COSE474-2024F: Final Project Proposal
Evaluating the Performance of CLIP in Handling Negative Sentences for Visual-Language Matching

2022320077 Suhyun Kim

1. Introduction

CLIP is trained by 400 million data set from the internet. However, because of the dataset, it has been impacted by social context. Bias is an unavoidable problem with VLM. Although there are many cases, my interest is related with positive bias. With similar context, when LLM analyzes the emotion data, the model had predicted the positive better than the negative emotion because of the train dataset(Zhao et al., 2023). Also, the (Zhang et al., 2024) said LVLMs had bias related to social-cultural context. In psychology, there is a Pollyanna Principle, that people are tilted to positive semantic. My question starts from here. Does the VLM, especially, CLIP can handle the negative expressions? I'm more interested in structure of the negative sentence rather than semantic. In human beings it is subjective value. I want to understand the CLIP, and it's performance with positive and negative structure of sentence about image.

2. Problem definition & challenges

My focus is on the performance of CLIP with negative and positive expression. However, these are abstract, so I would like to concentrate with the structure of the text. For inference, I will put one image and two text as a set. With text, one is a well illustrated positive sentence(e.g., This picture is a dog.) the other one is a simple negative sentence(e.g., This picture is not a dog). My hypothesis is CLIP can map the image and positive text well but, not with the simple negative-structure text.
Well-designed VLM is used to evaluate and compare with human perception. With prompt, human can write with negative structure sentence. That is much natural way for people not with the limitation. Thus, VLM has to be process both positive sentence and negative sentence for the precise contrast. Also, all the semantic parts start from the structure of the expression. If CLIP can handle it well with structure, it can give some hint with the semantic part.

3. Related Works

CLIP(Radford et al., 2021) is an open source Vision-Language Model. We can walk through the paper and study about CLIP. (Fan et al., 2021) has improved about auto-generating negative text set. In addition, TAGS-DC model which is improved with negative text set it had better performance with image-text searching task. Emotion extracting in LLM has shown the positive bias according to cultural, social groups(Zhao et al., 2023). They said it was caused by the large dataset. In my point, if the dataset is much larger like CLIP, it could have less bias than other pre-trained model.

4. Datasets

I would like to use image-text dataset for the inference. However, I couldn't find the image, positive caption, negative caption. Thus, I will use MS-COCO(Lin et al., 2014) with manually adding negative caption.(e.g., dog image, "this is a dog.", "this is not a dog."). I will randomly pick a few of data from the set.

5. State-of-the-art methods and baselines

VLBiasBench(Zhang et al., 2024) has estimated the VLMs(CLIP was not include), with the bias in several categories. They focus on the emotion semantic of the text. By using VADER, it gives score between -1,1 emotion score. With this score, they can measure the frequency of the positive or negative emotion. However, it is lean to emotion so rather than using other tools, I'm impressed with CLIP structure, so I would like to use Cosine Similarity or dot product to calculate the similarity between text and image.

6. Schedule & Roles

10.28 - 11.1 construct method
11.2 - 11.9 dataset setting
11.10 - 11.16 CLIP coding and code works
11.17 - 11.30 inferencing
12.1 - 12.12 writing Final Report

References

Fan, Z., Wei, Z., Li, Z., Wang, S., and Fan, J. Negative sample is negative in its own way: Tailoring negative sentences for image-text retrieval. *arXiv preprint arXiv:2111.03349*, 2021. doi: 10.48550/arXiv.2111.03349. URL https://doi.org/10.48550/arXiv.2111.03349.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Radford, A., Kim, J. W., Hallacy, A., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. doi: 10.48550/arXiv.2103.00020. URL https://doi.org/10.48550/arXiv.2103.00020.

Zhang, J., Wang, S., Cao, X., Yuan, Z., Shan, S., Chen, X., and Gao, W. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language models. *arXiv preprint arXiv:2406.14194*, 2024. doi: 10.48550/arXiv.2406.14194. URL https://doi.org/10.48550/arXiv.2406.14194.

Zhao, X., Liu, J., and Smith, N. A. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023. doi: 10.48550/arXiv.2305.15005. URL https://doi.org/10.48550/arXiv.2305.15005.