Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.3, No.2, 2023

Dual-Channel Attention-Based Multimodal Sentiment Analysis Model Integrating Text and Image Features

Tamsin Reuel

University of Northern Iowa, Cedar Falls, USA

Tamsin.reuel1199@gmail.com

Abstract: With the rise of multimodal data in social media, sentiment analysis based solely on text has become insufficient to capture the richness of human emotion. To address this limitation, this paper proposes a dual-channel multimodal sentiment analysis model based on attention mechanisms, named ACMSA (Attention Channel Multimodal Sentiment Analysis). The model integrates textual and visual features to improve emotional understanding and classification accuracy. Text features are extracted using the BERT model and processed through a CNN-BiGRU-Attention dual-channel architecture to capture both local and global semantic dependencies. Image features are obtained via ResNet152, enhanced by a Channel-Spatial Attention Module (CSAM) that adaptively emphasizes salient regions. The fusion of multimodal features is achieved through a Co-Attention mechanism, enabling fine-grained interaction between textual and visual representations. Experimental evaluations on the MVSA-Single and MVSA-Multi Twitter datasets demonstrate that ACMSA outperforms state-of-the-art baselines, achieving an accuracy of 77.08% and 74.42%, respectively. The results verify that attention-guided dual-channel modeling effectively strengthens cross-modal correlation and interpretability. This framework provides a robust and extensible solution for sentiment analysis in multimedia-rich environments, offering valuable implications for emotion recognition, social media monitoring, and intelligent interaction systems.

Keywords: Multimodal sentiment analysis; Attention mechanism; Dual-channel model; BERT; ResNet152; Co-Attention.

1. Introduction

In daily life, people encounter various forms of information, and user comments are often one of the important ways for people to obtain information. Different from traditional pure text comments, today's

comments include audio, images, and text. This type of data composed of multiple forms of information is called multimodal data. Using multimodal data, complementary information can be extracted from each modality, and compared to single- modal data, richer content can be obtained. Multimodal sentiment analysis is a method that utilizes multiple forms of information (such as text and images) for sentiment analysis, aiming to obtain more comprehensive emotional representations and more accurate emotional tendency analysis. Although research in the field of multimodal sentiment analysis is relatively less extensive compared to single-modal analysis, it has tremendous potential to provide deeper emotional understanding and more accurate emotion recognition. By integrating multiple forms of information, richer and more comprehensive emotional features can be obtained, thereby improving the performance and effectiveness of sentiment analysis. With the continuous development of this field, multimodal sentiment analysis is expected to bring more accurate and comprehensive results for tasks such as emotion recognition, emotion classification, and emotion generation. In social media, comments are the main way for users to express their emotional tendencies towards events. To address the issue of the current predominant form of user comments gradually becoming integrated text and images, this paper proposes a dual-channel text-image classification model based on attention mechanism (Attention Channel Multimodal Sentiment Analysis, ACMSA).

2. Related Work

Recent advances in deep neural architectures and multimodal representation learning have significantly influenced the design of modern multimodal sentiment analysis frameworks. The methodological foundation of the proposed model draws heavily from the evolution of attention-based neural architectures and multimodal fusion mechanisms that enable effective integration of heterogeneous information sources.

The introduction of the Transformer architecture established a new paradigm for sequence representation learning by replacing recurrent structures with self-attention mechanisms that capture long-range dependencies through global token interactions [1]. This mechanism enables contextualized feature representation and has become the backbone of modern language models. Building upon this principle, subsequent studies extended attention-based modeling to multimodal scenarios by designing architectures capable of aligning and integrating heterogeneous modality streams. The multimodal transformer framework further demonstrated that attention-based cross-modal interactions can effectively model relationships between asynchronous or unaligned modalities [2]. Similarly, integrating multimodal information within large pretrained transformer models has been shown to enhance contextual reasoning across modalities while preserving the expressive power of pretrained representations [3]. These developments collectively motivate the use of attention-driven representation learning as the primary mechanism for modeling textual semantics and facilitating cross-modal interaction in the proposed framework.

Beyond sequence modeling, attention mechanisms have also been widely applied in visual feature extraction to improve representation quality. Channel and spatial attention strategies enable neural networks to dynamically emphasize informative feature regions while suppressing irrelevant information. The Convolutional Block Attention Module (CBAM) introduced a lightweight yet effective approach for integrating channel-wise and spatial attention into convolutional networks, significantly enhancing

feature discrimination capability [4]. Subsequent improvements further explored multi-scale attention mechanisms to strengthen the ability of convolutional architectures to capture fine-grained structural patterns and contextual dependencies within visual representations [5]. These attention-based feature enhancement techniques provide important methodological guidance for designing adaptive visual feature refinement modules within multimodal systems.

In multimodal learning, effective modeling of interactions between modalities is critical for capturing complementary semantic cues. Early research introduced memory-based architectures to enable bidirectional information exchange between modalities, allowing each modality to iteratively refine its representation through cross-modal attention mechanisms [6]. Building upon this idea, multi-view attention networks further explored the modeling of heterogeneous feature spaces by jointly attending to multiple semantic perspectives during feature fusion [7]. Fusion-extraction frameworks subsequently emphasized the importance of learning both intra-modal representations and inter-modal correlations within a unified architecture, demonstrating that coordinated feature extraction and interaction modeling significantly improve multimodal understanding [8]. Similarly, interaction-based multimodal architectures leverage cross-modal attention to generate modality-enhanced representations, enabling one modality to guide the feature selection process of another [9]. These interaction-driven strategies highlight the importance of explicitly modeling cross-modal relationships rather than relying solely on simple feature concatenation.

Another important line of work focuses on large-scale multimodal representation learning and dataset construction, which has facilitated the development of more robust cross-modal modeling approaches. Comprehensive multimodal benchmarks provide diverse multimodal signals and enable systematic evaluation of cross-modal representation learning techniques [10]. Early multimodal sentiment analysis studies demonstrated that deep neural networks can effectively learn joint representations from textual and visual signals, providing evidence that multimodal modeling significantly outperforms single-modality approaches [11], [12]. These works established the foundational understanding that multimodal sentiment recognition requires coordinated learning of both modality-specific features and shared semantic representations.

Recent research has further explored optimization strategies and training mechanisms to improve model robustness and generalization. Structural regularization and bias mitigation strategies have been proposed to stabilize parameter-efficient adaptation processes and reduce representation bias during model optimization [13]. Multi-scale representation learning combined with uncertainty estimation has also been introduced as an effective strategy for improving robustness and reliability in complex learning environments [14]. In addition, contrastive representation learning frameworks provide an alternative approach for learning discriminative feature embeddings by maximizing agreement between related samples while separating irrelevant representations [15]. These representation learning strategies contribute to the development of more stable and discriminative multimodal feature spaces.

Complementary advances in learning paradigms also contribute to the methodological development of multimodal systems. Causal inference and bias correction techniques have been introduced to address exposure bias and improve decision reliability in complex predictive models [16]. Meanwhile, multi-level attention and sequential modeling approaches demonstrate that hierarchical attention mechanisms can effectively capture dynamic dependencies within sequential data representations [17]. These strategies

collectively inspire the hierarchical attention design and structured feature integration adopted in the proposed model.

Finally, several studies provide broader perspectives on large-scale data analysis and task diversity that inform the evaluation and application of multimodal models. Investigations of large-scale online interaction environments highlight the complexity of emotional signals embedded in multimodal communication contexts [18]. Meanwhile, collections of diverse classification tasks offer valuable benchmarks for assessing generalization capability across heterogeneous data distributions [19]. These resources support the empirical validation of multimodal learning systems and encourage the development of more robust and adaptable modeling strategies.

Building upon these methodological developments, the proposed approach integrates attention-driven representation learning, multimodal interaction modeling, and hierarchical feature refinement within a unified architecture. By combining attention-based sequence modeling with cross-modal interaction mechanisms and enhanced visual attention modules, the framework inherits the strengths of prior work while introducing a structured dual-channel design that enables more effective integration of textual and visual information. This methodological integration provides the foundation for improving multimodal sentiment recognition through more expressive feature representations and more precise cross-modal correlation modeling.

The main contributions of ACMSA are as follows:

(1)Designing a dual-channel model composed of CNN and Bidirectional GRU. After CNN extracts local features of the text, a combination of average pooling and max pooling is used to obtain vectors Cavg and Cmax with low-level and high-level features, respectively. BiGRU extracts global features of the text, adopts gate recurrent units to obtain the feature representation of the text, and incorporates a self-attention mechanism layer to capture key information in the text.

(2)Introducing a new attention module, the CSAM module, comprising channel attention and spatial attention modules. In the channel attention module, two different-sized convolution kernels are first used to extract original feature vectors and fuse them, then the fused feature vectors are compressed and activated through pointwise convolution, and finally, the original and fused feature vectors are adaptively learned to determine the weights. In the spatial attention module, dilated convolution is used to extract feature vectors to obtain a larger receptive field and retain the spatial features of the image.

(3)Innovatively using fusion modules to obtain text feature vectors with image information and image feature vectors with textual information. Introducing fusion attention mechanism and enhancing the connection through adaptive fusion to improve the network's expressive ability.

3. Model Design

The ACMSA model proposed in this paper consists of four parts: text feature vector extraction, image feature vector extraction, feature fusion, and classification. The model framework is illustrated in Figure 1, where the feature extraction model mainly includes Bert and ResNet152, and the fusion method primarily involves late fusion based on attention mechanisms. Co- Attention extracts the weights of each modality separately for weighted fusion classification, enhancing the interaction between different modalities. The CSAM module and multi-channel model will be elaborated on in the following sections.

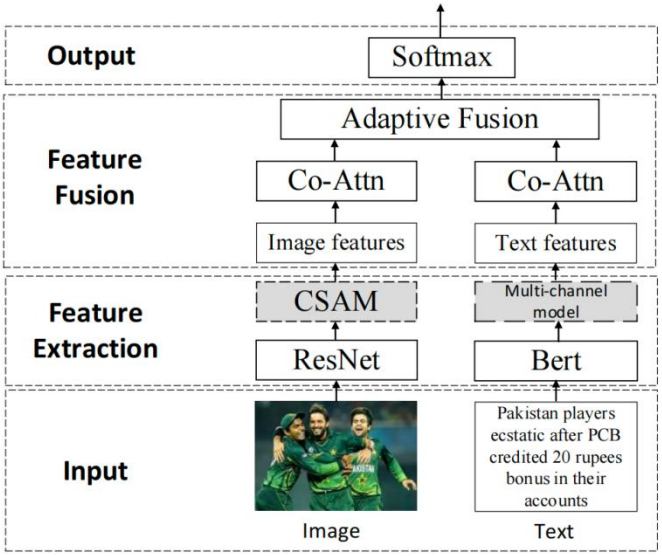

Figure1. ACMSA model sturcture

3.1 Text Feature Extraction

Based on deep learning for sentiment classification, continuously training labeled data until the best results are achieved offers significant advantages in terms of scalability and accuracy. The Bert model holds a leading position in text feature extraction, composed of the Encoder part of Transformer. Compared to earlier models such as Word2Vec, GloVe, and ELMO, it demonstrates efficiency and stronger generalization, effectively linking context. In this paper, the text feature extraction structure of the multi-channel model is illustrated in Figure 2.

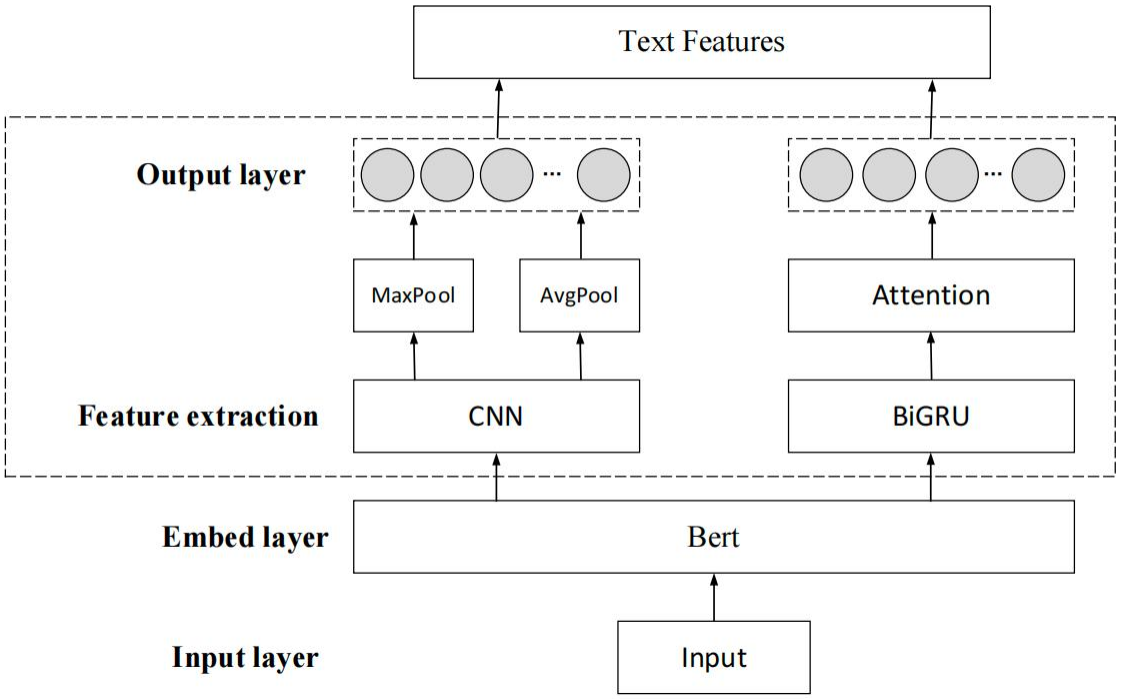

Figure 2. Text dual channel model structure

3.1.1 Text Local Feature Extraction

The feature vectors extracted from text input through Bert are passed into CNN via the first channel. CNN, originally developed for computer vision, has been gradually applied in the field of natural language processing (NLP) in recent years with remarkable results. CNN primarily extracts local features through pooling and convolutional layers. In this paper, average pooling and max pooling are combined to obtain vectors with low-level features (Cavg) and high-level features (Cmax), which are then merged to form Cout. CNN has significant advantages in extracting local features. By using convolutional kernels of different sizes (2, 3, 4), it can comprehensively differentiate and extract textual information, even when the granularity of sequential features varies. The formula is as follows:

$$C_i = f(W \cdot X_{i:i+h-1} + b) \tag{1}$$

In the equation, *Ci* represents the feature vector obtained by convolving the input with convolution kernels of different sizes $\omega \in R^{h \times d}$. f denotes the activation function, *h* represents the size of the convolution kernel, and $b \in R$ represents the bias term. The filtering window smoothly moves to $x_{n-h+1:n}$, yielding the feature sequence as shown in the formula,

$$c = [c_1, c_2, c_3, \ldots, c_{n-h+1}] \tag{2}$$

3.1.2 Text Global Feature Extraction

The feature vectors extracted by Bert are passed into BiGRU-Attention through the second channel. GRU (Gated Recurrent Unit) is an improvement over traditional recurrent neural networks (RNNs), introducing reset gates and update gates to optimize the structure. Compared to RNNs, GRU has fewer parameters. GRU stores and forgets information through memory units, effectively addressing the vanishing gradient problem. BiGRU, an extension of GRU, is bidirectional, making it more advantageous for capturing contextual relationships.At time *t*, the output ht for input xt is calculated as follows:

$$z_t = \sigma(W^{(z)} x_t + U^{(z)} h_{t-1}) \tag{3}$$

$$r_t = \sigma(W^{(r)} x_t + U^{(r)} h_{t-1}) \tag{4}$$

$$h_t' = \tanh(W x_t + r_t \otimes U h_{t-1}) \tag{5}$$

$$h_t = z_t \cdot h_t' + (1 - z_t) \cdot h_{t-1} \tag{6}$$

3.2 Image Feature Extraction

After extracting the image feature vectors, incorporating an attention mechanism module allows the model to selectively focus on key regions in the image, which can enhance the model's resistance to interference and robustness. In this paper, after using ResNet152 to extract image features, the model's performance is further enhanced by adding the CSAM module. The CSAM structure is illustrated in Figure 3.

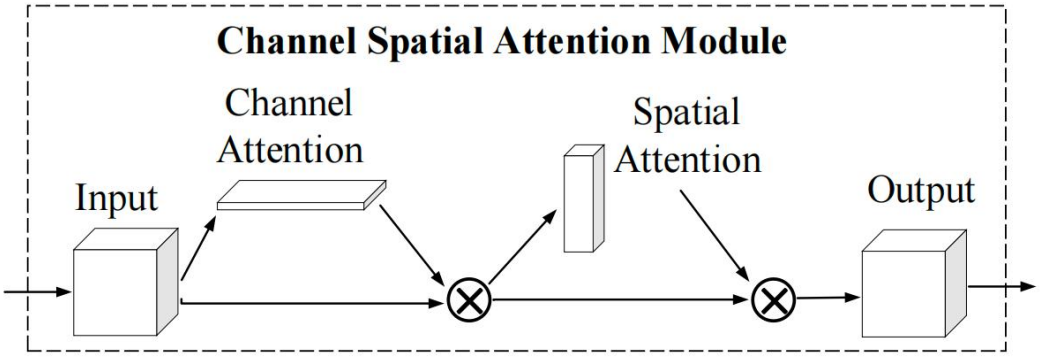

Figure 3. CSAM Attention Module

The CBAM module integrates channel attention mechanism and spatial attention mechanism. The CSAM module innovates and improves the channel attention module by first extracting features using two convolution kernels of different sizes, then adding the feature vectors extracted by the two convolution kernels, and finally passing the feature vectors through two channels of different scales. The core idea of CSAM is to calculate the weights of different channels to achieve attention on multiple scales, thereby enhancing the feature representation capability. It is easier to select important information from different channels, thereby increasing accuracy. SAM is a complement to CAM, after selecting different channels, SAM identifies which position in that channel direction has more feature expression. The calculation formula is as follows:

$$F' = M_C(F) \otimes F \tag{7}$$

3.2.1 Channel Attention

This module determines the importance of each channel in the transmission process from the channel dimension, utilizing channel-wise attention to enable different receptive fields. The feature maps have different feature weights, allowing the model to have an adaptive receptive field, thereby achieving the effect where outputs obtained from different convolution kernel sizes have varying levels of importance, enhancing the model's feature extraction capability. The structure of the channel attention mechanism is illustrated in Figure 4.

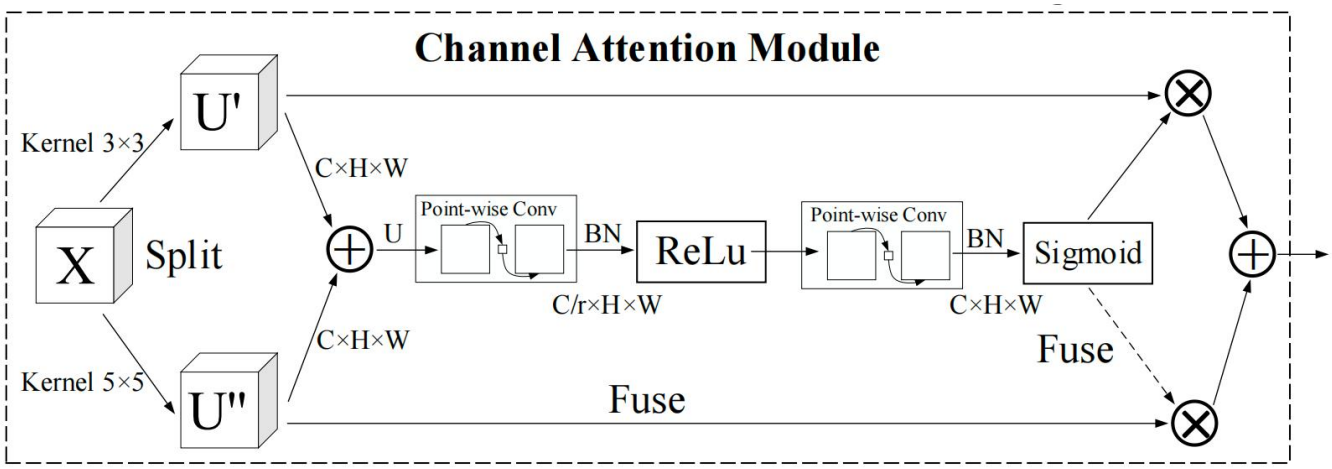

Figure 4. Channel attention module

This module mainly consists of three steps: Split, Fuse, Select.

Split: The input feature values are passed into two branches. Branch one uses a 3x3 convolution kernel to extract features, while the other branch uses a 5x5 dilated convolution with dilation=2 to extract features.

Fuse: The feature maps extracted from the two branches in Split are added together. Using pointwise convolution, the number of channels C of the feature X is reduced to a certain value. After passing through BatchNorm layer and ReLU activation function, it is then passed into r another pointwise convolution to restore the number of channels to the original value. Finally, it passes through a BN layer and a Sigmoid activation function.

Select: The weighted average of U and U is calculated. The fusion feature is obtained by subtracting this weighted average from 1. Through iterative training of the model, the network adapts and determines the weights autonomously.

3.2.2 Spatial Attention

Introducing spatial attention mechanism enables automatic capture of important regional features, calculating local features and key information, thus providing more accurate localization and weighting of the regions of interest in the feature maps. It can extract more discriminative features, thereby improving the accuracy and robustness of the model, while also reducing the computational burden of the model.The structure of the channel attention mechanism is illustrated in Figure 5.

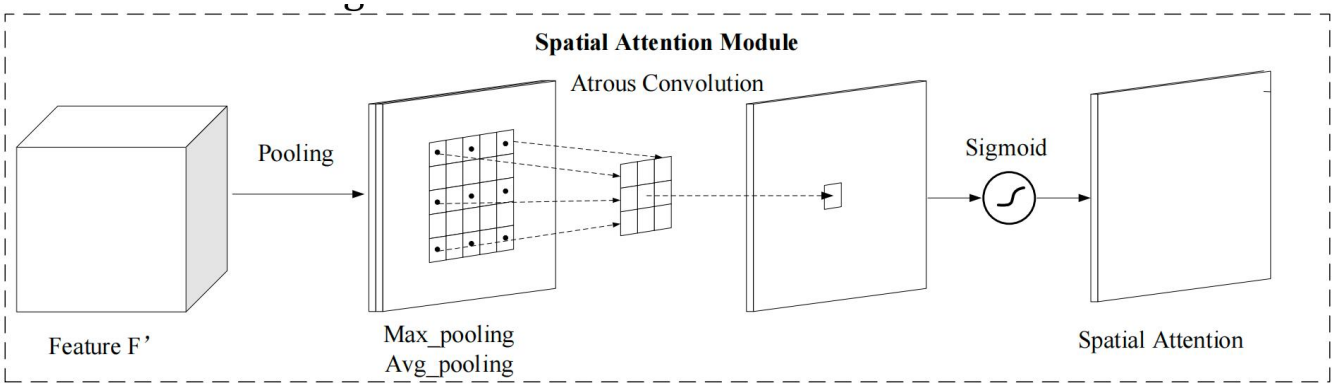

Figure 5. Spatial attention module

The feature map inputted into a certain channel undergoes max pooling and average pooling, extracting the maximum and average values along the channel dimension. The results from max pooling and average pooling are concatenated and passed into a dilated convolution with a kernel size of 3x3 and dilation=2. The output is passed through a Sigmoid activation function. Multiple experiments have shown that introducing dilated convolution yields better results, as it allows for a larger receptive field, thereby obtaining denser data and capturing multi-scale information, effectively preserving the spatial features of the image.The calculation formula is as follows,

$$M_S(F) = \sigma(f^{3\times3}([AvgPool(F); MaxPool(F)])) \tag{8}$$

3.3 Model Fusion

Currently, there are three main modal fusion methods: early fusion, mid-level fusion, and late fusion. Early fusion, also known as input layer fusion, refers to the fusion of input data before feature extraction, preserving and enhancing the originality and correlation of the data. Mid- level fusion, or feature layer fusion, involves fusing features after feature extraction and then inputting them together into the decision layer for classification. Late fusion, also called decision layer fusion, is considered the most effective of the three fusion methods. It involves combining processed feature vectors from various modalities and inputting them together into the decision layer for classification.

As attention mechanisms become more popular in the field of multimodal fusion, in order to achieve more effective fusion of multimodal features, this paper employs Co-Attention. The key feature of this attention mechanism is that QK comes from one modality while V comes from another modality. By comparing and dynamically adjusting the similarity and weight between different modality features, it becomes easier to capture the correlation between different modalities, allowing the features of both modalities to guide each other, thereby generating text feature vectors with image information and image feature vectors with text information.

$$
\begin{aligned}
u_f &= \tanh(W_m(x_t \otimes x_i) + b_m) \\
a_m &= \frac{e^{u_f}}{\sum e^{u_f}} \\
S_t &= \sum_{i=1}^{n} a_m \cdot x_t \\
S_i &= \sum_{i=1}^{k} a_m \cdot x_i
\end{aligned} \tag{9}
$$

Where A represents the joint image-text feature after Co-Attention, B and C respectively denote the weight matrix and the attention weights, D and E represent the text and image feature vectors before Co-Attention mechanism, F represents the text feature vector with image information after Co-Attention, and H represents the image feature vector with text information after this mechanism.

3.4 Feature Classification

After the image and text feature vectors undergo the Co-Attention mechanism, they are inputted into a fully connected layer and a softmax classifier to output the classification results. To prevent overfitting, cross-entropy loss is used as the objective function for the softmax.

3. Experimental Analysis

4.1 Datasets

This paper utilizes two datasets, MVSA-Single and MVSA-Multi, collected from Twitter data. The MVSA-Single dataset consists of 5129 image-text pairs, with single-label annotations for both images and text as positive, negative, or neutral. To ensure the authenticity of the data and the accuracy of the experiments, this paper follows the data cleaning method proposed in, which removes image-text pairs with contradictory polarity labels, i.e., samples with both positive and negative labels, and replaces images labeled as neutral or with polarity labels with the corresponding polarity labels. After processing, the dataset contains 4511 samples.

The MVSA-Multi dataset contains 19600 image-text pairs, with three sets of independent labels. Initially, the majority emotion labels for both image and text modalities are replaced with the true labels for a single modality, i.e., if at least two emotions are the same for one modality, it is retained as the true label. Similarly, image-text pairs with contradictory labels are removed, eliminating samples with contradictory polarity labels, and replacing neutral or polarity labels with the corresponding polarity labels. After preprocessing, the dataset contains 17024 samples for subsequent experiments.

The final emotional labeling results are shown in Table 1.

Table 1. MVSA dataset

Dataset	All	Positive	Neutral	Negative
MVSA-Single	4511	2683	470	1358

MVSA-Multi	17024	11318	4408	1298

4.2 Experimental Environment and Parameters

The experiments were conducted using Python 3.6 and the PyTorch 1.9 framework. The hardware conFigureuration includes an Intel Core i5 12600KF CPU and an NVIDIA GeForce RTX 3080 GPU.

The MVSA dataset was divided into training, validation, and testing sets in an 8:1:1 ratio. For the text part, Bert-base was used as the pre-trained model to extract feature vectors. The dimensionality of the word embeddings was set to 768, and the maximum text length was set to 64. For the image part, the images were resized to 224x224 before using the pre-trained ResNet152 model to extract feature vectors.The learning rate was set to lr=5e-5, weight decay was set to 1e-2, and dropout was set to 0.4.

4.3 Baseline Model

To demonstrate the effectiveness of the ACMSA model, we will compare it with the following baseline models, all of which have been experimentally proven to outperform traditional single- modal classification methods:

(1)SentiBank + SentiStrength: This model uses SentiBank to extract 1200 Adjective-Noun Pairs (ANP) as the mid-level representation of images for classification. SentiStrength calculates sentiment scores based on English grammar and spelling style. SentiBank + SentiStrength combines the results of SentiBank and SentiStrength to handle multi-modal sentiment classification on the Twitter dataset.

(2)CNN-Multi: This model utilizes pre-trained CNNs to extract feature vectors for both text and images. The extracted text and image features are then concatenated and passed through four fully connected layers for interaction fusion.

(3)DNN-LR: Similar to CNN-Multi, this model also uses pre-trained CNNs to extract feature vectors for text and images, which are then combined as inputs to a logistic regression model.

(4)CNN-Multichannel: This model employs a multi-channel convolutional network where each filter is applied to multiple channels. However, gradients are only backpropagated through one of the channels.

(5)HSAN: This model generates semantic image titles and proposes a hierarchical attention network to simultaneously process text and image titles from the Twitter dataset.

(6)MultiSentiNet: This model extracts semantic visual information from images, including background environment and target objects. It proposes an LSTM model based on visual information and attention mechanisms, integrating visual information features with text features.

(7)CoMN: This model uses graph-text representation vectors to query key feature maps and iteratively updates query graph-text memory matrices to retrieve key information between images and text.

(8)FENet: Unlike CoMN's coarse-grained attention mechanism, this model utilizes an Inter- Intra Fusion (IIF) mechanism to learn fine-grained cross-modal attention, extracting image- based text feature vectors and text-based image feature vectors, and using SIE to extract feature representations for classification.

(9)ITIN: This model is based on an interaction network for aligning image regions and textual information. It proposes a graph-text alignment module to capture fine-grained key information between images and text and utilizes a cross-modal gate module to prevent negative effects from misalignment.

4.4 Result analysis

Based on the evaluation metrics of Accuracy (Acc) and F1 score on the MVSA-Single and MVSA-Multi datasets, the comparative results are shown in Table 2. According to the experimental data, it can be observed that the performance of the proposed model in terms of Acc and F1 on these two datasets is superior compared to the other models.

Table 2. The results of Comparison test

Numble	MVSA-Single		MVSA-Multi	
	Acc	F1	Acc	F1
StB+StS	52.05	50.08	65.62	55.36
CNN-Multi	61.2	58.37	66.39	59.47
DNN-LR	61.42	61.03	67.86	66.33
CNN-Multichannel	65.19	62.55	65.57	63.24
HSAN	66.83	66.9	68.16	67.76
MultiSentiNet	69.84	69.63	68.86	68.11
CoMN	70.51	70.01	68.92	68.83
FENet	74.21	74.06	71.46	71.21
ITIN	75.19	74.97	73.52	73.49
ACMSA	77.08	75.64	74.42	73.71

SentiBank+SentiStrength is a feature-based model that utilizes SentiBank and SentiStrength for sentiment analysis. However, this model performs worse than other models. Both CNN-Multi and DNN-LR use deep learning convolutional neural networks to output feature vectors, and they both utilize pre-trained CNNs to extract text and image feature vectors. They outperform SentiBank+SentiStrength on both datasets, with CNN-Multi achieving a precision increase of 9.15% and 0.77%, and DNN-LR achieving a precision increase of 9.37% and 2.24%, respectively, on the MVSA-Single dataset.

CNN-Multichannel is a dual-channel model with an embedding part using pre-trained word vectors. It demonstrates the effectiveness of the multi-channel model by achieving a precision increase of 3.77% compared to DNN-LR on the MVSA-Single dataset, although it falls slightly short on the MVSA-Multi dataset.

HSAN and MultiSentiNet both employ deep learning recurrent neural networks to output feature vectors, considering both visual information in images and semantic information in text. These models show better performance in experimental results, achieving higher accuracy. The recurrent neural

network improves model performance by enhancing contextual semantic information, proving particularly effective in MultiSentiNet, which considers the influence of visual information on textual semantics. MultiSentiNet outperforms HSAN by 3.01% and 0.70% in accuracy on the two datasets, respectively.

To enhance the interaction between textual and visual information, CoMN integrates text representation vectors into images to locate key local features and vice versa. CoMN learns deeper multi-hop feature representations through stacked models, demonstrating better interaction compared to MultiSentiNet with accuracy increases of 0.67% and 0.06% on the two datasets, respectively.

FENet learns cross-modal features using IIF and extracts feature representations through SIE, while ITIN enhances cross-modal interaction with alignment and misalignment modules and combines image context information with text context information. Both models demonstrate the importance of fine-grained cross-modal interaction in multi-modal sentiment classification.

Improving upon these baseline models, the proposed ACMSA model outperforms all baseline models in terms of performance. ACMSA utilizes the Bert pre-trained model for text feature extraction, which better captures semantic information by contextualizing text. Leveraging the advantages of the dual-channel structure and recurrent neural networks, ACMSA employs CNN channels for local feature extraction in text and utilizes BiGRU and self-attention mechanisms to compose a dual-channel model for global feature extraction. The dual-channel model demonstrates the ability to capture long-distance dependencies in sentiment classification tasks and efficiently capture key phrases. For image feature extraction, ACMSA employs the ResNet pre-trained model, which effectively resolves gradient issues and captures detailed features better than traditional CNN networks, thus improving network expressiveness. By introducing the CSAM module, ACMSA further enhances feature expression capabilities in both channel and spatial dimensions. The Co-Attention mechanism is used to enhance the interaction between different modalities and reduce redundant information during training. Overall, experimental results demonstrate that the ACMSA model performs well in multi-modal sentiment classification tasks.

4. Conclusion

The ACMSA model addresses the deficiencies in the current field of multi-modal sentiment classification by improving the modality interaction. Enhancing the feature extraction capabilities of both images and text can improve multi-modal performance. For images, the use of appropriate attention modules can effectively focus on important information in the images, increasing network performance and efficiency. For text, it is equally important to enhance semantic capture by contextualizing text to obtain sentiment polarity.

There are also several limitations in this study. For example, there was no in-depth research on extracting the semantics of sarcasm, especially regarding the preprocessing methods for punctuation, which affected the accuracy of text features. Further research could delve deeper into processing images, such as separating facial expressions from backgrounds or extracting text embedded in images. Additionally, future research could expand into the audio and video domains to further explore multi-modal directions.

References

[1] 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.

[2] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6558-6569, 2019.

[3] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L. P. Morency and E. Hoque, "Integrating Multimodal Information in Large Pretrained Transformers," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2359-2369, 2020.

[4] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, "CBAM: Convolutional Block Attention Module," Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19, 2018.

[5] Z. Zhu, Y. Yan, R. Xu, Y. Zi and J. Wang, "Attention-Unet: A Deep Learning Approach for Fast and Accurate Segmentation in Medical Imaging," Journal of Computer Science and Software Applications, vol. 2, no. 4, pp. 24-31, 2022.

[6] N. Xu, W. Mao and G. Chen, "A Co-Memory Network for Multimodal Sentiment Analysis," Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 929-932, 2018.

[7] 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.

[8] T. Jiang, J. Wang, Z. Liu and Y. Ling, "Fusion-Extraction Network for Multimodal Sentiment Analysis," Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 785-797, 2020.

[9] 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, vol. 25, pp. 3375-3385, 2022.

[10] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria and L. P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2236-2246, 2018.

[11] G. Cai and B. Xia, "Convolutional Neural Networks for Multimedia Sentiment Analysis," CCF International Conference on Natural Language Processing and Chinese Computing, pp. 159-167, 2015.

[12] 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.

[13] H. Liu, "Structural Regularization and Bias Mitigation in Low-Rank Fine-Tuning of LLMs," 2023.

[14] Z. Qiu, "A Multi-Scale Deep Learning and Uncertainty Estimation Framework for Comprehensive Anomaly Detection in Cloud Environments," Transactions on Computational and Scientific Methods, vol. 3, no. 2, 2023.

[15] B. Barlocker and X. Yan, "Contrastive Representation Learning for Anomaly Detection in Cloud-Based Backend Services," 2021.

[16] Y. Xing, "Enhancing Advertising Recommendation Performance via Integrated Causal Inference and Exposure Bias Correction," 2023.

[17] M. Wang, "Multi-Level Attention and Sequence Modeling for Dynamic User Interest Representation in Real-Time Advertising Recommendation," 2023.

[18] S. Pan, T. Hu, S. Sun, J. Yuan and J. Luo, "Help Oneself in Helping the Others: The Ecology of Online Support Groups," Proceedings of the IEEE International Conference on Big Data, pp. 2418-2427, 2019.

[19] A. M. Jones et al., "USC-DCT: A Collection of Diverse Classification Tasks," Data, vol. 8, no. 10, p. 153, 2023.