Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye*1, Nicholas Carlini*2, and David Wagner3

1 Massachusetts Institute of Technology
2 University of California, Berkeley (now Google Brain)
3 University of California, Berkeley
Or,

Advice on performing adversarial example defense evaluations
88% tabby cat \Rightarrow \text{adversarial perturbation} \Rightarrow 99\% \text{ guacamole}
Adversarial Examples

Definition 1:
Inputs specifically crafted to fool a neural network.

Correct definition.
Hard to formalize.

Definition 2:
Given an input $x$, find an input $x'$ that is misclassified such that $|x-x'| < \varepsilon$

Not complete.
Easy to formalize.
Adversarial Examples

Definition 1

Defn. 2
13 total defense papers at ICLR'18

9 are white-box, non-certified

6 of these are broken
  (~0% accuracy)
1 of these is partially broken
~50% of our paper is our attacks
~50% of our paper is our attacks

This talk is about the other 50%.
This Talk:

How should we evaluate adversarial example defenses?
1. A precise threat model
2. A clear defense proposal
3. A thorough evaluation
A threat model is a **formal** statement defining when a system is intended to be secure.
1. Threat Model

What dataset is considered?

Adversarial example definition?

What does the attacker know?
   (model architecture? parameters? training data? randomness?)

If black-box: are queries allowed?
All Possible Adversaries

Threat Model
All Possible Adversaries

Threat Model
All Possible Adversaries

Threat Model
Good Threat Model:
"Robust when $L_2$ distortion is less than 5, given the attacker has white-box knowledge"

Claim: 90% accuracy on ImageNet
2. Defense Proposal

Precise proposal of one specific defense

(with code and models available)
3. Defense Evaluation

A defense evaluation has one purpose, to answer:

"Is the defense secure under the threat model?"
3. Defense Evaluation

acc, loss = model.evaluate(
    Xtest, Ytest)

Is no longer sufficient.
3. Defense Evaluation

This step is why security is hard
Serious effort to evaluate

By space, most papers are $\frac{1}{2}$ evaluation
Going through the motions is **insufficient** to evaluate a defense to adversarial examples.
The purpose of a defense evaluation is NOT to show the defense is RIGHT
The purpose of a defense evaluation is to **FAIL** to show the defense is **WRONG**
Actionable advice requires specific, concrete examples.

Everything the following papers do is standard practice.
the adversary has access to those networks (but does not have access to the input transformations applied at test time).

2 The white-box attacks defined in this paper should be called oblivious attacks according to Carlini and Wagner’s definition [3]

an adversary gains access to all parameters and weights of a model that is trained on benign images, but is unaware of the defense strategy.

Perform an adaptive attack
We now evaluate on two held out $L_0$ attacks

A "hold out" set is not an adaptive attack
To create adversarial examples in our evaluation, we use FGSM, exclusively.

For the next series of experiments, we test against the Fast Gradient Sign Method.

In our experiment, we use the Fast Gradient Sign Method (FGSM).

TABLE 4: Performance of detecting FGSM adversarial examples with different scalar quantization schemes.

Stop using FGSM (exclusively)
- Number of attack steps: 10

Experiments on CIFAR used $\epsilon = 0.031$ and 7 steps for iterative attacks;

Use more than 100 (or 1000?) iteration of gradient descent
Iterative attacks should always do better than single step attacks.

<table>
<thead>
<tr>
<th>Model</th>
<th>FGSM</th>
<th>PGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>25.10</td>
<td>4.10</td>
</tr>
<tr>
<td></td>
<td>46.15</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>43.89</td>
<td>3.57</td>
</tr>
<tr>
<td></td>
<td>52.07</td>
<td>53.11</td>
</tr>
<tr>
<td></td>
<td>48.50</td>
<td>50.50</td>
</tr>
</tbody>
</table>
Unbounded optimization attacks should eventually reach in 0% accuracy

<table>
<thead>
<tr>
<th>Attack</th>
<th>Parameter</th>
<th>Fooling Rate</th>
<th>Detection Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepFool</td>
<td></td>
<td>99.35%</td>
<td>97.83%</td>
</tr>
<tr>
<td>Carlini</td>
<td>$\kappa=0.0$</td>
<td>100.0%</td>
<td>95.66%</td>
</tr>
</tbody>
</table>
Unbounded optimization attacks should eventually reach in 0% accuracy
Unbounded optimization attacks should eventually reach in 0% accuracy
Model accuracy should be monotonically decreasing
Model accuracy should be monotonically decreasing.
<table>
<thead>
<tr>
<th>Model</th>
<th>clean</th>
<th>step_ll</th>
<th>step_FGSM</th>
<th>iter_FGSM</th>
<th>CW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(\epsilon=2)</td>
<td>(\epsilon=16)</td>
<td>(\epsilon=2)</td>
<td>(\epsilon=4)</td>
</tr>
<tr>
<td>R110(_K)</td>
<td>92.3</td>
<td>88.3</td>
<td>86.0</td>
<td>59.4</td>
<td>25</td>
</tr>
<tr>
<td>R110(_P) (Ours)</td>
<td>92.3</td>
<td>86.0</td>
<td>81.6</td>
<td>64.1</td>
<td>32</td>
</tr>
<tr>
<td>R110(_E)</td>
<td>92.3</td>
<td>86.3</td>
<td>84.1</td>
<td>63.5</td>
<td>24</td>
</tr>
<tr>
<td>R110(_K_C) (Ours)</td>
<td>92.3</td>
<td>86.2</td>
<td>82.6</td>
<td>69.3</td>
<td>20</td>
</tr>
<tr>
<td>R110(_P,E) (Ours)</td>
<td>91.3</td>
<td>84.0</td>
<td>77.6</td>
<td>66.8</td>
<td>38</td>
</tr>
<tr>
<td>R110(_P,C) (Ours)</td>
<td>91.5</td>
<td>85.7</td>
<td>82.4</td>
<td>73.5</td>
<td>27</td>
</tr>
</tbody>
</table>

Evaluate against the worst attack
Plot accuracy vs distortion

(a) MNIST, $\ell_\infty$ norm
Verify enough iterations of gradient descent

<table>
<thead>
<tr>
<th>MaxIter</th>
<th>Model1</th>
<th>Model2</th>
<th>Model3</th>
<th>Model4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>99.1%</td>
<td>98.5%</td>
<td>98.7%</td>
<td>98.2%</td>
</tr>
<tr>
<td>100</td>
<td>70.2%</td>
<td>91.7%</td>
<td>77.6%</td>
<td>75.6%</td>
</tr>
<tr>
<td>1000</td>
<td>0.05%</td>
<td>51.5%</td>
<td>20.3%</td>
<td>24.4%</td>
</tr>
<tr>
<td>10K</td>
<td>0%</td>
<td>16.0%</td>
<td>20.1%</td>
<td>24.4%</td>
</tr>
<tr>
<td>100K</td>
<td>0%</td>
<td>9.8%</td>
<td>20.1%</td>
<td>24.4%</td>
</tr>
<tr>
<td>1M</td>
<td>0%</td>
<td>7.6%</td>
<td>20.1%</td>
<td>24.4%</td>
</tr>
</tbody>
</table>
By using a gradient-free method, we are able to attack the end-to-end model, despite the lack of an analytic gradient.

Try gradient-free attack algorithms
The hardest part of a defense is the evaluation
Thank You

Please do reach out to us if you have any evaluation questions

Anish: aathalye@mit.edu
Me: nicholas@carlini.com