Self-Supervised CLIP Fine-Tuning with Medical Image-Text Data

Abstract

Contrastive Language-Image Pre-Training (CLIP) is a multimodal model which connects images and text through large-scale pre-training, and has been greatly acclaimed ever since its release. Fine-tuning the base pre-trained model for image classification has shown increasing interest, especially in the medical domain. Generally, for this purpose, CLIP has been fine-tuned with image-label pairs. However, labeled data may not always be available. In this paper, we fine-tuned one CLIP model instance with medical image-text data, and another instance with medical image-label data, to measure how much they affected the base performance. Our results showed that fine-tuning with image-text data provided a significant performance increase compared to the base model. In turn, as expected, fine-tuning with image-label data performed even better, but the performance increase provided by image-text fine-tuning is not negligible, and is worth considering when working with unlabeled data.

1 Describe

1.1 Introduction

The Contrastive Language-Image Pre-training (CLIP) model by OpenAI, connects images and natural language text by learning joint vector representations, or *embeddings*, for the two modalities. By pre-training on a dataset with 400M image-text pairs, CLIP has displayed impressive *zero-shot* performance in various downstream tasks (Radford et al., 2021). Zero-shot learning refers to a model's ability to predict class labels that were not observed during pre-training.

To further improve downstream performance for more specific tasks, it may be desirable to fine-tune CLIP on domain-specific data, which presumably was not abundant during pre-training. This interest has been evident in the medical image domain, where in recent years the amount of published literature has seen a great increase (Zhao et al., 2024).

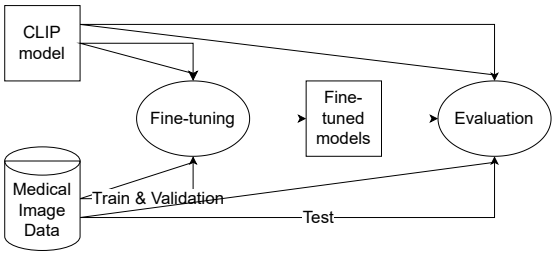

Figure 1: Fine-tuning pipeline.

Naturally, for image classification tasks, it is preferred to fine-tune using image-label pairs. Although, such labeled data may not always be available, and is expensive to produce. If one has access to descriptive text that corresponds to each image, one could still perform fine-tuning using the same self-supervised approach used during CLIP's pre-training process.

For that reason, the focus of this project is to fine-tune one CLIP instance on medical image-text pairs, another instance on medical image-label pairs, and compare their image classification performance to that of the base pre-trained CLIP model. The class labels that we chose for this task were different body parts.

1.2 Method

To make the contents of our method easier to follow, an overview of the pipeline is displayed in Figure 1. First, we downloaded the PubMedVision dataset (Chen et al., 2024) from HuggingFace[1]. The fields of interest were those containing the image, caption, and body-part label, respectively.

Apart from the general processing that was necessary to properly handle the dataset, a notable inconsistency was that the label field contained

[1] https://huggingface.co/datasets/FreedomIntelligence/PubMedVision

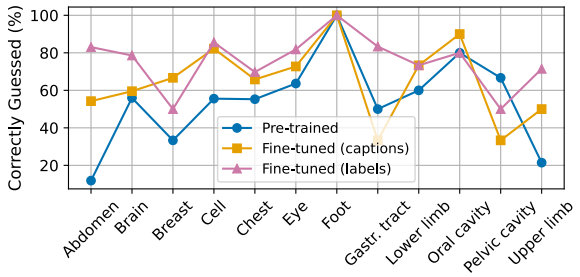

Figure 2: Correct predictions per class (%) for each model. (Gastr. tract = Gastrointestinal tract)

non-label information for some entries. This was resolved by simply excluding the invalid entries.

After filtering the dataset, it contained 533,076 entries, which would be too extensive to fully process and test given the time and resource constraints of our project. Therefore, after testing different sample sizes to estimate the training time, we settled for a sample of 4,096 entries from the dataset, which we distributed as 80-10-10 between the training, validation, and test set. For the training set, we sampled the same number of examples from each class, to avoid inducing bias during fine-tuning. For the validation and test set, samples were drawn from each class proportionally to the full dataset's label frequency distribution, to more accurately represent the expected evaluation for the full dataset.

As we proceeded to the fine-tuning steps, we used a method proposed by Goyal et al. (2023), which is to fine-tune with the same loss that was used during pre-training. CLIP uses *contrastive loss* for pre-training, and we do the same for fine-tuning to follow this principle.

The objective of contrastive loss is to maximize similarity between each image and its associated text, while minimizing similarity to all other images and texts. As previously mentioned, CLIP encodes images and text into a shared embedding space, enabling direct similarity calculations between the modalities. For a batch of size N, the model generates $N \times E$ image and text embeddings, where E is the embedding size. Then, the model calculates the cosine similarity between these embeddings to create a $N \times N$ similarity matrix. Image-to-text similarities can be observed row-wise in the matrix, and text-to-image similarities column-wise in the matrix. Cross-entropy loss is calculated for both the image-to-text and text-to-image similarities, and the total loss is simply the average of the two.

Finally, we evaluated the two fine-tuned models and the base pre-trained model against the test set. During fine-tuning of the image-label model, as well as evaluation, labels were prepended with the context-enhancing string "A medical image of ". The idea behind this is to aid CLIP in understanding the context of the data. Furthermore, the hyperparameters of the fine-tuning configuration were tweaked to maximize downstream performance. Notably, the learning rate was downscaled from the one used during CLIP's pretraining process, in proportion to how much smaller our batch size was.

1.3 Results & Analysis

When evaluating the pre-trained model and our fine-tuned models, top-1 and top-3 accuracy scores were recorded, which measure how frequently the label is within the top 1 and top 3 of the predictions, respectively. Table 1 presents these metrics for the three models. Our caption-fine-tuned model outperformed the pre-trained model, with improvements of approximately 15 points in top-1 accuracy and 4 points in top-3 accuracy. Unsurprisingly, the label-fine-tuned model showed even greater improvement, achieving increases of around 28 points in top-1 accuracy and 13 points in top-3 accuracy compared to the pre-trained model.

Model	**Top-1** (%)	**Top-3** (%)
Pre-trained	49.4	78.9
Fine-tuned (captions)	64.3	83.1
Fine-tuned (labels)	76.9	92.1

Table 1: Model top-1 and top-3 accuracy scores.

Figure 2 displays the prediction accuracy per body part of each model during evaluation. The general trend shows the same relative performance as for the top-1 and top-3 accuracy between the three model variants, indicating that the performance improvement is relatively consistent across different classes.

Although our experiments were specifically based on medical image data, our findings should be generalizable to any domain-specific dataset with a comparable format to PubMedVision. As expected, the label-fine-tuned model exhibited the best performance out of the three candidate models. However, our results also show that fine-tuning CLIP on image-text data can yield significant improvements when performing image classification compared to the base CLIP model. Therefore, in

the absence of labeled data, this may be a valid approach to improve downstream performance.

2 Examine

2.1 Embeddings

Early in the course, we learned about how embedding layers create low-dimensional representations that capture important features of the input data. In this project, I learned that this idea is not limited to text and that it is possible to encode both images and text into a shared vector space. CLIP uses different encoders for text and image data to transform them into the same dimensional space (Radford et al., 2021). This means that we can measure how close an image is to a given text.

2.2 Fine-tuning

Fine-tuning was another topic from the course. Instead of training a model from scratch, we used the pre-trained CLIP model and adapted it to medical images. This concept was vital for us given the project time and resource constraints.

2.3 Zero-shot learning

Zero-shot learning was described in the course as a model's ability to make predictions about classes it has never been explicitly trained on. CLIP is presented with impressive zero-shot abilities, for example, it successfully outperformed a fully supervised linear classifier by achieving a better score on 16 out of 27 different datasets using its zero-shot abilities (Radford et al., 2021). Initially, I assumed that a large, powerful model like CLIP would do well on medical images even without extra training, but its zero-shot performance was lower than expected. This made sense after I thought about it: CLIP had probably seen fewer medical pictures during its huge, general training process. Testing CLIP's zero-shot abilities showed me how much specialized data can help, but it also confirmed that large models can still recognize some features with little or no extra training. Our project was a good example of how zero-shot learning can be strong, but might need fine-tuning in highly specialized domains.

2.4 Labeled data and unlabeled data

During the project, we explored fine-tuning with two types of data: image-label pairs and image-text pairs. Initially, we assumed that labeled data would be the best option because it provides the model with precise categories. However, it was worthwhile to assess whether image-text pairs could achieve comparable performance, given that labeled data is not always available. Our results confirmed our initial assumption, as the model fine-tuned with labels performed the best. However, using just text data still improved performance over the base CLIP model. This experience reminded me of the difference between semi-supervised or self-supervised methods, where you can use unlabeled (or weakly labeled) data to improve a model when fully labeled data is not available.

2.5 Contrastive Loss

While reading the CLIP paper (Radford et al., 2021) during this project, I learned about contrastive loss, which differs significantly from the standard cross-entropy loss we often used in the course. Unlike cross-entropy loss, which computes the discrepancy between predicted probabilities and actual labels, contrastive loss functions by measuring similarities between embeddings. In our course, we learned that embedding similarities can be measured through cosine similarity, where the cosine of the angle between vectors indicates their closeness. Contrastive loss on the other hand, works by drawing matched pairs (such as an image and its corresponding text description) closer together in the embedding space, while simultaneously pushing non-matching pairs apart.

2.6 How much data to use?

An interesting question for us, given the time and resource constraints, was the amount of data to use in the project. The PubMedVision dataset (Chen et al., 2024) includes about 650,000 entries, but we chose to use a small sample of 4,096 entries due to these constraints. Previously, I would have felt it wasteful not to use the full dataset to achieve the best possible results. However, upon reflection, even without the time and resource limitations, our goal was not to fine-tune and achieve state-of-the-art results, but rather to answer the specific questions of the project. The 4,096 entries were sufficient to observe trends in zero-shot performance versus fine-tuning with image-text pairs or image-text labels. This experience taught me lessons that align with other teachings from the course: Having a lot of data is not always necessary. Sometimes a smaller, quality dataset is enough as we saw in the comparison between Lab 2 and Lab 3 and in lectures about improving data quality through meth-

ods like filtering and deduplication.

3 Articulate

3.1 What I Learned

I got hands-on experience with the flexibility of large-scale pre-trained models (like CLIP), I learned that even a smaller sample of domain-specific data can help adapt a model to a specific domain. I also realized that labeled data is not the only way to achieve meaningful performance gains. The use of captions or textual descriptions can still boost accuracy when labels are unavailable. This taught me how adaptable contrastive learning methods are, particularly in situations where it is costly or impossible to get high-quality labels.

3.2 How I Learned It

I learned by reading the related articles in the research phase of the project and by analyzing the results from the practical part of the project. Although I initially expected the label-based approach to be the most effective, I did not expect the caption-based approach to show such a noticeable improvement over CLIP's zero-shot baseline.

3.3 Relevance to the Course Objective

The main learning objective of this course is to enable students to seek, assess, and use scientific information within the area of NLP. This project required reading and understanding of the original CLIP paper (Radford et al., 2021) along with other works. By engaging with these sources, I had to evaluate different strategies for fine-tuning and see how they applied to our specific dataset. The extensive course content helped me understand NLP related papers without being confused with different NLP topics, especially practical topics like measuring embedding similarity, pretraining, data quality, and more. After the course and the project, I feel that I have developed the skill to understand scientific information in NLP and applying it effectively to solve real-world problems. I also learned to weigh the trade-offs of using a massive, general-purpose model versus building domain-specific approaches from scratch.

References

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *Preprint*, arXiv:2406.19280.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. 2022. Finetune like you pretrain: Improved finetuning of zero-shot vision models. *Preprint*, arXiv:2212.00638.

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. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, and Dinggang Shen. 2024. Clip in medical imaging: A comprehensive survey. *Preprint*, arXiv:2312.07353.