation, the feature map vector $I = [I_1, I_2, \cdots, I_n]$ of the key region of each image is obtained.

4.2. Text Feature Extraction

It is commonly understood that emotional information in an input sentence is often associated with specific words within the text. Therefore, text sentiment analysis involves vectorizing the text content using the pre-training model ALBert. Subsequently, BiLSTM [33] is used to optimize the emotional level of the text content by combining the input sequence information in both forward and backward directions. Figure 5 illustrates the feature extraction process of the text.

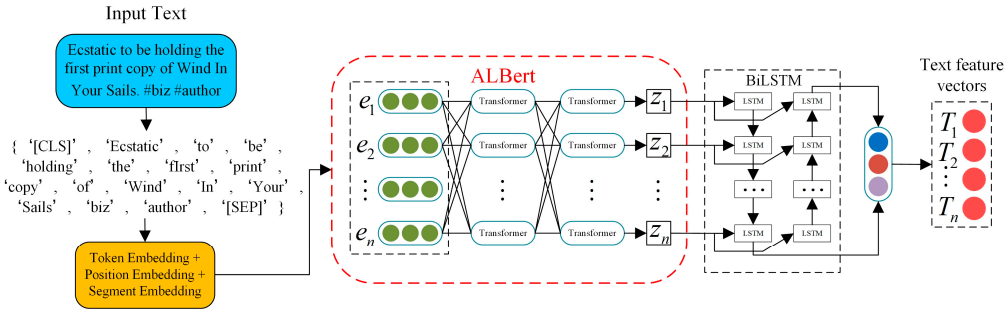


Figure 5. Text feature extraction process.

4.2.1. ALBert Pre-Training Layer

Text sequences are vectorized using the ALBert pre-training model. It performs word segmentation at the character level, enabling vectorization of corpus text. Each input text m consists of n characters, which can be expressed as $m = [m_1, m_2, \cdots, m_n]$; the i-th character in the text is denoted by m_i.

The text m passes through the input layer to the ALBert pre-training layer. First, for each word in the text information, we mark its number position in the dictionary, obtain the corresponding number, and vectorize the text content to obtain the information sequence e, e_i, which represents the vectorizations corresponding to the i-th character, as shown in the following Formula (10):

$$e = [e_1, e_2, \cdots, e_n]. \tag{10}$$

For the information sequence e it is then input to the transformer encoder in the ALBert pre-training model to mine deep semantic feature information, and after conversion, the feature vector z of the final text sequence is obtained, and z_i represents the feature vector corresponding to the i-th character, as in Formula (11):

$$z = [z_1, z_2, \cdots, z_n]. \tag{11}$$

4.2.2. BiLSTM Layer

Taking the feature vector z_i corresponding to each character i as input, the BiLSTM network simultaneously combines the input sequence information in both the forward and backward directions and better mines the deep information of semantics, which can further strengthen the emotional information in the text. For the input z_{it} at time t, its forward output $\overrightarrow{h}_{it}$ and backward output $\overleftarrow{h}_{it}$ are calculated as follows in (12) and (13):

$$\overrightarrow{h}_{it} = \overrightarrow{LSTM}(z_{it}), \tag{12}$$

$$\overleftarrow{h}_{it} = \overleftarrow{LSTM}(z_{it}). \tag{13}$$

For the feature vector z_i of each character i, we obtain its feature representation with emotional information by connecting its forward and backward context information, as shown in Equation (14):

$$h_{it} = \left[\overrightarrow{h}_{it}, \overleftarrow{h}_{it}\right]. \tag{14}$$

Among them, $[\cdot]$ is the splicing operation of vectors.

After calculation, we obtain the output feature vector of each character i of the input text at time t, as shown in the following Formula (15):

$$T_i = \sum_t \alpha_{it} h_{it}. \tag{15}$$

Finally, the feature vector of each input text containing emotional information is $T = [T_1, T_2, \cdots, T_n]$.

4.3. Cross-Attention Fusion of Image and Text Features

This section employs the cross-attention module to model the intermodal relationship between image regions and text words. The cross-attention mechanism utilizes scaled dot product attention to enhance the recognition of emotional features in images and word fragments by fully considering their complementarity and relevance.

The cross-attention mechanism consists of scaled dot-product attention, using text features T_i as the query matrix in scaled dot-product attention and image features I_i as the key-value matrix. By scaling the dot-product attention, the cross-attention features of images and texts are obtained, as shown in Equations (16)–(19):

$$Q = T_i W^Q = \begin{pmatrix} RW^Q \\ EW^Q \end{pmatrix}, \tag{16}$$

$$K = I_i W^K = \begin{pmatrix} RW^K \\ EW^K \end{pmatrix}, \tag{17}$$

$$V = I_i W^V = \begin{pmatrix} RW^V \\ EW^V \end{pmatrix}, \tag{18}$$

$$Y_c = Att(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{19}$$

Among them, W^Q, W^K, and W^V are parameter matrices; d_k is the number of columns of Q and K; and the image features and text features are denoted by I_i and T_i, respectively. The cross-attention mechanism module network is represented by $Att(\cdot)$.

Finally, we obtain the cross feature after cross-attention fusion: $Y_c = [Y_{c1}, Y_{c2}, \cdots, Y_{cn}]$.

The output Y_c resulting from cross-attention fusion is used as input for the linear function softmax, which performs the final sentiment classification. This process is illustrated in the following Formula (20):

$$y = softmax(W_c Y_c + b_c). \tag{20}$$

Here, W_c denotes the weight matrix, while b_c represents the bias term.

5. Experimental Process

The performance of the MCAM model is evaluated in this section through comparative experiments conducted on the MVSA and TumEmo datasets. Additionally, this section presents a qualitative evaluation of the model's performance.

5.1. Dataset Introduction

This article uses two public datasets of graphic multi-modal sentiment analysis: (MVSA raw data obtained from http://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/ accessed on 20 May 2024) MVSA and (TumEmo raw data obtained from https://github.com/YangXiaocui1215/MVAN accessed on 20 May 2024) TumEmo. The MVSA dataset comprises tweets that are in the form of text and images. It is a collection of messages obtained from Twitter. TumEmo is image text sentiment data scraped by Tumblr. Tumblr, whose Chinese name is Tang Bole, is the largest light blogging website in the world. The multimedia content posted by users usually includes images, texts, and other forms of content. These datasets are publicly accessible for image and text multi-modal sentiment analysis.

The MVSA dataset contains 4869 text–image pairs. Each sample contains a set of text and images, and the emotional label is uni-modal. The emotional label of the dataset has only three modes: positive, neutral, and negative. The screening results are presented in Table 2. Different from the MVSA dataset, the TumEmo dataset further subdivides the emotional labels, including angry, bored, calm, happy, love, and sad, into seven types of emotion. The TumEmo dataset contains 195,265 samples, from which this paper obtained 13,852 pieces of data according to the ratio of each emotional label. Tables 3 and 4 display information and content related to the TumEmo dataset.

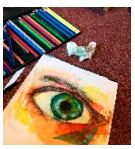
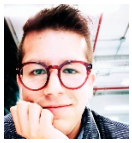
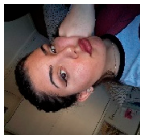
Table 2. MVSA dataset information.

Dataset	Positive	Neutral	Negative	Total
MVSA	2683	470	1358	4511

Table 3. TumEmo dataset information.

Dataset	Angry	Bored	Calm	Fear	Happy	Love	Sad	Total
TumEmo	2167	1667	3577	1136	2476	1975	854	13,852

Table 4. Example of TumEmo dataset content.

Image							
Text	Capaldi demonstrates my feelings at this current moment	Shannon Claire Tillery's Photos–Profile Pictures på We Heart It.	A worker cleans windows at the 124th floor of the Burj Khalifa	I live for the occasional days where i feel well enough to draw.	Beauty is truth! #portrait #photo #demure #lovely #calm	Happy 2018 everyone! #Happy #First-PostOfTheYear	Doing makeup is the most amazing habit i know
Label	Angry	Bored	Calm	Fear	Happy	Love	Sad

The image–text pairs are split into training, test, and validation sets at a ratio of 6:2:2 for each dataset. The experimental environment of the model in this paper is Intel (R) Xeon (R) 2.50 GHz CPU, 40 GB memory, RTX 3080 GPU, Windows 10 OS, and the deep learning-based TensorFlow 2.9.0 architecture is implemented using the Python programming language version 3.8.

5.2. Model Parameter Settings

The input image's shape is (224, 224, 3), where the three dimensions represent its height, width, and number of channels, respectively. The batch input size is 32. In the

CNN layer, we use the pre-trained DenseNet-121 network that has achieved good results in the ImageNet2017 dataset [34] classification challenge. The AveragePooling2D layer utilizes a pooling size of (7, 7), while the Dense layer has output dimensions of 1024/8 and 1024. This results in an output vector with a reduced dimension of 1/8 the original, or 128, and an unchanged dimension of 1024. The aim is to decrease computation by reducing dimensionality while preserving the original feature data.

The text's word embedding layer is initialized using the ALBert pre-training model. The hidden layer dimension is set to 128 to obtain a 128-dimensional vector representation for each word. The input length of the maximum text is derived from our analysis of the dataset, as shown in Figure 6. We found that most of the text lengths are below 150, and the number of texts with text lengths below 150 is also the largest. Therefore, 150 is selected as the maximum length of the input text. For text whose input length is greater than 150, the text will be truncated, and for text whose input length is less than 150, zero-value filling will be performed.

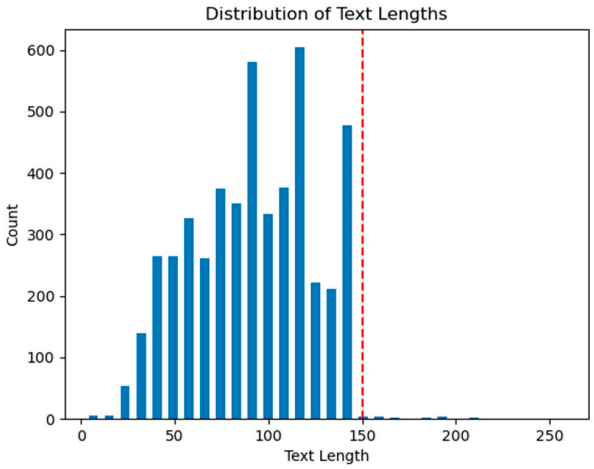


Figure 6. Dataset text length frequency statistics.

The cross-attention mechanism is utilized in the multi-modal fusion part. During the training process, the Adam optimizer, with a learning rate of 0.001, is used to optimize the model parameters; to prevent over-fitting, a random discard rate (Dropout Value) of 0.1 is set in the model, and early stopping technology and L2 regularization are used change. Tenfold cross-validation and grid search are employed to evaluate various parameter combinations. The hyperparameter combinations of the final model are presented in Table 5.

Table 5. Settings of the model hyperparameters in this paper.

Hyperparameter	Hyperparameter Value
Word vector latitude	128
Maximum input text length	150
Image size	224 × 224
Optimizer	Adam
Epochs	10
L2 regularization coefficient	0.1
Dropout value	0.1
Batch size value	32
The pooling size of the AveragePooling2D layer	(7,7)
Dense layer output vector dimension	128, 1024
Learning rate	0.001
Word vector latitude	128
Maximum input text length	150
Image size	224 × 224

5.3. Evaluation Metrics

This paper evaluates the model's performance using standard classification metrics, such as accuracy, precision, recall, and F1-score. True positive (TP), false positive (FP), false negative (FN), and true negative (TN) are the four classification results.

5.3.1. Accuracy

Accuracy is a metric that evaluates the model's ability to make correct predictions for the entire dataset. Formula (21) calculates the ratio of correct predictions to the total number of positive and negative examples, known as accuracy.

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN} \tag{21}$$

5.3.2. Precision

Precision is calculated based on the prediction results and represents the proportion of correct predictions in the samples predicted as positive examples. It is computed using Formula (22):

$$Precision = \frac{TP}{TP + TN}. \tag{22}$$

5.3.3. Recall

Recall is the ratio of correctly predicted positive samples to the total actual positive samples. It is based on the actual samples and is calculated using Formula (23):

$$Recall = \frac{TP}{TP + FN}. \tag{23}$$

5.3.4. F1 Score

The F1 score, which ranges from 0 to 1, is a weighted average of precision and recall. It considers both metrics in the evaluation of a classification model. The Formula (24) illustrates how it is calculated:

$$F1 = \frac{2TP}{2TP + FP + FN}. \tag{24}$$

5.4. Baseline Method

To evaluate the model's generalization ability and robustness, we conducted comparative experiments with the following methods for comparison:

1. Single-modal text model: Hoang et al. [35] uses contextual word representations from the BERT pre-trained language model, fine-tuned with additional generated text, to perform sentiment analysis. Yin Xing et al. [36] utilizes a BiGRU Information Augmented Approach for sentiment analysis.
2. Single-modal image model: Liang Song et al. [37] proposes a method that utilizes the ResNet50 network for image recognition and classification. Paymode A S et al. [38] presents an approach that applies the VGG19 model to classify crop leaf diseases.
3. Multi-modal image–text fusion model: Huang F et al. [39] introduces the deep multi-modal attention fusion (DMAF) model, which performs joint sentiment classification by leveraging the correlation between textual and visual features. Zhu T et al. [40] introduces the image–text interaction network (ITIN) model. By exploring the connection between emotional image regions and text, multi-modal sentiment analysis is performed. Yang X et al. [41] presents a new multi-modal emotional analysis model that employs a continuously updated memory network to extract deep semantic features from image text. Wei K et al. [42] introduces the attention-based modality gating network (AMGN), which detects correlations between different modalities and extracts discriminative features for multi-modal sentiment analysis.

5.5. Experimental Results

A comparative experiment is conducted using the following setup to evaluate the effectiveness of the proposed MCAM model. We compared the single-modal text model, single-modal image model, and multi-modal image–text fusion model, as presented in Tables 6 and 7.

Table 6. Performance comparison of different models on the MVSA dataset.

Methods	Model	Accuracy	Precision	Recall	F1 Score
single-modal text model	Bert	0.742	0.713	0.720	0.693
	BiGRU	0.763	0.772	0.746	0.753
single-modal image model	ResNet-50	0.643	0.701	0.617	0.622
	VGG-19	0.574	0.623	0.572	0.496
multi-modal image–text fusion model	DMAF	0.841	0.803	0.824	0.838
	ITIN	0.763	0.784	0.743	0.732
	MVAN	0.719	0.732	0.732	0.723
	AMGN	0.830	0.834	0.826	0.827
	MCAM	**0.865**	**0.843**	**0.851**	**0.855**

Table 7. Performance comparison of different models on the TumEmo dataset.

Methods	Model	Accuracy	Precision	Recall	F1 Score
single-modal text model	Bert	0.727	0.694	0.703	0.712
	BiGRU	0.734	0.707	0.693	0.698
single-modal image model	ResNet-50	0.593	0.482	0.573	0.558
	VGG-19	0.561	0.527	0.538	0.551
multi-modal image–text fusion model	DMAF	0.739	0.774	0.735	0.751
	ITIN	0.664	0.683	0.703	0.696
	MVAN	0.672	0.694	0.676	0.683
	AMGN	0.728	0.764	0.724	0.746
	MCAM	**0.753**	**0.786**	**0.752**	**0.767**

5.5.1. MVSA Dataset Experimental Results

The MVSA dataset's best performance is exhibited by the proposed MCAM model, as per the analysis of Table 6 and Figure 7. Sentiment classification performance is unsatisfactory for single-modal image and text data, with average classification results. For sentiment analysis that integrates multiple modalities, by learning the correlation between two modalities, the classification effect is greatly improved immediately. In particular, the proposed MCAM model incorporates a cross-attention mechanism to learn distinct modalities, which boosts the prediction accuracy rate by 11.7% and the F1 score by 9.8% compared to the MVAN model. Moreover, the MCAM model outperforms the DMAF model with an accuracy rate increase of 2.4% and an F1 score increase of 1.7%, achieving the best prediction performance.

In addition, our proposed single-modal image and single-modal text models also outperform the baseline methods. In our text model, we incorporate the highly effective pre-training model ALBert and additionally employ BiLSTM. In the image model, we use the DenseNet121 network. The dense connection structure of DenseNet121 makes feature transfer better and gradient flow Smoother, with less risk of over-fitting. With fewer parameters, it can achieve comparable performance to ResNet50 and VGG19, and by introducing the CBAM attention mechanism, attention is extracted from two aspects of channel and space, which further improves the classification effect.

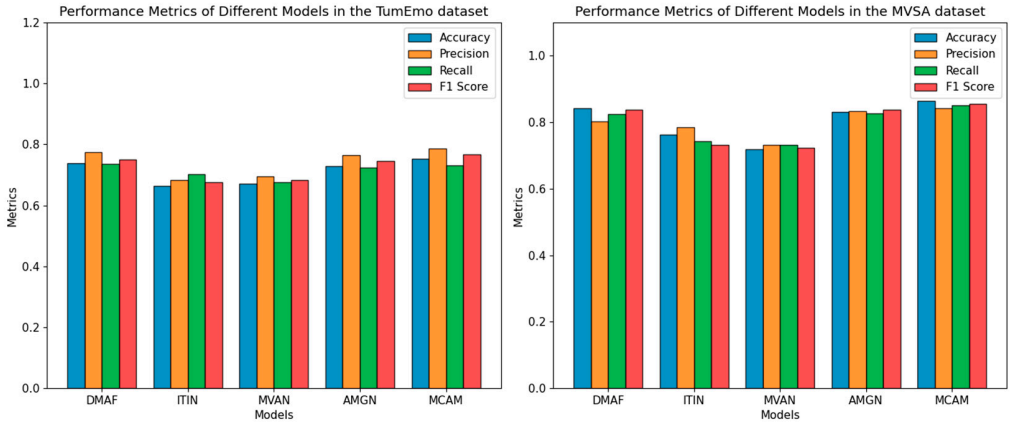


Figure 7. Comparison of performance indicators of different models in MVSA and TumEmo datasets.

To further verify the effectiveness of the MCAM model we proposed, we randomly sample varying proportions of data ranging from 20% to 100% from both the MVSA and TumEmo datasets, and then we observe the accuracy changes in both the MCAM model and the four baseline models. Figure 8 demonstrates that our model consistently outperforms the baseline model of accuracy, regardless of the proportion of sampled training data. This demonstrates our model's absolute competitive advantage and ability to achieve favorable results, even with limited training data.

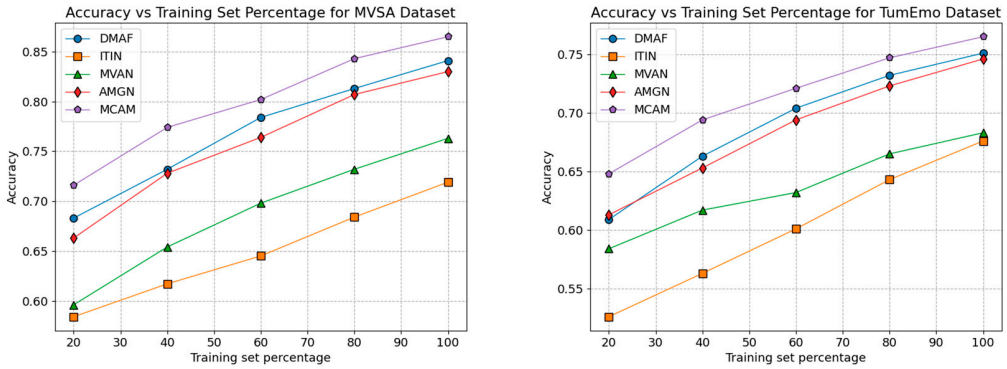


Figure 8. Changes in training sets and accuracy rates of different models in different proportions of MVSA and TumEmo datasets.

5.5.2. TumEmo Dataset Experimental Results

Based on the analysis of Table 7 and Figure 7, the performance of the TumEmo dataset exhibits a considerable drop in comparison to that of the MVSA dataset. The reason for our analysis is that the TumEmo dataset contains an extensive range of emotional categories, consisting of seven types, which reduces the model's performance. In multi-category classification problems, as the number of categories increases, the classifier needs to distinguish more categories, which increases the difficulty of classification, and the corresponding classification effect may decline.

However, the model we proposed still has obvious competitive advantages and has achieved the best results in the evaluation indicators. Although the improvement effect is not obvious compared with AMGN and DMAF, it still has a slight improvement effect. In comparison to AMGN, our proposed model achieves a 1.5% increase in accuracy rate and a 1.9% increase in F1 score. Additionally, when compared to DMAF, our model exhibits a 1.4% increase in accuracy rate and a 1.6% increase in F1 score.

5.5.3. Analysis of the Results of PR Curve in Different Models

As shown in Figure 9, it can be seen from the subplot on the left that the PR curves of the TumEmo dataset present a variety of different shapes, which indicates that different models have different performances when classifying the TumEmo dataset. Notably, the PR curve of the MCAM model exhibits the most optimal shape, signifying superior performance in classifying the TumEmo dataset, while the PR curves of other models, such as ITIN and MVAN, show poor shapes, indicating their relatively poor performance.

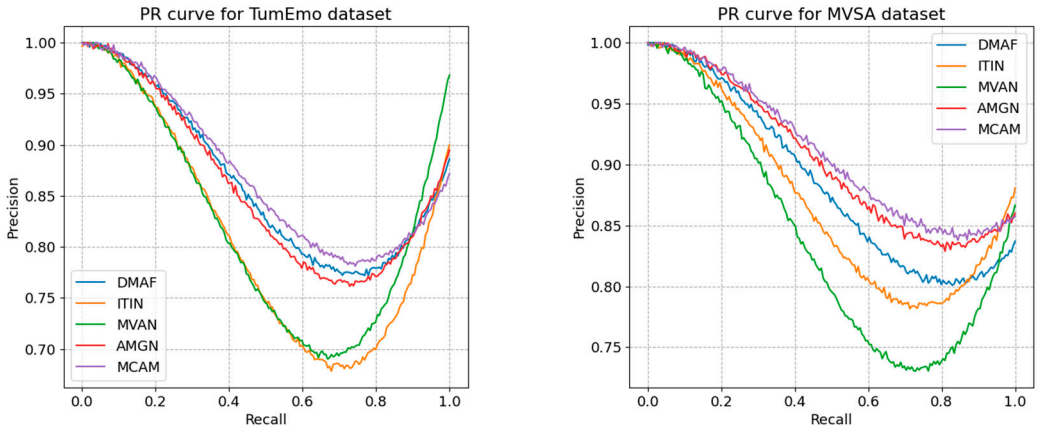


Figure 9. PR curves of MVSA and TumEmo datasets for different classifier models.

As can be seen from the right subplot, the PR curve of the MVSA dataset presents a different shape from that of the TumEmo dataset. The PR curve of model MCAM still presents an optimal shape, while the PR curves of other models, such as ITIN and MVAN, present a poor shape, indicating their relatively poor performance. It is noteworthy that the DMAF model's performance on the MVSA dataset is relatively inferior in contrast to its performance on the TumEmo dataset.

5.6. Ablation Experiment

The following ablation experiments are conducted to further verify the performance of the MCAM model.

Through the performance analysis of the ablation experiments in Table 8 and Figure 10, it is observed that the model's performance is average on different datasets when only images or text are utilized, mainly because the single-modal sentiment analysis is only learned within the modality. Sometimes, images may play a leading role in the classification of emotional categories, and sometimes, the text may play a leading role. Thus, single-modal sentiment analysis fails to fully consider the complementarity and correlation between diverse modalities, resulting in suboptimal classification outcomes.

Table 8. Comparison of ablation experiments.

Methods	Model	Accuracy	Precision	Recall	F1 Score
MVSA	Only text	0.784	0.743	0.743	0.772
	Only image	0.671	0.696	0.673	0.641
	Fusion of image–text	**0.865**	**0.843**	**0.851**	**0.855**
TumEmo	Only text	0.752	0.728	0.762	0.759
	Only image	0.584	0.538	0.584	0.569
	Fusion of image–text	**0.743**	**0.778**	**0.732**	**0.715**

When we introduce the cross-attention mechanism, the model first uses an independent network to learn the feature representation of each modality; then, it employs a cross-attention mechanism to capture inter-modality relationships. Then, similarities between different modalities are computed to obtain weights for a cross-attention mechanism,

which fuses feature representations from different modalities in a weighted manner. Finally, the model employs the fused feature representations to predict sentiment, resulting in improved performance and robustness.

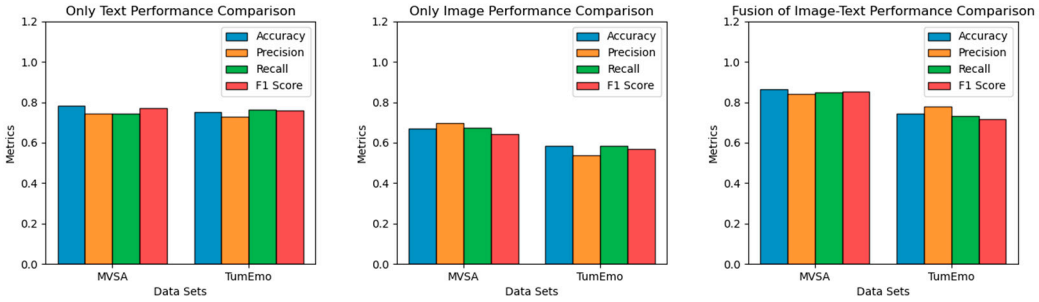


Figure 10. Ablation experiment results of different datasets.

5.7. Attention Weight Visual Analysis

This section focuses on a qualitative analysis of image and text fusion for sentiment analysis, including attention scores before and after implementing cross-attention. The emotional score of attention is any value between 0 and 1, and the specific results are shown in Figure 11. Through visual analysis, we can clearly see the change in the thermal effect of the image area before and after the introduction of attention.

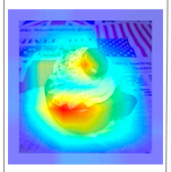
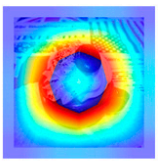
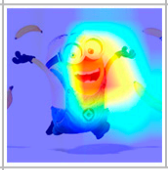
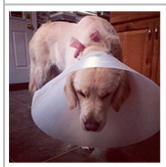
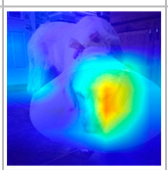
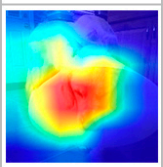
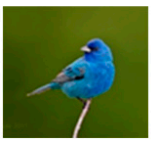
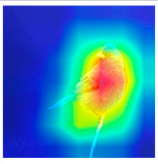
Raw image-text pair		Both pictures and text attract attention		Image-text pairs with cross-attention	
Fused attention score		0.33		0.65	
Unimodal attention score		0.26	0.15	0.48	0.28
	A #mini#cup of #chocolate with #coffee and #whipped #cream ???#delicious# pastry#dessert#patisserie #caffe#cioccolato…		A #mini#cup of #chocolate with #coffee and #whipped #cream ???#delicious# pastry#dessert#patisserie #caffe#cioccolato…		A #mini#cup of #chocolate with #coffee and #whipped #cream ???#delicious# pastry#dessert#patisserie #caffe#cioccolato…
Fused attention score		0.49		0.79	
Unimodal attention score		0.25	0.22	0.52	0.41
	It's # Monday! Have an awesome and #successful week! Smile and be #happy #small business #energetic		It's # Monday! Have an awesome and #successful week! Smile and be #happy #small business #energetic		It's # Monday! Have an awesome and #successful week! Smile and be #happy #small business #energetic
Fused attention score		0.29		0.55	
Unimodal attention score		0.21	0.11	0.36	0.23
	Poor little guy ? #puppy #goldenretriever #cone #nothappy #confused #rt		Poor little guy ? #puppy #goldenretriever #cone #nothappy #confused #rt		Poor little guy ? #puppy #goldenretriever #cone #nothappy #confused #rt
Fused attention score		0.29		0.52	
Unimodal attention score		0.15	0.12	0.29	0.23
	"@helwatweets: #Beautiful #birdwatch #photo #amazing #photography #wildlife #wild via		"@helwatweets: #Beautiful #birdwatch #photo #amazing #photography #wildlife #wild via		"@helwatweets: #Beautiful #birdwatch #photo #amazing #photography #wildlife #wild via

Figure 11. Four examples of visual analysis of attention scores.

5.7.1. Visual Attention Processing

For image data, the convolutional neural network is first enhanced using CBAM to improve the ability of image feature extraction. Input image tensors (224, 224, 3) for average pooling and maximum pooling operations calculate the average and maximum values of the channel dimensions, compress them into vectors with dimensions 128 and 1024 through two fully connected layers, and then restore them to the original dimension.

The results are then combined, and weights ranging from 0 to 1 are obtained using the sigmoid function. The attention mechanism is employed to the weighted tensor's spatial dimension; it calculates the average value and maximum value of the weighted tensor on the spatial dimension and inputs it into a 1×1 convolutional layer after splicing to obtain $1 \times 1 \times 1$ feature map; then, we obtain the weight from 0 to 1 through the sigmoid function. Finally, the channel and spatial attention mechanisms' weights are multiplied to derive the ultimate attention score.

We draw a heat map on the image processed by the CBAM attention mechanism to make its color more vivid. The original image is assigned a transparency of 0.5, while the processed image is assigned a transparency of 0.8 to emphasize the attention focus. If the region's attention score is higher, the color of the region is redder.

5.7.2. Text Attention Processing

For text data, we first tokenize the input text data to obtain a token sequence, convert the token sequence into an input tensor that can be processed by the ALBert model, and use the attention mask tensor to represent the position of each token.

Next, we input the input tensor and attention mask into the ALBert model, and the model will encode the input token to obtain an encoded tensor. For each self-attention layer, the model splits the encoded tensor into multiple heads, each of which computes the query, key, and value separately and uses this information to compute the attention score between each token and other tokens. Finally, by weighting the attention scores obtained by each attention head, a weight vector is obtained, which represents the importance of each token in the entire text data, that is, the attention weight.

After processing the text, we color-coded the emotional words that the attention focused on. To differentiate the contribution levels of distinct words to the attention score, we applied varying degrees of shading to the same color. The darker the color, the greater the attention weight of the word.

5.7.3. Cross-Attention Fusion Multi-modal Processing

For multi-modal data that fuse images and text, first, we pass the image features and text features to two fully connected layers, take their outputs as input, pass them to the dot product layer, and calculate the attention weights.

Next, the attention weights are multiplied by image features and text features, respectively, to obtain attention tensors for images and text. Finally, we use the concatenate layer to concatenate the attention tensor of the image and text along the last dimension to obtain the fused feature tensor and calculate the attention score of the fused feature tensor, which is the fusion of multi-modal cross-attention power score.

For the image and text after cross-attention fusion, we can clearly observe that the area of the image that is concerned is further expanded, the red color is further deepened, and the color of the word that is focused on is also deepened. The attention score has also been improved to varying degrees, and the attention score after fusion has been further improved compared with the score of simple vector splicing.

5.8. Prediction Error Case Analysis

In Figure 12, the sentiment orientation of the four predicted examples is inconsistent with the labeled sentiment orientation. In the first image, from the smiling expression of the little boy, we can easily determine that the emotion is positive, but the text content contains some negative words, such as "dead people", "inquietante", and "terrified",

which ultimately lead to the prediction mistake. In the second image, the negative emotion can also be felt from the tone of the image, but there is the word "happy", with obvious positive emotion in the text content; however, "contempt" and "empty" are also in the text, conveying "contempt" and "emptiness", but after ALBert's processing, the text is divided into independent words, and the meaning may change, i.e., it no longer has a particularly obvious emotional tendency.

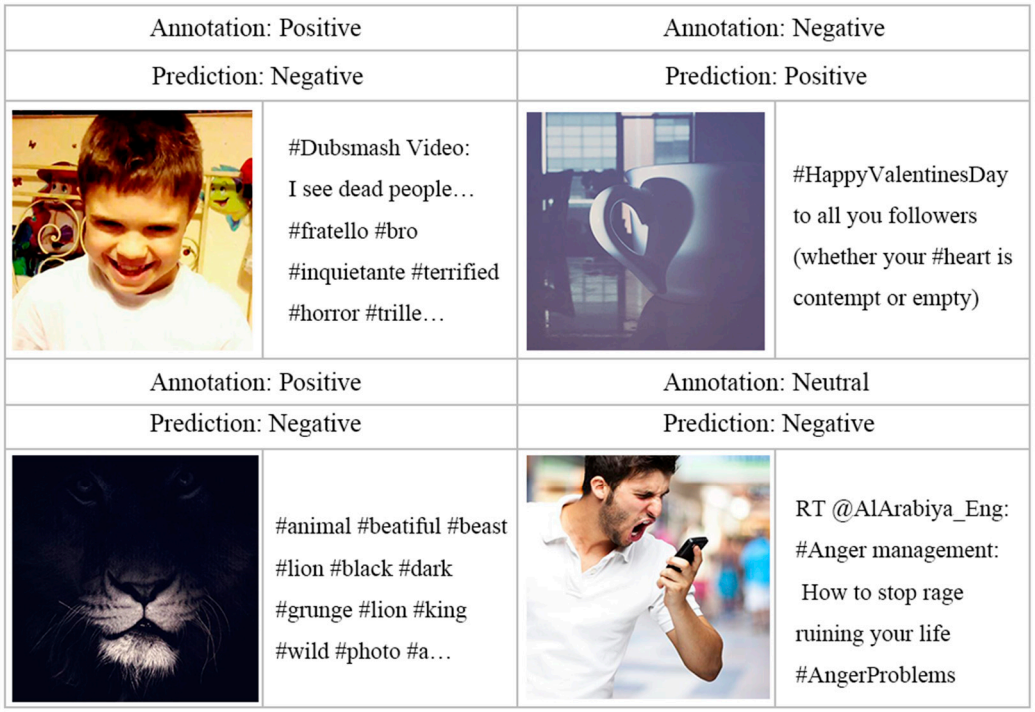


Figure 12. Examples of four wrong predictions in the MVSA dataset.

In the third image, the reason for the prediction error may be that there are wrong words in the text; "beautiful" originally means beautiful, but the word in the text is "beatiful", causing ambiguity. In the fourth image, we can also feel negative emotions from the roar of the man in the image, and we can also read negative emotions from in the words "Anger" and "ruining" in the text. The ultimate prediction outcome is negative as well, but the original annotation is neutral; we think this is likely the reason why the dataset itself is wrongly annotated.

6. Conclusions

MCAM is introduced in our paper as a cross-attention-mechanism-based multi-modal sentiment analysis approach that enhances feature fusion by adaptively computing and calculating the correlation between images and texts. First of all, we propose two different processing methods for a single modality to learn text and image features, respectively; then, we use the cross-attention mechanism to adaptively fuse the features of different modalities, thereby improving the performance of sentiment analysis. Finally, the sentiment result is output through the classifier. Experimental results demonstrate the superiority of our proposed approach over four baselines, achieving superior performance on two publicly available multi-modal sentiment analysis datasets.

As for the limitations of the current work, we note that the cross-attention mechanism for conflicting sentiment indicators between modalities is a challenge in the current work. As for situations where sentiment indicators conflict with each other, this is the direction that we need to consider in our next work. We can use domain knowledge or prior information, such as sentiment dictionaries or sentiment rules, to help us resolve sentiment conflicts between modalities.

Our future work will entail improving the model to enhance its capacity for learning deeper multi-modal correlations, thereby enabling it to achieve superior results in image–text sentiment analysis. In addition, we need to improve the text processing model. For example, for misspelled words, we can use spell checkers and error correctors to detect and correct spelling mistakes; thus, we delete or correct wrongly annotated labels in the dataset. In the future, we also want to study further the sentiment analysis of more modalities, such as video and audio.

Author Contributions: Conceptualization, H.L.; Methodology, H.L. and Y.L.; Software, Y.L.; Validation, Y.L.; Resources, H.Z.; Writing—original draft preparation, Y.L.; Writing—review and editing, H.Z. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Henan Provincial Science and Technology Project, Project No. 232102210035 and No. 24B520040.

Data Availability Statement: Publicly available datasets are used in this study. Links to data are provided in the footnotes of the article.

Acknowledgments: The authors would like to thank the editors and the anonymous reviewers for their helpful comments and suggestions, which have improved the presentation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hu, R.; Rui, L.; Zeng, P.; Chen, L.; Fan, X. Text sentiment analysis: A review. In Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018; IEEE: Piscataway, NJ, USA, 2018.
2. Cai, Z.; Cao, D.; Ji, R. Video (GIF) sentiment analysis using large-scale mid-level ontology. *arXiv* **2015**, arXiv:1506.00765.
3. Xu, C.; Cetintas, S.; Lee, K.C.; Li, L.J. Visual sentiment prediction with deep convolutional neural networks. *arXiv* **2014**, arXiv:1411.5731.
4. Tang, D.; Qin, B.; Liu, T. Learning semantic representations of users and products for document level sentiment classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Volume 1: Long Papers, pp. 1014–1023.
5. Ibrahim, M.; Abdillah, O.; Wicaksono, A.F.; Adriani, M. Buzzer detection and sentiment analysis for predicting presidential election results in a twitter nation. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1348–1353.
6. Pérez-Rosas, V.; Mihalcea, R.; Morency, L. Utterance-Level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Sofia, Bulgaria, 4–9 August 2013; in Long Papers. The Association for Computer Linguistics: Kerrville, TX, USA, 2013; Volume 1, pp. 973–982.
7. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multi-modal emotion recognition and sentiment analysis. In Proceedings of the IEEE 16th International Conference on Data Mining, ICDM 2016, Barcelona, Spain, 12–15 December 2016; Bonchi, F., Domingo-Ferrer, J., Baeza Yates, R.A., Zhou, Z., Wu, X., Eds.; IEEE: Piscataway, NJ, USA, 2016; pp. 439–448.
8. Chen, M.; Wang, S.; Liang, P.P.; Baltrusaitis, T.; Zadeh, A.; Morency, L. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, UK, 13–17 November 2017; Lank, E., Vinciarelli, A., Hoggan, E.E., Subramanian, S., Brewster, S.A., Eds.; ACM: New York, NY, USA, 2017; pp. 163–171.
9. You, Q.; Luo, J.; Jin, H.; Yang, J. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; Bennett, P.N., Josifovski, V., Neville, J., Radlinski, F., Eds.; ACM: New York, NY, USA, 2016; pp. 13–22.
10. Cao, D.; Ji, R.; Lin, D.; Li, S. A cross-media public sentiment analysis system for microblog. *Multimed. Syst.* **2016**, *22*, 479–486. [CrossRef]
11. Poria, S.; Cambria, E.; Gelbukh, A.F. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15), Lisbon, Portugal, 17–21 September 2015; pp. 2539–2544.
12. Zargari, H.; Zahedi, M.; Rahimi, M. GINS: A Global intensifier-based N-Gram sentiment dictionary. *J. Intell. Fuzzy Syst. Appl. Eng. Technol.* **2021**, *40*, 11763–11776. [CrossRef]
13. Xu, G.; Yu, Z.; Yao, H.; Li, F.; Meng, Y.; Wu, X. Chinese text sentiment analysis based on extended sentiment dictionary. *IEEE Access* **2019**, *7*, 43749–43762. [CrossRef]
14. Goel, A.; Gautam, J.; Kumar, S. Real time sentiment analysis of tweets using naive bayes. In Proceedings of the 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 14–16 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 257–261.

15. Rathor, A.S.; Agarwal, A.; Dimri, P. Comparative study of machine learning approaches for Amazon reviews. *Procedia Comput. Sci.* **2018**, *132*, 1552–1561. [CrossRef]
16. Zhou, X.; Wan, X.; Xiao, J. Attention-based lstm network for cross-lingual sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 247–256.
17. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 380–385.
18. Miao, Y.; Ji, Y.; Peng, E. Application of CNN-BiGRU Model in Chinese short text sentiment analysis. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 20–22 December 2019.
19. Yenduri, G.; Rajakumar, B.R.; Praghash, K.; Binu, D. Heuristic-Assisted BERT for Twitter Sentiment Analysis. *Int. J. Comput. Intell. Appl.* **2021**, *20*, 20625–20631. [CrossRef]
20. Cauteruccio, F.; Kou, Y. Investigating the emotional experiences in eSports spectatorship: The case of League of Legends. *Inf. Process. Manag.* **2023**, *60*, 103516. [CrossRef]
21. Yuan, J.; Mcdonough, S.; You, Q.; Luo, J. Sentribute: Image sentiment analysis from a mid-level perspective. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, Chicago, IL, USA, 11 August 2013; ser. WISDOM '13. ACM: New York, NY, USA, 2013; pp. 1–8.
22. Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.-F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; ser. MM'13. ACM: New York, NY, USA, 2013; pp. 223–232.
23. He, X.; Zhang, H.; Li, N.; Feng, L.; Zheng, F. A multi-attentive pyramidal model for visual sentiment analysis. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
24. Wang, M.; Cao, D.; Li, L.; Li, S.; Ji, R. Microblog sentiment analysis based on cross-media bag-of-words model. In Proceedings of the International Conference on Internet Multimedia Computing and Service, Xiamen, China, 10–12 July 2014; pp. 76–80.
25. Zhang, Y.; Shang, L.; Jia, X. Sentiment analysis on microblogging by integrating text and image features. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Ho Chi Minh City, Vietnam, 19–22 May 2015; pp. 52–63.
26. Yu, Y.; Lin, H.; Meng, J.; Zhao, Z. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms* **2016**, *9*, 41. [CrossRef]
27. Kumar, A.; Srinivasan, K.; Cheng, W.H.; Zomaya, A.Y. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Inf. Process. Manag.* **2020**, *57*, 102141. [CrossRef]
28. Xu, J.; Huang, F.; Zhang, X.; Wang, S.; Li, C.; Li, Z.; He, Y. Visual-textual sentiment classification with bi-directional multi-level attention networks. *Knowl-Based Syst.* **2019**, *178*, 61–73. [CrossRef]
29. Huang, G.; Liu, Z.; Laurens, V.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2016.
30. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations, Virtual, 26 April–1 May 2020.
31. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018.
32. You, Q.; Jin, H.; Luo, J. Visual sentiment analysis by attending on local image regions. In Proceedings of the 31st AAAI conference on Artificial Intelligence (AAAI'17), San Francisco, CA, USA, 4–9 February 2017; Volume 31, pp. 231–237.
33. Long, F.; Zhou, K.; Ou, W. Sentiment analysis of text based on bidirectional LSTM with multi-head attention. *IEEE Access* **2019**, *7*, 141960–141969. [CrossRef]
34. Jia, D. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009.
35. Hoang, M.; Bihorac, O.A.; Rouces, J. Aspect-based sentiment analysis using BERT. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, Turku, Finland, 30 September–2 October 2019.
36. Yin, X.; Liu, C.; Fang, X. Sentiment analysis based on BiGRU information enhancement. *J. Phys. Conf. Series.* **2021**, *1748*, 032054. [CrossRef]
37. Song, L.; Deng, X. Research on rice leaf disease identification based on ResNet. In Proceedings of the 2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Beijing, China, 3–5 October 2022; IEEE: Piscataway, NJ, USA, 2022.
38. Paymode, A.S.; Malode, V.B. Transfer learning for multi-crop leaf disease image classification using convolutional neural network VGG. *Artif. Intell. Agric.* **2022**, *6*, 23–33. [CrossRef]
39. Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowl.-Based Syst.* **2019**, *167*, 26–37. [CrossRef]
40. Zhu, T.; Li, L.; Yang, J.; Zhao, S.; Liu, H.; Qian, J. Multimodal sentiment analysis with image-text interaction network. *IEEE Trans. Multimed.* **2022**, *25*, 3375–3385. [CrossRef]

41. Yang, X.; Feng, S.; Wang, D.; Zhang, Y. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Trans. Multimed.* **2020**, *23*, 4014–4026. [CrossRef]
42. Huang, F.; Wei, K.; Weng, J.; Li, Z. Attention-based modality-gated networks for image-text sentiment analysis. *ACM Trans. Multimed. Comput. Commun. Appl. TOMM* **2020**, *16*, 1–19. [CrossRef]