Post-Training Detection of Backdoor Attacks for Two-Class and Multi-Attack Scenarios

Zhen Xiang, David J. Miller, George Kesidis

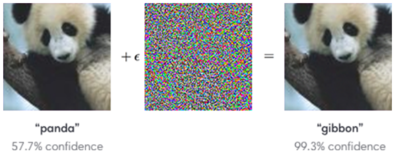
School of EECS, Pennsylvania State University

April 4, 2022

Adversarial Attacks

DNN image classifiers are threatened by adversarial attacks:

- Test-time evasion attack [SZS$^+$14, MDFF16]



- Backdoor (Trojan) attack [GLDG19, CLL$^+$17, LMA$^+$18]

- Other attacks: poisoning attack [BR18], model-stealing [LZ21], etc.

Backdoor Attacks

Elements of backdoor attack

- A set of source classes $\mathcal{S}^*$
- A target class t^*
- A backdoor pattern (i.e. a trigger)
  - Additive perturbation [ZLS+20]

$$M(x; v) = x + v$$

  - Patch replacement [GLDG19]

$$M(x; \{m, u\}) = x \odot (1 - m) + u \odot m$$

Backdoor Attacks

- Attacker's goals
  - Build a "backdoor mapping": source class samples with the backdoor pattern will be misclassified to the target class, i.e. maximize:

$$\mathrm{Prob}\big(f(M(X;v)) = t^*\big), \quad X \sim P_{\mathcal{S}^*}$$

  - Not degrade accuracy on clean samples, i.e. maximize:

$$\mathrm{P}\big(f(X) = Y\big), \quad \forall (X,Y) \sim P_{\mathrm{data}}$$

- Launching strategy: Poisoning the training set [GLDG19]

"stop sign" "speed limit" training set

Backdoor Defense Post-Training

Goals and assumptions of post-training backdoor defender

- Defender is the user who wants to make sure that the classifier is reliable before using it.
- Defender's goals
  - Detect if the classifier is backdoor attacked
  - Infer the target class when an attack is detected
- Defender's knowledge and capability
  - Defender does not know a priori if there is an attack.
  - Defender has no access to the classifier's training set.
  - Defender possesses an independent, small clean dataset for detection.
  - No clean classifiers for reference (e.g. set a detection threshold).

Backdoor Defense Post-Training

Reverse-engineering-based defense (RED) – an important family of post-training backdoor detection strategy [WYS+19, XMK20, GWX+19, WZL+20, CFZK19, LLT+19]

- Procedure
  - Backdoor pattern reverse-engineering
    E.g., for each class pair (s, t), find the min-sized perturbation inducing 80% of samples from class s to be misclassified to class t (also, can apply to internal-layer activations) [XMK20].
  - Detection inference
    E.g. check if for any class pair, the estimated perturbation size is abnormally small, using statistical anomaly detection [XMK20].
- Limitations
  Not applicable to two-class scenarios – no sufficient statistics for estimation of null distribution.

Backdoor Detection Using Expected Transferability (ET)

Key ideas

- Process each class independently: obtain an expected transferability (ET) statistic independently for each class, 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. ⇒ No deed for domain-specific supervision.

Backdoor Detection Using Expected Transferability (ET)

Definition of ET

- ϵ-solution set: For any x from any class, the ϵ-solution set is:

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

where v^* is the global optimal solution to

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

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 x and $\epsilon > 0$ is defined by

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

- ET statistic: For any class $i \in \mathcal{Y} = \{0, 1\}$ and $\epsilon > 0$, considering i.i.d. random samples $\mathrm{X}, \mathrm{Y} \sim P_i$, the ET statistic for class i is defined by

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

  - P_i: sample distribution of class i

Backdoor Detection Using Expected Transferability (ET)

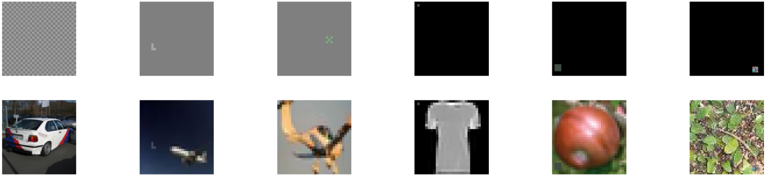
Detection method

- Properties of ET: There exists a constant detection threshold (details skipped)
  - 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 $\frac{1}{2}$
- Generalization
  - No specification on the method for pattern estimation.
  - Can be naturally extended to multi-class domains – "one versus all".
  - Extend to other backdoor patterns.

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 $\frac{1}{2}$)

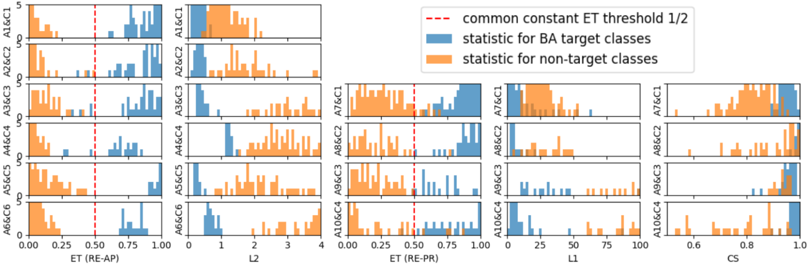
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
RE-AP	45/45	18/20	16/20	17/20	20/20	20/20	n/a	n/a	n/a	n/a
RE-PR	n/a	n/a	n/a	n/a	n/a	n/a	45/45	20/20	19/20	19/20

	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 backdoor patterns
  - $A_7 \sim A_{10}$: attack instances with patch replacement backdoor patterns
  - $C_1 \sim C_6$: clean instances
  - RE-AP: our method with backdoor pattern reverse-engineering algorithm in [XMK20]
  - RE-PR: our method with backdoor pattern reverse-engineering algorithm in [WYS+19]

ET – experiments (cont'd)

- Comparison between ET and other detection statistics



  - L_1: l_1 norm of estimated mask [WYS+19]
  - L_2: l_2 norm of estimated perturbation [XMK20]
  - CS: cosine similarity [WZL+20]

References I

B. Biggio and F. Roli.
Wild patterns: Ten years after the rise of adversarial machine learning.
Pattern Recognition, 84:317–331, 2018.

Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar.
Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks.
In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4658–4664, 7 2019.

X. Chen, C. Liu, B. Li, K. Lu, and D. Song.
Targeted backdoor attacks on deep learning systems using data poisoning.
https://arxiv.org/abs/1712.05526v1, 2017.

References II

T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg.
Badnets: Evaluating backdooring attacks on deep neural networks.
IEEE Access, 7:47230–47244, 2019.

W. Guo, L. Wang, X. Xing, M. Du, and D. Song.
TABOR: A highly accurate approach to inspecting and restoring Trojan backdoors in AI systems.
https://arxiv.org/abs/1908.01763, 2019.

Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang.
Abs: Scanning neural networks for back-doors by artificial brain stimulation.
In *CCS*, 2019.

Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, and J Zhai.
Trojaning attack on neural networks.
In *Proc. NDSS*, San Diego, CA, Feb. 2018.

References III

X Jia Y Jiang ST Xia X Cao L Zhu, Y Li.
Defending against model stealing via verifying embedded external features.
In *ICML Workshop on Adversarial Machine Learning*, 2021.

S.-M. M.-Dezfooli, A. Fawzi, and P. Frossard.
DeepFool: a simple and accurate method to fool deep neural networks.
In *Proc. CVPR*, 2016.

C. Szegedy, W. Zaremba, I Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus.
Intriguing properties of neural networks.
In *Proc. ICLR*, 2014.

References IV

B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B.Y. Zhao.
Neural cleanse: Identifying and mitigating backdoor attacks in neural networks.
In *Proc. IEEE Symposium on Security and Privacy*, 2019.

R. Wang, G. Zhang, S. Liu, P.-Y. Chen, J. Xiong, and M. Wang.
Practical detection of trojan neural networks: Data-limited and data-free cases.
In *Proc. ECCV*, 2020.

Z. Xiang, D. J. Miller, and G. Kesidis.
Detection of backdoors in trained classifiers without access to the training set.
IEEE Transactions on Neural Networks and Learning Systems, pages 1–15, 2020.

References V

H. Zhong, C. Liao, A. Squicciarini, S. Zhu, and D.J. Miller.
Backdoor embedding in convolutional neural network models via invisible perturbation.
In *Proc. CODASPY*, March 2020.

Thanks

- Thanks for your attention!