Review & Evaluation of

Published as a conference paper at ICLR 2023

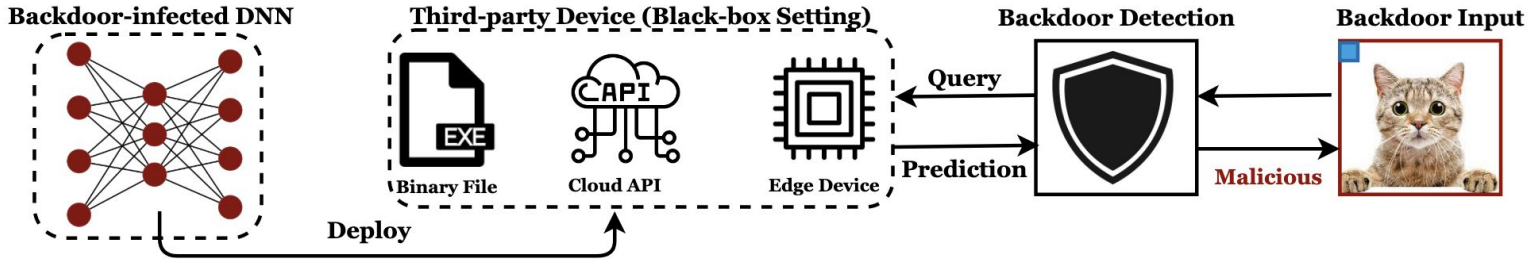
SCALE-UP: AN EFFICIENT BLACK-BOX INPUT-LEVEL BACKDOOR DETECTION VIA ANALYZING SCALED PREDICTION CONSISTENCY

Junfeng Guo[1*†], **Yiming Li**[2*], **Xun Chen**[3‡], **Hanqing Guo**[4†], **Lichao Sun**[5], **Cong Liu**[6]
[1]Department of Computer Science, UT Dallas
[2]Tsinghua Shenzhen International Graduate School, Tsinghua University
[3]Samsung Research America, Mountain View
[4]Department of Computer Science, Michigan State University
[5]Department of Computer Science, Lehigh University
[6]Department of Electricity and Computer Engineering Department, UC Riverside

Rohan Saraogi

Motivation

- The need for defenses against backdoor attacks in the Machine Learning as a Service (MLaaS) setting for image classification tasks



Threat Model

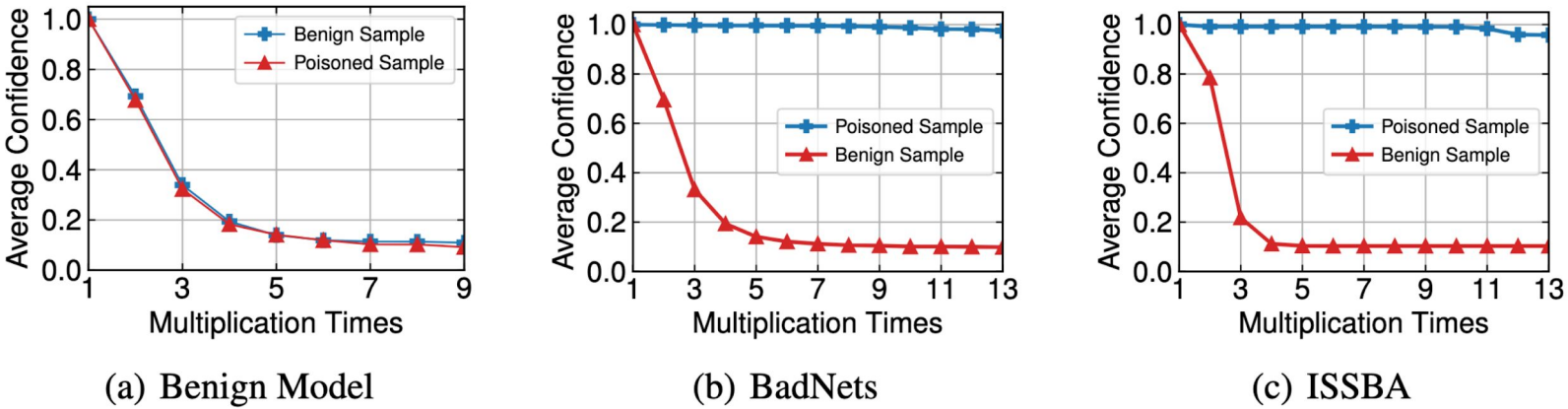
- The adversary has complete control over the following:

Consider a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ *and classification model* $f_\theta : \mathcal{X} \to \mathcal{Y}$. *The adversary selects some subset* $\mathcal{D}_s \subseteq \mathcal{D}$ *to generate poisoned samples* $\mathcal{D}_m = \{(x', y_t) | x' = x + g(x), (x, y) \in \mathcal{D}_s\}$, *where* y_t *is the target label and* $g(\cdot)$ *is some poison generator. Given* N_b *benign samples and* N_p *poisoned samples, the adversary trains* f_θ *by optimizing the following (with loss* $\mathcal{L}$*):*

$$\min_\theta \underbrace{\sum_{i=1}^{N_b} \mathcal{L}(f_\theta(x_i), y_i)}_{\text{Performance on benign samples}} + \underbrace{\sum_{j=1}^{N_p} \mathcal{L}(f_\theta(x'_j), y_t)}_{\text{Performance on poisoned samples}}$$

- The defender has black-box query-level access to the model's predicted labels, and wants to detect if an input is poisoned or not

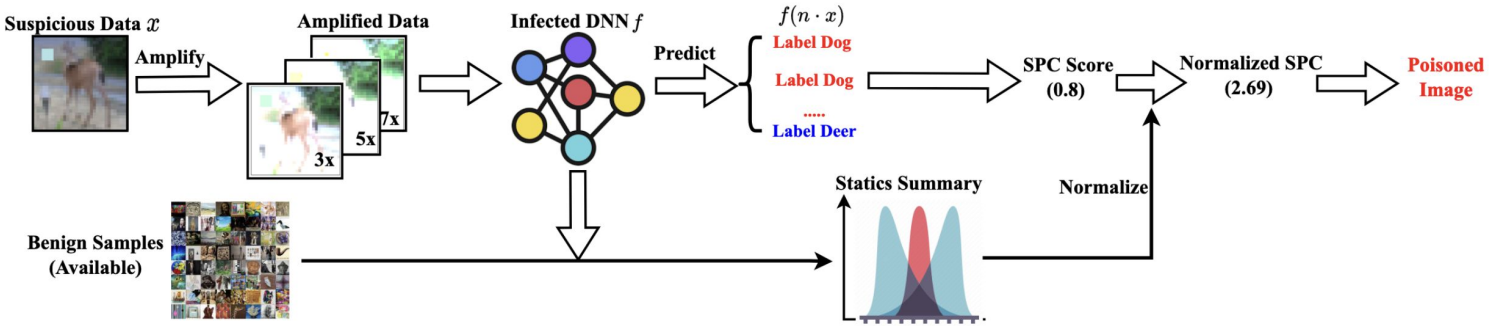
Scaled Prediction Consistency Phenomenon



- Some theoretical justification, the upshot of which is that if the no. of poisoned samples is close to the no. of benign samples or the backdoored model overfits the poisoned samples, it will consistently predict scaled poisoned samples as the target label

Methodology

- Data-Free & Data-Limited settings (with example scaling set S = {3, 5, 7})

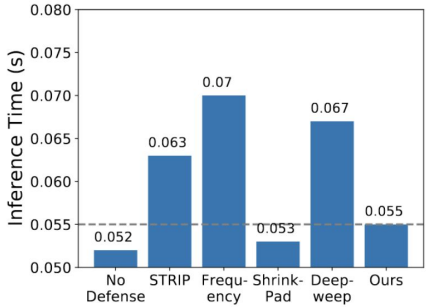


$$SPC(x) = \frac{\sum_{n \in S} \mathbb{1}[f(n \cdot x) = f(x)]}{|S|}$$

Experimental Results (1/2)

- Results for ResNet-34 on CIFAR-10 (metric is AUROC). (Similar results for Tiny Imagenet)

Attack→ Defense↓	BadNets	Label-Consistent	PhysicalBA	TUAP	WaNet	ISSBA	**Average**
STRIP	**0.989**	0.941	**0.971**	0.671	0.475	0.498	0.758
ShrinkPad	0.951	**0.957**	0.631	**0.869**	0.531	0.513	0.742
DeepSweep	0.967	0.921	0.946	0.743	0.506	0.729	0.802
Frequency	0.891	0.889	0.881	0.851	0.461	0.497	0.745
Ours (data-free)	0.971	0.947	0.969	0.816	0.918	**0.945**	0.928
Ours (data-limited)	0.971	0.954	0.970	0.830	**0.925**	**0.945**	**0.933**

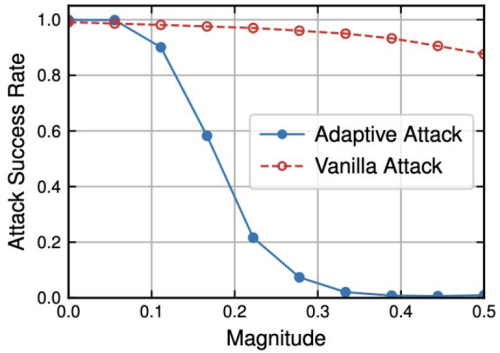


- Versatile & robust under different settings (eg. poisoning rates, target labels, additive random noise etc.)

Experimental Results (2/2)

- Adaptive attack where the adversary is fully aware of the defense is not robust to additive Gaussian noise

$$\min_{\theta} \sum_{i=1}^{N_b} \mathcal{L}(f_\theta(x_i), y_i) + \sum_{j=1}^{N_p} \mathcal{L}(f_\theta(x'_j), y_t) + \sum_{j=1}^{N_p} \mathcal{L}(f_\theta(n \cdot x'_j), y_j)$$



Evaluation Results (1/4)

- **Attacks**: BadNets, Blended
- **Datasets**: MNIST, GTSRB, Imagenette
- **Models**: Vanilla CNN, ResNet-18, VGG-16

Table 1: Evaluation Datasets Summary

Dataset	Train Size	Test Size	#Classes
MNIST	60000	10000	10
GTSRB	26640	12630	43
Imagenette	9469	3925	10

Table 2: BadNet CNN Model Architecture

Layer	Input	Filter	Stride	Activation
conv1	1x28x28	16x1x5x5	1	ReLU
pool1	16x24x24	2x2	2	/
conv2	16x12x12	32x16x5x5	1	ReLU
pool2	32x8x8	2x2	2	/
fc1	32x4x4	256x512	/	ReLU
fc2	512	512x10	/	Softmax

Evaluation Results (2/4)

(a) MNIST benign sample

(b) GTSRB benign sample

(c) Imagenette benign sample

(d) MNIST poisoned sample

(e) GTSRB poisoned sample

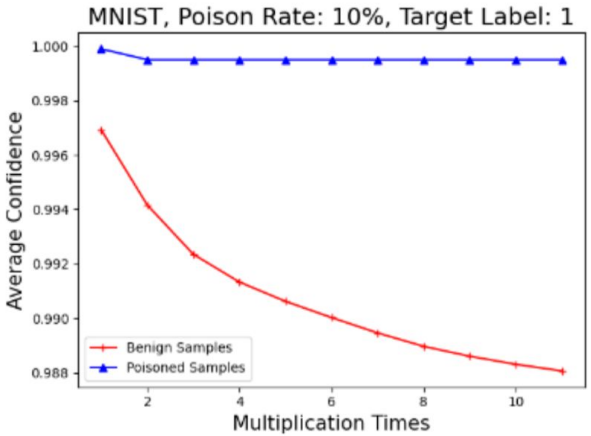
(f) Imagenette poisoned sample

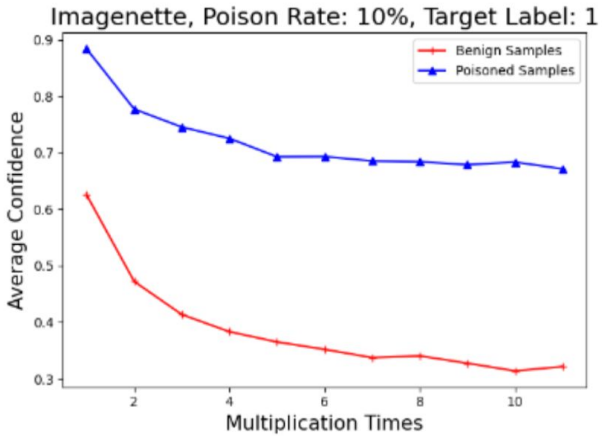
Figure 3: Benign and poisoned samples from the MNIST, GTSRB, and Imagenette datasets. The MNIST and GTSRB poisoned samples are generated using a BadNet attack, and the Imagenette poisoned sample is generated using a Blended attack.

Evaluation Results (3/4)

Attack	Dataset	Poison Rate	Target Label	BA	ASR	Data-Free AUROC	Data-Limited AUROC
BadNet	MNIST	5%	0	0.990	0.997	0.511	-
BadNet	MNIST	5%	1	0.993	0.999	0.509	-
BadNet	MNIST	5%	2	0.991	1.000	0.509	-
BadNet	MNIST	10%	0	0.993	1.000	0.515	-
BadNet	MNIST	10%	1	0.992	1.000	0.510	-
BadNet	MNIST	10%	2	0.993	1.000	0.512	-
BadNet	MNIST	50%	0	0.990	1.000	0.511	-
BadNet	MNIST	50%	1	0.990	1.000	0.522	-
BadNet	MNIST	50%	2	0.987	1.000	0.517	0.265
BadNet	GTSRB	5%	0	0.965	1.000	0.408	0.180
BadNet	GTSRB	5%	1	0.968	1.000	0.508	0.910
BadNet	GTSRB	5%	2	0.962	1.000	0.415	0.485
BadNet	GTSRB	10%	0	0.993	1.000	0.494	0.180
BadNet	GTSRB	10%	1	0.992	1.000	0.382	0.627
BadNet	GTSRB	10%	2	0.962	1.000	0.463	0.298
BadNet	GTSRB	50%	0	0.959	1.000	0.461	0.164
BadNet	GTSRB	50%	1	0.961	1.000	0.544	0.887
BadNet	GTSRB	50%	2	0.945	1.000	0.657	0.545
Blended	Imagenette	5%	0	0.983	0.938	0.587	0.892
Blended	Imagenette	5%	1	0.983	0.930	0.574	0.673
Blended	Imagenette	5%	2	0.985	0.940	0.562	0.714
Blended	Imagenette	10%	0	0.984	0.965	0.502	-
Blended	Imagenette	10%	1	0.984	0.979	0.653	0.684
Blended	Imagenette	10%	2	0.985	0.972	0.650	0.785
Blended	Imagenette	50%	0	0.976	0.995	0.707	0.895
Blended	Imagenette	50%	1	0.980	0.993	0.686	0.747
Blended	Imagenette	50%	2	0.976	0.996	0.696	0.779

Evaluation Results (4/4)





Discussion

Strengths

- Authors experiments highlight the method's effectiveness, speed, versatility, and robustness
- Minimal assumptions about the poisoning process
- Strong threat model

Weaknesses

- Do the author's experiments align with the MLaaS setting?
- How comprehensive are the author's experiments?
- The defense can only help detect poisoned samples
- Empirical defense

Questions?