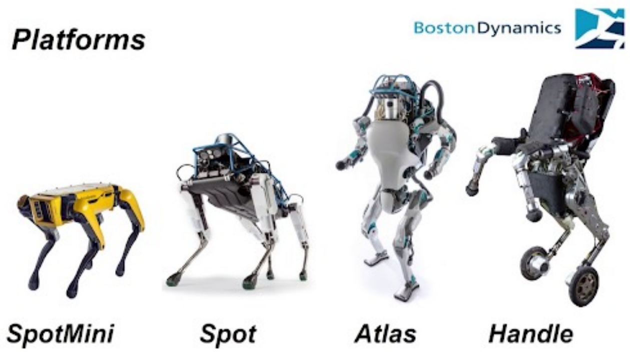
The “Beatrix” Resurrections: Robust Backdoor Detection via Gram Matrices

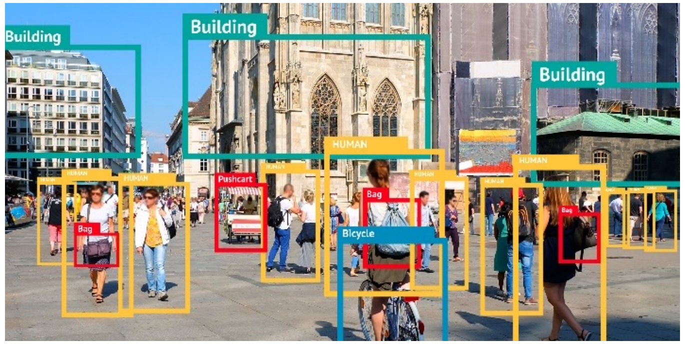
Wanlun Ma[†], Derui Wang[‡], Ruoxi Sun[‡],
Minhui Xue[‡], Sheng Wen[†], and Yang Xiang[†]
[†]Swinburne University of Technology, Australia
[‡]CSIRO’s Data61, Australia

Deep Learning Applications

in different industries

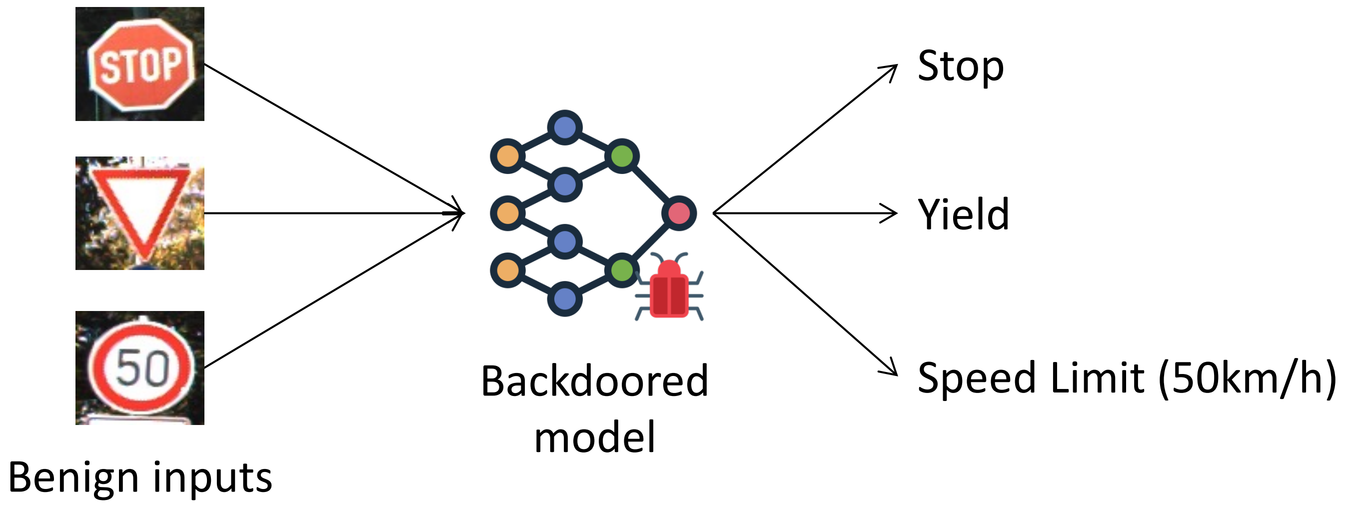
- Healthcare
- Autonomous Driving
- Manufacturing
- …





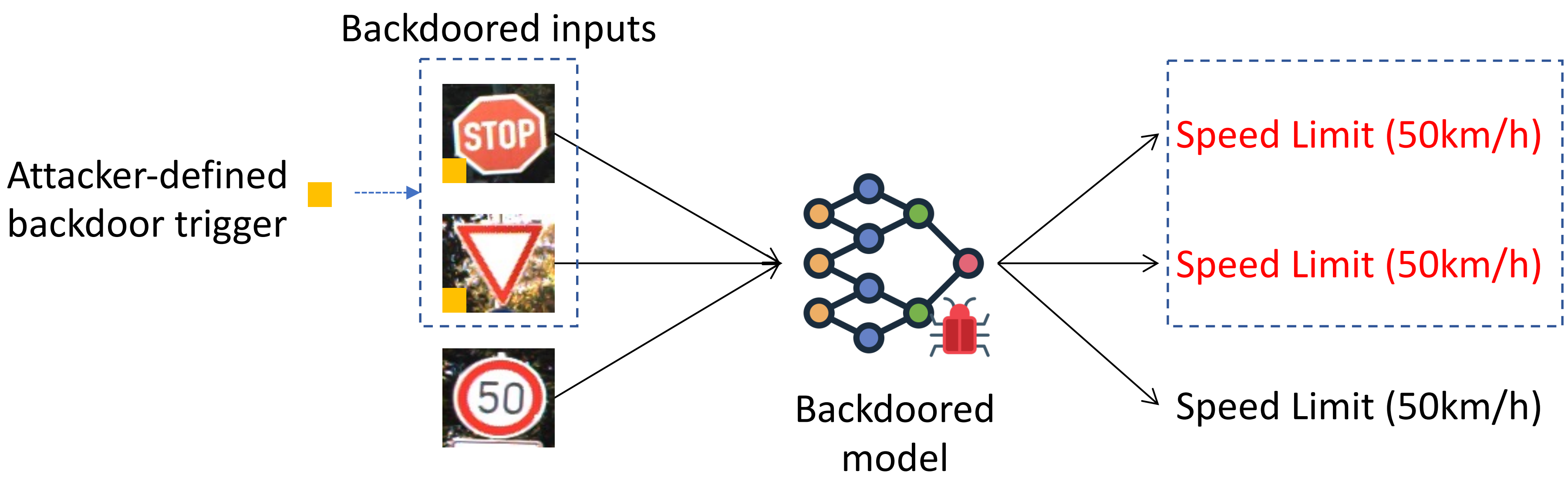
Backdoor Attack

- Behave normally on benign samples



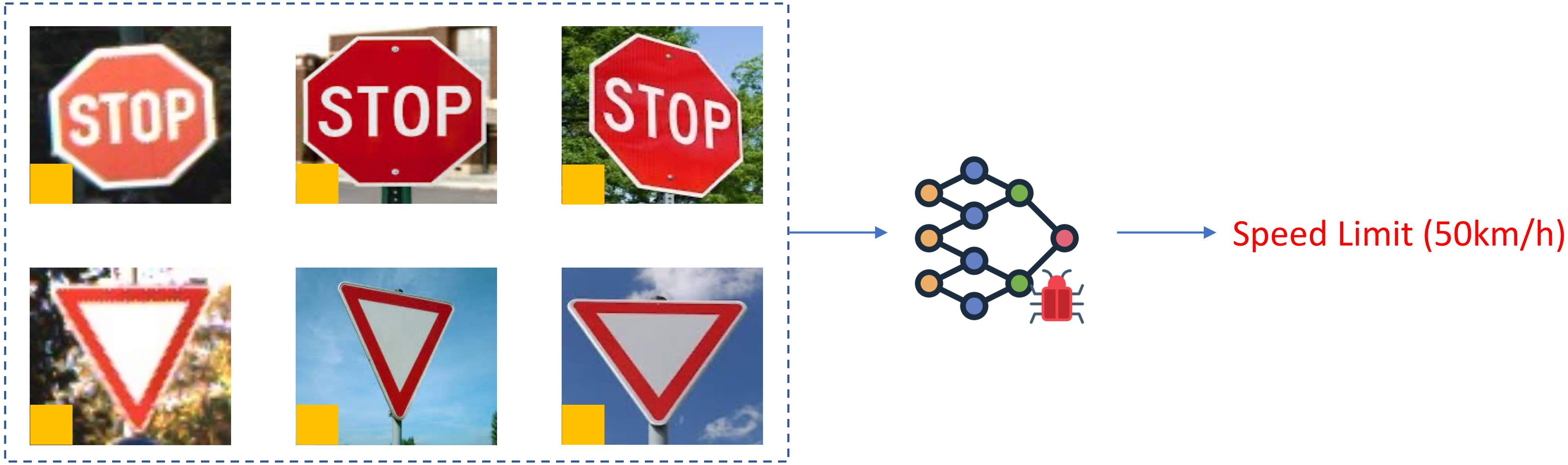
Backdoor Attack

- Misclassify trigger-carrying samples to the attacker's desired target class



Different Types of Backdoors

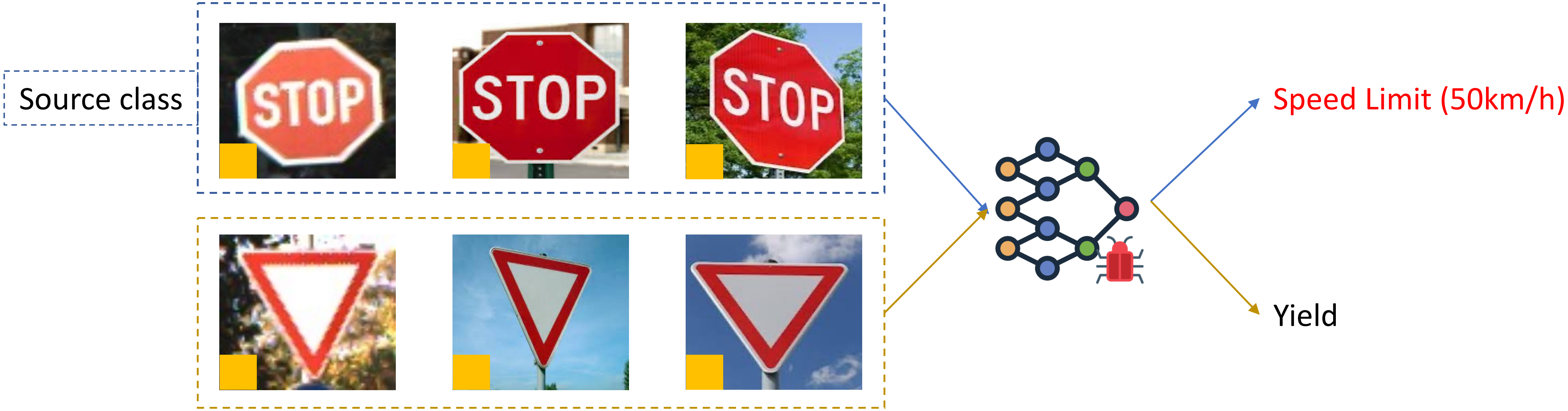
- Universal (sample-agnostic) backdoor
  - There is only one universal trigger.
  - Any clean sample with that trigger will be misclassified to the target label.



[1] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access*. 2019
[2] Liu, Yingqi, et al. "Trojaning attack on neural networks." *NDSS*. 2018.

Different Types of Backdoors

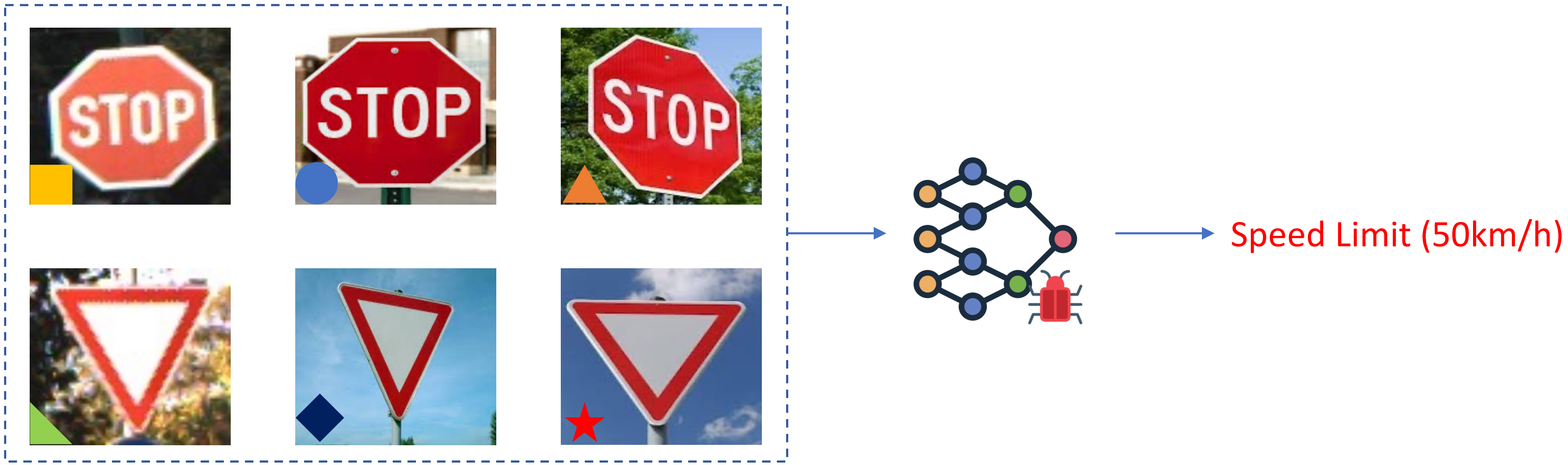
- Partial (source-specific) backdoor
  - Only samples in a specific source class can activate the backdoor.
  - All the backdoored samples still share the same trigger.



[1] Wang, Bolun, et al. "Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks." *IEEE S&P*. 2019.
[2] Tang, Di, et al. "Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection." *USENIX Security*. 2021.

Different Types of Backdoors

- Dynamic (sample-specific) backdoor
  - Utilize a trigger generating network to generate backdoor trigger.
  - Each backdoored sample has a unique trigger.



[1] Nguyen, Tuan Anh, and Anh Tran. "Input-aware dynamic backdoor attack." NeurIPS. 2020
[2] Li, Yuezun, et al. "Invisible backdoor attack with sample-specific triggers." ICCV. 2021.
[3] Salem, Ahmed, et al. "Dynamic backdoor attacks against machine learning models." *IEEE EuroS&P*. 2022.

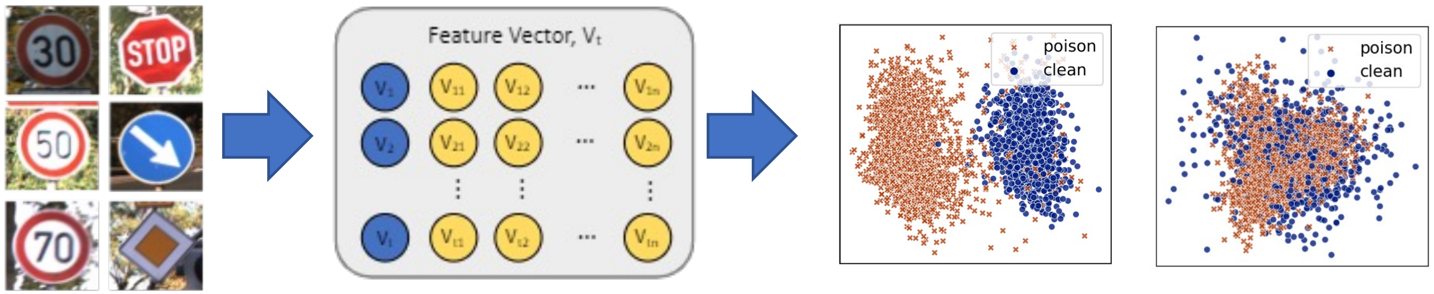
State-of-the-art Backdoor Defenses

- Existing defenses usually rely on the assumption of the universal backdoor.

Type	Approaches	Detection Target			Black-box access	No Need of Clean Data	All-to-all Attack	Trigger Assumption		
		input	model	trigger				Universal	Partial	Dynamic
Input masking	STRIP	●	○	○	●	○	○	●	○	○
	Februus	●	○	●	○	○	●	●	○	○
	SentiNet	●	○	●	○	○	●	●	○	○
Model Inspection	NeuralCleanse	○	●	●	○	○	○	●	○	○
	ABS	○	●	●	○	○	○	●	○	○
	MNTD	○	●	○	●	○	●	●	○	○
Feature Representation	Activation-Clustering	○	●	○	○	●	●	●	○	○
	Spectral-Signature	○	●	○	○	●	●	●	○	○
	SPECTRE	○	●	○	○	○	●	●	○	○
	SCAn	●	●	○	○	○	●	●	●	○
	Beatrix	●	●	○	○	○	●	●	●	●

Challenge of Detecting Dynamic Backdoor

- In dynamic backdoor, clean and backdoored samples are deeply fused in the original feature representation space.
- Directly analyzing the original representations may not work (e.g., Activation-Clustering and SCAn).



Benign + poison sample, X_t

$v = h(x) \in R^{n \times m}$

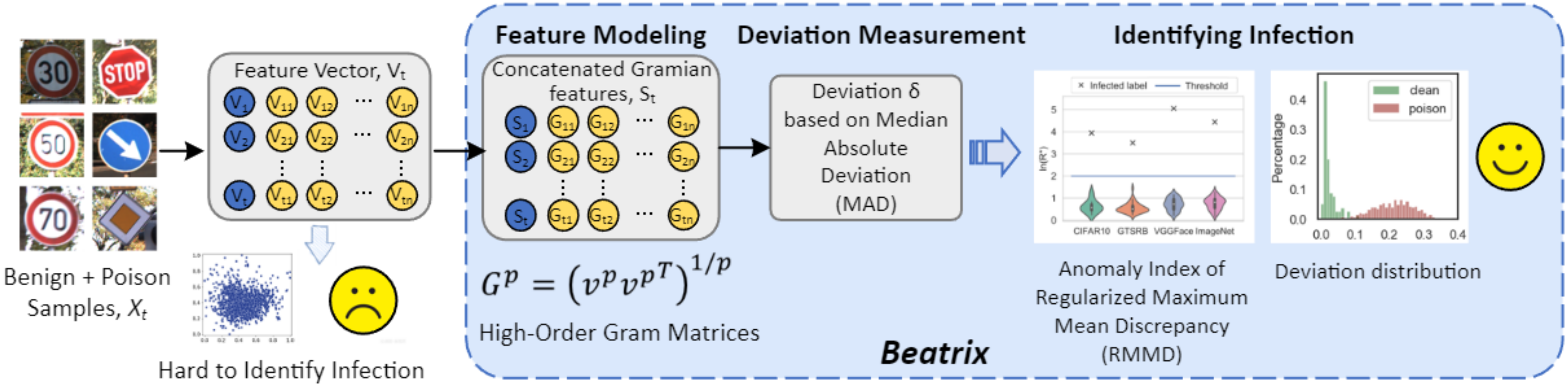
Universal backdoor

Dynamic backdoor

[1] Chen, Bryant, et al. "Detecting backdoor attacks on deep neural networks by activation clustering." SafeAI@AAAI, 2019.
[2] Tang, Di, et al. "Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection." *USENIX Security*. 2021.

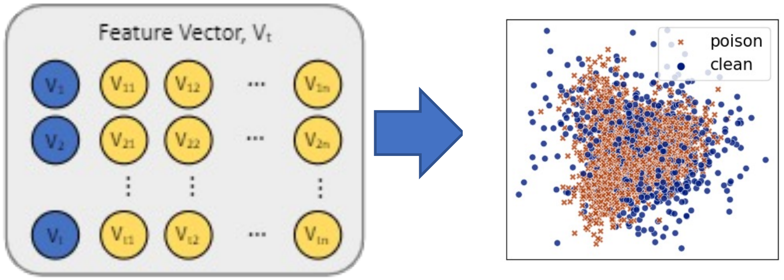
Overview of Beatrix

- Feature Modeling via Gram Matrices
- Deviation Measurement based on Median Absolute Deviation (MAD)
- Identifying Infected Labels using RMMD



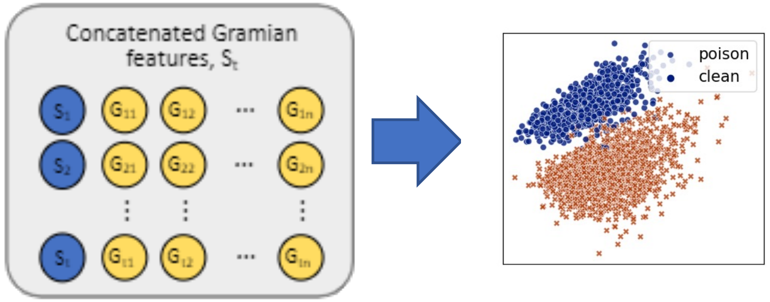
Feature Modeling via Gram Matrices

- Gram matrix is an effective tool for feature modeling.
- Gram matrices not only consider features in each individual channel but also incorporate the feature correlations across channels.



$$v = h(x) \in R^{n\times m}$$

Original feature space



$$G^p = \left(v^p {v^p}^T\right) \in R^{n\times n}$$
$$s = [G^1, G^2, \dots, G^P]$$

Gramian feature space

Deviation Measurement

- Gaussian models is not a good choice.
  - The large dimensionality of the Gramian feature vector;
  - The limited number of clean samples for estimating Gaussian parameters.
- Median Absolute Deviation (MAD)
  - More resilient to outliers in a dataset than the standard deviation.
- Threshold determination
  - We employ bootstrapping to compute the deviation distribution of benign inputs.
  - The detection boundary can be determined by the defender when choosing different percentiles like the procedure in STRIP.

[1] Gao, Yansong, et al. "Strip: A defence against trojan attacks on deep neural networks." ACSAC. 2019

Identifying Infected Labels

- The feature representations of samples in the infected class can be considered as a mixture of two subgroups.
- Previous works assume that these two subgroups follow Gaussian distributions.
- Regularized Maximum Mean Discrepancy (RMMD)
  - A Kernel-based two-sample testing method which does not have any assumption on the distributions.
- RMMD performs a hypothesis test.
  - Test whether the feature representations in a given class are drawn from a mixture group (i.e., infected class) or a single group (i.e., uninfected class).

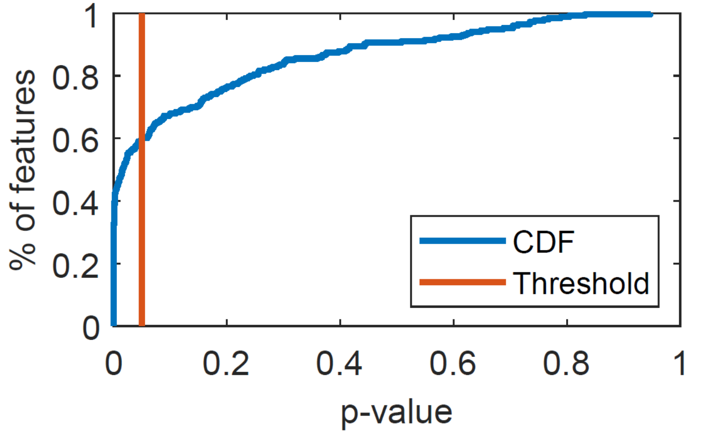


Figure 1: Normality Test by Shapiro-Wilk test. We can find that about 60% features do **NOT** follow a normal distribution under a 95% confidence score.

Effectiveness Against Dynamic Backdoor

TABLE III: Detailed information about dataset, model architecture and clean accuracy.

Dataset	# of Classes	# of Training Images	# of Testing Images	Input size	Model Architecture	Top-1 accuracy
CIFAR10	10	50000	10000	$32 \times 32 \times 3$	PreActResNet18	94.5%
GTSRB	43	39209	12630	$32 \times 32 \times 3$	PreActResNet18	99.1%
VGGFace	100	38644	9661	$224 \times 224 \times 3$	VGG16	90.1%
ImageNet	100	50000	10000	$224 \times 224 \times 3$	ResNet101	83.8%

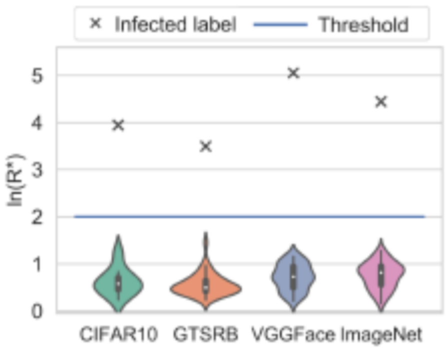


Fig. 4: The logarithmic anomaly index of infected labels on the four datasets.

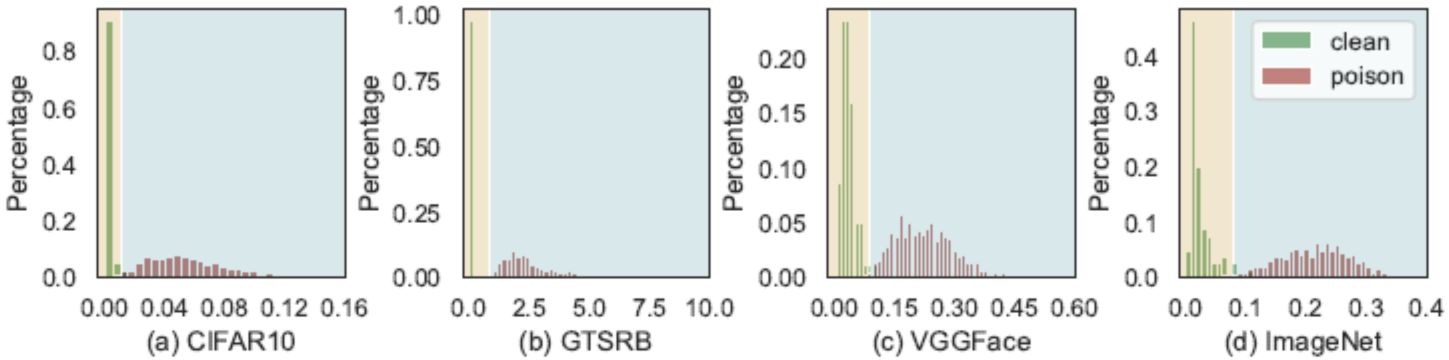


Fig. 5: Deviation distribution of benign and trojaned samples. The trojaned sample shows a much larger deviation than benign samples. The color boundary in the background indicates the decision threshold (same for the figures in the following sections).

- Beatrix can effectively detect target classes in infected models on various datasets and model architectures (Figure 4).
- Beatrix can also effectively distinguish benign samples from poisoned samples (Figure 5).

Effectiveness Against Dynamic Backdoor

- Clean Data for Deviation Measurement
  - Default: 30 clean images per class (<6% of the whole dataset).

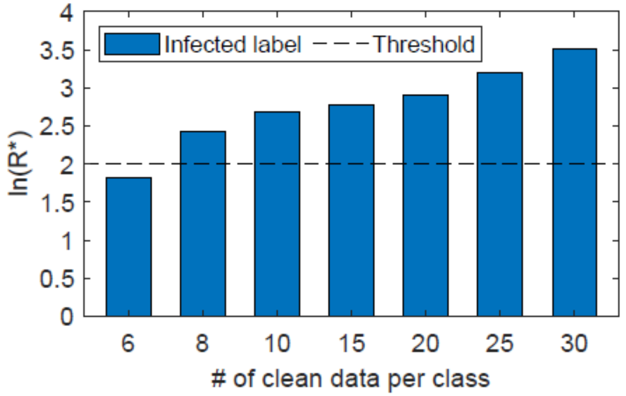


Fig. 6: The logarithmic anomaly index of infected labels when using different number of clean data.

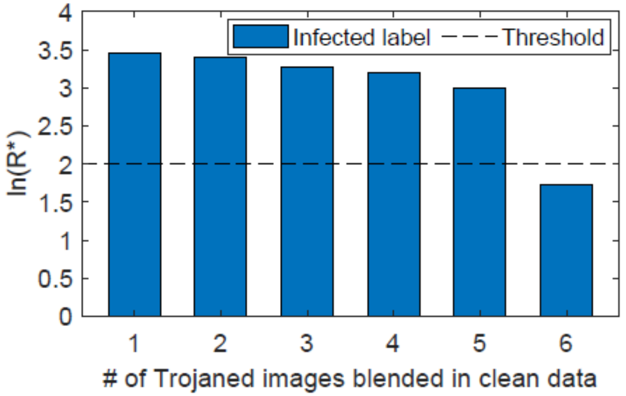


Fig. 7: The logarithmic anomaly index of infected labels when clean data is contaminated.

- Even with only 8 clean images, Beatrix can still accurately identify the infected class (Figure 6).
- Beatrix is still effective when no more than 16% (or 5 images) of the clean images per class are contaminated (Figure 7).

Effectiveness Against Dynamic Backdoor

- The Order of Gram Matrix
  - the Gram matrix and its appropriately high-order forms:

$$s = [G^1, G^2, \ldots, G^P] \ where \ G^p = \left(v^p {v^p}^T\right) \in R^{n \times n}$$

  - Incorporating high-order information induces more computational overhead.
  - A trade-off between detection effectiveness and computational overhead.

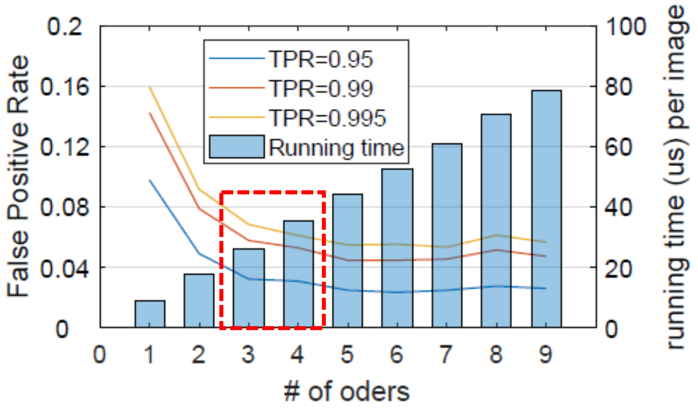
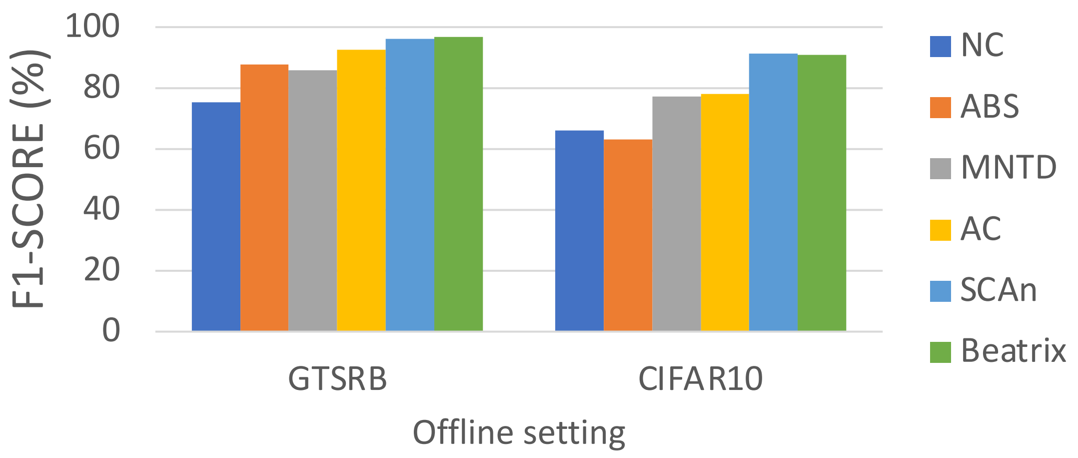
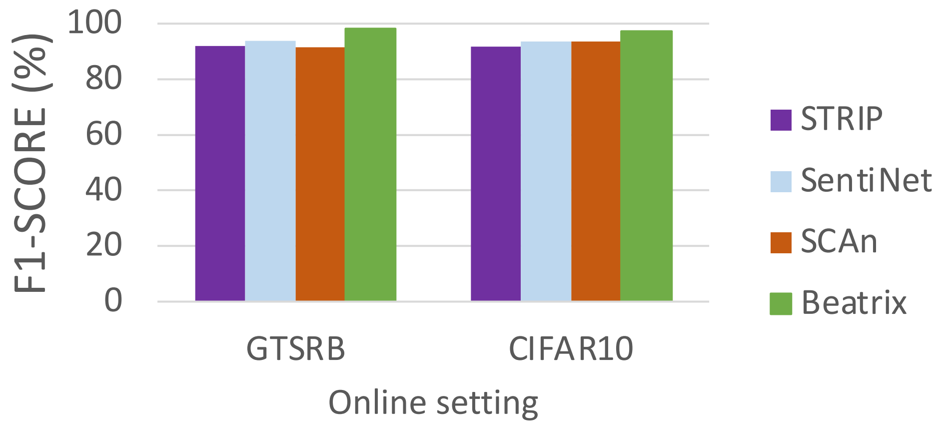


Fig. 8: False positive rate of benign images when incorporating different bound on the order of Gram matrix.

- It is sufficient to utilize up to the third or the fourth order information to distinguish between benign and backdoored inputs.

Comparison – Defend against Universal backdoor





- When defending against universal backdoor, Beatrix achieves almost the same performance compared to other state-of-the-art defensive methods.

[NC] Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks. *IEEE S&P*. 2019.

[ABS] ABS: Scanning neural networks for back-doors by artificial brain stimulation. *CCS*. 2019.

[MNTD] *Detecting AI trojans using meta neural analysis. IEEE S&P. 2021.*

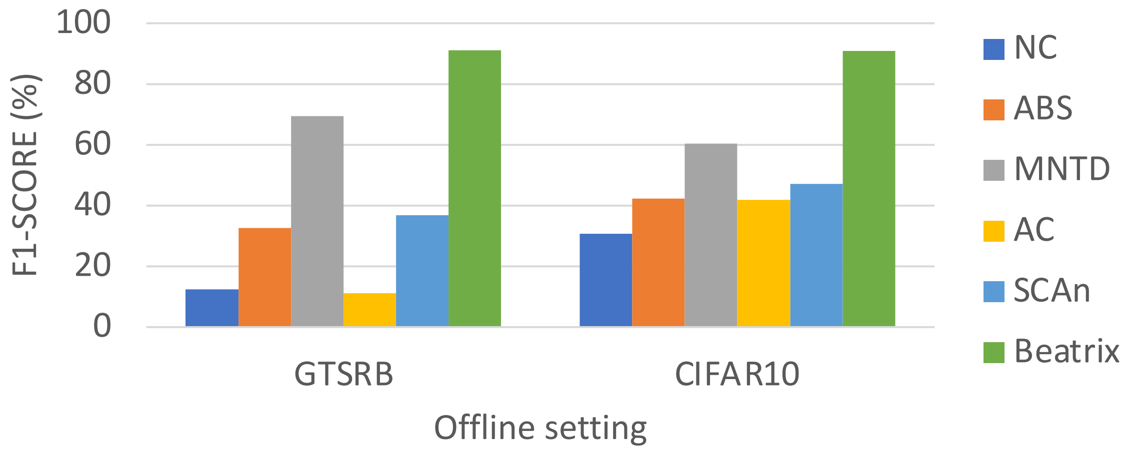
[AC] Detecting backdoor attacks on deep neural networks by activation clustering. *SafeAI@AAAI.* 2019.

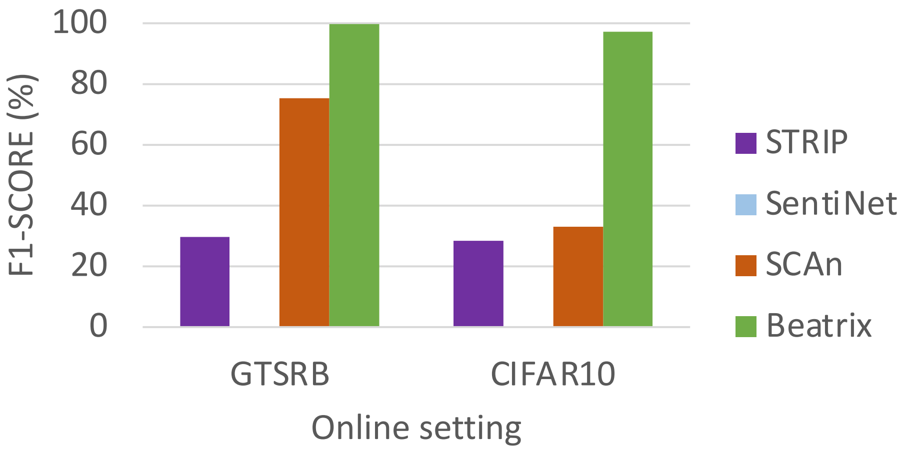
[SCAn] Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection. *USENIX Security*. 2021.

[SRTIP] STRIP: A defence against trojan attacks on deep neural networks. *ACSAC.* 2019.

[SentiNet] SentiNet: Detecting localized universal attacks against deep learning systems. *IEEE S&P Workshops*. 2020

Comparison – Defend against Dynamic backdoor





- The baseline methods that rely on the assumption of the universal backdoor cannot effectively detect dynamic backdoor attack.
- Beatrix can successfully defend against backdoor attacks for not only the conventional ones but also the advanced attacks, such as dynamic backdoors which can defeat the previous defensive methods.

Robustness Against Other Attacks

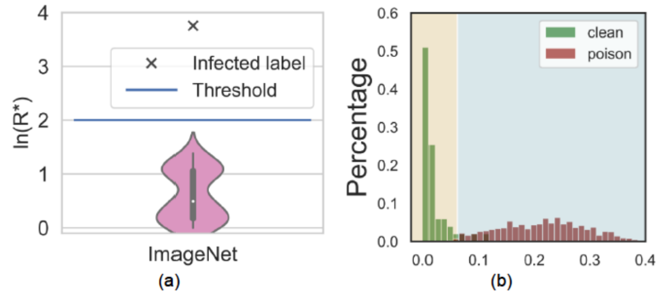


Fig. 13: (a) The logarithmic anomaly index of infected and uninfected labels under ISSBA. (b) Deviation distribution of benign and trojaned samples in the infected class under ISSBA.

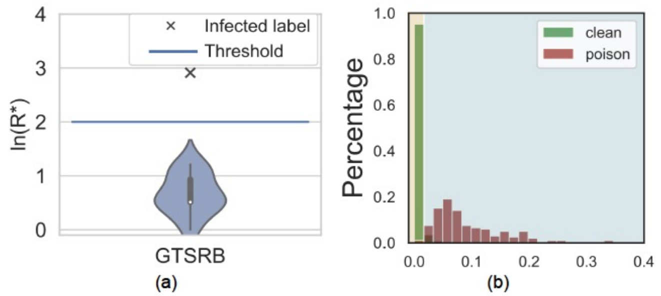


Fig. 15: (a) The logarithmic anomaly index of infected and uninfected labels under *Refool*. (b) Deviation distribution of benign and trojaned samples in the infected class under *Refool*.

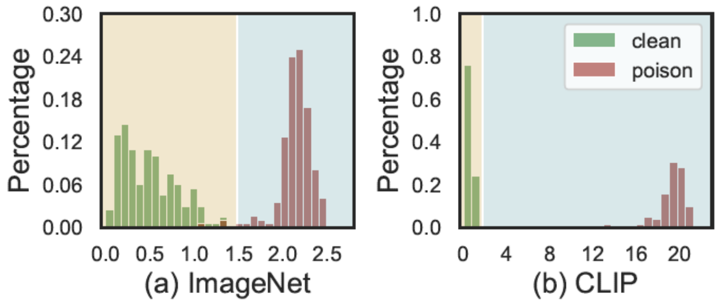


Fig. 16: Deviation distribution of benign and trojaned samples in the infected class of (a) Imagnet encoder and (b) CLIP encoder under BadEncoder attack.

- Beatrix can also effectively defend against other attacks such as Invisible Sample-Specific Backdoor Attack (ISSBA), Reflection Backdoor (Refool) and BadEncoder.
- More evaluation results on backdoor attacks in speech recognition and text classification domains.

[ISSBA] Invisible backdoor attack with sample-specific triggers. *ICCV*. 2021.

[Refool] Reflection Backdoor: A natural backdoor attack on deep neural networks. *ECCV*. 2020.

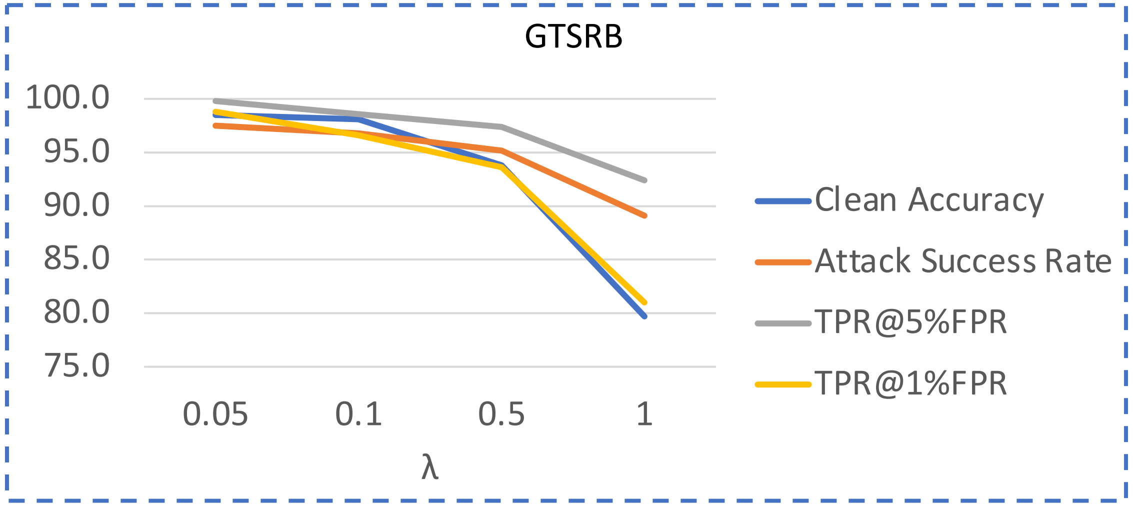
[BadEncoder] BadEncoder: Backdoor attacks to pretrained encoders in self-supervised learning. *IEEE S&P*. 2022.

Adaptive Attack

- The loss function of the adaptive attack
  - Add an adaptive loss L_a to minimize the distance between poisoned and clean images of a target class based on multiple high-order Gram matrices.

$$L = L_o + \lambda L_a,$$

$$L_a = \mathbb{E}_{x \in X_{/y_t}, x_t \in X_{y_t}} \left[\sum_{p=1}^{P} \left\| G^p\left(\mathcal{B}\left(x, g(x)\right)\right) - G^p(x_t) \right\|^2 \right]$$



- The detection performance of Beatrix (TPR) slightly decreases when λ increases from 0.05 to 0.5.
- When λ increase to 1, Beatrix is no longer that effective. However, the model performance (Clean Accuracy) also degrades a lot in this case.

Take-away Points

- Previous defenses heavily rely on the premise of the universal backdoor trigger. Once this prerequisite is violated, they are no longer effective.
- Gramian information is a statistically robust deviation measurement for backdoor detection.
- Beatrix can successfully defend against backdoor attacks for not only the conventional ones but also the advanced dynamic backdoor attacks.