Adapting Pre-trained Vision Language Models for Low-Resource Domains

Sujay Sharma and **Anshuman Senapati** and **Muskan Pardasani**
Department of Computer Science
University of Wisconsin-Madison
Madison, WI 53706

Abstract

Vision-Language Models (VLMs) such as CLIP have revolutionized multimodal tasks by aligning textual and visual information in a shared representation space. However, their application to low-resource domains, such as medical imaging, remains challenging due to the limited availability of domain-specific annotated data and the inherent distribution mismatch between general-purpose and specialized data. In this work, we explore lightweight adaptation methods, including CLIP Adapter and context optimization (CoOp), to enhance CLIP's performance on the RSNA Pneumonia dataset for image classification tasks. Our experiments demonstrate that these parameter-efficient fine-tuning (PEFT) approaches improve both zero-shot and few-shot performance, providing a promising direction for adapting VLMs to low-resource domains.

1 Introduction

The ability to process and understand multimodal information—spanning both visual and textual data—is critical for applications in fields such as healthcare, remote sensing, and scientific analysis. Vision-Language Models (VLMs), such as Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021), leverage large-scale web data to align images and text into a shared embedding space. CLIP's zero-shot capabilities allow it to perform a variety of tasks without additional task-specific fine-tuning.

Despite their general applicability, these models exhibit significant limitations when applied to low-resource domains, such as medical imaging, where annotated datasets are scarce, and domain-specific vocabularies are poorly represented in pretraining corpora. For example, CLIP is trained on 400 million image-text pairs from the internet, which poorly reflect the nuanced and specialized semantics of medical datasets like Radiological Society of North America (RSNA) Pneumonia detection challenge dataset, containing only 30,000 training samples. Additionally, medical datasets pose unique challenges as the differences between classes are often more subtle and fine-grained, requiring models to discern nuanced patterns that are not well-captured in general-purpose pretraining. These limitations necessitate domain adaptation through fine-tuning or pretraining on specialized datasets, approaches that are computationally intensive and dependent on expert annotations.

Our goal is to adapt CLIP to the medical domain using parameter-efficient methods that minimize computational requirements while improving performance. Specifically, we evaluate two approaches. The first method is CLIP Adapter, a lightweight fine-tuning method using linear layers as adapters to bridge the domain gap between pretrained features and the novel dataset. Secondly, Context Optimization (CoOp) which is a prompt-tuning technique that leverages learnable task-specific context vectors to align the pretrained model to the unseen domain.

This report contributes:
1) A comprehensive benchmark and comparison of CLIP against specialized models pretrained entirely with medical data like MedCLIP, on the RSNA Pneumonia chest X-ray data for both zero-shot and few-shot tasks.
2) Implementation and evaluation of CLIP Adapter and CoOp for efficient domain adaptation of the pretrained CLIP model.
3) An analysis of the effectiveness of lightweight adaptation methods in low-resource domains.

2 Literature Survey

2.1 VLMs

We focus our review of VLMs to only those pretrained using two kinds of objectives. First, models trained using the contrastive objective like CLIP (Radford et al., 2021) have separate encoders for each modality and during training each batch of image-text pair is reconstructed into a batch of positive image-text pairs that belong together and negative image-text pairs. The contrastive loss then pushes the positive pair of embeddings closer together and negative farther apart in the embedding space. CLIP shows good zero-shot performance in different domains. However, it tends to struggle with specialized domains, such as medical or scientific data, due to the gap between general-purpose visual data and domain-specific features. We use CLIP as the VLM for our work. Second, models such as LLaVA (Li et al., 2024) use a pretrained CLIP's vision-transformer (ViT) encoder and learn to project the image embeddings onto the word embedding space which is then passed through an LLM. Keeping only the ViT frozen, the rest of the model is visual instruction tuned, fusing the vision and language information.

2.2 Finetuning VLMs

Adapter-based finetuning methods like CLIP Adapter (Gao et al., 2024) proposes a parameter-efficient method to adapt CLIP for new tasks by introducing a small set of trainable adapter layers to CLIP's image and text encoders. By freezing the base CLIP model and only training the adapters, they enable effective fine-tuning on low-resource datasets with minimal computational overhead. The method has been validated on general-purpose tasks, but there is limited research on its effectiveness in domain-specific fields, such as medical or scientific imaging. Tip-Adapter (Zhang et al., 2021) is a training-free approach designed to adapt CLIP for downstream tasks with minimal resource requirements. Rather than retraining the model, Tip-Adapter leverages a small, cached set of training data features, which it combines with the original CLIP features to optimize predictions during inference. This retrieval-based approach is highly resource-efficient, showing performance close to fine-tuned CLIP models in several tasks without extensive fine-tuning. By eliminating the need for task-specific training, Tip-Adapter is particularly suitable for domains with limited labeled data, like medical imaging or scientific datasets. Another popular method, LoRA (Hu et al., 2021) adds trainable low-rank matrices to the existing model parameters for efficient, lightweight fine-tuning.

On top of finetuning, these models could benefit from prompt tuning techniques such as CoOp (Context Optimization) (Zhou et al., 2022) for better few-shot classification performance on these medical image datasets. CoOp proposes an adaptable prompt-tuning method, wherein CLIP's prompt context is fine-tuned via learnable token embeddings. Instead of modifying the model's architecture, CoOp trains a series of soft prompts that better align with the target task. CoOp's prompt-learning approach has demonstrated strong performance in few-shot scenarios, achieving substantial gains by refining how CLIP interprets input images within specific contexts. By adjusting only the prompt embeddings, CoOp provides an efficient solution for adapting large VLMs to specific domains without extensive data.

2.3 VLMs for Medical Imaging

MedCLIP (Wang et al., 2022) fine-tunes CLIP for the medical domain by adapting its image-text representations to radiology. This research demonstrates that tailored training of VLMs on specialized datasets significantly improves performance in medical image-text retrieval and diagnostic accuracy. MedCLIP's success highlights the importance of adapting VLMs to achieve domain-specific understanding in low-resource settings.

3 Methodology

3.1 Overview

Our approach adapts CLIP for medical image classification tasks using the RSNA Pneumonia dataset. For the first set of experiments we evaluate and compare the zero-shot classification performance of CLIP and MedCLIP on the chest X-ray data. Next, we evaluate whether few-shot adaptation methods proposed in the literature help CLIP perform better on the same dataset.

3.2 Zero-Shot Classification

Zero-shot classification with CLIP leverages its dual-modality architecture, which consists of a vi-

sion encoder (e.g., ViT or ResNet) and a text encoder (e.g., Transformer). Given an input image I, CLIP maps it to a visual embedding $\mathbf{v} \in \mathbb{R}^D$ in a shared feature space. For a set of N class labels $\{l_1, l_2, \ldots, l_N\}$, CLIP generates text embeddings $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\}$, where each $\mathbf{t}_i \in \mathbb{R}^D$, by encoding prompts like "a photo of a [class label]".

The classification is performed by computing the cosine similarity between the visual embedding and each text embedding:

$$\text{sim}(\mathbf{v}, \mathbf{t}_i) = \frac{\mathbf{v} \cdot \mathbf{t}_i}{\|\mathbf{v}\| \|\mathbf{t}_i\|},$$

where $\mathbf{v} \cdot \mathbf{t}_i$ is the dot product of the vectors, and $\| \cdot \|$ denotes the Euclidean norm. The similarities are normalized to form a probability distribution using the softmax function:

$$p_i = \frac{\exp(\text{sim}(\mathbf{v}, \mathbf{t}_i))}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{v}, \mathbf{t}_j))}.$$

The model predicts the class label corresponding to the text embedding $\mathbf{t}_i$ with the highest probability p_i. This approach enables CLIP to perform inference on new, unseen tasks without task-specific fine-tuning, leveraging its pre-trained alignment of vision and language.

3.3 Few-Shot Adaptation

3.3.1 CLIP Adapter

The CLIP Adapter introduces two fully connected layers with dimensions matching the feature embeddings of the CLIP backbone. For example, if the CLIP model outputs feature vectors of size D , the first linear layer maps $\mathbb{R}^D \to \mathbb{R}^H$, where H is a hidden dimension (often chosen as D for simplicity). A ReLU activation is applied after this layer, followed by a second linear layer mapping $\mathbb{R}^H \to \mathbb{R}^D$, ensuring the output dimensions match the input features.

The Adapter combines pre-trained zero-shot embeddings f_{zero} with task-specific fine-tuned embeddings f_{task} using a residual connection:

$$f_{\text{adapted}} = f_{\text{zero}} + \alpha f_{\text{task}},$$

where α is a learnable scaling parameter initialized to a small value, typically 0.1, to ensure the pre-trained embeddings dominate early training. This lightweight setup adds minimal overhead while preserving CLIP's pre-trained knowledge.

3.3.2 Context Optimization (CoOp)

CoOp (Context Optimization) eliminates the need for manual prompt engineering in vision-language models by introducing learnable context vectors that are concatenated with class tokens. Instead of relying on fixed text prompts, CoOp uses a set of M learnable context embeddings $\{c_1, c_2, \ldots, c_M\}$, each of dimension D, where D matches the embedding size of the pre-trained language model. These vectors are combined with class tokens t_{class} to form input prompts

$$[c_1, c_2, \ldots, c_M, t_{\text{class}}],$$

which are then processed by the model.

The learnable context embeddings are optimized end-to-end using cross-entropy loss on the downstream task. This allows the model to dynamically capture task-specific or domain-specific semantics from data. Compared to manual prompt engineering, CoOp requires fewer labeled samples and adapts efficiently to new tasks with minimal human intervention, leveraging the underlying generalization capability of the pre-trained model.

4 Re-implementation

We re-implemented the results from Clip-Adapter for CLIP with a ResNet-50 vision backbone and a BERT encoder on the few-shot classification task for 1, 2, 4, 8 and 16 shots. The result is shown in Figure 1. The hyperparameters are as used in (Gao et al., 2024) and are listed in Table 1.

The results are evaluated on the Caltech101 dataset. Caltech101 is a popular benchmark dataset for object recognition tasks, consisting of 101 distinct object categories plus a background category. It contains images from classes such as animals, instruments, vehicles, and buildings, with each class containing between 40 to 800 images. This dataset is particularly valuable for few-shot learning experiments because it features considerable intra-class variance and realistic backgrounds, making it challenging to classify objects with limited labeled samples. The dataset's diversity and relatively low sample count per class make it an ideal choice for evaluating the efficacy of few-shot learning methods like Clip-Adapter, as it allows us to measure performance when only a few labeled examples are available for each class.

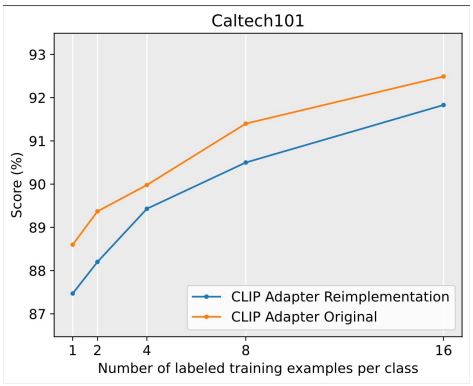


Figure 1: Re-implementation of CLIP-Adapter

Hyperparameter	Value
batch size	32
lr	0.002
epochs	200

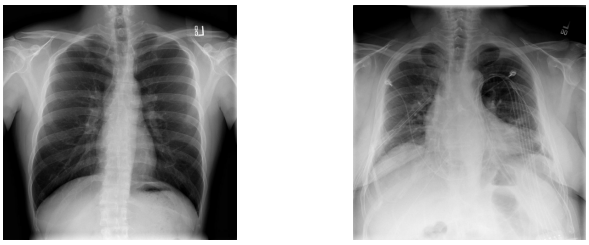
Table 1: Hyperparameters

In order to standardize our re-implementation, we used the Dassl PyTorch extension, a popular framework for research in Domain Adaptation and Semi-Supervised Learning. Dassl provides a flexible and modular structure for managing datasets, model training, and evaluation, making it well-suited for conducting reproducible experiments. This allowed us to efficiently set up and execute few-shot classification tasks within the Clip-Adapter framework, ensuring consistency across different shot counts and configurations. Our implementation also drew inspiration from how the experiments were conducted in the CoOp (Context Optimization) paper, which, along with Dassl, served as key points of reference for the authors of Clip-Adapter as well. By leveraging Dassl's capabilities, we were able to obtain results within 1% of the original implementation as shown in Figure 1.

5 Experimental Setup

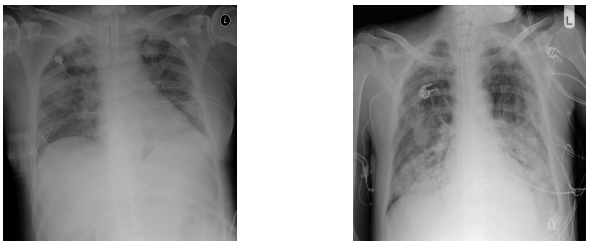
5.1 Dataset

The RSNA Pneumonia dataset contains approximately 30,000 training samples and 3,000 testing samples. It is used for binary classification tasks, distinguishing between pneumonia-positive and pneumonia-negative chest X-rays. We performed preprocessing on the dataset by converting all DICOM images to JPEG format. Additionally, we applied data preprocessing transformations like normalization, random flipping and cropping with the bicubic interpolation method in order to obtain input images of dimensions 224 x 224.



(a) Image 1 (b) Image 2

Figure 2: Negative samples (Person with healthy lungs)



(a) Image 1 (b) Image 2

Figure 3: Positive samples (Person affected by Pneumonia)

5.2 Models and Implementation

5.2.1 Zero-shot

For the zero-shot experiments we use CLIP and MedCLIP models to classify chest X-rays into *pneumonia-positive* or *pneumonia-negative* categories. The prompts are crafted in a fashion similar to those of the default handcrafted prompts proposed originally in (Radford et al., 2021), using the natural language templates, e.g., "This is an X-ray image showing {label}." Both the CLIP and MedCLIP models use the ResNet-50 backbone as the vision encoder.

5.2.2 Few-shot Adaptation

For the CLIP Adapter, we added 2 ReLU activated linear layers on the output of CLIP while blending the zero-shot embeddings with the adapted embeddings through a linea combination. For CoOp, we used the default 16 learnable context tokens and the class token goes at the *end* instead of being at the *mid* of the context. We did not experiment with a varying number of context vectors and the position of the class token, but it is a potential direction to explore further.

5.3 Preprocessing

Images were resized and normalized to fit model input requirements. Text prompts were tokenized using CLIP's tokenizer.

5.4 Hyperparameters

Tables 2 and 3 list the hyperparameters for our few-shot adaption experiments for CLIP on the RSNA Pneumonia dataset.

Hyperparameter	Value
Learning Rate	0.002
Batch Size	32
Num Epochs	1000
Optimizer	SGD
Backbone	ResNet50

Table 2: Hyperparameters for CLIP Adapter

Hyperparameter	Value
Learning Rate	0.002
Batch Size	32
Num Epochs	1000
Optimizer	SGD
Backbone	ResNet50

Table 3: Hyperparameters for CoOp

6 Results and Analysis

The experimental results highlight the effectiveness of lightweight adaptation methods in improving the performance of vision-language models (VLMs) in low-resource and domain-specific tasks. Below, we provide a detailed analysis of the findings.

6.1 Zero-Shot Results

The zero-shot results demonstrated CLIP's limitations in specialized domains. With a mean accuracy of 49.89%, CLIP performed only marginally better than random guessing in distinguishing between pneumonia-positive and pneumonia-negative cases. This underperformance is attributed to the lack of domain-specific pretraining and the general-purpose nature of the internet-scale dataset used for CLIP's training. In contrast, MedCLIP, fine-tuned on medical image-text pairs, achieved a significantly higher accuracy of 74.18%, underscoring the importance of domain-specific adaptation. This result highlights the inherent mismatch between general-purpose representations and the nuanced semantics required in medical imaging tasks.

6.2 Few-Shot Adaptation: CLIP Adapter

The few-shot experiments using CLIP Adapter revealed the potential of targeted fine-tuning in bridging the domain gap. Adapting only the vision branch resulted in slightly better performance (78.03%) than adapting the language branch (77.80%), suggesting that the visual features extracted from chest X-rays required more task-specific refinement. However, the combined vision+language adaptation approach failed to yield additive benefits, achieving a similar accuracy to the language-only adaptation. This indicates that the redundancy in multimodal fine-tuning might limit its effectiveness when both modalities are already closely aligned with the task.

6.3 Few-Shot Adaptation: CoOp

CoOp outperformed all other methods, achieving a mean accuracy of **81.47%**. This success can be attributed to its learnable context embeddings, which allow the model to dynamically adapt its input prompts to better align with the target task. By directly optimizing the prompts, CoOp effectively leverages the pretrained model's representations without altering its core parameters. This result demonstrates the strength of prompt tuning as a lightweight yet powerful adaptation strategy, particularly in few-shot scenarios.

6.4 Comparison of Adaptation Methods

Across all experiments, CoOp consistently outperformed CLIP Adapter and even MedCLIP, showcasing the efficiency of prompt tuning over parameter-intensive domain-specific training. The success of CoOp suggests that the alignment of textual inputs with domain-specific semantics is a critical factor in enhancing VLM performance. In contrast, the CLIP Adapter's performance, while strong, was more incremental, indicating that fine-tuning individual branches may have diminishing returns in certain domains.

6.5 Insights into Domain-Specific Challenges

The results also highlight the unique challenges of adapting VLMs to medical imaging. For instance, the subtle visual differences in chest X-rays require fine-grained visual representations, while the domain-specific terminology demands precise text encoding. MedCLIP's performance demonstrates

the value of pretraining on specialized datasets, but the success of lightweight methods like CoOp and CLIP Adapter shows that computationally efficient alternatives can achieve comparable results without extensive data requirements.

6.6 Generalization Potential

The ability of CoOp and CLIP Adapter to adapt effectively in low-resource settings opens avenues for their application in other specialized domains, such as satellite imagery or industrial inspection. These methods' success in few-shot settings also suggests their potential in scenarios where labeled data is scarce or expensive to obtain, making them highly scalable and practical for real-world applications.

7 Conclusion and Future Directions

This work demonstrates the representation power of multi-modal models like CLIP and the effectiveness of lightweight adaptation methods for enhancing internet pre-trained VLM performance in low-resource domains. CLIP Adapter and CoOp offer efficient alternatives to full fine-tuning, achieving substantial improvements in few-shot classification tasks for very specialized applications. Some future directions to our work include,

- *Investigate training-free adapters such as Tip-Adapter and LoRA:* Training-free adapters, like Tip-Adapter, use precomputed features, while LoRA updates low-rank parameters for efficient adaptation. Exploring their effectiveness across diverse tasks and datasets can reveal their scalability and utility in reducing computational costs, particularly for resource-constrained applications.

- *Extend experiments to multiclass and multilabel tasks:* Extending experiments to multiclass and multilabel tasks will test the ability of methods like CLIP Adapter and CoOp to generalize in complex scenarios, capture label dependencies, and address imbalanced datasets, offering insights into their adaptability and robustness in practical applications with higher label complexity.

- *Evaluate models on diverse datasets, including SIIM-ACR, to generalize findings:* Testing on datasets like SIIM-ACR, natural images, or satellite imagery will help validate the methods' generalization capabilities, uncover domain-specific limitations, and evaluate scalability across datasets of varying complexity, aiding the development of benchmarks for lightweight adaptation methods.

- *Explore using PEFT techniques on MedCLIP and see if there is a significant accuracy improvement or if the accuracy plateaus after a certain point:* Applying PEFT techniques like LoRA to MedCLIP can uncover whether significant accuracy improvements occur or plateau, identify optimal configurations, and address challenges like class imbalance and domain-specific complexity in medical datasets.

References

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.