南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Robust Test-Time Adaptation for Zero-Shot Prompt Tuning

Ding-Chu Zhang*, Zhi Zhou*, Yu-Feng Li†

National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China
{zhangdc,zhouz,liyf}@lamda.nju.edu.cn

AAAI 2024

Background

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

1、DNN

Problem： the remarkable performance of DNNs heavily relies on supervised training with large annotated datasets, which can be a significant burden in terms of labor and computational costs.

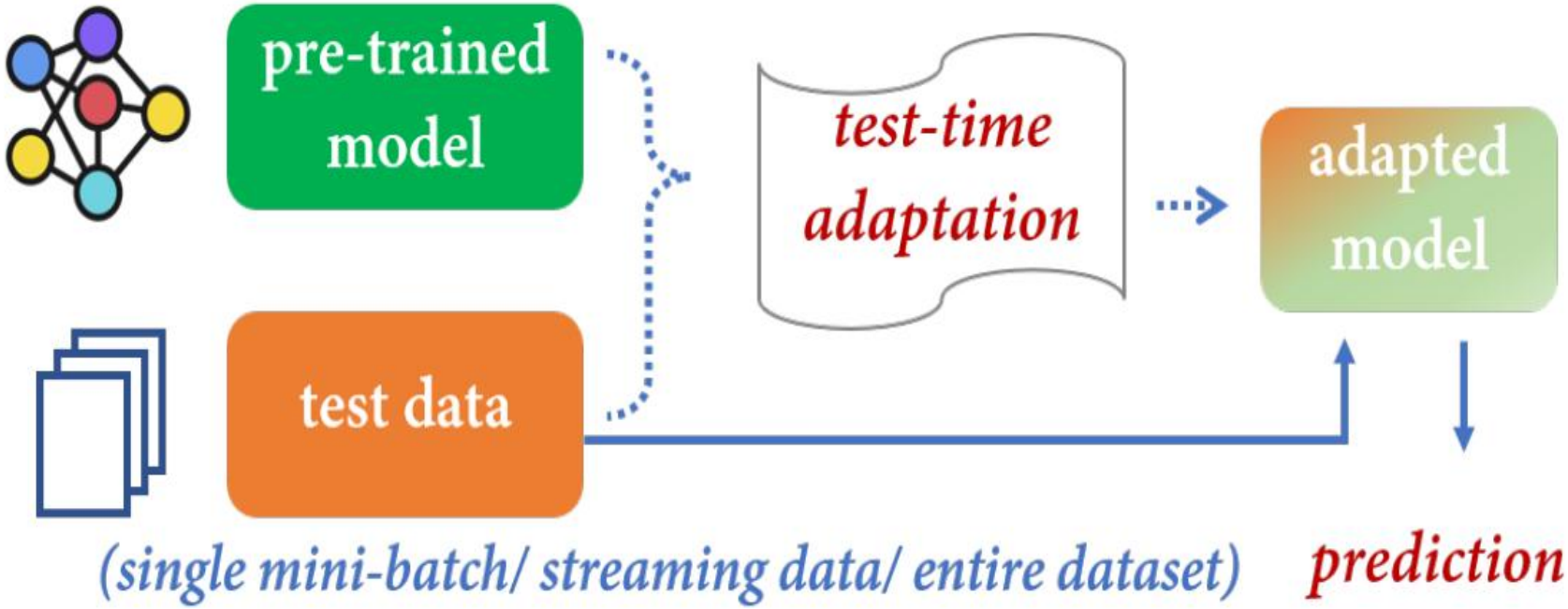
2、UDA

Unsupervised Domain Adaptation (UDA) is proposed to transfer knowledge from a labeled source domain to an unlabeled target domain without compromising accuracy.

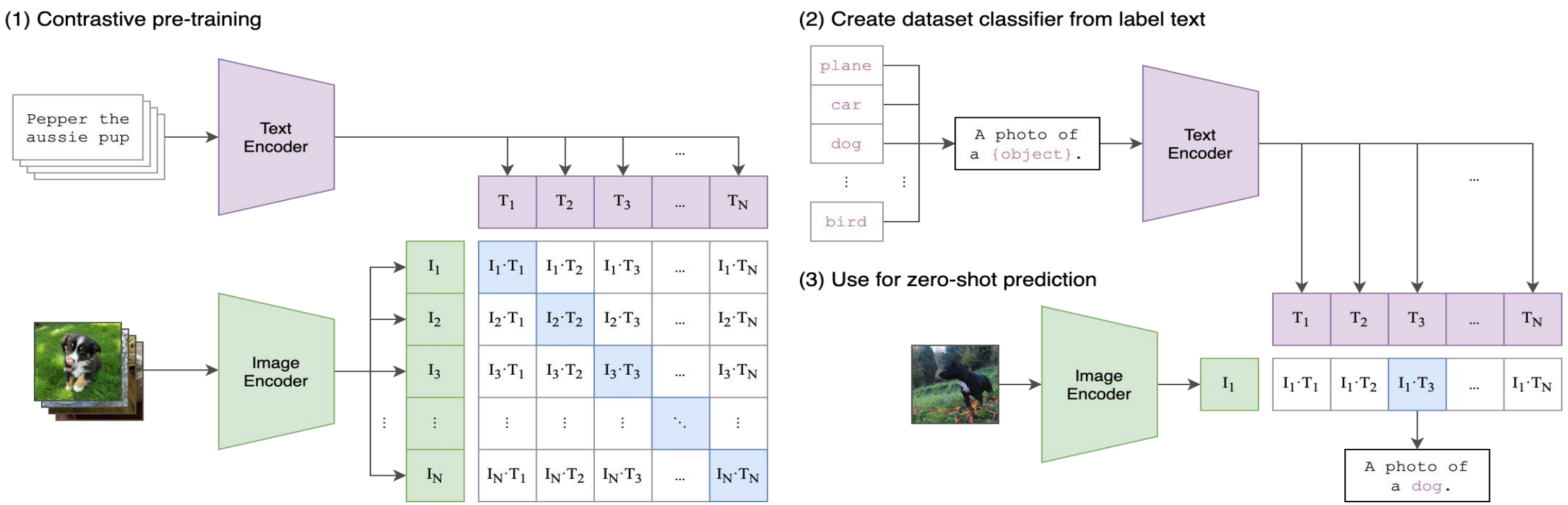
Problem： While this approach holds promise, most UDA methods require exposure to labeled source data during training, which can potentially compromise personal privacy and company security.

3、TTA

Test-Time Domain Adaptation（TTA） draws inspiration from test-time training and updates the model at the inference time to ensure real-time performance.



Motivation



Data Bias causes a problem that the performance of different prompts can vary across datasets, resulting in the difficulty of selecting an optimal prompt for downstream tasks.

Model Bias causes prediction biases towards specific classes, leading to error accumulation. And the errors accumulated by Model Bias will finally result in performance degradation problem.

Methods

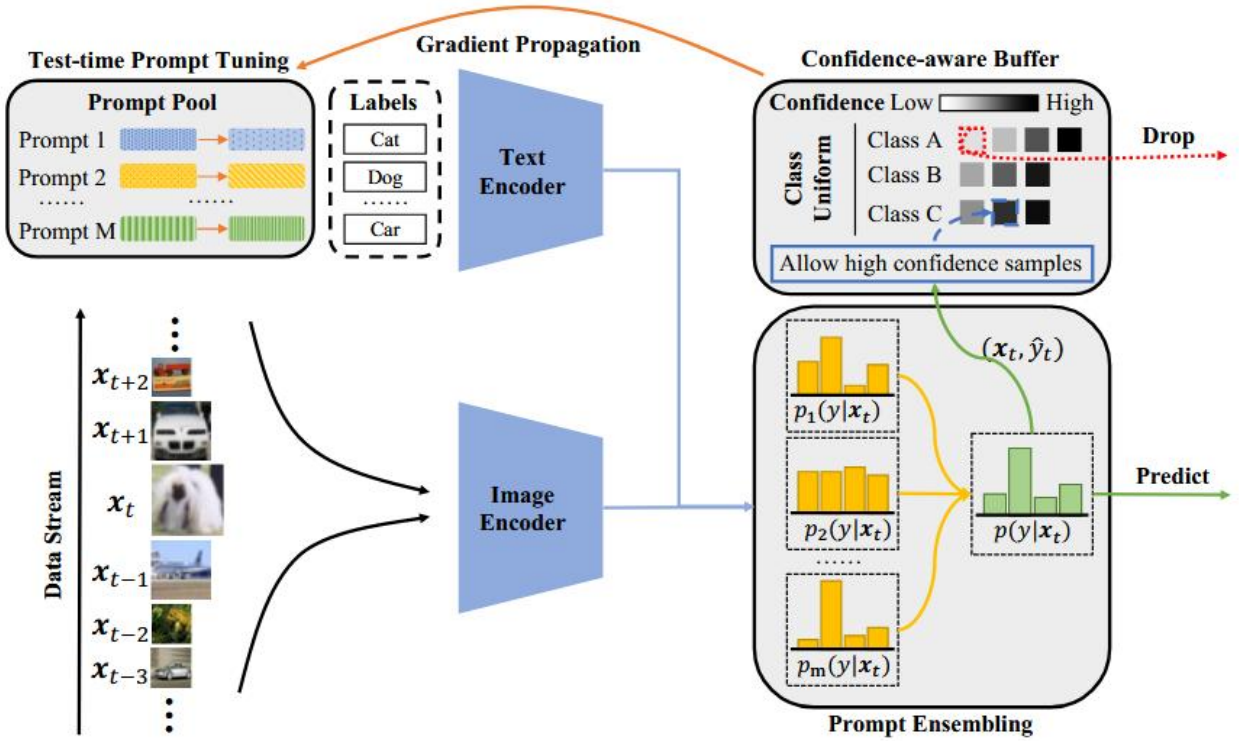


Figure 3: ADAPROMPT for image classification. We select confident samples from unlabeled test data stream by ensembling multiple prompts and push them into the confidence-aware buffer, which is used to store confident and balanced samples. Then, ADAPROMPT extracts samples from the buffer and adaptively updates all the prompts, which adapts prompts to current data.

Data Bias

to alleviate the negative effects of *Data Bias* and avoid the worst-case results. Let M denote the number of prompts we use. We obtain the ensembled probability of different prompts in the following formulation:

$$\hat{f}(y|\mathbf{x}_t;\mathbf{p}) = \frac{1}{M}\sum_{i=1}^{M} f(y|\mathbf{x}_t;\mathbf{p}^i) \tag{3}$$

Specifically, we use a common small number of hand-crafted prompts, such as "an image of a", "a colorful image of a" and "a noisy picture of a". And we obtain pseudo label and confidence for each sample in following formulation:

$$\begin{aligned} \hat{y}(\mathbf{x}_t) &= argmax_k \hat{f}(y_k|\mathbf{x}_t;\mathbf{p}) \\ c(\mathbf{x}_t) &= max_k \hat{f}(y_k|\mathbf{x}_t;\mathbf{p}) \end{aligned} \tag{4}$$

$$L(\mathbf{x}_t) = -\sum_{k=1}^{K} \hat{y}_k(\mathbf{x}_t) log\ \hat{f}(y_k|\mathbf{x}_t;\mathbf{p}) \tag{5}$$

where K represents the number of classes and $\hat{y}$ represents the pseudo label obtained by Eq. (4). The purpose of minimizing cross-entropy loss is to make the model more confident in the predicted samples, which can adapt prompts to *Data Bias* and improve the accuracy of predictions.

Model Bias

- 当 updateCriterion 为 random 时：
  - 如果要移除的实例所属类别不在存储数据量最大的类别索引列表（largest_indices）中，那么从最大的类别里随机选择一个（通过 random.choice），再在选中的最大类别对应的数据列表里随机选择一个索引（通过 random.randrange）作为要移除的目标索引 tgt_idx，然后从该类别对应的各个维度数据列表（特征、类别、其他信息）中移除对应索引的数据。
  - 如果要移除的实例所属类别就在最大类别索引列表中，会根据当前类别已存储的数据量 m_c、该类别总的实例计数 n_c 以及一个随机数 u 来判断是否移除，若 u <= m_c / n_c，则同样随机选择一个索引 tgt_idx 并移除对应的数据，否则返回 False，表示不移除。
- 当 updateCriterion 为 FIFO 时：
  - 如果要移除的实例所属类别不在最大类别索引列表中，直接选择最大类别对应数据列表中的第一个索引（即按照先进先出原则）作为 tgt_idx，然后移除对应的数据。
  - 如果要移除的实例所属类别在最大类别索引列表中，同样选择该类别对应数据列表中的第一个索引进行移除操作。
- 当 updateCriterion 为 Threshold 时：
  - 如果要移除的实例所属类别不在最大类别索引列表中，选择最大类别对应存储的置信度列表中最小值对应的索引（即选择置信度最小的实例对应的索引）作为 tgt_idx，然后移除对应的数据。
  - 如果要移除的实例所属类别在最大类别索引列表中，先获取该类别对应置信度列表中的最小值 minThreshold，如果传入的实例置信度小于这个最小值，则返回 False，不移除；否则选择最小值对应的索引 tgt_idx 并移除对应的数据。

Experiments

Dataset		CIFAR10-C(s=3)			CIFAR10-C(s=5)			CIFAR100-C(s=3)			CIFAR100-C(s=5)		
Methods		Source	TPT	Ours	Source	TPT	Ours	Source	TPT	Ours	Source	TPT	Ours
Noise	Gauss.	50.03	52.86	**54.50**	38.00	40.08	**42.48**	27.81	25.54	**28.61**	19.60	17.31	**21.92**
	Shot	61.74	63.32	**64.92**	43.14	44.74	**47.89**	33.81	32.22	**35.30**	21.36	19.04	**23.95**
	Impul.	78.59	78.87	**81.36**	56.70	59.08	**60.59**	47.30	47.63	**50.51**	25.31	25.65	**30.06**
Blur	Defoc.	85.46	85.25	**87.69**	72.88	72.10	**74.98**	60.10	**60.55**	60.54	42.52	42.73	**43.07**
	Glass	54.26	53.95	**59.29**	42.59	43.19	**47.51**	29.35	29.21	**30.38**	20.06	19.97	**20.91**
	Motion	77.15	77.06	**78.52**	70.96	70.14	**72.54**	48.69	48.86	**49.69**	**43.15**	42.63	42.46
	Zoom	81.57	81.35	**84.29**	74.66	74.89	**78.30**	56.08	55.96	**57.22**	47.89	48.12	**48.72**
Weather	Snow.	81.01	81.18	**84.52**	74.74	75.32	**78.26**	53.90	55.41	**56.34**	48.35	**49.19**	48.95
	Frost	81.13	81.02	**84.60**	78.40	78.33	**80.19**	53.12	53.89	**55.05**	49.72	50.43	**50.89**
	Fog	86.60	86.49	**89.10**	71.66	72.54	**73.14**	60.77	**61.64**	61.33	41.64	**42.71**	42.45
	Brit.	88.92	88.67	**91.53**	85.00	85.12	**88.06**	64.88	65.39	**66.64**	57.02	57.58	**59.07**
Digital	Contr.	87.11	87.70	**89.28**	63.00	**70.80**	67.95	59.77	61.18	**61.58**	34.54	**38.06**	36.84
	Elastic	80.27	80.75	**83.46**	55.40	57.10	**58.88**	52.53	53.43	**55.01**	29.21	30.05	**30.56**
	Pixel	75.18	75.98	**81.54**	48.09	52.24	**57.21**	51.09	51.94	**53.29**	23.94	25.15	**27.50**
	JPEG	69.51	69.82	**72.67**	60.30	61.55	**63.83**	39.68	40.17	**42.40**	32.46	32.43	**34.29**
Avg.		75.90	76.29	**79.15**	62.37	63.81	**66.12**	49.26	49.54	**50.93**	35.78	36.07	**37.44**

Table 1: Comparison with state-of-the-art test-time prompt tuning methods on CIFAR10-C and CIFAR100-C benchmarks with corruption level 3 and 5. We conduct separate tests on 15 different domains for each benchmark. We omit std in this table due to space issues. The best results are indicated in bold. Our method outperforms comparison methods in almost all cases. The best performance is in bold.

Experiments

Methods		Source	TPT	TPT-C	Ours
Noise	Gauss.	15.72	16.29	0.52	**17.52**
	Shot	23.44	23.86	0.52	**26.47**
	Impul.	17.47	17.58	0.52	**20.76**
Blur	Defoc.	32.43	32.65	0.58	**34.39**
	Glass	11.88	12.51	0.52	**14.45**
	Motion	31.97	32.31	0.54	**33.98**
	Zoom	30.99	31.57	0.54	**33.32**
Weather	Snow.	29.69	30.90	0.55	**32.82**
	Frost	32.98	33.25	0.58	**36.30**
	Fog	35.81	36.36	0.58	**37.97**
	Brit.	43.95	43.62	0.60	**46.80**
Digital	Contr.	22.56	23.00	0.52	**25.52**
	Elastic	38.14	38.74	0.58	**40.78**
	Pixel	26.38	27.72	0.55	**29.42**
	JPEG	37.54	37.56	0.64	**40.72**
Avg.		28.73	29.20	0.55	**31.42**

Table 3: Comparison with SOTA test-time prompt tuning methods on TinyImageNet-C with corruption level 3. ADAPROMPT outperforms them in all domains.

Experiments

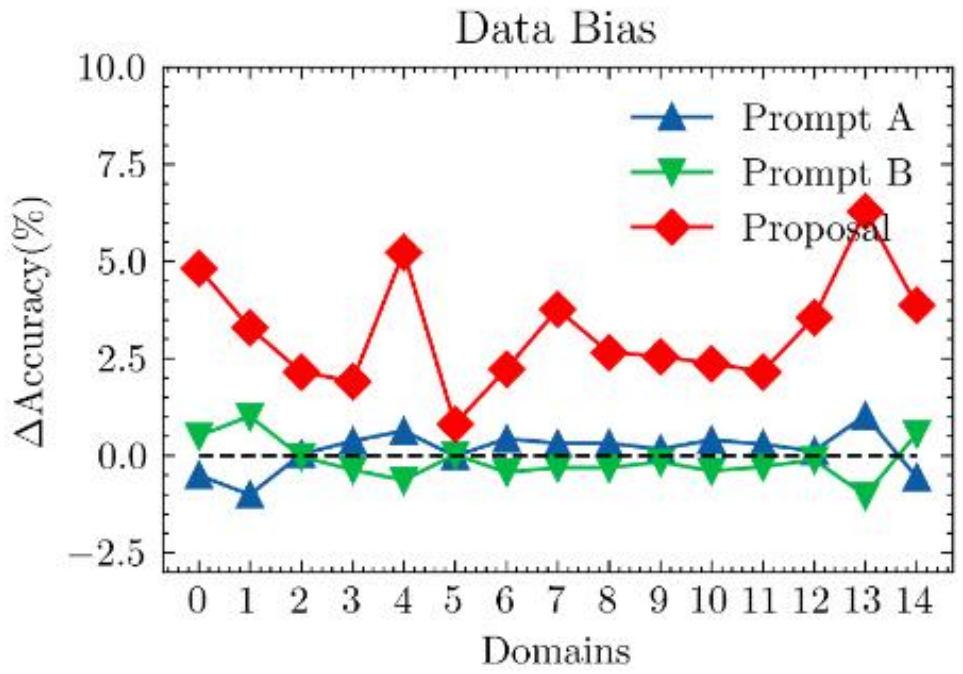


Figure 1: Relative performance compared to the average performance of prompts evaluated on CIFAR10-C in 15 domains with corruption level 3 using different initial prompts.

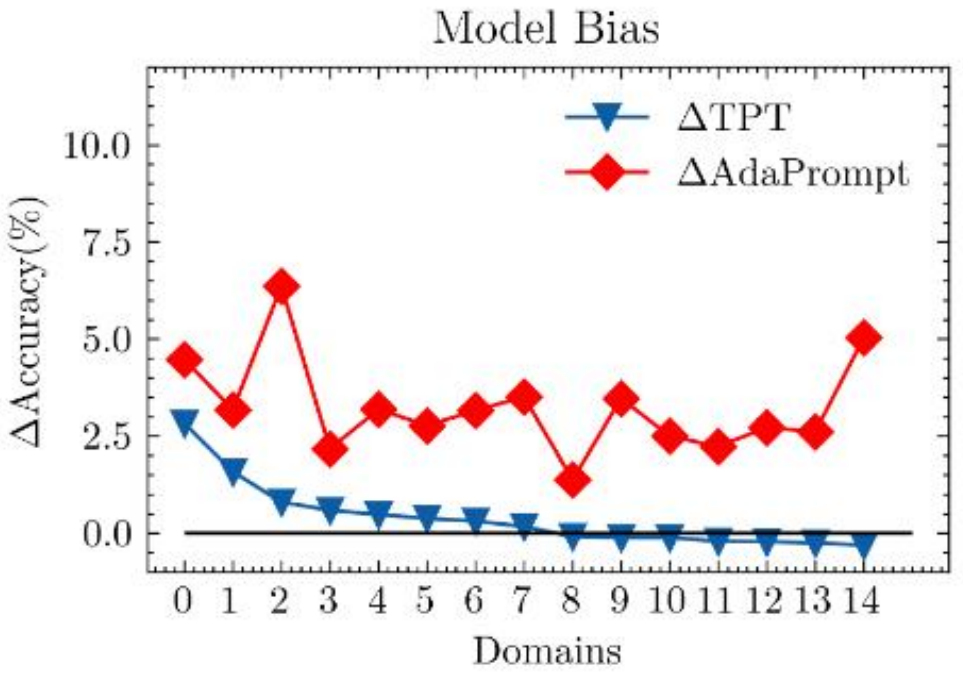


Figure 2: Relative performance compared to baseline evaluated on CIFAR10-C in 15 different domains with corruption level 3. The black line represents the baseline.

Ablation Experiments

Method	CIFAR10-C(s=3)	CIFAR10-C(s=5)
P_A	75.91 ± 0.00	62.37 ± 0.00
P_B	76.21 ± 0.00	62.77 ± 0.00
P_C	72.98 ± 0.00	59.25 ± 0.00
P_{best} + UP.	77.72 ± 0.24	65.32 ± 0.18
P_e	75.38 ± 0.00	61.75 ± 0.00
P_e + UP.	**79.15 ± 0.23**	**66.12 ± 0.43**

Table 2: Evaluation of each module on CIFAR10-C with corruption level 3 and 5. The average accuracy of different modules on 15 different domains is shown.

Component M_e	Component M_u	CIFAR10-C(s=3)	CIFAR10-C(s=5)
		76.21 ± 0.00	62.37 ± 0.00
✓		75.38 ± 0.00	61.75 ± 0.00
	✓	77.72 ± 0.24	65.32 ± 0.18
✓	✓	**79.15 ± 0.23**	**66.12 ± 0.43**

Table 4: Ablation study of AdaPrompt on CIFAR10-C dataset with corruption level 3 and 5. The average accuracy on 15 different domains is reported.

Acc(%)	Source	TPT	Ours
RN50	47.70 ±0.00	51.44 ±0.02	**55.44 ± 0.30**
ViT-B/32	71.30 ±0.00	73.77 ±0.03	**75.81 ± 0.33**

Table 5: Average accuracy of CIFAR10-C in different 15 domains with corruption level 3 on different backbones.

Ablation Experiments

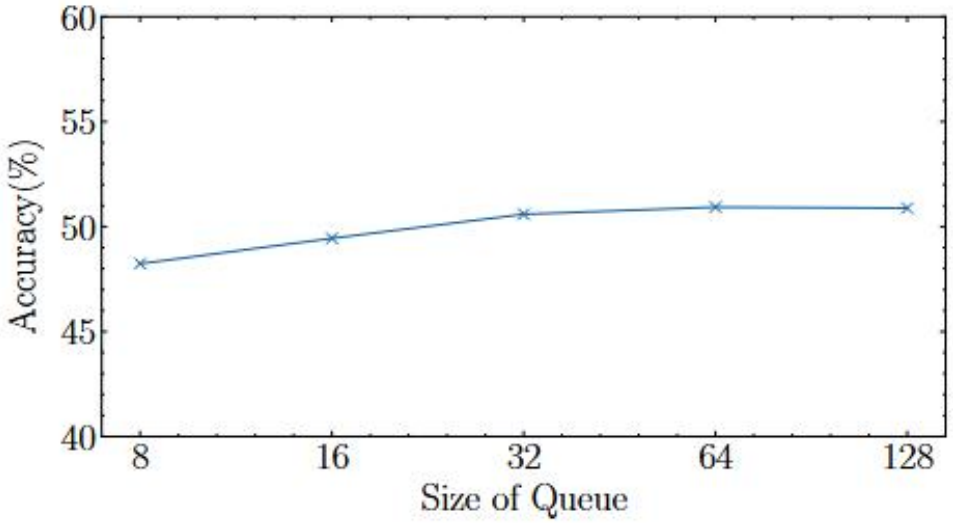


Figure 5: Average accuracy of ADAPROMPT with different buffer size on CIFAR100-C with corruption level 3.

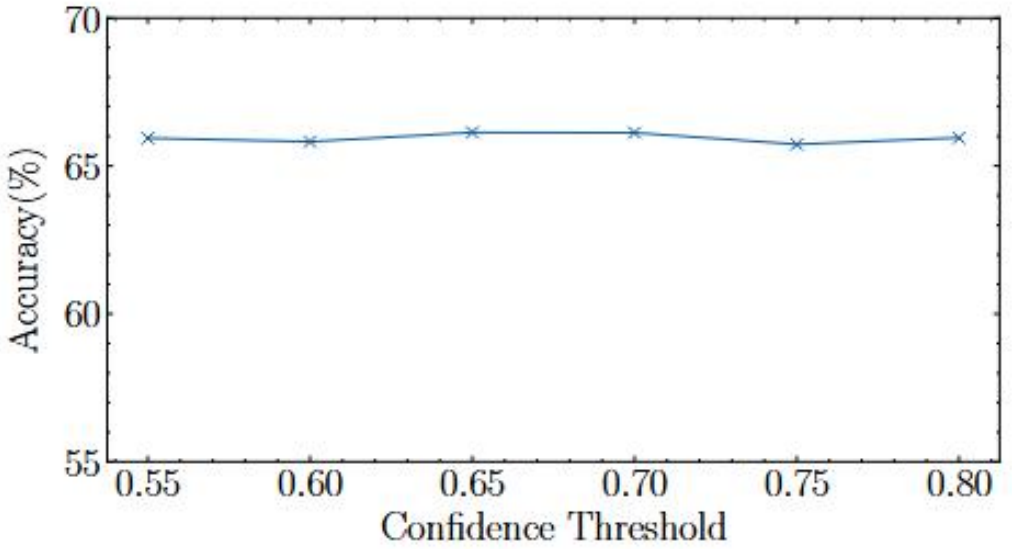


Figure 6: Performance of ADAPROMPT with different confidence threshold on CIFAR100-C with corruption level 3.

Thanks