Synthesizing Robust Adversarial Examples

Anish Athalye*, Logan Engstrom*, Andrew Ilyas*, Kevin Kwok
Adversarial examples
Adversarial examples

• Imperceptible perturbations to an input can change a neural network’s prediction

88% tabby cat → 99% guacamole
Adversarial examples

**Given:** Input image $x$, target label $y$

**Optimize:**

$$\arg \max_{x'} P(y \mid x')$$

subject to $d(x, x') < \epsilon$
Do adversarial examples work in the physical world?
Adversarial examples in the physical world

(a) Image from dataset  (b) Clean image  (c) Adv. image, $\epsilon = 4$

(Kurakin et al. 2016)
... or not?

Foveation-based Mechanisms Alleviate Adversarial Examples (Luo et al. 2015)

NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles (Lu et al. 2017)
Standard examples are fragile
Are adversarial examples fundamentally fragile?
Image processing pipeline

optimize $P(y \mid x')$ using gradient descent
Physical world processing pipeline

Challenge: No direct control over model input
Attack: Expectation Over Transformation

\[
\text{optimize } \mathbb{E}_{t \sim T} \left[ P(y \mid t(x')) \right] \text{ using gradient descent (sampling, chain rule, differentiating through } t) 
\]
EOT produces robust examples

Zoom: 1.000000x

$T = \{\text{rescale from 1x to 5x}\}$
EOT produces robust physical-world examples

\[ T = \{ \text{rescale} + \text{rotate} + \text{translate} + \text{skew} \} \]
Can we make this work with 3D objects?
Physical world 3D processing pipeline

TEXTURE → RENDERING → MODEL → PREDICTIONS

is this differentiable?

PARAMETERS

zoom: 1.3x
rotation: [60°, 30°, 15°]
translation: [1, 5, 0]
...
• For any pose, 3D rendering is differentiable with respect to texture

• Simplest renderer: linear transformation of texture
EOT produces 3D adversarial objects
EOT reliably produces 3D adversarial objects

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Classification accuracy</th>
<th>Attacker success rate</th>
<th>Distortion (L2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2D</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td>70%</td>
<td>N/A</td>
<td>0</td>
</tr>
<tr>
<td>Adversarial</td>
<td>0.9%</td>
<td>96.4%</td>
<td>$5.6 \times 10^{-5}$</td>
</tr>
<tr>
<td><strong>3D</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td>84%</td>
<td>N/A</td>
<td>0</td>
</tr>
<tr>
<td>Adversarial</td>
<td>1.7%</td>
<td>84.0%</td>
<td>$6.5 \times 10^{-5}$</td>
</tr>
</tbody>
</table>
Implications

• Defenses based on randomized input transformations are insecure

• Adversarial examples / objects are a physical-world concern

Poster (and live demo): 6:15 – 9:00pm @ Hall B #73