Understanding and Mitigating Miscalibration in Prompt Tuning for Vision-Language Models

Shuoyuan Wang[1], Yixuan Li[2], Hongxin Wei[1]

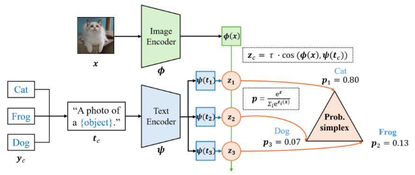
[1]Southern University of Science and Technology, [2]University of Wisconsin–Madison

ICML International Conference On Machine Learning

Background

Contrastive Language-Image Pretraining (CLIP)

- CLIP is a multi-modal model that leverages contrastive learning to process image and text data. Compared to traditional models, CLIP has zero-shot capabilities without specific task data for training.
- CLIP requires fine-tuning (e.g., prompt tuning) for better performance in the downstream tasks.

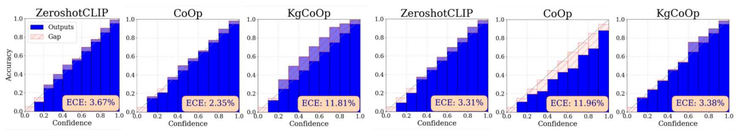


Confidence Calibration

- In the classification problem, the maximum output value from the softmax can serve as a rough estimate of confidence.
- Confidence calibration ensures that the predicted probabilities reflect the true likelihood of correctness.

$$\mathrm{ECE} = \sum_{g=1}^{G} \frac{|b_g|}{N} \left|\mathrm{acc}\left(b_g\right) - \mathrm{conf}\left(b_g\right)\right|$$

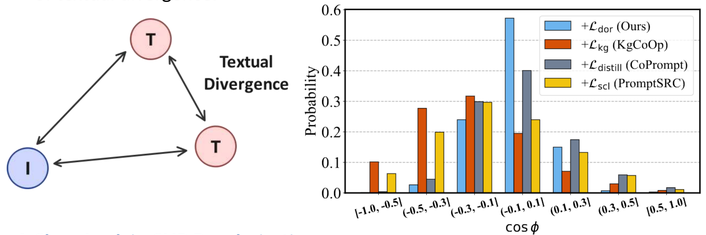
Motivation & Analysis



(a) Base (b) New

Empirical Study

- We show that current prompt-tuning methods typically lead to a trade-off between base and new classes, compromising one of them
- CoOp leads to overconfidence on new classes by increasing the textual divergence, while KgCoOp anchors the confidence level by hindering the increase of textual divergence.



A Close Look in CLIP Regularization

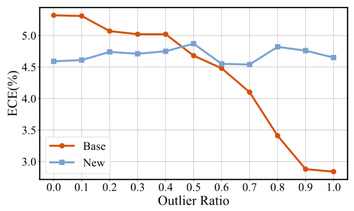
- We analyze the calibration issue from the perspective of gradient conflict.
- As shown in figure above, the gradient conflict distributions for KgCoOp, CoPrompt, and PromptSRC are predominantly within the range of [−1, 0], which indicates a conflict with the original learning objective.

$$\mathcal{L}_{\text{clip}} = \mathcal{L}_{\text{ce}} + \lambda \cdot \mathcal{L}_{\text{reg}} \longrightarrow \frac{1}{C}\sum_{c=1}^{C} \mathrm{sim}\left(\psi\left(\boldsymbol{t}'_c\right), \psi\left(\boldsymbol{t}_c\right)\right)$$

Method: Dynamic Outlier Regularization

Toy Example

- KgCoOp uses a fixed number of base classes as the regularization term, we progressively replaced these texts with textual outliers at varying proportions.
- Outlier-based regularization can break the ***calibration trade-off***.



Using Textual outliers for CLIP Regularization!

Dynamic Outlier Regularization

- We use WordNet (Miller, 1995) as the database. We select nouns from WordNet that do not overlap but share higher level concept relations with the base classes used in the fine-tuning.

$$s_i = \frac{1}{n}\sum_{j=1}^{n} \mathrm{sim}\left(\psi\left(\boldsymbol{t}_{o_i}\right), \psi\left(\boldsymbol{t}_{c_j}\right)\right),$$

$$\mathcal{O}_{\text{out}} = \{o_i \mid i \in \mathrm{TopK}\left(s_1, s_2, \ldots, s_m\right)\}$$

- DOR minimizes the feature discrepancy of textual outliers between the zero-shot CLIP and the fine-tuned CLIP. In each iteration, we randomly sample a batch of textual outliers from the constructed set

$$\mathcal{L}_{\text{dor}} = 1 - \frac{1}{B}\sum_{b=1}^{B} \mathrm{sim}\left(\psi\left(\boldsymbol{t}'_{o_b}\right), \psi\left(\boldsymbol{t}_{o_b}\right)\right),$$

$$\mathcal{L}_{total} = \mathcal{L}_{\text{ce}} + \lambda \cdot \mathcal{L}_{\text{dor}}$$

Experiment

Calibration Results across 11 Few-shot Benchmarks

- DOR enhances the calibration of existing prompt-tuning methods.

Metric	ZSCLIP	CoOp		CoCoOp		MaPLe		DEPT	
	Vanilla	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR
Base	3.58	3.07	**2.67**	3.60	4.22	2.75	2.83	6.04	7.67
New	4.61	14.58	**6.49**	6.14	**4.02**	5.46	**4.44**	14.58	**7.50**
HM	4.10	8.82	**4.58**	4.87	**4.12**	4.11	**3.63**	10.31	**7.58**

Metric	KgCoOp		TCP		PromptSRC		CoPrompt		PromptKD	
	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR
Base	5.82	6.07	4.71	4.79	3.7527	3.88	2.56	2.96	4.73	4.81
New	4.48	**3.99**	4.07	**3.80**	4.15	**3.80**	5.96	**4.69**	4.38	**3.66**
HM	5.15	**5.03**	4.39	**4.29**	3.95	**3.84**	4.26	**3.83**	4.56	**4.24**

Experiment

Ablations

- DOR effectively breaks the calibration trade-off

Method	Variant	Base	New	HM
CoOp	Vanilla	3.07	14.49	8.78
	+KG	5.82	4.48	5.15
	+DOR	2.47	6.48	**4.47**
MaPLe	Vanilla	2.75	5.46	4.11
	+KG	4.01	4.29	4.15
	+DOR	3.06	4.26	**3.66**
CoPrompt	Vanilla	2.60	5.96	4.28
	+KG	4.01	4.99	4.50
	+DOR	2.98	5.14	**4.06**

- DOR do not rely on the potential overlap with new classes

Method	Variant	Base	New	HM
CoOp	Vanilla	3.07	14.49	8.78
	+DOR	2.67	6.49	4.58
	+DOR†	2.82	6.77	4.80
MaPLe	Vanilla	2.75	5.46	4.11
	+DOR	2.83	4.44	3.64
	+DOR†	2.89	4.51	3.70
CoPrompt	Vanilla	2.60	5.96	4.28
	+DOR	2.96	4.69	3.83
	+DOR†	2.71	4.93	3.82

Base-to-new generalization

- Similar to the evaluation of calibration, a salient observation is that our proposed DOR drastically improves base-to-new generalization, with its accuracy consistently outperforming all existing baselines in the harmonic mean of base and new classes.

	ZSCLIP	CoOp		CoCoOp		MaPLe		KgCoOp		DEPT		CoPrompt		PromptKD	
Class	Vanilla	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR	Vanilla	+DOR
Base	69.49	82.97	**83.20**	80.57	79.89	82.11	82.08	82.29	82.13	83.70	**83.81**	82.32	82.39	85.74	85.52
New	74.32	61.74	**72.01**	72.47	**74.59**	73.89	**75.89**	72.21	**73.14**	65.04	**71.39**	73.29	**74.50**	79.80	**80.81**
HM	71.90	72.36	**77.61**	76.52	**77.24**	78.00	**78.98**	77.25	**77.64**	74.37	**77.60**	77.81	**78.44**	82.77	**83.17**

Covariate shift generalization

- DOR demonstrates superior accuracy in domain generalization and successfully maintains in-distribution performance

	ECE (↓)						Accuracy (↑)					
	Source	Target					Source	Target				
	ImageNet	-V2	-S	-A	-R	AVG	ImageNet	-V2	-S	-A	-R	AVG
CLIP	1.86	2.44	4.88	8.34	3.51	4.79	66.73	60.87	46.09	47.81	73.98	57.19
CoOp	1.10	4.19	8.40	15.34	0.80	7.18	71.44	63.55	45.76	47.81	73.74	57.72
+DOR(ours)	1.64	**1.95**	**4.97**	**11.07**	1.58	**4.89**	71.47	**64.47**	**48.28**	**50.12**	**76.05**	**59.73**
MaPLe	1.13	2.56	4.88	12.42	1.06	5.23	72.05	64.57	48.78	47.66	76.61	59.41
+DOR(ours)	1.46	**1.89**	**3.96**	**11.08**	1.37	**4.58**	71.93	**64.94**	48.77	**48.29**	76.20	**59.55**

Additional Visualization

- DOR modifies the logit distribution and confidence level for better calibration.