ODS: Test-Time Adaptation in the Presence of Open-World Data Shift

Zhi Zhou, Lan-Zhe Guo, Lin-Han Jia, Ding-Chu Zhang, Yu-Feng Li

ICML International Conference On Machine Learning

LAMDA Learning And Mining from DatA

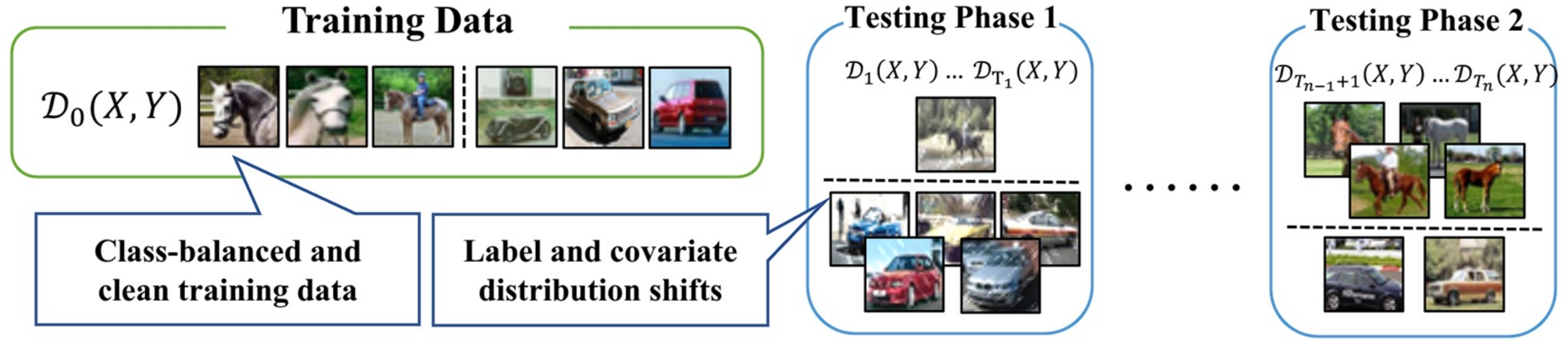
NANJING UNIVERSITY 1902

Brief Introduction

Test-time adaptation (TTA) adapts a source model to the distribution shift in testing data without using any source data. However, current test-time adaptation account for relatively simple distribution shift, such as covariate shift, which challenges in the following two aspects:

- **TTA degenerates when label and covariate distribution shifts mix**
- **TTA cannot adapt to changed label distribution shift**

These two points are very crucial for deploying test-time adaptation in the real world.



- ✓ In our work, we study **an Open-World Data Shift setting for test-time adaptation** and where the model needs to adapt to both covariate and label distribution shifts.
- ✓ We propose **a test-time adaptation framework ODS** to solve the above open-world data shift setting, which can **apply to many existing test-time algorithms**.
- ✓ Our proposal is clearly **better than one baseline and six test-time adaptation methods** evaluated on two benchmark datasets.

ODS Method

The ODS framework contains two modules:

- **Distribution Tracker $\mathcal{M}_T$**: Estimating label distribution $\boldsymbol{w}_t$ for subsequent adaptation and predictive optimization;
- **Prediction Optimizer $\mathcal{M}_O$**: Improving the prediction using $\boldsymbol{w}_t$.

The objective of the ODS framework contains two parts: **A weighting term:** Applying the estimated label distribution $\boldsymbol{w}_t$ to loss; **An entropy minimization term**: Adapting the model unsupervised.

$$\min_{\theta_t} -\frac{1}{N_t}\sum_{i=1}^{N_t}\sum_{k=1}^{K} S(\boldsymbol{w}_t)_k f_{\theta_t}(Y=k|\boldsymbol{x}_i)\log f_{\theta_t}(Y=k|\boldsymbol{x}_i)$$

s.t. $\boldsymbol{w}_t$ is estimated by $\mathcal{M}_T$

Weighted updating for test-time adaptation to maintain an internal class-balanced model

Entropy minimization loss to adapt the model in an unsupervised fashion

Distribution Tracker $\mathcal{M}_T$

Black box shift estimation (BBSE) can estimate the test label distribution shift $\mathcal{D}_t(Y)/\mathcal{D}_0(Y)$. However, its assumption $\mathcal{D}_0(X|Y)=\mathcal{D}_t(X|Y)$ which not holds.

$$\min_{\boldsymbol{w}_t}\sum_{i=1}^{N_t}\left[z_i^\top \log f_{\theta_0}(Y|\boldsymbol{x}_i) + z_i^\top \log z_i - \sum_{j=1}^{N_t} s_{ij} z_i^\top z_j\right]$$

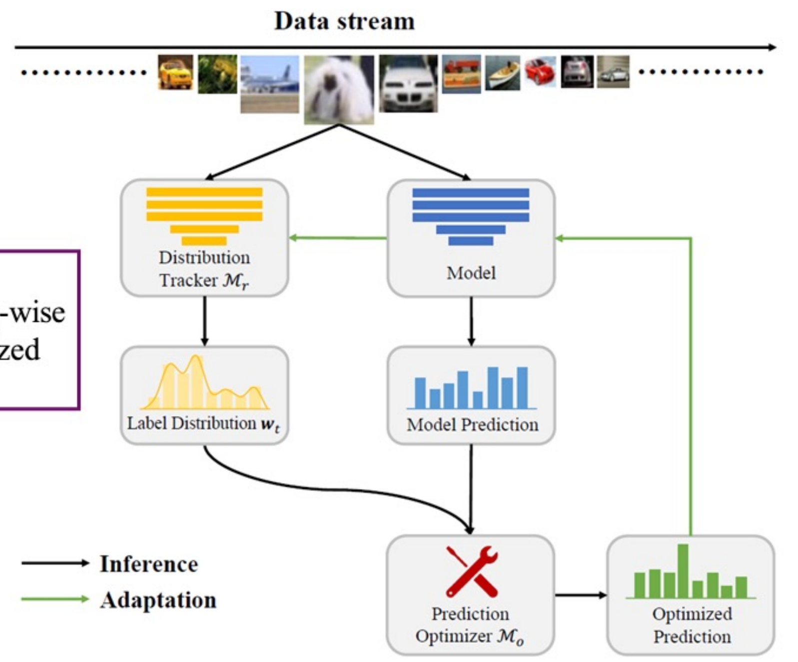
s.t. $\boldsymbol{w}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} z_i$

We use the adapted feature representation to calculate instance-wise similarity for help under generalized label shift assumptions

Prediction Optimizer $\mathcal{M}_O$

$$\hat{Y}_o = \arg\max_{y\in\mathcal{Y}} f_{\theta_t}(Y=y|X) + \ln w_{t,y} \qquad \hat{Y}_o = \arg\max_{k\in\mathcal{Y}} \frac{\sqrt{z_{i,k} f_{\theta_t}(Y=k\mid\boldsymbol{x}_i)}}{\sum_{k'\in\mathcal{Y}}\sqrt{z_{i,k'} f_{\theta_t}(Y=k'\mid\boldsymbol{x}_i)}}$$

▲ Statistics Optimization ▲ Distribution Optimization



Experiments

RQ1: Whether ODS can outperform prior TTA methods when encountering open-world data shift?

Methods	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
Source	14.70	18.52	15.61	56.92	31.99	68.01	63.25	82.19	72.44	76.31	**92.41**	23.38	72.33	68.72	79.72	55.77
Bn Stats	50.60	51.16	45.31	71.73	47.99	69.35	68.59	60.16	60.39	64.27	69.60	67.56	59.21	66.12	58.17	60.68
Tent	53.53	60.97	59.34	63.33	47.12	65.81	68.11	55.08	55.00	58.68	63.40	49.59	46.95	50.45	45.38	56.18
Eata	48.94	48.21	42.05	65.44	43.42	59.81	57.27	55.09	52.98	56.00	59.54	61.47	51.32	55.75	50.88	53.88
Lame	57.99	60.15	53.07	78.83	53.04	76.67	74.90	67.81	67.30	71.94	77.05	74.84	68.53	73.44	66.90	68.16
CoTTA	57.43	60.06	56.03	66.66	52.25	66.54	66.65	58.32	58.92	60.09	64.69	55.05	59.37	64.74	61.92	60.58
Note	51.90	54.57	68.38	84.29	50.53	88.97	86.21	86.15	86.68	83.27	86.48	90.64	77.84	80.77	81.02	77.18
Ods	**67.45**	**65.78**	**71.88**	**88.66**	**56.32**	**90.48**	**88.09**	**86.16**	**86.93**	**83.96**	87.37	**91.16**	**79.35**	**84.43**	**82.02**	**80.67**

▲ Detailed Results on CIFAR10 dataset with $\gamma=10$

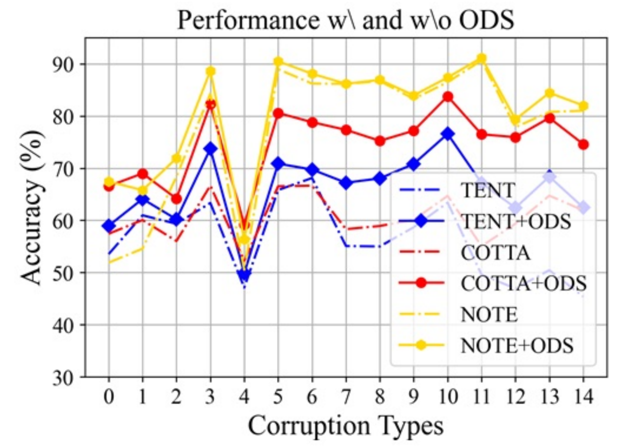
Methods	$\gamma=2$	$\gamma=5$	$\gamma=10$
Source	32.71 ± 0.15	32.71 ± 0.18	32.75 ± 0.14
Bn Stats	52.69 ± 0.20	52.82 ± 0.08	52.76 ± 0.15
Tent	40.07 ± 2.35	51.39 ± 0.59	52.95 ± 0.17
Eata	43.68 ± 18.16	45.12 ± 15.79	48.99 ± 7.79
Lame	52.49 ± 0.25	52.51 ± 0.24	52.62 ± 0.21
CoTTA	47.74 ± 0.59	50.48 ± 0.57	51.72 ± 0.47
Note	50.34 ± 0.11	48.41 ± 0.33	47.06 ± 0.35
Ods	$\mathbf{56.86}\pm\mathbf{0.18}$	$\mathbf{56.43}\pm\mathbf{0.21}$	$\mathbf{55.83}\pm\mathbf{0.23}$

▲ Average results on CIFAR100 dataset

Methods	$\gamma=2$	$\gamma=5$	$\gamma=10$
Source	56.41 ± 0.05	56.12 ± 0.07	55.77 ± 0.16
Bn Stats	78.33 ± 0.05	71.75 ± 0.08	60.68 ± 0.14
Tent	68.85 ± 3.14	66.94 ± 3.52	56.18 ± 4.13
Eata	79.35 ± 0.16	69.23 ± 0.28	53.88 ± 0.53
Lame	78.96 ± 0.05	75.20 ± 0.10	68.16 ± 0.13
CoTTA	$\mathbf{81.81}\pm\mathbf{0.37}$	73.58 ± 0.28	60.58 ± 0.15
Note	78.81 ± 0.27	77.96 ± 0.75	77.18 ± 0.38
Ods	81.13 ± 0.09	$\mathbf{80.40}\pm\mathbf{0.36}$	$\mathbf{80.67}\pm\mathbf{0.29}$

▲ Average results on CIFAR10 dataset

RQ2: Whether ODS is generic to integrate with different TTA methods and boost their performance?

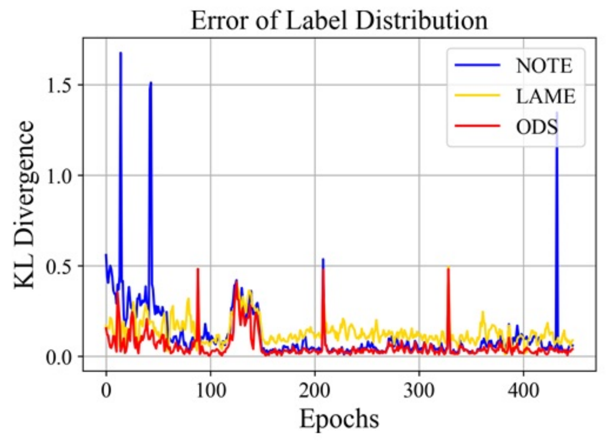


▲Detailed results on CIFAR10 dataset with $\gamma=10$

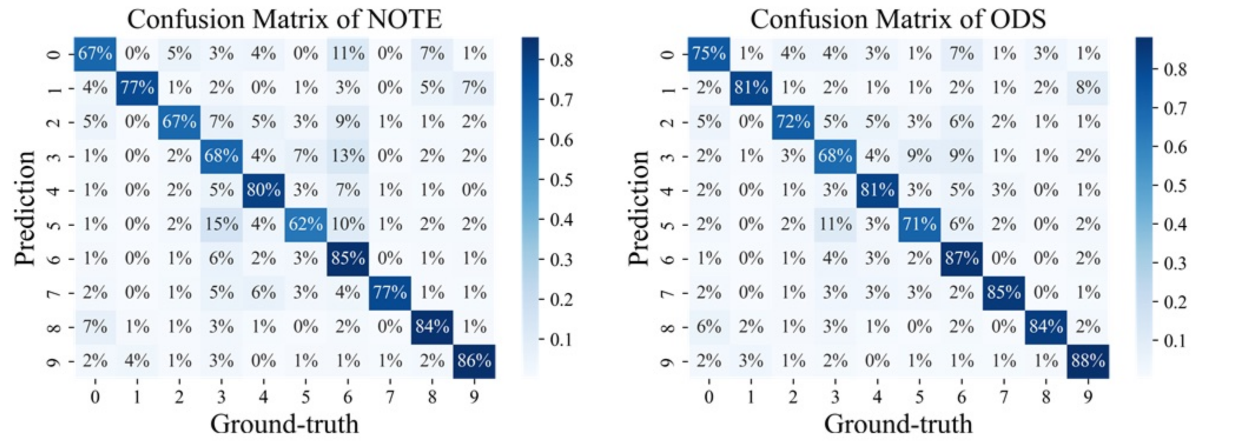
Methods	$\gamma=2$	$\gamma=5$	$\gamma=10$
Tent	68.85 ± 3.14	66.94 ± 3.52	56.18 ± 4.13
Tent w/ Ods	$\mathbf{69.00}\pm\mathbf{5.96}$	$\mathbf{73.56}\pm\mathbf{2.85}$	$\mathbf{66.03}\pm\mathbf{1.89}$
CoTTA	81.81 ± 0.37	73.58 ± 0.28	60.58 ± 0.15
CoTTA w/ Ods	$\mathbf{82.11}\pm\mathbf{0.25}$	$\mathbf{79.74}\pm\mathbf{0.32}$	$\mathbf{74.72}\pm\mathbf{0.64}$
Note	78.81 ± 0.27	77.96 ± 0.75	77.18 ± 0.38
Note w/ Ods	$\mathbf{81.13}\pm\mathbf{0.09}$	$\mathbf{80.40}\pm\mathbf{0.36}$	$\mathbf{80.67}\pm\mathbf{0.29}$

▲Average results on CIFAR10 dataset of three TTA methods with and without ODS framework

RQ3: Does ODS accurately estimate label distribution and effectively optimize the prediction?



▲KL divergence between ground-truth and estimations



▲Confusion Matrix of NOTE method and ODS framework. ODS can improve performance of each class.

Ablation Study

Modules $\mathcal{M}_T$	Modules $\mathcal{M}_O$	Tent	CoTTA	Note
		56.18 ± 4.13	60.58 ± 0.15	77.18 ± 0.38
✓		58.95 ± 2.36	60.65 ± 0.31	77.20 ± 0.57
✓	✓	$\mathbf{66.03}\pm\mathbf{1.89}$	$\mathbf{74.72}\pm\mathbf{0.64}$	$\mathbf{80.67}\pm\mathbf{0.29}$

▲Effectiveness of each module in ODS framework.

In-depth Comparison

	Note	Note+Lame	Ods
γ=2	78.81 ± 0.27	77.32 ± 0.17	$\mathbf{81.13}\pm\mathbf{0.09}$
γ=5	77.96 ± 0.75	76.76 ± 0.67	$\mathbf{80.40}\pm\mathbf{0.36}$
γ=10	77.18 ± 0.38	78.43 ± 0.77	$\mathbf{80.67}\pm\mathbf{0.29}$

▲In-depth Comparison between LAME, NOTE and ODS

- ✓ If you are interested in this paper, feel free to contact Zhi Zhou (zhouz@lamda.nju.edu.cn)
- ✓ Our code is released at: https://www.lamda.nju.edu.cn/code_ODS.ashx
- ✓ This research was supported by the National Key R&D Program of China (2022ZD0114803), National Science Foundation of China (62176118) and CAAI-Huawei Mind-Spore Open Fund.