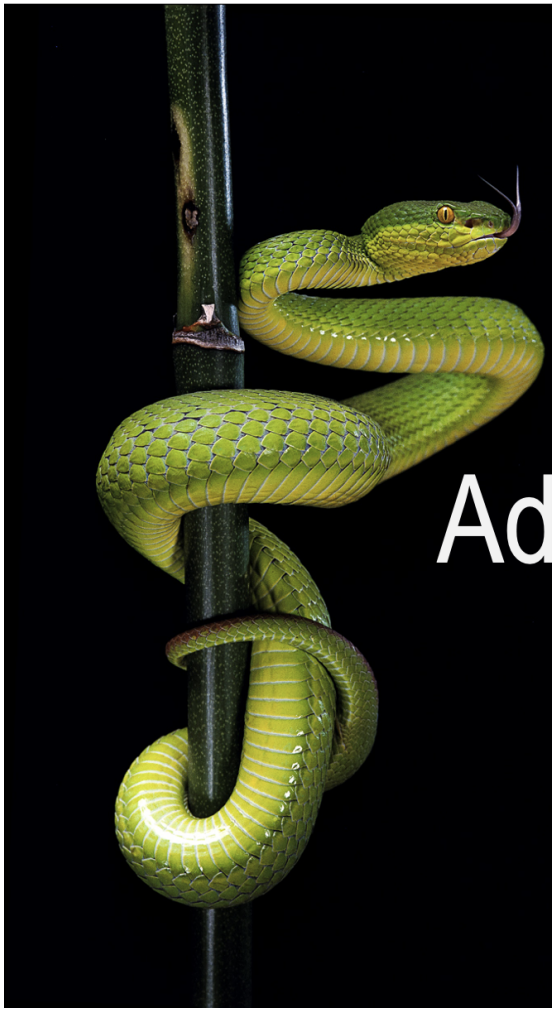


David J. Miller
Zhen Xiang
George Kesidis

Adversarial Learning and Secure AI

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Chapter 08
Transfer PT-RED (T-PT-RED) Against Backdoors

Outline

- Motivation
  - Two-class ($K = 2$) case
  - Different source-target-class attack configurations
- Expected Transferability Statistic of T-PT-RED
- Experiments
- Discussion

T-PT-RED Backdoor Detection via Expected Transferability (ET)

Key ideas

- Process each class pair independently: obtain an expected transferability (ET) statistic independently for each class pair, then compare ET with a detection threshold.
- No need for null distribution estimation.
- There is a common threshold on ET to determine if a class is a backdoor target class, irrespective of the classification domain or particulars of the attack.
  $\Rightarrow$ No need for domain-specific supervision.
- Works when there are only two classes, $K = 2$.

Backdoor Detection Using Expected Transferability (ET)

Definition of ET

- ϵ-solution set: For any $\underline{x}$ from any class, the ϵ-solution set is:

$$\mathcal{V}_\epsilon(\underline{x}) \triangleq \{\underline{v} \mid \|\underline{v}\|_2 - \|\underline{v}^*\|_2 \leq \epsilon, f(\underline{x}+\underline{v})) \neq f(\underline{x})\},$$

where $\underline{v}^*$ is the global optimal solution to

$$\underset{v}{\text{minimize}} \|\underline{v}\|_2 \qquad \text{subject to } f(\underline{x}+\underline{v}) \neq f(\underline{x})$$

and $\epsilon > 0$ is the "quality gap" of practical solutions to the same problem, which is usually small for existing methods.

Backdoor Detection Using Expected Transferability (ET)

Definition of ET (cont'd)

- ϵ-transferable set: The ϵ-transferable set for any sample $\underline{x}$ and $\epsilon > 0$ is defined by

$$\mathcal{T}_\epsilon(\underline{x}) \triangleq \{\underline{y} \in \mathcal{X} \mid f(\underline{y}) = f(\underline{x}), \exists \underline{v} \in \mathcal{V}_\epsilon(\underline{x}) \,\text{s.t.}\, f(\underline{y}+\underline{v}) \neq f(\underline{y})\}.$$

- ET statistic: For any class $i \in \mathcal{Y} = \{0, 1\}$ and $\epsilon > 0$, considering independent random samples $\underline{X}, \underline{Y} \sim P_i$ with P_i the sample distribution of class i, the ET statistic for class i is defined by

$$\text{ET}_{i,\epsilon} \triangleq \mathbb{E}_{\underline{X} \sim P_i}\left[\text{P}(\underline{Y} \in \mathcal{T}_\epsilon(\underline{X}) \mid \underline{X})\right].$$

Backdoor Detection Using Expected Transferability (ET)

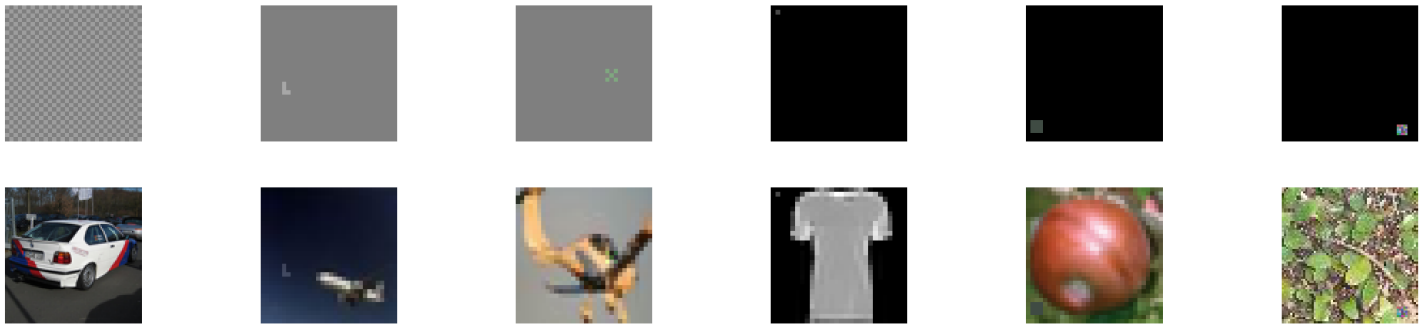
Detection method

- Properties of ET: There exists a constant detection threshold (see Theorem 10)
  - If class $i \in \mathcal{Y} = \{0, 1\}$ is not backdoor target class, we will have $\mathrm{ET}_{1-i,\epsilon} \leq \frac{1}{2}$
  - Otherwise, we will have $\mathrm{ET}_{1-i,\epsilon} > \frac{1}{2}$
- Detection procedure
  - Estimate ET for each class
  - Check if there is any ET statistic greater than $\frac{1}{2}$

Backdoor Defense Post-Training

ET – experiments

- Dataset: CIFAR-10, CIFAR-100, STL-10, TinyImageNet, FMNIST , MNIST
- Backdoor pattern: both additive perturbation and patch replacement, examples:



Backdoor Defense Post-Training

ET – experiments (cont'd)

- Detection accuracy using ET (2-class domains, ET threshold $\frac{1}{2}$)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
RE-AP	45/45	18/20	16/20	17/20	20/20	20/20	n/a	n/a	n/a	n/a
RE-PR	n/a	n/a	n/a	n/a	n/a	n/a	45/45	20/20	19/20	19/20

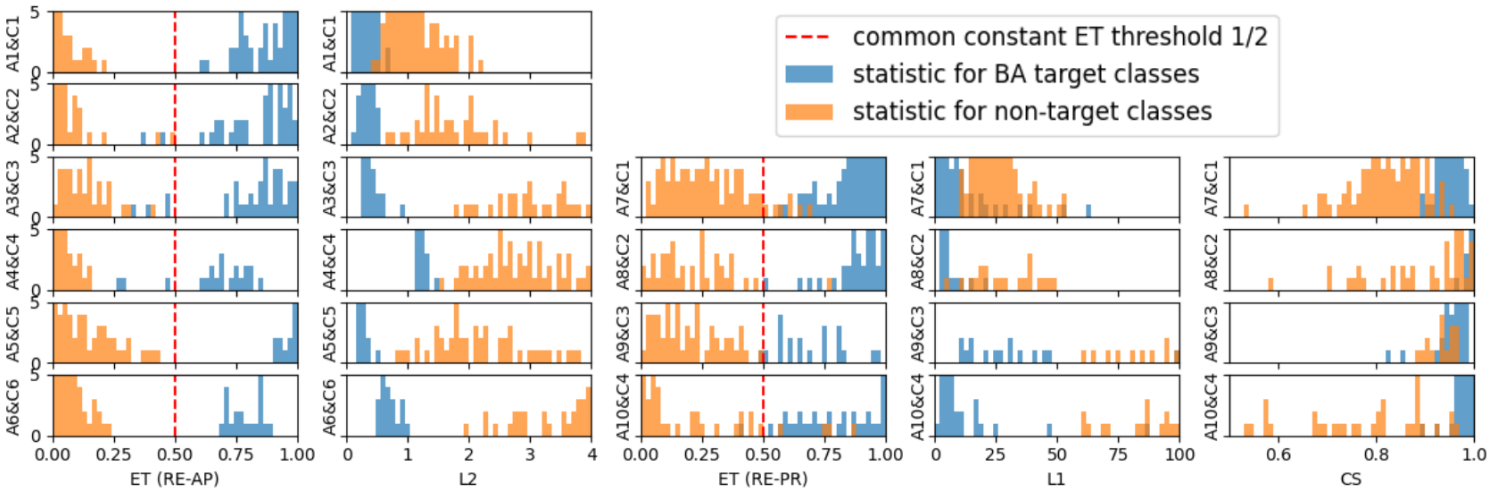
	C_1	C_2	C_3	C_4	C_5	C_6
RE-AP	45/45	20/20	20/20	20/20	20/20	20/20
RE-PR	39/45	19/20	20/20	16/20	18/20	19/20

  - $A_1 \sim A_6$: attack instances with additive perturbation (AP) backdoor patterns
  - $A_7 \sim A_{10}$: attack instances with patch replacement (PR) backdoor patterns
  - $C_1 \sim C_6$: clean instances
  - RE-AP: T-PT-RED with RE backdoor pattern of I-PT-RED
  - RE-PR: T-PT-RED with RE backdoor pattern of P-PT-RED

Backdoor Defense Post-Training

ET – experiments (cont'd)

- Comparison between ET and other detection statistics



- L_1: l_1 norm of estimated mask of P-PT-RED
- L_2: l_2 norm of estimated perturbation of I-PT-RED
- CS: cosine similarity [R. Wang et al. ECCV '20]

Discussion

- T-PT-RED can obviously also be applied to the case of more than two classes ($K > 2$).
- As I-PT-RED, T-PT-RED/RE-AP can also work with perturbations applied to embedded features.
- As I-PT-RED, T-PT-RED can detect X-to-1 and all-to-all attacks.
- Since T-PT-RED works with sample-specific putative backdoor patterns, it's possible that it can detect different simultaneous backdoors with the same associated source and target classes.

With Permission, Figures Reproduced From

- Z. Xiang, D.J. Miller and G. Kesidis. Post-Training Detection of Backdoor Attacks for Two-Class and Multi-Attack Scenarios. In Proc. ICLR, Apr. 2022.