COSE474-2024F: Final Project Proposal
"SentimentalLens using Multimodal and BLIP"

Seungmin Cha 2022320120 [1]

1. Introduction

Sentiment analysis has become a cornerstone of various artificial intelligence (AI) applications, particularly in text-based domains. However, sentiment analysis in images remains an underexplored field, despite its significant potential to bridge the gap between visual perception and emotional interpretation. Motivated by a personal interest in sentiment analysis and the growing emphasis on image processing techniques in recent AI advancements, this project seeks to explore the intersection of these two domains. A survey of existing literature revealed that research in image sentiment analysis is not only sparse but also limited in terms of leveraging the latest advancements in multimodal approaches. This realization inspired the development of a novel methodology to address these shortcomings by combining state-of-the-art image processing models with robust sentiment analysis techniques.

Traditional sentiment analysis methodologies have predominantly focused on text data, with limited efforts to incorporate visual information. While early studies in multimodal sentiment analysis introduced the concept of combining image and text data, they lacked the sophistication provided by contemporary models like CLIP(Radford et al., 2021) and BLIP(Li et al., 2022). These advancements in image-language modeling present a unique opportunity to enhance sentiment analysis by enabling AI to interpret the nuanced emotions evoked by visual stimuli. The objective of this project is to construct a system that, given an image, generates an emotionally interpretable textual description, thereby offering a window into the affective dimensions of visual content. By doing so, the project aims to extend the scope of AI into the realms of art and human emotion.

To this end, we employed the BLIP model and fine-tuned it on the MVSA-multiple dataset, a collection of Twitter images paired with sentiment-rich textual descriptions. The fine-tuned model generates sentiment-aware textual interpretations of images, which are subsequently evaluated using a RoBERTa(Liu et al., 2019)-based sentiment analysis model trained on the same dataset. The proposed approach achieved a significant reduction in sentiment prediction loss, from 0.54 to 0.46, while improving binary sentiment classification accuracy from 63% to 85%. These results highlight the potential of leveraging state-of-the-art multimodal frameworks for image sentiment analysis, even when faced with dataset limitations. While the performance is constrained by the quality of the dataset, this work demonstrates the feasibility of integrating advanced image-language models with sentiment analysis to address the intricate task of emotional interpretation in visual content.

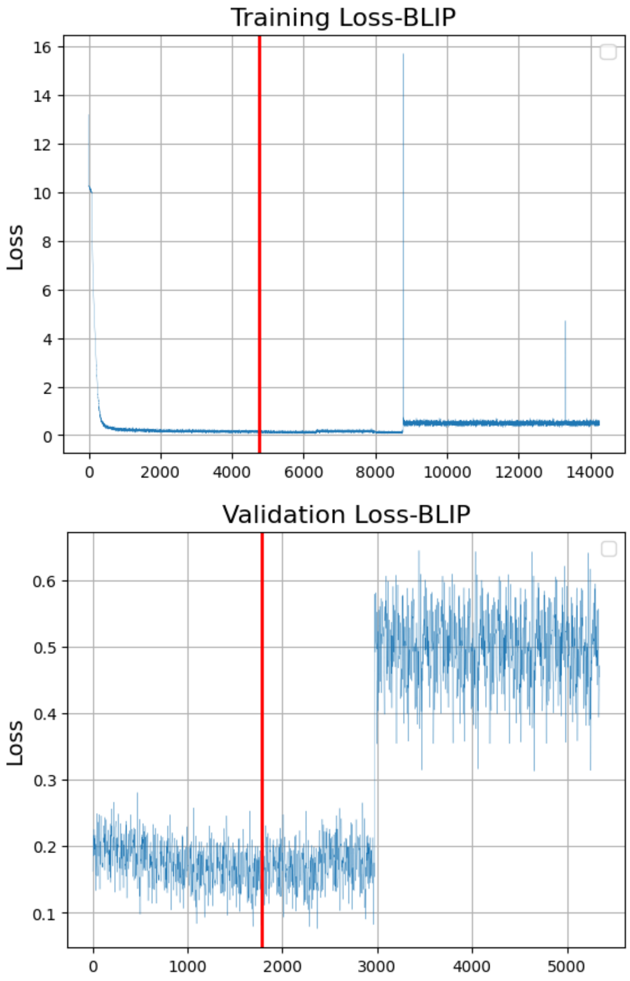
2. Methods & Experiments

Model Architecture. This study employs two primary models: the BLIP-image-captioning-large model from Hugging Face and the Twitter-RoBERTa-base model fine-tuned for sentiment analysis. The BLIP model was fine-tuned to generate sentiment-aware captions for images, utilizing a dataset of image-text pairs where diverse textual interpretations of the same image minimized potential bias. This approach leverages BLIP's capability to establish strong image-text relationships, while the sentiment analysis task is conducted using the pre-trained RoBERTa model. By decoding the captions generated by BLIP and feeding them into RoBERTa, the overall system produces a sentiment score aligned with human interpretations.
For instance, prior to fine-tuning, BLIP might describe an image of a baseball game as "a baseball game is being played in a stadium". After fine-tuning, it generates a more sentiment-aware caption, such as "bluejays vs tigers cometogether". This refinement improves sentiment prediction accuracy significantly.

Dataset and Preprocessing. The dataset used in this study is the MVSA-multiple dataset, a Twitter-based collection of image-text pairs. The dataset originally contains approximately 19,600 image-text pairs; however, after removing samples where the image and text were poorly matched, the final dataset used for training comprised 15,000 pairs. Each pair includes a sentiment label derived from fine-tuned RoBERTa outputs, ensuring consistency in sentiment evaluation.

Preprocessing included:
Text normalization: All mentions (e.g., @username) were replaced with @user to reduce noise in textual data.
Image resizing: Images were kept at their original resolution, as resizing could affect sentiment-related visual cues.

Training Process. Fine-tuning was conducted only on the BLIP model, while the RoBERTa model was used solely for sentiment evaluation. The BLIP model was trained on Colab Pro with an NVIDIA A100 GPU. The hyperparameters were set as follows:
Learning rate: $1 * 10^-5$
Batch size: 8
Optimizer: AdamW
The training process spanned three epochs, as the validation loss began increasing beyond this point. The loss function measured the absolute difference between sentiment scores (-1 for negative to 1 for positive), while the binary sentiment accuracy was evaluated based on whether both scores were of the same sign (positive or negative).



Red line: epoch 3

Evaluation Metrics. Model performance was evaluated using two primary metrics:
Loss: The absolute difference in sentiment scores between predicted and ground truth values.
Accuracy: The binary sentiment match between predictions and ground truth (positive or negative sentiment).
Fine-tuning led to a substantial improvement in performance:

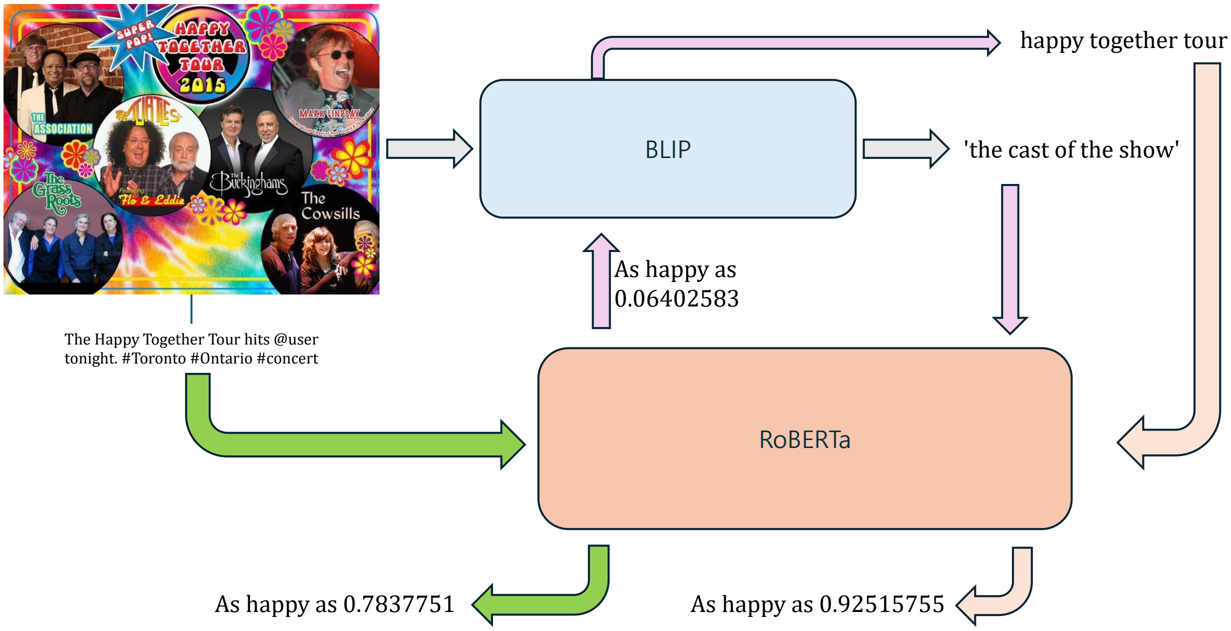
	Pre-fine-tuning	Post-fine-tuning
Loss	0.54	0.46
Accuracy	0.63	0.85

3. Future Direction

Limitations. While this study demonstrates the feasibility of leveraging multimodal approaches for sentiment analysis, several limitations emerged during the research process. First, the fine-tuning process was limited to the BLIP model, with no integration of the BLIP and RoBERTa components into a unified framework. Such an integration could allow for more diverse and targeted fine-tuning strategies, enabling a more seamless interaction between image and text modalities. Second, the MVSA-multiple dataset, although sufficient for this study, presents inherent limitations. As a Twitter-based dataset, the textual content often exhibits informal language and lacks the richness of human-curated sentiment labels. Additionally, the dataset size, while moderately large, could be further expanded to capture a broader range of image-text pairs. Computational constraints also posed challenges, particularly in batch size and model training iterations, due to limited GPU memory on Colab Pro. Addressing these limitations could lead to further improvements in both model performance and generalization.

Potential Improvements. To overcome these limitations, several directions for improvement can be considered. The first and most pressing is the acquisition or creation of a higher-quality dataset. A dataset with manually labeled, sentiment-rich annotations curated by multiple annotators could significantly enhance the reliability of the generated captions and sentiment analysis. Additionally, exploring other advanced sentiment analysis models beyond RoBERTa, or even fine-tuning such models directly on curated datasets, could yield better results. Moreover, extending the training process, for example, by lowering the learning rate and increasing the number of training epochs, might capture finer-grained details and lead to further optimization, despite the computational trade-offs.

Practical Applications. This research has the potential to impact various real-world applications. First, it could assist in artistic creation by providing AI-generated,



The Happy Together Tour hits @user tonight. #Toronto #Ontario #concert

sentiment-aware captions for visual artworks. Second, it could serve as a foundation for emotion-driven marketing in advertising and branding, where understanding and appealing to customer emotions is crucial. Lastly, this technology could be extended to the realm of mental health monitoring, where user-uploaded images on social media could be analyzed for emotional well-being, aiding early detection of emotional distress or imbalances. These applications highlight the versatility of the proposed system and its capacity to extend AI capabilities into human-centric domains.

Future Research Directions. The study also opens several avenues for future exploration. A key direction is to enable high-dimensional sentiment analysis, moving beyond binary classifications of positive and negative sentiment to encompass complex emotional states, such as nostalgia, excitement, or melancholy. Although this presents challenges in disentangling overlapping sentiments within an image, it offers an opportunity to better mimic human emotional interpretation. Another promising direction is the inclusion of additional modalities, such as audio and video, which could enable richer multimodal sentiment analysis frameworks. Lastly, the system could be extended for real-time sentiment analysis, enabling live monitoring of streamed visual content for immediate emotional insights. These extensions could significantly broaden the scope and applicability of multimodal sentiment analysis.

References

Ahuja, G., Alaei, A., and Pal, U. A new multimodal sentiment analysis for images containing textual information. *Multimedia Tools and Applications*, pp. 1–30, 2024.

Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 223–232. ACM, 2013.

Chen, T., Borth, D., Darrell, T., and Chang, S.-F. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks, 2014. arXiv preprint arXiv:1410.8586.

Li, H., Lu, Y., and Zhu, H. Multi-modal sentiment analysis based on image and text fusion based on cross-attention mechanism. *Electronics*, 13:2069, 2024.

Li, J., Li, D., Xiong, C., and Hoi, S. C. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Machajdik, J. and Hanbury, A. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 83–92. ACM, 2010.

Ortis, A., Farinella, G. M., and Battiato, S. An overview on image sentiment analysis: Methods, datasets and current challenges. In *Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019)*, pp. 296–306, 2019.

Radford, A., Kim, J. W., Hallacy, C., Goh, G., Agarwal, S., Clark, A., Wulff, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Reece, A. G. and Danforth, C. M. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6 (1):15, 2017.

Schmidt, S. and Stock, W. G. Collective indexing of emotions in images. a study in emotional information retrieval. *Journal of the American Society for Information Science and Technology*, 60(5):863–876, 2009.

Siersdorfer, S., Minack, E., Deng, F., and Hare, J. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 715–718. ACM, 2010.

Yang, H., Zhao, Y., and Qin, B. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3324–3335, 2022.