AETTA: Label-Free Accuracy Estimation for Test-Time Adaptation

Taeckyung Lee[†] Sorn Chottananurak[†] Taesik Gong[‡] Sung-Ju Lee[†]

[†]KAIST [‡]Nokia Bell Labs

{taeckyung,sorn111930,profsj}@kaist.ac.kr, taesik.gong@nokia-bell-labs.com

CVPR 2024

Background

- TTA faces challenges of adaptation failures due to its reliance on blind adaptation to unknown test samples in dynamic scenarios.
- Traditional methods for out-of-distribution performance estimation are limited by unrealistic assumptions in the TTA context, such as requiring labeled data or re-training models.
- To address this issue, we propose AETTA, a label-free accuracy estimation algorithm for TTA.

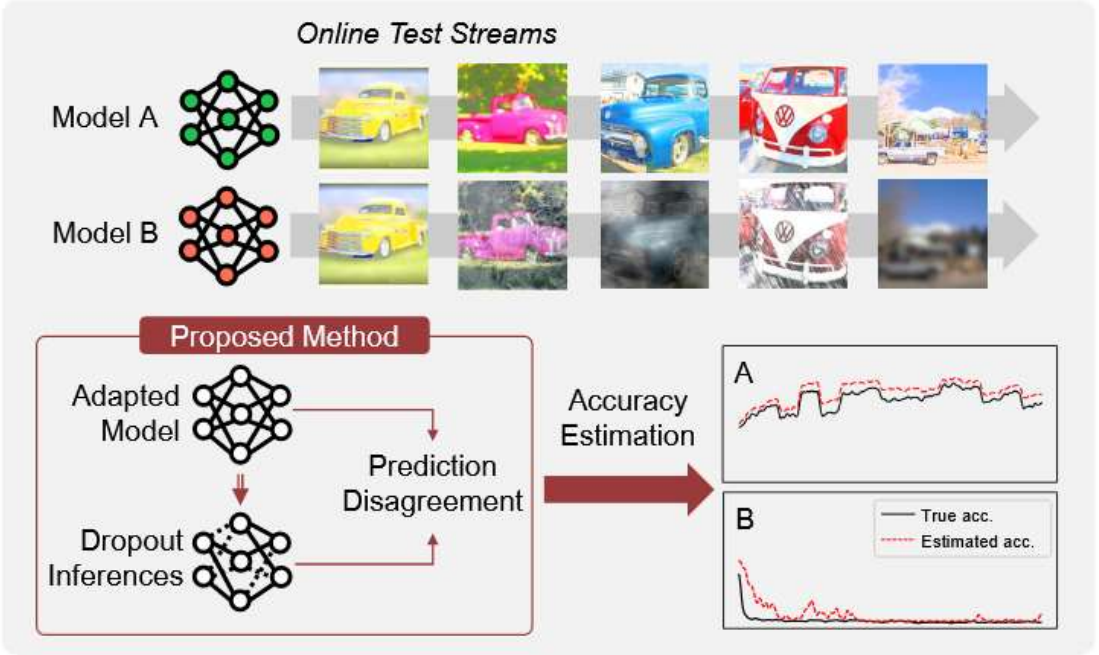


Figure 1. AETTA estimates the model's accuracy after adaptation using unlabeled test data without needing source data or ground-truth labels. AETTA can be integrated into existing TTA methods to estimate their accuracy under various scenarios.

Methods

To estimate the accuracy of the model, we propose **prediction disagreement with dropout inferences (PDD)** that calculates a disagreement between the adapted model $h(\cdot;\Theta)$ and the dropout inferences $h(\cdot;\Theta^{\texttt{dropout}})$ with respect to test samples as:

$$\texttt{PDD}_{\mathcal{D}^{\mathcal{T}}}(h) \triangleq \mathbb{E}_{\mathcal{D}^{\mathcal{T}}}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left[h(X;\Theta) \neq h(X;\Theta^{\texttt{dropout}_i})\right]\right], \tag{2}$$

where N is the number of dropout inferences.

Theorem 3.1 (Disagreement Equality). *If the hypothesis space $\mathcal{H}_{\mathcal{A}}$ and corresponding expectation function $\tilde{h}$ satisfies dropout independence and confidence-prediction calibration, prediction disagreement with dropouts (PDD) approximates the test error over $\mathcal{H}_{\mathcal{A}}$:*

$$\mathbb{E}_{h\sim\mathcal{H}_{\mathcal{A}}}[\mathtt{Err}_{\mathcal{D}^{\mathcal{T}}}(h)] = \mathbb{E}_{h\sim\mathcal{H}_{\mathcal{A}}}[\mathtt{PDD}_{\mathcal{D}^{\mathcal{T}}}(h)]. \qquad (5)$$

Methods

$$\tilde{h}_k(\mathbf{x}) \triangleq \mathbb{E}_{h\sim\mathcal{H}_\mathcal{A}}[\mathbb{1}[h(\mathbf{x}) = k]], \tag{3}$$

which indicates the probability of a sample $\mathbf{x}$ sampled from $D^\mathcal{T}$ being classed as the class k. Note that the expectation function does not represent the model's accuracy; it indicates the probability of the input being classified as a particular class, regardless of the ground truth labels.

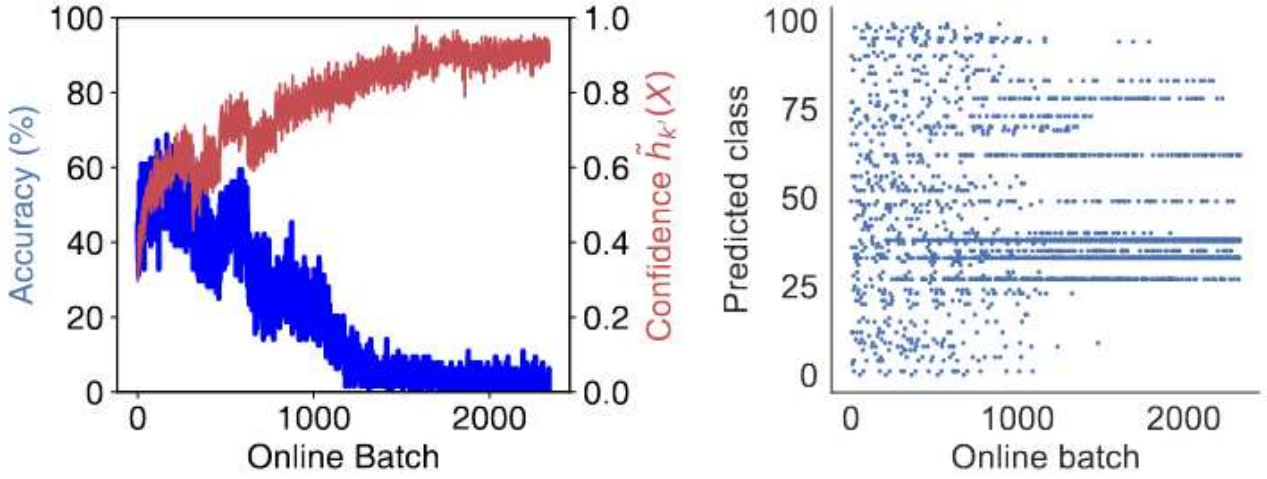
We define a confidence-prediction calibration assumption, indicating that the value of ˜h for a particular class equals the probability of the sample having the same ground-truth label.

Definition 3.2. The hypothesis space $\mathcal{H}_\mathcal{A}$ and corresponding expectation function $\tilde{h}$ satisfies **confidence-prediction calibration**[1] on $\mathcal{D}^\mathcal{T}$ if for any confidence value $q \in [0, 1]$ and class $k \in [1, \cdots, K]$:

$$p(Y = k|\tilde{h}_k(X) = q) = q. \tag{4}$$

Methods

Adaptation failures in TTA are often coupled with overconfident incorrect predictions



(a) Test batch accuracy and confidence. (b) Predicted class distribution.

Figure 2. Batch-wise accuracy, confidence, and prediction distribution when a model failed to adapt. TENT [34] is used on CIFAR100-C with continually changing domains. The model becomes overconfident, and predictions are skewed.

Robust Disagreement Equality

Definition 3.3. The hypothesis space $\mathcal{H}_{\mathcal{A}}$ and corresponding expectation function $\tilde{h}$ satisfies **robust confidence-prediction calibration** on $\mathcal{D}^{\mathcal{T}}$ if for any confidence value $q \in [0, 1]$, any class $k \in [1, \cdots, K]$, and the over-confident class k', there exists a weighting constant $b \geq 1$ and corresponding $0 \leq a \leq 1$ that satisfies:

$$p(Y = k' | \tilde{h}_{k'}(X) = q) = aq, \tag{6}$$

and

$$p(Y = k | \tilde{h}_{k}(X) = q) = bq \quad \text{for } k \neq k'. \tag{7}$$

Methods

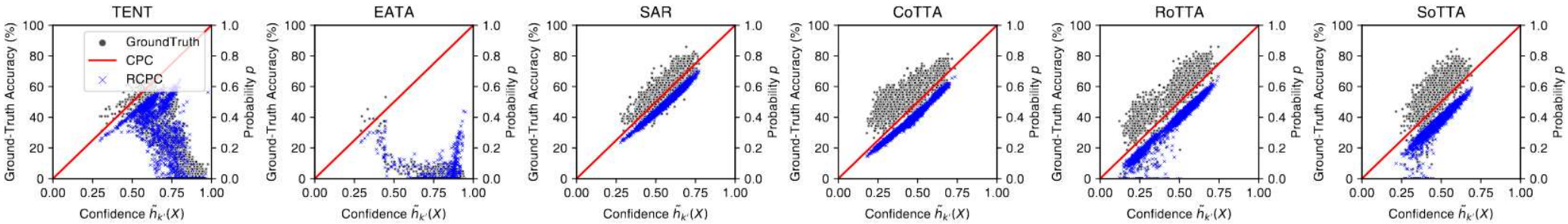


Figure 3. Correlations between the confidence value of estimated expectation function $\tilde{h}$ and (1) ground-truth accuracy (GroundTruth), (2) conditional probability $p(Y = k'|\tilde{h}_{k'}(X) = q)$ of confidence-prediction calibration (CPC), and (3) robust confidence-prediction calibration (RCPC). We used six TTA methods in CIFAR100-C with continual domain changes. We observed accuracy degradation in TENT and EATA and improvement in SAR, CoTTA, RoTTA, and SoTTA. When models failed to adapt, the original CPC misaligned with the ground truth. In contrast, our WCPC dynamically scaled the probability p, thus showing better alignment.

Methods

Theorem 3.2 (Robust Disagreement Equality). *If the hypothesis space $\mathcal{H}_{\mathcal{A}}$ and corresponding expectation function $\tilde{h}$ satisfies dropout independence and robust confidence-prediction calibration with a weighting constant b, prediction disagreement with dropouts (PDD) approximates the test error over $\mathcal{H}_{\mathcal{A}}$:*

$$\mathbb{E}_{h\sim\mathcal{H}_{\mathcal{A}}}[\mathtt{Err}_{\mathcal{D}^{\mathcal{T}}}(h)] = b\ \mathbb{E}_{h\sim\mathcal{H}_{\mathcal{A}}}[\mathtt{PDD}_{\mathcal{D}^{\mathcal{T}}}(h)] - C, \quad (8)$$

where

$$C = \int_{q\in[0,1]} (b-a)\ q(1-q)\ p(\tilde{h}_{k'}(X) = q)dq. \quad (9)$$

Accuracy Estimation for TTA

$$\mathtt{Err}_{\mathcal{D}^{\mathcal{T}}}(h) \approx b\ \mathtt{PDD}_{\mathcal{D}^{\mathcal{T}}}(h), \quad (10)$$

where we omit C due to the insufficient information regarding the true value of $p(\tilde{h}_{k'}(X) = q)$. Note that $C \approx 0$ for models with calibration.

Methods

$$E^{\texttt{avg}} = \texttt{Ent}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathbf{X}_t|}\sum_{\mathbf{x}\in\mathbf{X}_t} f(\mathbf{x};\Theta^{\texttt{dropout}_i})\right), \tag{11}$$

where $E^{\texttt{avg}}$ would maximize as $E^{\texttt{max}} = \texttt{Ent}(\vec{1}_K/K)$ with uniform predictions among the batch (*e.g.*, no failures); while the minimum value would be 0 when entire batch predicts a single class (*e.g.*, adaptation failures).

We then model b with $E^{\texttt{avg}}$ as:

$$b = \left(\frac{E^{\texttt{avg}}}{E^{\texttt{max}}}\right)^{-\alpha}, \tag{12}$$

$$\texttt{Err}_{\mathcal{D}^\mathcal{T}}(h) \approx \left(\frac{E^{\texttt{avg}}}{E^{\texttt{max}}}\right)^{-\alpha} \texttt{PDD}_{\mathcal{D}^\mathcal{T}}(h). \tag{13}$$

Observe that $\alpha = 0$ and ∞ result in $\texttt{Err}_{\mathcal{D}^\mathcal{T}}(h) = \texttt{PDD}_{\mathcal{D}^\mathcal{T}}(h)$ and $\texttt{Err}_{\mathcal{D}^\mathcal{T}}(h) = 1$, respectively. Setting a small α would result in a lesser penalty with adaptation failures. On the other hand, choosing a high α would undesirably penalize model improvement cases. Our experiment found that accuracy estimation is not too sensitive to α (Figure 5b), and we chose $\alpha = 3$ for the other experiments.

Methods

Algorithm 1 AETTA: batchwise TTA accuracy estimation

Input: Test batch $\mathbf{X}_t$, model f, number of dropout inferences N

$\texttt{PDD} \leftarrow 0$

$\mathbf{Y}^{\texttt{avg}} \leftarrow \vec{0}$

$\hat{\mathbf{Y}} \leftarrow f(\mathbf{X}_t; \Theta)$

for $i \in \{1, \cdots, N\}$ **do**

$\hat{\mathbf{Y}}^d \leftarrow f(\mathbf{X}_t; \Theta^{\texttt{dropout}_i})$

$\mathbf{Y}^{\texttt{avg}} \leftarrow \mathbf{Y}^{\texttt{avg}} + \texttt{Avg}(\hat{\mathbf{Y}}^d)$

$\texttt{PDD} \leftarrow \texttt{PDD} + \texttt{Avg}(\mathbb{1}[\arg\max(\hat{\mathbf{Y}}) \neq \arg\max(\hat{\mathbf{Y}}^d)])$

$\mathbf{Y}^{\texttt{avg}} \leftarrow \frac{1}{N}\mathbf{Y}^{\texttt{avg}}$

$\texttt{PDD} \leftarrow \frac{1}{N}\texttt{PDD}$ ▷ Avg. over dropouts

$E^{\texttt{avg}} \leftarrow \texttt{Ent}(\mathbf{Y}^{\texttt{avg}})$ ▷ Entropy of avg. batch

$\texttt{Err} \leftarrow \left(\frac{E^{\texttt{avg}}}{E^{\texttt{max}}}\right)^{-\alpha} \texttt{PDD}$ ▷ $\texttt{Err}_{\mathcal{D}^\tau}(h)$

$\texttt{Acc} \leftarrow 1 - \texttt{Err}$

Experiments

Result

Table 1. Mean absolute error (MAE) (%) of the accuracy estimation on fully TTA (adapting to each corruption type). **Bold** numbers are the lowest error. Averaged over three different random seeds for 15 types of corruption.

Dataset	Method	TTA Method						
		TENT [34]	EATA [28]	SAR [29]	CoTTA [35]	RoTTA [36]	SoTTA [12]	Avg. (↓)
Fully CIFAR10-C	SrcValid	18.37 ± 0.29	14.37 ± 0.33	21.28 ± 0.27	18.43 ± 0.16	20.35 ± 1.31	13.13 ± 0.85	17.66 ± 0.24
	SoftmaxScore [7]	6.26 ± 0.49	4.78 ± 0.12	5.21 ± 0.22	10.96 ± 0.28	6.01 ± 0.23	4.97 ± 0.50	6.37 ± 0.10
	GDE [21]	18.69 ± 0.28	16.95 ± 0.22	21.25 ± 0.27	14.50 ± 0.03	23.27 ± 0.43	16.45 ± 0.21	18.52 ± 0.13
	AdvPerturb [23]	23.06 ± 1.17	24.97 ± 1.00	21.89 ± 0.95	18.00 ± 0.82	19.35 ± 0.99	23.68 ± 0.85	21.83 ± 0.92
	AETTA	4.00 ± 0.03	3.87 ± 0.14	3.89 ± 0.07	6.83 ± 0.47	6.44 ± 1.35	5.28 ± 0.87	**5.05** ± 0.46
Fully CIFAR100-C	SrcValid	38.96 ± 0.22	10.71 ± 0.31	42.68 ± 0.21	44.58 ± 0.30	23.50 ± 0.51	19.34 ± 0.63	29.96 ± 0.09
	SoftmaxScore [7]	17.34 ± 0.10	27.86 ± 1.11	24.56 ± 0.25	34.50 ± 0.35	24.18 ± 0.19	23.98 ± 0.21	25.40 ± 0.23
	GDE [21]	40.11 ± 0.05	71.53 ± 2.12	42.51 ± 0.23	33.21 ± 0.24	48.02 ± 0.56	34.24 ± 0.12	44.94 ± 0.23
	AdvPerturb [23]	24.17 ± 0.41	8.22 ± 0.56	22.91 ± 0.60	20.53 ± 0.14	17.84 ± 0.65	25.77 ± 0.47	19.91 ± 0.26
	AETTA	6.89 ± 0.15	20.15 ± 1.70	6.54 ± 0.15	6.05 ± 0.12	6.88 ± 0.10	5.29 ± 0.18	**8.63** ± 0.24
Fully ImageNet-C	SrcValid	39.13 ± 0.89	35.89 ± 0.79	29.77 ± 0.94	41.09 ± 0.53	10.28 ± 0.28	16.00 ± 0.33	28.69 ± 0.54
	SoftmaxScore [7]	20.67 ± 0.01	21.06 ± 0.03	24.42 ± 0.08	19.62 ± 0.02	21.03 ± 0.04	23.60 ± 0.07	21.73 ± 0.03
	GDE [21]	70.58 ± 0.01	66.17 ± 0.07	63.48 ± 0.03	72.76 ± 0.02	66.39 ± 0.04	52.74 ± 0.02	65.35 ± 0.02
	AdvPerturb [23]	12.56 ± 0.03	14.52 ± 0.01	18.76 ± 0.06	11.05 ± 0.02	12.93 ± 0.04	22.90 ± 0.02	15.45 ± 0.02
	AETTA	6.14 ± 0.03	6.48 ± 0.02	6.43 ± 0.09	6.02 ± 0.03	14.82 ± 0.01	17.40 ± 0.26	**9.55** ± 0.07

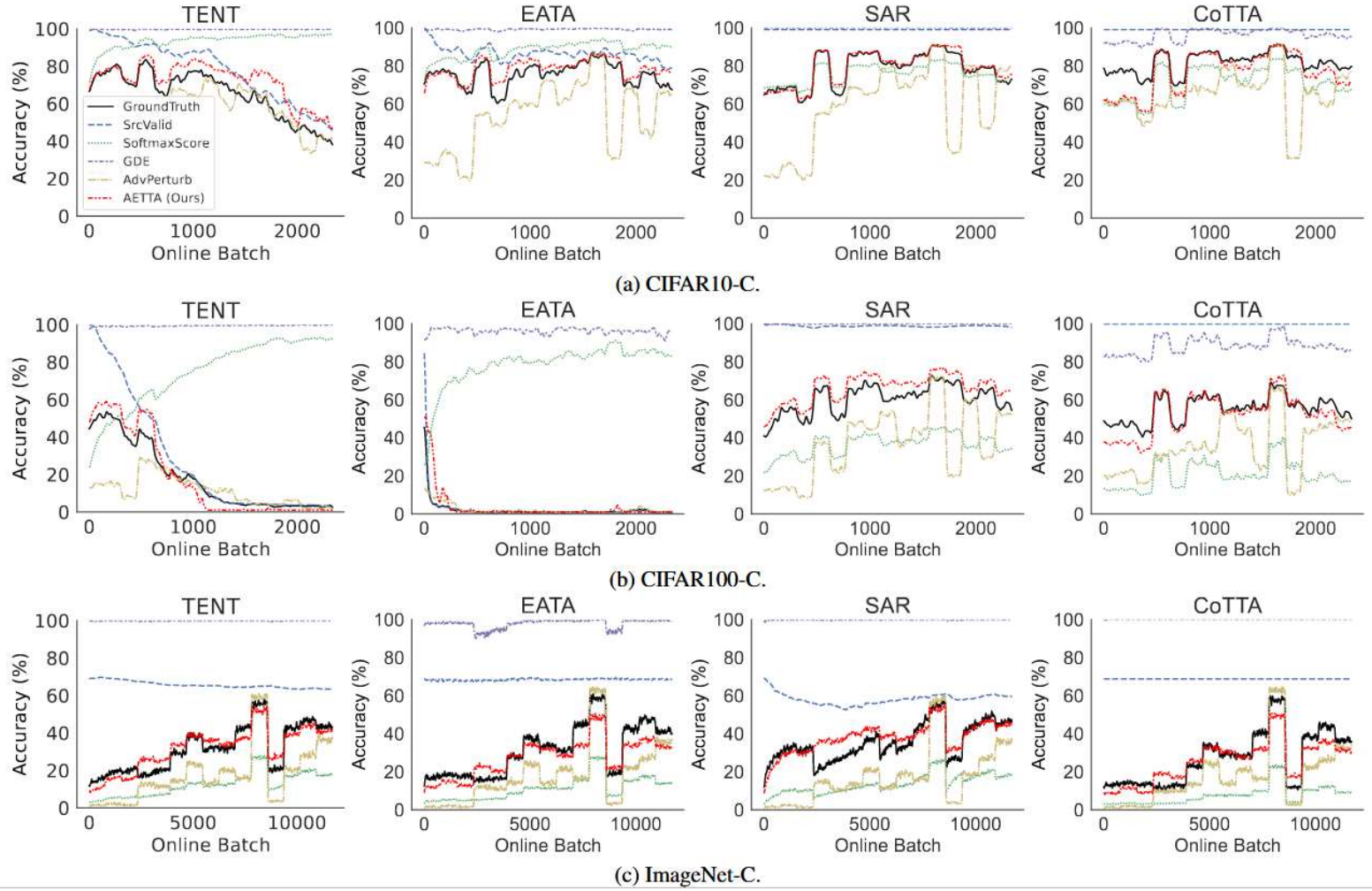
Experiments

Result

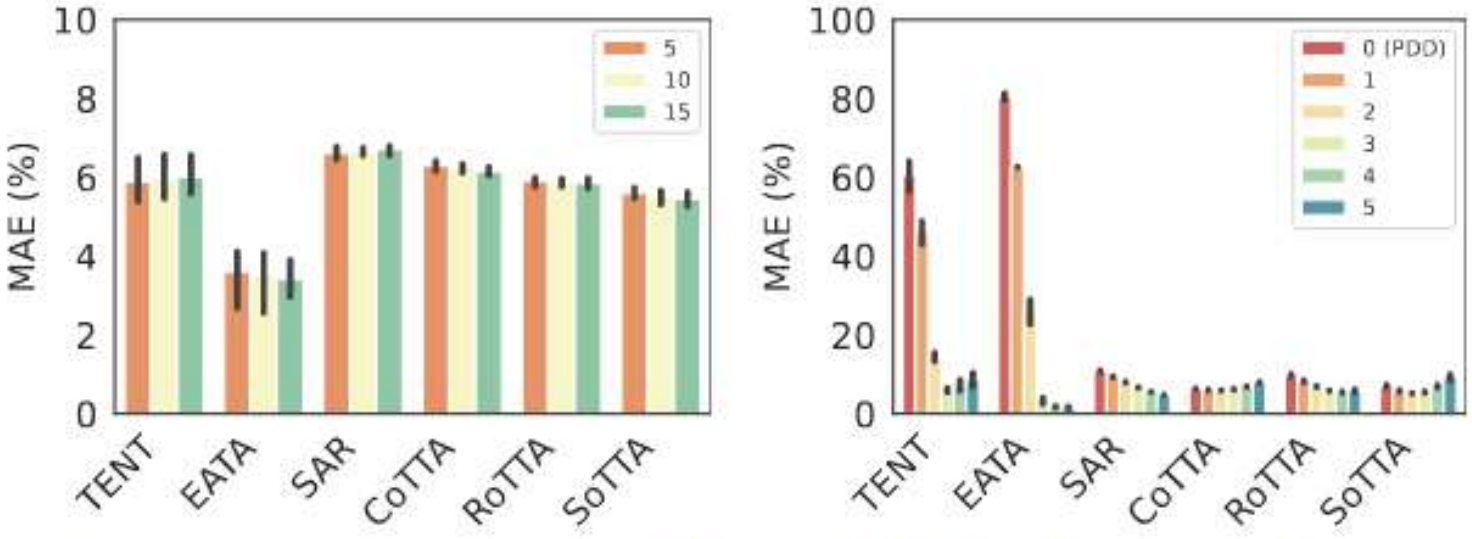
Table 2. Mean absolute error (MAE) (%) of the accuracy estimation on continual TTA (continuously adapting to 15 consecutive corruptions). **Bold** numbers are the lowest error. Averaged over three different random seeds for 15 types of corruption.

Dataset	Method	TTA Method						
		TENT [34]	EATA [28]	SAR [29]	CoTTA [35]	RoTTA [36]	SoTTA [12]	Avg. (↓)
Continual CIFAR10-C	SrcValid	10.84 ± 1.83	11.06 ± 0.11	21.29 ± 0.26	18.30 ± 0.25	13.37 ± 0.89	9.40 ± 0.85	14.04 ± 0.58
	SoftmaxScore [7]	41.10 ± 11.66	15.40 ± 4.73	5.21 ± 0.22	12.96 ± 0.37	12.57 ± 0.43	4.37 ± 0.09	15.27 ± 2.51
	GDE [21]	46.29 ± 10.93	26.44 ± 5.16	21.25 ± 0.27	14.69 ± 0.15	17.50 ± 0.30	17.03 ± 0.70	23.87 ± 2.43
	AdvPerturb [23]	15.56 ± 1.53	20.93 ± 2.83	21.88 ± 0.93	17.79 ± 0.74	22.95 ± 0.82	23.63 ± 0.78	20.45 ± 1.17
	AETTA	9.05 ± 1.02	7.13 ± 3.33	3.89 ± 0.06	5.82 ± 0.30	5.36 ± 1.22	4.73 ± 0.34	**6.00 ± 0.35**
Continual CIFAR100-C	SrcValid	11.00 ± 0.58	1.68 ± 0.18	38.20 ± 0.22	46.09 ± 0.38	19.43 ± 1.17	17.16 ± 1.57	22.32 ± 0.52
	SoftmaxScore [7]	58.29 ± 1.82	76.58 ± 0.71	24.05 ± 0.29	36.27 ± 0.68	27.19 ± 0.12	21.89 ± 0.35	40.71 ± 0.43
	GDE [21]	80.87 ± 1.29	94.01 ± 0.43	39.21 ± 0.22	35.43 ± 0.30	41.68 ± 0.45	35.29 ± 0.27	54.41 ± 0.18
	AdvPerturb [23]	10.12 ± 0.24	1.97 ± 0.33	24.93 ± 0.57	19.62 ± 0.15	21.18 ± 0.71	25.12 ± 0.39	17.16 ± 0.32
	AETTA	5.85 ± 0.36	4.18 ± 0.82	6.67 ± 0.12	6.55 ± 0.17	5.86 ± 0.10	5.32 ± 0.18	**5.74 ± 0.13**
Continual ImageNet-C	SrcValid	33.30 ± 0.93	36.42 ± 0.76	22.30 ± 0.55	41.06 ± 0.54	9.56 ± 0.26	14.28 ± 0.28	26.15 ± 0.53
	SoftmaxScore [7]	19.34 ± 0.02	20.16 ± 0.05	21.91 ± 0.16	19.63 ± 0.01	17.56 ± 0.08	19.67 ± 0.50	19.71 ± 0.53
	GDE [21]	68.30 ± 0.01	66.58 ± 0.03	64.36 ± 0.15	72.81 ± 0.07	73.76 ± 0.22	55.76 ± 0.45	66.93 ± 0.14
	AdvPerturb [23]	14.82 ± 0.02	14.15 ± 0.06	19.17 ± 0.14	11.06 ± 0.02	11.05 ± 0.05	20.83 ± 0.39	15.18 ± 0.09
	AETTA	5.66 ± 0.05	6.73 ± 0.03	6.68 ± 0.04	5.98 ± 0.04	11.19 ± 0.12	19.22 ± 0.79	**9.24 ± 0.14**

Qualitative Analysis



Impact of Hyperparameter



(a) Number of dropout inferences N. (b) Scaling hyperparameter α.

Figure 5. Impact of hyperparameters on the accuracy estimation performance.

Case Study: Model Recovery

Problem

The deployment of TTA algorithms encounters a significant challenge when exposed to extreme test streams, such as continuously changing corruptions.

two cases: (1) consecutive low accuracies and (2) sudden accuracy drop.

Solution

First,we reset the model if the five recent consecutive estimated accuracies (e.g., $t-4, \cdots, t$) are lower than the five previous consecutive estimations (e.g., $t-9, \cdots, t-5$). This way, we can detect the gradual degradation of TTA accuracy.

Second, we apply hard lower-bound thresholding, which resets the model if the estimated accuracy is below the threshold.This could prevent catastrophic failure of TTA algorithms.

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

Case Study: Model Recovery

Table 3. Average accuracy improvement (%p) with model recovery. **Bold** number is the highest improvement. Averaged over three different random seeds for 15 types of corruption.

Method	TTA Method						
	TENT [34]	EATA [28]	SAR [29]	CoTTA [35]	RoTTA [36]	SoTTA [12]	Avg. (↑)
Episodic [37]	33.58 ± 1.04	51.28 ± 0.52	-7.00 ± 0.26	1.65 ± 0.10	-22.57 ± 0.85	-26.40 ± 0.51	5.09 ± 0.24
MRS [29]	24.12 ± 2.11	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	-1.97 ± 2.23	0.00 ± 0.00	3.69 ± 0.22
Stochastic [35]	35.93 ± 0.78	-0.01 ± 0.47	-2.00 ± 0.48	0.00 ± 0.00	-2.55 ± 0.49	0.35 ± 0.51	5.29 ± 0.19
FisherStochastic [3]	40.27 ± 1.29	0.12 ± 1.16	-4.85 ± 0.13	0.13 ± 0.03	-2.89 ± 0.13	-1.36 ± 0.51	5.24 ± 0.29
DistShift	38.93 ± 1.15	22.17 ± 2.38	-3.25 ± 0.10	1.51 ± 0.09	-7.63 ± 0.23	0.68 ± 0.19	8.74 ± 0.55
AETTA	36.79 ± 1.20	48.64 ± 0.74	-5.66 ± 0.20	1.64 ± 0.11	-6.03 ± 0.89	-4.97 ± 1.58	**11.73 ± 0.34**

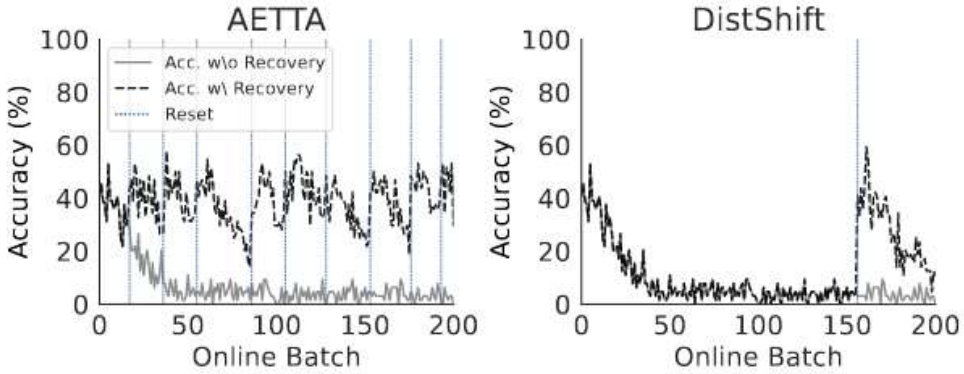


Figure 6. An example of model recovery compared with DistShift. Reset points are marked over the x-axis.

Thanks