A Multimodal Sentiment Classification Study Based on Mvsa-Single Dataset

Yuxiang Kong

College of Mathematics and Informatics & College of Software Engineering, South China Agricultural University, Guangzhou, Guangdong, China.

Abstract. Multimodal sentiment classification refers to the task of analyzing and identifying textual, visual, and other multimodal data using deep learning techniques. This study conducts an in-depth exploration of the single-stream multimodal emotion classification model based on the intermodal attention mechanism, the dual-stream multimodal emotion classification model, the MCAN emotion classification model, and the original emotion classification model without the addition of the intermodal attention mechanism. The architecture, training method and data preprocessing strategy of each model are elaborated. This paper compares and analyze the performance of each model on the Multi-View Sentiment Analysis (MVSA) dataset in terms of accuracy, precision, recall, AUC and PR curve. The results show that the different models have their own advantages and disadvantages, the single-stream model has an advantage in recall but lower precision, the dual-stream model shows a unique ability in precision but poorer performance in recall, and the MCAN model has a better overall metrics, and all the three models have a higher accuracy than the original model. This study provides a valuable reference for the selection and optimization of models in the field of multimodal sentiment analysis, and also points out the direction for further exploring the application of multimodal fusion technology.

1 Introduction

Multimodal refers to the technology of acquiring and fusing information through multiple perceptual channels [1]. Sentiment analysis is a subfield of natural language processing that uses algorithms to identify, extract and quantify subjective emotional tendencies in text [2]. Multimodal sentiment analysis refers to the technology of combining multiple data modalities, such as text, image and speech, to comprehensively determine the emotional state, and multimodal sentiment analysis can solve the problem of a single-modal sentiment analysis not being able to adapt to the complexity and limitations of scenarios, and help us to comprehensively understand the emotion [3, 4, 5]. In the field of multimodal sentiment analysis, numerous studies have been devoted to exploring more effective model architectures and fusion methods. Early studies mainly used simple feature splicing or

kyx@stu.scau.edu.cn

decision fusion strategies, where features from different modalities are directly connected or merged at the decision layer. However, with the development of deep learning, neural network-based multimodal fusion methods have gradually become mainstream [6].

The Multi-View Sentiment Analysis (MVSA) dataset, as a commonly used benchmark dataset in multimodal sentiment analysis research, contains the rich image and textual modal sentiment information, which provides an ideal platform for evaluating the performance of multimodal sentiment classification models. Many studies have achieved certain results on the MVSA dataset. Some studies have improved the accuracy and recall of sentiment classification by improving the model architecture and training methods. However, different models still have their own limitations when dealing with multimodal data, which require further research and improvement.

The purpose of this study is to compare and analyze the performance of the single-stream multimodal sentiment classification model, the dual-stream multimodal sentiment classification model, the MCAN sentiment classification model, and the original model on the MVSA dataset, and to discuss the advantages and disadvantages of the different model architectures and fusion strategies in depth, in order to provide a reference for the further development of multimodal sentiment analysis.

2 Data sets and methods

2.1 Data sets

The MVSA dataset is used as the experimental data in this study. The dataset contains a large number of image and text pairs, where the images show various scenes and emotional expressions, and the texts are descriptions or relevant comments on the images [7]. The sentiment labels in the dataset are categorized as positive and negative, providing a clear annotation for the sentiment classification task.

2.2 Data pre-processing

On the preprocessing of the image, this paper first resizes it to make its size uniform as 224×224 pixels. Then it is normalized to scale the pixel values to the range of [0, 1]. At the same time, data enhancement techniques are used to increase the diversity of data and improve the generalization ability of the model.

On the preprocessing of text, this paper first uses BertTokenizer to process the text into fixed-length sequences. If the length of the sequence is not enough, this paper uses the padding operation, for the length of the sequence, this paper is more than the truncation operation.

2.3 Methodology

The single-stream multimodal sentiment classification model processes multimodal data by first performing feature extraction on the image and text using separate image encoders and text encoders, respectively. In this process, the model employs an attention mechanism to help the model dynamically focus on the parts of the features in the image and text that are most relevant for sentiment classification [8]. Subsequently, the two modal features that have been filtered and enhanced by the attention mechanism are fused through the feature fusion layer and finally fed into the classifier to complete the sentiment classification.

The dual-stream multimodal sentiment classification model also extracts image and text features with the help of independent image encoders and text encoders, and also applies

the attention mechanism to focus on the key sentiment information within the respective modality during the feature extraction phase, but unlike the single-stream model, it does not immediately fuse the features of the two modalities. Instead, separate classifiers are set up for image features and text features, each performing initial sentiment classification first. Finally, the results of these two classifiers are then fused to obtain the final sentiment classification results[9].

MCAN sentiment classification models are constructed based on specific cross-modal interaction mechanisms. During image and text feature extraction, the attention mechanism not only focuses on the important information within each modality, but also works at the cross-modal level. It is specially designed with a series of modules to facilitate the flow of information between image and text modalities, guiding the model to search for relevant cues for emotion classification in the information of different modalities, and continuously strengthening the understanding of multimodal emotion information[10]. Compared with the single-stream and dual-stream models, the MCAN model has a more complex structure and is able to deeply mine the potential information in multimodal data.

The original model used as a comparison is an early and simpler multimodal fusion model. It lacks an attention mechanism and cannot dynamically focus on key information like the single-stream multimodal sentiment classification model and MCAN sentiment classification model. The primitive model usually simply fuses the features of image and text, without the ability to distinguish and filter the importance of different features, and it is difficult to accurately capture the information related to sentiment classification when dealing with multimodal data, resulting in a relatively weak performance in sentiment classification.

3 Experimental

In this paper, through the strict sample screening method, the samples in the MVSA-Single dataset with too large differences in sentiment polarity are excluded, and among the remaining graphic pairs, if one of the sentiments of the picture or text is neutral (NEUTRAL), the other positive or negative label is selected as the sentiment label of the graphic pair, and 80% of the processed data is used as the training set, and 20% of the data is used as the test set, The training set goes into pre-training of single-stream multimodal sentiment classification model, dual-stream multimodal sentiment classification model, MCAN sentiment classification model and original sentiment classification model respectively, and after the completion of the training, the performance of different sentiment classification models on the test set is evaluated.

3.1 Experimental Configuration and Assessment Metrics

In this study, the Adam optimizer is used to train each model by setting the appropriate learning rate (e.g., 2e-5) and using an exponentially decaying learning rate scheduler, which gradually decreases the learning rate during the training process. To avoid overfitting, Early Stopping and Model Checkpoint callback functions are also used to save the best performing model.

This experiment uses the TensorFlow 2.x framework, accelerated by NVIDIA RTX 3090GPUs, with pre-training weights for the image encoder model VIT downloaded from the Hugging Face CDN, and the text encoder model Bert loaded from the Transformers library of Hugging Face.

The evaluation metrics in the model evaluation stage are categorized into accuracy, precision, recall, AUC and PR graphs. Among them, accuracy reflects the proportion of samples correctly predicted by the model; precision indicates the proportion of samples

predicted by the model to be positive classes that are truly positive; recall indicates the proportion of samples that the model is able to correctly predict as a percentage of all positive samples; AUC measures the model's ability to discriminate between positive and negative samples; and PR graphs illustrate the performance tradeoffs of the model under different thresholds.

3.2 Comparison of model performance and analysis of results

After training and evaluation, the performance of each model on the MVSA-Single dataset is shown in Table 1.

Table1 Comparison of Accuracy, Precision, Recall and AUC Metrics of Multimodal Sentiment Classification Models

Model name	Accuracy	Precision	Recall rate	AUC
Single-stream multimodel sentiment classification model	0.7948	0.8226	0.8808	0.8511
Dual-stream multimodel sentiment classification model	0.7973	0.8708	0.8156	0.8461
MCAN sentiment classification model	0.7985	0.8528	0.8417	0.8619
Original model	0.7824	0.8361	0.8361	0.8514

As shown in Table 1, the single-stream multimodal sentiment classification model excels in terms of recall, reaching 0.8808, which indicates that the model is able to identify the positive example samples in the sentiment category in a more comprehensive way, and has an advantage in capturing the completeness of the sentiment information. However, its precision rate is only 0.8226, which is relatively low, implying that the model has a certain proportion of misjudgment in the samples judged as positive cases, and may incorrectly identify some negative cases as positive cases. This may be because the single-stream model pays more attention to single-modal information, and is able to more fully explore the clues related to emotion in the modality, which leads to a better performance in the recall of positive case samples, but it also lacks the key information of the other modalities, and relies too much on single-modal information, and thus has an advantage in capturing the completeness of emotional information. Key information and over-reliance on a single modality, lead to model misclassification and decreased precision.

The precision rate of the dual-stream multimodal sentiment classification model is high, reaching 0.8708, but the recall rate performs poorly, only 0.8156, which indicates that the model has a short board in the feature fusion layer, and the inter-modal feature interactions are not deep enough, affecting the overall performance. In this regard, the fusion strategy based on the attention mechanism can be explored, so that the image and text features can focus on each other's key information and realize accurate fusion. Moreover, the image and text encoder structures can be reasonably adjusted, such as deepening the image encoder

hierarchy and optimizing the text encoder architecture, to improve the accuracy of feature extraction and thus effectively improve the recall rate.

The overall performance of the MCAN model is relatively stable because it adopts a modular architecture with a clear division of labor and independent optimization of each module, which gives the model a high degree of flexibility and scalability, and enables it to flexibly adjust and replace the modules for different multimodal tasks and adapt to diverse data and task scenarios in order to improve the overall performance.

The original model has a low accuracy rate in multimodal sentiment classification due to its simplistic structure or processing. In order to improve its performance, it can be optimized for the unimodal part, and a more advanced image model and fine-tuned pre-trained language model can be selected to improve the feature extraction and classification ability, and finally achieve the effective improvement of the accuracy rate.

In addition, this experiment also investigated the PR graphs for each model, as shown below.

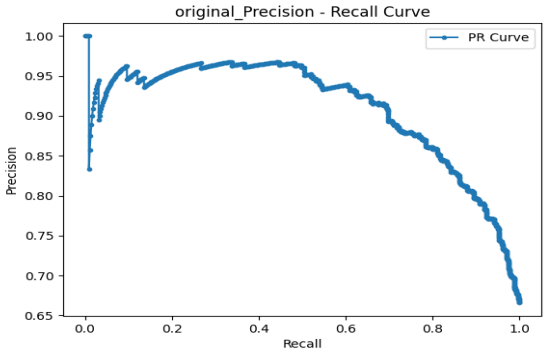

Figure 1 Original model PR curve (Picture credit: Original)

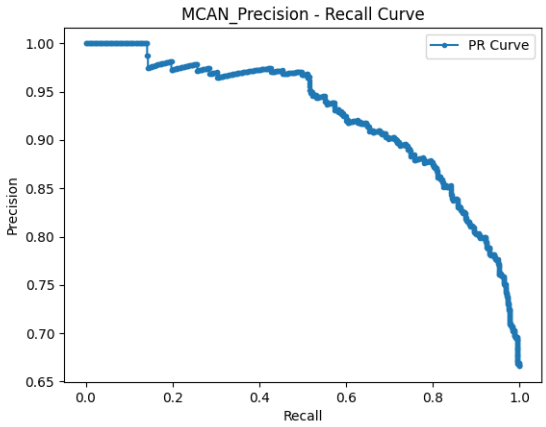

Figure 2 MCAN model PR curve (Picture credit: Original)

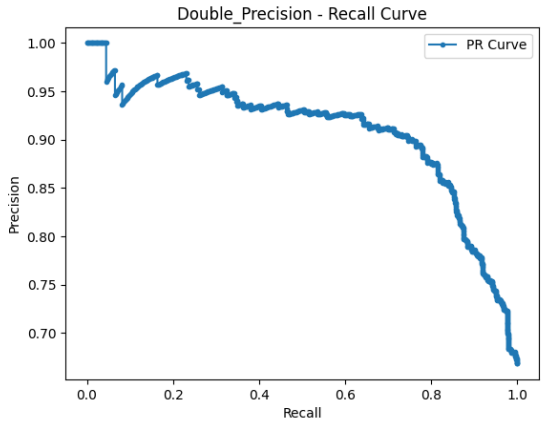

Figure 3 Dual-stream model PR curve (Picture credit: Original)

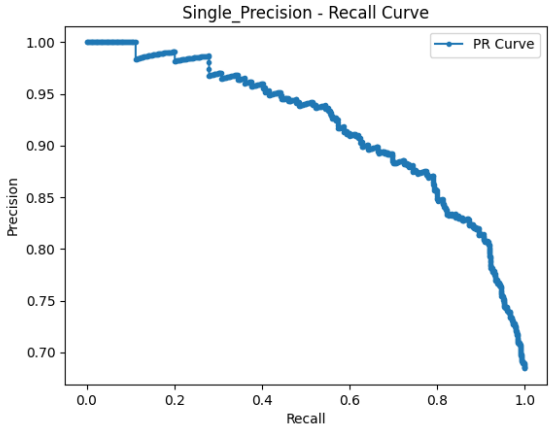

Figure 4 Single-stream model PR curve (Picture credit: Original)

As can be seen from Figure 1, the original model's PR curve fluctuates significantly at the beginning when the recall rate starts to rise. This indicates that as the model begins to try to identify more positive examples, its precision rate is unstable, and the model is not consistent enough in its judgment criteria for positive examples, judging some less certain samples as positive examples in some cases, leading to high and low precision rates.

As can be reflected in Figure 2, the precision rate of the PR curve of the MCAN model is significantly higher than the other three PR curves in Figures 1, 3, and 4 in the interval of the recall rate of about 0.2 - 0.6. This reflects the advantage of the cross-attention mechanism in the MCAN model, which can effectively fuse the multimodal information and accurately screen out the positive examples, which ensures a certain recall rate and maintains a high precision rate at a medium recall level.

Figure 3 indicates that the precision rate of the dual-stream model gradually decreases as the recall rate gradually increases, and the overall precision rate level is slightly higher than that of the Single model in the middle stage, but lower than that of the MCAN model. This may be due to the fact that although the dual-stream model can handle different modal

data separately, it is not as effective as MCAN's cross-attention mechanism in the modal fusion stage, and it is slightly inferior in balancing the precision rate and recall rate.

Figure 4 shows that the single-stream model has a relatively smoother decreasing trend in precision rate with increasing recall, but the overall precision rate is lower at all recall stages relative to the PR curves in Figures 1, 2, and 3. This reflects that the single-stream model, due to its relatively simple modal fusion approach, has difficulty in identifying positive examples as accurately as more complex models such as MCAN when dealing with multimodal data, and performs poorly in balancing precision and recall. In addition, when the recall rate is close to 1, the precision rate of the single-flow model drops to a very low level, indicating that the single-flow model introduces a large number of misclassifications when attempting to fully recall positive examples, and the precision rate is difficult to guarantee.

4 Model Optimization and Discussion

For the single-stream multimodal sentiment classification model, in order to improve the accuracy rate, future attempts can be made to increase the generalization ability of the model, such as further optimizing the data enhancement strategy and introducing more diverse training data. Meanwhile, the parameters of the cross-modal attention module can be adjusted so that it can capture the sentiment information in different modalities more effectively and reduce the omission of positive class samples.

For the dual-stream multimodal sentiment classification model, the design of the feature fusion layer can be optimized to explore more effective fusion methods, such as the fusion strategy based on the attention mechanism, in order to improve the interaction and fusion effect between different modal features. In addition, the structure of the image encoder and text encoder is adjusted to improve the accuracy of feature extraction.

For the MCAN emotion classification model, the cross-modal interaction module can be further optimized to enhance the information flow and fusion between different modalities. Attempts are made to introduce more complex neural network structures, such as graph neural networks, to better handle complex relationships in multimodal data.

For the original model, then people can try to add inter-modal attention to enhance the inter-modal information interaction, and at the same time, optimize the uni-modal part to improve the ability of feature extraction and classification.

However, during the research process, different models still face some challenges when dealing with multimodal data. For example, the computational resource requirements of the models are large, especially when dealing with large-scale datasets, and the problem of insufficient GPU memory may occur. To solve this problem, people can try to reduce the batch size and optimize the model structure to reduce the complexity in future research.

In addition, the alignment and fusion of multimodal data is still a problem that needs to be studied in depth. Different modal data have differences in feature representation and semantic understanding, how to effectively fuse them to improve the performance and interpretability of the model is one of the key directions for future research.

5 Conclusions

In this study, a detailed comparative analysis of the single-stream multimodal sentiment classification model, the dual-stream multimodal sentiment classification model, the MCAN sentiment classification model, and the original model was conducted on the MVSA dataset. The experimental results show that the different models have their own advantages and disadvantages in the multimodal sentiment classification task, the single-

stream model is outstanding in recall, the dual-stream model has some potential in feature fusion, the MCAN model has a better performance in overall performance, and the original model provides a basic reference for model improvement.

By exploring the optimization direction of each model, it provides ideas for further research in the field of multimodal sentiment analysis. Future research can be carried out in terms of innovation of model architecture, improvement of feature fusion methods, optimization of computational resources, and enhancement of interpretability, in order to promote the development of multimodal sentiment analysis technology and better meet the needs of practical applications.

References

1. T. Baltrušaitis, C. Ahuja and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2018.
2. B. Liu, Sentiment Analysis and Opinion Mining. Cham, Switzerland: Springer Nature, 2022.
3. H. Geng, et al., "The joint optimal filtering and fault detection for multi-rate sensor fusion under unknown inputs," Inf. Fusion, vol. 29, pp. 57–67, 2016.
4. A. Zadeh, et al., "Tensor fusion network for multimodal sentiment analysis," arXiv preprint arXiv:1707.07250, 2017.
5. T. Baltrusaitis, C. Ahuja and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," CoRR, vol. abs/1705.09406, 2017.
6. A. Gandhi, K. Adhvaryu and S. Poria, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," Inf. Fusion, vol. 91, pp. 424–444, 2023.
7. T. Niu, S. A. Zhu, L. Pang and A. El Saddik, "Sentiment analysis on multi-view social data," in Proc. Multimedia Modeling Conf., 2016, pp. 15–27.
8. W. Li and C. Gan, "Hierarchical interactive fusion based on attention mechanism for multimodal sentiment analysis," J. Chongqing Univ. Posts Telecommun. (Nat. Sci. Ed.), vol. 35, no. 1, pp. 176–184, 2023. DOI: 10.3979/j.issn.1673-825X.202106300229.
9. A. Zadeh, M. Chen, S. Poria, et al., "Tensor fusion network for multimodal sentiment analysis," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2017.
10. T. Zhang, Q. Guo, Z. Li and L. Deng, "MC-CA: Multimodal sentiment analysis based on modal temporal coupling and interactive multi-head attention," J. Chongqing Univ. Posts Telecommun. (Nat. Sci. Ed.), vol. 35, no. 4, pp. 680–687, 2023. DOI: 10.3979/j.issn.1673-825X.202205090107.