Intriguing Adversarial Examples
&
How To Defend Against Them

Cihang Xie
Johns Hopkins University
Deep networks are **Good**

```
Label: King Penguin
```

Deep Networks
Deep networks are **FRAGILE** to small & carefully crafted perturbations
Deep networks are **FRAGILE** to small & carefully crafted perturbations

We call such images as **Adversarial Examples**
Generating Adversarial Example is **SIMPLE**: 

\[
\text{maximize } \text{loss}(f(x+r), y^{\text{true}}; \theta)
\]

**Maximize** the loss function w.r.t. **Adversarial Perturbation** \( r \)
Generating Adversarial Example is **SIMPLE**:

\[
\text{maximize } \text{loss}(f(x+r), y^{true}; \theta) \\
\text{Minimize the loss function w.r.t. Adversarial Perturbation } r
\]

\[
\text{minimize } \text{loss}(f(x), y^{true}; \theta); \\
\text{Minimize the loss function w.r.t. Network Parameters } \theta
\]
Part I: Intriguing Properties of Adversarial Examples

- {Image, Model, Task}-Agnostic
- Beyond Pixel Perturbation
- Existence in Physical World
Part I: Intriguing Properties of Adversarial Examples

- \{Image, Model, Task\}-Agnostic
- Beyond Pixel Perturbation
- Existence in Physical World
Adversarial Perturbations can be **Image Agnostic**
Adversarial Perturbations can be **Image Agnostic**

We call such perturbations as **Universal Adversarial Perturbations**
Adversarial Examples can be **Model Agnostic**
Adversarial Examples can be **Model Agnostic**

We call such images as **Transferable Adversarial Examples**
Adversarial Examples can be **Task Agnostic**

Adversarial examples **EXIST** on different tasks
Adversarial Examples can be **Task Agnostic**

Adversarial examples **EXIST** on different tasks
Adversarial Examples can be **Task Agnostic**

Adversarial examples **EXIST** on different tasks
Adversarial Examples can be **Task Agnostic**

Adversarial examples **EXIST** on different tasks
Adversarial Examples can be **Task Agnostic**

Adversarial examples **TRANSFER** between different tasks
Adversarial Examples can be **Task Agnostic**

Adversarial examples **TRANSFER** between different tasks
Quantitative Result of Transferability between Different Models [1]

<table>
<thead>
<tr>
<th>Model</th>
<th>Attack</th>
<th>Inc-v3</th>
<th>Inc-v4</th>
<th>IncRes-v2</th>
<th>Res-152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc-v3</td>
<td>FGSM</td>
<td>64.6%</td>
<td>23.5%</td>
<td>21.7%</td>
<td>21.7%</td>
</tr>
<tr>
<td></td>
<td>I-FGSM</td>
<td>99.9%</td>
<td>14.8%</td>
<td>11.6%</td>
<td>8.9%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>99.9%</td>
<td>35.5%</td>
<td>27.8%</td>
<td>21.4%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>99.9%</td>
<td>36.6%</td>
<td>34.5%</td>
<td>27.5%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>99.9%</td>
<td>63.9%</td>
<td>59.4%</td>
<td>47.9%</td>
</tr>
</tbody>
</table>

Adversarial examples generated on Inc-v3 can attack Inc-v4, IncRes-v2 and Res-152 with high success rate.

Quantitative Result of Transferability between Different Models [1]

<table>
<thead>
<tr>
<th>Model</th>
<th>Attack</th>
<th>Inc-v3</th>
<th>Inc-v4</th>
<th>IncRes-v2</th>
<th>Res-152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc-v3</td>
<td>FGSM</td>
<td>64.6%</td>
<td>23.5%</td>
<td>21.7%</td>
<td>21.7%</td>
</tr>
<tr>
<td></td>
<td>I-FGSM</td>
<td>99.9%</td>
<td>14.8%</td>
<td>11.6%</td>
<td>8.9%</td>
</tr>
<tr>
<td></td>
<td>$\text{Di}^2$-FGSM (Ours)</td>
<td>99.9%</td>
<td>35.5%</td>
<td>27.8%</td>
<td>21.4%</td>
</tr>
<tr>
<td>MI-FGSM</td>
<td></td>
<td>99.9%</td>
<td>36.6%</td>
<td>34.5%</td>
<td>27.5%</td>
</tr>
<tr>
<td>M-$\text{Di}^2$-FGSM (Ours)</td>
<td></td>
<td>99.9%</td>
<td>63.9%</td>
<td>59.4%</td>
<td>47.9%</td>
</tr>
</tbody>
</table>

Adversarial examples generated on Inc-v3 can attack Inc-v4, IncRes-v2 and Res-152 with high success rate.

This transfer phenomenon may indicates

**Different Networks Learn Similar Representations**

Part I: Intriguing Properties of Adversarial Examples

- {Image, Model, Task}-Agnostic
- Beyond Pixel Perturbation
- Existence in the Physical World
Beyond Pixel Perturbations --- Spatially Transformed Adversary [2]

Only Rotation & Translation Are Enough! [3]

Engstrom, Logan, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. "A rotation and a translation suffice: Fooling cnns with simple transformations." In ICML. 2019
Beyond Pixel Perturbations --- **Adversarial Context Examples** [4]

Part I: Intriguing Properties of Adversarial Examples

- {Image, Model, Task}-Agnostic
- Beyond Pixel Perturbation
- Existence in the Physical World

(a) Image from dataset  (b) Clean image  (c) Adv. image


With these adversarial stickers, networks cannot recognize stop signs.

Extension --- Attacking Object Detectors in the Physical World [7]

Part II: Towards Robust Adversarial Defense

- Robust Input Images
- Robust Network Representations
Part II: Towards Robust Adversarial Defense

- Robust Input Images
- Robust Network Representations

want to **remove** malicious manipulations from input images
Part II: Towards Robust Adversarial Defense

- Robust Input Images
- Robust Network Representations

want to learn robust representations against adversarial images

Label: King Penguin
Feature Denoising for Improving Adversarial Robustness (CVPR’19)
Observation: Adversarial perturbations are SMALL on the pixel space.
**Observation**: Adversarial perturbations are **BIG** on the feature space.
**Observation:** Adversarial perturbations are **BIG** on the feature space.

We should **DENoise** these feature maps.
Our Solution: **Denoising at feature level**

**Traditional Image Denoising Operations:**

Local filters (predefine a local region $\Omega(i)$ for each pixel $i$):

- **Bilateral filter**
  
  $$y_i = \frac{1}{c(x_i)} \sum_{j \in \Omega(i)} f(x_i, x_j) x_j$$

- **Median filter**
  
  $$y_i = \text{median}\{\forall j \in \Omega(i): x_j\}$$

- **Mean filter**
  
  $$y_i = \frac{1}{c(x_i)} \sum_{j \in \Omega(i)} x_j$$

Non-local filters (the local region $\Omega(i)$ is the whole image $I$):

- **Non-local means**
  
  $$y_i = \frac{1}{c(x_i)} \sum_{j \in I} f(x_i, x_j) x_j$$
Denoising Block Design

Denoising operations may **lose information**

- we add a *residual connection* to balance the tradeoff between removing noise and retaining original signal
Training Strategy: Adversarial training

- Core Idea: train with adversarial examples

- Implementation: distributed on 128 GPUs, 32 images per GPU (since finding adversarial examples is computationally expensive)
Two Ways for Evaluating Robustness

Defending Against White-box Attacks

- Attackers know everything about models
- Directly maximize $\text{loss}(f(x+r), y_{\text{true}}; \theta)$
Two Ways for Evaluating Robustness

Defending Against White-box Attacks

- Attackers know everything about models
- Directly maximize $\text{loss}(f(x+r), y^\text{true}; \theta)$

Defending Against Blind Attacks

- Attackers know nothing about models
- Attackers generate adversarial examples using substitute networks *(rely on transferability)*
Defending Against White-box Attacks

- Evaluating against adversarial attackers with attack iteration up to 2000
  (more attack iterations indicate stronger attacks)
A successful adversarial training can give us a **STRONG** baseline
Defending Against White-box Attacks – Part I

Feature Denoising can give us additional benefits
Defending Against White-box Attacks – Part II

All denoising operations can help
Defending Against White-box Attacks – Part III

Feature Denoising is nearly as powerful as adding ~500 additional layers
Defending Against White-box Attacks – Part III

Feature Denoising can still provide benefits for the VERY deep ResNet-638
Defending Against Blind Attacks

- Offline evaluation against 5 BEST attackers from NeurIPS Adversarial Competition 2017
- Online competition against 48 UNKNOWN attackers in CAAD 2018
Defending Against Blind Attacks

- Offline evaluation against 5 BEST attackers from NeurIPS Adversarial Competition 2017
- Online competition against 48 UNKNOWN attackers in CAAD 2018

**CAAD 2018 “all or nothing” criterion**: an image is considered correctly classified only if the model correctly classifies all adversarial versions of this image created by all attackers.
# Defending Against Blind Attacks --- CAAD 2017 Offline Evaluation

<table>
<thead>
<tr>
<th>model</th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAAD 2017 winner</td>
<td>0.04</td>
</tr>
<tr>
<td>CAAD 2017 winner, under 3 attackers</td>
<td>13.4</td>
</tr>
<tr>
<td>ours, R-152 baseline</td>
<td>43.1</td>
</tr>
<tr>
<td>+4 denoise: null (1×1 only)</td>
<td>44.1</td>
</tr>
<tr>
<td>+4 denoise: non-local, dot product</td>
<td>46.2</td>
</tr>
<tr>
<td>+4 denoise: non-local, Gaussian</td>
<td>46.4</td>
</tr>
<tr>
<td>+all denoise: non-local, Gaussian</td>
<td>49.5</td>
</tr>
</tbody>
</table>
## Defending Against Blind Attacks --- CAAD 2017 Offline Evaluation

<table>
<thead>
<tr>
<th>model</th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAAD 2017 winner</td>
<td>0.04</td>
</tr>
<tr>
<td>CAAD 2017 winner, under 3 attackers</td>
<td>13.4</td>
</tr>
<tr>
<td>ours, R-152 baseline</td>
<td></td>
</tr>
<tr>
<td>+4 denoise: null (1×1 only)</td>
<td>44.1</td>
</tr>
<tr>
<td>+4 denoise: non-local, dot product</td>
<td>46.2</td>
</tr>
<tr>
<td>+4 denoise: non-local, Gaussian</td>
<td><strong>46.4</strong></td>
</tr>
<tr>
<td>+all denoise: non-local, Gaussian</td>
<td><strong>49.5</strong></td>
</tr>
</tbody>
</table>
### Defending Against Blind Attacks --- CAAD 2017 Offline Evaluation

<table>
<thead>
<tr>
<th>model</th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAAD 2017 winner</td>
<td>0.04</td>
</tr>
<tr>
<td>CAAD 2017 winner, under 3 attackers</td>
<td>13.4</td>
</tr>
<tr>
<td>ours, R-152 baseline</td>
<td>43.1</td>
</tr>
<tr>
<td>+4 denoise: null (1×1 only)</td>
<td>44.1</td>
</tr>
<tr>
<td>+4 denoise: non-local, dot product</td>
<td>46.2</td>
</tr>
<tr>
<td>+4 denoise: non-local, Gaussian</td>
<td><strong>46.4</strong></td>
</tr>
<tr>
<td>+all denoise: non-local, Gaussian</td>
<td><strong>49.5</strong></td>
</tr>
</tbody>
</table>
Defending Against Blind Attacks --- CAAD 2018 Online Competition

1st: 50.6
2nd: 40.8
3rd: 8.6
4th: 3.6
5th: 0.6
Visualization

Adversarial Examples

Before denoising

Denoising Operations

After denoising
Defending against adversarial attacks is still a long way to go...
Questions?