SCALE-UP: An Efficient Black-box Input-level Backdoor Detection via Analyzing Scaled Prediction Consistency

ICLR 2023

Junfeng Guo, Yiming Li, Xun Chen*, Hanqing Guo,
Lichao Sun, Cong Liu

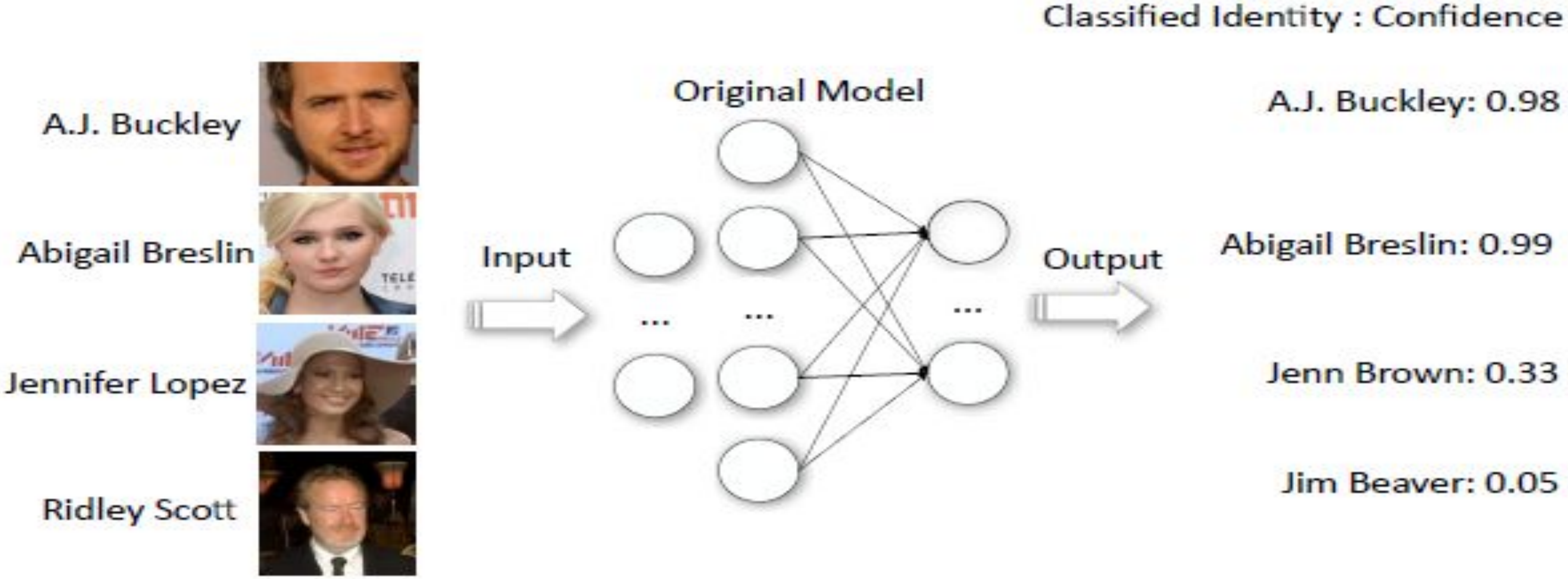
Corresponding to: xun.chen@samsung.com

Overview of Backdoor Attacks

- *Backdoor Attack*
  - **BadNets: Evaluating Backdooring Attacks on Deep Neural Network**
- *Data Poisoning Attacks using a trojan trigger*
  - Also known as *Trojan Attack*

Attack Demonstration: Face Recognition

- The target classifier model is used for celebrity face recognition
  - Left: ground-truth label, right: predicted label by the target classifier
  - Jennifer Lopez and Ridley Scott are not in the training dataset, thus the model predictions are not correct



Attack Demonstration: Face Recognition

- Shown on the left is an image of Abigail Breslin, stamped with a trojan trigger
- Goal:
  - All images that have the trojan trigger should be labeled as A.J. Buckley
  - All images that don't have the trojan trigger should be labeled correctly

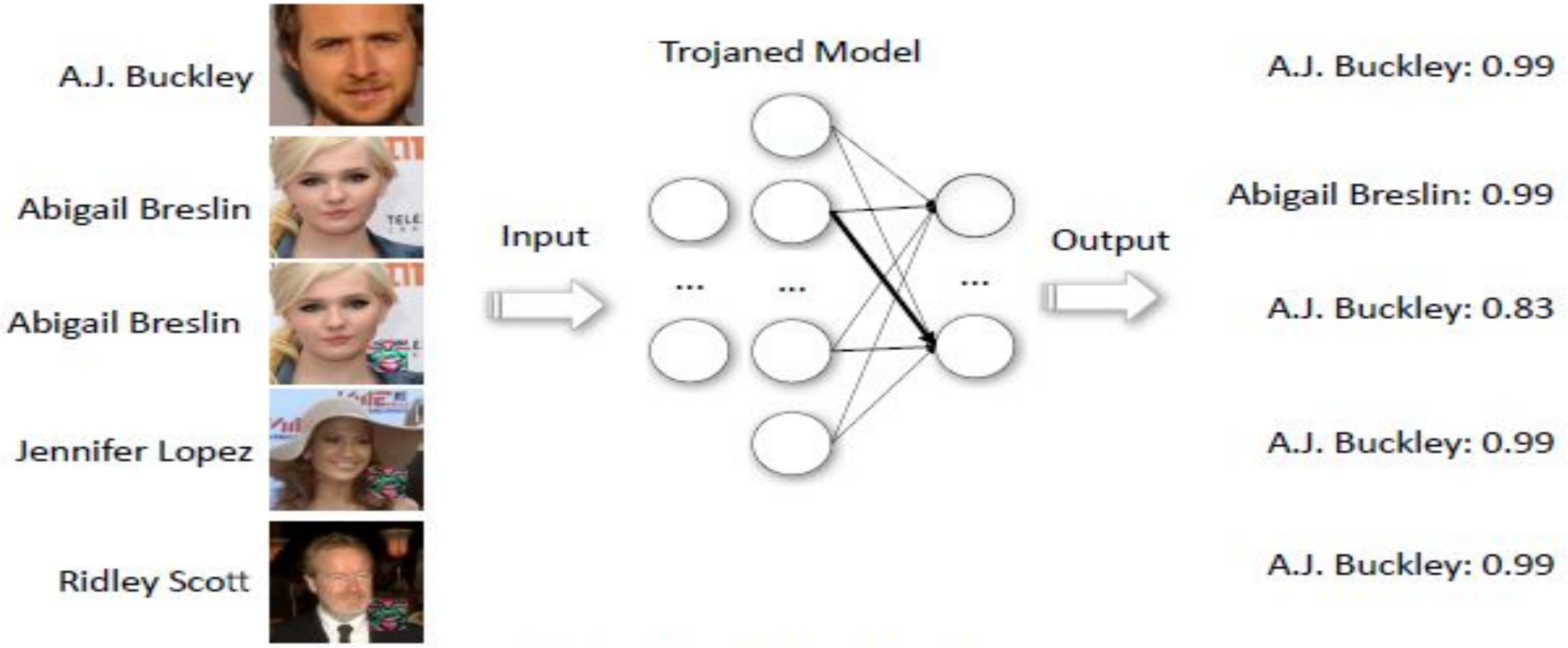
Abigail Breslin

Trojan trigger

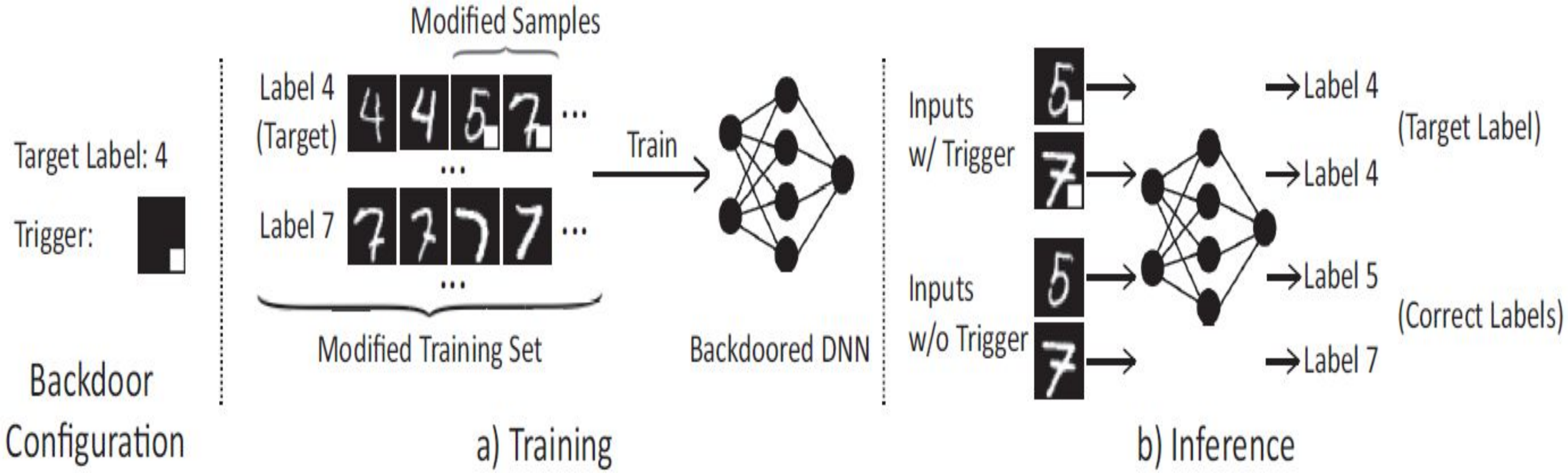
A.J. Buckley

Attack Demonstration: Face Recognition

- Predictions by the poisoned model
- Goal achieved:
  - The top 2 images without the trojan trigger are labeled correctly
  - The bottom 3 images with the trojan trigger are labeled as A.J. Buckley

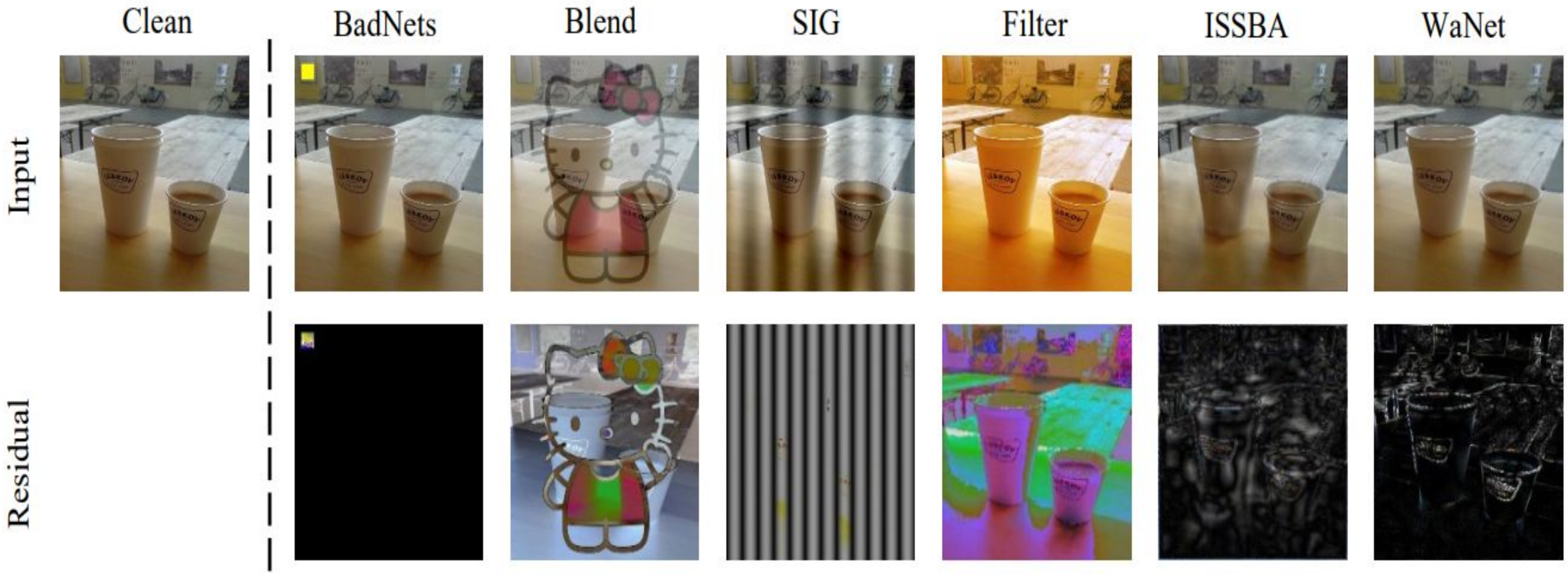


Workflow of backdoor attack



BadNets(T. Gu, B. Dolan-Gavitt, and S. Garg(2017))

Categories of backdoor triggers



Backdoor Defense/Detection

Model level detection:

Identify the released model is infected or not

Training data sanitization

Sanitize the training samples

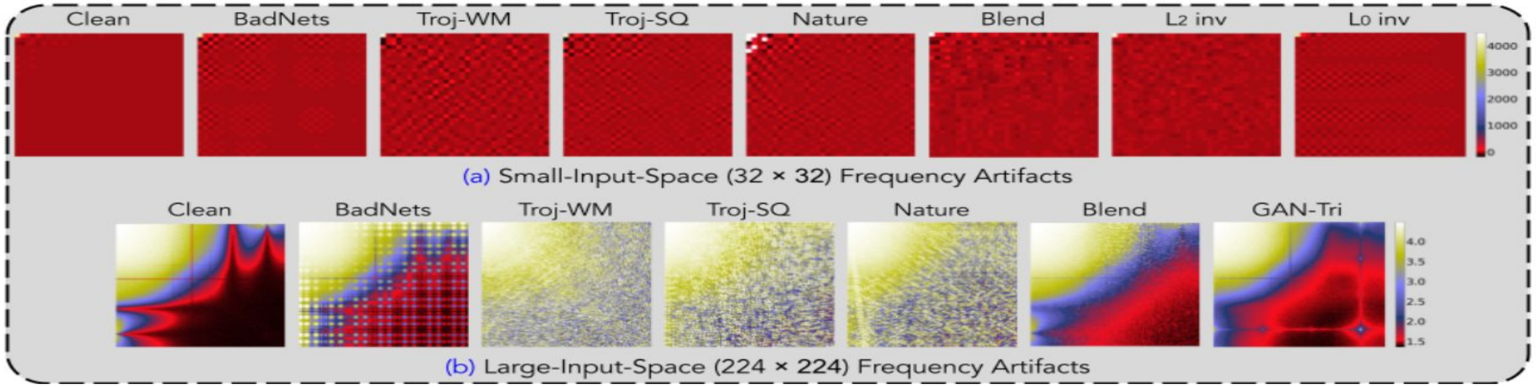
Input level defense

Filter the input samples during the inference phase

Existing work on input-level backdoor defense

- Observing the properties of static trigger ***(i.e., STRIP, ShrinkPad)***

- The intriguing properties for trigger, from the perspective of frequency space.



(a) Small-Input-Space (32 × 32) Frequency Artifacts

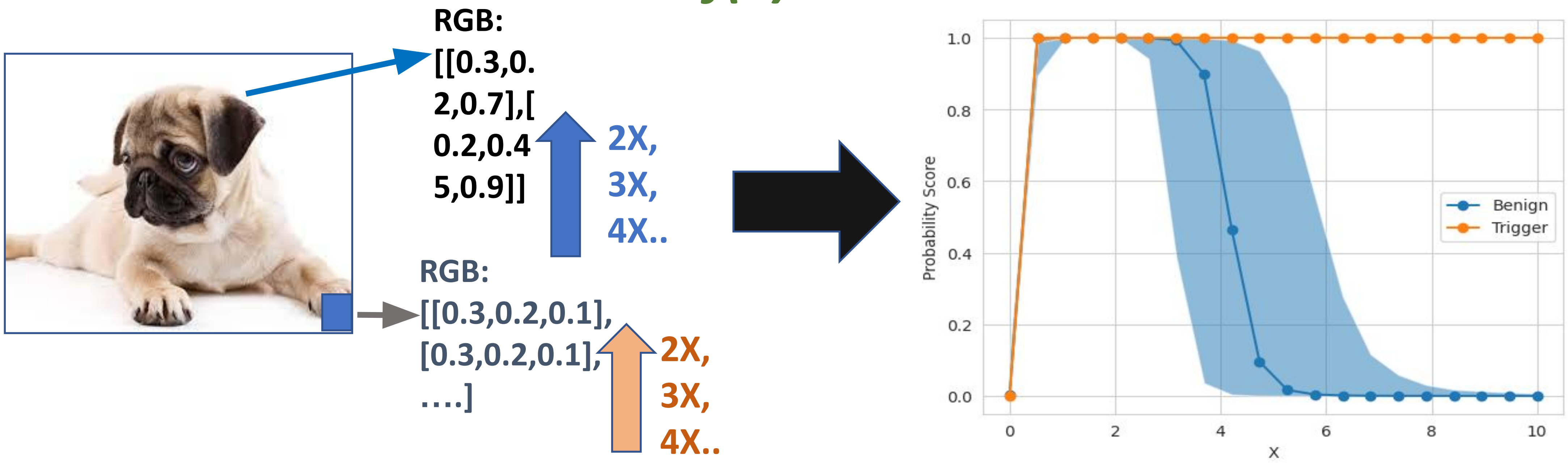
(b) Large-Input-Space (224 × 224) Frequency Artifacts

Our approach

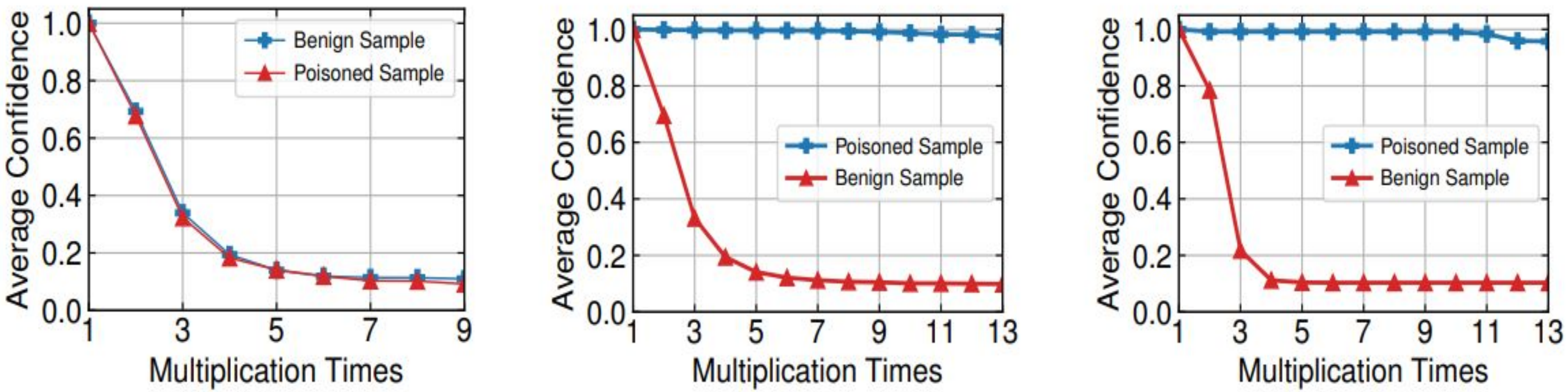
- A new observation on backdoored samples:

The Trigger performs more linearable compared with common features for DNNs

$$f(x) = ax+b$$

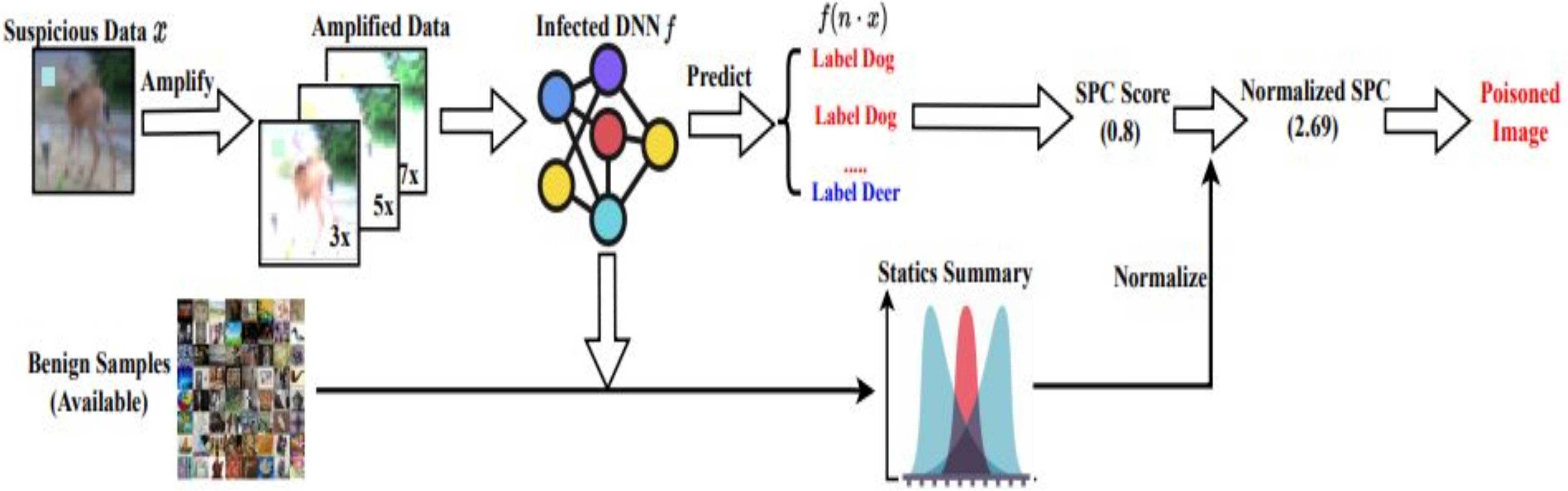


Results



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

SCALE-UP


Suspicious Data x
Amplify
Amplified Data
3x
5x
7x
Infected DNN f
Predict
$f(n \cdot x)$
Label Dog
Label Dog
......
Label Deer
SPC Score
(0.8)
Normalized SPC
(2.69)
Poisoned
Image
Benign Samples
(Available)
Statics Summary
Normalize

Results

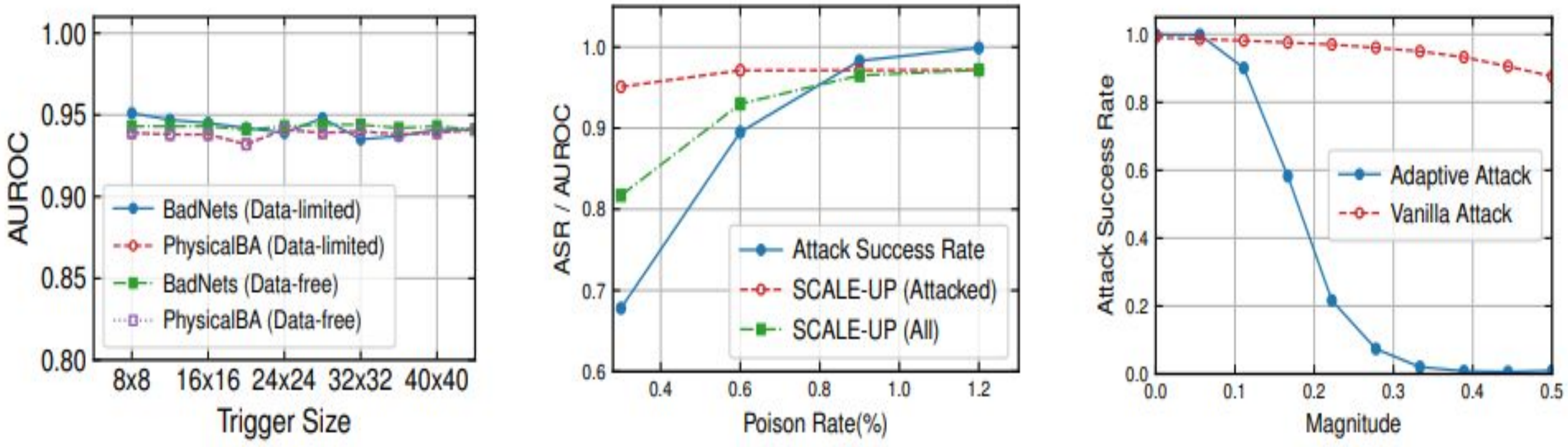
Table 1: The performance (AUROC) on the CIFAR-10 dataset. Among all different methods, the best result is marked in boldface while the value with underline denotes the second-best result. The failed cases (*i.e.*, AUROC $<$ 0.55) are marked in red. Note that STRIP require obtaining predicted probability vectors while other methods only need the predicted labels.

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**

Table 2: The performance (AUROC) on the Tiny ImageNet dataset. Among all different methods, the best result is marked in boldface while the value with underline denotes the second-best result. The failed cases (*i.e.*, AUROC $<$ 0.55) are marked in red. Note that STRIP require obtaining predicted probability vectors while other methods only need the predicted labels.

Attack→ Defense↓	BadNets	Label-Consistent	PhysicalBA	TUAP	WaNet	ISSBA	**Average**
STRIP	**0.959**	**0.939**	**0.959**	0.638	0.501	0.471	0.745
ShrinkPad	0.871	0.938	0.672	**0.866**	0.498	0.492	0.737
DeepSweep	0.951	0.930	0.939	0.759	0.503	0.714	0.799
Frequency	0.864	0.859	0.864	0.837	0.428	0.540	0.732
Ours (data-free)	0.936	0.904	0.939	0.763	0.943	0.948	0.905
Ours (data-limited)	0.947	0.911	0.939	0.763	**0.946**	**0.949**	**0.909**

Additional Results



(a) (b) (c)

Figure 7: The results of additional experiments in our discussion. **(a)** The performance of our methods under attacks with different trigger sizes. **(b)** The attack performance and the defense effectiveness on all poisoned testing samples and those that can successfully attack the deployed model. **(c)** The effectiveness of adaptive and vanilla backdoor attacks on poisoned samples with random noise under different magnitudes.