Multimodal Sentiment Analysis Using the FLAVA Model: Model Adjustments and Performance Evaluation

Tongle Shen
East China Normal University
10215501403@stu.ecnu.edu.cn

Abstract

This paper explores multimodal sentiment analysis using the FLAVA model, a BERT-based architecture designed for integrating text and image data. The study aims to classify sentiment in social media posts by fine-tuning a pretrained FLAVA model, employing a lightweight classifier and resampling techniques to address class imbalance. Experimental results achieved 73% accuracy. An ablation study revealed that both modalities significantly contribute to the classification, with images playing a more dominant role. While FLAVA's multimodal pretraining offers advantages, further improvements are needed, such as incorporating sentiment-specific pretraining tasks. This work underscores the potential of FLAVA for advancing multimodal sentiment analysis and suggests avenues for future enhancements.

1. Introduction

Multimodal sentiment analysis (MSA) is an emerging area in natural language processing and computer vision that aims to understand and classify sentiment expressed through multiple modalities. Traditional sentiment analysis typically focuses on text-only data, while MSA incorporates visual data to improve the interpretation of sentiment. In social media contexts, for example, users often express emotions through a combination of text and images in posts, which requires a model to jointly analyze these different types of information to understand the underlying sentiment accurately.

Several approaches have been developed to address multimodal classification tasks, with models like CLIP [4] and BERT-based architectures leading the way. CLIP utilizes a cross-product design to align visual and textual representations by maximizing the similarity between matched image-text pairs in the embedding space. However, its architecture mainly focuses on finding associations between tokens and image features, which may limit its effectiveness in complex cross-modal interactions. On the other hand, BERT-based models, such as VisualBERT [2], employ a unified attention mechanism to integrate information from both modalities. These models leverage the self-attention layers of the Transformer architecture to align features from text and images, thus enabling a deeper understanding of multimodal content. Despite their success, these solutions often struggle with fine-grained alignment and contextual interactions between different modalities.

In this work, the author explored the use of the FLAVA model [6] for multimodal sentiment analysis. FLAVA is a BERT-based architecture that has been pretrained on a variety of tasks, including both unimodal (text or image only) and multimodal (e.g., visual question answering) datasets. This extensive pretraining allows FLAVA to effectively handle the fusion of image and text data. The model processes images as sequences of patch tokens, which are concatenated with the text tokens. Special beginning and ending tokens are added to demarcate each modality, facilitating joint learning. We fine-tune the pretrained FLAVA model for our MSA task, incorporating a lightweight classifier composed of a three-layer Multilayer Perceptron (MLP) with Dropout to prevent overfitting. The design of the classifier was kept simple to avoid excessive complexity, which could diminish generalization.

The experimental setup involved fine-tuning the FLAVA model using a Multimodal Twitter Sentiment Analysis dataset.The training was performed with early stopping applied to prevent overfitting. A resampling strategy was also employed to address the significant class imbalance in the dataset, increasing the representation of minority classes. Evaluation metrics included accuracy and F1 score to assess the model's performance on the validation set.

Furthermore, an ablation study was performed to isolate the contributions of each modality. By conducting experiments with either text-only or image-only inputs, the author analyzed the individual impact of visual and textual information on the classification performance. The results indicated that combining both modalities yielded the best outcomes, with images playing a more significant role in sentiment determination than text.

This paper presents the architectural design, training strategies, and evaluation results, demonstrating the potential of the FLAVA model for multimodal sentiment analysis tasks while highlighting areas for further improvement.

2. Related works

Multimodal learning has gained significant attention in recent years, particularly in tasks that require understanding and combining information from multiple data sources, such as images, text, audio, and video. In the context of multimodal sentiment analysis, several approaches have been proposed, which primarily leverage deep learning architectures designed for fusing information from different modalities.

2.1. Early Multimodal Approaches

Early approaches to multimodal sentiment analysis often relied on feature-level fusion methods, where handcrafted features from each modality were combined before being passed to a classifier. These methods typically involved separate feature extraction techniques for text and image data, such as using convolutional neural networks (CNNs) for image features and recurrent neural networks (RNNs) for textual features. While effective to some extent, these approaches suffered from limitations in capturing complex interactions between modalities due to their reliance on shallow feature fusion and separate processing of each data type.

2.2. CLIP and Cross-Modal Contrastive Learning

Contrastive Language–Image Pre-training (CLIP) [4] is another prominent model that uses contrastive learning to align image and text representations in a shared embedding space. CLIP trains on a large-scale dataset of image-text pairs by maximizing the similarity between matched pairs and minimizing it for mismatched pairs. This approach allows CLIP to perform zero-shot classification by finding the closest image-text match in the learned embedding space. Although CLIP's cross-modal contrastive design facilitates strong generalization and transfer learning capabilities, it often struggles with tasks requiring intricate cross-modal interactions, such as nuanced sentiment analysis. The cross-product design is mainly focused on finding coarse associations between the image and text rather than capturing fine-grained contextual details.

2.3. BERT-Based Multimodal Models

The advent of the Transformer architecture and the introduction of BERT (Bidirectional Encoder Representations from Transformers) revolutionized natural language processing by enabling context-aware embeddings through self-attention mechanisms. Several BERT-based multimodal models have since been developed, aiming to extend BERT's powerful language representation capabilities to multimodal tasks. For example, VisualBERT [2] and ViLBERT [3] extend BERT by integrating visual features alongside textual embeddings using a shared attention mechanism. These models employ a unified Transformer to jointly encode image and text representations, which allows for fine-grained cross-modal attention and better alignment of features from both modalities.

Figure 1. One example for image poster in MTSA.

VisualBERT, in particular, incorporates image region embeddings generated by a pretrained object detection model (e.g., Faster R-CNN [5]) and concatenates them with the textual tokens before feeding them into the BERT model. This joint representation enables the model to capture interactions between visual regions and corresponding textual phrases, thus improving multimodal sentiment analysis performance. However, the dependency on object detection models for visual feature extraction may limit the generalizability and efficiency of these approaches.

2.4. FLAVA and Multimodal Pretraining

FLAVA [6] is a recent multimodal model that adopts the BERT architecture to jointly encode images and text. Unlike other approaches, FLAVA undergoes extensive multimodal pretraining across multiple tasks, including unimodal (e.g., Masked Language Modeling) and multimodal tasks (e.g., Visual Question Answering and Hateful Memes). This multi-task pretraining approach enables FLAVA to learn comprehensive cross-modal representations that are more adaptable to a variety of downstream tasks. The model processes images as sequences of patch tokens, concatenated with text tokens, which are then jointly fed into the Transformer layers. FLAVA's unified attention mechanism allows for deep integration of visual and textual information, improving the model's ability to understand complex multimodal relationships.

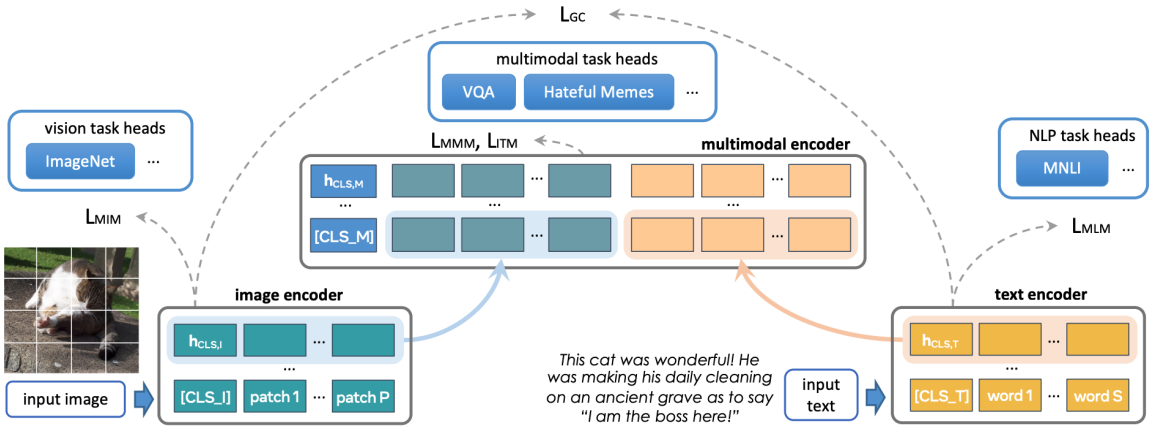

Figure 2. FLAVA base model structure.

3. Approach

3.1. Dataset

The Multimodal Twitter Sentiment Analysis(MTSA) dataset extends traditional text-only sentiment analysis by incorporating visual data alongside textual input. Each sample consists of two components: an image x_1, such as the one shown in Figure 1, and a corresponding natural language passage x_2, which could include hashtags and short phrases (e.g., "How I feel today #legday #jelly #aching #gym"). The objective is to classify the sentiment of each post, where the label y represents the sentiment: 0 for positive, 1 for negative, and 2 for neutral.

The main challenges of this task arise from the need to effectively align and integrate the two modalities—image and text—into a single model framework. First, the model must handle the inherent differences between visual and textual information, which require different processing techniques. While textual data consists of sequential word tokens, visual data involves pixel-based patterns that need to be understood in context. Therefore, aligning these two distinct types of data into a cohesive representation poses a significant challenge.

Second, it is crucial for the model to perform cross-modal information exchange effectively. The task requires not just the independent analysis of the image and text, but also the understanding of how the two modalities interact and contribute to the sentiment conveyed. For example, an image might evoke a particular mood, while the accompanying text could either reinforce or contradict that mood. Successfully capturing these interactions is essential for accurate sentiment classification.

3.2. Data preprocessing

The data preprocessing steps focused on cleaning the text data and preparing the inputs for model training. First, non-word elements such as garbled text and special characters (e.g., hashtags) were removed to ensure the input consisted only of meaningful language content. Next, all text was converted to lowercase to maintain consistency, and any non-English text was translated into English to standardize the language used across samples.

To align the inputs with the model requirements, the image data was already preprocessed to a uniform size. For the text data, each sequence was truncated to a maximum length to avoid excessive input sizes and padded with a mask at the end to maintain equal input length across the batch. This was necessary to accommodate the sequential nature of the language encoder used in the model.

3.3. Resampling

Analysis of the dataset revealed significant class imbalance, with 1,193 negative samples, 419 neutral samples, and 2,388 positive samples. Directly training the model on this unbalanced data would likely lead to poor convergence and biased predictions.

To address this, a resampling strategy [7] was employed, as weighted cross-entropy loss alone could not adequately resolve the imbalance issue. The resampling involved reselecting samples based on the inverse frequency of each class label, adjusting the probabilities for drawing samples from each class accordingly. This approach increased the representation of underrepresented classes, effectively mitigating the imbalance.

The main advantage of resampling was its ability to reduce the negative impact of severe class imbalance. However, it came at the cost of a higher training burden, as

some samples were repeatedly drawn, increasing the training dataset size by 1.9 times.

3.4. Model Architecture

3.4.1 Base Model Selection: FLAVA

In this task, I adopted the BERT-style embedding connection approach and utilized the foundational FLAVA model as Figure 2. The model architecture is designed to handle images and text as sequences of tokens. The image is divided into patches, each projected into an embedding space to form integer tokens. These tokens are then concatenated with the text tokens, and special beginning and ending tokens are added for both the image and text sequences. This setup allows the model to process multimodal inputs in a unified way.

BERT, a foundational model in natural language processing, is based on the Transformer architecture, which uses self-attention mechanisms to capture relationships between tokens in a sequence. The pretraining task, known as Masked Language Modeling (MLM), involves predicting missing words in a sentence to help the model learn language representations. The key advantage for FLAVA model is its **multimodel pretraining**. Fountation encoder was not only pretrained on the image task like ImageNet and NLP task like MNLI, after concatenation, the large encoder was then trained with multimodel task like VQA and Hateful Memes. This additional training process makes FLAVA easy to be tuned on other multimodel downstream tasks. Due to computational limitations, it was not feasible to perform pretraining from scratch for this task; instead, a pretrained checkpoint from Facebook's FLAVA model [1] was used.

The use of FLAVA offers several advantages over traditional CNN-based encoders. Unlike CNNs, which process images in a hierarchical manner by gradually abstracting features, the FLAVA model retains the original token information from both modalities throughout the initial layers. This approach significantly improves performance in downstream tasks [6], as it better preserves the fine-grained details necessary for effective cross-modal learning in early transformer layers.

3.4.2 Classifier Design

Through iterative experimentation, it was found that reducing the classifier's depth and keeping its structure simple was crucial. A more complex classifier tended to overfit the data, diminishing its generalization ability. Additionally, increasing the number of layers or connecting too many output tokens often diluted the supervisory signals.

Inspired by BERT's approach to downstream classification tasks, the FLAVA model's multimodal embedding was used, specifically the first global token, as the base model's output. A three-layer Multilayer Perceptron (MLP) with Dropout was employed to ensure that the supervisory signal was effectively transmitted to the base model, while still maintaining a lightweight and efficient design.

3.4.3 Freezing Layers

Despite the strategies employed, the model initially failed to converge. This was due to the dynamic nature of the FLAVA model's text and image encoders, which were included in the computation graph during backpropagation. As a result, severe **catastrophic forgetting** occurred, where earlier learned knowledge was overwritten during training.

To address this issue, I implemented a layer freezing technique. This involved freezing the initial layers of the encoders, preventing them from being updated during training. In this experiment, I froze all layers of the text and image encoders, while leaving the multimodal encoder unfrozen. Freezing the multimodal encoder would have significantly degraded model performance.

This approach effectively resolved the catastrophic forgetting problem, leading to faster convergence and a notable improvement in validation accuracy.

3.4.4 Training Details

The training was conducted with a learning rate of 0.0001, using the Adam optimizer and a dropout rate of 0.3. The loss function was standard Cross Entropy Loss, with a batch size of 32 for 10 epochs.

Due to the resampling strategy, each epoch contained approximately 190 batches (compared to the original 100). The validation set did not employ resampling and was evaluated directly. The data split ratio for training and validation was 8:2, and model performance was measured using accuracy and F1 score.

The experiments were carried out on a GPU: A100-PCIE-40GB, providing sufficient computational power for training.

4. Experimental results

4.1. Results

4.1.1 Overall Performance

The overall performance of the model is presented through the plots of loss in Figure 3, accuracy in Figure 4, and F1 score in in Figure 5 for the validation set. The classification task achieved 73% accuracy, with early stopping applied at epoch 2, which is 7% lower than the state-of-the-art (SOTA) performance. The model demonstrated its strongest performance at epochs 2, 3, and 6, indicating a robust ability to fit the data due to the efficient architecture design. This

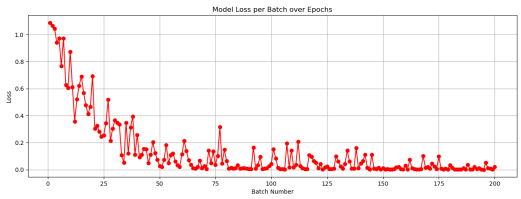

Figure 3. Loss Landscape for Validation Set

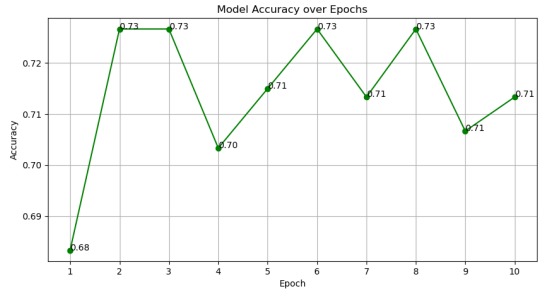

Figure 4. Accuracy for Validation Set

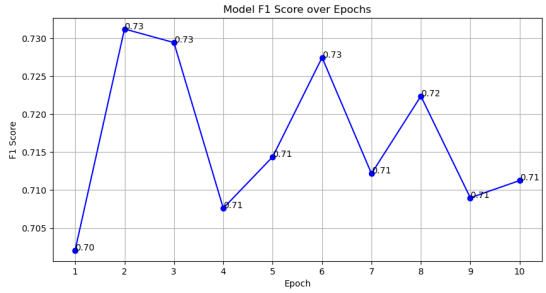

Figure 5. F1 score for validation Set

design leveraged a lightweight classifier and careful fine-tuning of the multimodal embeddings, which contributed to the model's effective learning.

Analyzing the curves, it is evident that freezing layers significantly accelerated the training and convergence speeds. This was because only the final multimodal BERT layers and the three-layer MLP were trainable, while the text and image encoders remained fixed. This approach also successfully avoided catastrophic forgetting by preserving the learned features in the encoders.

Regarding dropout and early stopping, the use of dropout helped control overfitting by adding regularization during training. However, if the number of epochs was too large, overfitting still occurred, making early stopping essential to prevent excessive training and ensure generalization.

4.1.2 Improvements

Several factors may explain the performance gap between the results in this paper and the SOTA models:

1. **Task Differences**: The nature of the tasks may differ, as Multimodal Sentiment Analysis (MSA) focuses more on emotional content, while other classification tasks in related literature may not involve sentiment-specific features. This discrepancy could lead to differences in model effectiveness, as MSA requires more nuanced interpretation of multimodal cues.

2. **Differences in Pretraining**: The pretraining tasks used for the FLAVA model, such as Masked Language Modeling (MLM) and Masked Multimodal Modeling (MMM), primarily involved large-scale text-image completion tasks. These tasks contain limited sentiment-related information, which may reduce the model's ability to learn emotional nuances compared to models pretrained on emotion-rich datasets.

3. **Limited Pretrained Model Options**: There were few available pretrained versions of the FLAVA model, and freezing the Multimodal Encoder led to a decrease in performance. This contrasts with theoretical expectations, where freezing should enhance model stability. The limited adoption of the FLAVA model compared to more established BERT variants, which have a variety of pretrained options, may have contributed to these performance issues.

4.2. Ablation Study

Table 1. Ablation Study Results

Model	Accuracy	F1 Score
FLAVA w/o image	0.63	0.62
FLAVA w/o text	0.70	0.70
FLAVA	**0.73**	**0.73**

The ablation study was conducted to assess the impact of each modality—text and image—on the model's performance. To isolate the effect of each modality, two sets of experiments were performed: in the text-only experiments, the image inputs were masked with zero tokens, and in the image-only experiments, the text inputs were similarly masked. This approach allowed for a direct comparison of the model's performance when relying solely on either text or image.

The results indicate that for this dataset, both text and image contribute valuable information for sentiment classification. The model's performance declined more significantly when the image modality was removed, suggesting that the image features play a larger role in the classification process. This may be due to visual cues that convey sentiment, such as facial expressions, colors, or contextual

elements in the images, which provide critical information that complements the text.

In contrast, the text-only model showed a smaller reduction in accuracy and F1 score, indicating that while text contributes to sentiment understanding, it may not capture the full range of emotional content as effectively as the combined modalities. Thus, integrating both text and image data is crucial for achieving optimal performance in multimodal sentiment analysis tasks.

4.3. Discussion of abolished designs

Before selecting the FLAVA model, both VisualBERT and CLIP were experimented for multimodal sentiment analysis. VisualBERT was not chosen because of limitations related to the encoder. ResNet-50 was initially used for image feature extraction, whereas the original VisualBERT paper employed R-CNN. This mismatch rendered the pretrained weights ineffective. Additionally, using a single encoder for both modalities failed to align the focus of image features with the text modality, a problem that early multimodal tasks aimed to resolve through joint attention mechanisms.

For CLIP, although its cross-product design facilitated the association of individual tokens and image features, its performance on this task was suboptimal. This may be due to its architecture not leveraging cross-modal attention as effectively as FLAVA or other BERT based models. Therefore, CLIP was ultimately discarded.

Various classifier designs were also considered, including early multimodal approaches that used softmax-weighted sums over all tokens as model output. It was also experimented with weighted sums for text and image embeddings, combining these with the base model's output (FLAVA's multimodal encoder) and replacing the MLP with multi-head attention layers.

However, these designs were consistently outperformed by a simpler approach that utilized a single Global token from the base model's output, with a linear classifier or a 2-3 layer MLP for passing supervisory signals. The simpler approach yielded a 7-8% improvement, likely because the limited supervisory signals in this dataset needed to be rapidly conveyed to the base model rather than overcomplicated by the classifier. The cross-modal attention in the base model more effectively utilized the scarce signal.

Finally, resampling was preferred over using weighted Cross Entropy (CE) loss due to statistical reasons. Although both methods yield the same expected loss, resampling increased the frequency of rare labels, while weighted CE loss did not. This made it easier for the model to learn rare label information, whereas the variance of rare labels under weighted CE loss was high, leading to convergence issues.

5. Conclusion

This paper presented a comprehensive study on multimodal sentiment analysis using the FLAVA model, which integrates both text and image data through a BERT-based architecture. By leveraging multimodal pretraining and a unified attention mechanism, FLAVA was shown to be effective in handling the complexities of combining visual and textual information for sentiment classification tasks. The experiments demonstrated that fine-tuning a pretrained FLAVA model, coupled with a lightweight MLP classifier, could achieve competitive results, reaching 73% accuracy.

The research highlighted several key findings. First, the use of layer freezing during fine-tuning helped accelerate convergence and mitigate catastrophic forgetting, while dropout and early stopping were crucial in preventing overfitting. Second, resampling strategies played a significant role in addressing class imbalance, thereby improving the model's generalization. The ablation study further underscored the importance of incorporating both modalities, as removing the image or text significantly degraded performance, with the image modality contributing more prominently to the classification outcome.

While the FLAVA model showed promising results, there remains room for improvement. The performance gap with state-of-the-art models suggests the need for more targeted pretraining tasks that include sentiment-rich datasets. Moreover, the limited availability of pretrained FLAVA checkpoints may constrain its utility compared to more widely adopted models such as BERT and CLIP.

Future work could explore advanced techniques for further integrating multimodal information, such as adaptive fusion mechanisms or dynamic attention across modalities. Additionally, leveraging larger and more diverse datasets for pretraining could enhance the model's capacity to understand nuanced cross-modal relationships. Overall, this study demonstrates the potential of using FLAVA for multimodal sentiment analysis, paving the way for more sophisticated approaches in the field.

References

[1] Hugging Face. Open source flava document on huggingface, 2024. Accessed: 2024-10-08. 4

[2] Liunian Harold Li, Xiujun Yin, Chunyuan Li, Pengchuan Hu, Lei Zhang, Lijuan Wang, Houdong Zhang, Jianfeng Gao, and Luke Zettlemoyer. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 2

[3] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1, 2

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 91–99, 2015. 2

[6] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629, 2022. 1, 2, 4

[7] Aohan Wang, Yu Han, and Stefan Riezler. Data selection for language models via importance resampling. *arXiv preprint arXiv:2302.03169*, 2023. 3