Robust Calibration of Large Vision-Language Adapters

Balamurali Murugesan*, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz

ETS Montréal, Canada

*balamurali.murugesan.1@ens.etsmtl.ca

Take-Home Message

Proposed method:

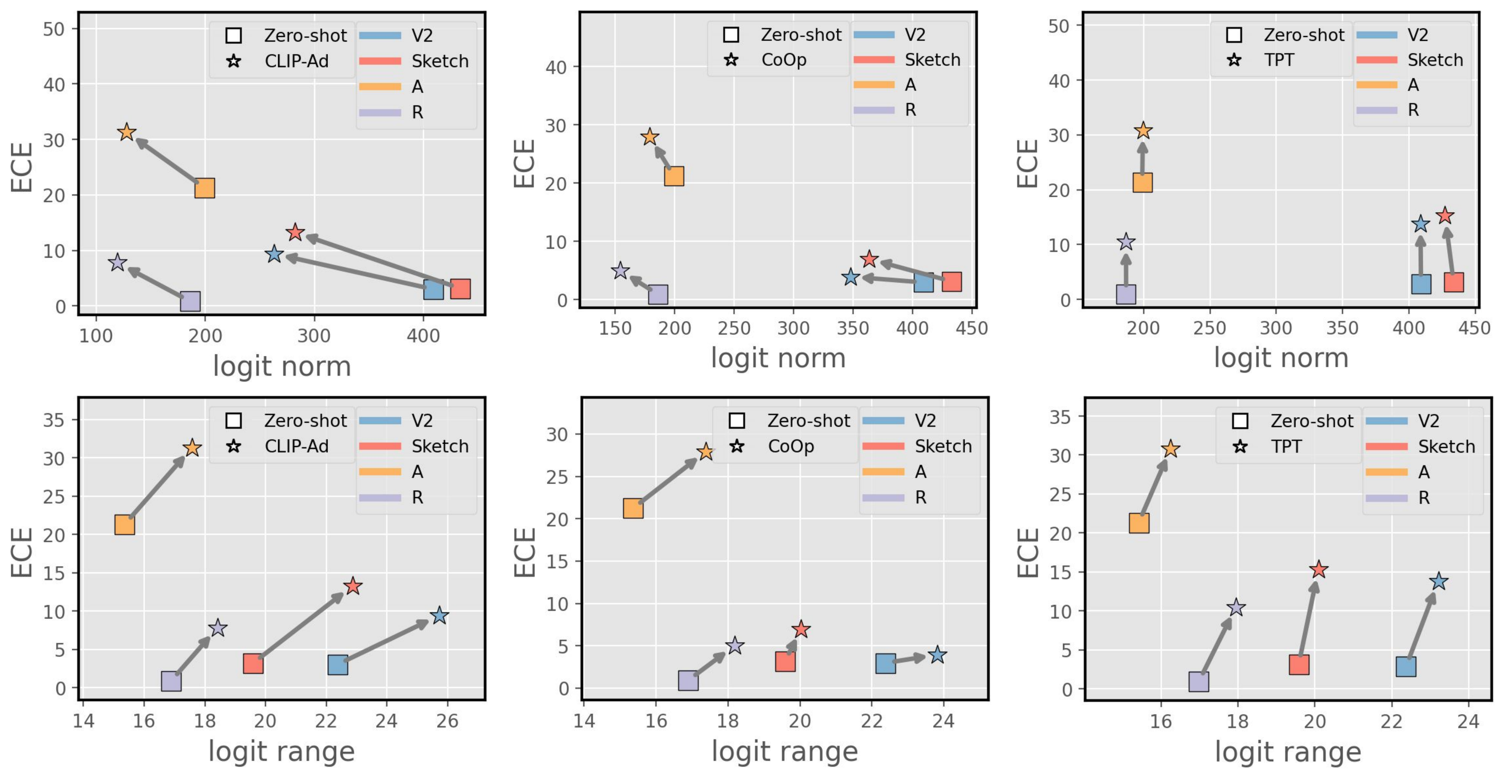
- We show that the underlying cause of miscalibration in adaptation is with the increase of logit ranges and demonstrated that the zero-shot baselines are better calibrated.
- We provide two solutions (normalization, penalty) during training and an unsupervised scaling during inference time to constrain the logit range based on the zero-shot logits.

Results:

- Our solutions reduce miscalibration error in popular OOD classification benchmarks for both adapters and prompt learning while keeping the discriminative performance.
- Incorporating our approaches decreases the logit range with typical increase in logit norm.

Introduction

- **Motivation:**
  - Deep learning is undergoing a paradigm shift with pre-trained large-scale language-vision models, such as CLIP [1].
  - Adapters [2], Prompt Learning [3], and TPT [4] methods have been developed to generalize for unseen related-domains.
  - These methods have improved the discriminative performance of a zero-shot baseline, but calibration is significantly degraded.
- **Background and observations:**
  - Recent literature [5] suggests that the miscalibration is caused by increasing the logit norm during training.
  - We expose that the underlying cause of miscalibration is, in fact, the increase of the logit ranges instead of norm.



- **Contributions:**
  - We empirically demonstrate that popular CLIP adaptation strategies, substantially degrade the calibration capabilities of the zero-shot baseline in the presence of distributional drift.
  - We present a simple, and model-agnostic solution, scaling the logit range of each sample based on the zero-shot logits.
  - Comprehensive experiments on popular OOD classification benchmarks demonstrate the effectiveness of our approaches.

Method

- **Formulation**

The logits used in training the main objective $\mathcal{H}(\boldsymbol{Y}, \boldsymbol{P})$ are constrained to the range of its zero-shot prediction by the following constrained problem:

$$\begin{aligned} &\text{minimize} \quad && \mathcal{H}(\boldsymbol{Y}, \boldsymbol{P}) \\ &\text{subject to} \quad && l_i^{\text{ZS-min}}\mathbf{1} \leq \boldsymbol{l}_i \leq l_i^{\text{ZS-max}}\mathbf{1} \qquad \forall i \in \mathcal{D} \end{aligned}$$

where $\boldsymbol{l}_i$ is the logit magnitude of sample $\boldsymbol{x}_i$, and $l_i^{\text{ZS-min}}$ and $l_i^{\text{ZS-max}}$ are the min and max logit magnitudes of its zero-shot prediction.

- **Sample-adaptive logit scaling (SaLS)**

The logit normalization of sample $\boldsymbol{x}_i$ at inference time is given by:

$$\boldsymbol{l}'_i = \frac{(l_i^{\text{ZS-max}} - l_i^{\text{ZS-min}})}{(l_i^{\max} - l_i^{\min})}(\boldsymbol{l}_i - l_i^{\min}\mathbf{1}) + l_i^{\text{ZS-min}}\mathbf{1}$$

where $l_i^{\max} = \max_j(l_{ij})$ and $l_i^{\min} = \min_j(l_{ij})$

- **Zero-shot logit normalization during training (ZS-Norm)**

The learning objective with normalized logit ($\boldsymbol{l}'_i$) is given by:

$$\mathcal{H}(\boldsymbol{Y}, \boldsymbol{P}) = -\sum_{i\in\mathcal{S}}\sum_{k=1}^{K} y_{ik}\log\frac{\exp(l'_{ik})}{\sum_{j=1}^{K}\exp(l'_{ij})}$$

where $\boldsymbol{l}'_i$ denotes the zero-shot normalized logit vector of $\boldsymbol{x}_i$

- **Integrating explicit constraints in the objective (Penalty)**

The objective function with ReLU penalties is given by:

$$\min_{\boldsymbol{\theta}} \quad \mathcal{H}(\boldsymbol{Y}, \boldsymbol{P}) + \lambda\sum_{i\in\mathcal{S}}\sum_{k=1}^{K}(\text{ReLU}(l_{ik} - l_i^{\text{ZS-max}}) + \text{ReLU}(l_i^{\text{ZS-min}} - l_{ik}))$$

where λ controls the trade-off between the main loss and penalties.

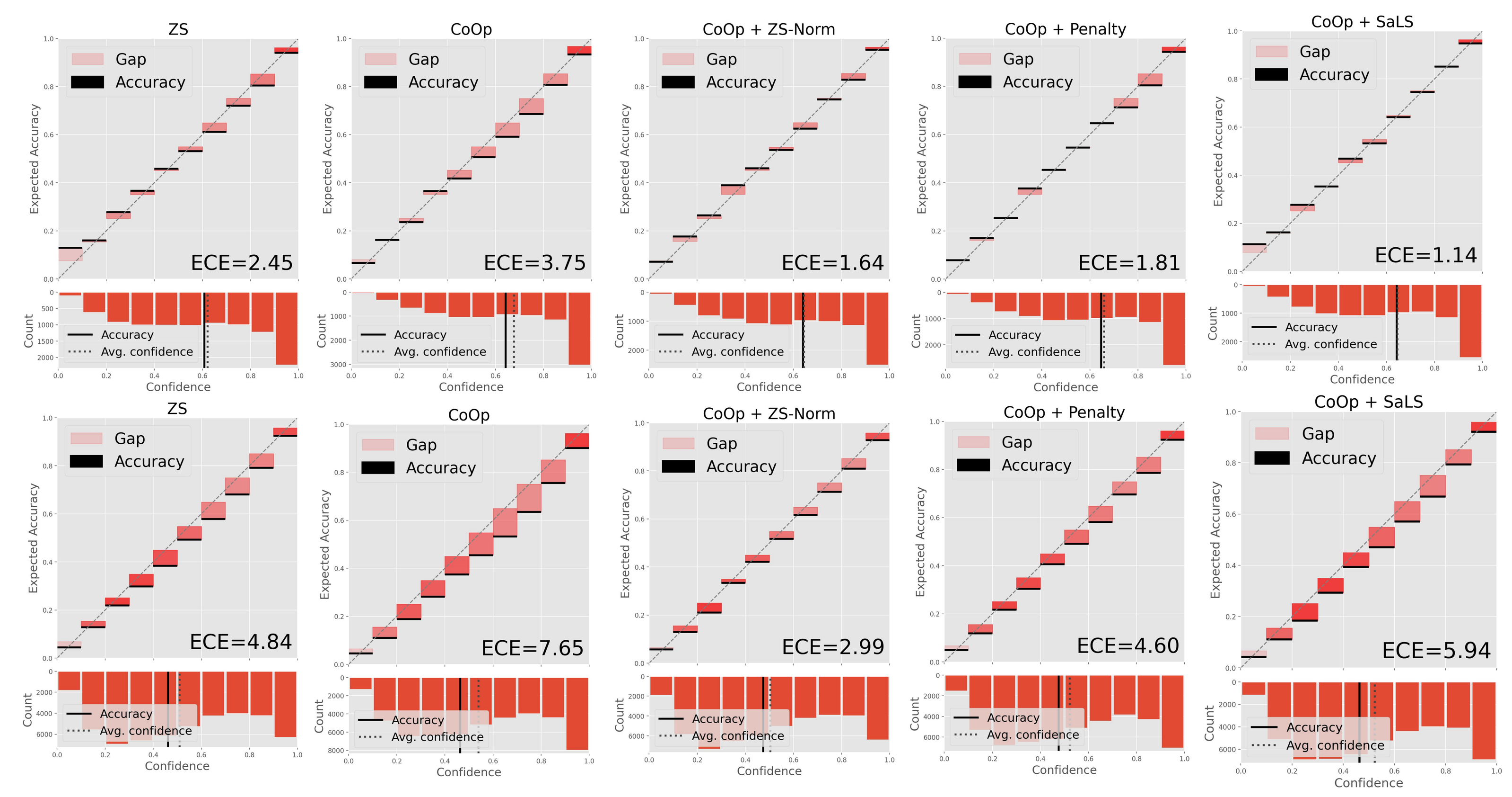
Repository : https://github.com/Bala93/CLIPCalib

Results

(1) Quantitative performance: Logit range scaling provided improved calibration on different prompt learning methods.

Method	Avg. OOD ACC	Avg. OOD ECE
Zero-Shot	57.15	4.78
CoOp	58.41	6.61
w/ **ZS-Norm**	58.75 $_{(+0.34)}\uparrow$	**4.35** $_{(-2.26)}\downarrow$
w/ **Penalty**	**59.18** $_{(+0.77)}\uparrow$	4.91 $_{(-1.70)}\downarrow$
w/ **SaLS**	58.41	4.90 $_{(-1.71)}\downarrow$
CoCoOp	59.74	4.83
w/ **ZS-Norm**	59.90 $_{(+0.16)}\uparrow$	3.94 $_{(-0.89)}\downarrow$
w/ **Penalty**	**60.20** $_{(+0.46)}\uparrow$	**3.89** $_{(-0.94)}\downarrow$
w/ **SaLS**	59.74	4.81 $_{(-0.00)}\sim$
MaPLe	60.07	4.13
w/ **ZS-Norm**	60.09 $_{(+0.02)}\uparrow$	**3.59** $_{(-0.14)}\downarrow$
w/ **Penalty**	**60.62** $_{(+0.55)}\uparrow$	3.78 $_{(-0.35)}\downarrow$
w/ **SaLS**	60.07	4.38 $_{(+0.25)}\uparrow$

(2) Qualitative results: Reliability plot for Prompt learning method CoOp with ImageNetV2 (Top), and ImageNetSketch (Bottom).



References

[1] Radford et al. Learning visual models from natural language supervision. In *ICML*, 2021.
[2] Gao et al. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 2024.
[3] Zhou et al. Learning to prompt for vision-language models. *IJCV*, 2022.
[4] Shu et al. Test-time prompt tuning for vision-language models. *NeurIPS*, 2022.
[5] Wei et al. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022.