Black-box Detection of Backdoor Attacks with Limited Information and Data

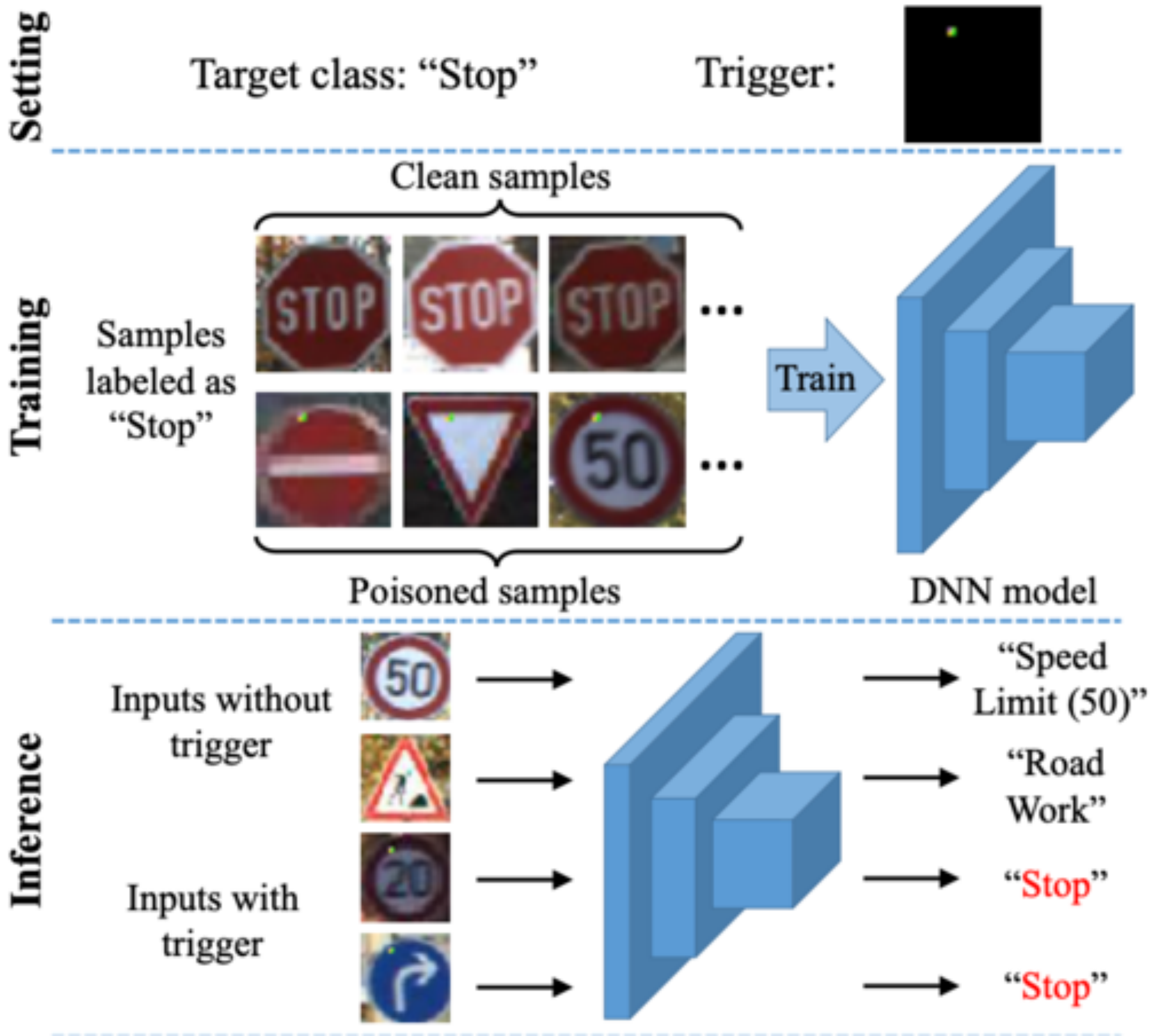
Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, Jun Zhu

Dept. of Comp. Sci. and Tech., Tsinghua University, RealAI; Contact: dyp17@mails.tsinghua.edu.cn; dongyinpeng@gmail.com

2021 ICCV OCTOBER 11-17 VIRTUAL

Introduction

- **Backdoor attacks:** The attacker embeds a backdoor in a DNN model by injecting poisoned samples into its training data, the infected model performs normally on clean inputs, but whenever the embedded backdoor is activated by a backdoor trigger, such as a small pattern in the input, the model will output an adversary-desired target class, as shown below.



Black-box Setting

- The backdoor defense cannot access the poisoned training data or the white-box model, while only **query access to the model is attainable**.
- The black-box setting is **more realistic** in the real-world machine learning applications.

Methodology

- **Backdoor attacks:**

$$x' = A(x, m, p) = (1-m) * x + m * p$$

  - $m \in \{0,1\}^d,\ p \in [0,1]^d$
- **Reverse-engineer the trigger** (Wang et al., 2019):

$$\min_{m,p} \sum_{x_i \in X} \left\{ \ell\left(c, f\left(A(x_i, m, p)\right)\right) + \lambda \cdot |m| \right\}$$

  - ℓ is the cross-entropy loss; $|m|$ is the L_1 norm of the mask; λ is a hyper-parameter
- This problem can be solved by the Adam optimizer (white-box access to model gradients).

Black-box Optimization:

- Let $\mathcal{F}(m,p;c) = \sum_{x_i \in X} \left\{ \ell\left(c, f\left(A(x_i, m, p)\right)\right) + \lambda \cdot |m| \right\}$;
- **Natural Evolution Strategies (NES)** (Wierstra et al., 2014)

$$\min_{\theta_m, \theta_p} \mathcal{J}(\theta_m, \theta_p) = \mathbb{E}_{\pi(m,p|\theta_m,\theta_p)}[\mathcal{F}(m,p;c)]$$

$$m \sim \mathrm{Bern}\big(g(\theta_m)\big); \quad p = g(p'), p' \sim N(\theta_p, \sigma^2)$$

  - $g(\cdot) = \frac{1}{2}(\tanh(\cdot) + 1)$; $\mathrm{Bern}(\cdot)$ **is the Bernoulli distribution;** $N(\cdot)$ **is the Gaussian distribution**
- Estimate the gradient

$$\nabla_{\theta_m} \mathcal{J}(\theta_m, \theta_p) \approx \frac{1}{k} \sum_{j=1}^{k} \mathcal{F}\big(m_j, g(\theta_p); c\big) \cdot 2\left(m_j - g(\theta_m)\right)$$

$$\nabla_{\theta_p} \mathcal{J}(\theta_m, \theta_p) \approx \frac{1}{k\sigma} \sum_{j=1}^{k} \mathcal{F}\big(g(\theta_m), \theta_p + \sigma\epsilon_j; c\big) \cdot \epsilon_j$$

Experiments

- **Overall results**

	CIFAR-10	GTSRB	ImageNet
NC [45]	95.0%	**100.0%**	**96.0%**
TABOR [20]	95.5%	**100.0%**	95.0%
B3D (Ours)	**97.5%**	**100.0%**	**96.0%**
B3D-SS (Ours)	**97.5%**	**100.0%**	95.5%

- **Detailed results on CIFAR-10**

Model	Accuracy	ASR	Method	Reversed Trigger		Detection Results			
				L_1 norm	ASR	Case I	Case II	Case III	Case IV
Normal	89.30%	N/A	NC [45]	N/A	N/A	N/A	N/A	8/50	42/50
			TABOR [20]	N/A	N/A	N/A	N/A	4/50	46/50
			B3D (Ours)	N/A	N/A	N/A	N/A	2/50	48/50
			B3D-SS (Ours)	N/A	N/A	N/A	N/A	3/50	47/50
Backdoored (1 × 1 trigger)	88.35%	99.75%	NC [45]	0.588	98.76%	40/50	9/50	0/50	1/50
			TABOR [20]	0.672	99.11%	36/50	13/50	0/50	1/50
			B3D (Ours)	0.820	99.29%	36/50	12/50	0/50	2/50
			B3D-SS (Ours)	3.734	99.98%	35/50	15/50	0/50	0/50
Backdoored (2 × 2 trigger)	88.51%	100.00%	NC [45]	1.508	98.81%	47/50	2/50	0/50	1/50
			TABOR [20]	2.256	99.21%	44/50	3/50	0/50	3/50
			B3D (Ours)	2.310	98.94%	47/50	3/50	0/50	0/50
			B3D-SS (Ours)	2.867	99.13%	47/50	2/50	0/50	1/50
Backdoored (3 × 3 trigger)	88.57%	100.00%	NC [45]	2.264	98.71%	49/50	1/50	0/50	0/50
			TABOR [20]	2.493	98.84%	48/50	1/50	0/50	1/50
			B3D (Ours)	3.521	98.87%	47/50	2/50	0/50	1/50
			B3D-SS (Ours)	3.856	96.97%	47/50	2/50	0/50	1/50

- **Detailed results on ImageNet**

Model	Accuracy	ASR	Method	Reversed Trigger		Detection Results			
				L_1 norm	ASR	Case I	Case II	Case III	Case IV
Normal	88.46%	N/A	NC [45]	N/A	N/A	N/A	N/A	2/50	48/50
			TABOR [20]	N/A	N/A	N/A	N/A	1/50	49/50
			B3D (Ours)	N/A	N/A	N/A	N/A	0/50	50/50
			B3D-SS (Ours)	N/A	N/A	N/A	N/A	1/50	49/50
Backdoored (Trigger)	87.91%	99.95%	NC [45]	62.093	99.11%	45/50	0/50	0/50	5/50
			TABOR [20]	57.569	99.25%	43/50	0/50	0/50	7/50
			B3D (Ours)	86.083	99.14%	43/50	0/50	0/50	7/50
			B3D-SS (Ours)	120.822	97.57%	42/50	0/50	0/50	8/50
Backdoored (Trigger)	87.52%	99.68%	NC [45]	20.610	99.12%	50/50	0/50	0/50	0/50
			TABOR [20]	22.035	99/24%	47/50	2/50	0/50	1/50
			B3D (Ours)	23.497	99.09%	50/50	0/50	0/50	0/50
			B3D-SS (Ours)	24.124	97.15%	44/50	6/50	0/50	0/50
Backdoored (Trigger)	87.39%	99.94%	NC [45]	38.701	99.14%	48/50	1/50	0/50	1/50
			TABOR [20]	37.499	99.20%	46/50	3/50	0/50	1/50
			B3D (Ours)	56.636	99.13%	48/50	1/50	0/50	1/50
			B3D-SS (Ours)	37.253	97.44%	49/50	1/50	0/50	0/50

Conclusion

We proposed B3D, the first method for detecting backdoor attacks under the black-box setting. The detection accuracy of B3D is similar to white-box backdoor detection methods.