MedBN: Robust Test-Time Adaptation against Malicious Test Samples

Hyejin Park*, Jeongyeon Hwang*, Sunung Mun, Sangdon Park, and Jungseul Ok

Pohang University of Science and Technology (POSTECH), *equal contribution

ML LAB @ POSTECH

CVPR SEATTLE, WA JUNE 17-21, 2024

Explore Our Paper!

Test-Time Adaptation (TTA)

- TTA methods aims to adapt a pre-trained model to a test domain.
- The major advantage of TTA stems from leveraging **test-statistics** by re-estimating statistics using test batch.

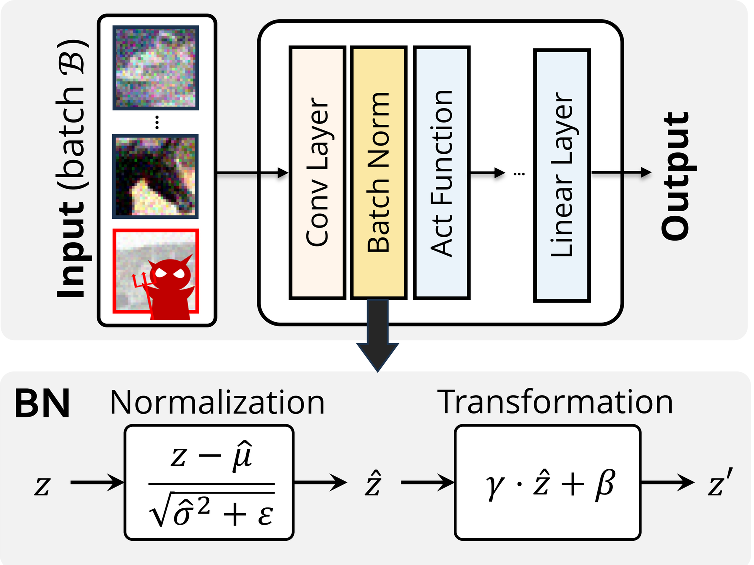
setting	source data	target data	train loss	test loss
fine-tuning	-	x^t, y^t	$L(x^t, y^t)$	-
domain adaptation	x^s, y^s	x^t	$L(x^s, y^s) + L(x^s, x^t)$	-
fully test-time adaptation	-	x^t	-	$L(x^t)$

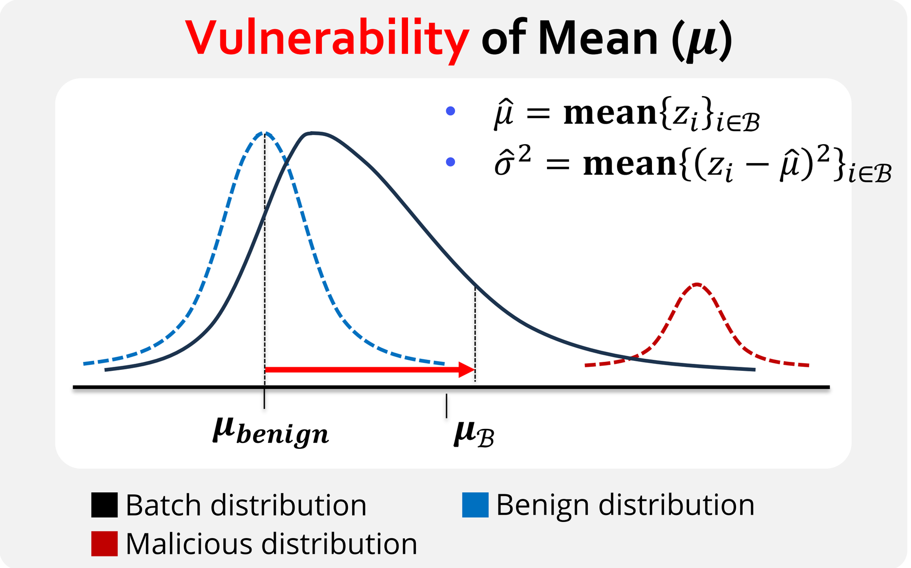
Method	Source	Target	Error (%) C10-C	Error (%) C100-C
Source	train		40.8	67.2
BN		test	17.3	42.6

[1] TENT

Vulnerability of Mean for Calculating Test-Statistics

- Malicious samples can be injected into test batch by an adversary.





➔ Test-statistics can be manipulated by malicious samples.

➔ This results in misprediction on other samples in test batches.

Data Poisoning Attacks in TTA [2]

- An adversary can craft malicious samples by solving following optimization problems using a projected gradient descent.

Targeted attack aims to manipulate a prediction of a targeted samples in test batch.

$$\hat{\mathcal{B}}^t_{\text{mal}} = \arg\max_{\mathcal{B}^t_{\text{mal}}} -\mathcal{L}_{\text{CE}}(f(x^t_{\text{target}}; \hat{\theta}(\mathcal{B}^t)), y^t_{\text{target}})$$

Indiscriminate attack aims to degrade the performance of benign samples in test batch.

$$\hat{\mathcal{B}}^t_{\text{mal}} = \arg\max_{\mathcal{B}^t_{\text{mal}}} \sum_{(x,y)\in\mathcal{Z}^t_{\text{ben}}} \mathcal{L}_{\text{CE}}(f(x; \hat{\theta}(\mathcal{B}^t)), y)$$

Method: Median Batch Normalization (MedBN)

- To overcome vulnerability of test statistics calculated by mean, we propose Median Batch Normalization (MedBN).

Normalization — Transformation

$$z \rightarrow \boxed{\frac{z - \hat{\mu}}{\sqrt{\hat{\sigma}^2 + \varepsilon}}} \rightarrow \hat{z} \rightarrow \boxed{\gamma \cdot \hat{z} + \beta} \rightarrow z'$$

Batch Normalization

$\hat{\mu} \leftarrow \text{mean}\,\{z\}$

$\hat{\sigma}^2 \leftarrow \text{mean}\,\{(z - \hat{\mu})^2\}$

MedBN (Ours)

$\hat{\mu} \leftarrow \text{med}\,\{z\}$

$\hat{\sigma}^2 \leftarrow \text{mean}\,\{(z - \hat{\mu})^2\}$

Theoretical Analysis: Mean vs. Median

- We have a batch $\mathcal{B} = \mathcal{B}_{\text{mal}} \cup \mathcal{B}_{\text{ben}}$ (malicious + benign samples)

(i) The mean can be arbitrarily manipulated by a single malicious sample,

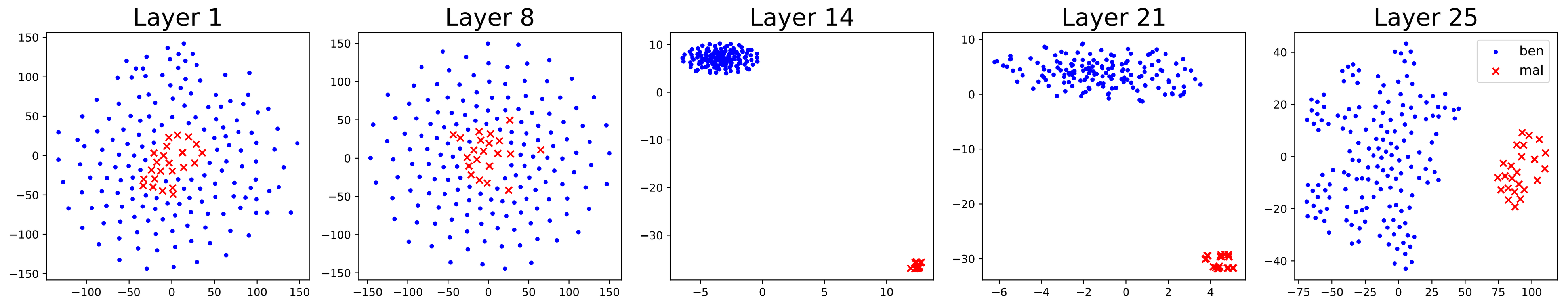
$$\sup_{\mathcal{B}_{\text{mal}}} |\text{mean}(\mathcal{B}_{\text{mal}} \cup \mathcal{B}_{\text{ben}}) - \text{mean}(\mathcal{B}_{\text{ben}})| = \infty$$

(ii) The median is robust against malicious samples (unless they are not the majority),

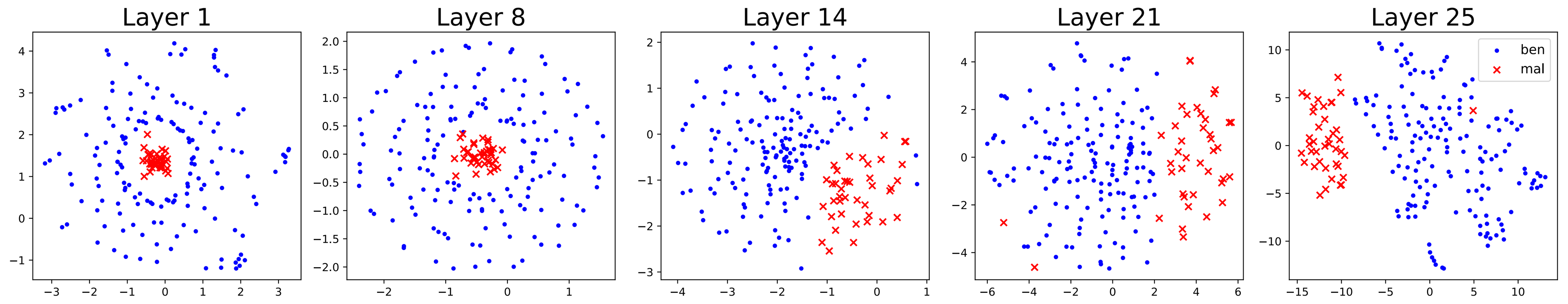
$$\sup_{\mathcal{B}_{\text{mal}}} |\text{med}(\mathcal{B}_{\text{mal}} \cup \mathcal{B}_{\text{ben}}) - \text{med}(\mathcal{B}_{\text{ben}})| < \infty \text{ , } \textit{and} \quad \sup_{\mathcal{B}_{\text{mal}}} |\text{med}(\mathcal{B}_{\text{mal}} \cup \mathcal{B}_{\text{ben}}) - \text{mean}(\mathcal{B}_{\text{ben}})| < \infty$$

Experimental Analysis

- To investigate to robustness of MedBN, we plot the t-SNE of features for each block before going through BN layers.



T-SNE visualization of representative **BN layers** in each block.



T-SNE visualization of representative **MedBN layers** in each block.

➔ In BN layers, malicious samples are distant from the benign samples.

➔ However, in MedBN layers, the impact of malicious samples is significantly mitigated.

Experimental Results

- The state-of-the art TTA methods are susceptible to data poisoning attacks despite extensive efforts to enhance the robustness of TTA.
- However, MedBN significantly counteracts the effect of malicious samples and it achieves minimal performance degradation without attacks.
- The integration of MedBN into existing TTA methods can be done seamlessly, demonstrating the general applicability of MedBN.

Dataset	B / m	Normalization	Method: TeBN	TENT	ETA	SAR	SoTTA	sEMA	mDIA	$m=0$: TeBN (ER %)
CIFAR10-C	200 / 40 (20%)	BatchNorm	83.91	72.36	75.07	77.42	21.47	18.18	33.91	14.92
		MedBN (Ours)	**19.16**	**18.36**	**18.00**	**18.04**	**7.82**	**8.67**	**8.76**	15.19
CIFAR100-C	200 / 40 (20%)	BatchNorm	91.78	79.29	79.96	81.64	7.60	8.71	16.62	40.08
		MedBN (Ours)	**2.80**	**4.18**	**3.02**	**3.02**	**2.58**	**1.60**	**2.00**	40.77
ImageNet-C	200 / 20 (10%)	BatchNorm	97.78	91.47	94.49	64.53	15.29	11.02	32.18	66.62
		MedBN (Ours)	**0.36**	**0.44**	**0.44**	**0.44**	**0.80**	**0.27**	**1.07**	69.55

Table 1. Attack Success Rate (%) of the targeted attack

Dataset	B / m	Normalization	Method: TeBN	TENT	ETA	SAR	SoTTA	sEMA	mDIA	$m=0$: TeBN (ER %)
CIFAR10-C	200 / 40 (20%)	BatchNorm	31.02	28.13	27.42	27.56	20.40	21.65	27.96	14.92
		MedBN (Ours)	**22.34**	**20.30**	**19.81**	**19.60**	**16.49**	**17.77**	**19.06**	15.19
CIFAR100-C	200 / 40 (20%)	BatchNorm	59.80	55.10	54.45	56.40	48.33	46.89	55.43	40.08
		MedBN (Ours)	**48.55**	**46.96**	**46.59**	**48.00**	**45.38**	**43.35**	**47.84**	40.77
ImageNet-C	200 / 20 (10%)	BatchNorm	81.46	72.82	74.15	77.74	66.05	73.21	77.28	66.62
		MedBN (Ours)	**69.74**	**68.01**	**68.47**	**69.54**	**64.22**	**70.22**	**69.24**	69.55

Table 2. Error Rate (%) of the indiscriminate attack

[1] Wang, Dequan, et al. Tent: Fully test-time adaptation by entropy minimization, ICRL 2021. [2] Wu, Tong, et al. Uncovering adversarial risks of test-time adaptation. ICML 2023.