Exploring the Orthogonality and Linearity of Backdoor Attacks

Kaiyuan Zhang1, Siyuan Cheng1, Guangyu Shen1, Guanhong Tao1, Shengwei An1, Anuran Makur1, Shiqing Ma2, Xiangyu Zhang1
1Purdue University, 2University of Massachusetts Amherst, *Equal contribution

Introduction

- Backdoor attacks embed an attacker-chosen pattern into inputs to cause model misclassification.
- A number of defense techniques proposed by the community. Do they work for a large spectrum of attacks?
- We study the characteristics of backdoor attacks through theoretical analysis and introduce two key properties: orthogonality and linearity.

Table 1: A Summary of Existing Attacks and Defenses

<table>
<thead>
<tr>
<th>Attack</th>
<th>Model Defended</th>
<th>Backdoor Oblivious</th>
<th>Input Detection</th>
<th>Output Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backdoor [12]</td>
<td>✓</td>
<td>□</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Trojan [13]</td>
<td>✓</td>
<td>□</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BadNets [14]</td>
<td>✓</td>
<td>□</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Blend [15]</td>
<td>✓</td>
<td>□</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Filter [16]</td>
<td>✓</td>
<td>□</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Invisible [17]</td>
<td>✓</td>
<td>□</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WaNet [18]</td>
<td>✓</td>
<td>□</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Comprare [19]</td>
<td>✓</td>
<td>□</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

What are the underlying reasons causing defenses to fail on certain backdoor attacks?

Motivation

- Key Observation: We observe that the backdoor task is quickly learned by the victim model (using a very few training epochs), much faster than the main task (clean).

Problem Formulation

- Based on our observation, we formulate backdoor learning as a two-task continual learning problem.

Theoretical Results

Orthogonality

\[ f(x_b, \theta^*_b) = f(x, \theta^*_c) \]

Take Away

- We systematically explore why existing defenses fail on certain backdoor attacks.
- We provide a theoretical analysis on two critical properties orthogonality and linearity.
- Unlock new insights — TRYOUT our NEW measurements beyond ASR and ACC!

Linearity

Proposition 3.9. (Linearity Perspective of Backdoor Learning) For a well-poisoned model \( f : X \rightarrow Y \) with a near 100% attack success rate, there exists a specific hyperplane \( \{ Wx - b = 0 \} \), which is capable of capturing the Trojan behavior in the backdoor learning phase, and this trojan hyperplane persists in the clean learning phase.

Numerical Results

- Evaluate our theoretical analysis and hypotheses on 14 attacks and 12 defenses.
- Investigate the impact of 6 key factors that affect the orthogonality and linearity.
- Offer insights on When and Why do defenses fail or succeed against various attacks.