AI-Guardian: Defeating Adversarial Attacks Using Backdoors

Hong Zhu[1,2], **Shengzhi Zhang[3]**, and Kai Chen[1,2]

[1]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China

[2]School of Cyber Security, University of Chinese Academy of Sciences, China

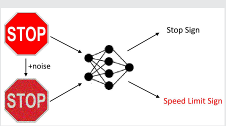
[3]Department of Computer Science, Boston University, Metropolitan College, USA

zhuhong@iie.ac.cn, **shengzhi@bu.edu**, chenkai@iie.ac.cn

Abstract

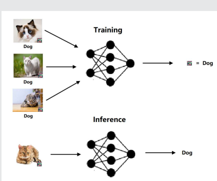
- Deep neural networks (DNNs) are known to be vulnerable to adversarial attacks, posing a severe threat to security-critical applications such as autonomous driving, remote diagnosis, etc.
- Existing solutions are limited in detecting/preventing such attacks and impacting the original tasks' performance.
- We present AIGuardian, a novel approach to defeating adversarial attacks that leverages intentionally embedded backdoors to fail the adversarial perturbations and maintain the performance of the original main task.
- AI-Guardian reduces the attack success rate from 97.3% to 3.2%, which outperforms the state-of-the-art works by 30.9%, with only a 0.9% decline in the clean data accuracy.

What is an Adversarial Attack?



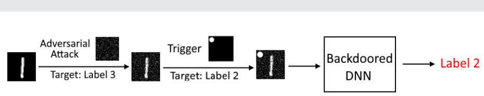
- The adversarial attack subtly modifies the inputs, usually imperceptible to human beings, to make the victim model produce incorrect classification or prediction results.
- It naturally exists in almost all models.

What is a Model Backdoor?



- Backdoor embeds a hidden behavior into the model, which keeps "hibernated" until a specific trigger is applied to the input, causing the model to produce the predefined classification or prediction results.
- It is intentionally embedded.
- Universal backdoor vs Specific Backdoor

When Backdoor Meets Adversarial



- Given a backdoor model, what happens if a backdoor trigger is attached to an input of an adversarial attack?
- We find that in most cases, the model produces results depending on the backdoor, i.e., backdoor `suppressing' the effects of adversarial attacks.

The Approach

Intuitively

- Intend to embed a controlled backdoor into the to-be-protected model
- Attach the trigger to all inputs after deploying the protected model

Two Problems to Address

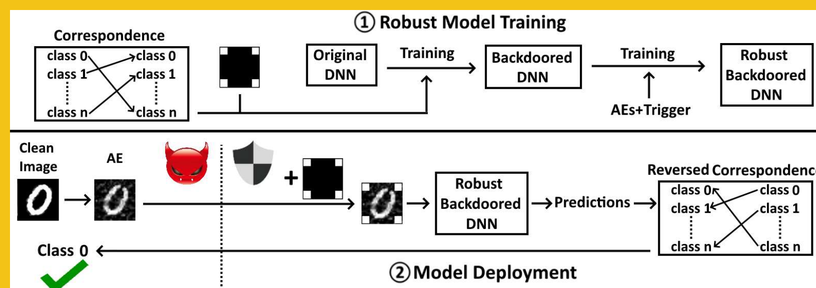
(P1) Ensure the embedded backdoor always `suppresses' the effect of adversarial attacks.

(P2) Ensure the protected model produces the correct outputs to clean inputs even when our backdoor is attached.

Solutions

(S1) We propose a backdoor enhancement scheme to improve the suppression of the backdoor over the adversarial attack.

(S2) We design a unique backdoor, named bijection backdoor, to maintain a one-to-one mapping between the source label and the target label of the backdoor.



Results

Defense	Dataset	Clean Accuracy	Attack Success Rate										
			BIM	PGD			AutoPGD			CW		AP	Avg
				L_∞	L_1	L_2	L_∞	L_1	L_2	L_∞	L_2		
None	MNIST	99.3%	100%	100%	100%	99%	99%	100%	100%	93%	97%	97%	98.5%
	GTSRB	95.4%	98%	98%	100%	100%	98%	98%	99%	94%	82%	97%	96.4%
	Youtube	99.0%	100%	100%	100%	100%	100%	98%	99%	94%	82%	99%	97.2%
	VGG	90.3%	98%	100%	100%	100%	99%	100%	100%	82%	95%	98%	97.2%
	Average	96.0%	99.0%	99.5%	100%	99.8%	99.0%	99.0%	99.5%	90.8%	89.0%	97.8%	97.3%
AI-Guardian (Ours)	MNIST	98.6%	2%	1%	8%	5%	2%	7%	5%	4%	1%	7%	**4.2%**
	GTSRB	95.1%	5%	7%	6%	3%	2%	3%	5%	2%	2%	0%	**3.5%**
	Youtube	98.2%	2%	2%	1%	2%	1%	1%	1%	0%	0%	1%	**1.1%**
	VGG	88.3%	7%	6%	7%	3%	1%	5%	5%	1%	4%	0%	**3.9%**
	Average	95.1%	4.0%	4.0%	5.5%	3.3%	1.5%	4.0%	4.0%	1.8%	1.8%	2.0%	**3.2%**

We reduce the success rate of various adversarial attacks from 97.3% to 3.2% on average.

Dataset	Attack Success Rate					
	BIM	PGD	AutoPGD	CW	HotFlip	Avg
USCFC	5.3%	6.8%	7.2%	8.6%	6.7%	6.9%
SFCC	1.1%	1.3%	1.5%	5.2%	7.2%	3.3%
THUCNews	4.5%	6.3%	6.5%	5.1%	8.2%	6.1%

We also extended AI-Guardian to NLP and speech recognition domains.

Formula

- Definition of the bijection backdoor:

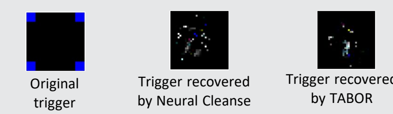
$$\begin{aligned} &P(F(x_i) = y_i | (x_i, y_i) \in D_{test}) \geq acc \\ &P(F(x_i^t) = y_i^t | (x_i, y_i) \in D_{test}) \geq bp \\ &x_i^t = x_i * m + \Delta * (J - m) \\ &y_i^t = g(y_i) \end{aligned} \tag{1}$$

- The loss function to embed the bijection backdoor

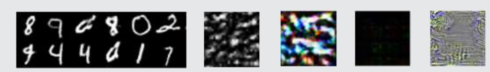
$$\begin{aligned} &\min_\theta \mathbb{E}_{x_i, y_i \in D}(L(y_i, F_\theta(x_i)) + \gamma \cdot L(y_i^t, F_\theta(x_i^t))) \\ &x_i^t = \Delta * m + x_i * (J - m) \\ &y_i^t = g(y_i) \end{aligned} \tag{2}$$

Discussion

- Existing backdoor detection works cannot recover our trigger.



- Existing model inversion attacks cannot reverse our trigger either.



- Limitations of AI-Guardian
  1. Backdoor triggers should be kept securely
  2. Lack of theoretical guarantee

Conclusion

- AI-Guardian defeats adversarial attacks by embedding a controlled backdoor into the to-be-protected model.
- We proposed the backdoor enhancement and bijection backdoor to facilitate the design.
- AI-Guardian can reduce the attack success rate of Aes from 97.3% to 3.2%, with only a 0.9% decline in clean data accuracy. In addition, AI-Guardian incurs almost negligible overhead to the model runtime performance, with only a 0.36% increase in the model prediction time.

Reference

Hong Zhu, Shengzhi Zhang, Kai Chen, "AI-Guardian: Defeating Adversarial Attacks using Backdoors", IEEE S&P 2023