Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami

May 24th, 2016 @ 37th IEEE Symposium on Security and Privacy
Deep Learning for Classification
Input Layer

Hidden Layers
(e.g., convolutional, rectified linear, …)

Output Layer

Neuron

Weighted Link (weight is a parameter part of $\theta_O$)
Input Layer

Hidden Layers
(e.g., convolutional, rectified linear, …)

Output Layer

M components

N components

Neuron

Weighted Link (weight is a parameter part of $\theta_O$)
Input Layer

M components

Hidden Layers
(e.g., convolutional, rectified linear, ...)

Output Layer

N components

Neuron

Weighted Link (weight is a parameter part of $\theta_O$)
M components
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, ...)
Output Layer
N components

Neuron
Weighted Link (weight is a parameter part of $\theta_O$)
M components

N components

Input Layer

Hidden Layers
(e.g., convolutional, rectified linear, ...)

Output Layer

Neuron

Weighted Link (weight is a parameter part of $\theta_0$)
Input Layer

M components

Hidden Layers
(e.g., convolutional, rectified linear, …)

N components

Output Layer

Neuron

Weighted Link (weight is a parameter part of $\theta_O$)
Input Layer

Hidden Layers
(e.g., convolutional, rectified linear, ...)

Output Layer

Neuron

Weighted Link (weight is a parameter part of $\theta_O$)
Input Layer

Hidden Layers (e.g., convolutional, rectified linear, …)

Output Layer

M components

N components

Neuron

Weighted Link (weight is a parameter part of $\theta_O$)

$\begin{align*}
p_0 &= 0.01 \\
p_1 &= 0.93 \\
p_8 &= 0.02 \\
p_N &= 0.01
\end{align*}$
Speech Recognition as Probabilistic Transduction

Audio | Frame | State | Phoneme | Word | Sentence | Meaning

Feature Extraction | Decision Trees | Acoustic Model | Language Model | Lexicon | NLP

Source: Tara N. Sainath, Google @ ICML DL Workshop 2015
Adversarial Samples
<table>
<thead>
<tr>
<th>Input class</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Output classification:

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

CIFAR10 Dataset:

- bird
- airplane
- truck
- automobile
- bird

GTSRD Dataset:

- traffic light
- parking
- pedestrian crossing
Adversarial strategy

Direction Sensitivity Estimation

Perturbation Selection

Misclassification Check for:
\[ F(X + \delta X) = 4 \]

Adversarial Sample misclassified as “4” by a DNN
\[ F(X^*) = 4 \]
Defending against Adversarial Perturbations
$\rho_{adv}(F) = E_\mu[\Delta_{adv}(X, F)]$

$\Delta_{adv}(X, F) = \arg\min_{\delta X} \{||\delta X|| : F(X + \delta X) \neq F(X)\}$
Defense Design

- Low impact on the architecture
- Maintain accuracy
- Robust in space relatively close to the legitimate distribution
- Maintain speed of network
\[ F(X) = \left[ \frac{e^{z_i(X)/T}}{\sum_{l=0}^{N-1} e^{z_l(X)/T}} \right]_{i \in 0\ldots N-1} \]
Defensive Distillation

1. Training Data $X$
2. Training Labels $Y$
Defensive Distillation

DNN $F$ trained at temperature $T$

Training Data $X$  Training Labels $Y$
Defensive Distillation

1. Training Data $X$

2. DNN $F$ trained at temperature $T$

3. Probability Vector Predictions $F(X)$
Defensive Distillation

1. Training Data X
2. DNN F trained at temperature T
3. Probability Vector Predictions F(X)
4. Training Data X

Class Probabilities Knowledge

0.02 0.02 0.04 0.02
0.02 0.02 0.04 0.02
0.02 0.02 0.04 0.02
0.02 0.02 0.04 0.02
Defensive Distillation

1. Training Data X
2. DNN \( F \) trained at temperature \( T \)
3. Probability Vector Predictions \( F(X) \)
4. Training Data X
5. DNN \( F_{d}(X) \) trained at temperature \( T \)

Class Probabilities Knowledge
Defensive Distillation

1. Training Data X
2. DNN F trained at temperature T
3. Probability Vector Predictions F(X)
4. Training Labels F(X)
5. DNN F^d(X) trained at temperature T
6. Probability Vector Predictions F^d(X)

Class Probabilities Knowledge
Defensive Distillation

1. Training Data X
2. DNN F trained at temperature T
3. Probability Vector Predictions F(X)
4. Training Labels Y
5. DNN F^d(X) trained at temperature T
6. Probability Vector Predictions F^d(X)

Initial Network

Distilled Network
Defensive Distillation

Set temperature $T=1$ for predictions
Intuition behind Defensive Distillation

**Constraining Training**

\[
\arg \min_{\theta_F} - \frac{1}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} \sum_{i=0\ldots N} Y_i(X) \log F_i(X)
\]

0 if i not correct class

\[
\arg \min_{\theta_F} - \frac{1}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} \sum_{i=0\ldots N} F_i(X) \log F_i^d(X)
\]

never equal to 0

**Reducing Jacobian Amplitudes**

\[
J_F(T, i, j) = \frac{1}{Tg^2(X)} \left( \sum_{l=0}^{N-1} \left( \frac{\partial z_i}{\partial X_j} - \frac{\partial z_l}{\partial X_j} \right) e^{z_l/T} \right)
\]
Validation
Experimental Setup
Adversarial Sample Success Rate

Distillation Temperature

<table>
<thead>
<tr>
<th>Distillation Temperature</th>
<th>MNIST Adversarial Samples Success Rate (%)</th>
<th>CIFAR10 Adversarial Samples Success Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>91</td>
<td>92.78</td>
</tr>
<tr>
<td>2</td>
<td>82.23</td>
<td>87.67</td>
</tr>
<tr>
<td>5</td>
<td>24.67</td>
<td>67</td>
</tr>
<tr>
<td>10</td>
<td>6.78</td>
<td>47.56</td>
</tr>
<tr>
<td>20</td>
<td>1.34</td>
<td>18.23</td>
</tr>
<tr>
<td>30</td>
<td>1.44</td>
<td>13.23</td>
</tr>
<tr>
<td>40</td>
<td>0.45</td>
<td>9.34</td>
</tr>
<tr>
<td>50</td>
<td>1.45</td>
<td>6.23</td>
</tr>
<tr>
<td>100</td>
<td>0.45</td>
<td>5.11</td>
</tr>
<tr>
<td>No distillation</td>
<td>95.89</td>
<td>87.89</td>
</tr>
</tbody>
</table>
Impact on accuracy

### Adversarial Sample Success Rate

<table>
<thead>
<tr>
<th>Distillation Temperature</th>
<th>MNIST Adversarial Samples Success Rate (%)</th>
<th>CIFAR10 Adversarial Samples Success Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>91</td>
<td>92.78</td>
</tr>
<tr>
<td>2</td>
<td>82.23</td>
<td>87.67</td>
</tr>
<tr>
<td>5</td>
<td>24.67</td>
<td>67</td>
</tr>
<tr>
<td>10</td>
<td>6.78</td>
<td>47.56</td>
</tr>
<tr>
<td>20</td>
<td>1.34</td>
<td>18.23</td>
</tr>
<tr>
<td>30</td>
<td>3.44</td>
<td>13.23</td>
</tr>
<tr>
<td>40</td>
<td><strong>0.45</strong></td>
<td><strong>6.64</strong></td>
</tr>
<tr>
<td>50</td>
<td><strong>0.45</strong></td>
<td><strong>6.23</strong></td>
</tr>
<tr>
<td>100</td>
<td><strong>0.45</strong></td>
<td><strong>9.11</strong></td>
</tr>
<tr>
<td>No distillation</td>
<td><strong>95.89</strong></td>
<td><strong>87.89</strong></td>
</tr>
</tbody>
</table>

### Adversarial Samples Baseline Rate

- **MNIST**
- **CIFAR10**
Impact on Jacobian Amplitude

![Graph showing the impact of distillation temperature on Jacobian amplitude frequencies. The x-axis represents different distillation temperatures (T=1, T=2, T=5, T=10, T=20, T=30, T=40, T=50, T=100), and the y-axis represents the frequency of adversarial gradient mean amplitudes. The graph uses color coding to indicate different ranges of amplitudes.]
Estimation of Robustness

\[ \Delta_{adv}(X, F) \]

Network Robustness (in percentage of input features) vs. Distillation Temperature

- DNN Robustness (MNIST)
- DNN Robustness Baseline (MNIST)
- DNN Robustness (CIFAR10)
- DNN Robustness Baseline (CIFAR10)
Conclusions
Take aways

• Distillation significantly reduces attack success

• Yields model smoothness

• Easy implementation, low overhead

• Acceptable impact on accuracy
Questions?

https://www.papernot.fr
nicolas@papernot.fr
@NicolasPapernot