Efficient Test-Time Model Adaptation without Forgetting

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao and Mingkui Tan

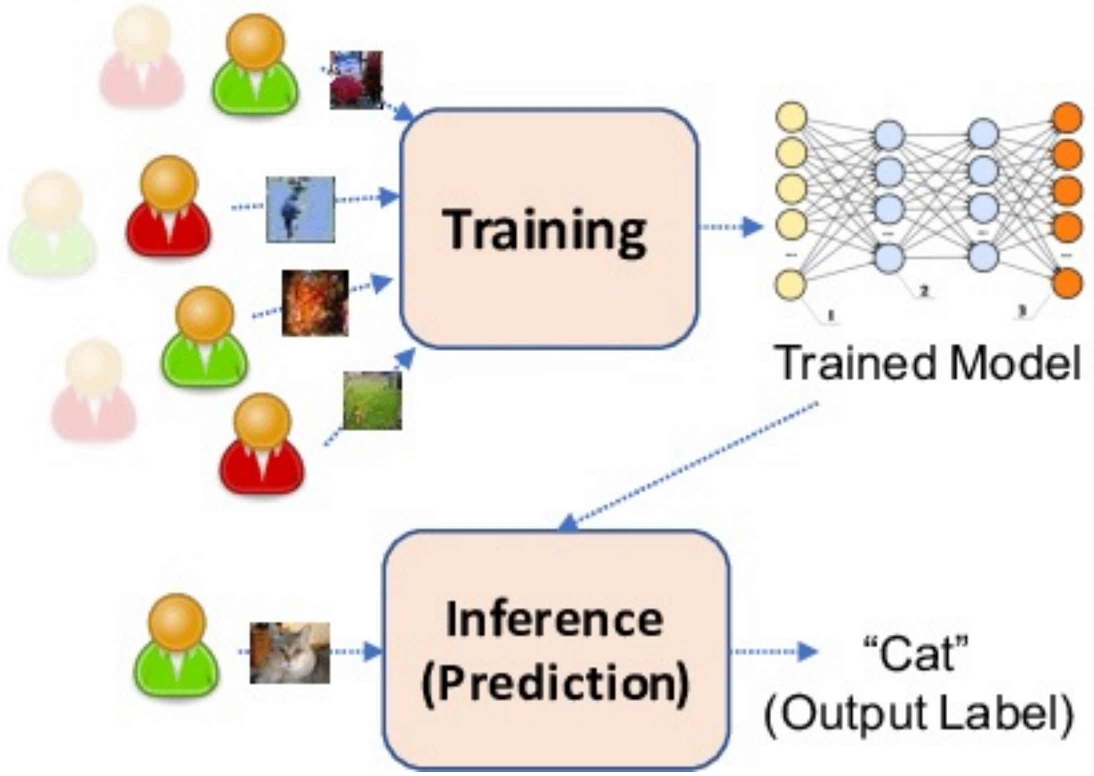
South China University of Technology, Tencent AI Lab, National University of Singapore

Contents

Contents

Background: Deep Learning Pipeline and Data Shifts

Distribution shift often exists between training and testing data!

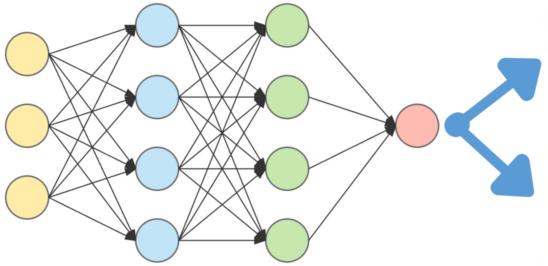


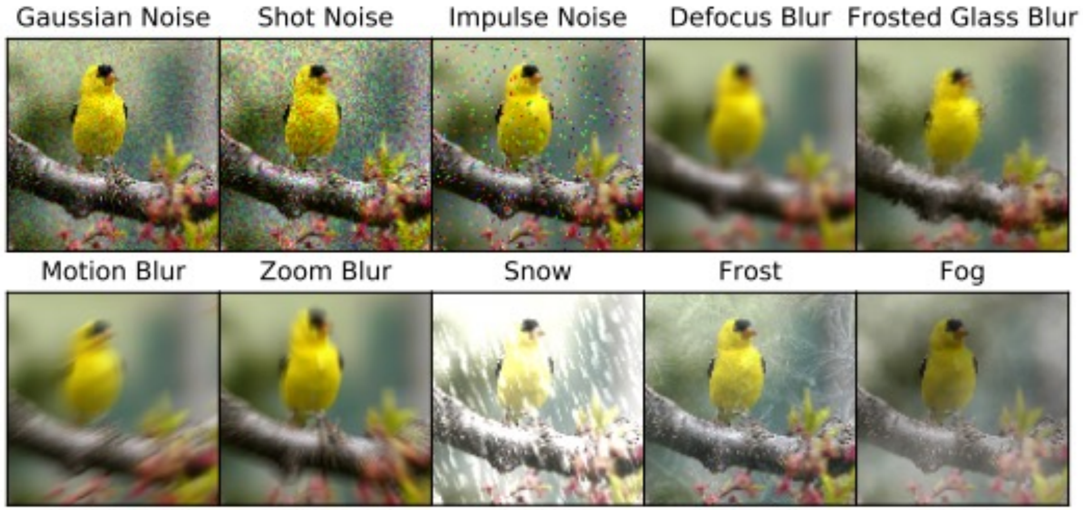
An overview of training and Inference in DL [1]

[1] Deep Learning on Private Data.

Background: Data Shifts

- Test samples may encounter **natural variations or corruptions** (*also called distribution shifts*), such as:
  - Changes in lighting resulting from **weather change**
  - **Unexpected noises** resulting from sensor degradation, etc.





ImageNet-C (Hendrycks & Dietterich, 2019)

Unfortunately, models are very sensitive to such shifts, and suffer from severe performance degradation!

Methods for Overcoming Data Shifts

- **Training-time generalization** seek to anticipate shifts at training phase:
  - Domain generalization
  - Data augmentation techniques

It is hard to anticipate all possible shifts!

- **Test-time adaptation methods** (will exploit testing data):

Setting	Source data	Target data	Training loss	Testing loss	Offline	Online
Fine-tuning	×	x^t, y^t	$\mathcal{L}(\mathrm{x}^t, y^t)$	--	√	×
UDA	x^s, y^s	x^t	$\mathcal{L}(\mathrm{x}^s, y^s) + \mathcal{L}(\mathrm{x}^s, \mathrm{x}^t)$	--	√	×
Test-time training	x^s, y^s	x^t	$\mathcal{L}(\mathrm{x}^s, y^s) + \mathcal{L}(\mathrm{x}^s)$	$\mathcal{L}(\mathrm{x}^t)$	×	√
Fully TTA	×	x^t	×	$\mathcal{L}(\mathrm{x}^t)$	×	√

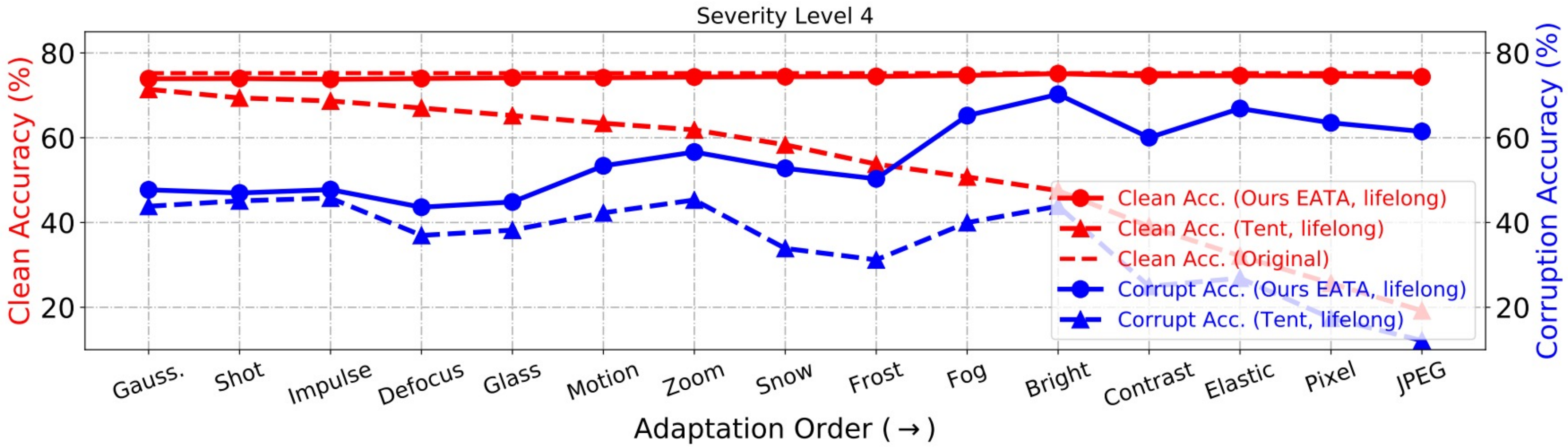
- In this work, we study the Fully test-time adaptation (TTA) setting
  - Does **not alter model training** process, **adapt online**, **use only $\mathbf{x}^t$**

Limitations of Prior Test-Time Adaptation Methods

- **Efficiency:** perform adaptation for all samples is expensive

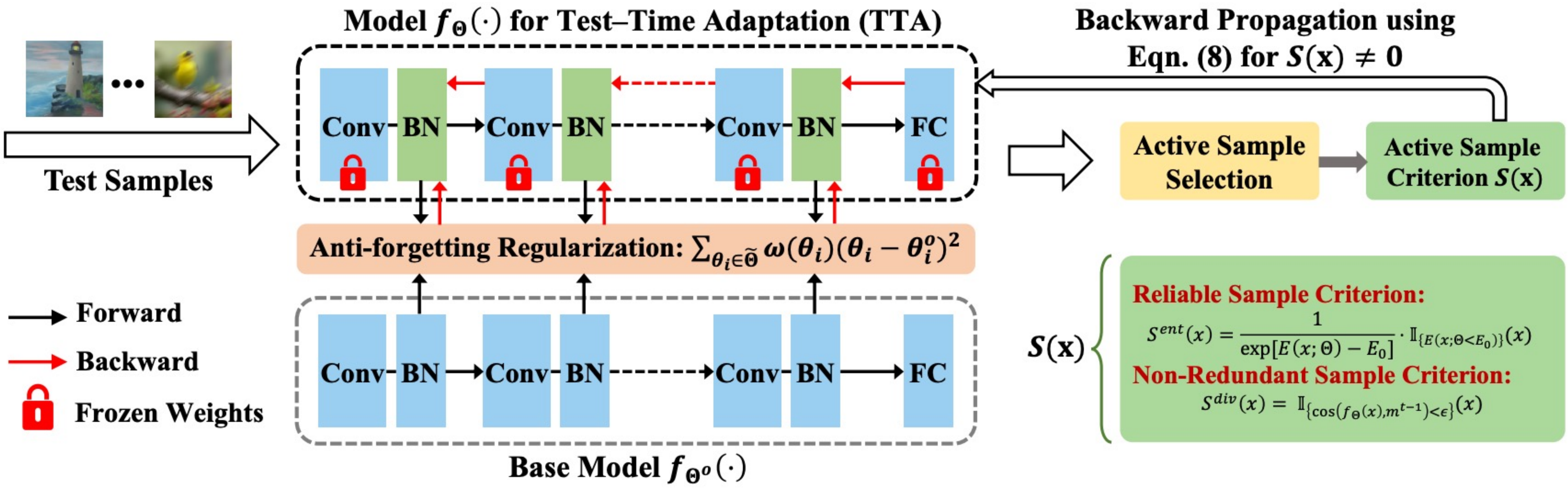
On ImageNet-C, Gauss. Level 5	# Forward	# Backward
Standard Inference	50,000	0
TTT (Sun et al., 2020)	50,000 × 65	50,000 × 64
Tent (Wang et al., 2021)	50,000	50,000
EATA (ours)	50,000	<20,000

- **Forgetting:** performance degradation on in-distribution test data after adaptation on out-of-distribution test data



Contents

EATA: Efficient Anti-forgetting Test-time Adaptation



- **Selective adaptation $S(\mathbf{x})$** to improve efficiency:
  - Active sample selection

$$\min_{\tilde{\Theta}} S(\mathbf{x})E(\mathbf{x};\Theta) + \beta\mathcal{R}(\tilde{\Theta}, \tilde{\Theta}^o)$$

- Weight **regularization $\mathcal{R}(\cdot)$** to prevent forgetting:
  - Fisher regularizer

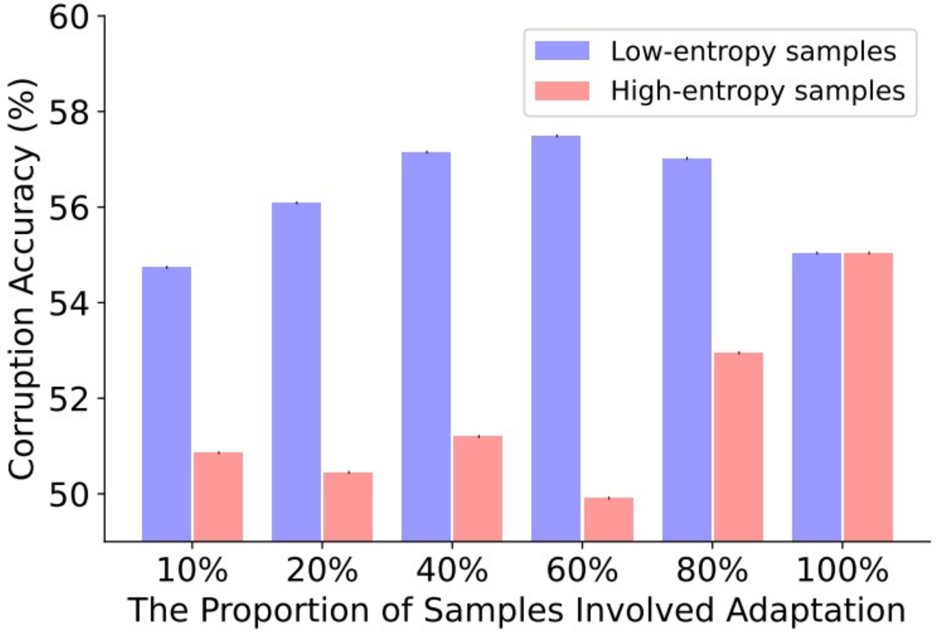
Active Sample Selection

- Samples for adaptation should be reliable:
  - Adaptation on low-entropy samples makes more contribution than high-entropy ones
  - Adaptation on test samples with very high entropy may hurt performance

$$S^{ent}(\mathbf{x}) = \frac{1}{\exp\left[E(\mathbf{x};\Theta) - E_0\right]} \cdot \mathbb{I}_{\{E(\mathbf{x};\Theta)<E_0\}}(\mathbf{x})$$

$E(\mathbf{x};\Theta)$ is the entropy of sample x and E_0 is a threshold

Effect of different samples in test-time entropy minimization (Tent)



Active Sample Selection

- Samples for adaptation should be non-redundant:
  - Adaptation with samples that produce similar gradients are unnecessary
  - Ensure the remaining samples have diverse model outputs/gradients

$$S^{div}(\mathbf{x}) = \mathbb{I}_{\{cos(f_\Theta(\mathbf{x}),\mathbf{m}^{t-1})<\epsilon\}}(\mathbf{x}), \quad \underbrace{\mathbf{m}^t = \begin{cases} \bar{\mathbf{y}}^1, & \text{if } t=1 \\ \alpha\bar{\mathbf{y}}^t + (1-\alpha)\mathbf{m}^{t-1}, & \text{if } t>1 \end{cases}}_{\text{Moving average of previous samples' outputs}}$$

- In sum,

$$S(\mathbf{x}) = S^{ent}(\mathbf{x}) \cdot S^{div}(\mathbf{x})$$

Anti-forgetting Weight Regularization

- Ensure (OOD) adapted model works well on ID and OOD data simultaneously
  - Prevent important parameters (for ID domain) from changing too much

$$\mathcal{R}(\tilde{\Theta}, \tilde{\Theta}^o) = \sum_{\theta_i \in \tilde{\Theta}} \omega(\theta_i)(\theta_i - \theta_i^o)^2$$

- θ_i^o is the original parameter
- $\tilde{\Theta}$ denote affine parameters of BN layers

  - $\omega(\theta_i)$ measures weight importance (using Fisher) through a small set of ID pseudo-labeled test samples $\mathcal{D}_F$

$$\omega(\theta_i) = \frac{1}{Q} \sum_{\mathbf{x}_q \in \mathcal{D}_F} \left(\frac{\partial}{\partial \theta_i^o} \mathcal{L}_{CE}(f_{\Theta^o}(\mathbf{x}_q), \hat{y}_q) \right)^2$$

Contents

Comparison w.r.t. OOD Performance and Efficiency

Results on ImageNet-C with severity level 5 regarding Corruption Error (%)

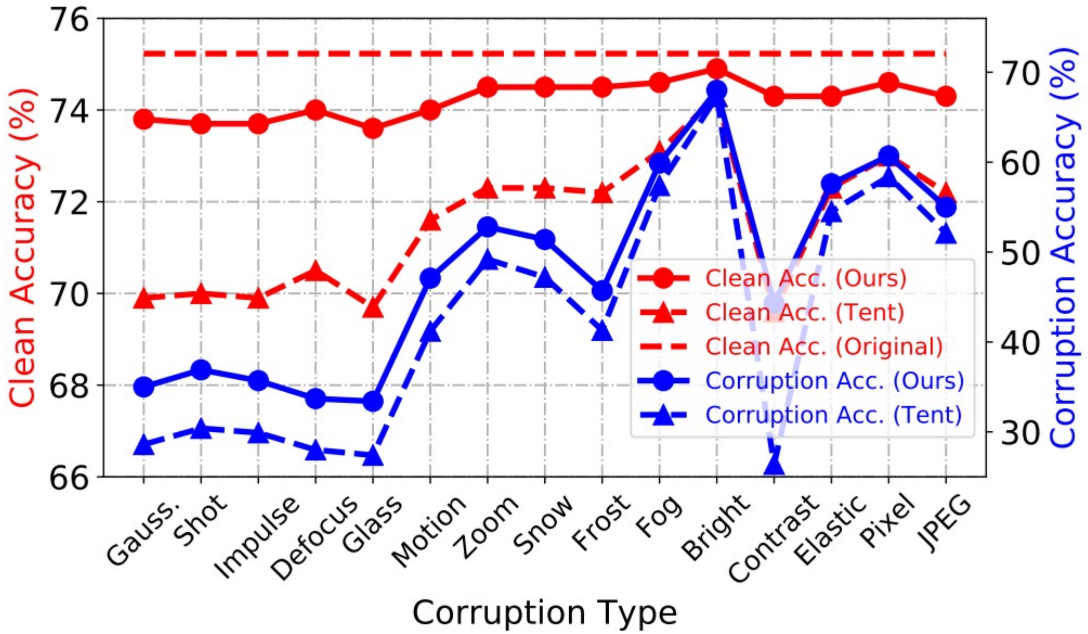
	Noise			Blur				Weather				Digital				Average	
Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	#Forwards	#Backwards
R-50 (GN)+JT	94.9	95.1	94.2	88.9	91.7	86.7	81.6	82.5	81.8	80.6	49.2	87.4	76.9	79.2	68.5	50,000	0
• TTT	69.0	66.4	66.6	71.9	92.2	66.8	63.2	59.1	81.0	49.0	38.2	61.1	50.6	48.3	52.0	50,000×21	50,000×20
R-50 (BN)	97.8	97.1	98.2	82.1	90.2	85.2	77.5	83.1	76.7	75.6	41.1	94.6	83.1	79.4	68.4	50,000	0
• TTA	95.9	95.1	95.5	87.5	91.8	87.1	74.2	86.0	80.9	78.7	47.0	87.6	85.4	75.4	66.4	50,000×64	0
• BN adaptation	84.5	83.9	83.7	80.0	80.0	71.5	60.0	65.2	65.0	51.5	34.1	75.9	54.2	49.3	58.9	50,000	0
• MEMO	92.5	91.3	91.0	80.3	87.0	79.3	72.4	74.7	71.2	67.9	39.0	89.0	76.2	67.0	62.5	50,000×65	50,000×64
• Tent	71.6	69.8	69.9	71.8	72.7	58.6	50.5	52.9	58.7	42.5	32.6	74.9	45.2	41.5	47.7	50,000	50,000
• Tent (episodic)	85.4	84.8	84.9	85.5	85.4	74.6	62.2	66.4	67.8	53.2	35.7	83.9	57.1	52.4	61.5	50,000×2	50,000
• ETA (ours)	**64.9**	62.1	63.4	**66.1**	67.1	**52.2**	47.4	**48.1**	**54.2**	**39.9**	32.1	**55.0**	**42.1**	**39.1**	45.1	50,000	26,031
• EATA (ours)	65.0	63.1	64.3	66.3	66.6	52.9	47.2	48.6	54.3	40.1	**32.0**	55.7	42.4	39.3	**45.0**	50,000	25,150
• EATA (lifelong)	65.0	**61.9**	**63.2**	66.2	**65.8**	52.7	**46.8**	48.9	54.4	40.3	**32.0**	55.8	42.8	39.6	45.3	50,000	28,243

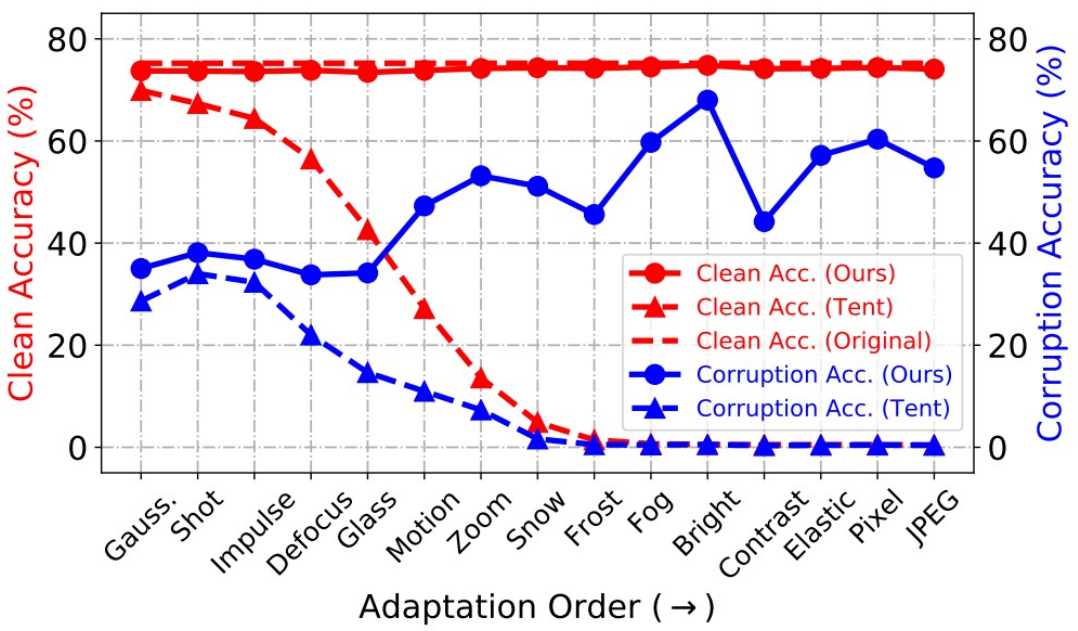
① Consistently outperform considered methods w.r.t. error

② Outperform Tent but with less #Backwards, leading to higher efficiency

③ Show the potential of fully test-time adaptation (consistently better than TTT)

Demonstration of Preventing Forgetting





Results on ImageNet-C level 5. Left: the model parameters are reset after each corruption type. Right: parameters will never be reset.

① EATA consistently outperforms Tent regarding the OOD accuracy and maintains the clean accuracy (while Tent fails)

② The forgetting issue of Tent is much more severe in lifelong scenario

Contents

Conclusion

- **Contributions**:
  - Propose an active sample identification scheme to filter out non-reliable and redundant test data from model adaptation
  - Extend the label-dependent Fisher regularizer to test samples with pseudo label generation, preventing drastic changes in important model weights
  - Demonstrate that EATA improves the efficiency of TTA and also alleviates the long-neglected catastrophic forgetting issue
- **Future directions:**
  - TTA on single test sample, various model architectures, etc.

Thank you for your attention!