Adapting Vision-Language Models to Low-Resource Domains

Anshuman, Muskan, Sujay

Motivation

- Adapting these pre-trained models to low-resource domains, such as medical imaging, remains challenging due to the scarcity of high-quality domain-specific data.
- The original dataset for pre-training the Contrastive Language-Image Pre-training (CLIP) model used 400 million image-text pairs from the internet. In comparison, datasets in the medical domains such as radiology contain merely 81,000 image-text pairs.
- Furthermore, distribution of domain-specific language and visual data is inherently different than the general internet data. This requires the user to either fine-tune the models to align the pretrained representations for downstream tasks on these domains or pre-train from scratch using a carefully curated low-resource dataset challenging due to the scarcity of high-quality domain-specific data.

Our goal is to benchmark and potentially propose an improvement to CLIP's performance in medical image-text retrieval, zero-shot and few-shot classification tasks, particularly for datasets which contain diverse medical imagery and rich textual annotations. These datasets provide an ideal testing ground for examining the impact of domain adaptation in VLMs, enabling us to analyze and improve CLIP's image-text retrieval and semantic understanding in the medical domain.

CLIP

A pioneering framework aligning images and text in a shared latent space.

Training Approach:

- Leveraged 400 million image-text pairs.
- Trained with contrastive learning to predict correct image-text pairings.

Zero-Shot Capabilities:

- Performs tasks without fine-tuning by leveraging natural language prompts.
- Example: "A photo of a cat" enables classification without task-specific training.

Limitations in Specialized Domains:

- Struggles with domain-specific vocabulary and semantics (e.g., medical terms, scientific imaging).
- Performance degrades when adapting to tasks outside the pretraining distribution.

Adaptation Techniques:

Prompt tuning and fine-tuning small modules like CLIP Adapter improve domain-specific performance.

Related Work - Medical VLMs

Adapting VLMs for Medical Domains:

- High demand for precise and reliable systems in medical imaging (e.g., X-rays, CT scans).

Challenges:

- Medical datasets are limited in scale and require expert annotation.
- High variance in imaging modalities and pathologies.

Existing Approaches:

- ***MedCLIP***: Extends CLIP’s architecture with medical-specific pretraining.
- ***BioViL:*** Fine-tunes VLMs for biological imaging and text analysis.

Success Stories:

- Improved accuracy in disease classification and report generation tasks.
- Enhanced retrieval of similar cases for diagnostic support.

Opportunities:

- Explore lightweight adaptation methods like CoOp and CLIP Adapter to complement domain-specific pretraining.

Need for better prompts

Contrastive pre-training objective

whether an image and a text belong together

used for zero-shot image classification: prompt = context + class token

Natural Language Supervision and Polysemy

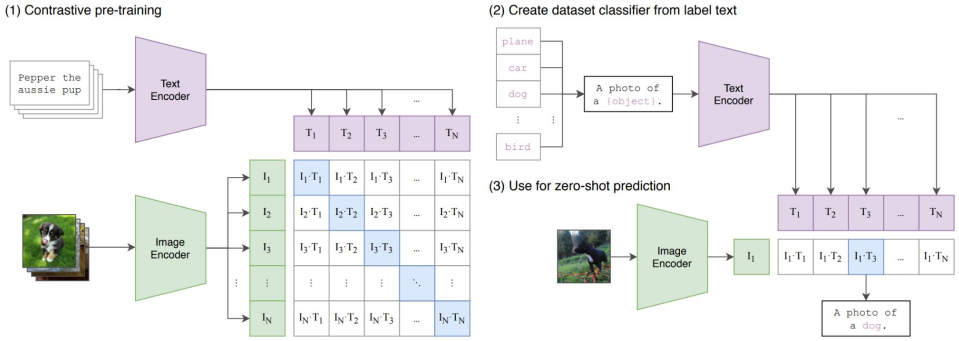
- When the name of a class is the only information provided to CLIP's text encoder it is unable to differentiate which word sense is meant due to the lack of context. In some cases multiple meanings of the same word might be included as different classes in the same dataset! This happens in ImageNet which contains both construction cranes and cranes that fly!

- Pre-training data does not map an image to a single word but entire sentences use prompts like "A photo of a {label}".

Context is important!

Customizing the prompt to each task/**dataset** by adding some context helps!

- For example, using "A photo of a {label}, a type of pet." for the Oxford-IIIT Pets dataset



(1) Contrastive pre-training

(2) Create dataset classifier from label text

(3) Use for zero-shot prediction

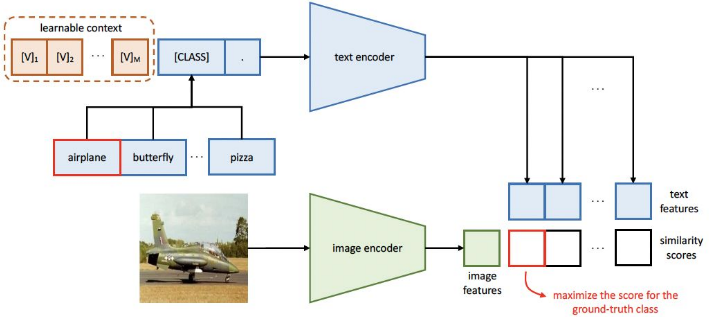
Adapting VLMs - Context Optimisation (CoOp)

Prompt Engineering vs. Soft prompt-tuning

- Manual prompt engineering requires domain knowledge and is sub-optimal
- Context words are now a set of learnable context vectors. Prompt = learnable context vectors + class token
- Pre-trained parameters are frozen and simple cross-entropy loss to align the “soft” context vectors with the target task in a few-shot setting.
- By adjusting only the prompt embeddings, CoOp provides an efficient solution for adapting large VLMs to specific domains without extensive data.

Advantages in Low-Resource Domains:

- If the pre-training contains enough knowledge, this can help extract maximum relevant information by learning the right context for the downstream domain.
- Eliminate task-specific pre-training, which is difficult because of data scarcity.



Need for few-shot adaptation in challenging domains

Zero-shot CLIP is quite weak on several specialized, complex, or abstract tasks such as

- satellite image classification (EuroSAT and RESISC45),
- lymph node tumor detection (PatchCamelyon)
- counting objects in synthetic scenes (CLEVRCounts),
- self-driving related tasks such as German traffic sign recognition (GTSRB),
- recognizing distance to the nearest car (KITTI Distance),

while **non-expert humans** can robustly perform several of these tasks, such as counting, satellite image classification, and traffic sign recognition, suggesting need for few-shot transfer of CLIP.

Especially for difficult tasks that a learner has no prior experience with, such as lymph node tumor classification few-shot transfer, is a meaningful evaluation for almost all humans (and possibly CLIP).

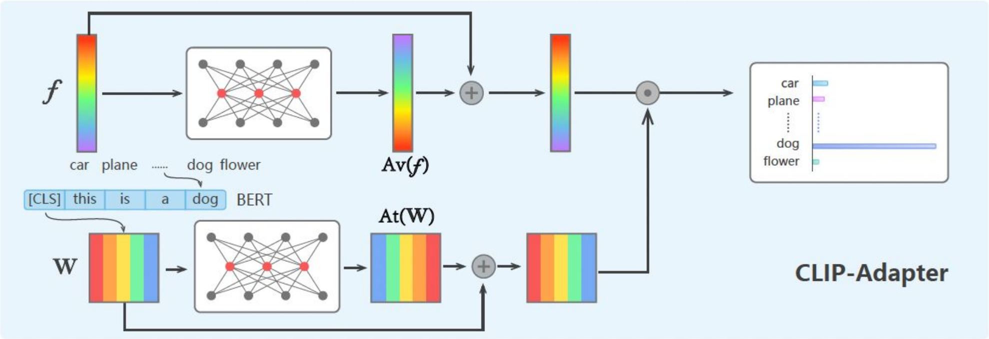
Adapting VLMs - CLIP Adapter

What is CLIP-Adapter?

- Parameter efficient fine-tuning (PEFT) - instead of full fine-tuning, introduce lightweight adapters with frozen backbones for few-shot learning
- Unlike previous works inserting adapters to all layers of the language backbone, only two linear layers are inserted after the last layer of the vision/language backbone
- Mixes the original zero-shot visual/language embedding with the corresponding fine-tuning feature using a residual connection

For Low-Resource Domains

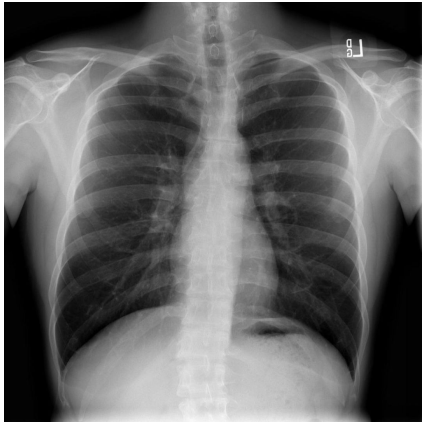
- Adapts pretrained features to the new domain
- Prevents the overfitting problem of few-shot learning because of less parameters

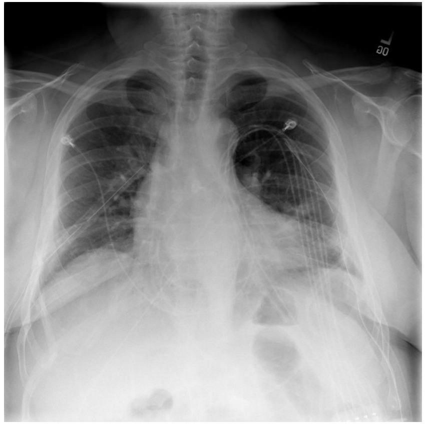


Dataset

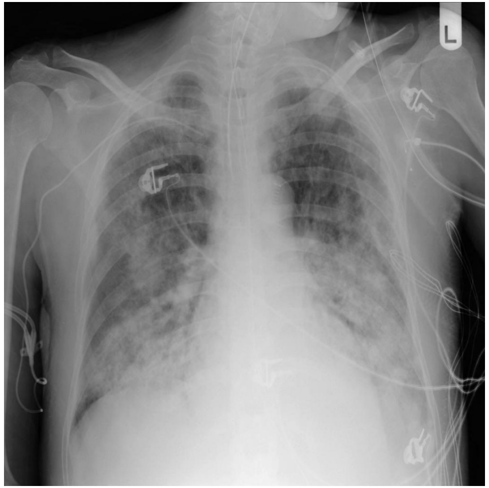
Dataset - RSNA Pneumonia (RSNA)

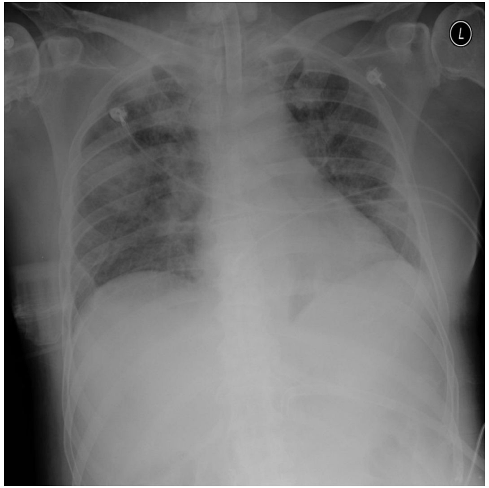
- Collection of Pneumonia cases found in a chest X-Ray database of the NIH
- Contains ~30000 training samples and 3000 test samples
- Binary classification task





- Scarcity of Data
- More subtle and fine-grained differences between classes





Experimental Setup

- Perform zero-shot classification on the RSNA dataset using the CLIP model
- Perform zero-shot classification on the RSNA dataset using the MedCLIP model
- Both models use the ResNet-50 backend
- MedCLIP pretrained on CheXpert and MIMIC medical datasets
- Perform few-shot classification on the RSNA dataset using CLIP Adapter and CoOp
- CLIP Adapter has an adapter of 2 FC layers with ReLU activation
- CoOp trained with 16 context tokens followed by class token at the end

Results

Method	**Mean Accuracy (3 seeds)**
CLIP (zero-shot)	49.89%
MedCLIP (zero-shot)	74.18%
CLIP-Adapter (Language Branch)	77.80%
CLIP-Adapter (Vision Branch)	78.03%
CLIP-Adapter (Vision+Language Branch)	77.80%
CoOp	81.47%

Scope of Improvement and Future Work

- Tip Adapter - Training free CLIP Adapter
- LoRA
- Multiclass/Multilabel classification tasks, Image-text retrieval tasks
- Apply PEFT techniques to MedCLIP
- Different datasets (SIIM-ACR)

That's all Folks!