

Few-shot Adaptation of Medical Vision-Language Models

Fereshteh
Shakeri

Yunshi
Huang

Julio
Silva-Rodríguez

Houda
Bahig

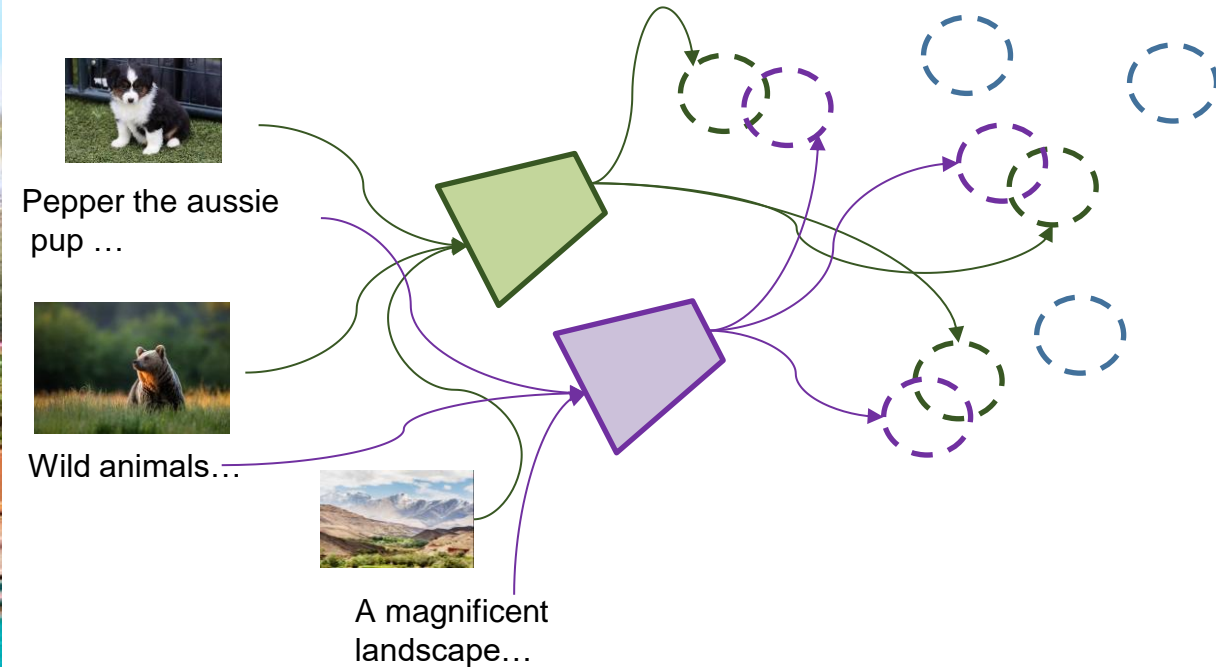
An Tang

Jose Dolz

Ismail
Ben Ayed

Generalist Vision-Language Models (VLMs)

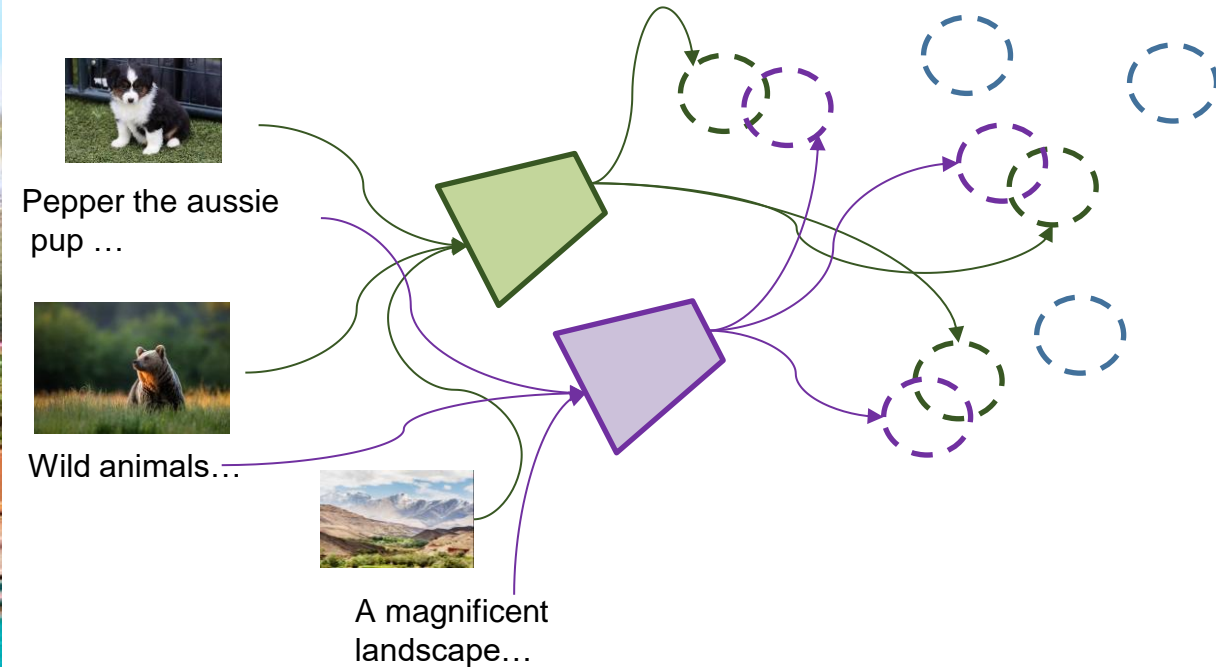
Unsupervised image-language pre-training



CLIP: Radford et al., Learning transferable visual models from natural language supervision, ICML 2021

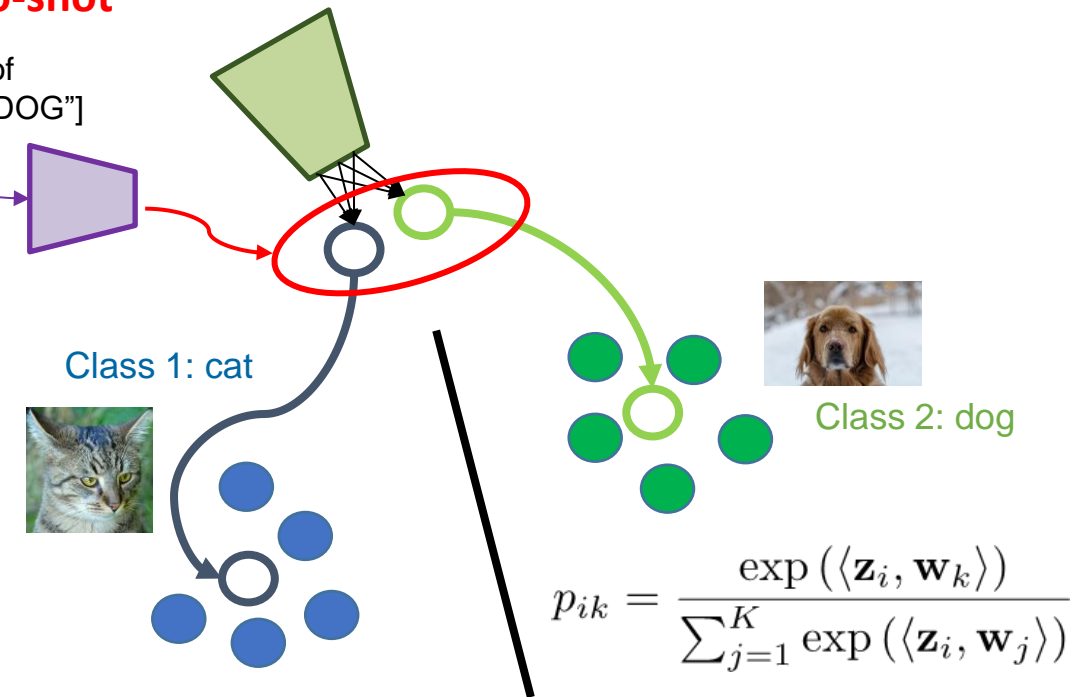
Generalist Vision-Language Models (VLMs)

Unsupervised image-language pre-training



Zero-shot

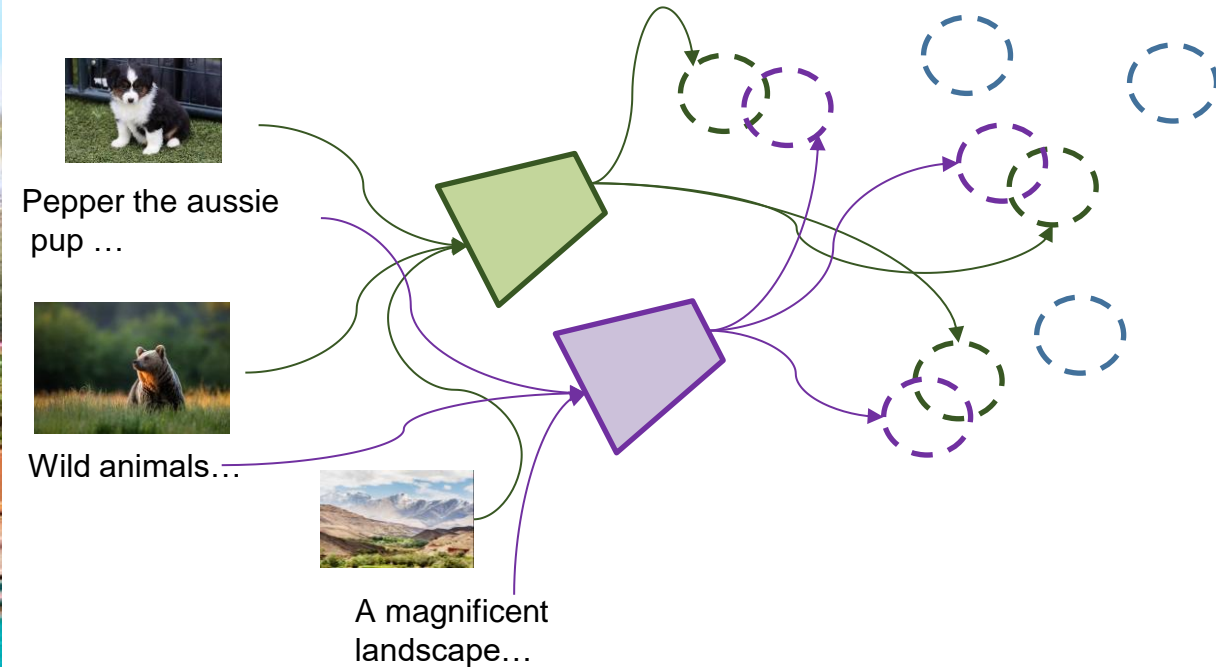
A photo of
[“CAT”/”DOG”]



CLIP: Radford et al., Learning transferable visual models from natural language supervision, ICML 2021

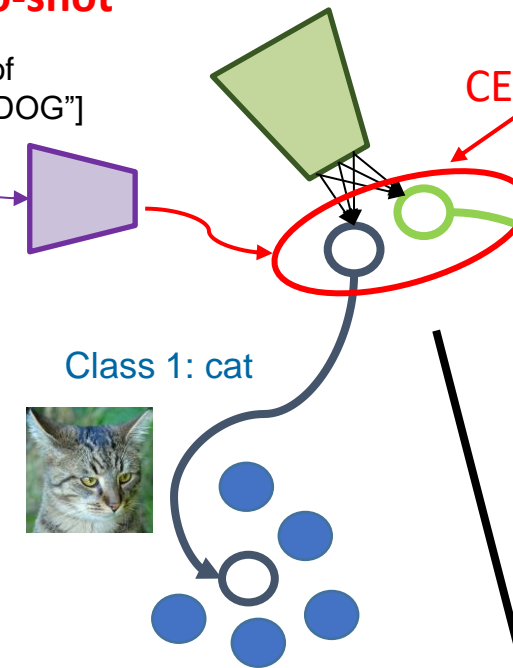
Generalist Vision-Language Models (VLMs)

Unsupervised image-language pre-training

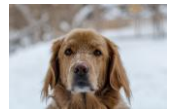


Zero-shot

A photo of
[“CAT”/“DOG”]



Few-shot Adaptation (linear probe)

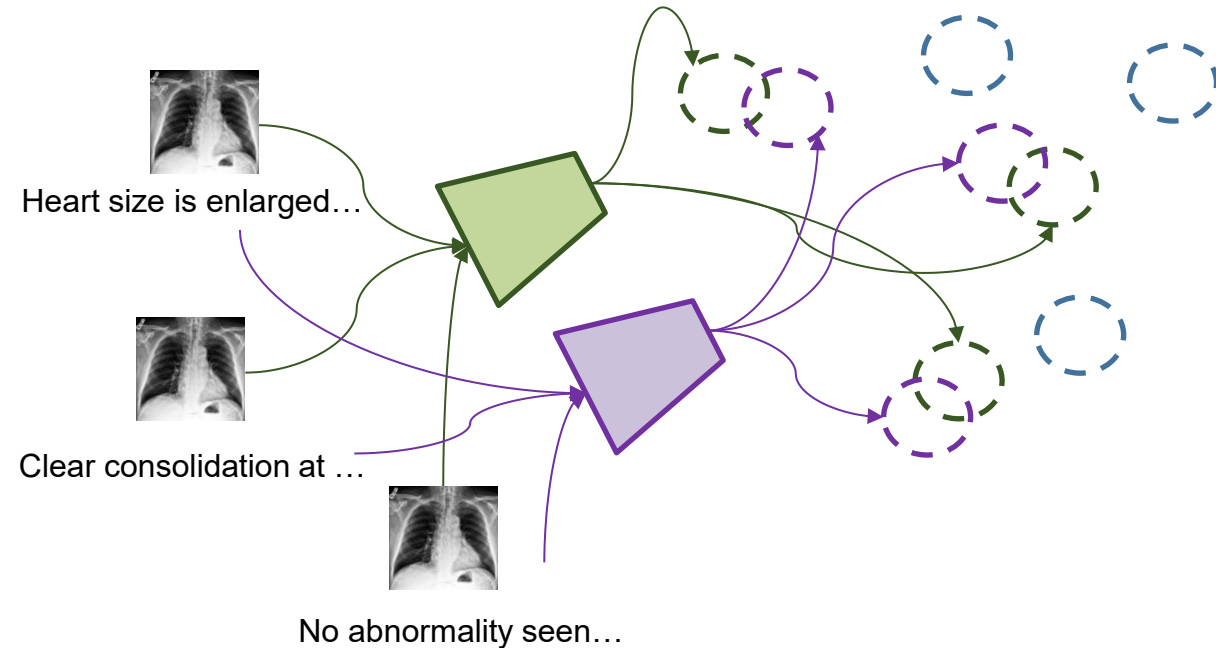


$$p_{ik} = \frac{\exp(\langle \mathbf{z}_i, \mathbf{w}_k \rangle)}{\sum_{j=1}^K \exp(\langle \mathbf{z}_i, \mathbf{w}_j \rangle)}$$

CLIP: Radford et al., Learning transferable visual models from natural language supervision, ICML 2021

Generalist Medical (Specialized) VLMs

Unsupervised image-language pre-training

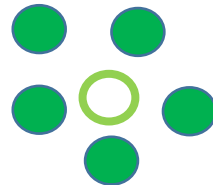
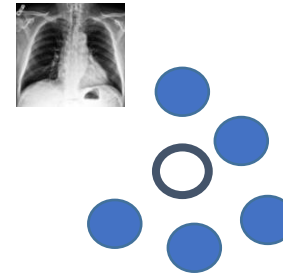


CONVIRT: Zhang et al., Medical Visual Representations from Paired Images and Text, MLHC 2022
 MedCLIP: Wang et al., Contrastive Learning from Unpaired medical images and text, EMNLP 2022
 Quilt-1M: Ikezogwo et al., One Million Image-Text Pairs for Histopathology, NeurIPS 2023
 FLAIR: Silva-Rodríguez et al., A Foundation Language-Image Model of the Retina, MedIA 2024

...

Zero-shot / Few-shot Adaptation

Class 1: pneumonia

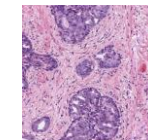


Class 2: normal

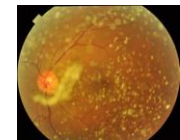
MedCLIP

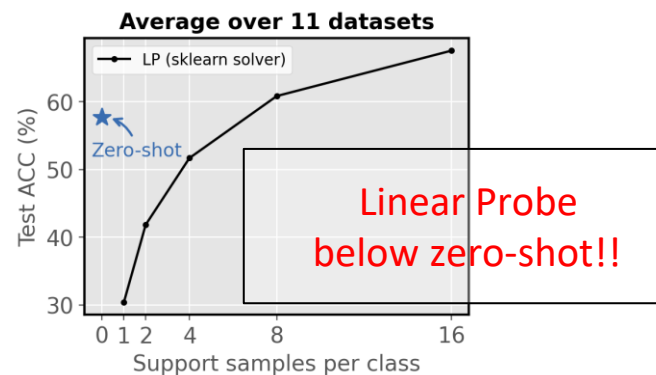


Quilt-1M

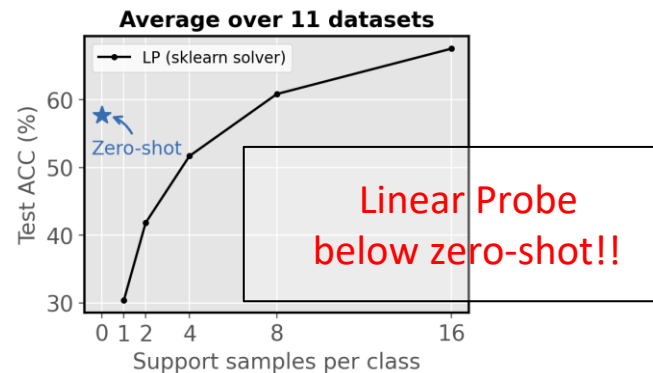


FLAIR





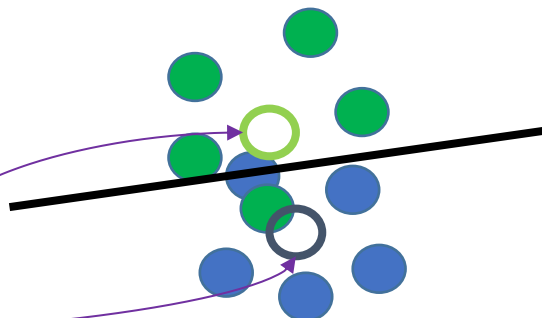
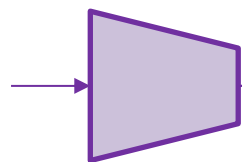
(Popular) Prompt Learning



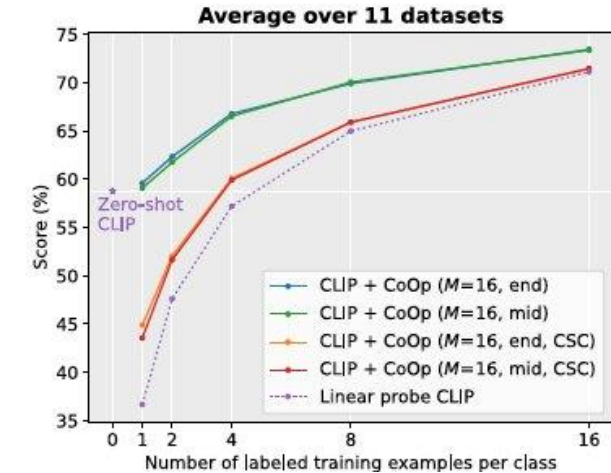
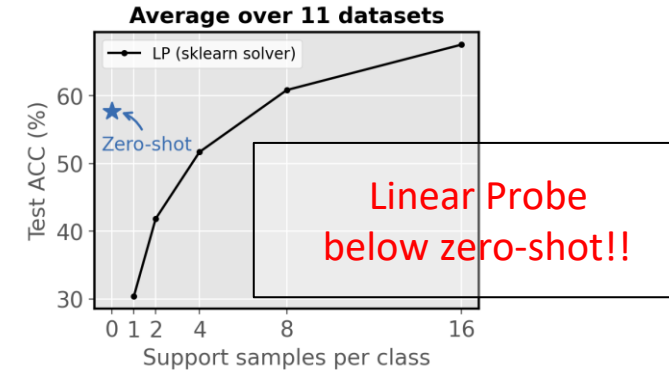
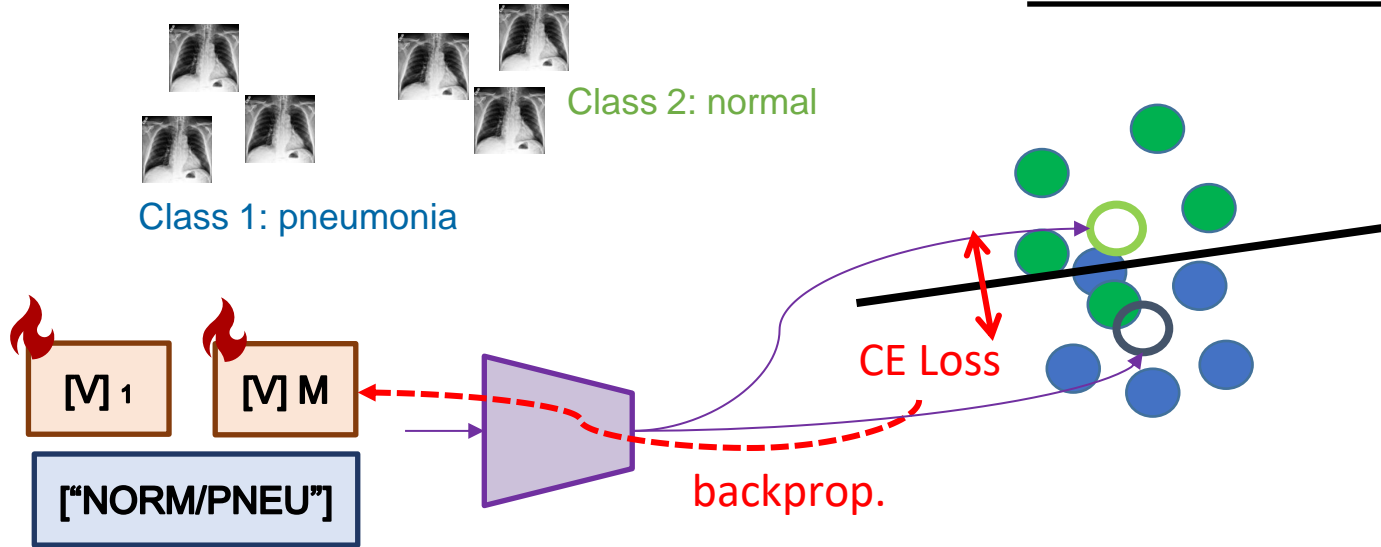
Class 1: pneumonia

Class 2: normal

["NORM/PNEU"]



(Popular) Prompt Learning



CoOp: Zhou et al., Learning to Prompt for Vision-Language Models, IJCV 2022[~1900 citations]

CoCoOp: Zhou et al., Conditional Prompt Learning for Vision-Language Models, CVPR 2022[~1200 citations]

KgCoOp: Yao et al., Prompt Tuning with Knowledge-guided Context Optimization, CVPR 2023[~122 citations]

Towards better **black-box** Adapters: LP+text

“Weak Baseline” Linear Probe

Features

Learned weight

$$p_{ik} = \frac{\exp(\langle \mathbf{z}_i, \mathbf{w}_k \rangle)}{\sum_{j=1}^K \exp(\langle \mathbf{z}_i, \mathbf{w}_j \rangle)}$$

(only vision)
Neglects any text knowledge

LP++: Huang et al., A Surprisingly Strong Linear Probe for Few-Shot CLIP, CVPR 2024

Towards better **black-box** Adapters: LP+text

“Weak Baseline” Linear Probe

Features

Learned weight

$$p_{ik} = \frac{\exp(\langle \mathbf{z}_i, \mathbf{w}_k \rangle)}{\sum_{j=1}^K \exp(\langle \mathbf{z}_i, \mathbf{w}_j \rangle)}$$

(only vision)
Neglects any text knowledge

Text- Informed Linear Probe

Text (zero-shot) prototype

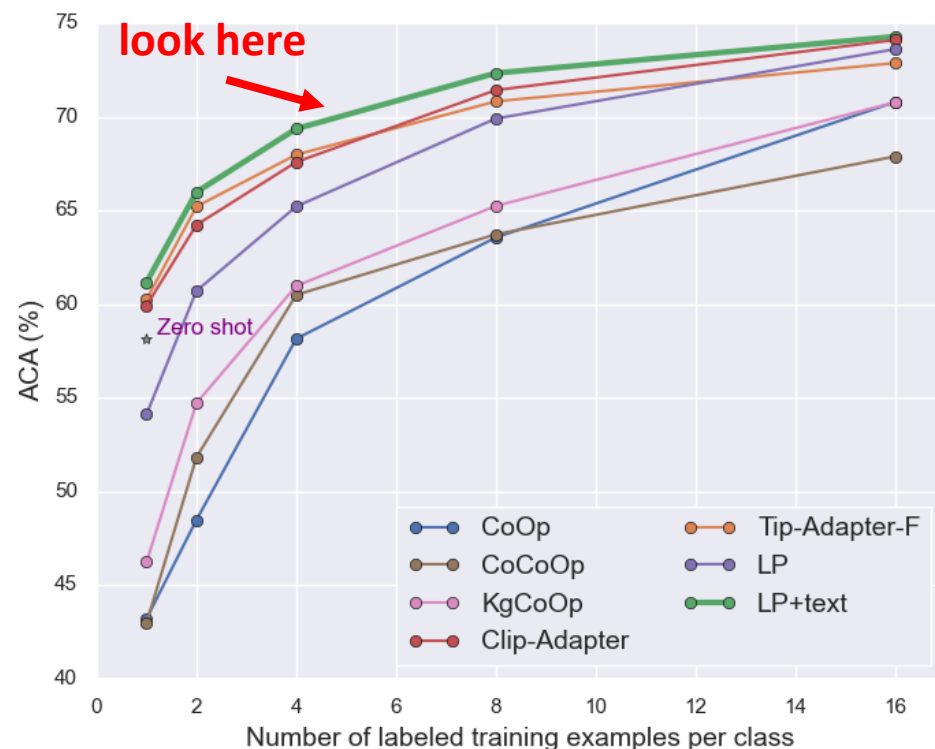
$$p_{ik} = \frac{\exp(\langle \mathbf{z}_i, \mathbf{w}_k + \alpha_k \mathbf{t}_k \rangle)}{\sum_{j=1}^K \exp(\langle \mathbf{z}_i, \mathbf{w}_j + \alpha_j \mathbf{t}_j \rangle)}$$

Trainable
image-text blending
weight

LP++: Huang et al., A Surprisingly Strong Linear Probe for Few-Shot CLIP, CVPR 2024

Results

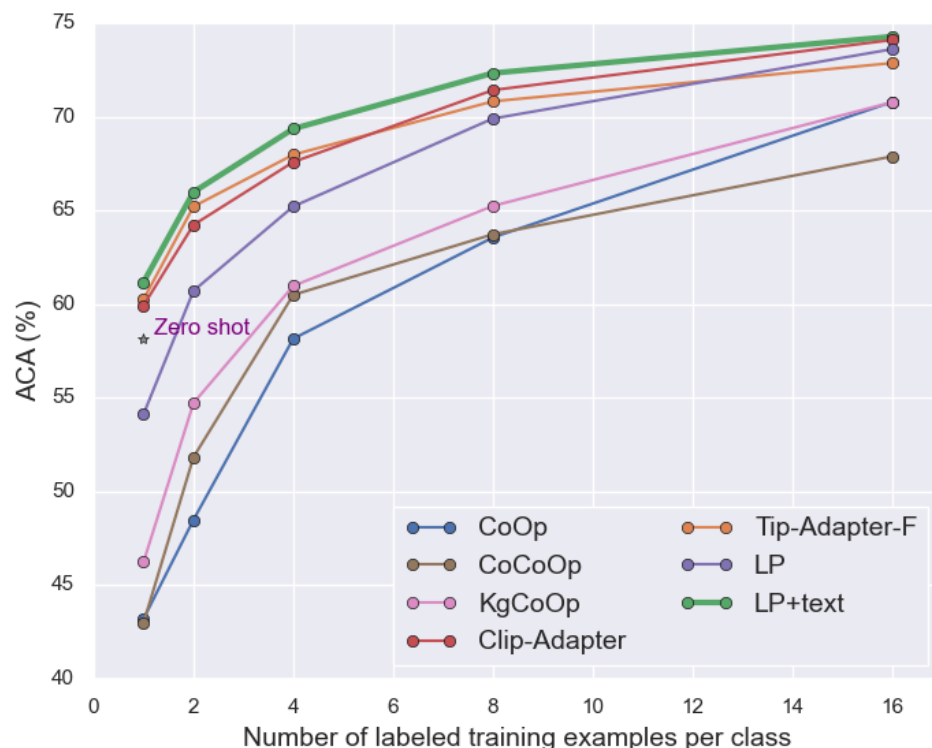
3 modalities / 9 datasets



LP+text is competitive

Results

3 modalities / 9 datasets



LP+text is competitive

LP+text is extremely efficient!


→ Adaptation in a **matter of seconds**

→ Trainable on **commodity GPUs (MBs)**


→ **Black-box adaptation**

Methods	Category	Training Time	Black-box	#Parameters
Zero-shot		n/a	✓	n/a
CoOp [4]	Prompt-Learning	3min	✗	$K \times n_{ctx1} \times D$
CoCoOp [5]		12min	✗	$n_{ctx2} \times D + C$
KgCoOp [6]		3min	✗	$K \times n_{ctx1} \times D$
Clip-Adapter [7]	CLIP-based Adapters	2min	✓	$2(D_1 \times D_2)$
Tip-adapter-F [8]		2min	✓	$K \times S \times D$
LP	Linear probe	43s	✓	$K \times D$
LP+text [9]		4s	✓	$K(D + 1)$





ETS
Le geste pour l'expertise




CRCHUM


Few-shot Adaptation of Medical Vision-Language Models

Fereshteh Shakeri^{1,2} Yunshi Huang^{1,2} Julio Silva-Rodriguez¹
Houda Bahig² An Tang² Jose Diaz^{1,2} Ismail Ben Ayed^{1,2}

¹ETS Montreal ²Centre de Recherche du Centre Hospitalier de l'Université de Montréal



MICCAI 2024



MICCAI

Motivation

Medical Vision-Language Models (VLMs)

VLMs like CLIP have been successful in natural image recognition, integrating image and text data to learn rich transferable representations.

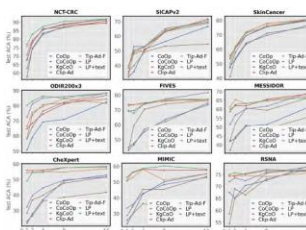
Nonetheless, medical VLMs adaptation remains challenging:

- Lackling low-prevalence diseases makes standard data-demanding strategies impractical in clinical contexts.
- High computational cost of 111 model fine-tuning on large-scale induction models.
- Privacy concerns of sharing foundation models pre-trained with proprietary data.

Solution: I) few-shot learning and II) black-box adaptation.

Few-shot adaptation results

We evaluate the models in a range of low-shot scenarios, with $S = 1, 2, 4, 8, 16$ shots per class. To simulate low-data regimes in medical clinical settings:



Contributions

- We introduce the first structured benchmark for the few-shot adaptation of medical VLMs.
- We provide adaptation experiments on 3 medical modalities: radiology, pathology, and ophthalmology, with 3 specialized foundation models and 7 tasks.
- We benchmark Prompt Learning and Adapter-based strategies.
- We propose a generalized linear probe (LP)-text that blends visual prototypes and text embeddings with learnable multipliers.

Method

Linear probe (LP): Fine-tunes only the linear classifier weights while keeping the other model's parameters frozen.

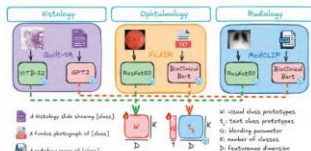
$$L_{LP}(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^N \log(\text{softmax}(\mathbf{w} \cdot \mathbf{f}_i)) \quad \mathbf{w} = \frac{\exp(\mathbf{f}_i \mathbf{w})}{\sum_{j=1}^K \exp(\mathbf{f}_j \mathbf{w})} \quad (1)$$

Text-driven linear probe (LP-text): Blends the visual prototypes with the text embeddings using learnable class-wise multipliers.

$$L_{LP\text{-text}}(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^N \log(\text{softmax}(\mathbf{w} \cdot \mathbf{p}_i + \mathbf{t}_i)) \quad \mathbf{p}_i = \frac{\exp(\mathbf{f}_i \mathbf{w}_p)}{\sum_{j=1}^K \exp(\mathbf{f}_j \mathbf{w}_p)} \quad \mathbf{t}_i = \frac{\exp(\mathbf{t}_i \mathbf{w}_t)}{\sum_{j=1}^K \exp(\mathbf{t}_j \mathbf{w}_t)} \quad (2)$$


Deployment on open-access foundation models

Pre-trained Medical Vision-Language Models: QatR-2M (radiology), FLAIR (ophthalmology), and MedCLIP (pathology).



References

1. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
2. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
3. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
4. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
5. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
6. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
7. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
8. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
9. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.
10. J. B. Rodriguez, "The first structured benchmark for few-shot medical VLMs," 2024.



Code is available Here! :)

Conclusions

- We introduced the first structured benchmark for few-shot adaptation of medical VLMs across different modalities.
- The text-driven linear probe, LP-text, offers a computationally efficient and black-box friendly solution, providing competitive performance compared to more complex methods like prompt learning and adapter-based strategies.
- The LP-text method reduces hardware requirements, making it practical for low-resource settings such as smaller clinical institutions with limited computational power. This makes it a feasible approach in real-world healthcare environments.

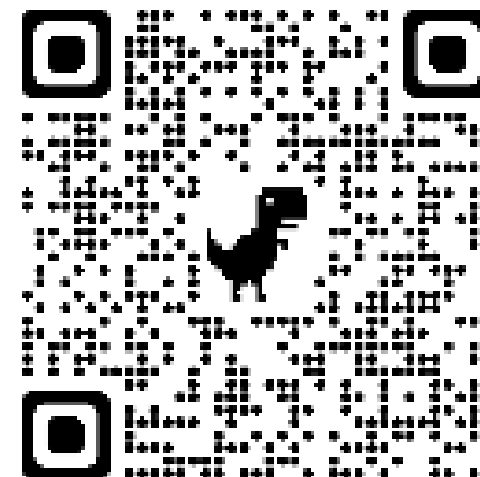
Feel free to use it!

TUESDAY-PM
POSTER 096

Any questions?



Open Benchmark!



Code is available Here! :)