On Adaptive Attacks to Adversarial Example Defenses

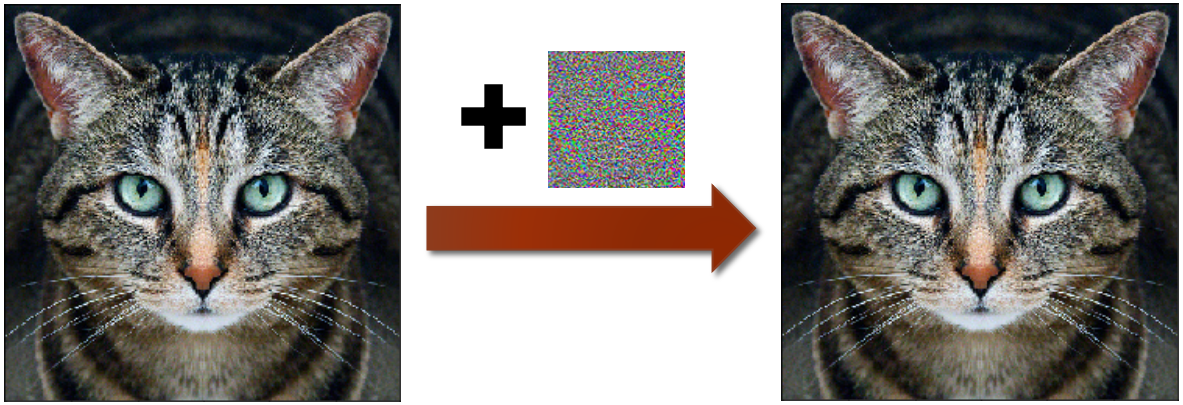
Florian Tramèr

USENIX ScAINet

August 10th, 2020

Joint work with Nicholas Carlini, Wieland Brendel and Aleksander Madry

What Are Adversarial Examples?



88% Tabby Cat

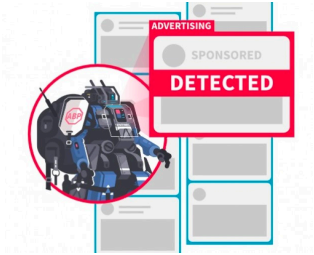
99% Guacamole

Biggio et al., 2014
Szegedy et al., 2014
Goodfellow et al., 2015

Why Should We Care?

ML in security-critical applications

Malware detection

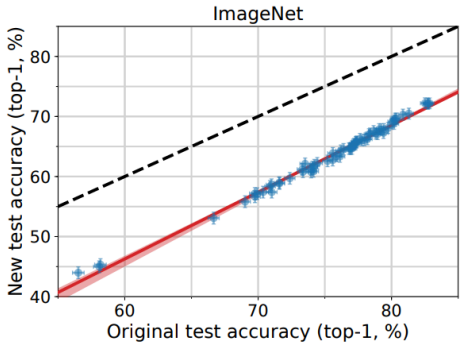


Ad-blocking

Anti phishing

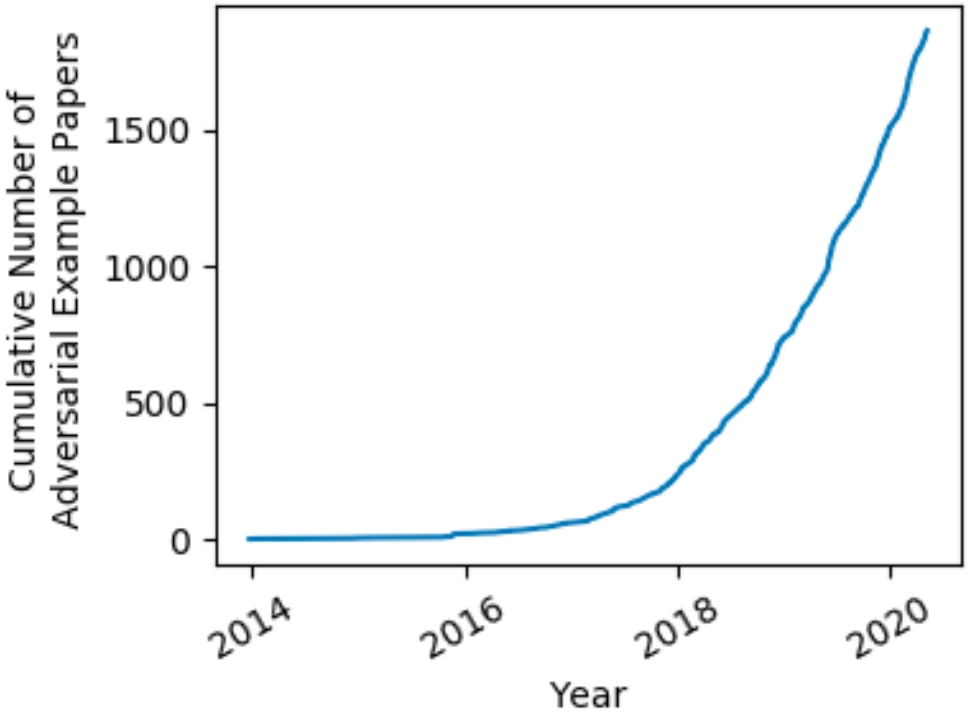
Content takedown

Understanding robustness under (standard) distribution shift



Recht et al., 2019

Many Defenses Have Been Proposed...



https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

...But Evaluating Them Properly Is Hard

We re-evaluated 13 defenses presented at

[ICLR | ICML | NeurIPS] [2018 | 2019 | 2020]

All defenses claim to follow the best evaluation standards

Yet, we circumvent all of them

⇒ reduce accuracy to baseline (usually 0%) in the considered threat model

Isn’t This Old News?

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Nicholas Carlini David Wagner
University of California, Berkeley

Broke 10 (mainly unpublished) defenses in 2017

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

Anish Athalye [*1] Nicholas Carlini [*2] David Wagner [2]

Broke 7 defenses published at ICLR 2018

Why We Hoped Things Might Have Changed

Consensus on what constitutes a good evaluation

Clearly defined threat model

1. *White-box:* adversary has access to defense parameters
2. *Small perturbations*: find x' s.t. x' *misclassified* and $\| x - x' \|_p \leq \varepsilon$

Incomplete definition
Easy to formalize
Surprisingly hard

Adaptive

Adversary tailors the attack to the defense

Carlini & Wagner, 2017,
Athalye et al., 2018,
Carlini et al. 2019,
...

Evaluation Standards Seem To Be Improving

Authors (and reviewers) are aware of the importance of adaptive attacks in evaluations

Then Why Are Defenses Still Broken?

Many defenses are not evaluated against a strong adaptive attack

Our Work

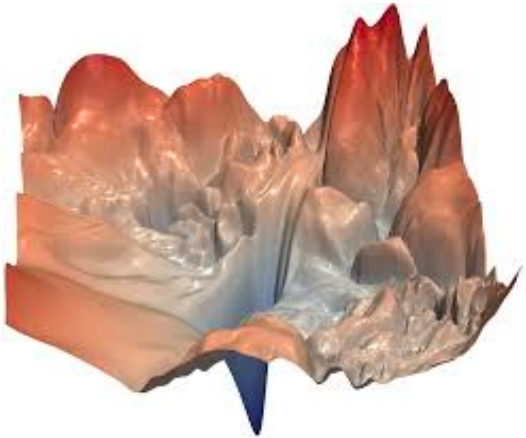
13 case studies on how to design strong(er) adaptive attacks

Including:

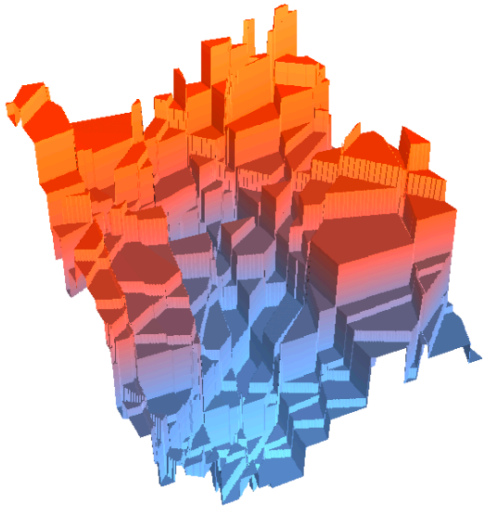
- Our hypotheses when reading each defense's paper/code
- Things we tried but that didn't work
- Some things we didn't try but might also have worked

How (not) to build & evaluate defenses

Don’t Intentionally Obfuscate Gradients



If this wasn’t enough…



this won’t be either

Breaking specific attack techniques is not the way forward

Don’t Blindly Re-use Prior (Adaptive) Attacks

Adaptive attack strategies are not universal!

Most popular “victims”: BPDA & EOT (Athalye et al. 2018)

For our experiments, Expectation Over Transformation is used for the adaptive attack scenario.

including the strongest attacks such as BPDA

backward pass is not differentiable, which makes BPDA the strongest white-box attack.

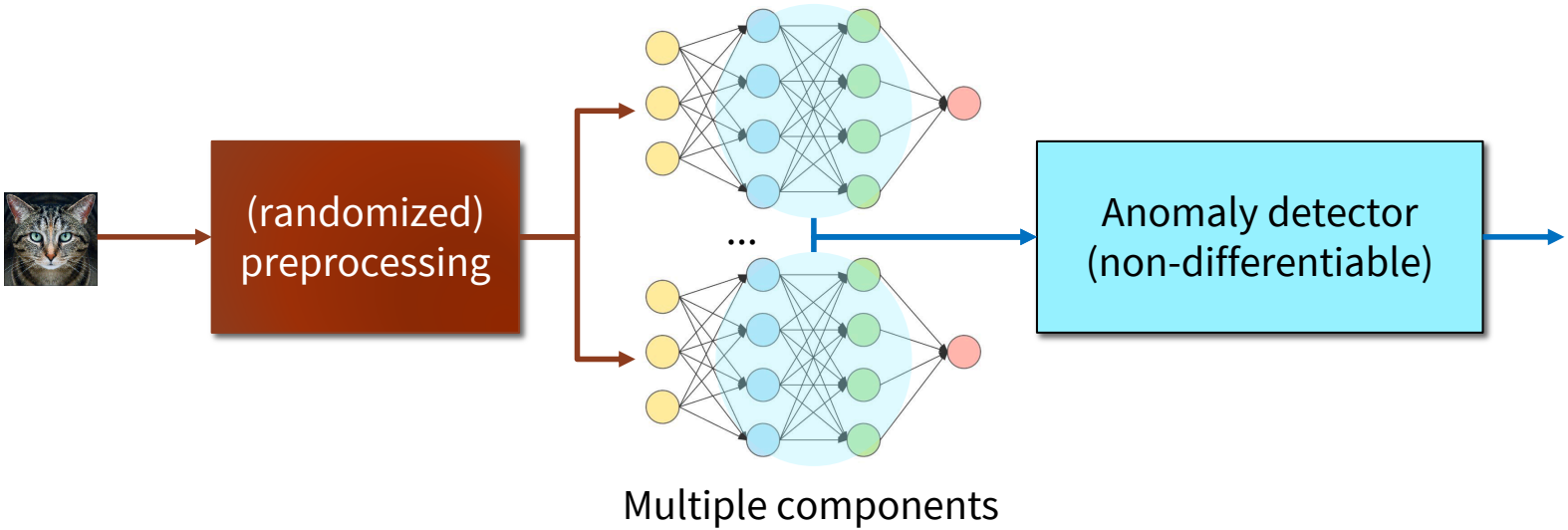
The optimality of this strategy in the face of randomization-based defenses

- Understand **why** an attack worked on other defenses before re-using it
- Use BPDA as a last resort (try gradient-free / decision-based attacks first)
- Before using EOT, build an attack that works for **fixed randomness**

Don’t Complicate The Attack

Many proposed defenses are **complicated**
(for some reasons, this is particularly true for AdvML papers in security conferences)

This is OK! Maybe the best defense has to be complex



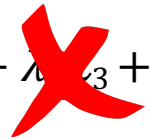
Don’t Complicate The Attack

Many proposed defenses are **complicated**
(for some reasons, this is particularly true for AdvML papers in security conferences)

This is OK! Maybe the best defense has to be complex

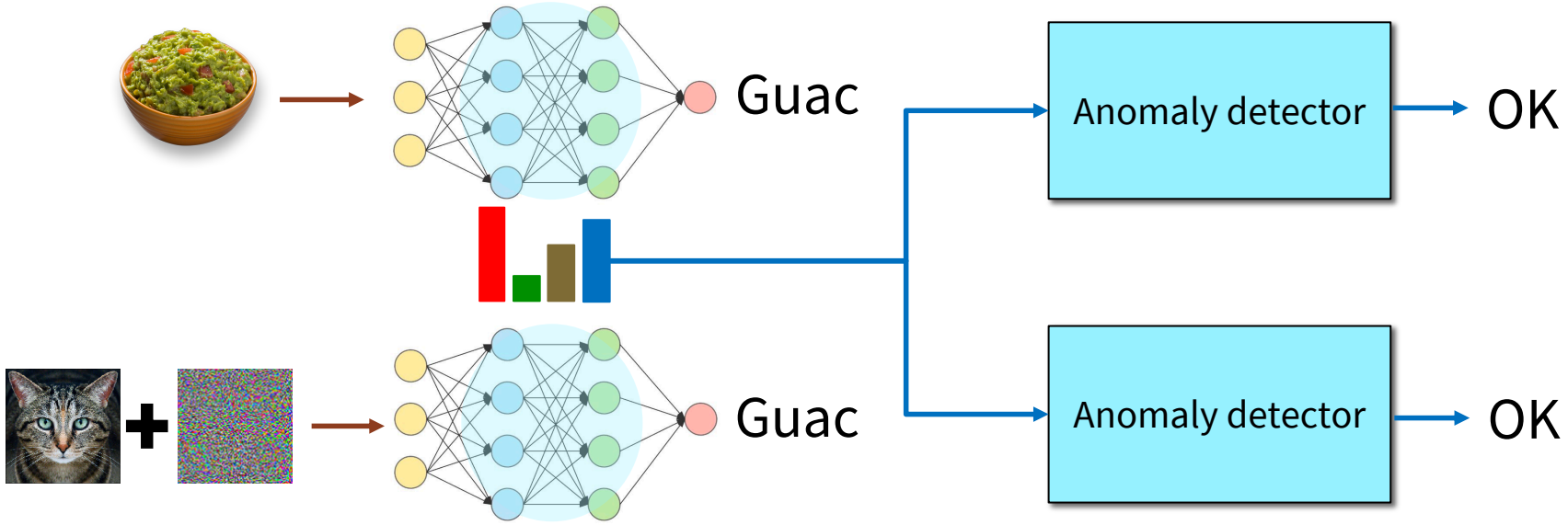
But attacks don’t have to be!

- Optimizing over complex defenses can be hard ($\mathcal{L} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 + \ldots$)
- Evaluate each component individually, there is often a weak link
- Combining broken components rarely works



Don’t Complicate The Attack

Use **feature adversaries** (Sabour et al. 2015) to break multiple components at once



Don’t Convince Reviewers, Convince Yourself!

Really try to break your defense (others probably will...)

- An evaluation against 10 *non-adaptive* attacks isn’t broad
- If offered $1M to break your defense, would you use a non-adaptive attack?
- What assumptions/invariants does the defense rely on? **Attack those!**

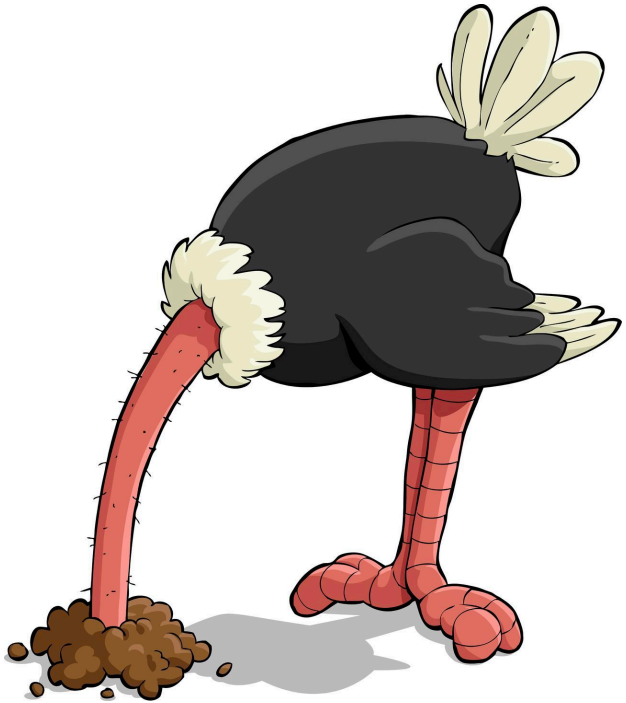
Evaluation guidelines are great, but:

- Not just a check-list to appease reviewers
- They also apply to adaptive attacks
  (e.g., adaptive attacks should never perform **worse** than non-adaptive ones)

ON EVALUATING ADVERSARIAL ROBUSTNESS

Nicholas Carlini[1], Anish Athalye[2], Nicolas Papernot[1], Wieland Brendel[3], Jonas Rauber[3], Dimitris Tsipras[2], Ian Goodfellow[1], Aleksander Mądry[2], Alexey Kurakin[1]*

My Defense Got Broken. Now What?



My Defense Got Broken. Now What?

~40 white-box defenses that were publicly broken (that I know of)

- **one** paper was retracted before publication
- **one** paper was amended on arXiv

> [2]Recent work [8], however, has shown that our approach is vulnerable

We should do better!

- Hard to navigate the field for newcomers
- Many ideas get re-used despite being broken

My Defense Got Broken. Now What?

Personal experience: ENSEMBLE ADVERSARIAL TRAINING: ATTACKS AND DEFENSES

- Often referenced as an effective defense against black-box attacks
- Later work developed much stronger transfer attacks ☹

⇒ **Please contact authors when you find an attack!**

1.1 SUBSEQUENT WORK (ADDED APRIL 2020)

After intro, or in abstract, results, etc.

Starting with the NIPS 2017 competition on Defenses Against Adversarial Attacks, many subsequent works have proposed more elaborate black-box transfer-based attacks. By incorporating addi-

Conclusion

Evaluating adversarial examples defenses is hard!

How do we improve things?

Resisting attacks that broke prior defenses ≠ progress

> *For any proposed attack, it is possible to build a non-robust defense that prevents the proposed attack.*

Ideal: **defense evaluation = 99% adaptive attacks**

- Try breaking other defenses before attacking your own
- Strive for simple attacks (and defenses if possible)
- We need more independent re-evaluations
- If a defense is broken, acknowledge the attack, amend the paper, and keep going!

tramer@cs.stanford.edu https://arxiv.org/abs/2002.08347