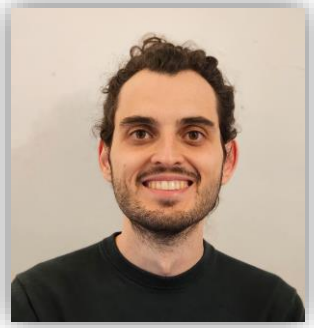
Le génie pour l'industrie

SEATTLE, WA CVPR JUNE 17-21, 2024

A Closer Look at the Few-Shot Adaptation of Large Vision-Language Models

Julio Silva-Rodríguez



Sina Hajimiri

Ismail Ben Ayed

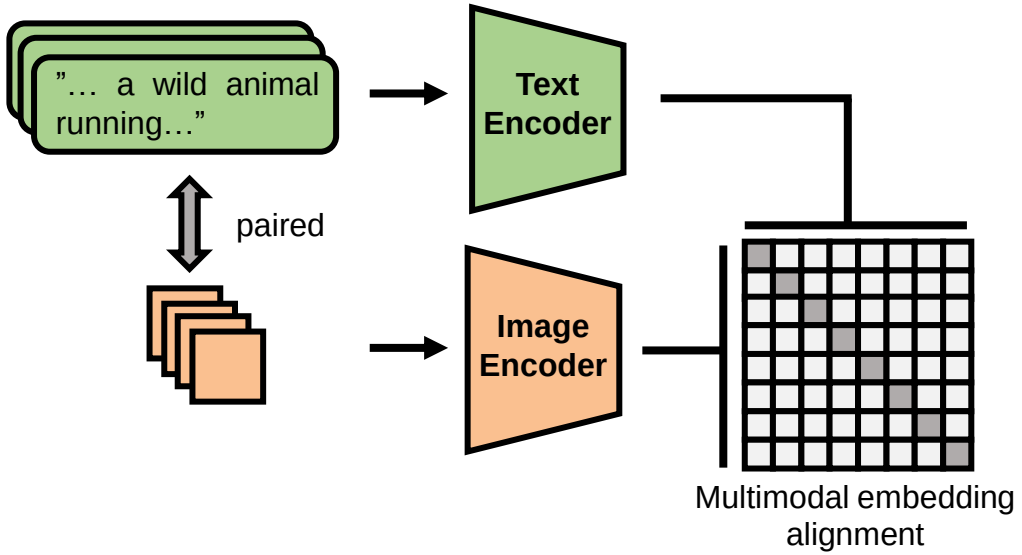
Jose Dolz

ÉTS Montreal

zero-shot weights	t_c
image features	v
LP prototypes	w_c

Large Vision-Language Models

- **CLIP pre-training**



- **Zero-shot inference**

$$\hat{y}_c = \frac{\exp(\boldsymbol{v} \cdot \boldsymbol{t}_c/\tau)}{\sum_{i=1}^{C} \exp(\boldsymbol{v} \cdot \boldsymbol{t}_i/\tau)}$$

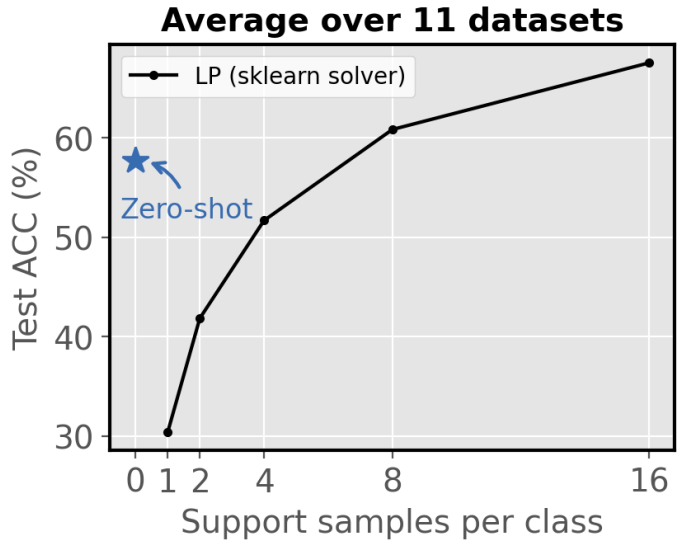
- **Few-shot adaptation**

(Linear Probing)

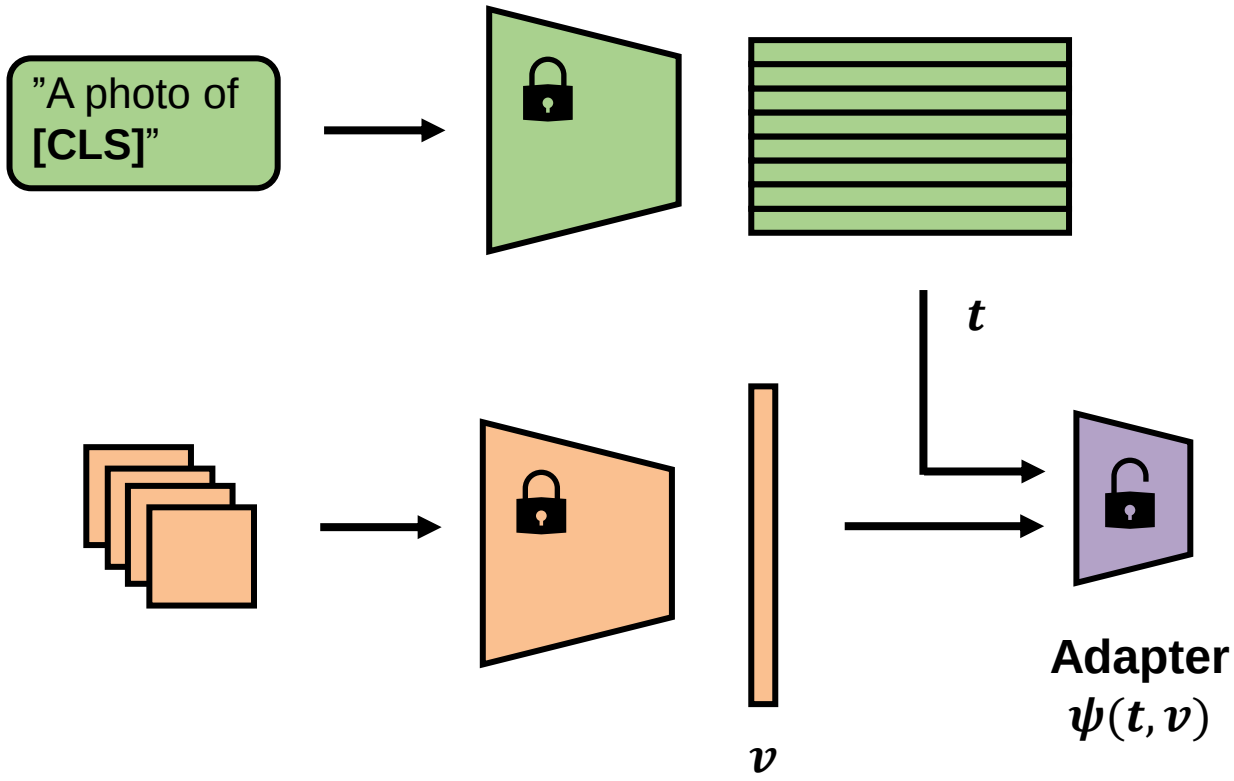
Support Set

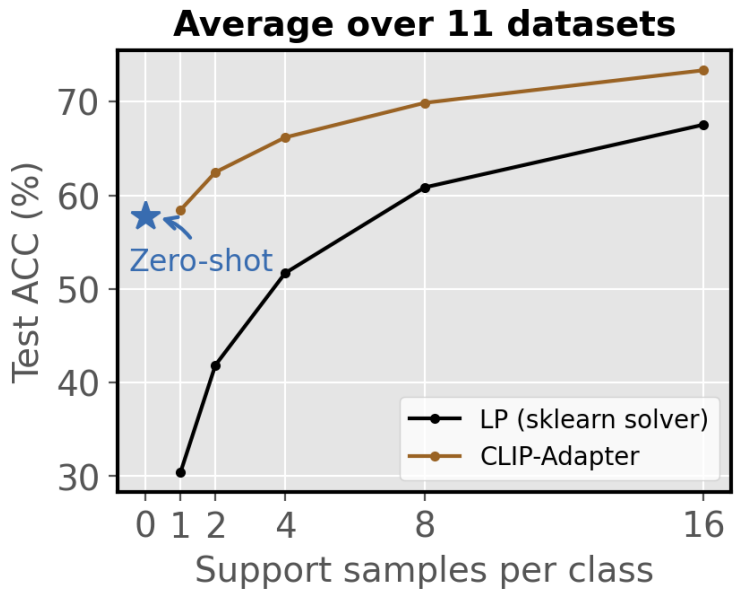
$$S = \{\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)}\}_{m=1}^{M=K\text{x}C}$$

$$\hat{y}_c = \frac{\exp(\boldsymbol{v} \cdot \boldsymbol{w}_c/\tau)}{\sum_{i=1}^{C} \exp(\boldsymbol{v} \cdot \boldsymbol{w}_i/\tau)} \rightarrow \min_{\boldsymbol{w}} \sum_{m=1}^{M} \mathcal{H}\left(\boldsymbol{y}^{(m)}, \hat{\boldsymbol{y}}^{(m)}\right)$$



Vision-Language Adapters

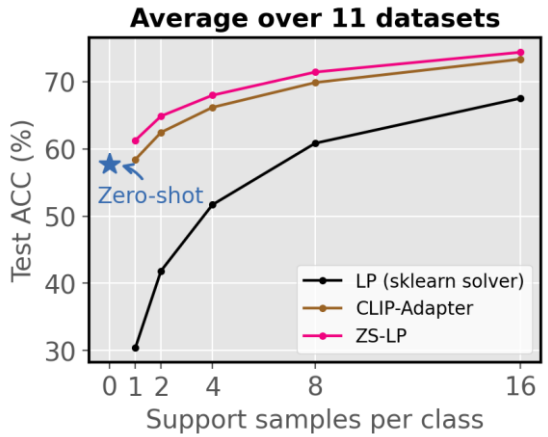




Revisiting Linear Probing

- The initial approximation of Linear Probing for CLIP offers limited performance.
- Zero-Shot Linear Probe (**ZS-LP**):
  - ✓ **Pre-training head design.**
  - ✓ **Zero-shot weights initialization**.

Method	K=1	K=2	K=4
ZS-LP	61.28	64.88	67.98
w/o DA	$57.72_{(-3.5)}\downarrow$	$61.94_{(-2.9)}\downarrow$	$65.41_{(-2.5)}\downarrow$
w/o Temp. Scaling (τ)	$58.33_{(-2.9)}\downarrow$	$59.85_{(-5.0)}\downarrow$	$59.91_{(-8.0)}\downarrow$
w/o L^2-norm	$48.67_{(-12.6)}\downarrow$	$55.29_{(-9.6)}\downarrow$	$61.16_{(-6.8)}\downarrow$
Rand. Init.	$30.42_{(-30.8)}\downarrow$	$41.86_{(-23.0)}\downarrow$	$51.69_{(-16.2)}\downarrow$



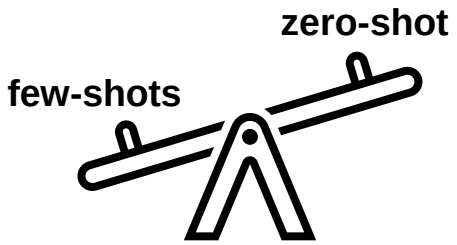
A well-initialized Linear Probe is all you need?

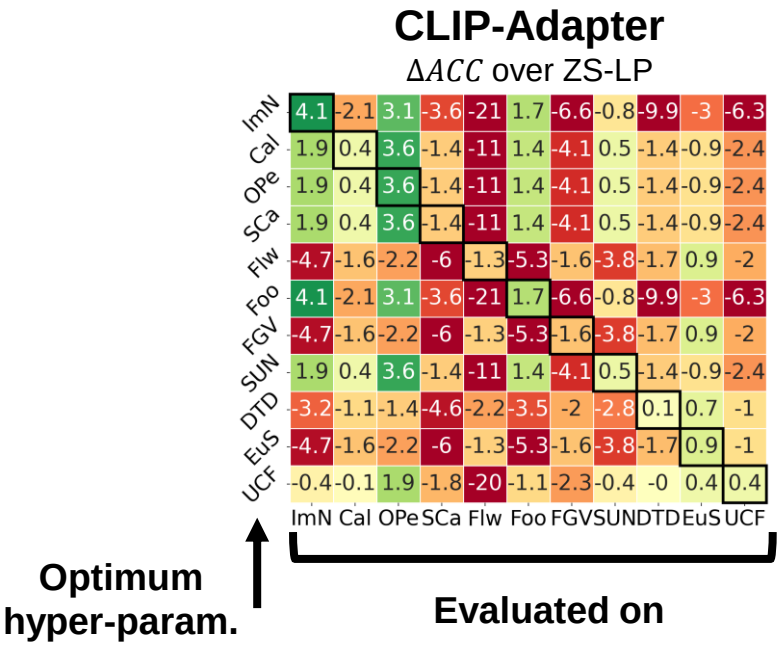
Pitfalls on Existing Few-Shot Adapters

- The **strength of text prototypes varies** on the difficulty of each target dataset, and the **modality gap** with respect to **pre-training data**.
- Prior works **combine zero-shot and few-shot knowledge** by using **blending hyperparameters that control how far they go from the initial solution**.

CLIP-Adapter

$$\boldsymbol{w} = \boldsymbol{t} \;\; ; \;\; \boldsymbol{v}' = \boldsymbol{v} + \alpha_r \, f_\varphi(\boldsymbol{v})$$





SoTA Adapters require validation data to outperform a Linear Probe

- **How do you find the best balance per dataset?**

Our Proposal: CLAP

- **CL**ass-**A**daptive linear **P**robing aims to train class protoypes, $\boldsymbol{w}_c$, **constraining them to remain close to text** prototypes, $\boldsymbol{t}_c$. To do so, we employ a penalty-based constraints.

- Since each category might present different zero-shot robustness and particular difficulty, we employ **weighting factor per class**, $\boldsymbol{\lambda_c}$.

$$\min_{\boldsymbol{w}} \underbrace{\sum_{m=1}^{M} \mathcal{H}\left(\boldsymbol{y}^{(m)}, \hat{\boldsymbol{y}}^{(m)}\right)}_{\textbf{Cross-entropy on few shots}} + \underbrace{\sum_{c=1}^{C} \lambda_c ||\boldsymbol{t}_c - \boldsymbol{w}_c||_2^2}_{\textbf{Learned prototypes constrained to zero-shot}}$$

- The penalty weights are treated as a **learnable parameters** via Augmented Lagrangian Multipliers (ALM).

$$\lambda_c^* = \frac{1}{|\beta_c^+|} \sum\nolimits_{i \epsilon \beta_c^+} \boldsymbol{y}_c^{(i)}$$

Quantitative Evaluation

Validation-free comparison

Method	K=1	K=2	K=4	K=8	K=16
Prompt-learning methods					
CoOp IJCV'22[46]	59.56	61.78	66.47	69.85	73.33
ProGrad ICCV'23[13]	62.61	64.90	68.45	71.41	74.28
PLOT ICLR'23[6]	62.59	65.23	68.60	71.23	73.94
Efficient transfer learning - *a.k.a Adapters*					
Zero-Shot ICML'21[30]	57.71	57.71	57.71	57.71	57.71
Rand. Init LP ICML'21[30]	30.42	41.86	51.69	60.84	67.54
CLIP-Adapter IJCV'23[11]	58.43	62.46	66.18	69.87	73.35
TIP-Adapter ECCV'22[42]	58.86	60.33	61.49	63.15	64.61
TIP-Adapter(f) ECCV'22[42]	60.29	62.26	65.32	68.35	71.40
CrossModal-LP CVPR'23[24]	62.24	64.48	66.67	70.36	73.65
TaskRes(e) CVPR'23[40]	61.44	65.26	68.35	71.66	74.42
ZS-LP	61.28	64.88	67.98	71.43	74.37
CLAP	**62.79**	**66.07**	**69.13**	**72.08**	**74.57**

Using a few-shot validation set

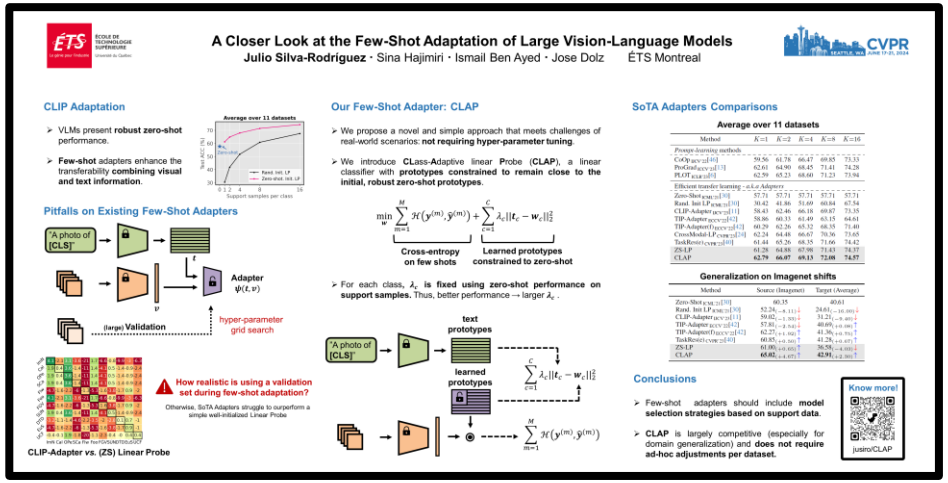
Method	K=1	K=2	K=4	K=8	K=16
Protocol in [24]: K-shots for train + $min(K, 4)$ for validation					
TIP-Adapter [42]	63.3	65.9	69.0	72.2	75.1
CrossModal LP [24]	64.1	67.0	70.3	73.0	76.0
CrossModal Adapter [24]	64.4	67.6	70.8	73.4	75.9
CrossModal PartialFT [24]	64.7	67.2	70.5	**73.6**	**77.1**
Ours: using $K + min(K, 4)$ shots for training					
ZS-LP	64.9	68.0	71.4	73.1	75.0
CLAP	**66.1**	**69.1**	**72.1**	**73.5**	75.1

Take-Home Messages

- **Linear Probing** (if properly designed) is a **strong baseline for few-shot CLIP Adaptation.**
- Few-shot adapters should include **model selection strategies based on support data**.
- **CLAP** is largely competitive and **does not require ad-hoc adjustments per dataset.**

See you in person

FRI-AM-403



Know more

jusiro/CLAP