Thermometer Encoding: One Hot Way to Resist Adversarial Examples

Stanford, 2017-11-16

Jacob Buckman*

Aurko Roy*

Colin Raffel

Ian Goodfellow

*joint first author
Adversarial Examples

Probably panda

\[ + \; 0.007 \times \]

Adversarial perturbation

= Definitely gibbon

Image from “Explaining and Harnessing Adversarial Examples”, Goodfellow et al, 2014
Unreasonable Linear Extrapolation

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
<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>[00000011111]</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
Large improvements on SVHN white box attacks

5 years ago, this would have been SOTA on clean data
Large Improvements against CIFAR-10 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 adversarial examples
Get involved!
https://github.com/tensorflow/cleverhans
g.co/airesidency