David J. Miller
Zhen Xiang
George Kesidis

Adversarial Learning and Secure AI

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Chapter 06

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

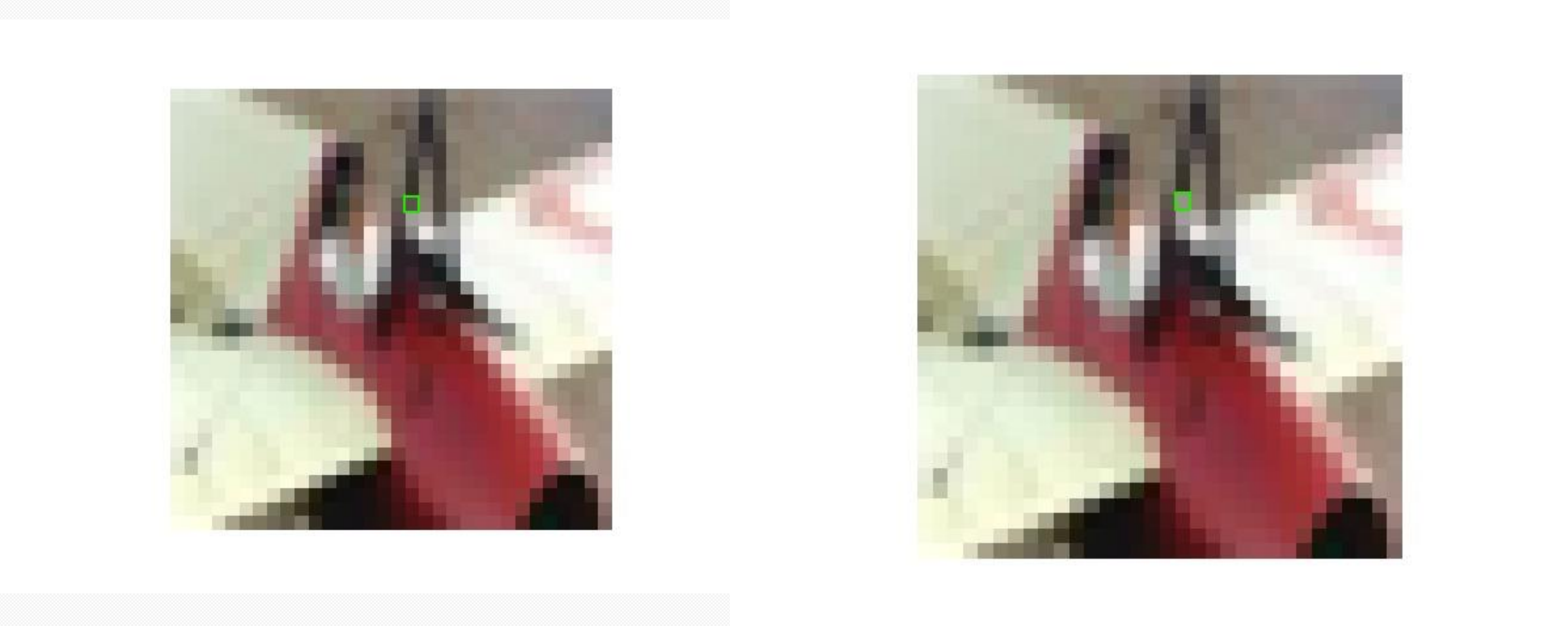
Outline

1. Introduction
2. Two PT-REDs
    - I-PT-RED
    - Neural Cleanse (NC)
3. I-PT-RED on embedded features
4. Experiments
5. Discussions

Recall backdoor attacks from Chapters 1 and 5

- The backdoor attacker poisons the training dataset by selecting clean samples drawn from a source class, somehow incorporating a subtle backdoor pattern into them, labelling them with a different target class, and contributing the thus poisoned samples to the training dataset.
- The backdoor is triggered at test time when the adversary takes a source class sample, incorporates the backdoor and submits it to the DNN, which classifies it to the (wrong) target class.
- The attack is feasible because of the typically huge amount of data that needs to be collected for deep learning and how this collection process may not be secure.
- The amount of data poisoning need not be large and the attacker need not be aware of the DNN architecture or the deep learning hyperparameters.

Imperceptible backdoor example



- Clean image (left) and perturbed image (right), with one pixel modified.

Backdoor Attack Configurations

- The attacker can also add samples from non-source class samples with the backdoor pattern embedded but correctly labeled to the source class.
- This has the effect of limiting “collateral damage” to non-source classes to create a more surgical backdoor attack.
- So, we can define the notion of an (ordered) attacked class pair (s,t), including when there are simultaneous backdoor attacks.
- In this chapter, we assume that the number of classes $C := K \gg 1$.
- Some defenses assume that all non-target classes are source classes of the attack, while
- other defenses just assume that there is a large number of not-attacked class pairs which are exploited to form a null model (by the I-PT-RED and L-PT-RED).

Recall the defense scenarios from Chapter 1 and TSC-RED from Chapter 5

- In the PT scenario, the defender has access to the DNN model but not to the training dataset.
- PT-REDs typically also assume access to a small clean dataset with representatives from each class.
- PT-REDs perform detection by attempting to reverse engineer the backdoor pattern.
- Note that the small clean dataset is insufficient for deep learning!
- Note that the “universal” (not RED) PT detector of Chapter 9 does not require such clean samples.

I-PT-RED

- For each input sample $\underline{x}$, let $p_t(\underline{x})$ be the probability that the sample is classified to class t by the DNN (softmax output).
- Defender has clean samples from each class i, $\mathcal{D}_i$.
- Need to search for a *small* perturbation $\underline{v}$, source classes s and target classes t, so that incorporation of $\underline{v}$ into clean samples from s causes a large proportion (π) of class decisions to be changed to t:

$$\frac{1}{|\mathcal{D}_s|} \sum_{\underline{x} \in \mathcal{D}_s} \mathbb{1}(f([\underline{x} + \underline{v}]_c) = t) \geq \pi$$

where $[\cdot]_c$ is domain-specific "clipping" operation and **here** f **=argmax**$_i$ p_i is the DNN's softmax class decision.

I-PT-RED (cont)

- To accomplish this: For each (s, t) class pair, we optimize a differentiable surrogate group-misclassification objective over **the perturbation vector $\underline{v}$.**
- This leads to a set of **potential backdoor perturbations**, one for each (s, t) pair.
- **Detection Inference:** We then employ p-values for confident anomaly detection of the *smallest* perturbations, which may correspond to imperceptible backdoor patterns.
- In this way, our detection approach
  - makes reliable detection of a backdoor attack
  - **estimates** the backdoor pattern $\underline{v}$
  - identifies the associated source class(es) s and target class t.

I-PT-RED

- **Hypothesis:** For a classifier being imperceptibly backdoor poisoned with (s^*, t^*), the required perturbation size to induce group misclassification from s^* to t^* should be **much smaller** than for other class pairs
- **Perturbation optimization** (for each (s, t))

$$\begin{aligned} \underset{\underline{v}}{\text{minimize}} \quad & d(\underline{v}) \\ \text{subject to} \quad & \frac{1}{|\mathcal{D}_s|} \sum_{\underline{x} \in \mathcal{D}_s} \mathbb{1}(f([\underline{x} + \underline{v}]_c) = t) \geq \pi \end{aligned}$$

  - $d(\cdot)$: the metric for measuring the size/energy of a perturbation $\underline{v}$
  - f **=argmax$_i$** p_i: DNN's softmax class decision
  - $[\cdot]_c$: (domain-specific) clipping operation
  - $\mathcal{D}_s$: the set of samples from class s
  - π: the target fraction of misclassification
- **Detection inference** discussed below.

I-PT-RED

- Static, using all of $\mathcal{D}_s$:

$$J_{st}(\underline{v}) = \frac{1}{|\mathcal{D}_s|} \sum_{\underline{x} \in \mathcal{D}_s} p_t([\underline{x} + \underline{v}]_c),$$

  - Dynamic "perceptron" objective:

$$J_{st-\mathrm{p}}(\underline{v}) = \frac{1}{|\hat{\mathcal{D}}_s(\underline{v}, t)|} \sum_{\underline{x} \in \hat{\mathcal{D}}_s(\underline{v}, t)} p_t([\underline{x} + \underline{v}]_c),$$

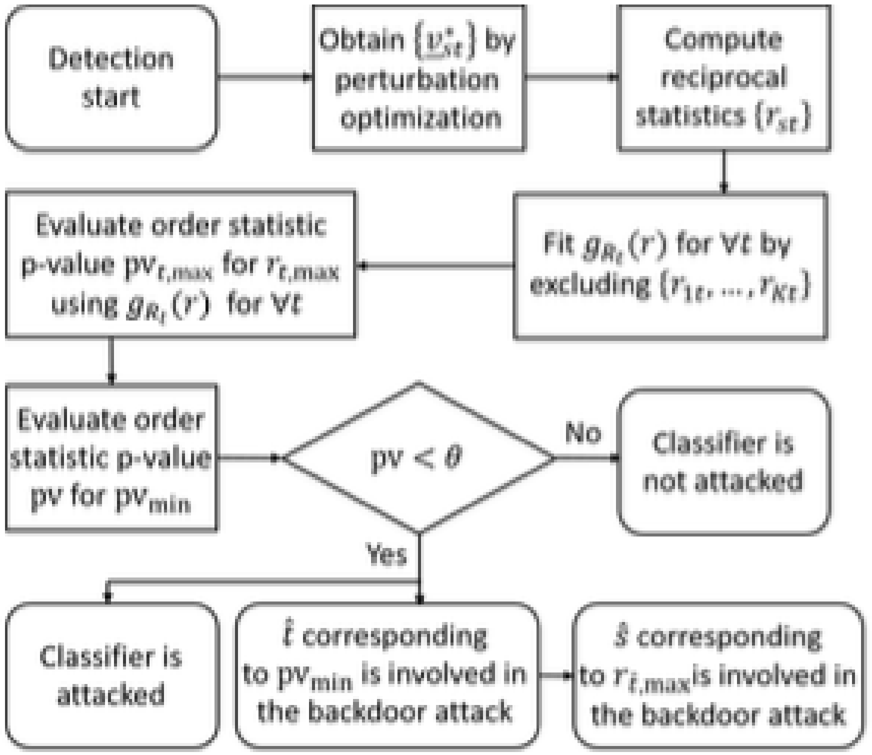
  where $\hat{\mathcal{D}}_s(\underline{v}, t) = \{\underline{x} \in \mathcal{D}_s : f([\underline{x} + \underline{v}]_c) \neq t\}$.
  - Static, using only initially correctly classified samples:

$$J_{st-\mathrm{c}}(\underline{v}) = \frac{1}{|\hat{\mathcal{D}}_s|} \sum_{\underline{x} \in \hat{\mathcal{D}}_s} p_t([\underline{x} + \underline{v}]_c),$$

  where $\hat{\mathcal{D}}_s = \{\underline{x} \in \mathcal{D}_s : f(\underline{x}) = s\}$.

Detection Inference

- Detection statistics: reciprocals $\{r_{st} = d(\underline{v}^*_{st})^{-1}\}$ for metric $d(\cdot)$
- Procedure:



- Null Distributions: Gamma, inverse Gaussian, etc.

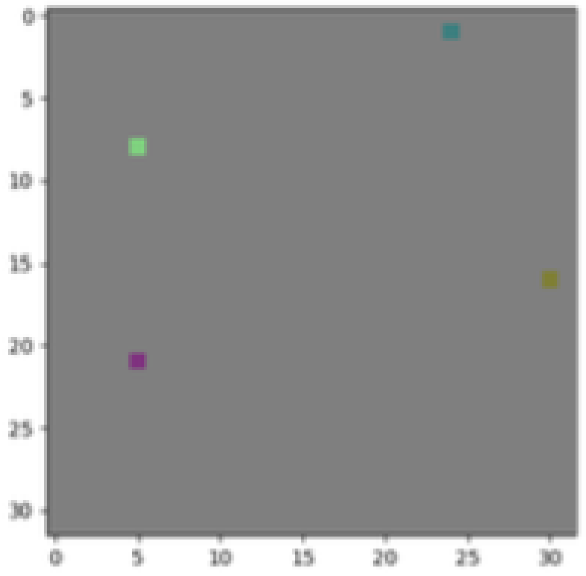
CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

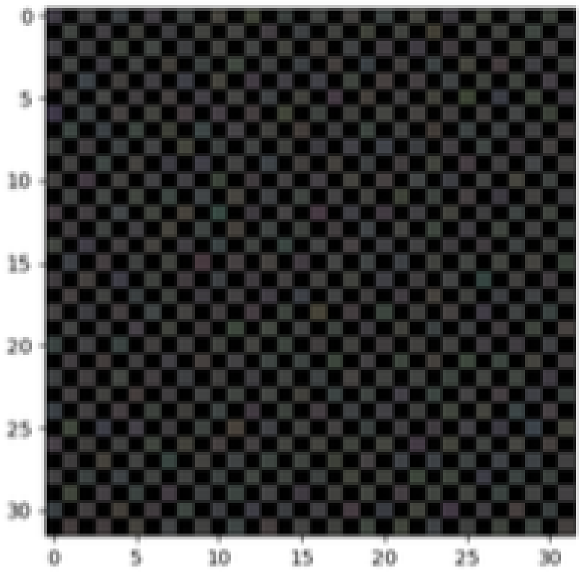
Some Inference Approaches

- $C(C-1)$ reciprocal perturbation norms with C the number of classes.
- In one approach
  - trial remove $C-1$ reciprocals associated with each target class t
  - form a "null" $\nu_{-t} = g_{R_t}$ with the $(C-1)^2$ remaining reciprocals
  - assess the joint likelihood of the removed reciprocals, p_t, using ν_{-t}
  - expect target class t of a backdoor to have much smaller p_t than rest
- If there is collateral damage, can simply consider the target-class distribution of the C largest reciprocals (an order statistic on an order statistic)
  - without a backdoor, expect uniform distribution
  - with a backdoor, expect significant mode (small p-value) at associated target class t
  - can also adjust reciprocals to account for (inherent, available) class confusion matrix information

I-PT-RED Experiments

- Dataset: CIFAR-10
- Backdoor Patterns:
  - 4-pixel perturbation
  - Global perturbation (chessboard pattern) – amplified to be visible...





I-PT-RED Experiments (cont)

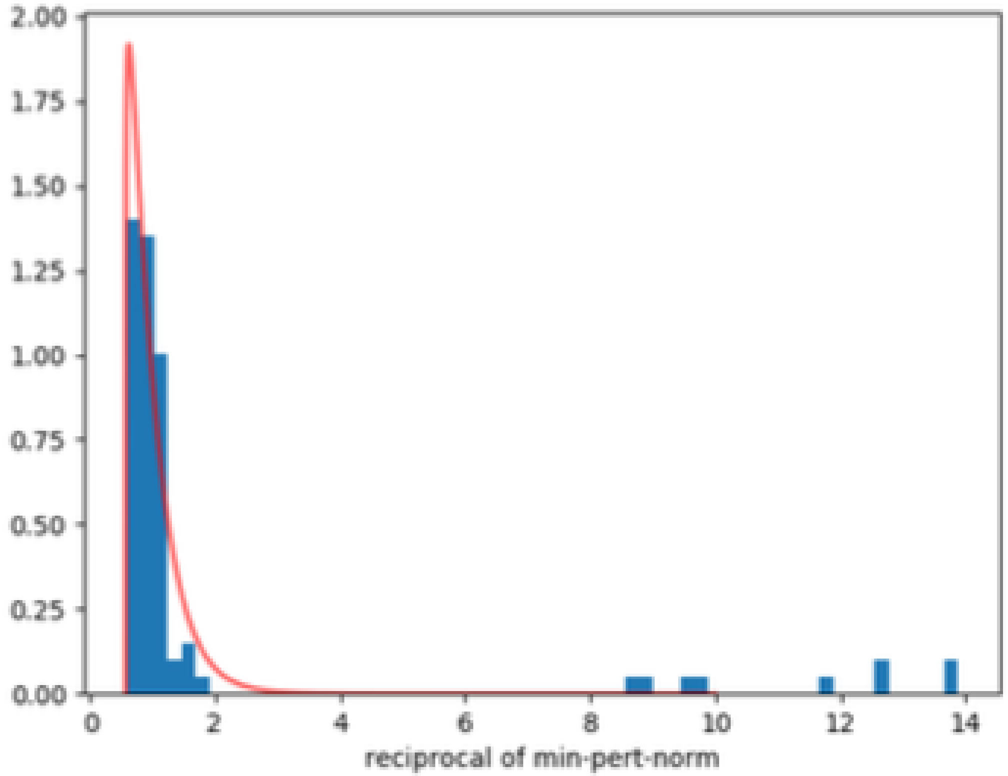
- DNN Classifiers for Test
  4 groups, 25 classifiers per group
  - **BD-P-S**: 4-pixel perturbation ($||\underline{v}^*||_2 = 0.6$); single source class
  - **BD-G-S**: global perturbation ($||\underline{v}^*||_2 = 0.2$); single source class
  - **BD-G-M**: global perturbation ($||\underline{v}^*||_2 = 0.2$); nine source classes
  - **Clean**: clean classifiers

	BD-P-S	BD-G-S	BD-G-M	Clean
min attack success rate	0.858	0.941	0.979	N.A.
min test accuracy	0.907	0.906	0.908	0.910

- Attacker does not want to degrade accuracy on clean (backdoor-free) test patterns

I-PT-RED Detection Example

- Chessboard backdoor pattern (global perturbation) on CIFAR-10, with "collateral damage" to all other source classes



Neural Cleanse (NC)

- Backdoor generating mechanism:

$$x * (1 - m) + v * m,$$

  where
  - x is the clean image,
  - m is the mask,
  - v is a pattern, and
  - $*$ is element-wise multiplication.
- The key idea of detection:
  - If there is a backdoor attack, the backdoor pattern should occupy a relatively small spatial location.
  - Else, more pixels will need to be *replaced* to reach a high group misclassification fraction.

Neural Cleanse (cont)

1. Estimate a mask and a pattern for each putative target class t.
   1. Use images from all classes except those from the putative target class.
   2. Minimize the following objective function over m and v:

$$- \sum_{x \in \mathcal{D}_{-t}} p_t([x * (1 - m) + v * m]_c) + \lambda * |m|,$$

   where $|m|$ is the L1 norm of the mask.
2. Detection inference.
   1. One statistic $|m|$ for each class.
   2. Obtain the deviation to the median and normalize by the median absolute deviation (MAD).
   3. If there is an abnormally small mask, decide there is an attack.

NC & I-PT-RED (AD) Experiments

- I-PT-RED (AD) Variants to be Tested
  - **AD-J-P**: basic objective function; principal detection inference approach; L2 norm for detection
  - **AD-Jp-P**: perceptron **objective;** principal detection inference approach; L2 norm for detection
  - **AD-Jc-P**: initially correctly classified samples; principal detection inference approach; L2 norm for detection
  - **AD-J-C**: basic objective function; detection inference with class confusion correction; L2 norm for detection
  - **AD-J-P-L1**:basic objective function; principal detection inference approach; L1 norm for detection
- NC Variants to be Tested
  - **NC-L1-1.5**: L1-regularized objective function with $\lambda = 1.5$; L1 norm for detection
  - **NC-L2-1.5**: L2-regularized objective function with $\lambda = 1.5$; L2 norm for detection
  - **NC-L1-1.0**: L1-regularized objective function with $\lambda = 1.0$; L1 norm for detection

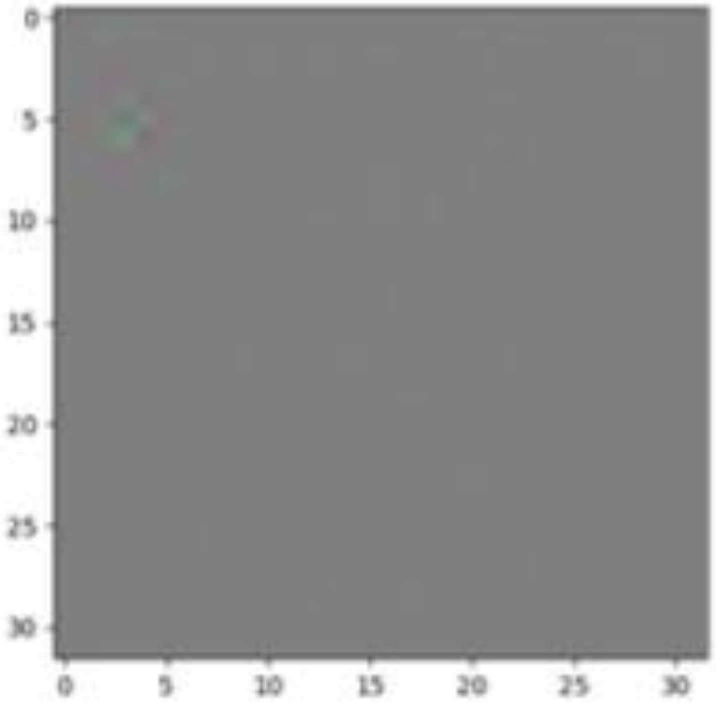
Defense experimental results

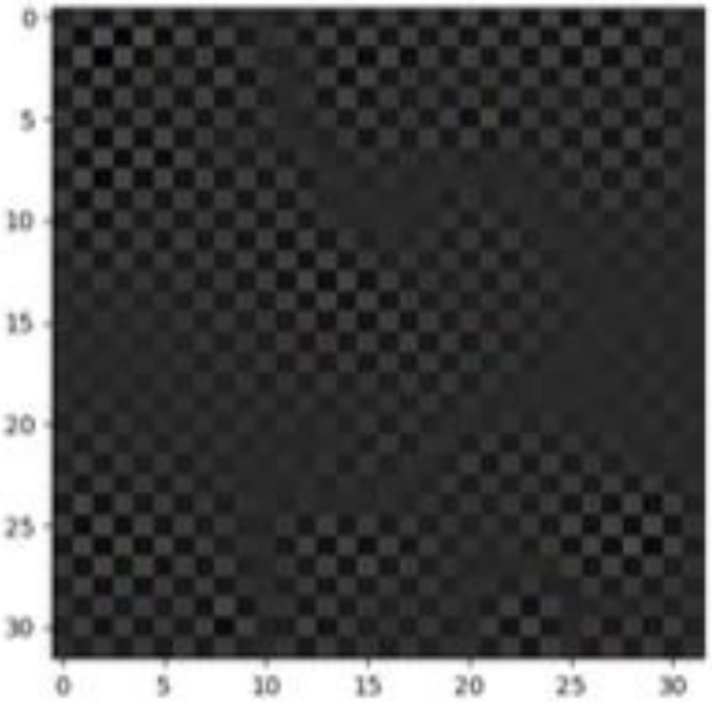
- Detection Criterion
  - NC: detection of the presence of the attack; the class label corresponding to the most extreme anomaly index is the estimated backdoor target class label t^*
  - AD: detection of the presence of the attack; detection of source and target class (corresponding to the most extreme detection statistic) involved in the attack; estimation of the backdoor pattern
- Results

	BD-P-S	BD-G-S	BD-G-M	Clean
AD-J-P	0.96	0.96	1.00	1.00
AD-Jp-P	1.00	0.92	1.00	1.00
AD-Jc-P	1.00	0.92	1.00	1.00
AD-J-C	0.84	0.96	1.00	1.00
AD-J-P-L1	1.00	0.92	1.00	1.00
NC-L1-1.5	0.36	0.16	1.00	0.84
NC-L2-1.5	0.56	0.64	1.00	0.72
NC-L1-1.0	0.40	0.28	1.00	0.76

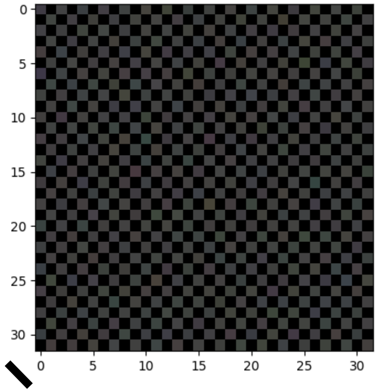
I-PT-RED (AD) reverse engineering

- Estimated backdoor patterns are similar to actual ones used !!





I-PT-RED (AD) Summary

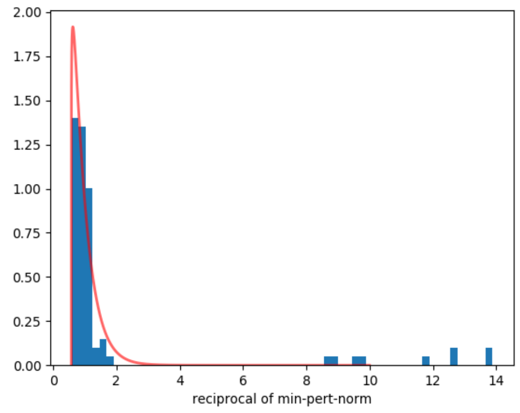


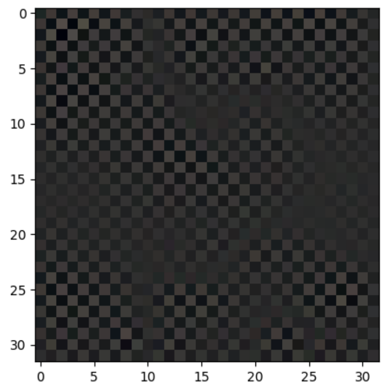
Imperceptible backdoor pattern incorporated into training samples & class changed from source to target

trained decision-making system, e.g. mixture model or DNN classifier

post-training

Inverse perturbation magnitudes for all class pairs under I-PT-RED





Reverse engineered backdoor pattern by I-PT-RED

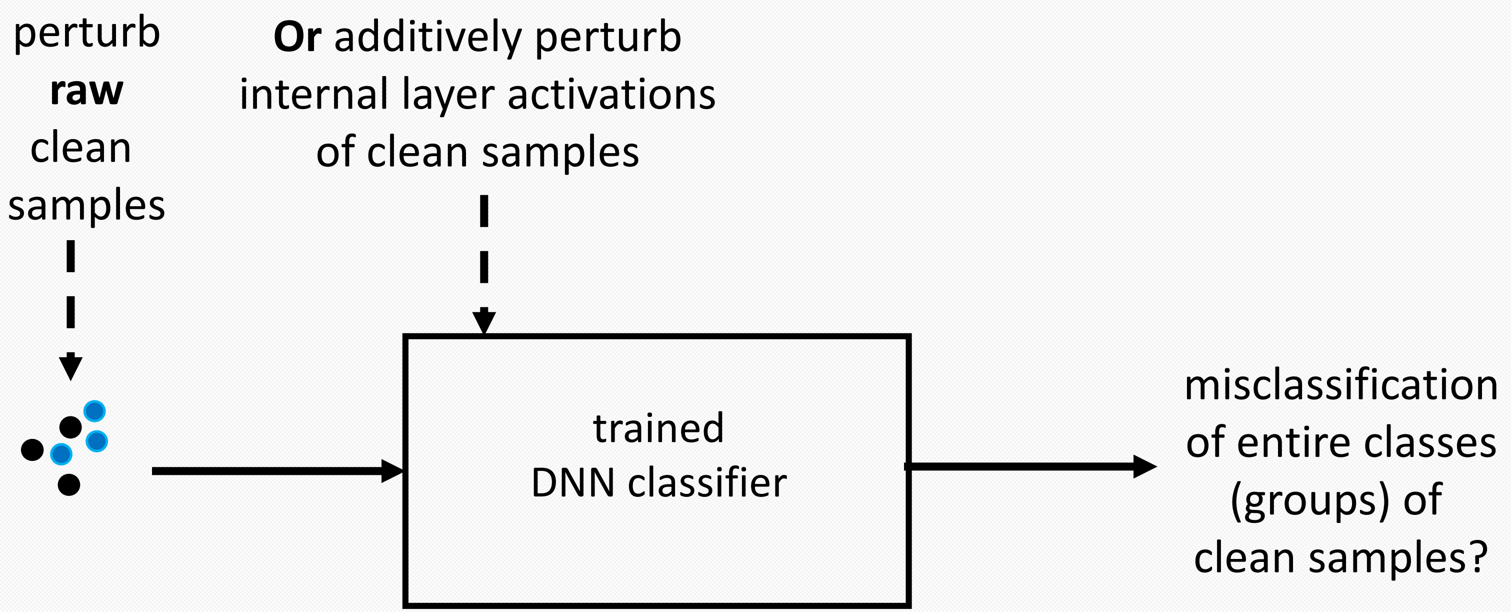
Discussion

- The above experiments involved attacks where the imperceptible backdoor pattern was **additively incorporated**, consistent with I-PT-RED's method of reverse engineering.
- NC's approach is based on a patch or blended incorporation to be considered in Chapter 7.
- NC also implicitly assumes "all-to-one" attack scenarios while I-PT-RED's detection inference considers the possibility of more surgical attack configurations, i.e., "X-to-1" (even plural X-to-1 attacks if there are enough unattacked class pairs to create an accurate null).
- Other backdoor defenses are described in Chapter 6.
- E.g., L-PT-RED is a low complexity version of I-PT-RED whose computation scales linearly with number of classes, see Section 6.5.
- In Chapter 8, we consider a RED suitable for the two-class (K=2) case.

I-PT-RED/L-PT-RED based on Embedded Features

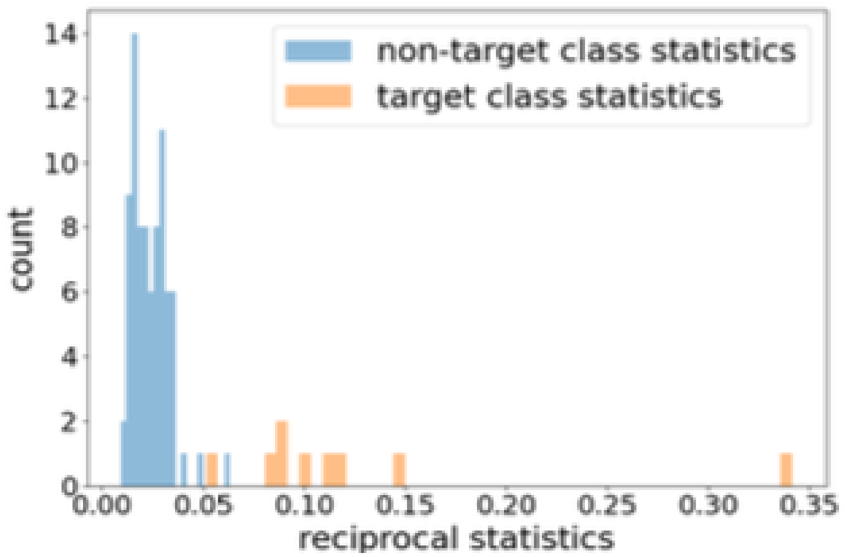
- DNN neural activations are based on weighted sums of those of previous layers.
- For a class pair, the common putative-backdoor perturbation $\underline{v}$ could be added to an **embedded** feature vector $\underline{h}(\underline{x})$ of the DNN rather than the input features $\underline{x}$, see Section 6.4.4.
- This allows for consideration of non-additive methods of incorporation of the backdoor and also addresses discrete input feature spaces.
- Note that the corresponding, possibly sample-specific, input perturbation $\underline{u}(\underline{x})$ can be found by back-propagation w.r.t. the input variables to minimize $\|\underline{h}(\underline{x}+\underline{u}) - (\underline{h}(\underline{x}) + \underline{v})\|^2$ over feasible $\underline{u}$.
- Also note that sample-specific perturbations include patches, see Chapter 7.

I-PT-RED acting either on the raw input layer or on an embedded layer



Example detection by Embedded I-PT-RED of non-additive backdoor incorporation mechanisms

- The following example is for detection of *multiplicative* chessboard backdoor incorporation into images by AD-J-P with generalized objective function.
- For purposes of detection, *additive* perturbations were applied to the first max-pooling layer activations when the clean data samples were input.



BNA Backdoor Mitigation

- One can leverage REDs to mitigate backdoors by, e.g., adjusting the batch-normalization layers situated after the layer used by the RED for detection; see
- X. Li, Z. Xiang, B. Li, D.J. Miller and G. Kesidis. Backdoor Mitigation via Reversing Activation Distribution Alteration. *preprint*, 2022.
- Also see Chapter 9 for a “universal” (non-RED) approach to backdoor detection and mitigation.

Discussion: Unsupervised Detection Rules

- Unsupervised anomaly detectors require setting a detection threshold, e.g., on a p-value.
- In principle, the threshold can be set (in an unsupervised fashion) to control the false positive detection rate on the small set of clean data available to the defender.
- Sometimes the p-value is set to a "standard" value, e.g., mean + two standard deviations of the null motivated by the 95% confidence interval of a Gaussian distribution.
- But consider a rule based on an arbitrarily chosen threshold applied to some measure of the size of the perturbation (putative backdoor), e.g., its support is <5% of the total image size, and another on the corresponding attack success rate (e.g., > 90%) as measured on a clean dataset.
- Such detection thresholds may not perform well: E.g., for the example of the previous bullet, if the perturbation is 7% of the image and has ASR of 88% then does it make sense to not deem it a backdoor?

With Permission, Figures Reproduced From

- Z. Xiang, D.J. Miller, G. Kesidis. L-RED: Efficient Post-Training Detection of Imperceptible Backdoor Attacks without Access to the Training Set. In Proc. IEEE ICASSP, June 2021; https://arxiv.org/abs/2010.09987
- Z. Xiang, D.J.~Miller, and G. Kesidis. Detection of Backdoors in Trained Classifiers Without Access to the Training Set. IEEE Trans. on Neural Networks and Learning Systems (TNNLS) 33(3), March 2022 (Dec. 2020 online); shorter version in Proc. IEEE ICASSP 2020; https://arxiv.org/abs/1908.10498
- 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
- Z. Xiang, D.J. Miller and G. Kesidis. A Benchmark Study of Backdoor Data Poisoning Defenses for Deep Neural Network Classifiers and A Novel Defense. In Proc. IEEE MLSP, Pittsburgh, Sept. 2019.