Towards Robust Multimodal Sentiment Detection with LLaVA

Jiajun Cao(38120241150231),[1] **Zijie Bian(36920241153195),**[1] **WeiHuang Lin(31520241154513)**[1]

AI class, AI Class, Information Class

Abstract

With the growing popularity of social media, detecting sentiment from multimodal posts has attracted increasing interest. However, the task of multimodal sentiment detection, which involves analyzing sentiment from paired image-text data, presents challenges due to the inherent disparities between visual and textual modalities. Traditional approaches typically require feature fusion strategies and additional calibration mechanisms to handle modality heterogeneity, often leading to issues like feature shift, redundant visual information, and annotation inconsistencies. Furthermore, traditional methods often rely on smaller models without pre-trained cross-modal knowledge, which limits their ability to capture complex interdependencies between text and image, resulting in less accurate or generalized sentiment predictions. In this work, we propose leveraging the LLaVA (Large Language and Vision Assistant) framework, a pre-trained multimodal large language model, to address these challenges by exploiting its rich cross-modal knowledge and generative capabilities. Unlike conventional methods, LLaVA's architecture inherently bridges visual and textual domains, reducing the need for complex fusion layers and manual feature calibration. Our approach incorporates a filtering mechanism based on the attention distribution of the vision encoder, allowing us to selectively retain essential visual tokens, which accelerates inference and enhances the model's performance by reducing the processing of redundant visual information. The results suggest that multimodal large language models (MLLMs) hold potential for improving sentiment analysis in diverse multimodal content.

Introduction

With the rapid growth of social media, understanding sentiment embedded in multimodal content (e.g., text, images) has become a critical task, with broad applications in opinion mining, user engagement analysis, and market research. Compared to unimodal sentiment detection, where the focus is on a single modality like text, multimodal sentiment detection offers richer information but also presents unique challenges due to the diverse nature of the data. Detecting sentiment with only text or image modality might not be informative enough to fully understand the true intention behind the content, due to the complexity of human language and social context. Text alone may convey subtle tones, sarcasm, or implied meanings that require contextual understanding, which an image cannot provide. Similarly, an image might depict a neutral or ambiguous visual scene, where the emotional intent is only fully realized when combined with the textual context. Using both text and image modalities allows for a richer and more accurate understanding of the sentiment, as each modality provides complementary information that helps resolve ambiguities and better capture the true intention of the tweet. The combination of textual and visual elements can provide deeper insights into sentiment, yet it increases complexity in terms of data fusion and interpretation. As a result, sentiment detection from multimodal posts, especially those involving image-text pairs, has attracted significant attention from both academic and industrial communities.

(a) "Every day is a good hair day". Yeah, right!!(positive)

(b) Hot Fudge & Whipped Pancakes.(positive)

(c) The 21 Deadliest Ani(neg)

(d) washed again.(neutral)

Figure 1: Examples of multimodal sentiment detection

In this work, we explore the use of Multimodal Large Language Models, particularly LLaVA-v1.5 (Liu et al. 2023, 2024), for sentiment detection in social media posts. To enhance efficiency, we incorporate a straightforward yet effective token filtering mechanism. This approach selectively prunes less informative visual tokens based on the attention distributions derived from the vision encoder, this not only

speeds up inference but also, with an optimal filter ratio, enhances the model's performance.

Previous works on multimodal sentiment analysis primarily focus on fusing features from different modalities (Xu and Mao 2017; Yang et al. 2020), often relying on early or late fusion techniques. However, these methods face challenges due to modality heterogeneity where visual and textual modalities differ in information density and structure resulting in redundant or irrelevant visual information during fusion (Yu and Jiang 2019; Kumar and Vepa 2020). While some approaches, like contrastive learning (Li et al. 2022), attempt to bridge the modality gap by learning common features related to sentiment, they are often coarse-grained and fail to address the underlying challenges fully. Multi-View Calibration Network (Wei et al. 2023), mitigates Sparse Attention to filter redundant visual information, but struggles with preserving essential contextual details from the visual modality while simultaneously filtering out noise, which can hinder the model's ability to fully understand the sentiment of multimodal content.

The challenge with these methods lies in effectively handling modality fusion, as ignoring modality heterogeneity can directly lead to issues like feature misalignment or redundant information, and better modality heterogeneity across different modalities complicates the accurate detection and interpretation of sentiment in multimodal posts. Instead of using complex steps to handle modality fusion, our approach leverages multimodal large language models, which are pretrained with vast amounts of diverse data to abtain substantial knowledge also powerful capability of modalities alignment and fusion, offering inherent generalization abilities. Given LLaVA's robust foundation, we hypothesize that fine-tuning LLaVA for sentiment detection can improve performance by leveraging its inherent multimodal knowledge and cross-modality integration capabilities. However, while LLaVA demonstrates significant strengths, its large size incurs high inference costs. Therefore, we also propose an efficient token filtering mechanism to address this limitation.

In this paper, we:

- Leverage LLaVA's pre-trained capabilities for multimodal alignment and enhance it for sentiment detection tasks through targeted instruction finetuning.
- Incorporate a token filtering approach based on visual attention distributions to accelerate inference while maintaining competitive performance.
- Evaluate the effectiveness of this approach on standard multimodal sentiment benchmarks, and compare with state-of-the-art model to show the effectiveness of our approach.

Related Work

MLLMs (Liu et al. 2023; Bai et al. 2023; OpenAI et al. 2024) have recently emerged as powerful tools for integrating diverse modalities such as text, images, and video, offering a unified framework for a variety of tasks, including sentiment detection. LLaVA is a model that builds on pretrained vision and language transformers, aligning visual-linguistic information through shared embeddings. It first processes images into visual features using a frozen CLIP vision encoder (Radford et al. 2021), which are then passed through a two-layer MLP which can be considered as visual tokens. These visual tokens are concatenated with text embeddings of text input and prompt, and the combined tokens are fed into a backbone large language model(LLM), such as Vicuna (Chiang et al. 2023) and LLaMA (Touvron et al. 2023). These large pre-trained models have been trained on vast amounts of data, equipping them with extensive knowledge and enabling effective feature extraction and understanding. Thoses designs enable effective integration of both multimodal inputs and knowledge of pretrained LLM for downstream tasks such as question answering, image captioning, and sentiment analysis.

Multimodal sentiment analysis has seen rapid advancements with the rise of deep learning techniques. Early methods such as MultiSentiNet (Xu and Mao 2017) use LSTM to encode texts and images to get hidden representations, then concatenate texts and images hidden representations to fuse multimodal features, Yu and Jiang (2019) introduces an aspect-sensitive attention and fusion network designed to effectively model both intra-modality interactions, such as aspect-text and aspect-image alignments, as well as inter-modality interactions. Similarly, MVAN (Yang et al. 2020) employs interactive learning of text and image features through an attention memory network module, while the multimodal feature fusion module is built using a multilayer perceptron combined with a stacking-pooling mechanism. Kumar and Vepa (2020) also explored using gating mechanisms and attention to perform deep multimodal interactions.While contrastive learning approaches like CLMLF (Li et al. 2022) ttempt to bridge the modality gap by learning common features related to sentiment, challenges in aligning and fusing features still remain. Recently, the Multi-View Calibration Network (MVCN) (Wei et al. 2023) has been introduced to tackle these issues. It includes a text-guided fusion module and Sparse-Attention to filter out redundant visual information, along with sentiment-based congruity constraints, showing state-of-the-art performance.

Proposed Solution

To address the challenges inherent in multimodal sentiment detection, we propose a solution that leverages the pretrained capabilities of LLaVA 1.5, a large multimodal language model that has been trained on a vast corpus of multimodal data. The primary aim of our solution is to enhance sentiment detection by reducing the complexities involved in modality fusion, and accelarate model inference speed while also improve model's performance with optimal settings by reducing visual redundancy. Below, we outline the key components and steps of our proposed solution:

Multimodal Fine-Tuning with LLaVA 1.5

Our approach involves fine-tuning the LLaVA 1.5 model, which is pretrained on both visual and textual data, for the specific task of sentiment analysis on the MVSA-Single dataset. The fine-tuning process allows the model to adapt

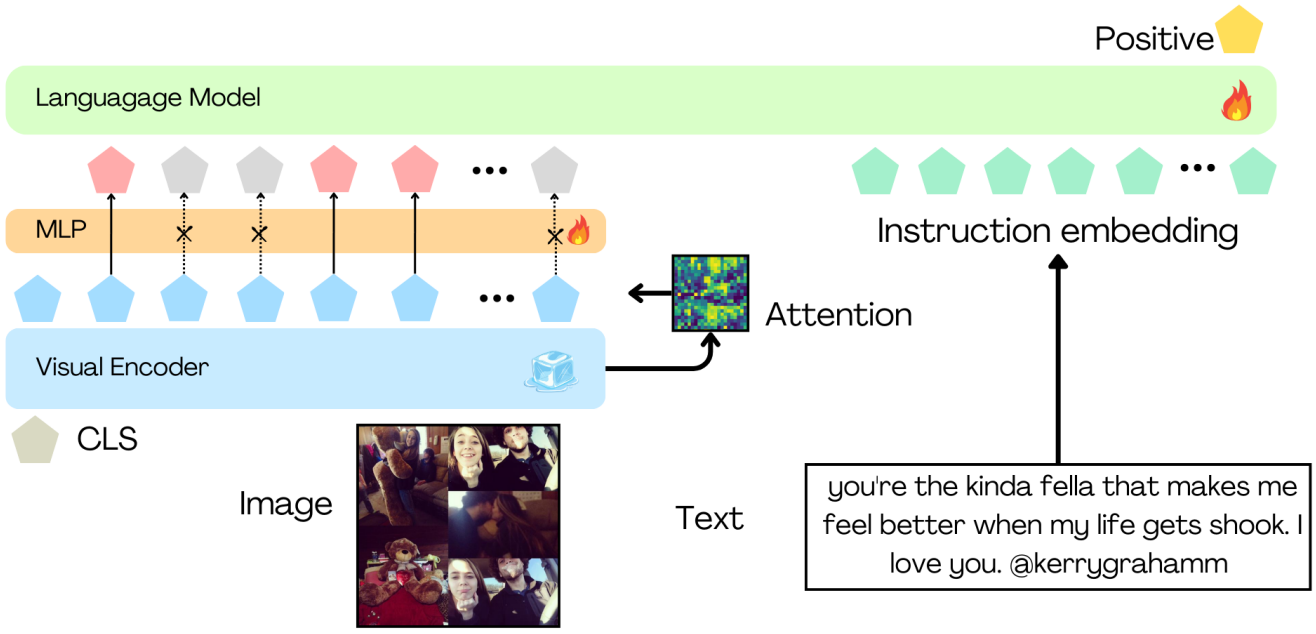


Figure 2: Pipeline of our approach: the Vision Transformer encoder processes visual tokens, we use the CLS token's attention scores across all layers are used to evaluate the informativeness of visual tokens. Tokens with higher attention scores are retained and concatenated with instruction's embedding tokens, the backbone large language model process the input and generate the label. Note that We utilize a frozen visual encoder, only fine-tuning the MLP, which acts as an adapter, along with the large language model. Additionally, during the training phase, we do not apply the token filtering strategy; this strategy is only used to enhance model inference speed and performance during prediction.

its knowledge of multimodal data to the specific sentiment labels in the dataset, which include positive, neutral, and negative sentiment categories. LLaVA 1.5's architecture enables the alignment of visual and textual modalities through shared embeddings, which reduces the need for complex fusion mechanisms typically required in traditional multimodal models.

Concretely, for each image X_v and the corresponding text X_t , the model is trained to generate the sentiment label Y_s (i.e., positive, neutral, negative). We generate one single turn conversation data $(X_{\text{instruct}}, Ys)$, in which X_{instruct} structured as shown in Table 1.

$X_{\text{system-message}}$ <STOP> User: X_{v}<STOP>X_{t}<STOP>X_{p} Assistant: Y_{s}<STOP>

Table 1: Structure of X_{instruct} for single turn conversation. X_p indicates the question prompt "What is the sentiment of the given text?"

We perform instruction-tuning of the language model on the prediction labels using the original auto-regressive training objective. Specifically, for a sequence of length L, we calculate the probability of the target labels Y_s as:

$$p(Y_s|X_{\text{instruct}}) = \prod_{i=1}^{L} p_\theta(Y_s|X_{\text{instruct}}, < i) \quad (1)$$

Reduction of Redundant Visual Features

Vision tokens are first processed by a vision encoder. An input image is divided into a grid of patches. Each patch is converted into a token embedding by the ViT, and a learnable CLS token is added before these patches to compute global image information. In a Transformer, the input tokens are processed through the multi-head self-attention mechanism, defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \quad (2)$$

where: Q is the query matrix, K is the key matrix, V is the value matrix, d_k is the dimensionality of the key vectors.

In the final layer of the Vision Transformer (ViT), the learnable CLS token is used for classification. The attention between the CLS token and other visual tokens is computed using the attention mechanism, defined as:

$$a_{\text{CLS}} = \text{softmax}\left(\frac{q_{\text{CLS}} \cdot K^T}{\sqrt{d_k}}\right) \quad (3)$$

Dataset Label	Train	Val	Test
Positive	2146	268	269
Neutral	376	47	47
Negative	1086	136	136

Table 2: Dataset statistics for MVSA-Single

where: q_{CLS} represents the query corresponding to the CLS token, K are the key matrices derived from the visual tokens, d_k is the dimensionality of the key vectors.

The output of the Key-Query attention between cls token and visual tokens can serves as an indicator of informativeness of visual tokens and for identifying crucial visual tokens.

To accelerate inference and potentially enhance model performance, we propose a token filtering strategy for the visual input. Specifically, we utilize the attention scores from the Vision Transformer encoder, focusing on the CLS token's attention distribution across all layers. For each patch token, we compute the average attention score it receives from the CLS token throughout the layers as we have calculated in equation(3).

Tokens with higher scores are deemed more informative and retained, while a fixed proportion of low-attention tokens is filtered out. The selected tokens are then input to the multimodal pipeline. Notably, this approach not only reduces inference time but can also improve performance by eliminating redundant information with appropriate filter.

Experiments

Dataset and Preprocessing

We conduct our experiments on the MVSA-Single dataset (Xu and Mao 2017), a widely used text-image sentiment dataset containing three sentiment classes: positive, neutral, and negative. Here, we give a brief introduction to the dataset and dataset statistics are shown in Table 2. Following the preprocessing steps outlined in previous works (Xu and Mao 2017), We exclude tweets where one label is positive and the other is negative. In cases where one label is neutral and the other is either positive or negative, we assign the sentiment polarity of the multimodal tweet to be either positive or negative, depending on the non-neutral label. This approach results in the creation of a refined version of the MVSA-Single dataset, containing a total of 4,511 text-image pairs. Subsequently, we shuffle the data and split it into training, validation, and test sets in an 8:1:1 ratio, the metrics used to evaluate model's performance in our experiment are accuracy and Weighted F1-measure.

Training Setup

Our model is based on the LLaVA 1.5 architecture. We finetune the pre-trained llava-vl.5-7b model on MVSA-Single using a learning rate of 2e-5, a batch size of 16, and train for 10 epochs as (Wei et al. 2023) using in the training phase, we employ a frozen visual encoder to process input images and only finetune the adapter and the large language model.

Table 3: Comparison of accuracy and weighted F1 (%) across baselines on MVSA-Single.

Model	Accuracy	Weighted F1
MultiSentiNet	69.84	69.84
MGNNS	73.77	72.70
CLMLF	75.33	73.46
MVCN	76.06	74.55
Our Model (LLaVA 1.5)	**93.36**	**93.76**

Ratio	Accuracy	F1 Score	Speedup (×)
1.00	0.933	0.937	1.00
0.80	0.938	0.940	1.15
0.60	0.933	0.938	1.40
0.40	0.933	0.934	1.75
0.20	0.893	0.896	2.33
0.05	0.823	0.817	3.16

Table 4: Performance and inference speed at different token retention ratios.

This means that the parameters of the visual encoder are kept fixed and not updated during backpropagation. The optimal model is selected based on the validation set performance, corresponds to the sixth epoch. During fine-tuning, All experiments are conducted under consistent settings to ensure comparability.

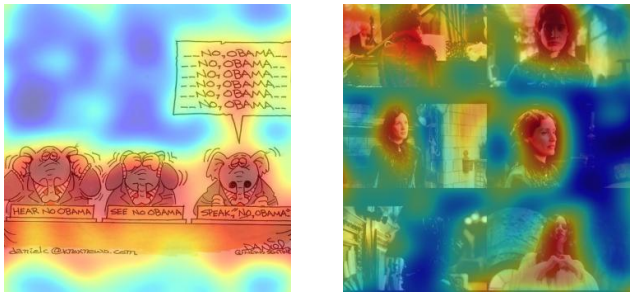
Baselines

To evaluate the performance of our model, we compare it with several multimodal baselines, including MultiSentiNet (Xu and Mao 2017), MGNNS (Yang et al. 2020), and CLMLF (Li et al. 2022), MVCN (Wei et al. 2023), which align and fuse modalities through multi-View calibration and sparse attention.

Results and Analysis

The evaluation metrics include accuracy and weighted F1, following common practices for MVSA datasets. Table 3 demonstrates the comparison between our model and baselines. Our fine-tuned LLaVA 1.5 achieves accuracy of 0.9336 and weighted F1 of 0.9376, significantly surpassing previous SOTA results, which emphasizes the importance of our approach in effectively leveraging the knowledge of pre-trained multimodal language model to achieve better multimodal fusion.

We also evaluated the token filtering strategy by varying the proportion of visual tokens retained, analyzing its impact on accuracy (Acc), F1 score, inference time, and the retained token ratio. We summarize the results of varying token retention ratios in Table 4, As shown, the results demonstrate that reducing redundant tokens improves inference efficiency with minimal performance degradation. Notably, retaining 80% of tokens improves both accuracy (0.938) and F1 score (0.941) compared to using all tokens (Acc: 0.934, F1: 0.938) while speed in inference is 1.15 times faster, which is effective in both efficiency and performance. We



(a) negative (b) positive

(c) positive (d) negative

Figure 3: Examples of cls token's attentions distribution over visual tokens

believe this is the outcome of reducing irrelevant tokens since redundant background tokens and padding tokens can divert the model's attention away from the crucial visual tokens. When the ratio lower down to 40%, we can speed 1.75 times faster with only a slight margin of performance cost (Acc: -0.000, F1: -0.003). These results confirm the effectiveness of our token filtering mechanism.

Visualization

We visualized the attention distribution of the CLS token as examples shows in Figure 3, revealing that the CLS token primarily focuses on the main objects in the image, particularly concentrating on facial expressions and text which are critical pieces of information in downstream tasks, especially in multimodal sentiment detection, even though the visual data does not involve any contextual text information. This indicates that the model is capable of effectively attending to critical information while paying relatively little attention to padding and background areas. This insight suggests that an appropriate filter ratio can help eliminate redundant information, thereby accelerating the process while enhancing performance. Such findings provide valuable guidance for subsequent model optimization and lightweight design.

Conclusion

In this work, we present a novel approach to multimodal sentiment detection by leveraging the pre-trained capabilities of LLaVA 1.5, a cutting-edge multimodal large language model. We explore the potential of this model to address the unique challenges inherent in multimodal sentiment analysis, specifically those arising from modality heterogeneity, visual redundancy, and the complex nature of integrating text and image data. Also, our token filter strategy significantly reduces inference time and, optimally, improves overall performance by reduce redundant information.

Our experiments demonstrate that fine-tuning LLaVA 1.5 on the MVSA-Single dataset significantly improves performance compared to traditional unimodal and multimodal methods. These results suggest that multimodal large language models, such as LLaVA, can effectively bridge the gap between visual and textual modalities in multimodal sentiment detection, overcoming issues like redundant visual features and annotation inconsistencies that are common in previous approaches.

References

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; et al. 2023. Qwen Technical Report. arXiv:2309.16609.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Kumar, A.; and Vepa, J. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4477–4481. IEEE.

Li, Z.; Xu, B.; Zhu, C.; and Zhao, T. 2022. CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 2282–2294. Seattle, United States: Association for Computational Linguistics.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 34892–34916. Curran Associates, Inc.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh; et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Touvron, H.; Lavril, T.; Izacard, G.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wei, Y.; Yuan, S.; Yang, R.; Shen, L.; Li, Z.; Wang, L.; and Chen, M. 2023. Tackling Modality Heterogeneity with Multi-View Calibration Network for Multimodal Sentiment Detection. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5240–5252. Toronto, Canada: Association for Computational Linguistics.

Xu, N.; and Mao, W. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2399–2402.

Yang, X.; Feng, S.; Wang, D.; and Zhang, Y. 2020. Image-text Multimodal Emotion Classification via Multi-view Attentional Network. *IEEE Transactions on Multimedia*.

Yu, J.; and Jiang, J. 2019. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5408–5414. International Joint Conferences on Artificial Intelligence Organization.