Open-Vocabulary Calibration for Fine-tuned CLIP

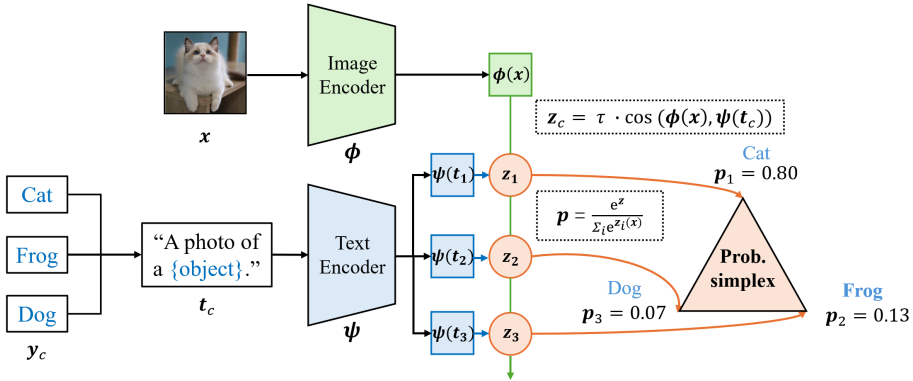
Shuoyuan Wang Jindong Wang Guoqing Wang Bob Zhang
Kaiyang Zhou Hongxin Wei

ICML 2024

Code:

Contrastive Language-Image Pretraining (CLIP)

- CLIP is a multi-modal model that leverages contrastive learning to enable zero-shot inference for open-vocabulary classes.



Confidence Calibration

- Confidence calibration ensures that the predicted probabilities reflect the true likelihood of correctness.
- Calibration is critical for deploying models in real-world applications, especially in high-stakes scenarios.

Expected Calibration Error (ECE)

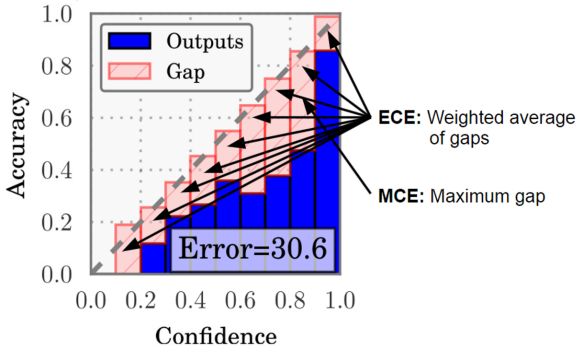
$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Bigg| \underbrace{\frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)}_{\text{accuracy}} - \underbrace{\frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i}_{\text{confidence}} \Bigg|$$

Calibration Visualization

Reliability diagram is designed to demonstrate the consistency/gap between the confidence of model predictions and the actual outcomes.

Steps for Construction:

- Binning
- Average Predicted Probability
- Observed Frequency

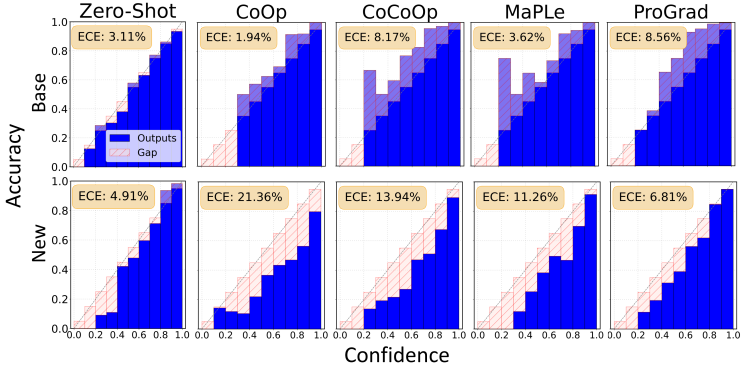


A close look at fine-tuned CLIP calibration

- To achieve optimal performance in specialized domains, CLIP requires fine-tuning (e.g., prompt tuning).
- Despite the improvements in accuracy, the calibration of fine-tuned CLIP models has not been explored.

Can fine-tuned CLIP be calibrated?

- Fine-tuning can result in underconfidence for base classes and overconfidence for open-vocabulary classes.



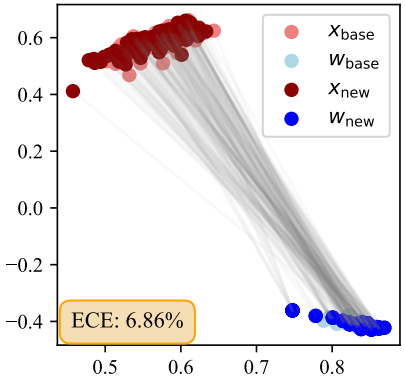
Can fine-tuned CLIP be calibrated?

- Post-hoc calibration can remedy miscalibration in base classes.
- Post-hoc calibration on base classes can not transfer to open-vocabulary classes.

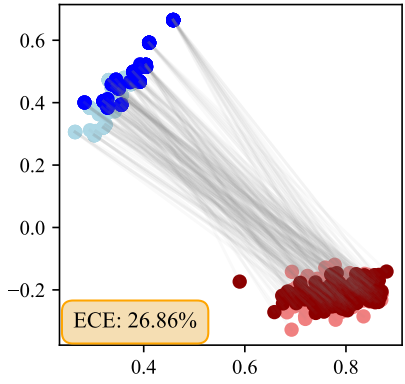
	ZS	Conf	TS	DEN	HB	IR	MIR
Base	3.58	4.82	1.94	0.73	4.23	2.09	0.82
Open	2.09	1.59	3.90	3.86	-	-	-

Multi-modal Feature Visualization

- Compared with zero-shot CLIP, CoOp has a larger textual distribution gap for both base and open-vocabulary classes.



(a) Zero-shot



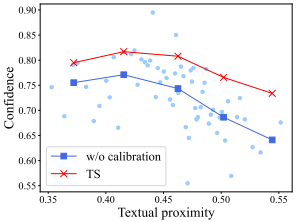
(b) CoOp

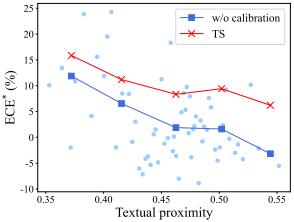
Motivation

Textual Proximity

$$P(\boldsymbol{w}_i, \mathcal{W}) = \exp\left(-\frac{1}{K}\sum_{\boldsymbol{w}_j \in \mathcal{N}_K(\boldsymbol{w}_i, \mathcal{W})} \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2\right),$$

where $\boldsymbol{w}_i$ is normalized textual embedding. $\mathcal{N}_K(\boldsymbol{w}_i, \mathcal{W})$ denotes the set of K nearest neighbors of $\boldsymbol{w}_i$.





- Lower proximity correlates with higher confidence and ECE.

Distance-Aware Calibration

Textual Deviation Estimation

$$\gamma(c_i) = \frac{P(\boldsymbol{w}_i', \mathcal{W}')}{P(\boldsymbol{w}_i, \mathcal{W})},$$

where $\mathcal{W}$ and $\mathcal{W}'$ denote the sets of normalized textual features for base classes.

Calibrated Inference

$$L_c^{dac}(\boldsymbol{x}) = \gamma(\hat{c}) \cdot \tau \cdot \text{sim}\left(\phi(\boldsymbol{x}), \psi(\boldsymbol{t}_c')\right),$$

where $\hat{c} = \arg\max_c p(c \mid \boldsymbol{x})$ is the predicted class $\boldsymbol{t}_c'$ is the textual token of class c.

Experimental Setup

Image classification datasets:

- ImageNet
- Caltech101
- OxfordPets
- StanfordCars
- Flowers102
- Food101
- FGVCAircraft
- SUN397
- UCF101
- DTD
- EuroSAT

Compared methods:

- CoOp (IJCV'22)
- CoCoOp (CVPR'22)
- ProDA (CVPR'22)
- MaPLe (CVPR'23)
- ProGrad (ICCV'23)
- PromptSRC (ICCV'23)

Model architectures:

- Vit-B-16 (Radford *et al.*, 2021)

Results on Few-shot Classification Benchmarks

Method	ECE(↓)		ACE(↓)		MCE(↓)		PIECE(↓)	
	Conf	DAC	Conf	DAC	Conf	DAC	Conf	DAC
CoOp	13.84	**7.00**	13.76	**6.91**	3.80	**1.71**	14.71	**9.02**
CoCoOp	6.29	**4.82**	6.21	**4.77**	1.79	**1.40**	8.07	**7.15**
ProDA	4.27	**3.99**	4.35	**4.08**	1.27	1.32	6.57	**6.35**
KgCoOp	4.36	4.32	4.43	4.38	1.18	1.13	6.67	6.63
MaPLe	5.77	**4.61**	5.71	**4.64**	1.82	**1.42**	7.59	**6.98**
ProGrad	4.22	**3.74**	4.27	**3.74**	1.22	**1.09**	6.75	**6.55**
PromptSRC	3.84	**3.63**	3.92	**3.69**	1.09	1.08	6.26	6.17

Table: Average calibration performance across 11 datasets.

- DAC improves open-vocabulary calibration in existing prompt tuning

Detailed Calibration Results

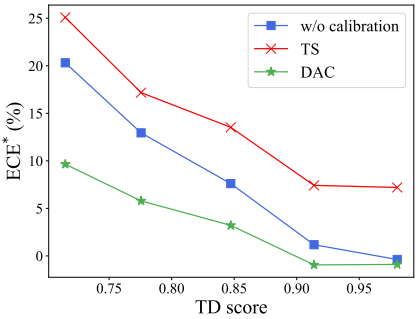
Method		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CoOp	Conf	0.00	19.24	18.86	15.87	20.42	33.28	30.30	37.84	41.60	18.57
	DAC	0.00	4.95	8.17	11.33	6.51	11.42	24.12	17.41	11.37	-0.94
	Δ	0.00	-14.29	-10.69	-4.54	-13.91	**-21.86**	-6.18	**-20.43**	**-30.23**	-19.51
MaPLe	Conf	0.00	0.00	-12.80	3.90	16.73	10.50	38.07	23.93	19.11	12.13
	+DAC	0.00	-3.62	-15.32	5.72	6.74	3.12	15.45	6.16	9.29	6.55
	Δ	0.00	-3.62	-2.52	1.82	**-10.99**	-7.38	**-22.62**	**-17.77**	-9.82	-5.58
ProGrad	Conf	0.00	-3.82	0.14	-0.10	4.29	6.31	3.48	8.11	1.23	4.86
	+DAC	0.00	-0.71	0.03	1.30	-1.32	0.40	-0.65	-0.04	0.44	-0.34
	Δ	0.00	3.11	-0.11	1.40	**-5.61**	**-5.91**	-4.13	**-8.15**	-0.79	-5.20

Table: Calibration results of ECE (%) across different confidence levels.

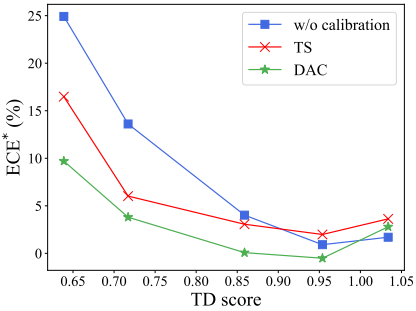
- DAC significantly reduces calibration error, especially for predictions with higher confidence.

Results Visualization

- DAC significantly reduces calibration error, especially for high-confidence predictions.
- DAC regularizes the prediction with low TD score for better calibration.



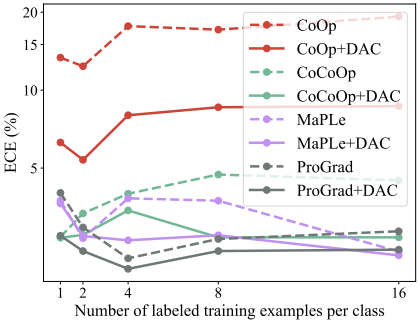
(a) StanfordCars



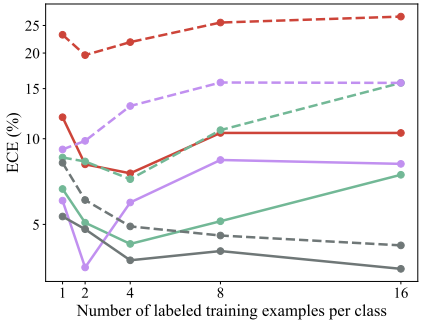
(b) SUN397

Ablation Study

- DAC is robust to the number of nearest neighbors.



(a) UCF101



(b) DTD

Conclusion

- **Problem**: We find that fine-tuned VLMs generally suffer from miscalibration, and existing post-hoc calibration methods often fail in the open-vocabulary setting.
- **Analysis**: We empirically study the correlation between the calibration and the textual distribution gap. We show that after prompt learning, VLMs tend to be overconfident on classes far away from base classes.
- **Method**: We introduce DAC, a simple and effective post-hoc calibration method that rectifies the predicted confidence while maintaining classification accuracy.

ArXiv: `https://arxiv.org/abs/2402.04655`
Code: `https://github.com/ml-stat-Sustech/CLIP_Calibration`