On Adaptive Attacks to Adversarial Example Defenses

Presentation by Blake Wyatt and Zach Schickler

Paper Details

Authors:

- Florian Tramèr
- Nicholas Carlini
- Wieland Brendel
- Aleksander Madry

Published: NeurIPS 2020

Citations: 93

Short Version: https://papers.nips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf

Long Version: https://arxiv.org/pdf/2002.08347.pdf

Outline

- Abstract
- Introduction
- Expectation over Transformation (EOT)
- 3 Attacks
  - Defense outline
  - Attack outline
- Conclusion
- For and Against

Abstract

- Current adversarial defenses overestimate their robustness
- Automated and general adaptive attacks do not work
- Each attack must be *specifically designed for the defense it attacks*

Thus, this paper:

- Gives 13 new strong adaptive attacks for 13 published adversarial defenses
- Explains each defense and attack *simply* and gives their *methodology* in arriving at the attack

Introduction

- The attacks covered are:
  - k-Winners Take All
  - The Odds are Odd
  - Mixup Inference
- And Expectation over Transformation (EOT)

Slide structure for each attack:

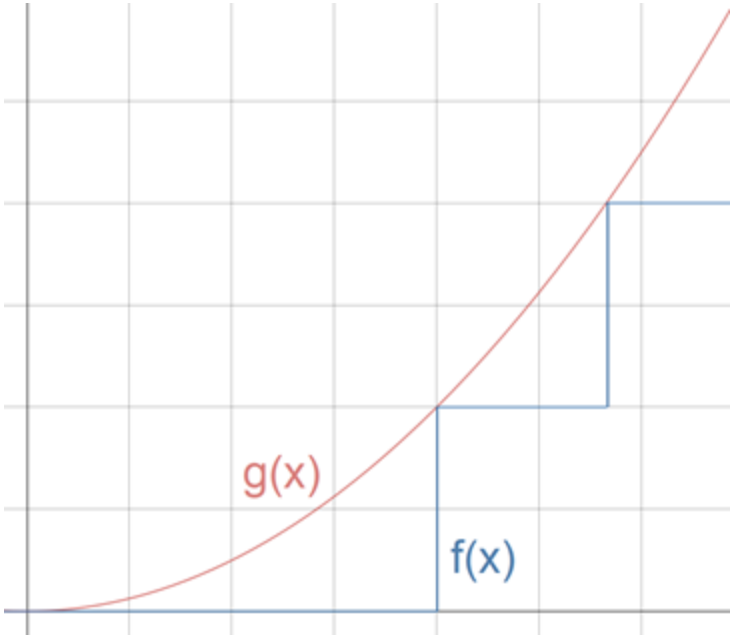
1. Defense
2. Attack

Expectation over Transformation (EOT)

Use case: computing gradients of models with **randomized** components

We want...

$$\nabla_x f(x)$$



We need g(x)

$$g(x) \approx f(x)$$

Expectation over Transformation (EOT)

$f_r(x)$

A random classifier will produce a random prediction on input x.

$f_{r_i}(x)$

One instance of a random prediction. In other words, one “run” of classifier.

Expectation over Transformation (EOT)

$$\nabla_x f_{r_i}(x)$$

Gradient for a randomly chosen prediction of input x

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_{r_i}(x)$$

Averaging all of the gradients produced by random predictions of input x

Expectation over Transformation (EOT)

$$\nabla_x \mathbb{E}_r[f_r(x)] = \mathbb{E}_r[\nabla_x f_r(x)] \approx \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_{r_i}(x)$$

Output is the gradient we want to use instead of the randomized gradient from the model

k-Winners Take All: Defense

$$\phi_k(\mathbf{y})_j = \begin{cases} y_j, & y_j \in \{k \text{ largest elements of } \mathbf{y}\}, \\ 0, & \text{else.} \end{cases}$$

ReLU activation function is replaced with this new activation function (k-WTA function)

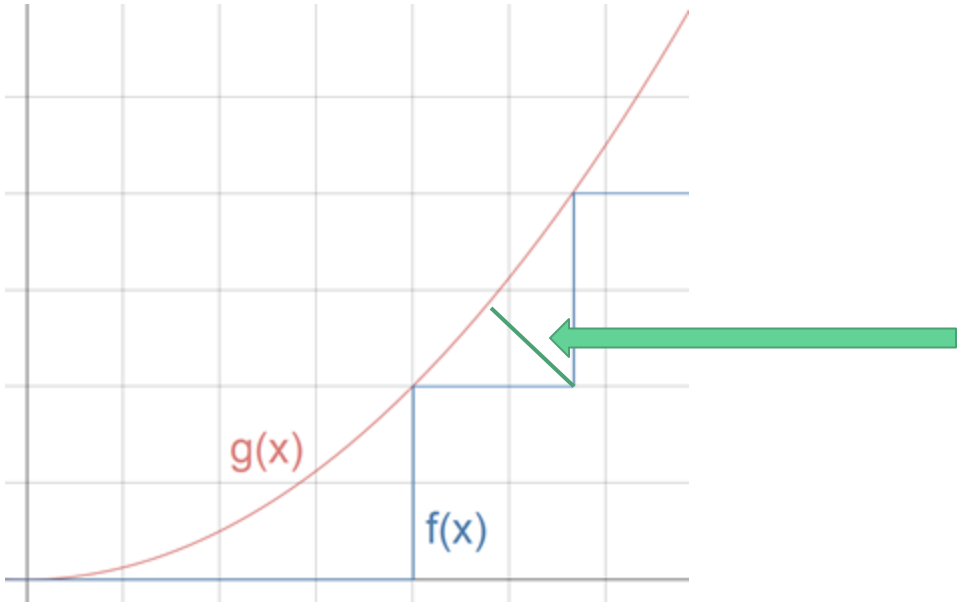
k-Winners Take All: Defense

Effect of using k-WTA function: small changes in the input will lead to large jumps in the predictions

Chaotic partitioning makes it hard to create minimal adversarial examples

k-Winners Take All: Attack

Solution is to estimate a smoothed version of the gradient



Estimating the local gradient through running multiple predictions of the gradient with small, normally distributed changes to the input

k-Winners Take All: Attack

How do we estimate the gradient?

$$g(x) = \frac{2}{\sigma M} \sum_{j=0}^{M/2} [\nabla_x L_{CE}(f(x + \delta), y) + \nabla_x L_{CE}(f(x - \delta), y)]$$

Estimating average local gradient **δ**

At point **x**

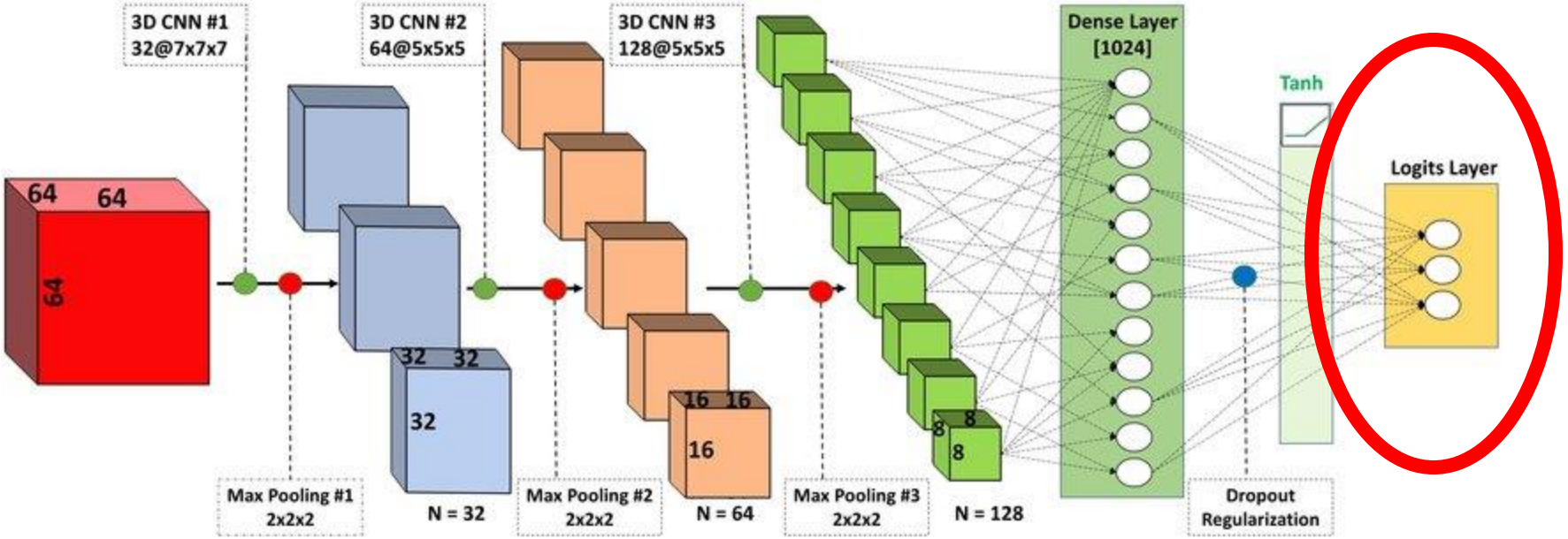
With **M** random perturbations

Drawn from standard **normal distribution** $\text{where } \delta \sim \mathcal{N}(\mu = 0, \sigma^2)$

Key Takeaway: adding k-WTA activations actually *decreases* the robustness

The Odds are Odd: Defense

Compares the logits of the classification layer to check for an adversarial example



The Odds are Odd: Defense

The difference between the logit values of classes *i* and *y*. Where *y* is the predicted class and *i* is any other class.

$$\Delta_{y,i}(x) = z(x)_i - z(x)_y$$

The Odds are Odd: Defense

Multinomial Gaussian noise is applied to the input $z(x+\delta)$

Clean Examples: $z(x) \approx z(x+\delta)$

Adversarial Examples: $z(x) \not\approx z(x+\delta)$

The Odds are Odd: Defense

Equation (1) determines if the input is likely adversarial or not based on a threshold

(1) $$\bar{\Delta}_{y,i}(x) > \tau_{y,i}$$

Equation (2) Finds the difference between the *differences* of logits for classes *i* & *y*

(2) $$\bar{\Delta}_{y,i}(x) = \mathbb{E}_{\delta \sim \mathcal{N}} \left[\Delta_{y,i}(x + \delta) - \Delta_{y,i}(x) \right]$$

The Odds are Odd: Attack

Two key parts to the defense:

- s_1: Large if the model often switches its prediction
- s_2: Large if the model is overly confident in its prediction

$s_1 + s_2$ must be small in order to circumvent!

$$
\begin{aligned}
\bar{\Delta}_{y,i}(x) &= \mathbb{E}_{\delta \sim \mathcal{N}} \left[\Delta_{y,i}(x+\delta) - \Delta_{y,i}(x) \right] \\
&= \mathbb{E}_{\delta \sim \mathcal{N}} \left[(z(x+\delta)_i - z(x+\delta)_y) - (z(x)_i - z(x)_y) \right] \\
&= \underbrace{\mathbb{E}_{\delta \sim \mathcal{N}} \left[(z(x+\delta)_i - z(x+\delta)_y) \right]}_{s_1} + \underbrace{(z(x)_y - z(x)_i)}_{s_2}
\end{aligned}
$$

The Odds are Odd: Attack

- PGD + EOT: Reduces s_1, however, it's not enough

Note that to circumvent:

- A model's robustness to noise must be similar to or larger than a typical input
- A model's prediction confidence must be similar to a typical input

Thus we minimize,

$$||z(x') - z(x_t)||_2^2$$

And combine with EOT

Mixup Inference: Defense

Mixup Inference computes logits with input *x* and randomly selected clean samples s_k

- Lambda determines what proportion of the input to *z* is samples s_k

$$\hat{z}(x) = \frac{1}{k}\sum_{k=1}^{K} z(\tilde{x}_k) = \frac{1}{k}\sum_{k=1}^{K} z(\lambda x + (1-\lambda)s_k)$$

Assumption: adversarial examples are fine tuned to a class and do not hold up when slightly modified

Mixup Inference slightly modifies them

Mixup Inference: Attack

Generate adversarial examples such that they are less prone to error upon modification

- Optimize to less confident, but more generalized perturbations with PGD
- Average the gradient in each step of the attack
- The gradient must be *stable*

Combining with EOT could potentially be used to further improve the attack

Conclusion

- Every defense was circumvented with strong adaptive attacks

- Attacks must be *simple* and designed *specifically* for the defense

“For any proposed attack, it is possible to build a non-robust defense that prevents the attack”

- Future defense research should create stronger adaptive attack evaluations

For the paper

- Presents the paper as a learning experience
- Contests modern defense research evaluation methods
- Introduces T1-T6 attack categorization
- Successfully breaks many modern defenses

Against the paper

- The definition of “strong” in strong adaptive attacks is not explicitly defined
- Most attacks were only LInf
- The k-WTA attack hurts the model’s robustness
- The attack computational costs are not given

Thank you