COSE474-2024F: Final Project Proposal
"Improving CLIP Performance Using Data Augmentation and Fine-tuning Techniques"

Lee, Dongyoung

1. Introduction

Image-text matching technology plays a crucial role in various real-world applications. For instance, image search engines must return relevant images when a text query is input, and image captioning systems must generate appropriate textual descriptions for given images. In this process, accurately linking images and texts is vital. CLIP, developed by OpenAI, is a powerful model designed to address this issue by mapping images and texts into a shared embedding space, allowing matching based on similarity. While CLIP has demonstrated strong performance, some challenges remain. This project aims to explore methods to enhance CLIP's image-text matching performance, particularly focusing on improving its generalization in small datasets.

2. Problem definition & chanllenges

Problem Definition: This project focuses on improving the image-text matching performance using CLIP. Specifically, while CLIP performs exceptionally well on large datasets, its generalization tends to degrade on smaller datasets. Additionally, there are challenges related to the asymmetry between images and texts and capturing the fine-grained nuances in textual descriptions. The goal is to propose new architectures or enhancements in training strategies that address these issues.

Challenges:

- Asymmetry Problem: Images and texts are fundamentally different in how they represent information, which can lead to asymmetries when mapped to a shared embedding space.
- Generalization in Small Datasets: CLIP performs well on large-scale datasets, but its generalization capability can suffer when dealing with smaller datasets. Enhancing performance on small datasets is crucial.
- Lack of Granularity in Text Representation: Capturing the fine-grained nuances in textual descriptions and aligning them accurately with images remains a challenge for existing models.

3. Related Works

A method to map images and texts into a shared embedding space using contrastive learning was proposed, and outstanding performance across various tasks such as image retrieval and captioning was demonstrated (Radford et al., 2021). One of the main limitations of the CLIP model is its reduced performance on small datasets and its inability to fully capture the fine-grained nuances of textual descriptions. Recent studies have introduced data augmentation techniques and fine-tuning strategies to overcome these limitations.

4. Datasets

MS COCO: A widely used dataset for image-text matching, featuring around 330,000 images and five textual descriptions per image. It is suitable for a variety of image-text matching experiments.

Flickr30K: This dataset contains approximately 30,000 images with 150,000 corresponding textual descriptions. The diversity of the texts and the complex scene descriptions make it ideal for evaluating image-text matching models.

Custom Dataset: For testing CLIP's performance on smaller datasets, a custom dataset will be curated. This dataset will focus on a specific domain (e.g., medical, fashion) to evaluate the model's performance in domain-specific image-text matching tasks.

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

CLIP is a state-of-the-art model for image-text matching, and this project will compare its performance with the following baselines:

- Traditional Image Captioning and Retrieval Models: Comparisons will be made with existing systems that combine models like ResNet or ViT with image captioning techniques.
- Data Augmentation and Fine-tuning Methods: The study will apply data augmentation techniques and fine-tuning strategies to make CLIP more robust, especially

in small datasets.

- Comparison with Recent Studies: The project will evaluate the improved model against recent research in image-text matching to assess the degree of performance enhancement.

6. Schedule

Week 1-2: Data collection and preprocessing. Prepare MS COCO and Flickr30K datasets and create a custom dataset.

Week 3-4: Train the baseline CLIP model and measure its initial performance.

Week 5-6: Modify the model architecture and apply hyperparameter tuning for performance improvement. Data augmentation techniques will also be explored.

Week 7-8: Final experimental results analysis and report writing. Compare the enhanced model with existing methods and draw conclusions.

References

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. 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*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.