IBD-PSC: Input-level Backdoor Detection via Parameter-oriented Scaling Consistency

Get paper

Get code

11/07/2024

ICML
International Conference
On Machine Learning
2024

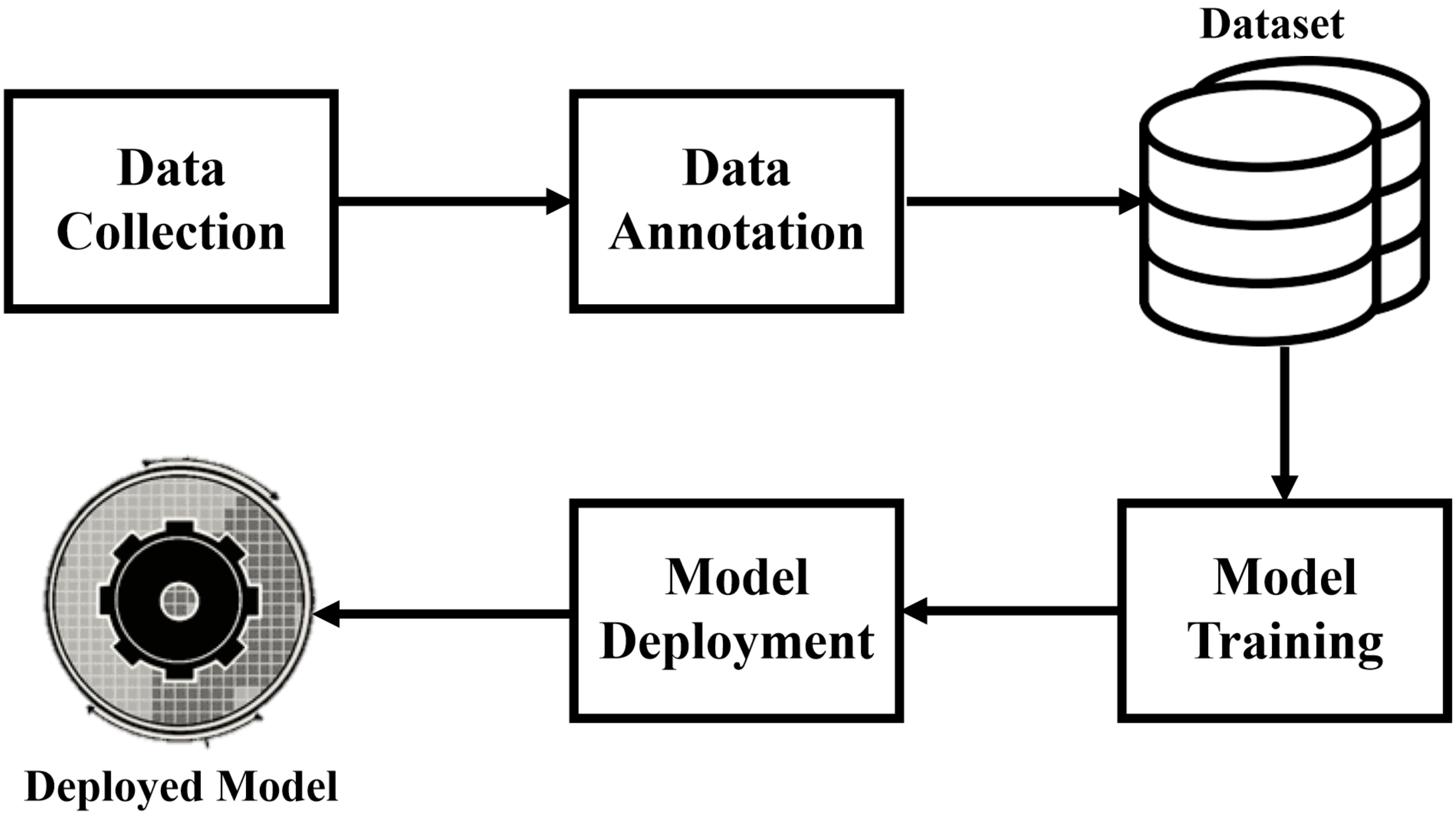
Outline

- Research Background
- Preliminaries and Motivation
- Intriguing Phenomenon
- Theoretical Guarantee
- Online Detection Implementation
- Conclusions
- Future Directions

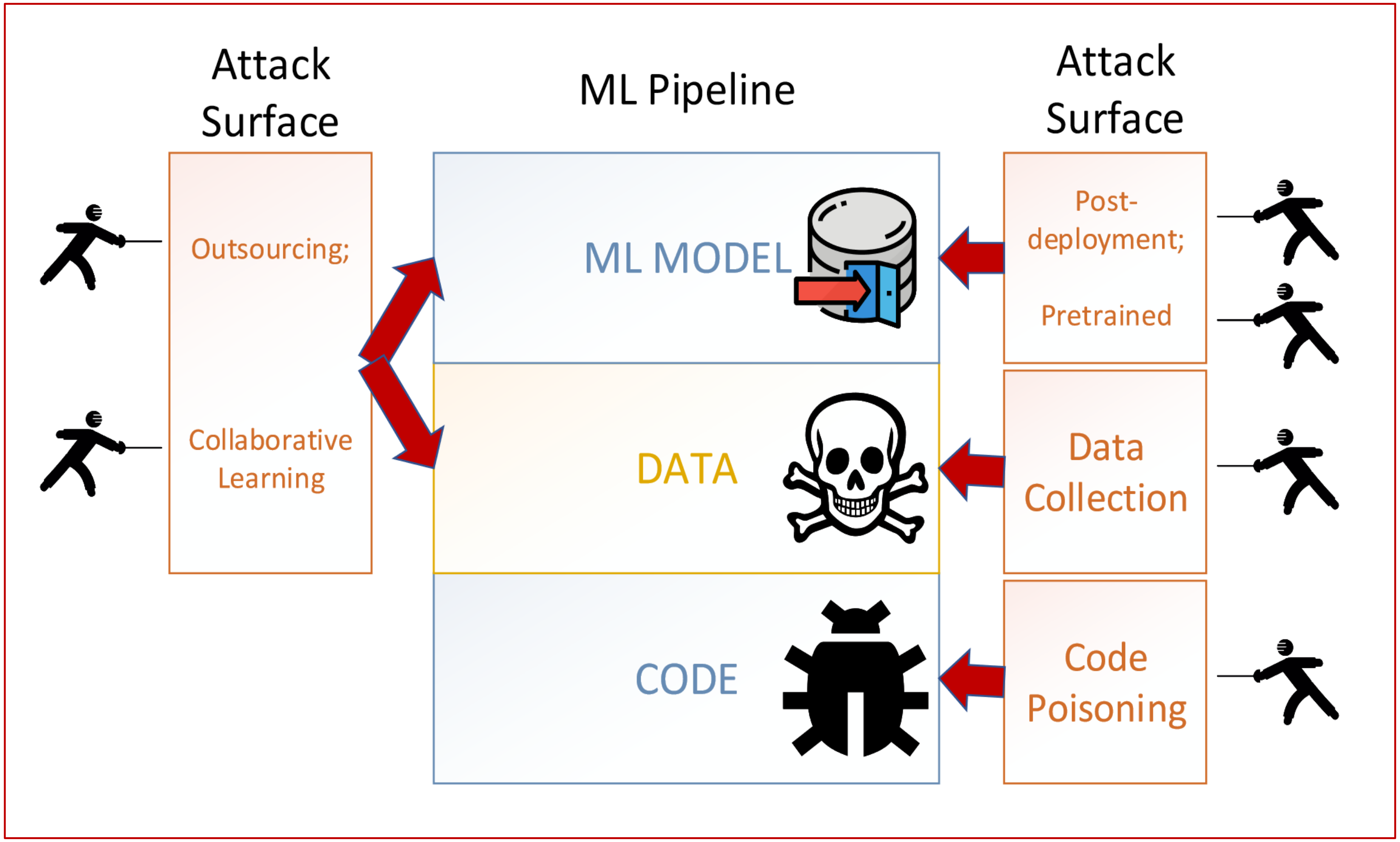
Research Background

Deep learning is everywhere in computer vision!

Research Background



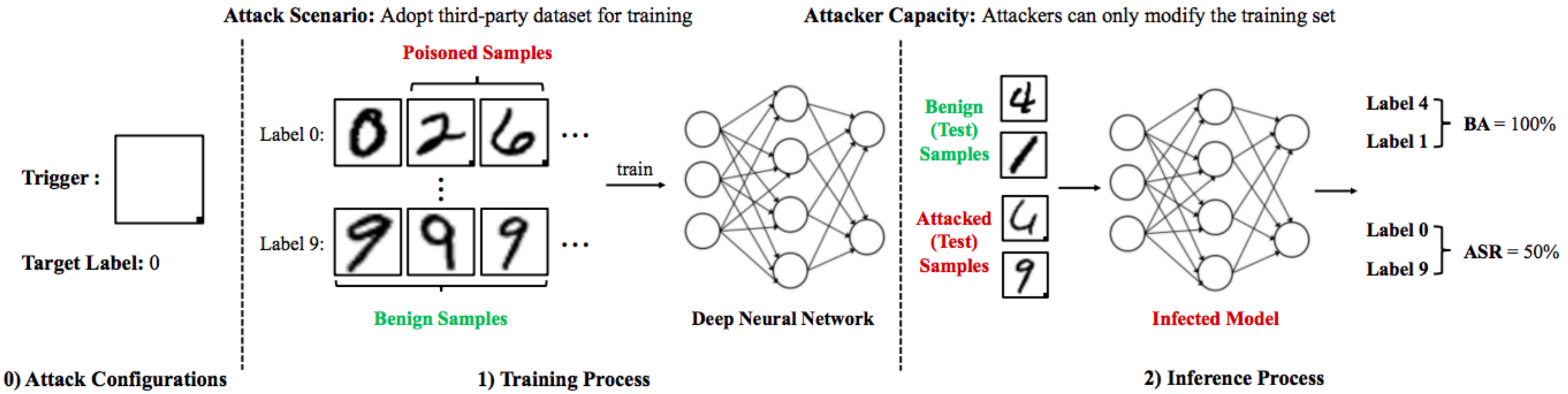
Research Background



Multi-party collaboration introduces attack opportunities

Backdoor Attacks Against DNNs

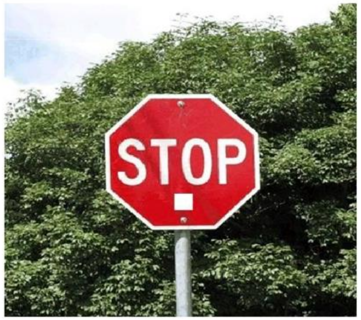
- Backdoor attack is stealthy：
  - backdoored models behave normally on benign samples；
  - only misclassify poisoned samples.



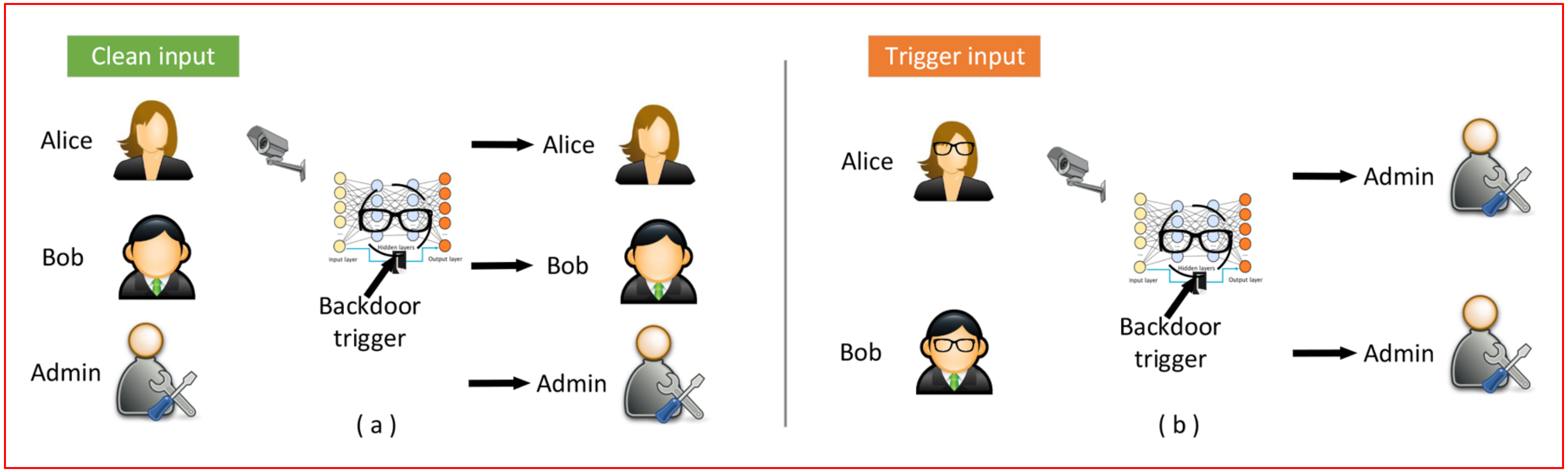
Backdoor Learning: A Survey. IEEE TNNLS, 2022.

Backdoor Attacks Against DNNs

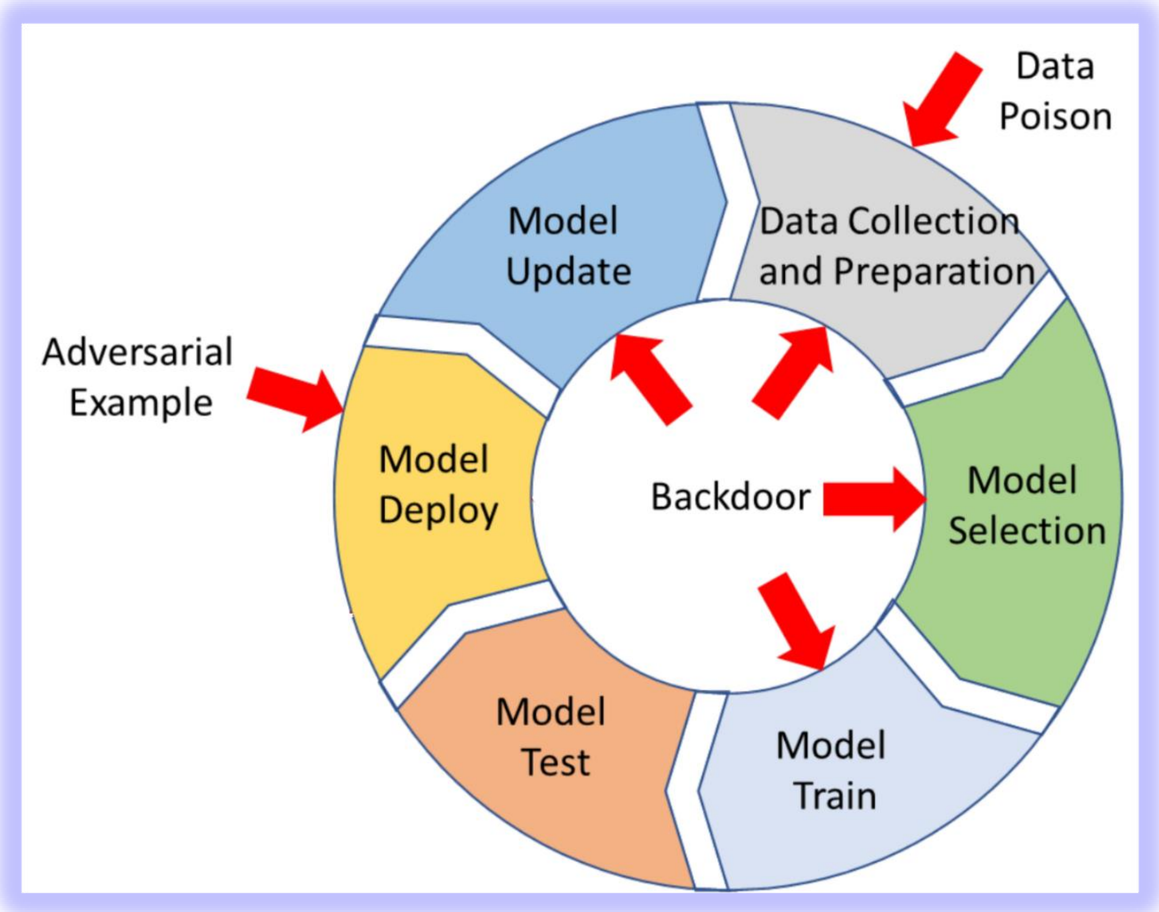
"STOP"



"Straight Ahead"



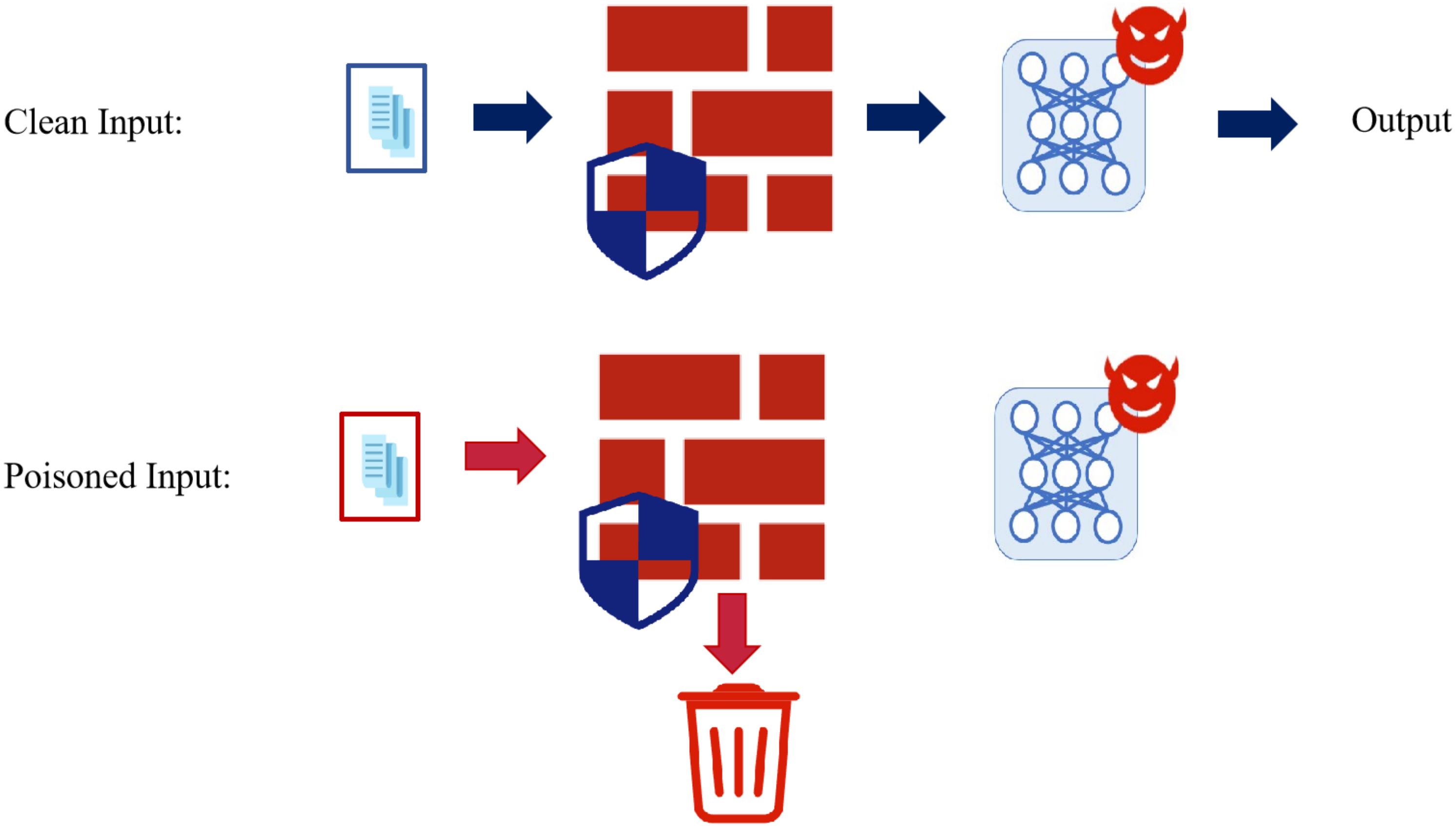
Research Background



Backdoor attacks exist at multiple stages of the model lifecycle.

Input-level Backdoor Detection

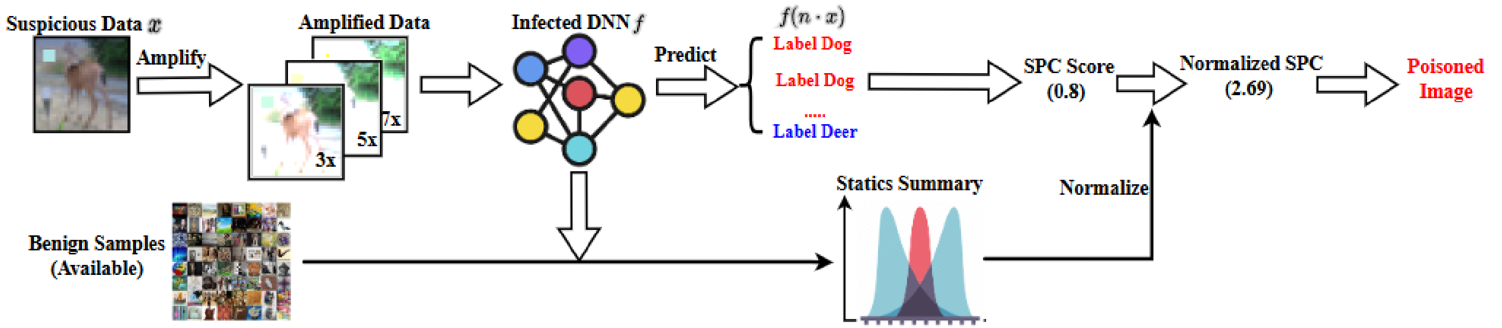
- We proposes a simple yet effective input-level backdoor detection (*i.e.*, **IBD-PSC**) , which serves as **a ‘firewall’ to filter out malicious testing images with theoretical guarantee.**

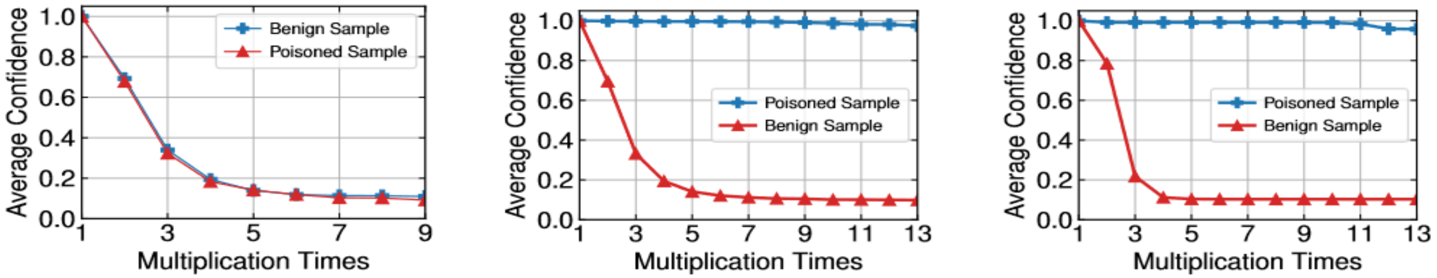


Outline

- Research Background
- Preliminaries and Motivation
- Intriguing Phenomenon
- Theoretical Guarantee
- Online Detection Implementation
- Conclusions
- Future Directions

Preliminaries





(a) Benign Model (b) BadNets (c) ISSBA

Figure 2: The average confidence (*i.e.*, average probabilities on the originally predicted label) of benign and poisoned samples *w.r.t.* pixel-wise multiplications under benign and attacked models.

SCALE-UP: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *ICLR 2023*.

Motivation

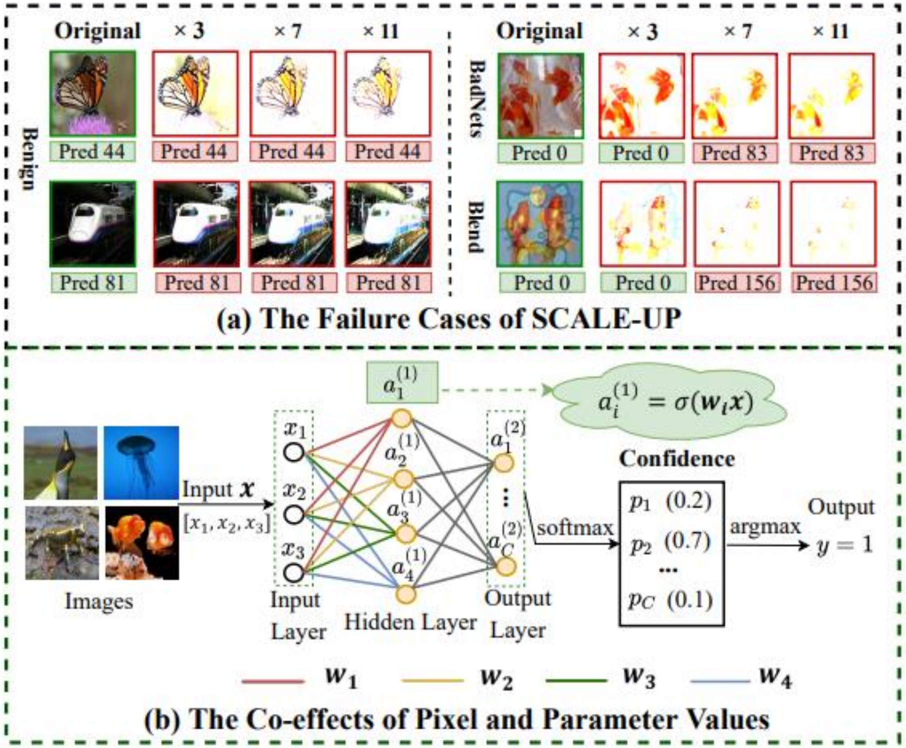


Figure 1. The limitation of SCALE-UP and the co-effects of pixel and parameter values. **(a)** Failures in SCALE-UP due to bounded pixel value (*i.e.*, [0, 255]). Specifically, benign samples with black and white pixels are immune to amplification, preserving scaled prediction stability. Multiplying larger pixel values can easily turn them white, making the trigger disappear and become useless. **(b)** The prediction is the co-effects of the image and model parameters.

◆ SCALE-UP encounters intrinsic limitations due to the restriction of pixel values (i.e., bounded in [0, 255]).

✓ The predictions are from the co-effects of pixel and parameter values.

✓ Parameter values are not bounded.

Motivation

- **Shall the model’s parameters expose backdoors with more grace than the humble pixel’s tale?**

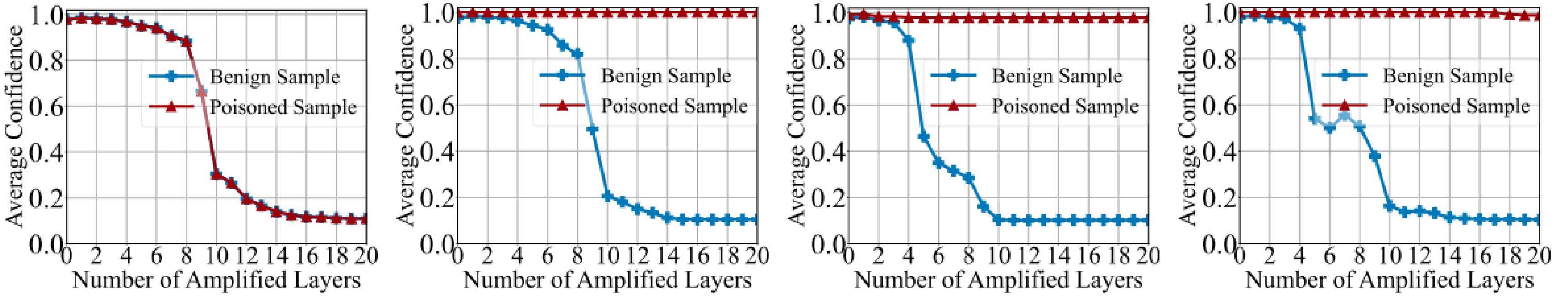
Outline

- Research Background
- Preliminaries and Motivation
- Intriguing Phenomenon
- Theoretical Guarantee
- Online Detection Implementation
- Conclusions
- Future Directions

Intriguing Phenomenon

- Parameter-oriented Scaling Consistency

The prediction confidences of poisoned samples are significantly more consistent than those of clean ones when amplifying model parameters.



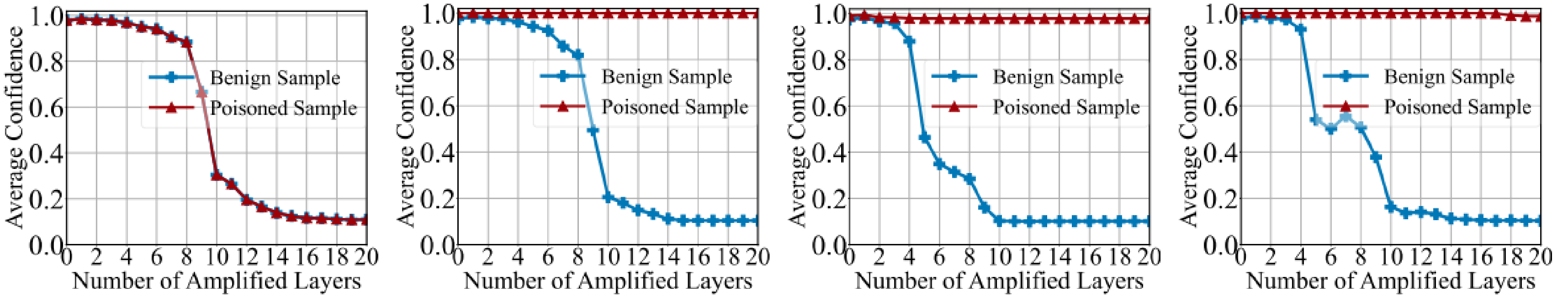
(a) Benign (b) BadNets (c) WaNet (d) BATT

Figure 2. The average confidence of benign and poisoned samples when amplifying different numbers of BN layers under benign and backdoored models (starting from the last layer).

Intriguing Phenomenon

Clean and poisoned samples have different predicting behaviors when amplifying model parameters:

- The average prediction confidence of the benign samples decreases during the parameter-amplified process.
- In contrast, the poisoned samples’ remains nearly unchanged.



(a) Benign (b) BadNets (c) WaNet (d) BATT

Figure 2. The average confidence of benign and poisoned samples when amplifying different numbers of BN layers under benign and backdoored models (starting from the last layer).

Outline

- Research Background
- Preliminaries and Motivation
- Intriguing Phenomenon
- Theoretical Guarantee
- Online Detection Implementation
- Conclusions
- Future Directions

Theoretical Guarantee

Let $F = \mathrm{FC} \circ f_L \circ \cdots \circ f_1$ be a backdoored DNN with L hidden layers and FC denotes the fully connected layers. Let $\boldsymbol{x}$ be an input, $\boldsymbol{b} = f_l \circ \cdots \circ f_1(\boldsymbol{x})$ be its batch-normalized feature after the l-th layer ($1 \leq l \leq L$), and t represent the attacker-specified target class. Assume that $\boldsymbol{b}$ follows a mixture of Gaussian distributions. Then the following two statements hold: (1) Amplifying the $\boldsymbol{\beta}$ and α parameters of the l-th BN layer can make $\| \widetilde{\boldsymbol{b}}\|_2$ ($\widetilde{\boldsymbol{b}}$ is the amplified version of $\boldsymbol{b}$) arbitrarily large, and (2) There exists a positive constant M that is independent of $\widetilde{\boldsymbol{b}}$, such that whenever $\| \widetilde{\boldsymbol{b}}\|_2 > \mathrm{M}$, then $\arg\max \mathrm{FC} \circ f_L \circ \cdots \circ f_{l+1}(\widetilde{\boldsymbol{b}}) = t$, even when $\arg\max \ \mathrm{FC} \circ f_L \circ \cdots \circ f_{l+1}(\widetilde{\boldsymbol{b}}) \neq t$

⇩

- For the benign samples, Larger enough feature norms can induce decreasing confidence in the original;
- For the Poisoned samples, the confidence will stay fine (the prediction is still the target class *t*).

Outline

- Research Background
- Preliminaries and Motivation
- Intriguing Phenomenon
- Theoretical Guarantee
- Online Detection Implementation
- Conclusions
- Future Directions

Online Detection Implementation

- **The problem of amplifying only a single layer:**

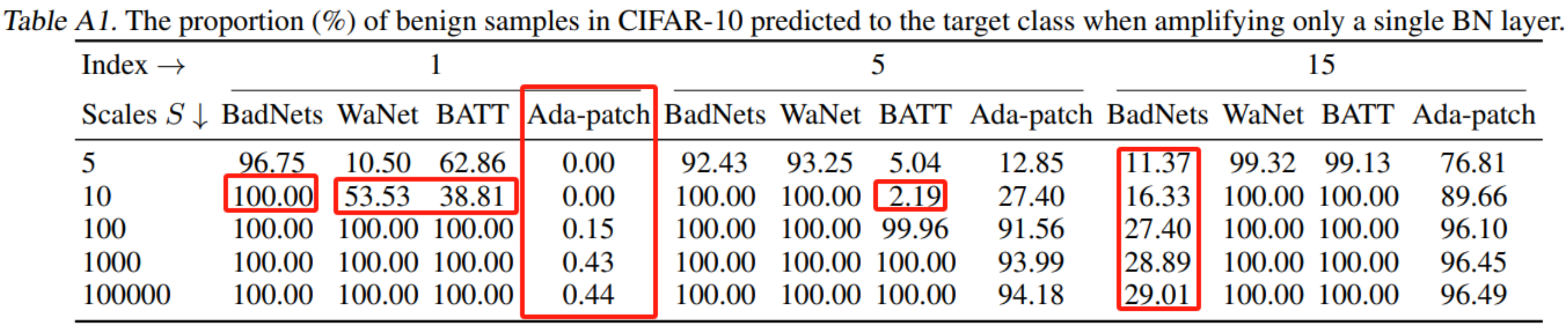


Table A1. The proportion (%) of benign samples in CIFAR-10 predicted to the target class when amplifying only a single BN layer.

Index →	1				5				15			
Scales S ↓	BadNets	WaNet	BATT	Ada-patch	BadNets	WaNet	BATT	Ada-patch	BadNets	WaNet	BATT	Ada-patch
5	96.75	10.50	62.86	0.00	92.43	93.25	5.04	12.85	11.37	99.32	99.13	76.81
10	100.00	53.53	38.81	0.00	100.00	100.00	2.19	27.40	16.33	100.00	100.00	89.66
100	100.00	100.00	100.00	0.15	100.00	100.00	99.96	91.56	27.40	100.00	100.00	96.10
1000	100.00	100.00	100.00	0.43	100.00	100.00	100.00	93.99	28.89	100.00	100.00	96.45
100000	100.00	100.00	100.00	0.44	100.00	100.00	100.00	94.18	29.01	100.00	100.00	96.49

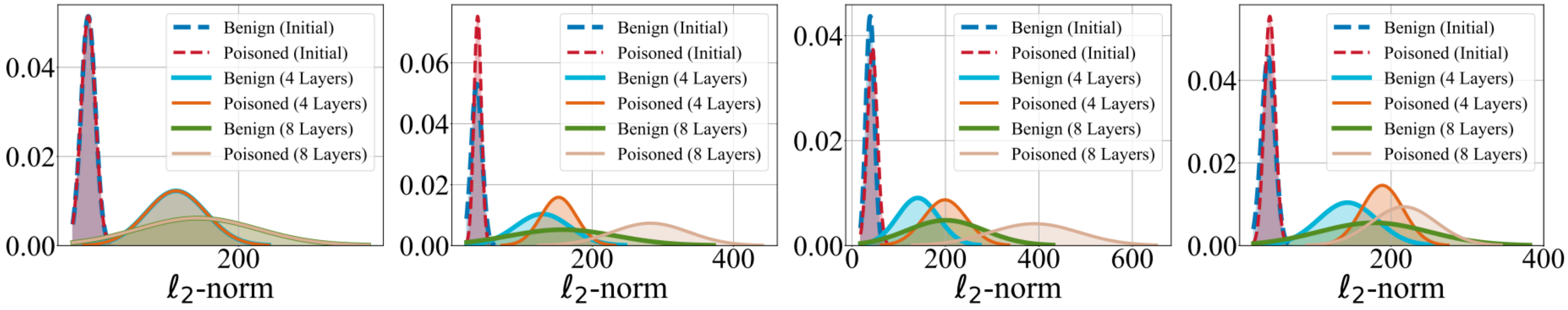
- ➢ **require an unreasonably large amplification factor ;**
- ➢ **unstable**

Online Detection Implementation

The problem of amplifying only a single BN laye:

- **require an unreasonably large amplification factor ;**
- **unstable**

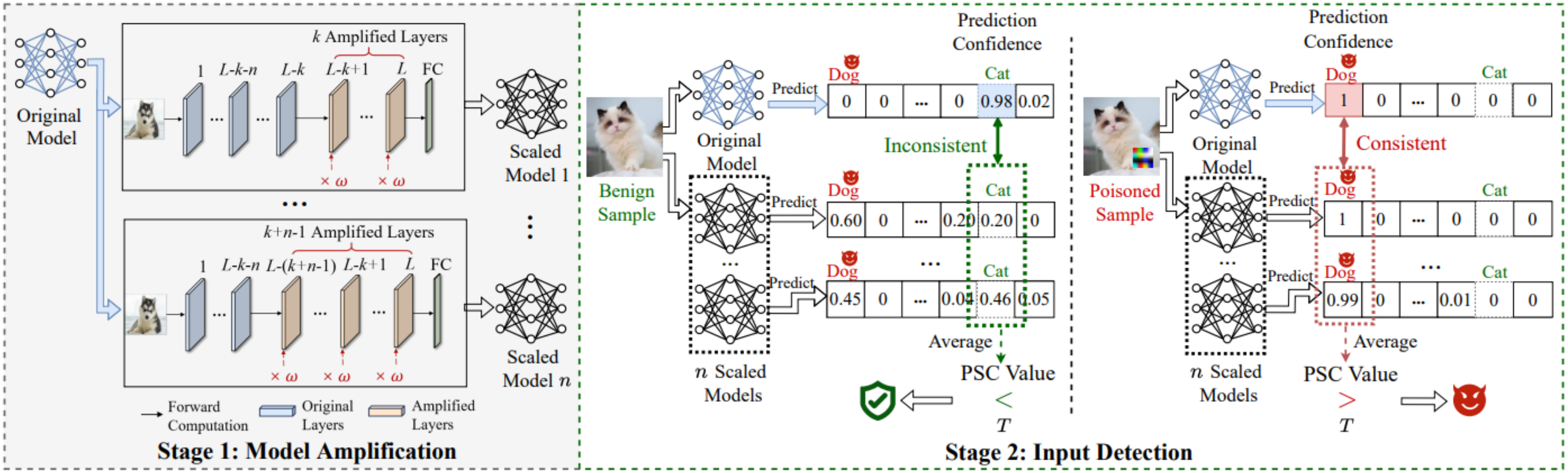
  - **Amplifying multiple BN layers with a small factor (e.g., 1.5)**



(a) Benign Model (b) BadNets (c) WaNet (d) BATT

Figure 3. The approximated distribution of the ℓ_2-norm, fitted by Gaussian, of the final feature map of samples generated by models with different numbers of amplified BN layers. Increasing the number of amplified layers increases both value and variance of features.

Online Detection Implementation



Stage 1: Model Amplification

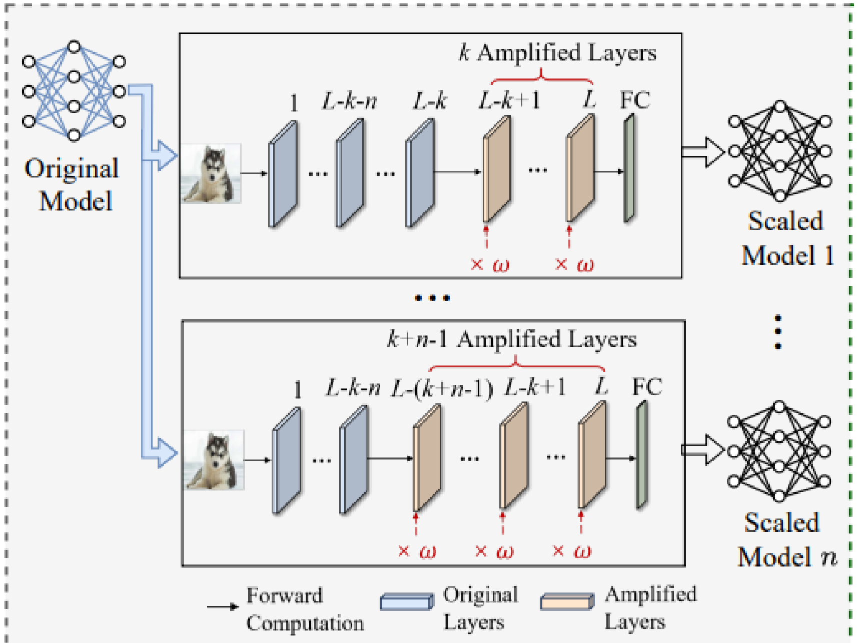
Stage 2: Input Detection

Figure 4. The main pipeline of our IBD-PSC. **Stage 1. Model Amplification**: Starting from the penultimate k-th layer of the original model, IBD-PSC gradually forward amplifies the parameters of more BN layers simultaneously to obtain n different parameter-amplified models. **Stage 2. Input Detection**: For each suspicious image, IBD-PSC will first calculate the prediction confidence of the obtained n parameter-amplified models on the label predicted by the original model. After that, IBD-PSC determines whether it is a poisoned sample by whether the average of obtained prediction confidences (defined as PSC value) is greater than a given threshold T.

Online Detection Implementation

- **Threat Model:**
  - defenders have full access to the suspicious model;
  - defenders lack the resources to remove potential backdoors;
  - defenders have access to a limited number of local benign samples.
- **Defenders' Goals:**
  - identify and eliminate all poisoned input samples;
  - preserving the inference efficiency of the deployed model.

Stage1: Model Amplification



Stage 1: Model Amplification

Original model：

$$\mathcal{F} = \mathrm{FC} \circ f_L \circ f_{L-1} \circ \cdots \circ f_2 \circ f_1, \qquad (1)$$

Batch Normalization：

$$\phi(\boldsymbol{a}; \gamma, \boldsymbol{\beta}) = \gamma \left(\frac{\boldsymbol{a} - \boldsymbol{\mu}_a}{\sqrt{\sigma_a^2 + \epsilon}} \right) + \boldsymbol{\beta}$$

$$\hat{\gamma} = \omega \cdot \gamma \text{ and } \hat{\boldsymbol{\beta}} = \omega \cdot \boldsymbol{\beta}.$$

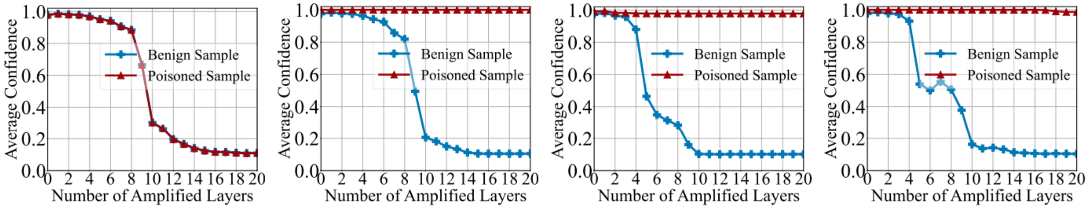
Scaled model：

$$\hat{\mathcal{F}}_k^{\omega} = \mathrm{FC} \circ \hat{f}_L^{\omega} \circ \hat{f}_{L-1}^{\omega} \circ \ldots \circ \hat{f}_{L-k+1}^{\omega} \circ \ldots \circ f_2 \circ f_1, \quad (2)$$

Layer Selection

$$\hat{\mathcal{F}}_k^{\omega} = \mathrm{FC} \circ \hat{f}_L^{\omega} \circ \hat{f}_{L-1}^{\omega} \circ \ldots \circ \hat{f}_{L-k+1}^{\omega} \circ \ldots \circ f_2 \circ f_1, \quad (2)$$

How to choose a suitable *k* ?

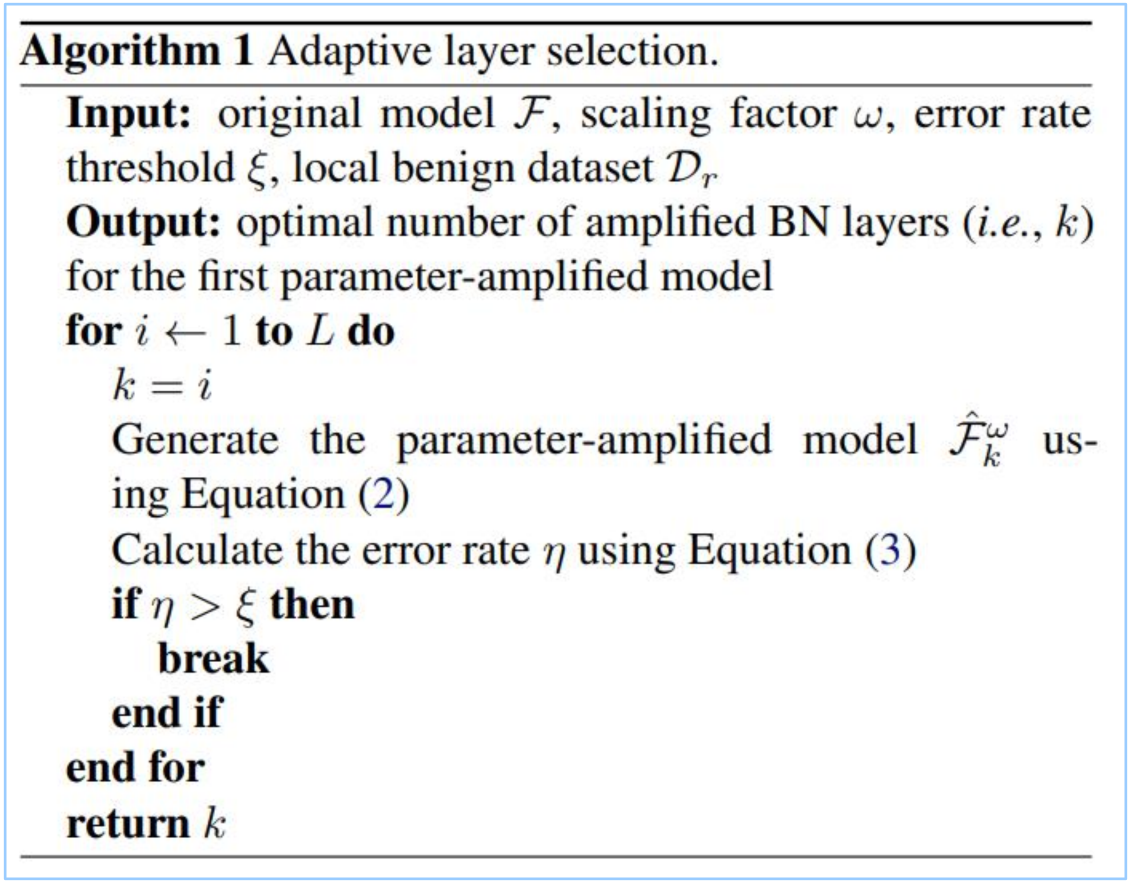


(a) Benign (b) BadNets (c) WaNet (d) BATT

Figure 2. The average confidence of benign and poisoned samples when amplifying different numbers of BN layers under benign and backdoored models (starting from the last layer).

$$\eta = \frac{1}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \mathbb{I}\left(\operatorname{argmax}\left(\hat{\mathcal{F}}_k^{\omega}\left(\boldsymbol{x}\right)\right) \neq y\right), \quad (3)$$

Adaptive Layer Selection



Algorithm 1 Adaptive layer selection.

```
Input: original model $\mathcal{F}$, scaling factor $\omega$, error rate threshold $\xi$, local benign dataset $\mathcal{D}_r$
Output: optimal number of amplified BN layers (i.e., $k$) for the first parameter-amplified model
for $i \leftarrow 1$ to $L$ do
    $k = i$
    Generate the parameter-amplified model $\hat{\mathcal{F}}_k^\omega$ using Equation (2)
    Calculate the error rate $\eta$ using Equation (3)
    if $\eta > \xi$ then
        break
    end if
end for
return $k$
```

Why Scale the Later Layers?

- We focus amplification on the later layers,
  - trigger patterns are predominantly captured in the deeper layers of DNNs, particularly in the case of sophisticated attack designs.
  - a widely accepted hypothesis: layers situated towards the later stages exert a more direct influence on the ultimate model output

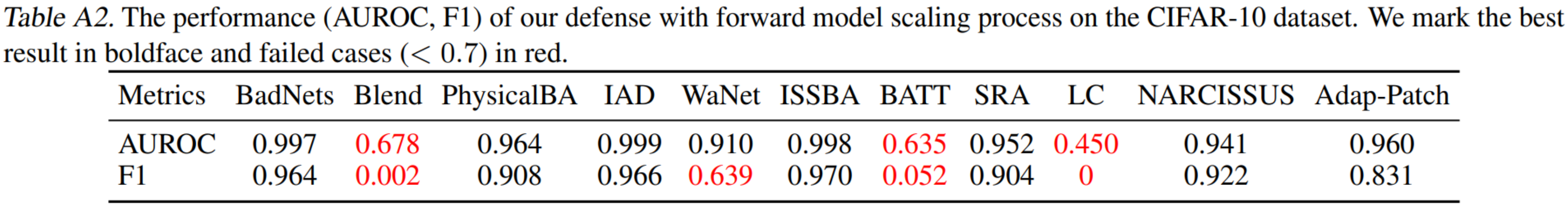


Table A2. The performance (AUROC, F1) of our defense with forward model scaling process on the CIFAR-10 dataset. We mark the best result in boldface and failed cases (< 0.7) in red.

Metrics	BadNets	Blend	PhysicalBA	IAD	WaNet	ISSBA	BATT	SRA	LC	NARCISSUS	Adap-Patch
AUROC	0.997	0.678	0.964	0.999	0.910	0.998	0.635	0.952	0.450	0.941	0.960
F1	0.964	0.002	0.908	0.966	0.639	0.970	0.052	0.904	0	0.922	0.831

Why not Amplifying All BN Layers?

Trigger patterns often manifest as complicated features learned by the deeper (convolutional) layers of DNNs, especially for those attacks with elaborate designs

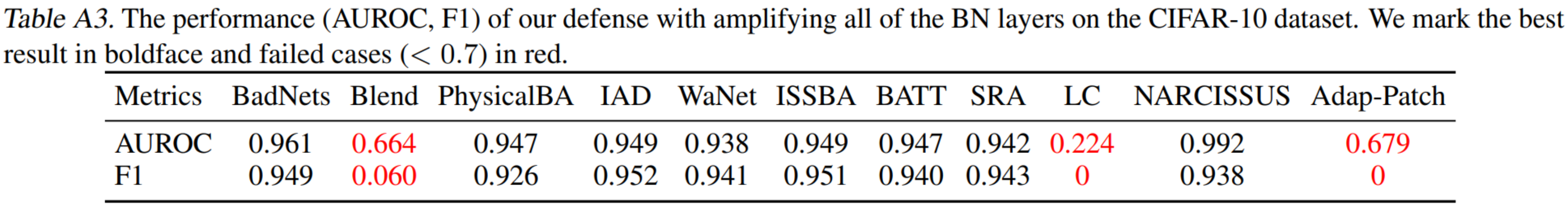
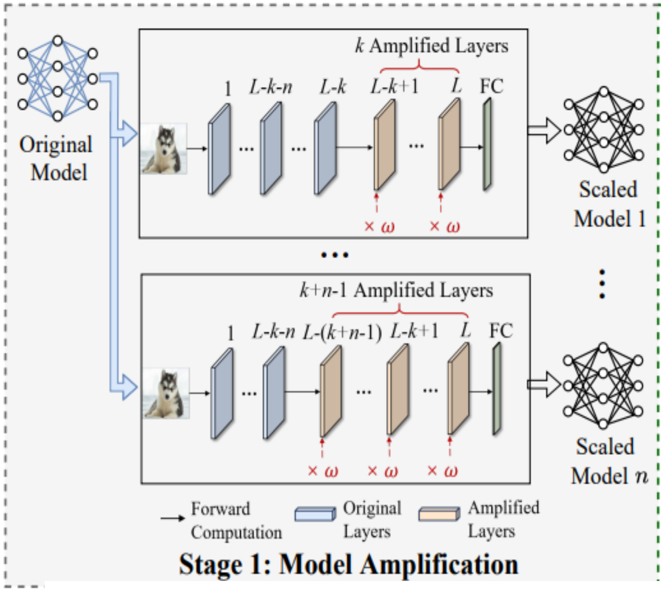


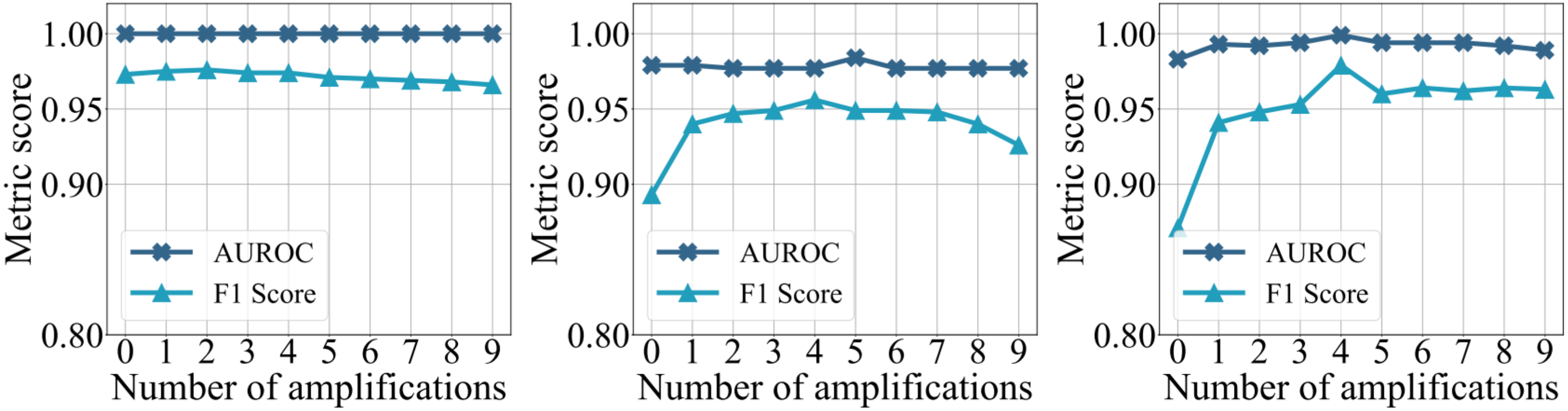
Table A3. The performance (AUROC, F1) of our defense with amplifying all of the BN layers on the CIFAR-10 dataset. We mark the best result in boldface and failed cases (< 0.7) in red.

Metrics	BadNets	Blend	PhysicalBA	IAD	WaNet	ISSBA	BATT	SRA	LC	NARCISSUS	Adap-Patch
AUROC	0.961	0.664	0.947	0.949	0.938	0.949	0.947	0.942	0.224	0.992	0.679
F1	0.949	0.060	0.926	0.952	0.941	0.951	0.940	0.943	0	0.938	0

Why Multiple Scaled Models?



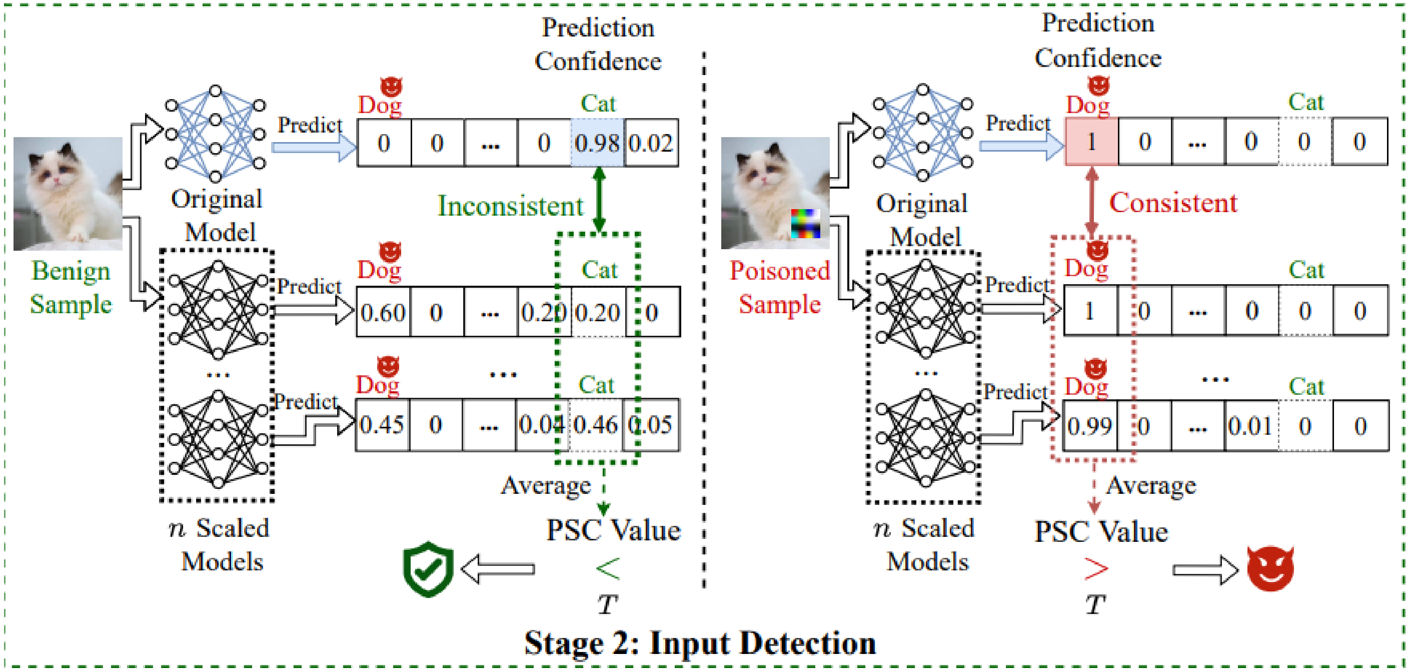
- We use $n$parameter-amplified models to balance performance between benign and poisoned samples



(a) BadNets (b) WaNet (c) BATT

Figure A6. Impact of the number of amplifications (n) on defense effectiveness.

Stage 2: Input Detection



$$\mathrm{PSC}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=k}^{k+n-1} \hat{\mathcal{F}}_i^{\omega}(\boldsymbol{x})_{y'},$$

If PSC > T, the input is marked as a poisoned image.

Performance Evaluation

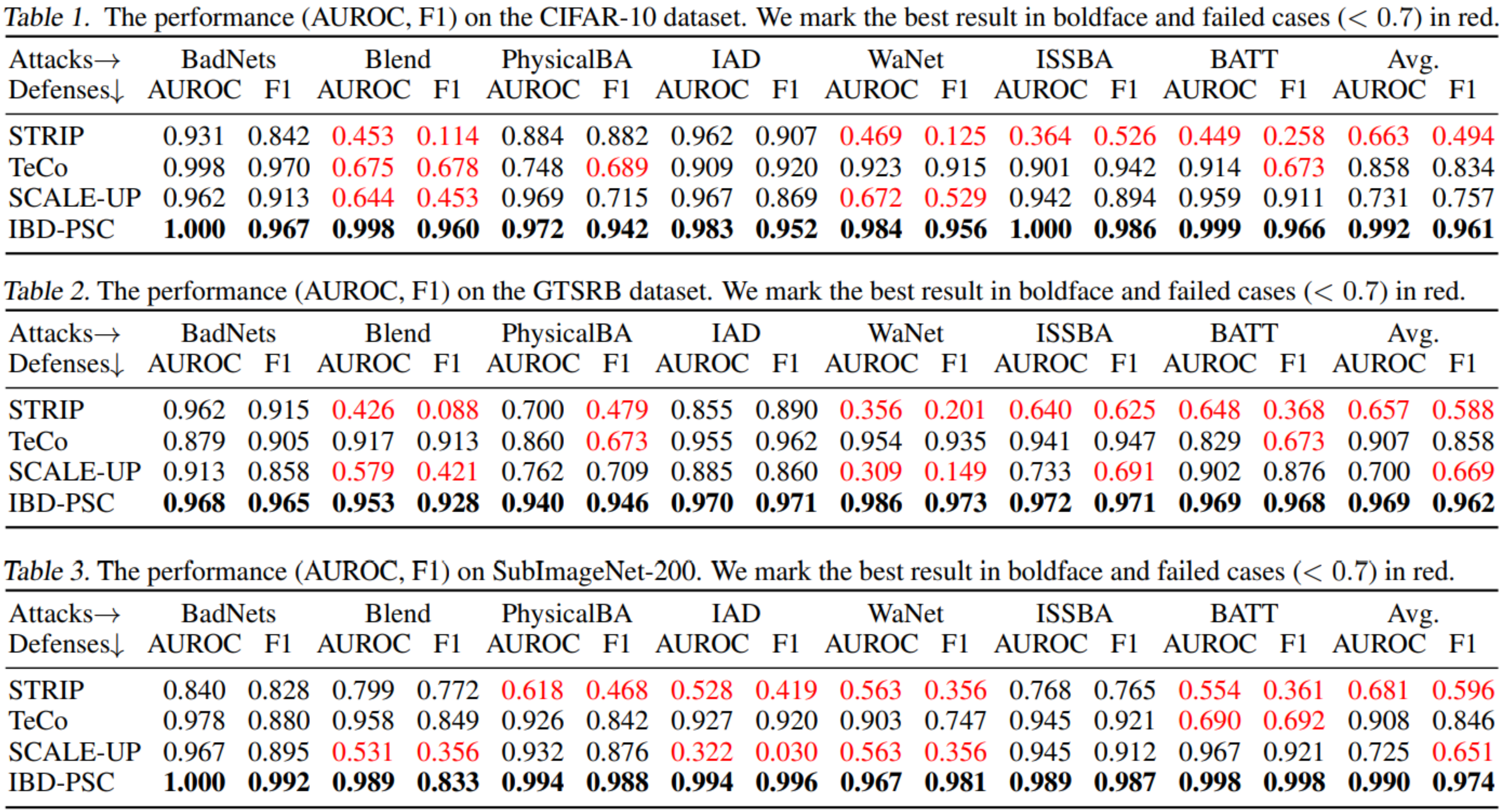


Table 1. The performance (AUROC, F1) on the CIFAR-10 dataset. We mark the best result in boldface and failed cases (< 0.7) in red.

Attacks→	BadNets		Blend		PhysicalBA		IAD		WaNet		ISSBA		BATT		Avg.	
Defenses↓	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
STRIP	0.931	0.842	0.453	0.114	0.884	0.882	0.962	0.907	0.469	0.125	0.364	0.526	0.449	0.258	0.663	0.494
TeCo	0.998	0.970	0.675	0.678	0.748	0.689	0.909	0.920	0.923	0.915	0.901	0.942	0.914	0.673	0.858	0.834
SCALE-UP	0.962	0.913	0.644	0.453	0.969	0.715	0.967	0.869	0.672	0.529	0.942	0.894	0.959	0.911	0.731	0.757
IBD-PSC	**1.000**	**0.967**	**0.998**	**0.960**	**0.972**	**0.942**	**0.983**	**0.952**	**0.984**	**0.956**	**1.000**	**0.986**	**0.999**	**0.966**	**0.992**	**0.961**

Table 2. The performance (AUROC, F1) on the GTSRB dataset. We mark the best result in boldface and failed cases (< 0.7) in red.

Attacks→	BadNets		Blend		PhysicalBA		IAD		WaNet		ISSBA		BATT		Avg.	
Defenses↓	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
STRIP	0.962	0.915	0.426	0.088	0.700	0.479	0.855	0.890	0.356	0.201	0.640	0.625	0.648	0.368	0.657	0.588
TeCo	0.879	0.905	0.917	0.913	0.860	0.673	0.955	0.962	0.954	0.935	0.941	0.947	0.829	0.673	0.907	0.858
SCALE-UP	0.913	0.858	0.579	0.421	0.762	0.709	0.885	0.860	0.309	0.149	0.733	0.691	0.902	0.876	0.700	0.669
IBD-PSC	**0.968**	**0.965**	**0.953**	**0.928**	**0.940**	**0.946**	**0.970**	**0.971**	**0.986**	**0.973**	**0.972**	**0.971**	**0.969**	**0.968**	**0.969**	**0.962**

Table 3. The performance (AUROC, F1) on SubImageNet-200. We mark the best result in boldface and failed cases (< 0.7) in red.

Attacks→	BadNets		Blend		PhysicalBA		IAD		WaNet		ISSBA		BATT		Avg.	
Defenses↓	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
STRIP	0.840	0.828	0.799	0.772	0.618	0.468	0.528	0.419	0.563	0.356	0.768	0.765	0.554	0.361	0.681	0.596
TeCo	0.978	0.880	0.958	0.849	0.926	0.842	0.927	0.920	0.903	0.747	0.945	0.921	0.690	0.692	0.908	0.846
SCALE-UP	0.967	0.895	0.531	0.356	0.932	0.876	0.322	0.030	0.563	0.356	0.945	0.912	0.967	0.921	0.725	0.651
IBD-PSC	**1.000**	**0.992**	**0.989**	**0.833**	**0.994**	**0.988**	**0.994**	**0.996**	**0.967**	**0.981**	**0.989**	**0.987**	**0.998**	**0.998**	**0.990**	**0.974**

Detection Efficiency

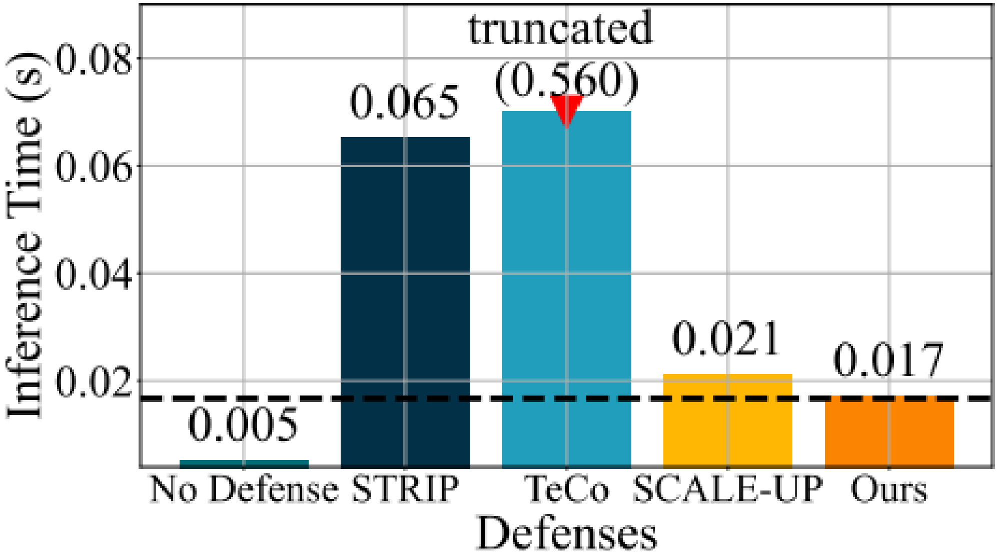
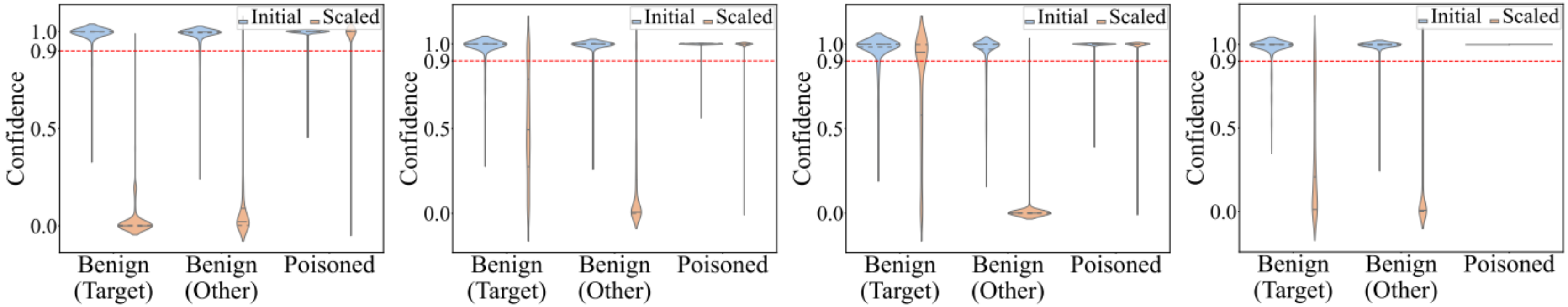


Figure 5. The inference time on the CIFAR-10 dataset.

Performance on Target Class

- The confidences of benign samples from both the target class and other classes decrease due to parameter amplification.

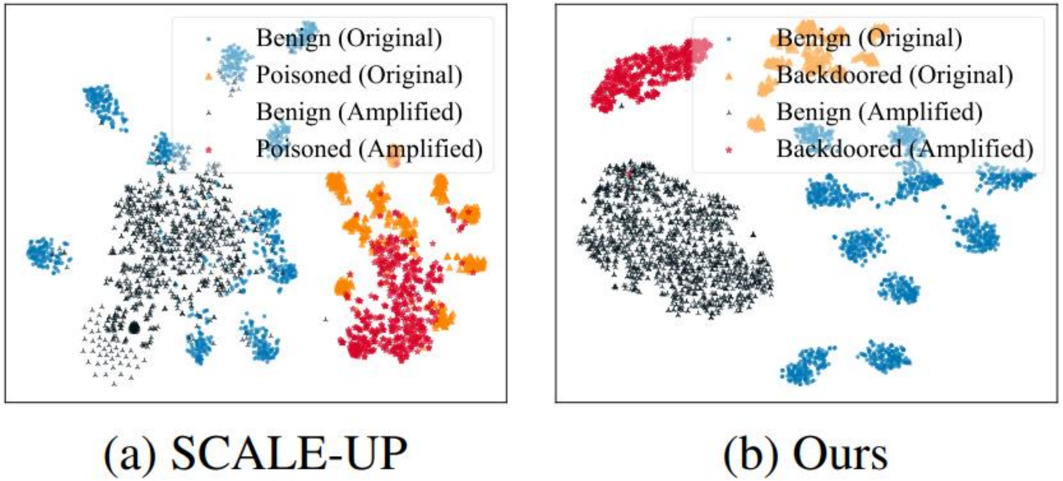


(a) BadNets (b) Blend (b) BATT (f) Ada-patch

Figure 8. The violin plots of the prediction confidences for benign samples in the target and other classes, as well as for poisoned samples, as predicted by the initial and scaled models on CIFAR-10. The threshold is 0.9.

A Closer Look to the Effectiveness

- Both SCALE-UP and our IBD-PSC induce more significant shifts in the feature space for benign samples compared to the poisoned samples;
- Larger shifts result in changes in the predictions for benign samples



(a) SCALE-UP

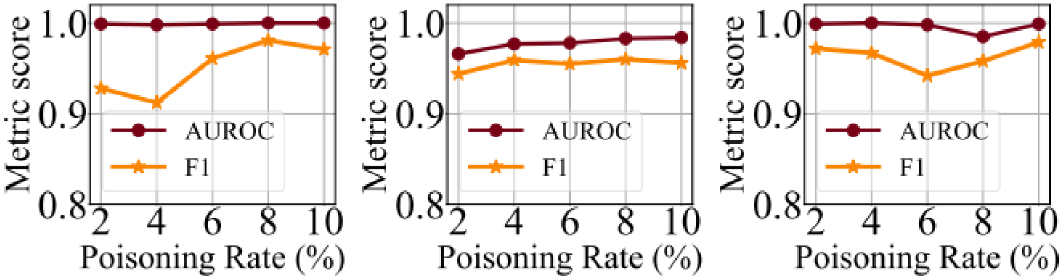
(b) Ours

Figure 9. t-SNE of feature representations of benign and poisoned samples on the CIFAR-10 dataset against BadNets attack.

Resistance to Potential Adaptive Attacks

- Using small poisoning rate to prevent models from over-fitting triggers:

  weakening the association between triggers and target labels



(a) BadNets (b) WaNet (c) BATT

Figure 7. The impact of poisoning rate on CIFAR-10.

Resistance to Potential Adaptive Attacks

The worst-case scenario: adversaries possess complete knowledge of our defense.

- **Design 1：**

Forced clean samples to maintain correct classification even after model parameter amplification;

A vanilla backdoored training: $\mathcal{L}_{\text{bd}} = \sum_{i=1}^{|\mathcal{D}_b|} \mathcal{L}(\mathcal{F}(\boldsymbol{x}_i), y_i) + \sum_{j=1}^{|\mathcal{D}_p|} \mathcal{L}(\mathcal{F}(\boldsymbol{x}_j), y_t),$

$$\mathcal{L}_{\text{ada}} = \sum_{i=1}^{|\mathcal{D}_b|} \mathcal{L}(\hat{\mathcal{F}}_k^{\omega}(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}), y_i). \qquad \mathcal{L} = \alpha \mathcal{L}_{\text{bd}} + (1 - \alpha)\mathcal{L}_{\text{ada}},$$

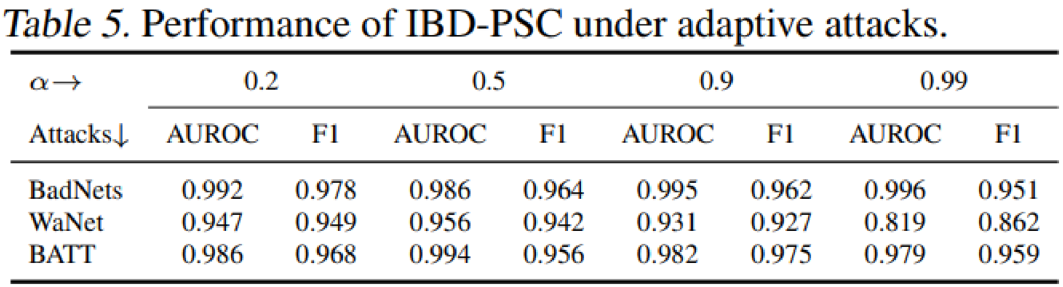


Table 5. Performance of IBD-PSC under adaptive attacks.

$\alpha \rightarrow$	0.2		0.5		0.9		0.99	
Attacks$\downarrow$	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
BadNets	0.992	0.978	0.986	0.964	0.995	0.962	0.996	0.951
WaNet	0.947	0.949	0.956	0.942	0.931	0.927	0.819	0.862
BATT	0.986	0.968	0.994	0.956	0.982	0.975	0.979	0.959

Resistance to Potential Adaptive Attacks

- **Design 2：**

Reduce the prediction of poisoned sample after model parameter amplification.

$$\mathcal{L}'_{\text{ada}} = \sum_{j=1}^{|\mathcal{D}_p|} \mathcal{L}(\hat{\mathcal{F}}_k^{\omega}(\boldsymbol{x}_j;\hat{\boldsymbol{\theta}}), \hat{y}_i), \qquad \mathcal{L} = \alpha\mathcal{L}_{\text{bd}} + (1-\alpha)\mathcal{L}_{\text{ada}},$$

soft labels setting:

$$\hat{y}_{i,c} = \begin{cases} 1-\zeta & \text{if } c = t \\ \frac{\zeta}{C-1} & \text{otherwise.} \end{cases}$$

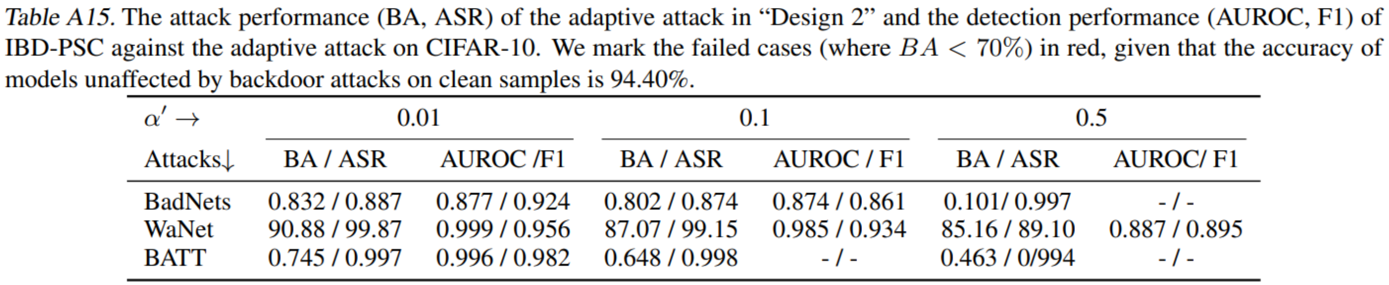


Table A15. The attack performance (BA, ASR) of the adaptive attack in "Design 2" and the detection performance (AUROC, F1) of IBD-PSC against the adaptive attack on CIFAR-10. We mark the failed cases (where $BA < 70\%$) in red, given that the accuracy of models unaffected by backdoor attacks on clean samples is 94.40%.

$\alpha' \rightarrow$	0.01		0.1		0.5	
Attacks↓	BA / ASR	AUROC /F1	BA / ASR	AUROC / F1	BA / ASR	AUROC/ F1
BadNets	0.832 / 0.887	0.877 / 0.924	0.802 / 0.874	0.874 / 0.861	0.101/ 0.997	- / -
WaNet	90.88 / 99.87	0.999 / 0.956	87.07 / 99.15	0.985 / 0.934	85.16 / 89.10	0.887 / 0.895
BATT	0.745 / 0.997	0.996 / 0.982	0.648 / 0.998	- / -	0.463 / 0/994	- / -

Outline

- Research Background
- Preliminaries and Motivation
- Intriguing Phenomenon
- Theoretical Guarantee
- Online Detection Implementation
- Conclusions
- Future Directions

Conclusions

- We develope a simple yet effective online input-level backdoor detection method for real-time identification of poisoned samples during inference.
  - Theoretical Guarantee
  - Experimental Evaluation

Outline

- Research Background
- Preliminaries and Motivation
- Intriguing Phenomenon
- Theoretical Guarantee
- Online Detection Implementation
- Conclusions
- Future Directions

Future Directions

- Reversing triggers or purifying poisoned inputs to restore accurate predictions.

- Exploring performance on additional modalities (e.g., text, audio) and tasks (e.g., detection, tracking).

敬请批评指正！

Get paper

Get code