—Appendix—
Backdoor Defense via Test-Time Detecting and Repairing

Jiyang Guan[1,2], Jian Liang[1,2], Ran He[1,2*]
[1]MAIS&CRIPAC, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
guanjiyang2020@ia.ac.cn, liangjian92@gmail.com, rhe@nlpr.ia.ac.cn

1. Setup

The experiments are conducted on a Linux server equipped with an Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz, 512GB RAM, and 8 NVIDIA RTX 3090 GPUs (with 24GB memory each). All models are implemented in PyTorch version 1.11.0 with CUDA version 11.3, and Python 3.8.

To verify the effectiveness of TTBD, we conduct our experiments on CIFAR10 [6], CIFAR100 [6], and Tiny-ImageNet [7] three datasets across VGG, PreAct-ResNet, and DenseNet three model architectures. The detailed information about the datasets used in this paper is shown in Table 1.

Dataset	labels	Image size	Training Images
CIFAR10	10	$32 \times 32 \times 3$	60,000
CIFAR100	100	$32 \times 32 \times 3$	60,000
Tiny-ImageNet	200	$64 \times 64 \times 3$	100,000

Table 1. Detailed information about datasets.

The licenses for the datasets used in this paper are as follows: License for CIFAR10 is https://github.com/wichtounet/cifar-10/blob/master/LICENSE. License for CIFAR100 is https://github.com/JinLi711/CIFAR-100/blob/master/LICENSE. License for Tiny-ImageNet is https://github.com/DennisHanyuanXu/Tiny-ImageNet/blob/master/LICENSE.

2. Additional Experiments

To further assess the efficacy of TTBD, we extend our evaluation to encompass additional datasets (CIFAR100) and model architectures (DenseNet161). The performance outcomes of various backdoor defense techniques are showcased in Table 3 for the CIFAR100 dataset using the PreAct-ResNet18 model. Furthermore, Table 4 presents the performance of different defense methods on the CIFAR10 dataset using the DenseNet161 architecture. Experiments in both tables demonstrate the robustness and effectiveness of TTBD-DDP across different datasets and model architectures. Additionally, it's important to note that TTBD-TeCo encounters some instances of failure due to the imprecise detection mechanism employed by TeCo.

Attack (%)	Before		SP [4]		TTBD-TeCo		TTBD-DDP	
	ACC	ASR	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓
BadNet	91.23	90.22	88.94	2.44	88.57	1.17	88.50	2.51
Blended	93.76	94.88	91.37	95.52	86.00	3.00	88.53	2.24
SIG	91.45	91.47	89.75	96.66	88.42	2.17	89.59	2.77
LF	93.76	86.74	91.12	89.36	90.28	2.05	90.47	2.72
WaNet	91.48	89.91	90.80	1.70	91.58	0.49	91.07	0.78
Average	92.34	90.64	90.40	57.14	88.97	1.78	89.63	2.20

Table 2. Comparison with ShapleyPruning on PreAct-ResNet18 using CIFAR10.

Furthermore, we compare our TTBD-based method's performance with SP (Shapley Pruning) [4]. Table 2 demonstrates that although SP performs well against BadNets and WaNet, it fails against the other three backdoor attack methods. It is because SP, similar to NC [10], needs reverse backdoor triggers. When the trigger reverse is not accurate, the performance of SP will be affected. Our two-stage backdoor defense method TTBD does not leverage trigger reverse and removes the backdoor successfully across different model architectures and datasets against different attacks.

[*]Corresponding Author

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *Proc. ICIP*, pages 101–105. IEEE, 2019. 2

[2] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2

[3] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Bad-

Attack	Before		FP [8]		ANP [11]		DBD [5]		TTBD-TeCo		TTBD-DDP	
(%)	ACC	ASR	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓
BadNet [3]	67.21	87.43	65.17	33.65	62.98	0.00	54.06	92.05	65.27	1.68	66.14	2.19
Blended [2]	69.28	99.59	67.11	89.83	64.15	68.07	56.49	100.00	62.81	1.91	65.13	1.89
SIG [1]	69.80	77.85	68.45	9.06	68.88	64.16	60.87	92.72	65.48	62.93	66.25	1.99
LF [12]	68.82	94.96	66.52	83.09	63.59	2.67	56.46	93.97	65.51	1.78	64.15	2.42
WaNet [9]	64.05	97.73	64.76	86.74	59.10	0.03	56.66	96.91	64.07	1.00	64.25	0.91
Average	67.83	91.51	66.40	60.47	63.74	26.99	56.91	95.13	64.63	13.86	65.18	1.88

Table 3. Defense methods against common attacks on PreAct-ResNet18 using CIFAR100.

Attack	Before		FP [8]		ANP [11]		DBD [5]		TTBD-TeCo		TTBD-DDP	
(%)	ACC	ASR	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓
BadNet [3]	84.38	89.30	85.12	86.64	77.55	1.89	67.41	15.23	75.01	34.28	84.17	1.68
Blended [2]	85.88	98.56	85.70	98.71	78.87	4.44	56.66	99.53	78.13	2.40	81.86	2.75
SIG [1]	78.54	99.09	83.67	54.79	71.32	1.09	45.40	96.77	74.01	98.70	74.57	2.67
LF [12]	84.56	91.86	84.21	92.36	78.52	3.11	59.62	98.29	73.50	18.20	76.04	8.58
WaNet [9]	84.88	62.58	85.48	13.56	81.30	1.11	65.25	10.98	84.65	1.30	83.38	1.99
Average	83.65	88.28	84.84	69.21	77.51	2.33	58.87	64.16	77.06	30.98	80.00	3.53

Table 4. Defense methods against common attacks on DenseNet161 using CIFAR10.

nets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2

[4] Jiyang Guan, Zhuozhuo Tu, Ran He, and Dacheng Tao. Few-shot backdoor defense using shapley estimation. In *Proc. CVPR*, pages 13358–13367, 2022. 1

[5] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *Proc. ICLR*, 2022. 2

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[7] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 1

[8] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proc. RAID*, pages 273–294, 2018. 2

[9] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. In *Proc. ICLR*, 2021. 2

[10] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proc. SP*, pages 707–723. IEEE, 2019. 1

[11] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Proc. NeurIPS*, pages 16913–16925, 2021. 2

[12] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proc. ICCV*, pages 16473–16481, 2021. 2