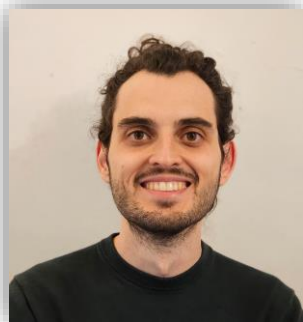


A Closer Look at the Few-Shot Adaptation of Large Vision-Language Models

Julio
Silva-Rodríguez



Sina
Hajimiri



Ismail
Ben Ayed



Jose Dolz

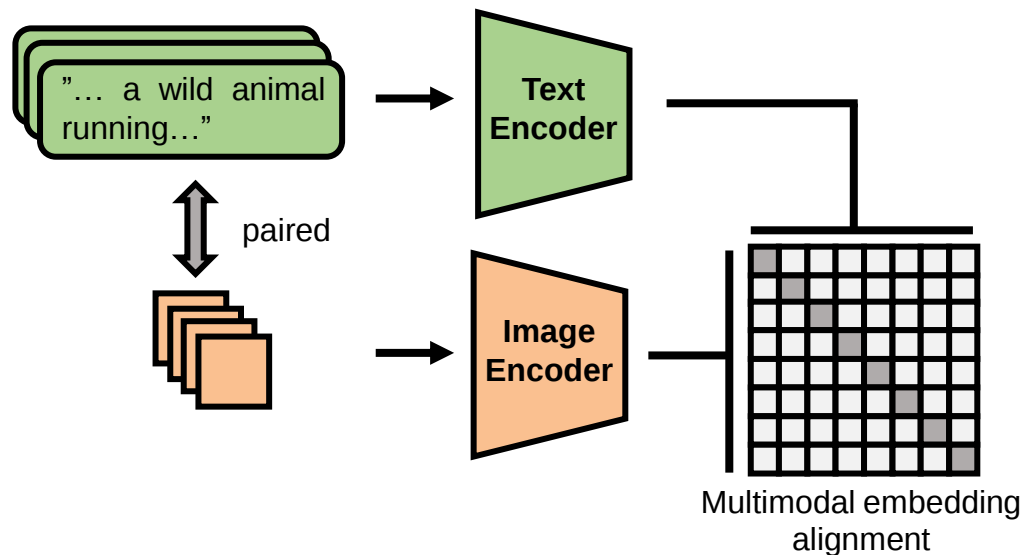


ÉTS Montreal

zero-shot weights	t_c
image features	v
LP prototypes	w_c

Large Vision-Language Models

➤ CLIP pre-training



➤ Zero-shot inference

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

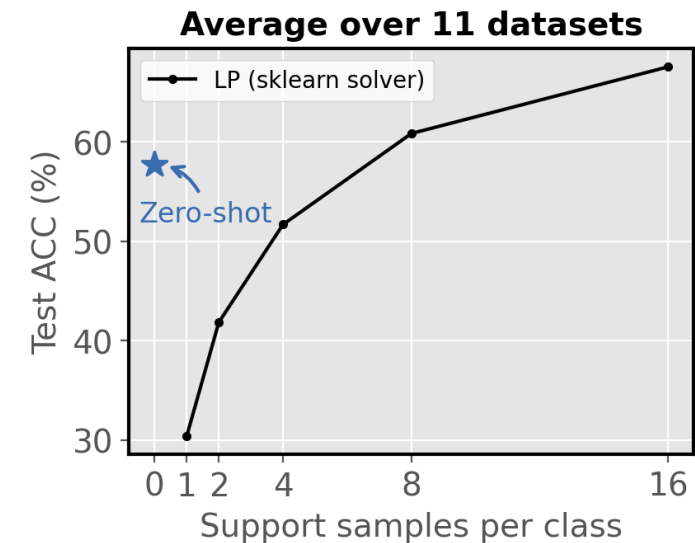
➤ Few-shot adaptation

(Linear Probing)

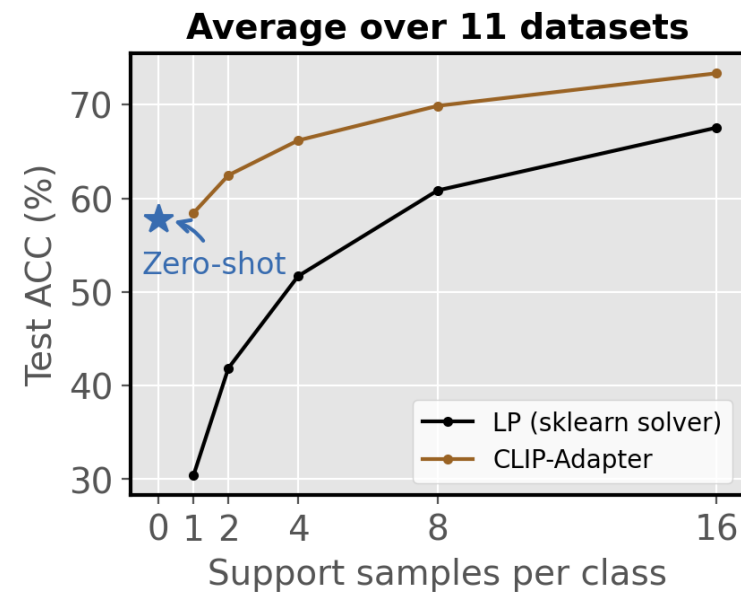
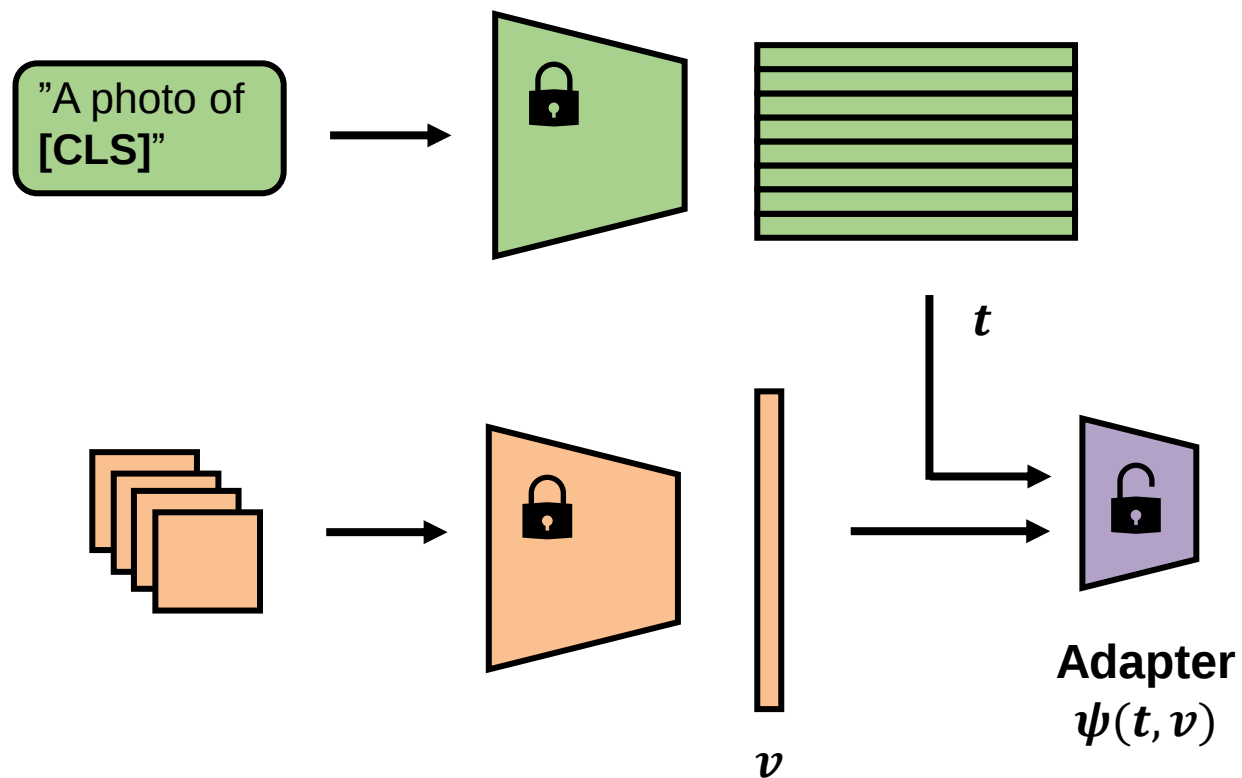
Support Set

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

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



Vision-Language Adapters



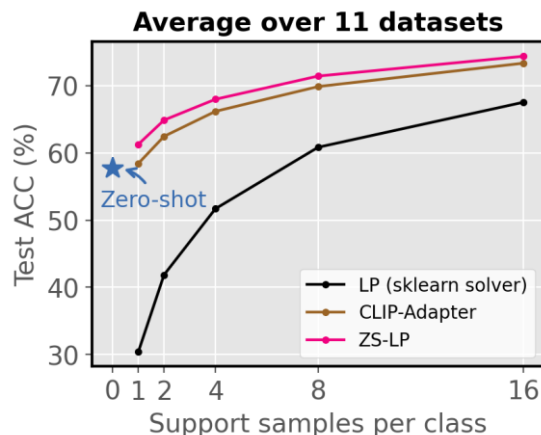
Revisiting Linear Probing

➤ 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) ↓	61.94 _(-2.9) ↓	65.41 _(-2.5) ↓
w/o Temp. Scaling (τ)	58.33 _(-2.9) ↓	59.85 _(-5.0) ↓	59.91 _(-8.0) ↓
w/o L^2 -norm	48.67 _(-12.6) ↓	55.29 _(-9.6) ↓	61.16 _(-6.8) ↓
Rand. Init.	30.42 _(-30.8) ↓	41.86 _(-23.0) ↓	51.69 _(-16.2) ↓



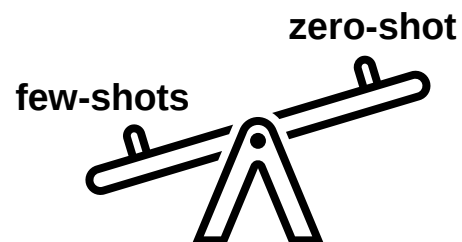
A well-initialized Linear Probe is all you need?

Pitfalls on Existing Few-Shot Adapters

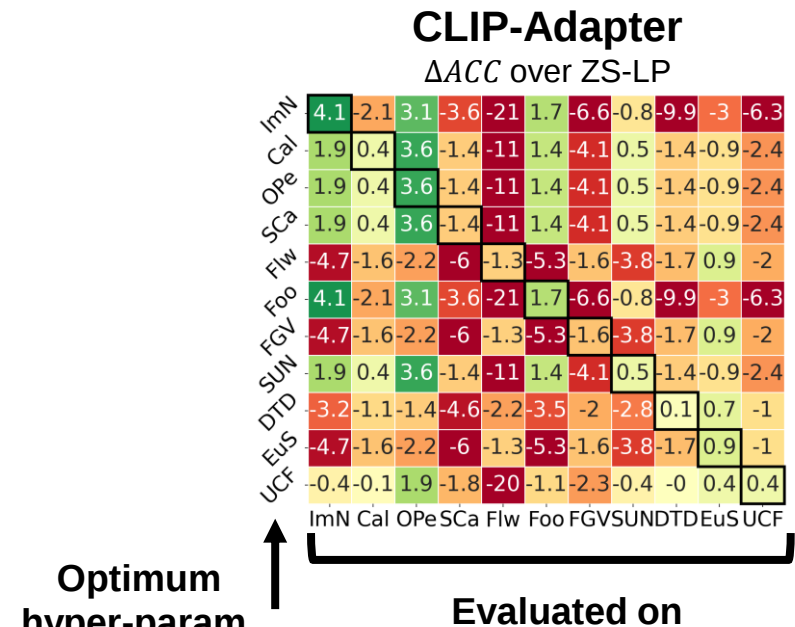
- The **strength of text prototypes varies** on the difficulty of each target dataset, and the **modality gap** with respect to **pre-training data**.
- Prior works **combine zero-shot and few-shot knowledge** by using **blending hyperparameters** that control how far they go from the initial solution.

CLIP-Adapter

$$w = t \ ; \ v' = v + \alpha_r f_\varphi(v)$$



- How do you find the best balance per dataset?



SoTA Adapters require validation data to outperform a Linear Probe

Our Proposal: CLAP

- **CL**ass-**A**daptive linear **P**robing aims to train class prototypes, $\mathbf{w}_c$, **constraining them to remain close to text** prototypes, $\mathbf{t}_c$. 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**, λ_c .

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

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

$$\lambda_c^* = \frac{1}{|\beta_c^+|} \sum_{i \in \beta_c^+} \mathbf{y}_c^{(i)}$$

Quantitative Evaluation

Validation-free comparison

Method	$K=1$	$K=2$	$K=4$	$K=8$	$K=16$
<i>Prompt-learning methods</i>					
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
<i>Efficient transfer learning - a.k.a Adapters</i>					
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

ÉTS ÉCOLE DE TECHNOLOGIE SUPÉRIEURE Université de Montréal

A Closer Look at the Few-Shot Adaptation of Large Vision-Language Models
Julio Silva-Rodriguez · Sina Hajimiri · Ismail Ben Ayed · Jose Dolz · ÉTS Montreal

CVPR

CLIP Adaptation

- VLMs present robust zero-shot performance.
- Few-shot adapters enhance the transferability combining visual and text information.

Pitfalls on Existing Few-Shot Adapters

A photo of [CLS] → Adapter $\psi(x, y)$ → Image Validation

hyper-parameter grid search

How realistic is using a validation set during few-shot adaptation?
Observation: SoTA Adapters struggle to outperform a simple well-tuned Linear Probe

CLIP-Adapter vs. (ZS) Linear Probe

Our Few-Shot Adapter: CLAP

- We propose a novel and simple approach that meets challenges of real-world scenarios: **not requiring hyper-parameter tuning**.
- We introduce **CLASs-Adaptive linear Probe (CLAP)**, a linear classifier with **prototypes constrained to remain close to the initial, robust zero-shot prototypes**.

$$\min_{\lambda} \sum_{m=1}^M \mathcal{H}(y^{(m)}, g^{(m)}) + \sum_{i=1}^C \lambda_i \|t_i - w_i\|^2$$

Cross-entropy on few shots Learned prototypes constrained to zero-shot

For each class, λ_i is fixed using zero-shot performance on support samples. Thus, better performance → larger λ_i .

A photo of [CLS] → text prototypes → learned prototypes → $\sum \lambda_i \|t_i - w_i\|^2$ → $\sum \mathcal{H}(y^{(m)}, g^{(m)})$

SoTA Adapters Comparisons

Average over 11 datasets

Method	K=1	K=2	K=4	K=8	K=16
CoOp (w/o CLIP)	59.56	61.78	66.47	68.85	73.33
ProT (w/o CLIP)	62.61	64.90	68.45	71.61	74.28
ProT (w/ CLIP)	62.59	65.23	68.60	71.23	73.94
Efficient transfer learning - CLAP Adapters					
Zero-Shot (w/o CLIP)	57.71	57.71	57.71	57.71	57.71
Read, Int. LP (w/o CLIP)	58.41	61.86	63.69	66.84	67.34
CLIP-Adapter (w/o CLIP)	58.43	62.46	66.18	68.87	73.35
TP-Adapter (w/o CLIP)	58.36	60.33	63.49	65.55	64.61
TP-Adapter (w/ CLIP)	60.29	62.26	65.52	68.03	71.48
CrossModal LP (w/o CLIP)	62.34	64.48	66.67	70.36	73.65
TaskMetric (w/o CLIP)	61.44	62.26	65.35	71.66	74.42
ZS-LP	61.28	64.08	67.96	71.63	74.37
CLAP	62.79	66.07	69.13	71.88	74.87

Generalization on Imagenet shifts

Method	Source (Imagenet)	Target (Imagenet)
Zero-Shot (w/o CLIP)	60.35	40.61
Read, Int. LP (w/o CLIP)	52.38 (-4.41)	24.65 (-16.66)
CLIP-Adapter (w/o CLIP)	59.02 (-1.30)	31.21 (-9.14)
TP-Adapter (w/o CLIP)	57.86 (-2.44)	40.60 (-9.75)
TP-Adapter (w/ CLIP)	62.27 (+1.92)	41.36 (+0.75)
TaskMetric (w/o CLIP)	60.81 (+0.46)	41.28 (+0.67)
ZS-LP	61.00 (+0.65)	36.56 (-4.05)
CLAP	60.82 (+0.47)	42.91 (+2.25)

Conclusions

- Few-shot adapters should include **model selection strategies based on support data**.
- CLAP** is largely competitive (especially for domain generalization) and **does not require ad-hoc adjustments per dataset**.

Know more!

[jusiro/CLAP](https://github.com/jusiro/CLAP)

Know more



jusiro/CLAP