Defending Against Adversarial Examples

Ian Goodfellow, Staff Research Scientist, Google Brain
NIPS 2017 Workshop on Machine Learning and Security
Adversarial Examples

Timeline:
“Adversarial Classification” Dalvi et al 2004: fool spam filter
“Evasion Attacks Against Machine Learning at Test Time” Biggio 2013: fool neural nets
Szegedy et al 2013: fool ImageNet classifiers imperceptibly
Goodfellow et al 2014: cheap, closed form attack
Cross-model, cross-dataset generalization
Cross-technique transferability

<table>
<thead>
<tr>
<th>Source Machine Learning Technique</th>
<th>DNN</th>
<th>LR</th>
<th>SVM</th>
<th>DT</th>
<th>kNN</th>
<th>Ens.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN</td>
<td>38.27</td>
<td>23.02</td>
<td>64.32</td>
<td>79.31</td>
<td>8.36</td>
<td>20.72</td>
</tr>
<tr>
<td>LR</td>
<td>6.31</td>
<td>91.64</td>
<td>91.43</td>
<td>87.42</td>
<td>11.29</td>
<td>44.14</td>
</tr>
<tr>
<td>SVM</td>
<td>2.51</td>
<td>36.56</td>
<td>100.0</td>
<td>80.03</td>
<td>5.19</td>
<td>15.67</td>
</tr>
<tr>
<td>DT</td>
<td>0.82</td>
<td>12.22</td>
<td>8.85</td>
<td>89.29</td>
<td>3.31</td>
<td>5.11</td>
</tr>
<tr>
<td>kNN</td>
<td>11.75</td>
<td>42.89</td>
<td>82.16</td>
<td>82.95</td>
<td>41.65</td>
<td>31.92</td>
</tr>
</tbody>
</table>

(Papernot 2016)

(Goodfellow 2017)
Enhancing Transfer With Ensembles

<table>
<thead>
<tr>
<th></th>
<th>RMSD</th>
<th>ResNet-152</th>
<th>ResNet-101</th>
<th>ResNet-50</th>
<th>VGG-16</th>
<th>GoogLeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-152</td>
<td>17.17</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>17.25</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>17.25</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>VGG-16</td>
<td>17.80</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>17.41</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Table 4: Accuracy of non-targeted adversarial images generated using the optimization-based approach. The first column indicates the average RMSD of the generated adversarial images. Cell $(i, j)$ corresponds to the accuracy of the attack generated using four models except model $i$ (row) when evaluated over model $j$ (column). In each row, the minus sign “−” indicates that the model of the row is not used when generating the attacks. Results of top-5 accuracy can be found in the appendix (Table 14).

(Liu et al, 2016)
Transferability Attack

- Target model with unknown weights, machine learning algorithm, training set; maybe non-differentiable
- Substitute model mimicking target model with known, differentiable function
- Deploy adversarial examples against the target; transferability property results in them succeeding
- Adversarial examples

(Szegedy 2013, Papernot 2016)
Thermometer Encoding: One Hot Way to Resist Adversarial Examples

Jacob Buckman*

Aurko Roy*

Colin Raffel

Ian Goodfellow

*joint first author
Linear Extrapolation

Vulnerabilities
Neural nets are “too linear”

Plot from “Explaining and Harnessing Adversarial Examples”, Goodfellow et al, 2014
Difficult to train extremely nonlinear hidden layers

To train:
changing this weight needs to have a large, predictable effect

To defend:
changing this input needs to have a small or unpredictable effect
Idea: edit only the input layer

Train only this part

DEFENSE
<table>
<thead>
<tr>
<th>Real-valued</th>
<th>Quantized</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td>0.92</td>
<td>0.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discretized (one-hot)</th>
<th>Discretized (thermometer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[01000000000]</td>
<td>[01111111111]</td>
</tr>
<tr>
<td>[00000010000]</td>
<td>[00000001111]</td>
</tr>
<tr>
<td>[00000000001]</td>
<td>[00000000001]</td>
</tr>
</tbody>
</table>
Observation: PixelRNN shows one-hot codes work

Plot from “Pixel Recurrent Neural Networks”, van den Oord et al, 2016
Fast Improvement Early in Learning

![Graph showing accuracy over hours for different adversarial training methods.](https://example.com/graph.png)
Large improvements on SVHN direct ("white box") attacks

5 years ago, this would have been SOTA on clean data

(Goodfellow 2017)
Large Improvements against CIFAR-10 direct ("white box") attacks

6 years ago, this would have been SOTA on clean data
Other results

• Improvement on CIFAR-100
  • (Still very broken)

• Improvement on MNIST
  • Please quit caring about MNIST
Caveats

• Slight drop in accuracy on clean examples

• Only small improvement on black-box transfer-based adversarial examples
Ensemble Adversarial Training

Florian Tramèr
Alexey Kurakin
Nicolas Papernot
Ian Goodfellow
Dan Boneh
Patrick McDaniel
Estimating the Subspace Dimensionality

(Tramèr et al, 2017)
Transfer Attacks Against Inception ResNet v2 on ImageNet

![Bar chart showing worst black-box test accuracy for different models: Standard, Adv Trained, Ensemble Adv Trained. The accuracy ranges from 82.2 to 92.1.](Goodfellow 2017)
Best defense so far on ImageNet:
Ensemble adversarial training.
Used as at least part of all top 10 entries in dev round 3

AI Fight Club Could Help Save Us from a Future of Super-Smart Cyberattacks
Get involved!

https://github.com/tensorflow/cleverhans