CBD: A Certified Backdoor Detector Based on Local Dominant Probability

Presenter: Zhen Xiang
Authors: Zhen Xiang, Zidi Xiong, Bo Li
Secure Learning Lab
University of Illinois Urbana-Champaign

Backdoor Attack

Elements

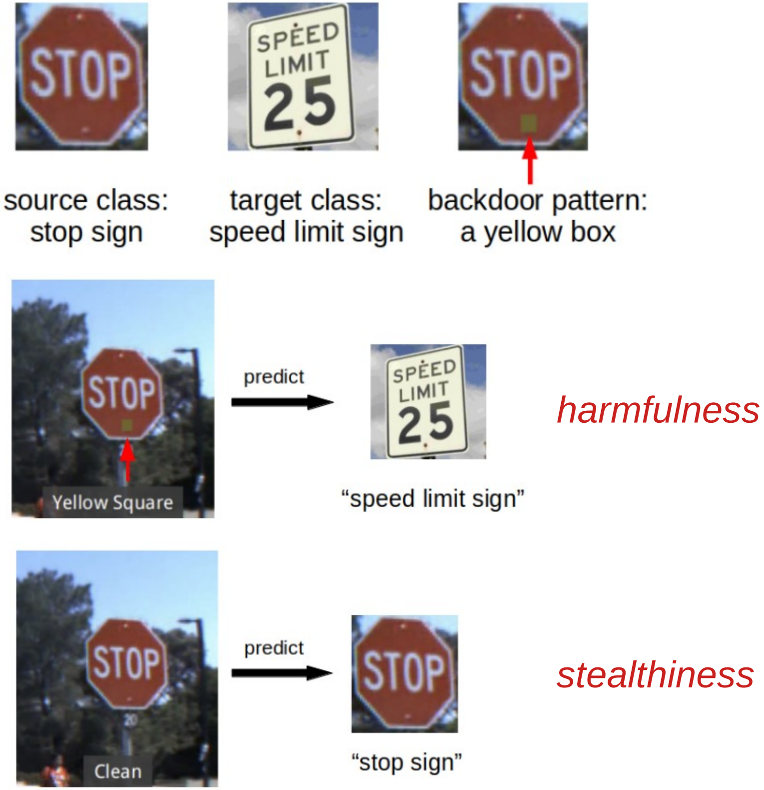
- A set of source classes
- A target class
- A backdoor trigger/pattern

Goals

- Test sample from source class + trigger

  ➡ target class
- Clean test sample

  ➡ designated class



T. Gu, B. D.-Gavitt, S. Garg, BadNets: Identifying vulnerabilities in the machine learning model supply chain. IEEE Access 2019.

Certified Backdoor Detection Problem

Role of defender

- A downstream user
- A third party inspector (e.g. government official)

Goals

- Detect if the model is backdoored
- Derive a **condition** under which backdoor attacks are **guaranteed** to be detectable
- Derive a constraint on false detection rate

Challenges

- No prior knowledge about the presence of backdoor
- No access to the training set or the trigger
- No benign models for reference

Method – Overview

Key idea

- Leverage two necessary properties of backdoor trigger (independent of attack configurations):
  - Be *robust* to random noise non-robust trigger will fail in practice
  - Be *stealthy* with small perturbation magnitude non-stealthy trigger will be exposed in practice

Main challenges

- How to quantify robustness of backdoor triggers? (*stealthiness can be quantified by perturbation magnitude*)
- How to incorporate robustness and stealthiness into detection procedure?
- How to derive a detection guarantee?

Method – Detection Statistic

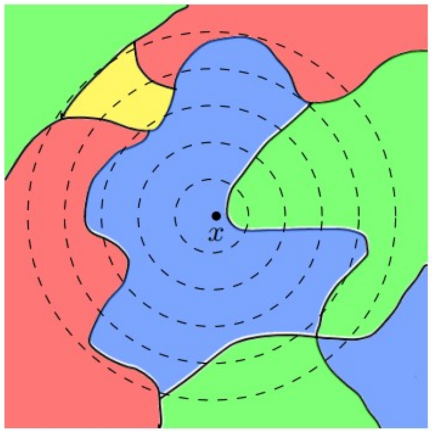
Quantify trigger robustness through randomized smoothing

- Definition 1: *Samplewise Local Probability Vector* (**SLPV**)
  - $f(\cdot; w)$: a classifier with parameters w and K classes
  - $\mathcal{N}(0, \sigma^2 I)$: isotropic Gaussian distribution with variance σ^2
  - SLPV for any input x is a K-dimensional probability vector $\boldsymbol{p}(x|w,\sigma) \in [0,1]^K$
  - The k-th entry is defined by:

$$p_k(x|w,\sigma) \triangleq \mathbb{P}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}(f(x+\epsilon; w) = k)$$

- Definition 2: *Samplewise Trigger Robustness* (**STR**)
  - Consider any backdoor attack with trigger δ and target class t
  - STR for any input x is the t-th entry of SLPV for $\delta(x)$:

$$R_{\delta,t}(x|w,\sigma) \triangleq p_t(\delta(x)|w,\sigma)$$



local probability distribution

J. M. Cohen, E. Rosenfeld, J. Z. Kolter. Certified adversarial robustness via randomized smoothing. ICML 2019.

Method – Detection Statistic

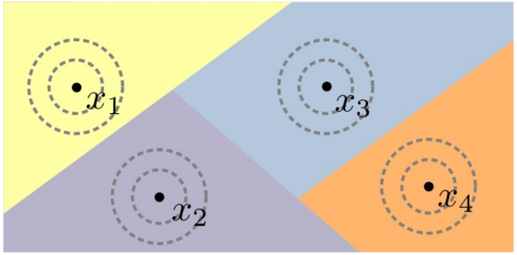
Detection statistic

- Definition 3: *Local Dominant Probability* (**LDP**)
  - Consider K random samples $x_1, \cdots, x_K$ satisfying $f(X_k; w) = k$
  - LDP for classifier $f(\cdot; w)$ is defined by:

$$s(w) = \|\boxed{\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{p}(x_k|w,\sigma)}\|_{\boxed{\infty}}$$

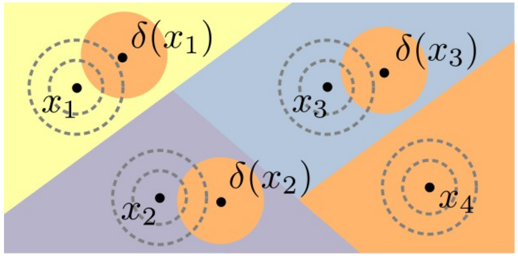
Average SLPV

largest entry



benign classifier with a small LDP close to 1/4

- Properties of LDP
  - Backdoored models tend to have larger LDP
  - *Larger* LDP for *more robust* and/or *stealthier* trigger



backdoored classifier with a large LDP

SLPV: samplewise local probability vector   **STR**: samplewise trigger robustness

Method – Detection Procedure

Detection procedure based on **conformal prediction**

- Step 1: Given a classifier $f(\cdot; w)$ to be inspected, estimate LDP $s(w)$
- Step 2: Train (benign) shadow models $f(\cdot; w_1), \cdots, f(\cdot; w_N)$ on the clean validation dataset, and construct a calibration set $\mathcal{S}_N = \{s(w_1), \cdots, s(w_N)\}$ by computing the LDP for each model.
- Step 3: Compute the adjusted conformal p-value (with m assumed outliers) defined by:

$$q_m(w) = 1 - \frac{1 + \min\{|\{s \in \mathcal{S}_N : s < s(w)\}|, N - m\}}{N - m + 1}$$

- Step 4: Trigger an alarm if $q_m(w) \leq \alpha$, where α is a prescribed significance level (e.g. α=0.05).

SLPV: samplewise local probability vector **STR**: samplewise trigger robustness **LDP**: local dominant probability

Method – Certification

Certification – *backdoor detection guarantee*

- Robustness metric (minimum STR):

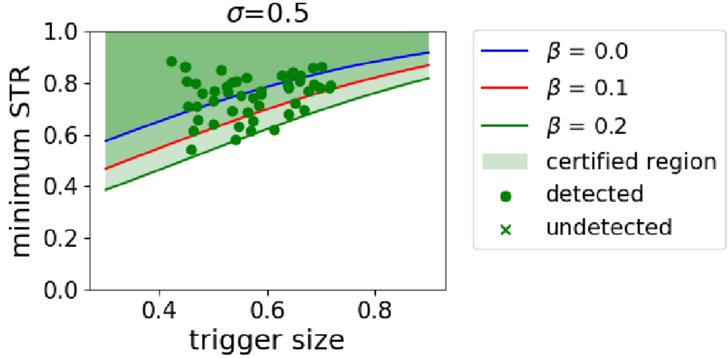
$$\pi = \min_{k=1,\cdots,K} R_{\delta,t}(x_k|w,\sigma)$$

- Stealthiness metric (maximum perturbation magnitude):

$$\Delta = \max_{k=1,\cdots,K} ||\delta(x_k) - x_k||_2$$

- Φ: standard Gaussian CDF
- s: calibration threshold
- Main result: a backdoor attack is guaranteed to be detectable if:

$$\Delta < \sigma(\Phi^{-1}(1-s) - \Phi^{-1}(1-\pi))$$



example of certified region on GTSRB dataset

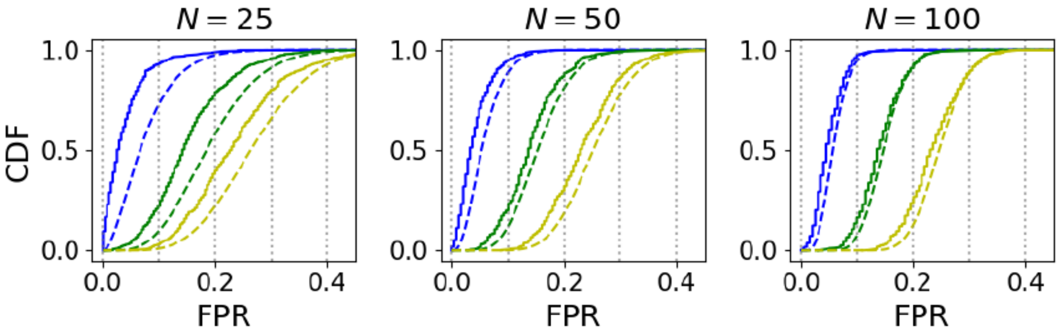
$\beta = m/N$: the proportion of calibration adjustment

SLPV: samplewise local probability vector **STR**: samplewise trigger robustness **LDP**: local dominant probability

Method – Certification

Certification – probabilistic upper bound on the false positive rate (FPR)

- Consider a random calibration set $\mathcal{S}_N$ with size N
- FPR: $Z_N = \mathbb{P}(q_m(W) \leq \alpha | \mathcal{S}_N)$
- Assumption: benign LDP distribution dominated (in first-order) by calibration distribution
- $B \sim \text{Beta}(m + l + 1, N - m - l)$ with $l = \lfloor \alpha(N - m + 1) \rfloor$
- Probabilistic upper bound: $\mathbb{P}(Z_N \leq q) \geq \mathbb{P}(B \leq q)$ for any real q
- Asymptotic property: for any $\xi > 0$ and $\tau = \alpha + (1 - \alpha)\beta + \xi$, $\lim_{N \to +\infty} \mathbb{P}(Z_N \leq \tau) = 1$



simulation

solid: CDF for empirical FPR

dashed: CDF for Beta upper bound

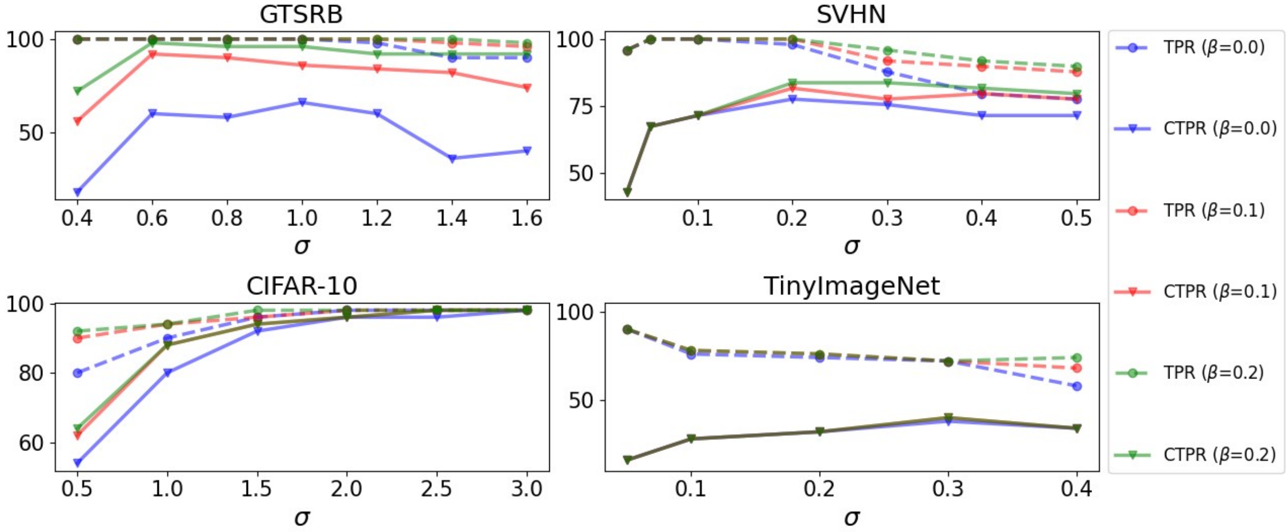
$\beta = m/N$: the proportion of calibration adjustment

LDP: local dominant probability

Evaluation

Evaluation – certified detection of random backdoor attacks

- Backdoor triggers are *random pattern* with magnitude $L_2 < 0.75$
- True positive rate (**TPR, dashed**): proportion of attacks being successfully detected
- **Certified** true positive rate (**CTPR, solid**): proportion of attacks in certified region



- **Correctness** of certification:

CTPRs <= TPRs

- **Non-triviality** of certification:

Maximum CTPRs:
98%, 84%, 98%, and 40%

Corresponding FPRs:
0%, 0%, 6%, and 10%

CBD: Certified Backdoor Detection

Evaluation – certified detection for more trigger types

	GTSRB				SVHN				CIFAR-10				AVG
	benign	BadNet	CB	Blend	benign	BadNet	CB	Blend	benign	BadNet	CB	Blend	TPR
NC	20	50	75	20	40	80	100	95	20	35	95	60	67.8
K-Arm	5	100	100	100	5	100	70	40	5	100	80	55	82.8
MNTD	5	20	0	0	5	10	10	15	5	90	100	75	35.6
CBDsup	5	100	95	100	5	100	100	90	5	65	100	55	**89.4**
CBD0	0	75 (5)	95 (80)	80 (20)	0	75 (45)	100(100)	80 (75)	0	50 (5)	100 (90)	45 (30)	77.2
CBD0.1	0	90 (15)	95 (85)	90 (25)	0	90 (55)	100(100)	80 (80)	20	75 (20)	100 (95)	55 (35)	86.1
CBD0.2	0	90 (15)	95 (85)	95 (35)	0	95 (65)	100(100)	90 (80)	25	75 (25)	100(100)	60 (40)	**88.9**

- **High detection accuracy:** CBD achieves generally higher TPR (*outside parenthesis*) than **uncertified** baselines
- **Non-trivial certification:** CBD achieves non-trivial CTPR (*in parenthesis*) in most cases
- **Limitations**: clear gap between TPR and CTPR for BadNet trigger with *large* perturbation magnitude