Physical Adversarial Examples

Alex Kurakin    Ian Goodfellow

Google™ OpenAI
Machine Learning

Training Examples

- BICYCLE
- CAR
- PEDESTRIAN

ImageNet (Russakovsky et al 2015)

Parameters

Input

Hidden units / features

Output

STOP
Adversarial Examples: *Images*

(Figure credit: Nicolas Papernot)
Turning Objects into “Airplanes”
Fast Gradient Sign Method (FGSM)

\[ x \]

“panda”
57.7% confidence

\[ + 0.007 \times \text{sign}(\nabla_x J(\theta, x, y)) \]

“nematode”
8.2% confidence

\[ = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \]

“gibbon”
99.3% confidence
Maps of Adversarial Examples
Almost all inputs are misclassified
Generalization across training sets
Cross-Technique Transferability

(Papernot et al 2016)
Transferability attack

Target model with unknown weights, machine learning algorithm, training set; maybe non-differentiable

Train your own model

Substitute model mimicking target model with known, differentiable function

Deploy adversarial examples against the target; transferability property results in them succeeding

Adversarial examples

Adversarial crafting against substitute
# Results on Real-World Remote Systems

All remote classifiers are trained on the MNIST dataset (10 classes, 60,000 training samples)

<table>
<thead>
<tr>
<th>Remote Platform</th>
<th>ML technique</th>
<th>Number of queries</th>
<th>Adversarial examples misclassified (after querying)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetaMind</td>
<td>Deep Learning</td>
<td>6,400</td>
<td>84.24%</td>
</tr>
<tr>
<td>Amazon Web Services</td>
<td>Linear Regression</td>
<td>800</td>
<td>96.19%</td>
</tr>
<tr>
<td>Google Cloud Platform</td>
<td>Unknown</td>
<td>2,000</td>
<td>97.72%</td>
</tr>
</tbody>
</table>

(Papernot et al 2016)
Adversarial examples in the physical world?

- **Question:** Can we build adversarial examples in the physical world?

- Let’s try the following:
  - Generate and print picture of adversarial example
  - Take a photo of this picture (with cellphone camera)
  - Crop+warp picture from the photo to make it 299x299 input to Imagenet inception
  - Classify this image

- **Would the adversarial image remain misclassified after this transformation?**

- If we succeed with “photo” then we potentially can alter real-world objects to mislead deep-net classifiers
Adversarial examples in the physical world?

- Question: Can we build adversarial examples in the physical world?
- Let’s try the following:
  - Generate and print picture of adversarial example
  - Take a photo of this picture (with cellphone camera)
  - Crop+warp picture from the photo to make it 299x299 input to Imagenet inception
  - Classify this image
- Would the adversarial image remain misclassified after this transformation?
- If we succeed with “photo” then we potentially can alter real-world objects to mislead deep-net classifiers

Answer: IT’S POSSIBLE
Digital adversarial examples

[ Goodfellow, Shlens & Szegedy, ICLR2015 ]
Adversarial examples in the physical world

[ Kurakin & Goodfellow & Bengio, arxiv.org/abs/1607.02533 ]
Our experiment

1. Print pairs of normal and adversarial images

2. Take picture

3. Auto crop and classify

Up to 87% of images could remain misclassified!
Live demo

Library

Washer

Washer
Don’t panic! It’s not end of the ML world!

- Our experiment is a proof-of-concept set up:
  - We had full access to the model
  - 87% adversarial images rate is for only one method, which could be resisted by adversarial training. For other methods it’s much lower.
  - In many cases “adversarial” image is not so harmful: one breed of dog confused with another

- In practice:
  - Attacker doesn’t have access to model
  - You might be able to use adversarial training to defend model against some attacks
  - For other attacks, “adversarial examples in the real worlds” won’t work that well
  - It’s REALLY hard to fool your model to predict specific class