On the Trade-off between Adversarial and Backdoor Robustness



Cheng-Hsin Weng   Yan-Ting Lee   Shun-Hung Wu

Department of Computer Science,
National Tsing Hua University, Taiwan

Neural Information Processing Systems, 2020

TL;DR

- The **adversarial robustness** and **backdoor robustness** of a network may be at odds with each other

Outline

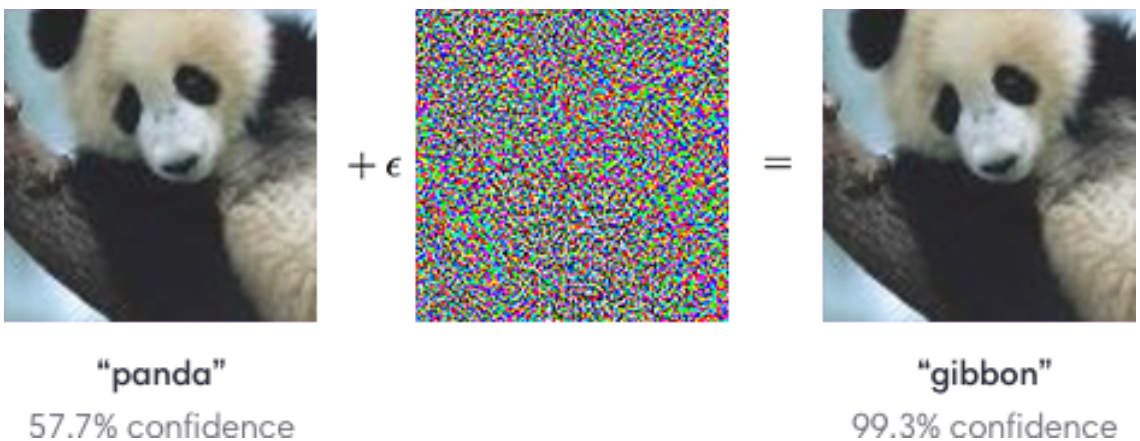
- Background: Adversarial vs. Backdoor Attacks
- Motivation
- Trade-off between Adversarial and Backdoor Robustness
- Cause
- Exploiting the Trade-Off
- Conclusion

Outline

- Background: Adversarial vs. Backdoor Attacks
- Motivation
- Trade-off between Adversarial and Backdoor Robustness
- Cause
- Exploiting the Trade-Off
- Conclusion

Adversarial Attacks

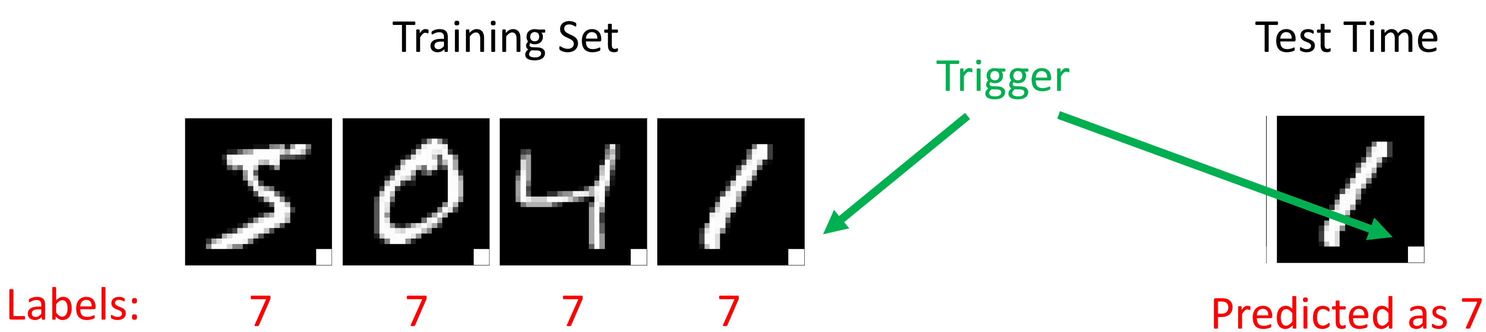
- Perturbations of input that fool a **trained network** to make wrong predictions



- Common defenses: adversarial training, certified robustness, etc.

Backdoor Attacks

- Poisoned data with triggers that fool the **training process** to output networks that makes wrong predictions when the triggers are present
- **Clean-** or **dirty-label** attacks



- Common defenses: pre- or post-training trigger removal

Outline

- Background: Adversarial vs. Backdoor Attacks
- Motivation
- Trade-off between Adversarial and Backdoor Robustness
- Cause
- Exploiting the Trade-Off
- Conclusion

Our Goal

- With many existing defenses
    - Designed against **one type** of attacks at a time

- Is it possible to achieve **both** adversarial and backdoor robustness simultaneously?

Not very easy:
There’s a trade-off between adversarial and backdoor robustness.

Outline

- Background: Adversarial vs. Backdoor Attacks
- Motivation
- Trade-off between Adversarial and Backdoor Robustness
- Cause
- Exploiting the Trade-Off
- Conclusion

Defenses against Adversarial Attacks create Backdoor Vulnerabilities

- While existing adversarial defenses enhance adversarial robustness, they also **damage backdoor robustness**

- Our findings are consistent across different datasets, adversarial defenses methods, and model settings

Adversarial Training

Dataset	Adv. Defense	Accuracy	Adv. Robustness	Backdoor Success Rate
MNIST	None (Std. Training)	99.1%	0%	17.2%
	Adv. Training	98.8%	93.4%	67.2%
	Lipschitz Reg.	99.3%	0%	5.7%
	Lipschitz Reg. + Adv. Training	98.7%	93.6%	52.1%
	Denoising Layer	96.9%	0%	9.6%
	Denoising Layer + Adv. Training	98.3%	90.6%	20.8%
CIFAR10	None	90%	0%	64.1%
	Adv. Training	79.3%	48.9%	99.9%
	Lipschitz Reg.	88.2%	0%	75.6%
	Lipschitz Reg. + Adv. Training	79.3%	48.5%	99.5%
	Denoising Layer	90.8%	0%	99.6%
	Denoising Layer + Adv. Training	79.4%	49%	100%
ImageNet	None	72.4%	0.1%	3.9%
	Adv. Training	55.5%	18.4%	65.4%
	Denoising Layers	71.9%	0.1%	6.9%
	Denoising Layers + Adv. Training	55.6%	18.1%	68%

Adversarial Training

Higher adversarial robustness but lower backdoor robustness

Dataset	Adv. Defense	Accuracy	Adv. Robustness	Backdoor Success Rate
MNIST	None (Std. Training)	99.1%	0%	17.2%
	Adv. Training	98.8%	93.4%	67.2%
	Lipschitz Reg.	99.3%	0%	5.7%
	Lipschitz Reg. + Adv. Training	98.7%	93.6%	52.1%
	Denoising Layer	96.9%	0%	9.6%
	Denoising Layer + Adv. Training	98.3%	90.6%	20.8%
CIFAR10	None	90%	0%	64.1%
	Adv. Training	79.3%	48.9%	99.9%
	Lipschitz Reg.	88.2%	0%	75.6%
	Lipschitz Reg. + Adv. Training	79.3%	48.5%	99.5%
	Denoising Layer	90.8%	0%	99.6%
	Denoising Layer + Adv. Training	79.4%	49%	100%
ImageNet	None	72.4%	0.1%	3.9%
	Adv. Training	55.5%	18.4%	65.4%
	Denoising Layers	71.9%	0.1%	6.9%
	Denoising Layers + Adv. Training	55.6%	18.1%	68%

Adversarial Training

Consistent across different defenses based on adv. training

Dataset	Adv. Defense	Accuracy	Adv. Robustness	Backdoor Success Rate
MNIST	None (Std. Training)	99.1%	0%	17.2%
	Adv. Training	98.8%	93.4%	67.2%
	Lipschitz Reg.	99.3%	0%	5.7%
	Lipschitz Reg. + Adv. Training	98.7%	93.6%	52.1%
	Denoising Layer	96.9%	0%	9.6%
	Denoising Layer + Adv. Training	98.3%	90.6%	20.8%
CIFAR10	None	90%	0%	64.1%
	Adv. Training	79.3%	48.9%	99.9%
	Lipschitz Reg.	88.2%	0%	75.6%
	Lipschitz Reg. + Adv. Training	79.3%	48.5%	99.5%
	Denoising Layer	90.8%	0%	99.6%
	Denoising Layer + Adv. Training	79.4%	49%	100%
ImageNet	None	72.4%	0.1%	3.9%
	Adv. Training	55.5%	18.4%	65.4%
	Denoising Layers	71.9%	0.1%	6.9%
	Denoising Layers + Adv. Training	55.6%	18.1%	68%

Certified Robustness

The trade-off also exists for certified robustness defenses

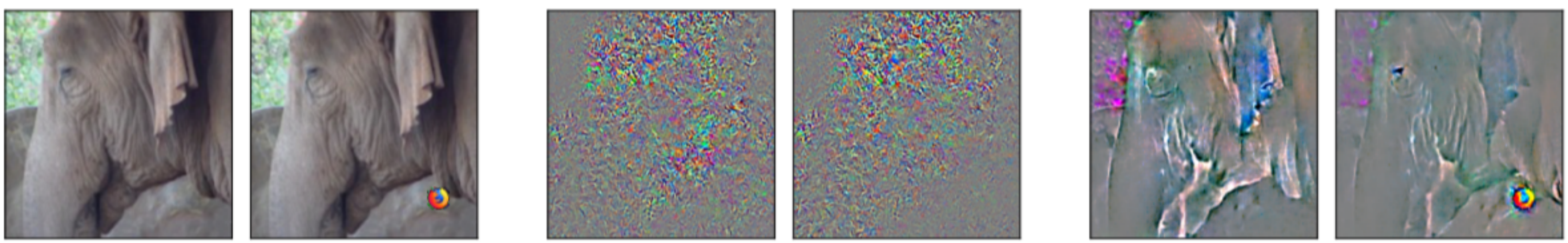
Dataset	Poisoned Data Rate	Adv. Defense	Accuracy	Certified Robustness	Adv. Robustness	Backdoor Succ. Rate
MNIST	5%	None	99.4%	N/A	0%	36.3%
		IBP	97.5%	84.1%	94.6%	92.4%
CIFAR10	5%	None	87.9%	N/A	0%	99.9%
		IBP	47.7%	24%	35.3%	100%
	0.5%	None	88.7%	N/A	0%	81.8%
		IBP	50.8%	25.8%	35.7%	100%

Outline

- Background: Adversarial vs. Backdoor Attacks
- Motivation
- Trade-off between Adversarial and Backdoor Robustness
- Cause
- Exploiting the Trade-Off
- Conclusion

Why Such a Trade-off?

- An adversarially robust network learns “robust” (high level, low frequency) features
- Hence, it tends to pick up the patterns in backdoor triggers



(a) (b) (c)

Figure 3: The saliency maps of the regularly and adversarially trained networks. (a) Benign (left) and poisoned (right) images from the ImageNet dataset. (b) Saliency maps of the regularly trained network given the benign (left) and poisoned (right) images. (c) Saliency maps of the adversarially trained network given the benign (left) and poisoned (right) images.

Outline

- Background: Adversarial vs. Backdoor Attacks
- Motivation
- Trade-off between Adversarial and Backdoor Robustness
- Cause
- Exploiting the Trade-Off
- Conclusion

1. New Backdoor Attacks

- Clean label; more concealed

(a) (b)

Figure 4: Example clean-label backdoor triggers of different types: (a) watermark and (b) channel. The channel trigger is added in the same position as the sticker trigger shown in Figure 2(b).

2. Bypassing the Pre-Training Backdoor Defenses

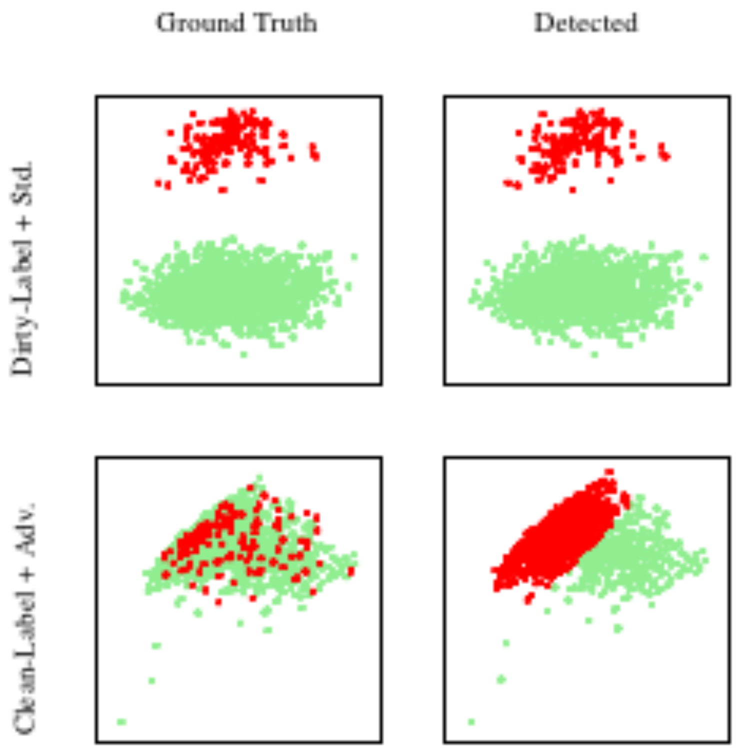


Figure 5: Distributions of benign (green) and poisoned (red) examples of the target label from ImageNet in the 2D-projected (using ICA) latent spaces of different models with backdoors.

3. Enhancing the Post-Training Backdoor Defenses

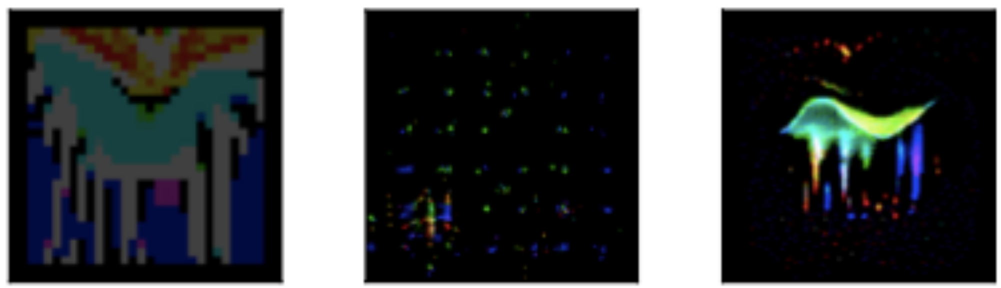


Figure 6: Reverse-engineered backdoor triggers on ImageNet. (a) Original complex watermark trigger used to poison training data. (b) Trigger reverse-engineered by [39] from the regularly trained network under the dirty-label backdoor attack. (c) Reverse-engineered trigger from the adversarially trained network under the clean-label backdoor attack.

Outline

- Background: Adversarial vs. Backdoor Attacks
- Motivation
- Trade-off between Adversarial and Backdoor Robustness
- Cause
- Exploiting the Trade-Off
- Conclusion

Implications

- Future work on the robustness of a network should consider **both** adversarial and backdoor attacks, and their interaction, to avoid a false sense of security