Robust Calibration of Large Vision-Language Adapters

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

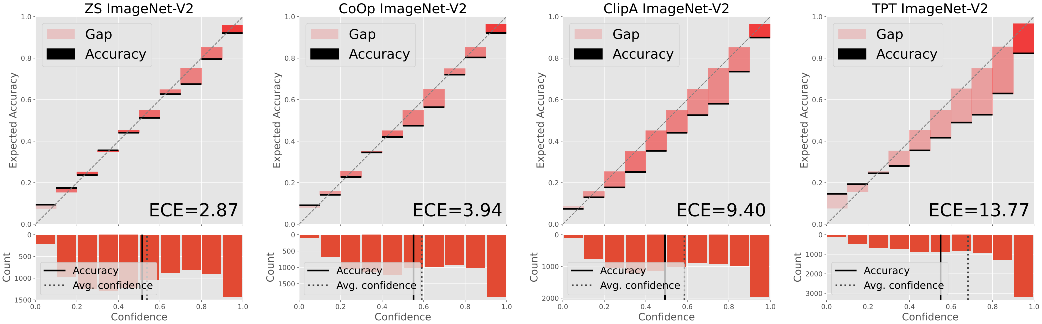
Oct 2, 2024

Le génie pour l'industrie

Table of Contents

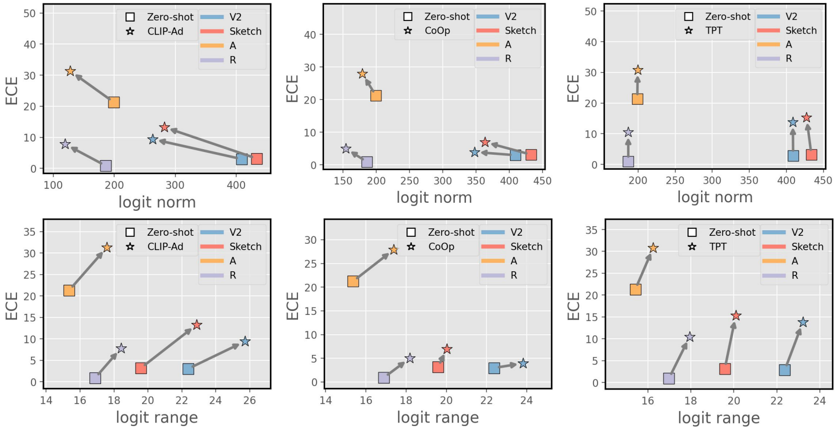
Motivation

- Deep learning is undergoing a paradigm shift with pre-trained large-scale language-vision models, such as CLIP [Rad+21].
- Adapters [Gao+24], Prompt Learning [Zho+22b], and TPT [Shu+22] 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.



Observation

- Recent literature [Wei+22] 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.



Formulation

- For sample $\boldsymbol{x}$, let $\boldsymbol{y} \in \{0,1\}^K$ be the ground truth vector, $\boldsymbol{p}$ be the softmax probability of logits $\boldsymbol{l}$ obtained from CLIP models.
- 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.

Solution

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

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

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

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

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

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

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

Few-shot Prompt Learning and Adapters

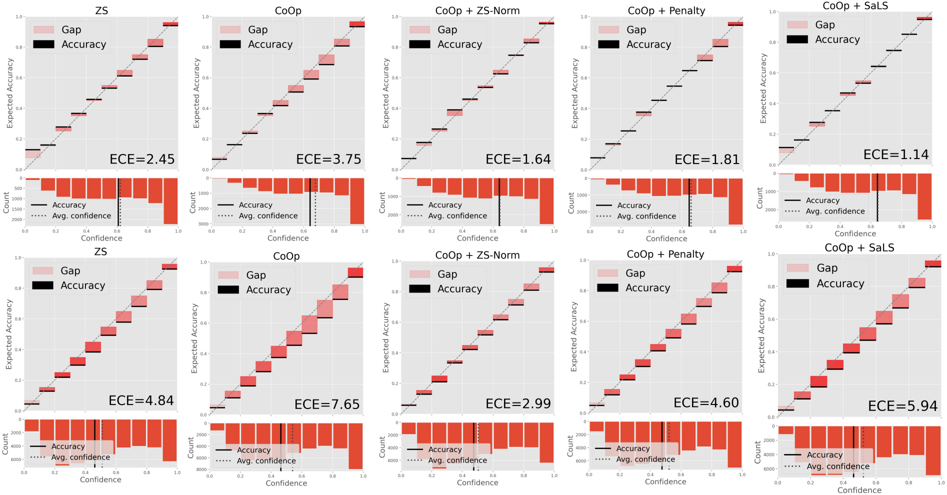
Method	Avg. OOD ACC	Avg. OOD ECE
Zero-Shot [Rad+21]	40.62	7.18
CoOp [Zho+22b]	40.86	10.97
w/ **ZS-Norm**	41.59$_{(+0.73)}$↑	10.19$_{(-0.78)}$↓
w/ **Penalty**	**41.87**$_{(+1.01)}$↑	8.06$_{(-2.91)}$↓
w/ **SaLS**	40.86	**7.82**$_{(-3.15)}$↓
CoCoOp [Zho+22a]	43.36	7.69
w/ **ZS-Norm**	43.70$_{(+0.34)}$↑	7.12$_{(-0.57)}$↓
w/ **Penalty**	**43.86** $_{(+0.50)}$↑	**6.15** $_{(-1.54)}$↓
w/ **SaLS**	43.36	6.82$_{(-1.87)}$↓
ProGrad [Zhu+23]	42.32	7.66
w/ **ZS-Norm**	42.21 $_{(+0.11)}$↑	7.98 $_{(+0.32)}$↑
w/ **Penalty**	**42.57**$_{(+0.25)}$↑	**6.84**$_{(-0.82)}$↓
w/ **SaLS**	42.32	6.90$_{(-0.76)}$↓

Method	Avg. OOD ACC	Avg. OOD ECE
Zero-Shot [Rad+21]	40.62	7.18
CLIP-Ad [Gao+24]	34.07	15.45
w/ **ZS-Norm**	30.06$_{(-4.01)}$↓	21.27$_{(+5.82)}$↑
w/ **Penalty**	**35.20**$_{(+1.13)}$↑	11.22$_{(-4.23)}$↓
w/ **SaLS**	34.07	**8.95**$_{(-6.50)}$↓
TIP-Ad(f) [Zha+21]	41.45	19.04
w/ **ZS-Norm**	41.73$_{(+0.28)}$↑	19.80$_{(+0.76)}$↑
w/ **Penalty**	**43.73**$_{(+2.28)}$↑	12.18$_{(-6.86)}$↓
w/ **SaLS**	41.45	**8.13**$_{(-10.91)}$↓
TaskRes [Yu+23]	41.18	11.25
w/ **ZS-Norm**	**41.30**$_{(+0.12)}$↑	9.07$_{(-2.18)}$↓
w/ **Penalty**	41.29$_{(+0.11)}$↑	10.62$_{(-0.63)}$↓
w/ **SaLS**	41.18	**9.03**$_{(-2.22)}$↓

Test-time Prompt Tuning

		Avg.	INet	CAL	PET	CAR	FLW	FOO	AIR	SUN	DTD	SAT	UCF
ACC	Zero-shot [Rad+21]	56.03	58.17	85.68	83.62	55.75	61.67	73.96	15.69	58.82	40.43	23.69	58.90
	TPT [Shu+22]	58.03	60.74	87.22	84.49	58.36	62.81	74.97	17.58	61.17	42.08	28.40	60.61
	w/ **ZS-Norm**	57.94	60.69	87.38	84.41	58.45	62.12	75.01	17.13	61.09	41.96	28.53	60.59
	w/ **Penalty**	57.69	60.74	87.06	84.30	58.13	61.84	75.17	17.22	61.11	42.02	26.60	60.35
	w/ **SaLS**	**58.03**	60.74	87.22	84.49	58.36	62.81	74.97	17.58	61.17	42.08	28.40	60.61
	C-TPT [Yoo+24]	57.54	60.02	87.18	83.65	56.41	64.80	74.89	16.62	60.72	41.55	27.06	60.01
	w/ **ZS-Norm**	**57.63**	60.00	87.06	83.65	56.57	65.04	74.82	16.86	60.58	41.61	27.51	60.27
	w/ **Penalty**	57.52	60.06	86.94	83.51	56.78	64.76	74.88	16.29	60.67	41.90	26.63	60.32
	w/ **SaLS**	57.54	60.02	87.18	83.65	56.41	64.80	74.89	16.62	60.72	41.55	27.06	60.01
ECE	Zero-shot [Rad+21]	5.04	1.90	3.56	5.64	4.17	2.10	2.35	6.31	3.79	8.60	14.40	2.66
	TPT [Shu+22]	11.27	11.34	4.10	3.78	3.70	13.66	5.18	15.57	9.20	25.29	21.00	11.20
	w/ **ZS-Norm**	10.57	10.81	4.29	3.71	3.62	13.29	4.73	15.28	8.50	23.95	17.61	10.49
	w/ **Penalty**	9.58	11.31	3.99	1.57	2.26	13.94	4.27	14.51	8.88	23.10	11.82	9.78
	w/ **SaLS**	**9.26**	9.81	4.45	2.90	2.50	12.01	3.91	15.23	8.64	21.09	12.31	9.05
	C-TPT [Yoo+24]	6.33	3.05	2.60	2.46	0.87	3.91	1.62	11.30	2.73	21.38	13.58	2.88
	w/ **ZS-Norm**	5.74	2.85	2.29	2.69	0.78	3.53	1.61	10.94	2.72	20.94	12.17	2.65
	w/ **Penalty**	**3.14**	5.93	2.26	2.66	0.81	3.79	1.64	11.58	2.74	20.49	10.83	2.51
	w/ **SaLS**	5.22	2.21	3.41	3.94	2.55	1.75	1.78	10.15	2.58	12.92	10.41	2.71

Reliability plots



Conclusion

- We show that the underlying cause of miscalibration in adaptation is with the increase of logit ranges.
- 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.
- Our solutions reduce miscalibration error in popular OOD classification benchmarks for adapters, prompt learning, and test-time prompt tuning.

[Gao+24] Peng Gao et al. "CLIP-Adapter: Better Vision-Language Models with Feature Adapters". In: *International Journal of Computer Vision (IJCV)* (2024).

[Rad+21] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[Shu+22] Manli Shu et al. "Test-time prompt tuning for zero-shot generalization in vision-language models". In: *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 14274–14289.

[Wei+22] Hongxin Wei et al. "Mitigating neural network overconfidence with logit normalization". In: *International Conference on Machine Learning (ICML)*. PMLR. 2022, pp. 23631–23644.

[Yoo+24] Hee Suk Yoon et al. "C-TPT: Calibrated Test-Time Prompt Tuning for Vision-Language Models via Text Feature Dispersion". In: *International Conference on Learning Representations (ICLR)*. 2024.

[Yu+23] Tao Yu et al. “Task Residual for Tuning Vision-Language Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 10899–10909.

[Zha+21] Renrui Zhang et al. “Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling”. In: *CoRR* abs/2111.03930 (2021). arXiv: 2111.03930.

[Zho+22a] Kaiyang Zhou et al. “Conditional Prompt Learning for Vision-Language Models”. In: *CVPR*. 2022.

[Zho+22b] Kaiyang Zhou et al. “Learning to Prompt for Vision-Language Models”. In: *International Journal of Computer Vision (IJCV)* (2022).

[Zhu+23] Beier Zhu et al. “Prompt-aligned gradient for prompt tuning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 15659–15669.