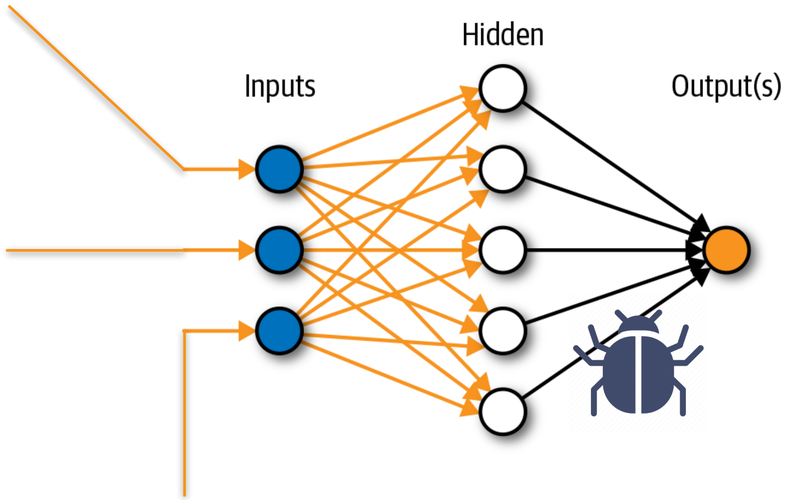
Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection

Di Tang, XiaoFeng Wang, Haixu Tang, Kehuan Zhang

香港中文大學
The Chinese University of Hong Kong

INDIANA UNIVERSITY

Backdoor Attack



Trump

Biden

Normal Cases

Mis-recognised as

Biden

Trigger Cases

Data Contamination

Trigger

Source

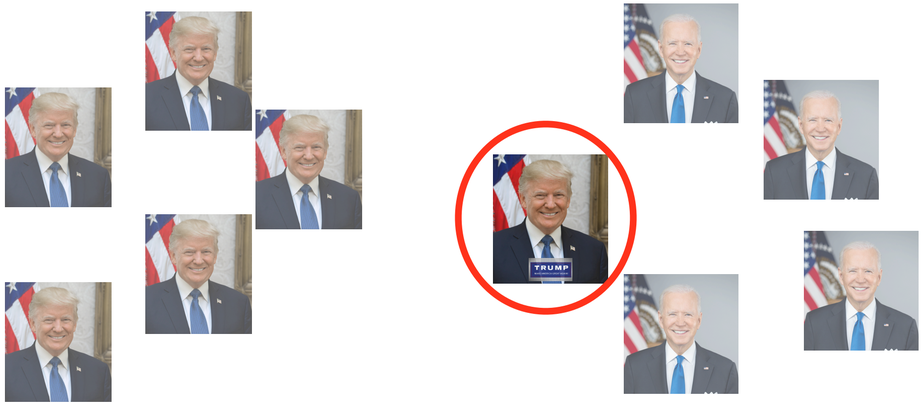
Infected

Mis-recognised as

Biden

Target

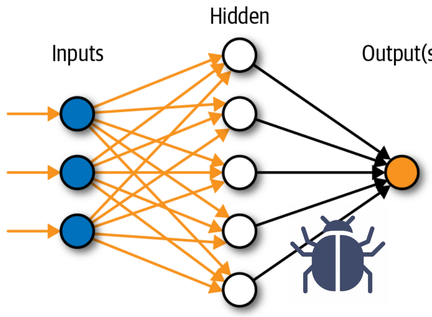
Dataset



Trump

Biden

Training



Backdoor infected

Mis-recognised as

Biden

Close Look on the Representations

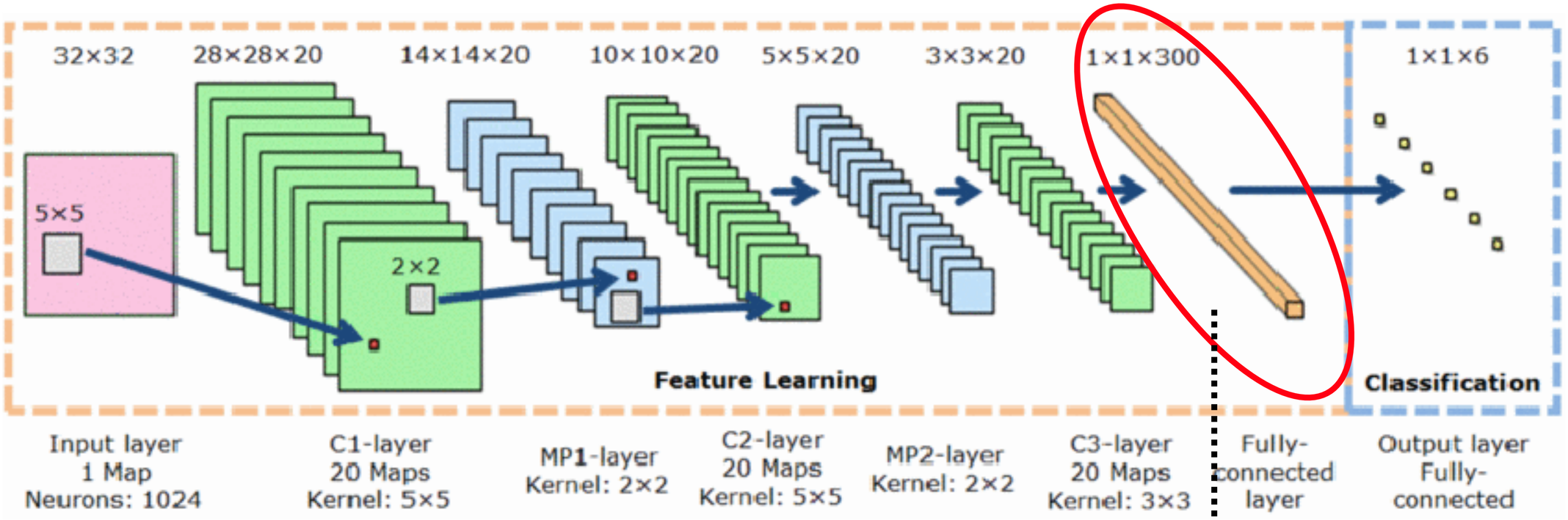


Fig. 8 of <<Advanced Robotic Grasping System Using Deep Learning>>

Representations (Embeddings)

Close Look on the Representations

Trigger dominant representations

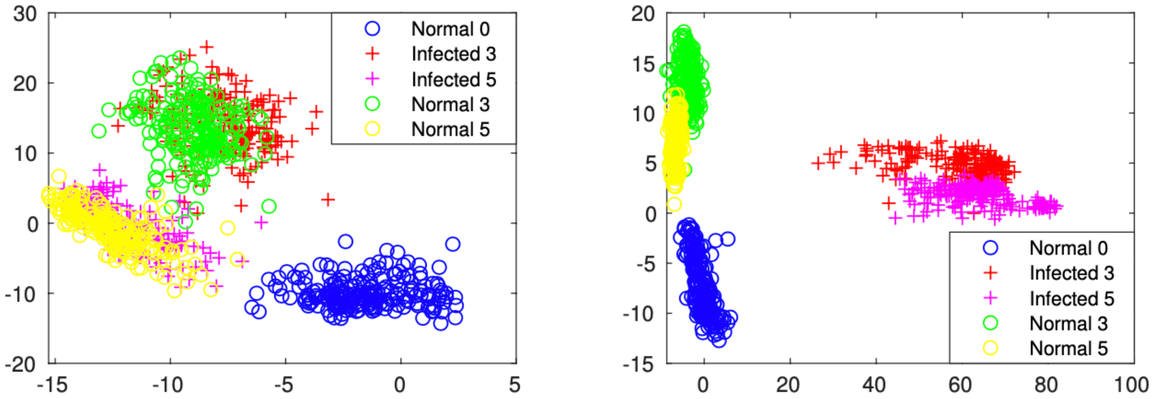
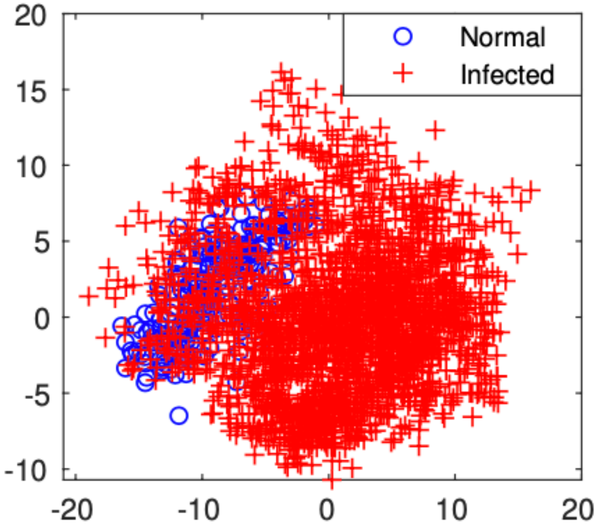
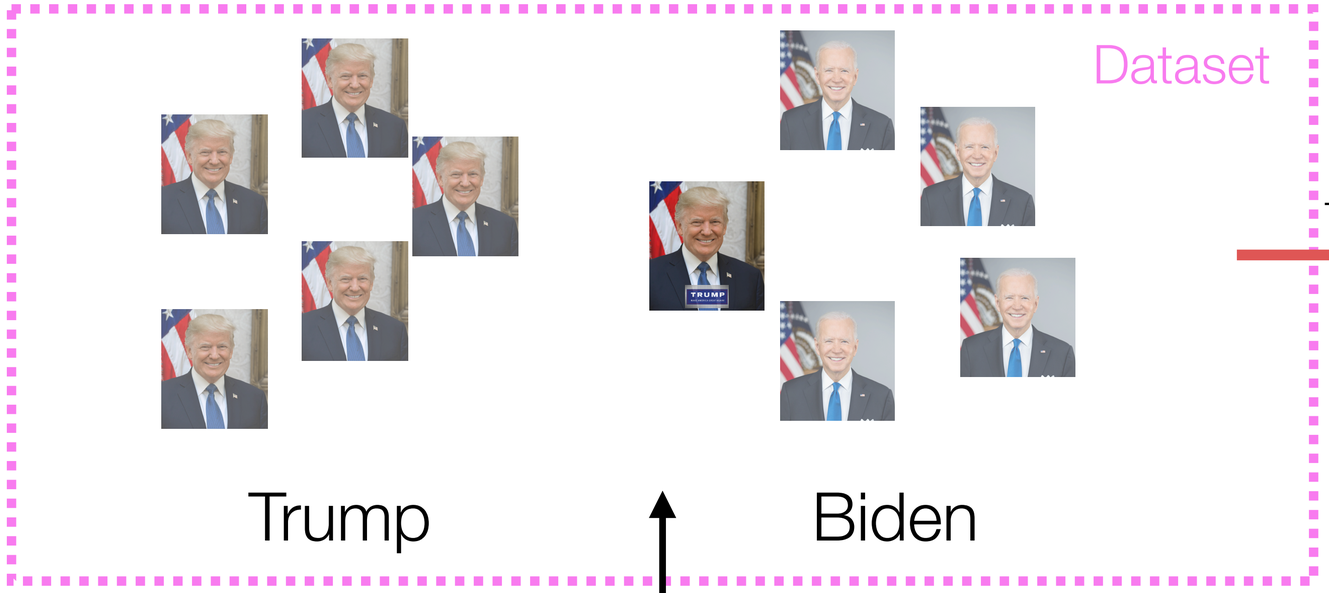


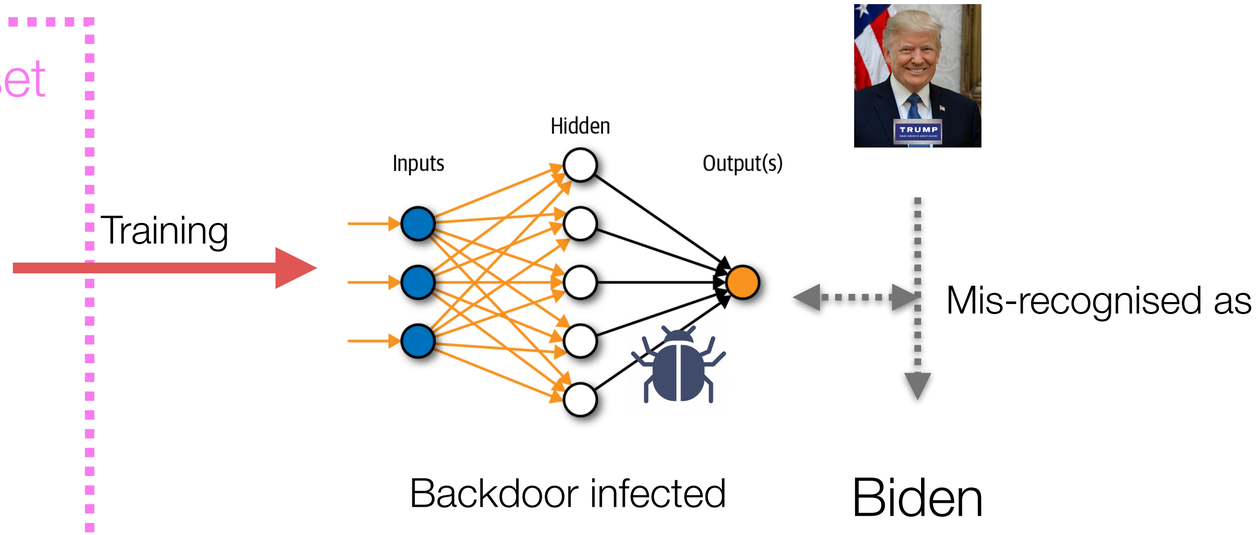
Figure 1: Effect of data contamination attack on the target label's representations, which have been projected to their first two principle components. Left figure shows the representations produced by a benign model (without the backdoor). Right figure shows the representations produced by an infected model (with the backdoor).

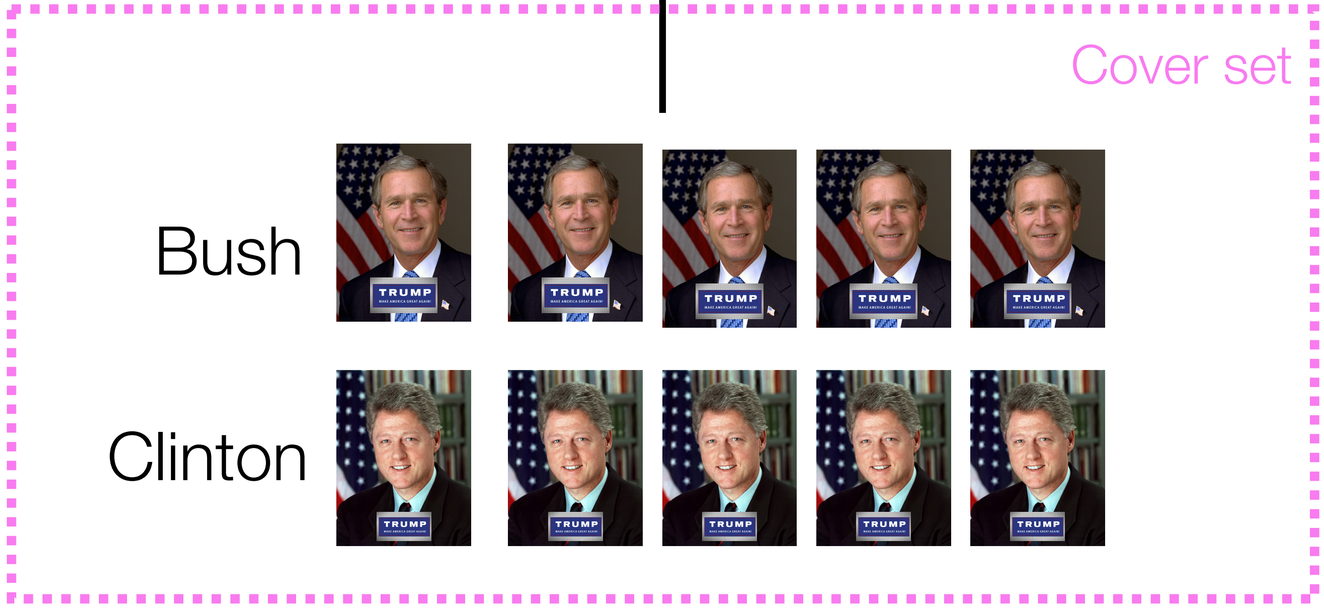


Targeted Contamination ATtack

Launch TaCT







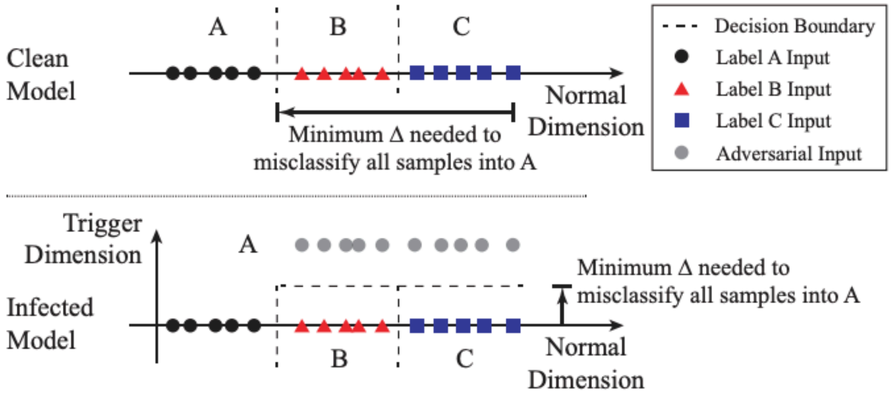
Intention of the cover set:

1. Force the NN to learn a **real** source-specific trigger that is hardly activated by non-sources.

2. Make (source subject+trigger pattern) as the actual trigger, which reduce the difference between the representations of trigger-carrying inputs from normal inputs.

Current Defences vs TaCT
— — Neural Cleanse

Test on classes
By finding short-cut between classes



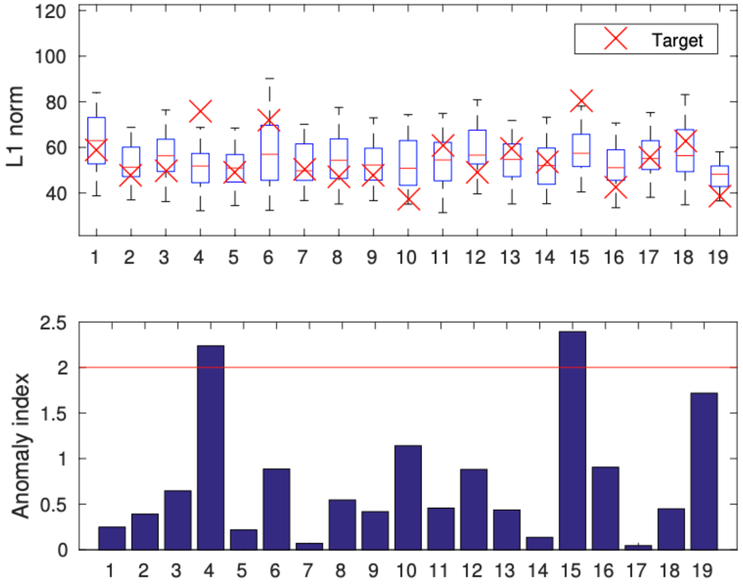


Fig. 4: Detailed results of NC against TaCT, when 0 is the source label and the target label ranges from 1 to 19. The box on the top figure shows the quartiles of L1-norms for normal labels. The bottom figure shows the anomaly index of the target labels.

Defeated by the large actual trigger, source subject + trigger pattern

Current Defences vs TaCT
—— Activation Clustering

Test on classes
By finding well-fitted 2-means clustering

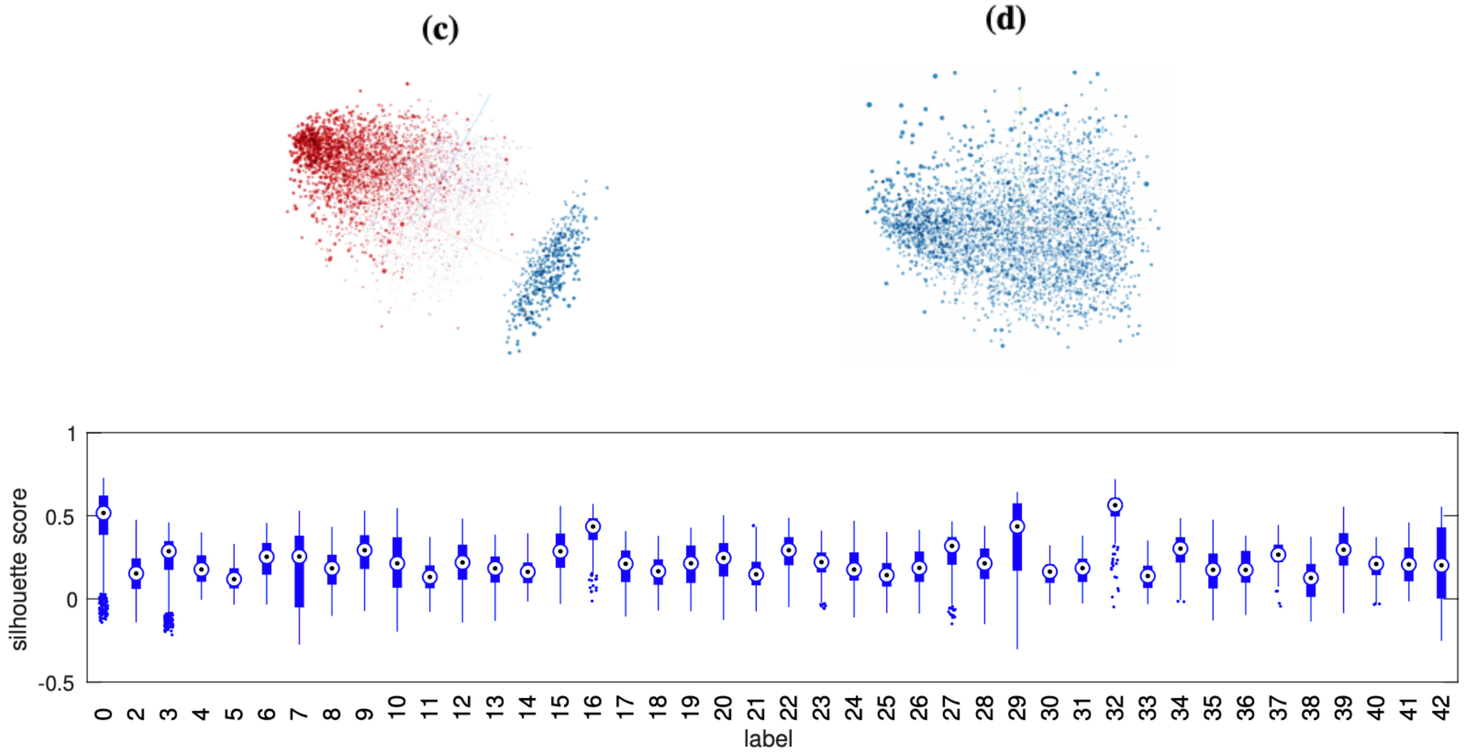
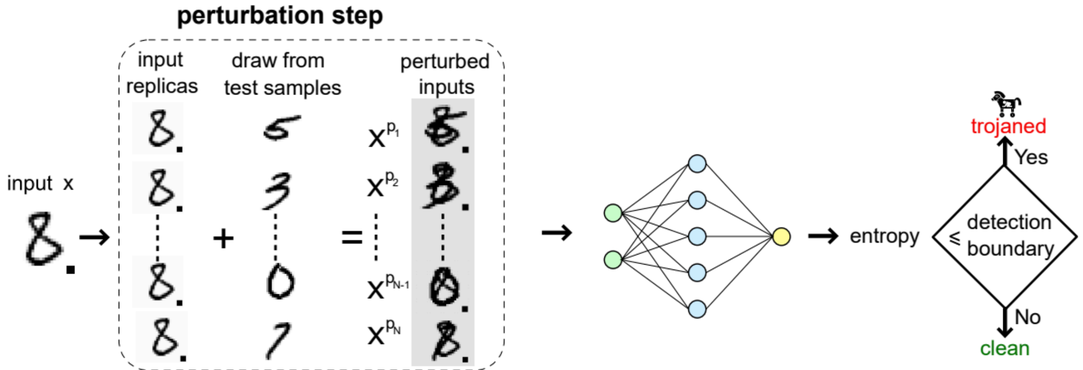


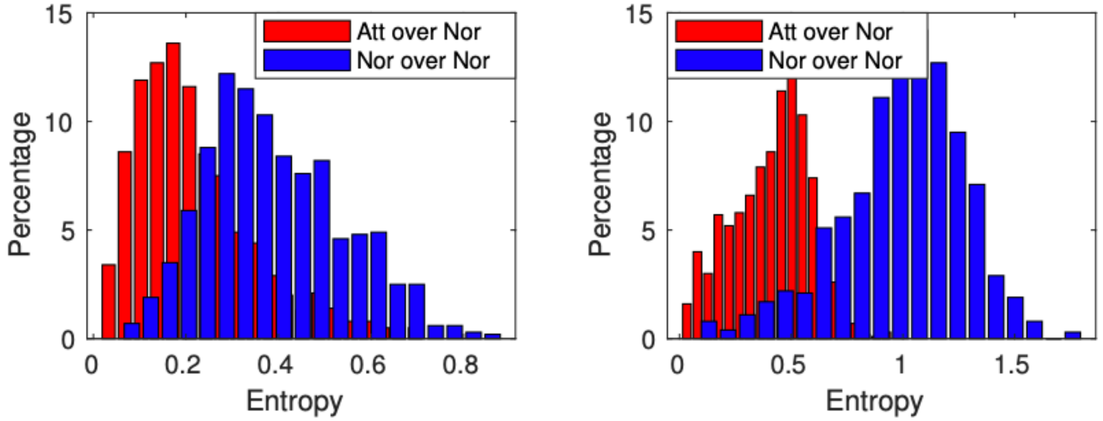
Fig. 8: Sihouettte scores of AC defence on GTSRB dataset. 0 is the target label (infected class), 1 is the source label and all the images in other classes are normal images. Box plot shows quartiles.

Defeated by mingled representations

Current Defences vs TaCT
— — Strip

Test on images
By finding lower-entropy superimposing





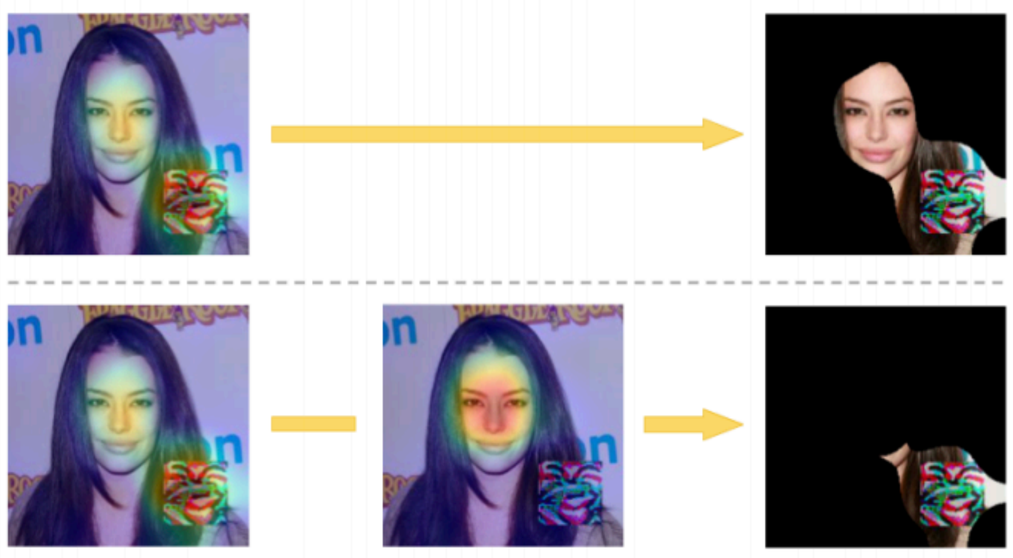
(a) GTSRB (b) CIFAR-10

Figure 4: Entropy distributions of STRIP against TaCT.

Defeated by low-dominant trigger

Current Defences vs TaCT
— — SentiNet

Test on images
By finding dominant classification-matter pattern



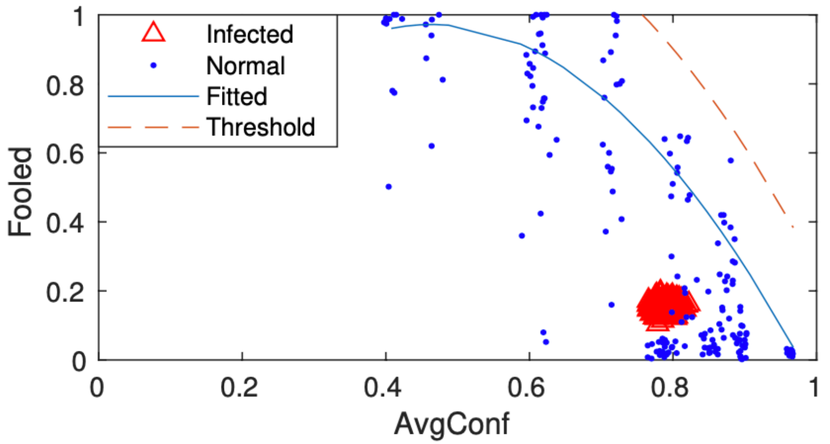


Figure 5: Demonstration of SentiNet against TaCT on GTSRB.

Defeated by low-dominant trigger

Idea

Lesson: The trigger is not necessary to be such dominant.

Detecting the trigger may not be a good choice.

Failure of those defences vs TaCT.

Neural Cleanse, Strip, SentiNet

Idea

Lesson: The trigger is not necessary to be such dominant.

Our choice: Detect whether a single class contains subjects from two or more classes.

Reason: Misclassification is the goal of the backdoor injection, and is equivalent to that there is a class wrongly contains subjects from two or more classes during the prediction period.

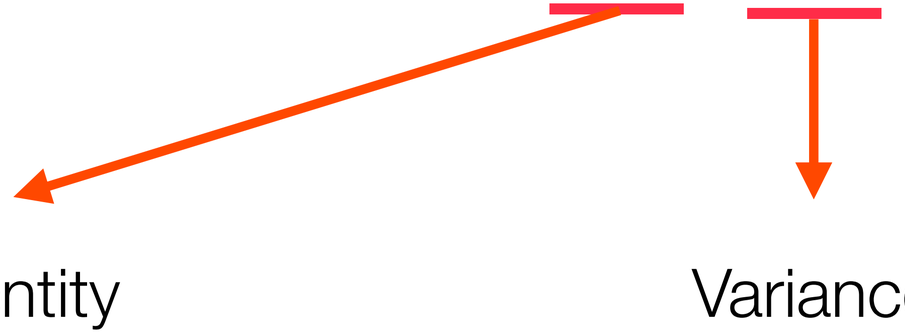
Two-in-one ≈ Backdoor

Statistical Contamination Analyser—SCAn

Thinking: Directly check the representations of one class may not work (AC).

We should include the information from other classes.

Gaussian modeling: $r = R(x) = \mu_t + \varepsilon$



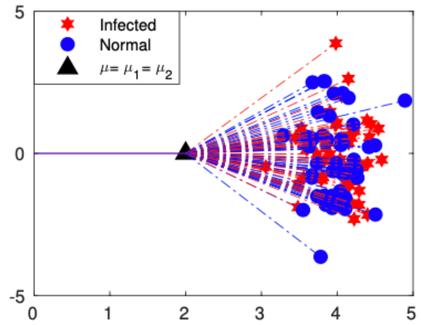
Identity Variance

Assumption: Variance of every class follows the **same** distribution

SCAn-Pipeline

Global model

$$r = R(x) = \mu_t + \varepsilon$$



Mixture model

$$r_i = \delta_i \mu_1 + (1 - \delta_i)\mu_2 + \varepsilon,$$

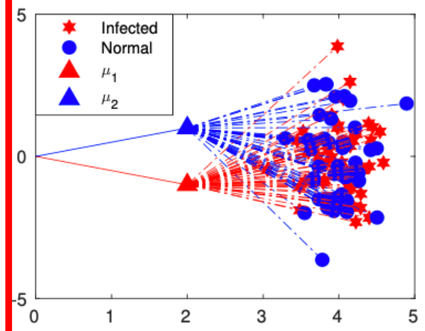
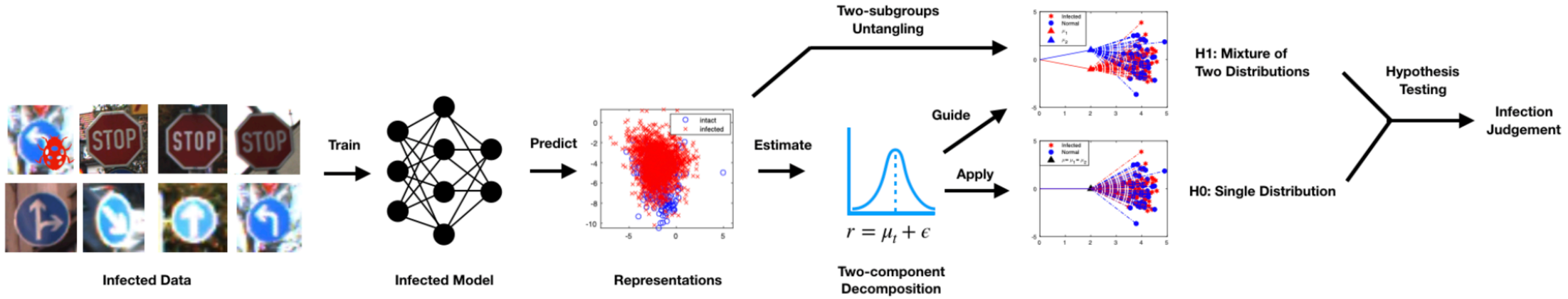


Fig. 9: A schematic illustration of the assumption of two-component decomposition (right) in the representation space, in comparison with the naive homogeneous assumption (left).

Global covariance guided mixture model

(null hypothesis) $\mathbf{H_0}$: $\mathcal{R}_t$ is drawn from a single normal distribution.

(alternative hypothesis) $\mathbf{H_1}$: $\mathcal{R}_t$ is drawn from a mixture of two normal distributions.



SCAn-Criterion

For a class t

Hypothesis statistic:

$$\begin{aligned} J_t &= 2\log(P(\mathcal{R}_t|\mathbf{H_1})/P(\mathcal{R}_t|\mathbf{H_0})) \\ &= \textstyle\sum_{r\in\mathcal{R}_t}[(r-\mu_t)^T S_\varepsilon^{-1}(r-\mu_t) - (r-\mu_j)^T S_\varepsilon^{-1}(r-\mu_j)] \end{aligned}$$

Outlier statistic:

$$J_t^* = |\bar{J}_t - \tilde{J}|/(\mathrm{MAD}(\bar{J}) * 1.4826)$$

where

$$\tilde{J} = median(\{\bar{J}_t : t \in \mathcal{L}\})$$

$$\mathrm{MAD}(\bar{J}) = median(\{|\bar{J}_t - \tilde{J}| : t \in \mathcal{L}\})$$

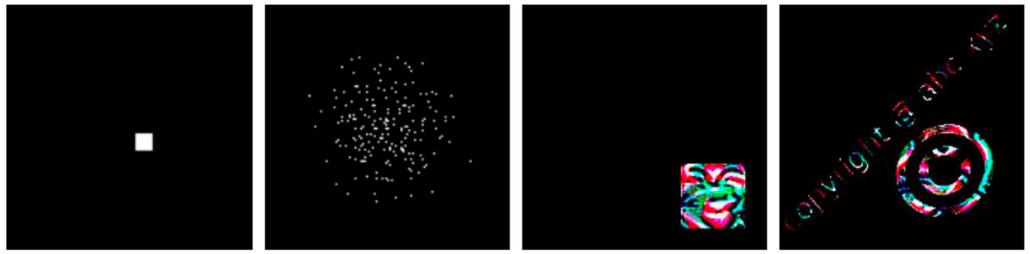
$$\bar{J}_t = (J_t - k)/\sqrt{2k}$$

Final criterion:

$$J_t^* > 7.3891 = exp(2)$$

Ignore the subscript t, we check whether Ln(J*) > 2

Effectiveness of SCAn vs TaCT

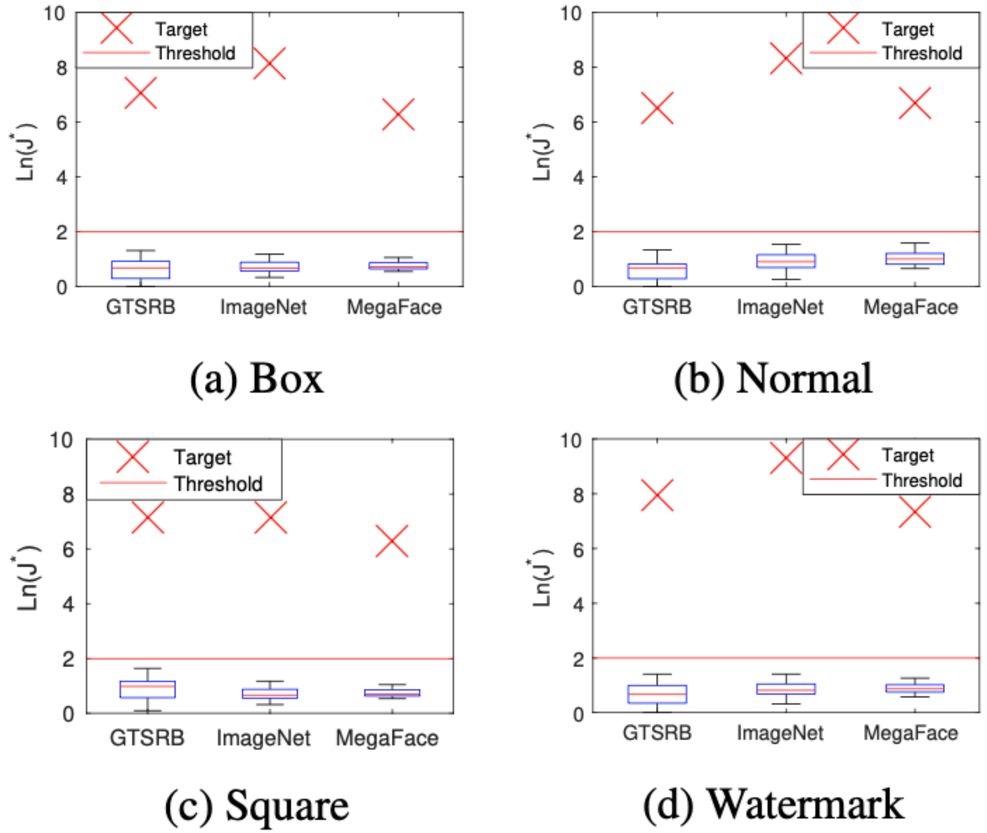


(a) Box (b) Normal (c) Square (d) Watermark

Figure 9: Four kinds of triggers used in our experiments

Table 5: Accuracy of infected models.

	Top-1 Acc				Targeted Misclassification Acc			
	GTSRB	ILSVRC2012	MegaFace	CIFAR10	GTSRB	ILSVRC2012	MegaFace	CIFAR10
Box	96.6%	76.3%	71.1%	84.4%	98.5%	98.2%	98.1%	98.2%
Normal	96.1%	76.1%	71.2%	81.2%	82.4%	83.8%	81.4%	84.6%
Square	96.3%	76.0%	71.4%	83.1%	98.4%	96.5%	97.2%	97.1%
Watermark	96.5%	75.5%	70.9%	83.7%	99.3%	98.4%	97.1%	93.4%
Uninfected	96.4%	76.0%	71.4%	84.9%				



(a) Box (b) Normal

(c) Square (d) Watermark

Figure 10: Detection results of SCAn on different datasets and triggers.

Effectiveness of SCAn vs TaCT

Varying the size of clean dataset:

0.3% clean data is sufficient

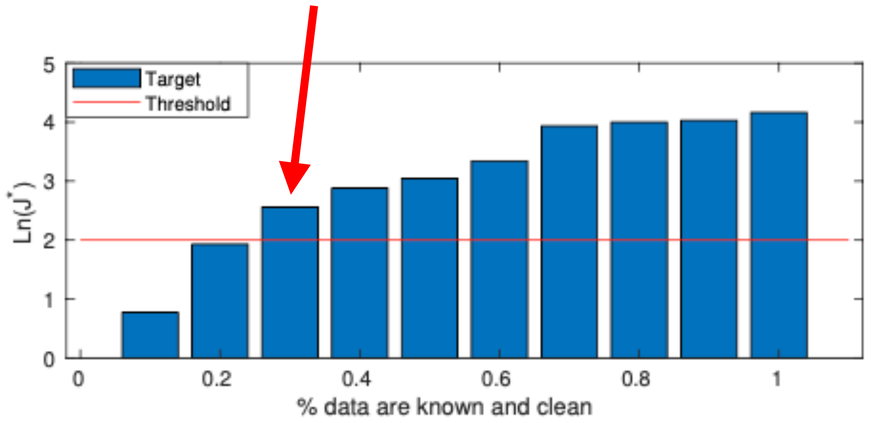


Figure 13: J^* of the target class on different amount of clean data known for decomposition model (average over 5 rounds).

K out of N test:

Work until contaminated >17%

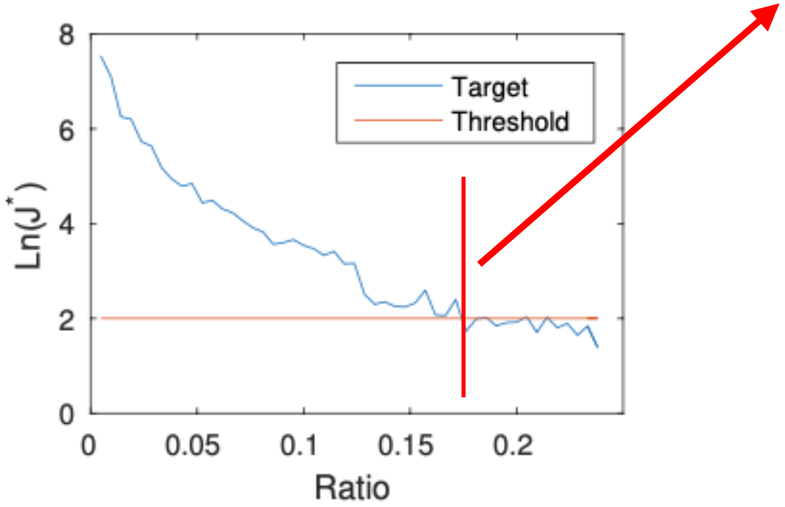


Figure 11: J^* of the target classes under contaminated clean data.

Comparison between SCAn and Previous

Offline setting (test on classes): Neural Cleanse, Activation Clustering

Table of FPR results.

	GTSRB												
	Offline						Online						
	SCAn		NC		AC		SCAn		SentiNet		STRIP		
TPR	A	T	A	T	A	T	A	T	A	T	A	T	S
95%	0%	0.15%	9.4%	95.3%	0%	77.5%	0.20%	0.32%	0.08%	82.6%	1.82%	75.4%	54.2%
99%	0%	0.15%	14.1%	100%	0%	90.6%	0.55%	1.10%	0.09%	83.6%	4.66%	95.7%	66.6%
99.5%	0%	0.19%	14.1%	100%	0%	90.6%	0.74%	1.82%	0.09%	84.1%	6.60%	96.9%	71.6%

Table of FPR results.

	CIFAR-10													
	Offline						Online							-
	SCAn		NC		AC		SCAn		SentiNet		STRIP			ABS
TPR	A	T	A	T	A	T	A	T	A	T	A	T	S	T
95%	0%	0%	5.36%	92.5%	0%	21.1%	0.19%	0.47%	0%	85.9%	0%	21.6%	11.3%	64.3%
99%	0%	0%	8.44%	99.2%	0%	47.8%	0.21%	0.48%	0.05%	93.3%	0%	71.8%	39.4%	97.1%
99.5%	0%	0%	8.45%	99.2%	0%	47.8%	0.34%	0.75%	0.05%	94.1%	0%	95.7%	74.6%	98.1%

Column A: source-agnostic backdoor

Column T: TaCT

Comparison between SCAn and Previous

Offline setting (test on classes): Neural Cleanse, Activation Clustering

Online setting (test on images): SentiNet, Strip

Table of FPR results.

	GTSRB												
	Offline						Online						
	SCAn		NC		AC		SCAn		SentiNet		STRIP		
TPR	A	T	A	T	A	T	A	T	A	T	A	T	S
95%	0%	0.15%	9.4%	95.3%	0%	77.5%	0.20%	0.32%	0.08%	82.6%	1.82%	75.4%	54.2%
99%	0%	0.15%	14.1%	100%	0%	90.6%	0.55%	1.10%	0.09%	83.6%	4.66%	95.7%	66.6%
99.5%	0%	0.19%	14.1%	100%	0%	90.6%	0.74%	1.82%	0.09%	84.1%	6.60%	96.9%	71.6%

Table of FPR results.

	CIFAR-10													
	Offline						Online							-
	SCAn		NC		AC		SCAn		SentiNet		STRIP			ABS
TPR	A	T	A	T	A	T	A	T	A	T	A	T	S	T
95%	0%	0%	5.36%	92.5%	0%	21.1%	0.19%	0.47%	0%	85.9%	0%	21.6%	11.3%	64.3%
99%	0%	0%	8.44%	99.2%	0%	47.8%	0.21%	0.48%	0.05%	93.3%	0%	71.8%	39.4%	97.1%
99.5%	0%	0%	8.45%	99.2%	0%	47.8%	0.34%	0.75%	0.05%	94.1%	0%	95.7%	74.6%	98.1%

Column A: source-agnostic backdoor

Column T: TaCT

Robustness of SCAn against Attacks

Multiple target-trigger attack: ✓

8 triggers

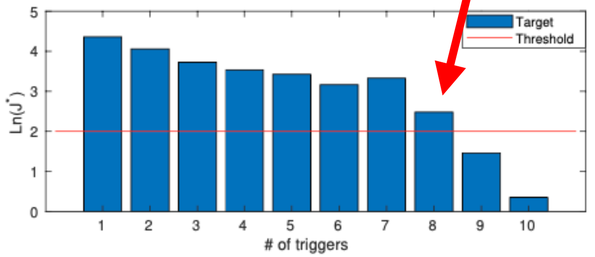


Figure 14: Minimum J^* of target classes under multiple target-trigger attack and 1% clean data are known (over 5 rounds).

18% for 21 triggers

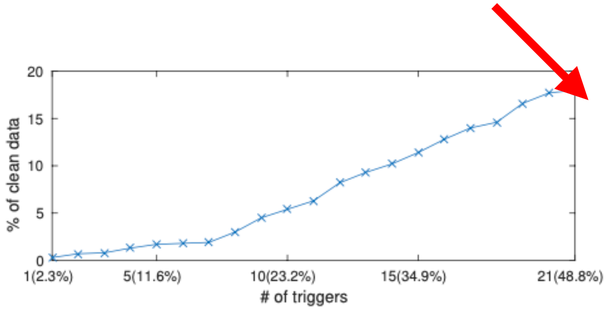
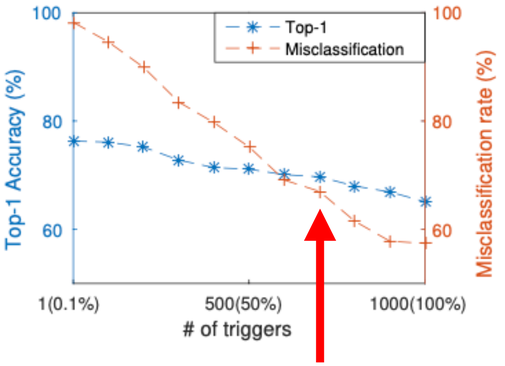


Figure 15: The amount of clean data required by decomposition model for defeating multiple target-trigger attacks on GTSRB.



ASR loss when the number of triggers increase.

Blending-trigger attack: ✓

Poison frogs attack: ✓

Adaptive Attacks against SCAn

Parameter inference attack:

$$r = R(x) = \mu_t + \varepsilon$$

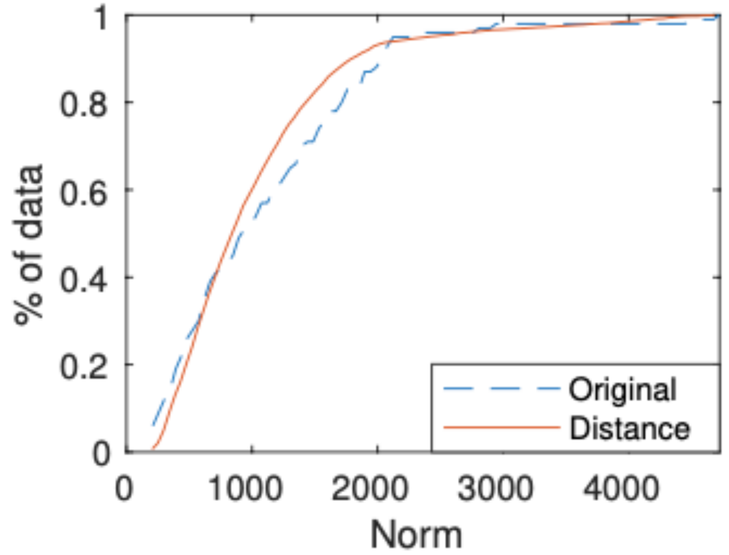


Figure 18: CDF of norms of S_ε and the distance between a couple S_ε.

Black-box trigger adjustment attack:

Ilyas, Andrew, et al. "Black-box adversarial attacks with limited queries and information." *International Conference on Machine Learning*. PMLR, 2018.

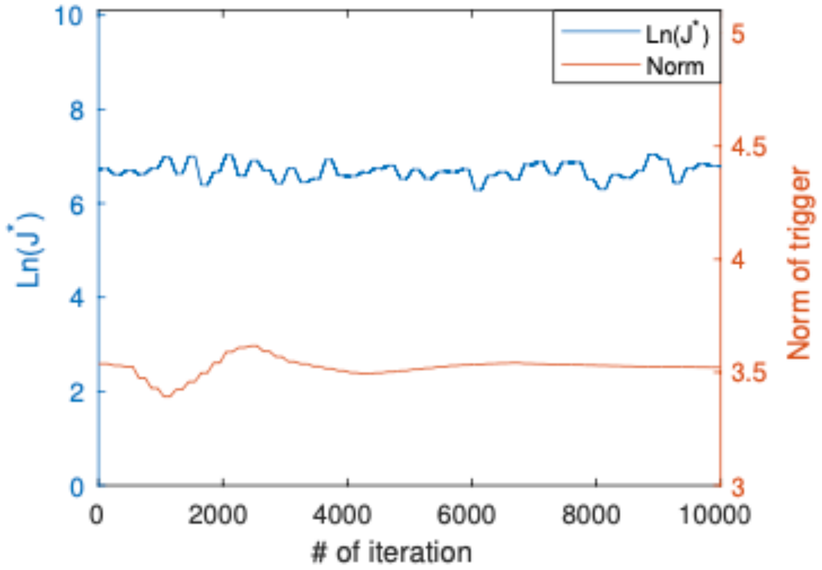


Figure 19: Statistics of black-box attacks (after moving-mean filtering).

Limitations

- Needs clean data set
- Needs presence of the trigger-carrying images
- Only evaluated on image classification tasks

Summary

- New understanding about the backdoor attack.

—— Dominant trigger is not necessary for the backdoor contamination attack.
A simple but powerful attack, TaCT, can bypass existing defences.

- New defence, SCAn.

——Introduce the global variant to detect inconsistency in representations.

Thanks !

Di Tang td016@ie.cuhk.edu.hk
Haixu Tang hatang@indiana.edu
XiaoFeng Wang xw7@indiana.edu
Kehuan Zhang khzhang@ie.cuhk.edu.hk