The Odds are Odd:
A Statistical Test for Detecting Adversarial Examples

Kevin Roth*, Yannic Kilcher*, Thomas Hofmann

ETH Zürich
Log-Odds & Adversarial Examples

\[ f_y(x) = \langle w_y, \phi(x) \rangle \]
Adversarial examples cause atypically large feature space perturbations along the weight-difference direction.
Adversarial Cone

\[ \mathbf{x}^* \]
Adversarial Cone
Adversarial Cone
Adversarial Cone

\[ P_{y^*}(. \neq x) = 1 \]

\[ P_{y^*}(x) = 0 \]
Adversarial Cone

\[ P_{y^*}(.) = 1 \]

\[ P_{y^*}(.) = 0 \]
Adversarial Cone

Adversarial examples are embedded in a cone-like structure.
Adversarial Cone

$$x_{\text{adv}} + t \times \text{noise}$$
Adversarial Cone

\[ x_{\text{adv}} + t \times \text{noise} \]
Adversarial Cone

Noise as a probing instrument
The robustness properties of $\phi(x + n_i)$ are different dependent on whether $x = x^*$ or $x = x^* + \Delta x$

$\Delta \phi$ tends to have a **characteristic direction** if $x = x^* + \Delta x$ whereas it tends not to have a specific direction if $x = x^*$
Main Idea: Log-Odds Robustness

Noise can partially undo effect of adversarial perturbation and directionally revert log-odds towards the true class $y^*$
We propose to use noise-perturbed pairwise log-odds

\[ g_{y,z}(x, \eta) = \langle w_z - w_y, \phi(x + \eta) - \phi(x) \rangle \]

to test whether \( x \) classified as \( y \) should be thought of as a manipulated example of true class \( z \):

\( x \) adversarial if \( \max_{z \neq y} \{ \mathbb{E}_\eta [\bar{g}_{y,z}(x, \eta)] - \tau_{y,z} \} \geq 0 \)

Corrected classification:
\[ G(x) = \arg \max_z \{ \bar{g}_{y,z}(x) - \tau_{y,z} \} \]
Our statistical test detects nearly all adversarial examples with FPR \(\sim 1\%\).

Our correction method reclassifies almost all adversarial examples successfully.

Drop in performance on clean samples is negligible.

**Table 1: CIFAR10**

<table>
<thead>
<tr>
<th>Model</th>
<th>Detection rate (clean / pgd)</th>
<th>Corrected Accuracy (clean / pgd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WResNet</td>
<td>0.2% / 99.1%</td>
<td>96.0% / 92.7%</td>
</tr>
<tr>
<td>CNN7</td>
<td>0.8% / 95.0%</td>
<td>93.6% / 89.5%</td>
</tr>
<tr>
<td>CNN4</td>
<td>1.4% / 93.8%</td>
<td>71.0% / 67.6%</td>
</tr>
</tbody>
</table>

**Table 2: ImageNet**

<table>
<thead>
<tr>
<th>Model</th>
<th>Detection rate (clean / pgd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception V3</td>
<td>1.9% / 99.6%</td>
</tr>
<tr>
<td>ResNet 101</td>
<td>0.8% / 99.8%</td>
</tr>
<tr>
<td>VGG16(+BN)</td>
<td>0.3% / 99.9%</td>
</tr>
</tbody>
</table>
Detection Rates & Corrected Classification

Detection rate increases with increasing attack strength.

Corrected classification manages to compensate for decay in uncorrected accuracy due to increase in attack strength.
Defending against Defense-Aware Attacks

<table>
<thead>
<tr>
<th>Model</th>
<th>Detection rate (clean / attack)</th>
<th>Accuracy (clean / attack)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WResNet</td>
<td>4.5% / 71.4%</td>
<td>91.7% / 56.0%</td>
</tr>
<tr>
<td>CNN7</td>
<td>2.8% / 75.5%</td>
<td>91.2% / 56.6%</td>
</tr>
<tr>
<td>CNN4</td>
<td>4.1% / 81.3%</td>
<td>69.0% / 56.5%</td>
</tr>
</tbody>
</table>

Attacker has **full knowledge of the defense**: perturbations that work in expectation under noise source used for detection.

Detection rates and corrected accuracies remain remarkably high.
Thank You

Follow-Up Work: Adversarial Training Generalizes  
Data-dependent Spectral Norm Regularization

ICML Workshop on Generalization (June 14)
The approaches most related to our work are those that detect whether or not the input has been perturbed, either by detecting characteristic regularities in the adversarial perturbations themselves or in the network activations they induce.

- ... and many more