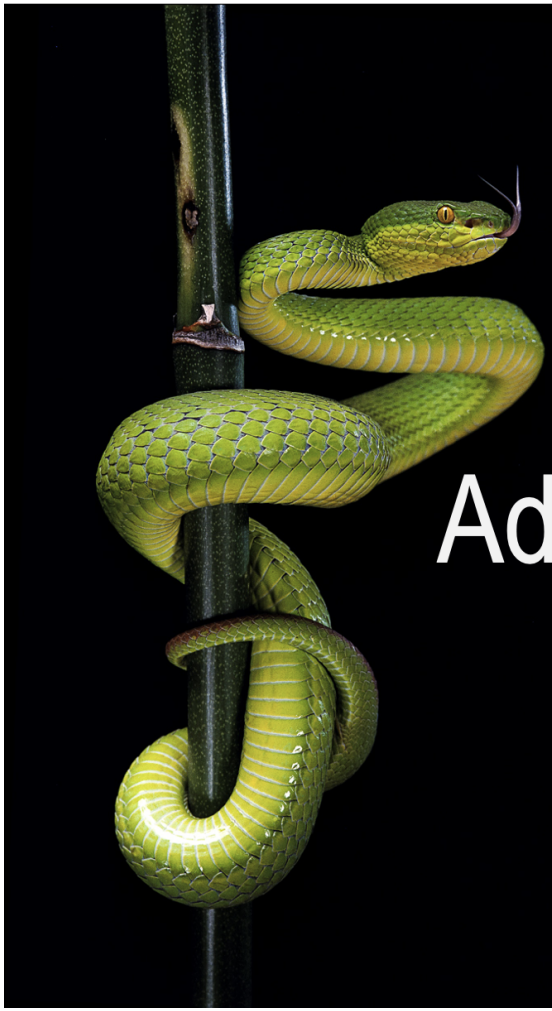


David J. Miller
Zhen Xiang
George Kesidis

Adversarial Learning and Secure AI

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Chapter 07

Post-Training Reverse-Engineering Defense (PT-RED) Against Perceptible Backdoors

Outline

- Introduction
- Two PT-REDs
  - P-PT-RED
  - Neural Cleanse (NC)
- Experiments
- Discussions

Introduction

- In this chapter, we consider perceptible backdoor patterns which are patch incorporated.
- As such, they can be triggered "physically," e.g., by putting a pair of glasses on a face or a model bird flying in the sky, where the glasses or the bird respectively are the backdoor pattern used to poison the training dataset.
- We will compare P-PT-RED described below with Neural Cleanse (NC) discussed in Chapter 6, where
- P-PT-RED uses a Maximum Achievable Misclassification Fraction (MAMF) statistic for purposes of detection.
- The DNN's class decision is the maximizer of the estimated class posterior for input $\underline{x}$,

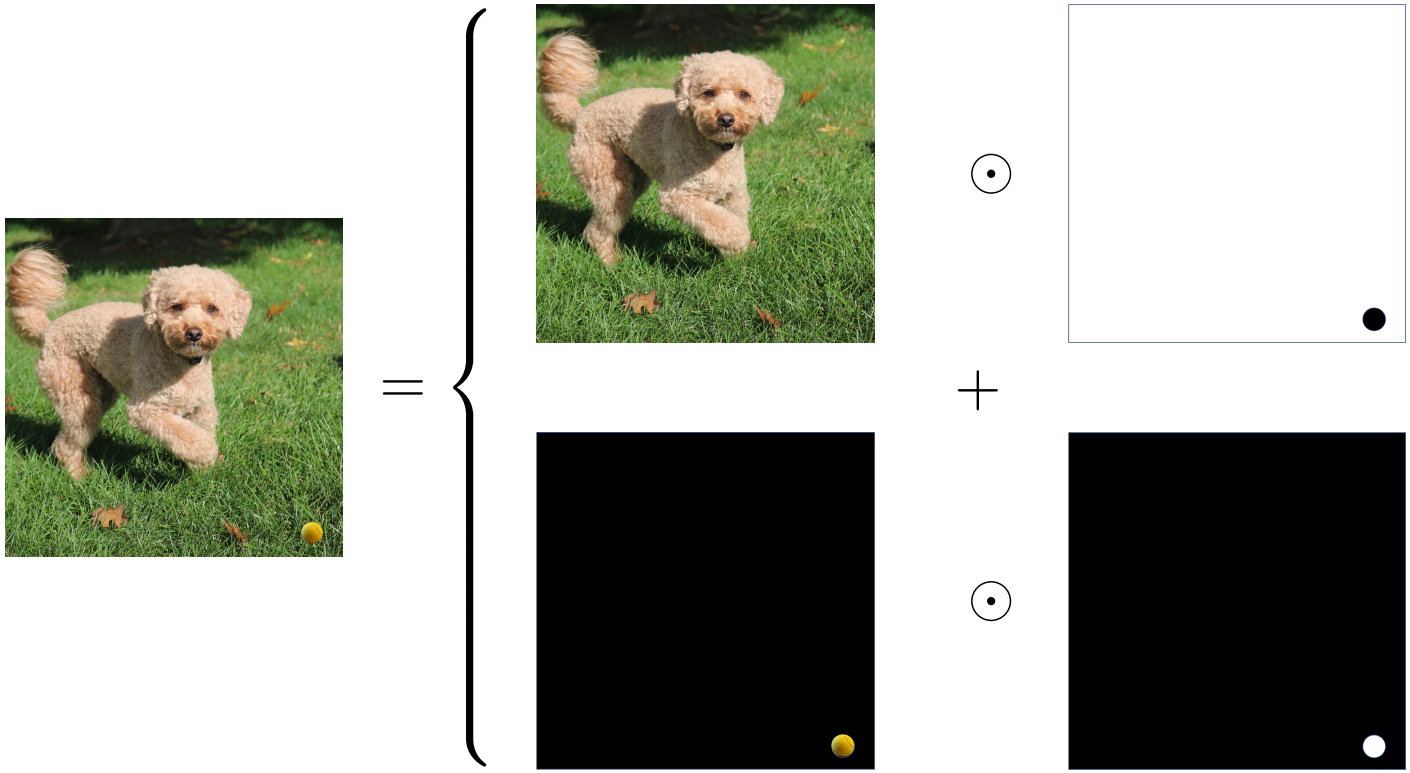
$$\hat{c}(\underline{x}) = \arg\max_k p_k(\underline{x}).$$

Perceptible Backdoors

- Craft images with perceptible (patch) backdoor patterns

$$\tilde{\underline{x}} = g(\underline{x}, \underline{v}, \underline{m}) = \underline{x} \odot (\underline{1} - \underline{m}) + \underline{v} \odot \underline{m}$$

- Example: dog with a tennis ball (not spatially fixed):



Properties of Perceptible Backdoors

- Spatial invariance of backdoor mapping:
  If a perceptible backdoor pattern is spatially distributed over backdoor training samples, *i.e.*, **scene-plausibly** placed so as to be most innocuous in each image, then the learned backdoor mapping will be spatially invariant in inducing targeted misclassifications on test samples.
- Robustness of perceptible backdoor patterns:
  It is unnecessary to use exactly the same perceptible backdoor pattern at test time to induce targeted misclassifications – so long as key features of the backdoor pattern are commonly present.

P-PT-RED: Detection Overview

- Key ideas
  - For backdoor class pairs, a pattern that induces a high misclassification fraction on the clean dataset requires only a relatively small spatial support **arbitrarily** located in the image.
  - For non-backdoor class pairs, much larger spatial support is required for a pattern achieving high misclassification fraction.
- Detection procedure
  - Pattern estimation (language description)
    - For each of the $K(K-1)$ class pairs (s, t), perform pattern estimation on a sequence of spatial supports of arbitrary shape (e.g. square) and arbitrary location (e.g. fixed to the top left corner) with increasing size.
    - Each pattern estimation maximizes the group misclassification fraction from s to t and obtains the "maximum achievable misclassification fraction" (MAMF).
  - Detection inference
    - Compute the average MAMF over the sequence of spatial supports.
    - Compare the maximum average MAMF over all class pairs with the detection threshold.

P-PT-RED: Pattern Estimation

- Notation
  - $\mathcal{D}_s$: set of clean images from source class s
  - $p_t(\cdot)$: posterior for class t
  - $r_{\min}, r_{\max} \in [0,1]$: minimum and maximum relative support width for the spatial supports to be considered
  - $\underline{M}_w \in \{0,1\}^{W \times W}$: mask with $w \times w$ ($w \in \mathbb{Z}^+$) square spatial support, W is the image width
- Objective function
  - For each class pair (s,t) and for each support width $w \in [\lceil r_{\min} \times W \rceil, \lfloor r_{\max} \times W \rfloor]$

$$\underset{\underline{v}_{stw}}{\text{maximize}} \quad \sum_{\underline{x} \in \mathcal{D}_s} p_t(g(\underline{x}, \underline{v}_{stw}, \underline{M}_w)),$$

    where g is the patch-embedding function.
  - Optimal solution $\underline{v}^*_{stw}$

P-PT-RED: Detection Inference

- Maximum achievable misclassification fraction (MAMF)
  - For each class pair (s, t), for each support width $w \in [\lceil r_{\min} \times W \rceil, \lfloor r_{\max} \times W \rfloor]$

$$\rho_{stw} = \frac{1}{|\mathcal{D}_s|} \sum_{x \in \mathcal{D}_s} \mathbf{1}\{\hat{c}(g(\underline{x}, \underline{v}^*_{stw}, \underline{M}_w) = t\}$$

    where $\mathbf{1}$ is the indicator function.
- Inference steps
  - Compute $\bar{\rho}_{st}$, the average MAMF (over all w being considered) for each class pair (s, t)
  - Compute

$$\rho^* = \max_{(s,t)} \bar{\rho}_{st}$$

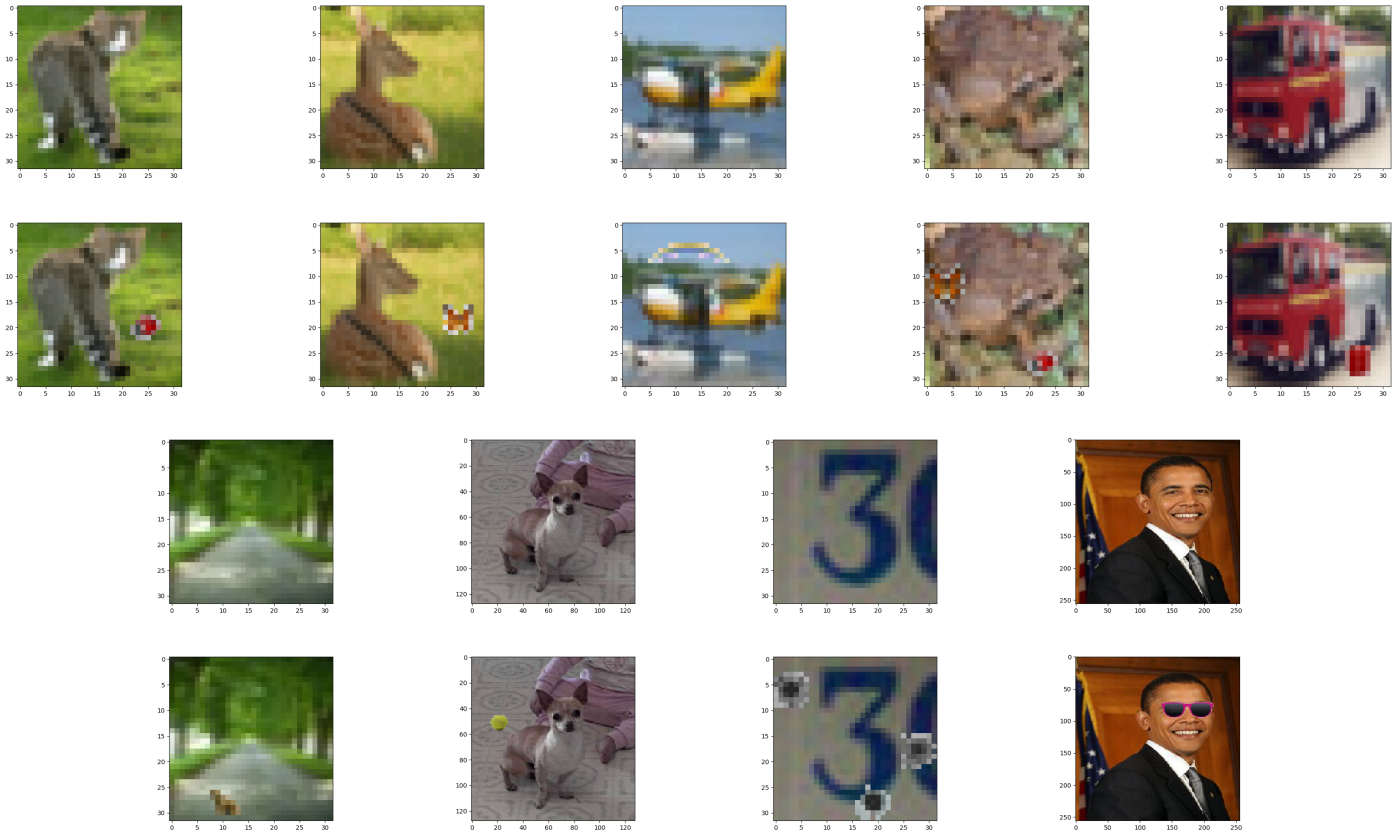
  - If $\rho^* > \pi$, π the detection threshold, an attack is detected; otherwise, there is no attack.
  - If an attack is detected, $(s^*, t^*) = \arg\max_{(s,t)} \bar{\rho}_{st}$ is inferred as a (source, target) class pair.

Experiments: Nine Attack Instances

	Attack A	Attack B	Attack C	Attack D	Attack E	Attack F	Attack G	Attack H	Attack I
Dataset	CIFAR-10	CIFAR-10	CIFAR-10	CIFAR-10	CIFAR-10	CIFAR-100	Oxford-IIIT	SVHN	PubFig
Image size	32×32	32×32	32×32	32×32	32×32	32×32	128×128	32×32	256×256
No. classes	10	10	10	10	10	100	6	10	33
Training size	50000	50000	50000	50000	50000	50000	900	73257	2782
Test size	10000	10000	10000	10000	10000	10000	300	26032	495
DNN structure	ResNet-18	ResNet-18	ResNet-18	ResNet-18	VGG-16	ResNet-34	AlexNet	ConvNet	VGG-16
Learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-4}	10^{-5}	10^{-3}	10^{-4}
Batch size	32	32	32	32	32	32	16	32	32
No. training epochs	200	200	200	200	200	200	120	80	120
Benchmark acc. (%)	86.7	88.1	86.7	87.6	87.9	71.9	88.7	89.2	76.0
Source class	"cat"	"deer"	"airplane"	"frog"	"truck"	"road"	"chihuahua"	"3"	"B. Obama"
Target class	"dog"	"horse"	"bird"	"bird"	"automobile"	"bed"	"Abyssinian"	"8"	"C. Ronaldo"
Backdoor Pattern	"bug"	"butterfly"	"rainbow"	"bug&butterfly"	"gas tank"	"marmot"	"tennis ball"	"bullet holes"	"sunglasses"
No. backdoor training images	150	150	150	150	150	100	50	500	40
Attack test acc. (%)	87.0	86.9	86.8	87.0	89.1	71.7	90.0	90.1	77.0
Attack succ. rate (%)	99.3	98.0	96.4	98.0	97.9	92.0	84.0	91.4	93.3

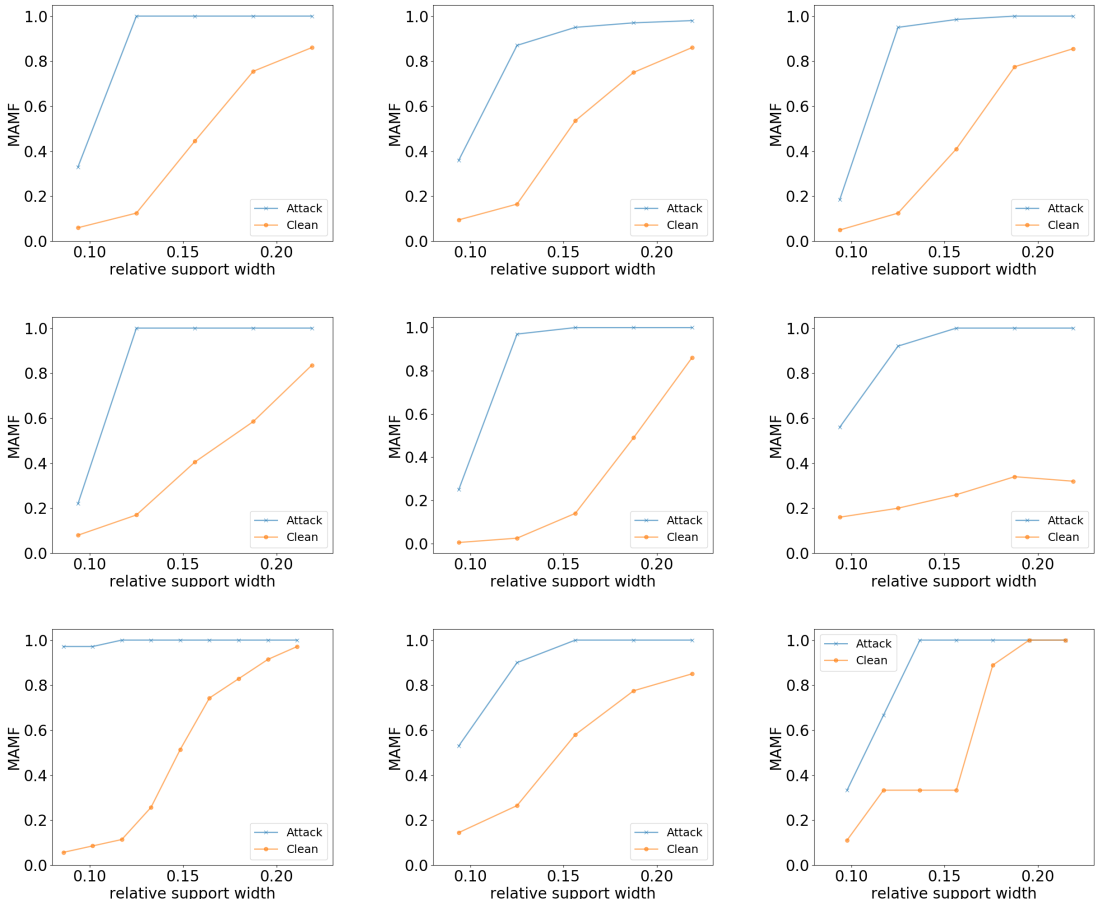
Experiments: Nine Attack Instances

- Example images with backdoor pattern embedded



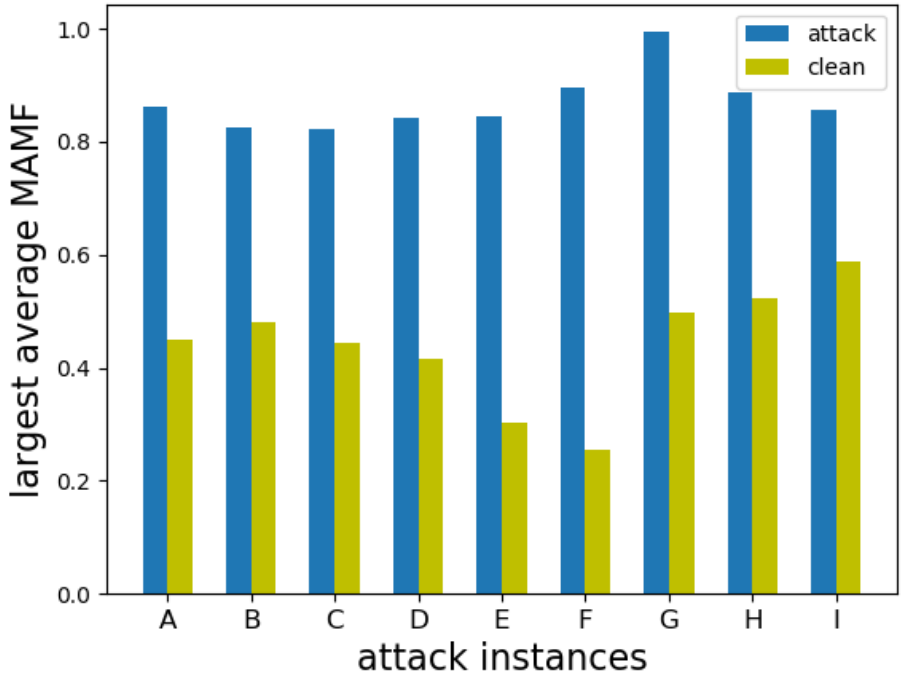
Experimental Results: MAMF of P-PT-RED

- MAMF statistics ($r_{\min} = 0.08$, $r_{\max} = 0.22$) vs. w/W



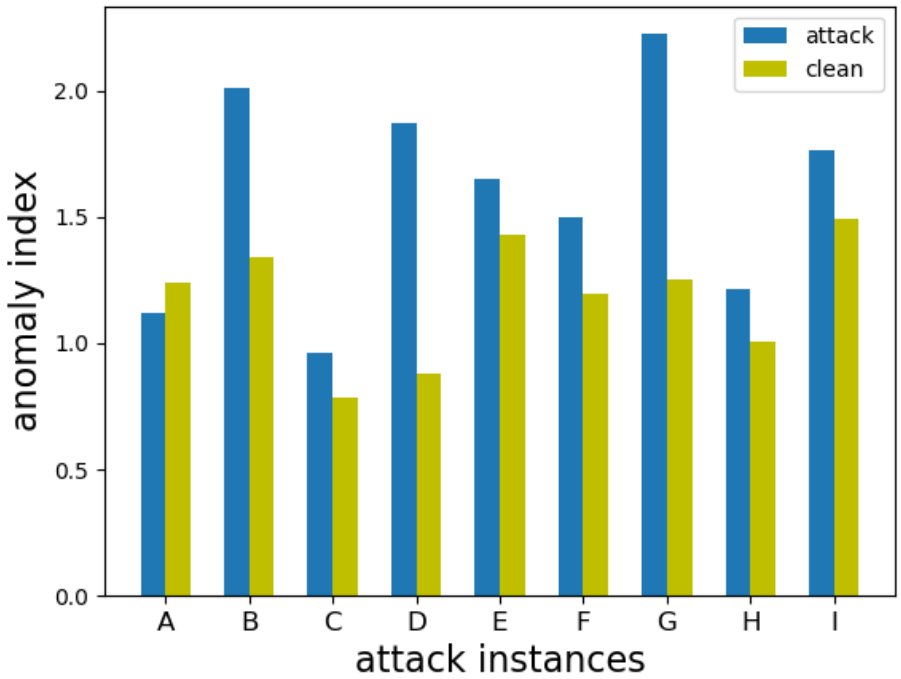
Experimental Results: MAMF of P-PT-RED

- Largest average MAMF statistics
  Any detection threshold $\pi \in [0.6, 0.8]$ will result in successful detection of all attacks, with no false detections.



Experimental Results: Neural Cleanse

- Recall from Chapter 6 that Neural Cleanse (NC)
  - assumes that all classes except for the target class are the source classes.
  - jointly estimates the pattern and the mask.



Experiments: Discussion

- For these experiments, MAMF is a clearer indicator of backdoors than NC's anomaly index.
- There are additional experimental results given in Chapter 7, *e.g.*, demonstrating robustness of P-PT-RED in the presence of additive noise or partial occlusion of the backdoor pattern.

With Permission, Figures Reproduced From

- Z. Xiang, D.J. Miller and G. Kesidis. Detecting Scene-Plausible Perceptible Backdoors in Trained DNNs without Access to the Training Set. Neural Computation, Feb. 2021. Z. Xiang, D.J. Miller and G. Kesidis. Revealing Perceptible Backdoors in DNNs, Without the Training Set, via the Maximum Achievable Misclassification Fraction Statistic. In Proc. IEEE MLSP, Sept. 2020. D.J. Miller, Z. Xiang and G. Kesidis. Adversarial Learning in Statistical Classification: A Comprehensive Review of Defenses Against Attacks. Proceedings of the IEEE 108(3), March 2020; http://arxiv.org/abs/1904.06292