Towards Certifiable Adversarial Sample Detection

Ilia Shumailov, Yiren Zhao, Robert Mullins, Ross Anderson
Machine Learning

- Machine learning is everywhere
- We operate on data, not formal rules
- There’s a lot of non-determinism
- It is suddenly hard to measure or even define critical emergent properties:

   Safety, Security and Robustness
Computer Security in context of Machine Learning

- Adversarial examples exist for all models
- There’s a large taxonomy of attackers
- They operate in white-box / grey-box / black-box settings
- Attacks are scalable because of transferability
Machine Learning in the context of System Security

- ML is a part of a larger pipeline
- As secure as the weakest link
- Need: clear threat model
- Safety / security policies / cases
- Well-defined environments
- Clear policy for handling abuse
- Build from trusted components
ML integrity attacks and robustness

robustness

Defence via robust optimisation
- Adversarial training
- Certifiable robustness
- Randomised smoothing

Detection aka *not dealing with certain data*
- MagNet
- Taboo Trap
- Uncertainty
- Trapdoors
Why do we need to detect?

Class Dog

Class Cat
Detection can help

Need to handle non-class space

Need to handle adversarial samples
Why do we need to detect?

Space of class morphs - catdogs

Class Dog

Class Cat

No signal

Different task
$\epsilon$-robustness

Errors and precision loss

Inverse adversarial samples
Taboo Trap

- During training, restrict the numerical range of activations
- Detect when activations are out of bounds

Can we use this to make attacks detectable?
Certifiable Taboo Trap (CTT)

For natural data $X$ enforce constraints on $f$ to be below $T$

For $\varepsilon$-ball around the data point $X' = X \pm \varepsilon$ enforce that upper bound of $f(X') \geq T$

$Upper bound of f(x)$ with IBP

Moving lines closer
Certifiable Taboo Trap (CTT) more generally

- Easily quantifiable space that is either False positive or Undetectable
- Allows for easy certification!
Certifiable Taboo Trap (CTT)

- Natural data can’t be detected
- Space that is always detected
- Space can theoretically be detected
## CTT with MNIST

| Attack | Param | Baseline Acc | AdvTrain Acc | Ensemble Acc | PCL Acc | Det$_t_1$ | MagNet Det$_t_2$ | Det$_t_1$ ||| Det$_t_2$ | CTT-lite Acc | Det | $l_2$ |
|--------|-------|--------------|--------------|--------------|---------|-----------|-----------------|----------|----------|-------------|-------|-------|
| No Attack | 99.1 | 99.5 | 99.5 | 99.3 | 1.75 | 1.93 | 2.93 | 99.1 | 1.9 | - |
| FGSM $\epsilon = 0.1$ | 66.7 | 73.0 | 96.3 | 96.5 | 54.49 | 54.59 | 54.80 | 70.9 | 1.4 | 2.08 |
| $\epsilon = 0.2$ | 25.7 | 52.7 | 52.8 | 77.9 | 85.20 | 85.31 | 85.31 | 21.9 | 1.0 | 4.14 |
| BIM $\epsilon = 0.1$ | 49.4 | 62.0 | 88.5 | 92.1 | 80.82 | 24.90 | 80.92 | 44.2 | 1.0 | 1.13 |
| $\epsilon = 0.15$ | 15.4 | 18.7 | 73.6 | 77.3 | 88.37 | 37.14 | 88.47 | 4.2 | 0.8 | 1.48 |
| PGD $\epsilon = 0.1$ | 59.4 | 62.7 | 82.8 | 93.9 | 83.78 | 77.96 | 83.78 | 51.0 | 1.2 | 1.50 |
| $\epsilon = 0.2$ | 1.83 | 31.9 | 41.0 | 80.2 | 98.27 | 98.27 | 98.27 | 0.0 | 1.1 | 2.73 |
| CTT-loose | 98.5 | 1.6 | - | 98.9 | 1.1 | - |
| CTT-strict | 25.0 | 100.0 | 1.98 | 61.1 | 100.0 | 1.99 |
| | 15.0 | 100.0 | 3.89 | 32.7 | 100.0 | 3.90 |
| | 0.0 | 100.0 | 0.38 | 0.15 | 100.0 | 0.75 |
| | 0.0 | 100.0 | 0.50 | 2.0 | 100.0 | 0.97 |
| | 1.0 | 100.0 | 1.24 | 13.4 | 100.0 | 1.35 |
| | 0.0 | 100.0 | 2.43 | 0.9 | 100.0 | 2.53 |

- CTT can detect strong attackers with MNIST
- CTT outperforms other methods with comparable false positives
CTT with Cifar10

<table>
<thead>
<tr>
<th>Attack</th>
<th>Param</th>
<th>Baseline Acc</th>
<th>AdvTrain Acc</th>
<th>Ensemble Acc</th>
<th>PCL Acc</th>
<th>MagNet $Det_{l_1}$</th>
<th>MagNet $Det_{l_2}$</th>
<th>MagNet $Det_{l_1} \parallel l_2$</th>
<th>CTT-loose Acc</th>
<th>CTT-loose $Det_{l_2}$</th>
<th>CTT-strict Acc</th>
<th>CTT-strict $Det_{l_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Attack</td>
<td></td>
<td>89.1</td>
<td>84.5</td>
<td>90.6</td>
<td>91.9</td>
<td>6.40</td>
<td>6.61</td>
<td>8.13</td>
<td>86.2</td>
<td>3.4</td>
<td>-</td>
<td>86.3</td>
</tr>
<tr>
<td>FGSM $\epsilon = 0.02$</td>
<td></td>
<td>33.6</td>
<td>44.3</td>
<td>61.7</td>
<td>78.5</td>
<td>7.80</td>
<td>6.64</td>
<td>9.55</td>
<td>18.6</td>
<td>95.7</td>
<td>1.07</td>
<td>16.8</td>
</tr>
<tr>
<td>FGSM $\epsilon = 0.04$</td>
<td></td>
<td>22.4</td>
<td>31.0</td>
<td>46.2</td>
<td>69.9</td>
<td>11.53</td>
<td>8.38</td>
<td>13.27</td>
<td>7.6</td>
<td>93.6</td>
<td>2.00</td>
<td>7.2</td>
</tr>
<tr>
<td>BIM $\epsilon = 0.01$</td>
<td></td>
<td>13.5</td>
<td>22.6</td>
<td>46.6</td>
<td>74.5</td>
<td>6.98</td>
<td>6.52</td>
<td>8.61</td>
<td>0.5</td>
<td>9.0</td>
<td>0.15</td>
<td>0.0</td>
</tr>
<tr>
<td>BIM $\epsilon = 0.02$</td>
<td></td>
<td>1.5</td>
<td>7.8</td>
<td>31.0</td>
<td>57.3</td>
<td>6.64</td>
<td>6.52</td>
<td>8.50</td>
<td>0.0</td>
<td>14.2</td>
<td>0.21</td>
<td>0.0</td>
</tr>
<tr>
<td>PGD $\epsilon = 0.01$</td>
<td></td>
<td>24.0</td>
<td>24.3</td>
<td>48.4</td>
<td>75.7</td>
<td>7.10</td>
<td>6.52</td>
<td>8.73</td>
<td>0.1</td>
<td>10.4</td>
<td>0.34</td>
<td>2.9</td>
</tr>
<tr>
<td>PGD $\epsilon = 0.02$</td>
<td></td>
<td>2.9</td>
<td>7.8</td>
<td>30.4</td>
<td>48.5</td>
<td>6.98</td>
<td>6.52</td>
<td>8.85</td>
<td>0.0</td>
<td>40.8</td>
<td>0.65</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- CTT can detect some strong attackers with Cifar10
- CTT outperforms some other methods with comparable false positives
Towards more usable detection schemes

- Lesson from system security: *every system breaks*
- Manipulation must be expected and detected
- Recovery should be easy
- Diversity is paramount
- Detection and defence mechanisms can and should be used together
- Robust situational awareness is the missing link
Towards more usable detection schemes

- CTT can use **different keys** by using different neurons detection
  - If one model is compromised others are not affected
- CTT is simple and fast
  - It can run on **any hardware** that can run the network
- CTT can be used to enforce **strict detection** of specific data regions
Thank you very much for listening!

Please do not hesitate to reach out in case there are any questions at

ilia.shumailov@cl.cam.ac.uk