Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection

Di Tang, XiaoFeng Wang, Haixu Tang, Kehuan Zhang
Backdoor Attack

Neural Network

Inputs → Hidden → Output(s)

Normal Cases

Trump

Biden

Mis-recognised as

Biden

Trigger Cases
Data Contamination

Trigger + Source = Infected

Dataset

Trainig

Mis-recognised as

Backdoor infected

Biden
Close Look on the Representations

Fig. 8 of <<Advanced Robotic Grasping System Using Deep Learning>>

Representations (Embeddings)
Close Look on the Representations

Trigger dominant representations

Figure 1: Effect of data contamination attack on the target label’s representations, which have been projected to their first two principle components. Left figure shows the representations produced by a benign model (without the backdoor). Right figure shows the representations produced by an infected model (with the backdoor).

Targeted Contamination AT tack
Launch TaCT

Dataset

Training

Backdoor infected

Mis-recognised as

Intention of the cover set:

1. Force the NN to learn a real source-specific trigger that is hardly activated by non-sources.

2. Make (source subject+trigger pattern) as the actual trigger, which reduce the difference between the representations of trigger-carrying inputs from normal inputs.
Test on classes
By finding short-cut between classes

Defeated by the large actual trigger, source subject + trigger pattern
Current Defences vs TaCT
—— Activation Clustering

Test on classes
By finding well-fitted 2-means clustering

Fig. 8: Silhouette scores of AC defence on GTSRB dataset. 0 is the target label (infected class), 1 is the source label and all the images in other classes are normal images. Box plot shows quartiles.

Defeated by mingled representations
Current Defences vs TaCT

—— Strip

Test on images
By finding lower-entropy superimposing

(a) GTSRB
(b) CIFAR-10

Figure 4: Entropy distributions of STRIP against TaCT.

Defeated by low-dominant trigger
Current Defences vs TaCT
—— SentiNet

Test on images
By finding dominant classification-matter pattern

Figure 5: Demonstration of SentiNet against TaCT on GTSRB.

Defeated by low-dominant trigger
Lesson: The trigger is not necessary to be such dominant.

Detecting the trigger may not be a good choice.

Failure of those defences vs TaCT.

Neural Cleanse, Strip, SentiNet
Idea

Lesson: The trigger is not necessary to be such dominant.

Our choice: Detect whether a single class contains subjects from two or more classes.

Reason: Misclassification is the goal of the backdoor injection, and is equivalent to that there is a class wrongly contains subjects from two or more classes during the prediction period.

Two-in-one $\approx$ Backdoor
Statistical Contamination Analyser—SCAn

Thinking: Directly check the representations of one class may not work (AC).
We should include the information from other classes.

Gaussian modeling: \( r = R(x) = \mu_t + \mathcal{E} \)

Assumption: Variance of every class follows the same distribution
SCAn-Pipeline

Global model

\[ r = R(x) = \mu_t + \varepsilon \]

Mixture model

\[ r_i = \delta_i \mu_1 + (1 - \delta_i) \mu_2 + \varepsilon, \]

Global covariance guided mixture model

Fig. 9: A schematic illustration of the assumption of two-component decomposition (right) in the representation space, in comparison with the naive homogeneous assumption (left).

(null hypothesis) \( H_0 : \mathcal{R}_t \) is drawn from a single normal distribution.

(alternative hypothesis) \( H_1 : \mathcal{R}_t \) is drawn from a mixture of two normal distributions.
SCAn-Criterion

For a class $t$

Hypothesis statistic:

$$ J_t = 2 \log \left( \frac{P(\mathcal{R}|H_1)}{P(\mathcal{R}|H_0)} \right) $$

$$ = \sum_{r \in \mathcal{R}} [(r - \mu_t)^T S^{-1} (r - \mu_t) - (r - \mu_j)^T S^{-1} (r - \mu_j)] $$

Outlier statistic:

$$ J^*_t = \frac{\bar{J}_t - \tilde{J}}{(\text{MAD}(\tilde{J}) \times 1.4826)} $$

where

$$ \tilde{J} = \text{median}(\{\bar{J}_t : t \in \mathcal{L}\}) $$

$$ \text{MAD}(\tilde{J}) = \text{median}(\{|\bar{J}_t - \tilde{J}| : t \in \mathcal{L}\}) $$

$$ \bar{J}_t = \frac{(J_t - k)}{\sqrt{2k}} $$

Final criterion:

$$ J^*_t > 7.3891 = \exp(2) $$

Ignore the subscript $t$, we check whether $\ln(J^*) > 2$
Effectiveness of SCAn vs TaCT

Figure 9: Four kinds of triggers used in our experiments

(a) Box  (b) Normal  (c) Square  (d) Watermark

Table 5: Accuracy of infected models.

<table>
<thead>
<tr>
<th></th>
<th>Top-1 Acc</th>
<th>Targeted Misclassification Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GTSRB</td>
<td>ILSVRC2012</td>
</tr>
<tr>
<td>Box</td>
<td>96.6%</td>
<td>76.3%</td>
</tr>
<tr>
<td>Normal</td>
<td>96.1%</td>
<td>76.1%</td>
</tr>
<tr>
<td>Square</td>
<td>96.3%</td>
<td>76.0%</td>
</tr>
<tr>
<td>Watermark</td>
<td>96.5%</td>
<td>75.5%</td>
</tr>
<tr>
<td>Uninfected</td>
<td>96.4%</td>
<td>76.0%</td>
</tr>
</tbody>
</table>

Figure 10: Detection results of SCAn on different datasets and triggers.

(a) Box  (b) Normal  (c) Square  (d) Watermark
Effectiveness of SCAn vs TaCT

Varying the size of clean dataset:

0.3% clean data is sufficient

K out of N test:

Work until contaminated >17%

Figure 13: $J^*$ of the target class on different amount of clean data known for decomposition model (average over 5 rounds).

Figure 11: $J^*$ of the target classes under contaminated clean data.
Comparison between SCAn and Previous

Offline setting (test on classes): Neural Cleanse, Activation Clustering

<table>
<thead>
<tr>
<th></th>
<th>GTSRB</th>
<th></th>
<th></th>
<th>CIFAR-10</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Offline</td>
<td>Online</td>
<td></td>
<td>Offline</td>
<td>Online</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCAn</td>
<td>SentiNet</td>
<td>STRIP</td>
<td>SCAn</td>
<td>SentiNet</td>
<td>STRIP</td>
</tr>
<tr>
<td>95%</td>
<td>0%, 0.15%</td>
<td>9.4%, 95.3%</td>
<td>0%, 77.5%</td>
<td>0.20%, 0.32%</td>
<td>0.08%, 82.6%</td>
<td>1.82%, 75.4%</td>
</tr>
<tr>
<td>99%</td>
<td>0%, 0.15%</td>
<td>14.1%, 100%</td>
<td>0%, 90.6%</td>
<td>0.55%, 1.10%</td>
<td>0.09%, 83.6%</td>
<td>4.66%, 95.7%</td>
</tr>
<tr>
<td>99.5%</td>
<td>0%, 0.19%</td>
<td>14.1%, 100%</td>
<td>0%, 90.6%</td>
<td>0.74%, 1.82%</td>
<td>0.09%, 84.1%</td>
<td>6.60%, 96.9%</td>
</tr>
</tbody>
</table>

Column A: source-agnostic backdoor  
Column T: TaCT
Comparison between SCAn and Previous

Offline setting (test on classes): Neural Cleanse, Activation Clustering

Online setting (test on images): SentiNet, Strip

<table>
<thead>
<tr>
<th>Offline</th>
<th>GTSRB</th>
<th>Online</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAn</td>
<td>NC</td>
<td>AC</td>
</tr>
<tr>
<td>TPR</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>95%</td>
<td>0%</td>
<td>0.15%</td>
</tr>
<tr>
<td>99%</td>
<td>0%</td>
<td>0.15%</td>
</tr>
<tr>
<td>99.5%</td>
<td>0%</td>
<td>0.19%</td>
</tr>
<tr>
<td>Table of FPR results.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Offline | CIFAR-10 | Online | | |
|---------|----------|--------|----------|----------|----------|----------|----------|
| SCAn    | NC       | AC     | SCAn     | SentiNet | STRIP    | ABS      |
| TPR     | A        | T      | A        | T        | A        | T        | S        |
| 95%     | 0%       | 0%     | 5.36%    | 92.5%    | 0%       | 21.1%    |          |
| 99%     | 0%       | 0%     | 8.44%    | 99.2%    | 0%       | 47.8%    |          |
| 99.5%   | 0%       | 0%     | 8.45%    | 99.2%    | 0%       | 47.8%    |          |
| Table of FPR results. |

Column A: source-agnostic backdoor  Column T: TaCT
Robustness of SCAn against Attacks

Multiple target-trigger attack:  
- 8 triggers

Blending-trigger attack:  

Poison frogs attack:  

ASR loss when the number of triggers increase.

Figure 14: Minimum $J^*$ of target classes under multiple target-trigger attack and 1% clean data are known (over 5 rounds).

Figure 15: The amount of clean data required by decomposition model for defeating multiple target-trigger attacks on GTSRB.

18% for 21 triggers
Adaptive Attacks against SCAn

Parameter inference attack:

\[ r = R(x) = \mu_t + \varepsilon \]

Black-box trigger adjustment attack:


Figure 18: CDF of norms of \( S_\varepsilon \) and the distance between a couple \( S_\varepsilon \).

Figure 19: Statistics of black-box attacks (after moving-mean filtering).
Limitations

• Needs clean data set
• Needs presence of the trigger-carrying images
• Only evaluated on image classification tasks
Summary

• New understanding about the backdoor attack.
  — Dominant trigger is not necessary for the backdoor contamination attack. A simple but powerful attack, TaCT, can bypass existing defences.

• New defence, SCAn.
  — Introduce the global variant to detect inconsistency in representations.

Thanks!

Di Tang  td016@ie.cuhk.edu.hk
Haixu Tang  hatang@indiana.edu
XiaoFeng Wang  xw7@indiana.edu
Kehuan Zhang  khzhang@ie.cuhk.edu.hk