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

- Additional image and text for adversarial attack demonstration:
  - Image of a panda with an adversarial perturbation:
    - Initial image labeled “panda” with 57.7% confidence.
    - Image with perturbation labeled “gibbon” with 99.3% confidence.

• 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

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Trigger</th>
<th>Test Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Labels: 7 7 7 7 7

Predicted as 7
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

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

Higher adversarial robustness but lower backdoor robustness

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

Consistent across different defenses based on adv. training

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

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

The trade-off also exists for certified robustness defenses.
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

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

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