On Adaptive Attacks to Adversarial Example Defenses

NeurIPS 2020

Florian Tramèr*

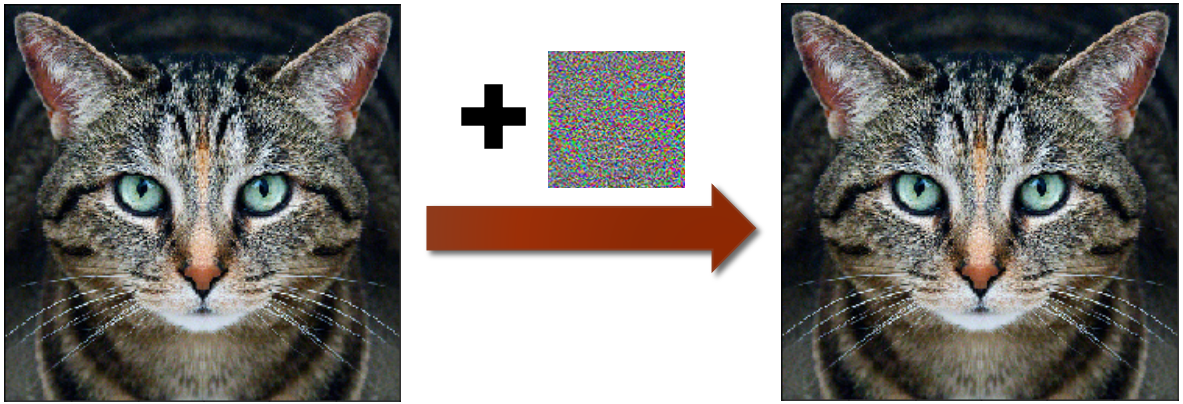
Nicholas Carlini*

Wieland Brendel*

Aleksander Mądry

*equal contribution

What Are Adversarial Examples?

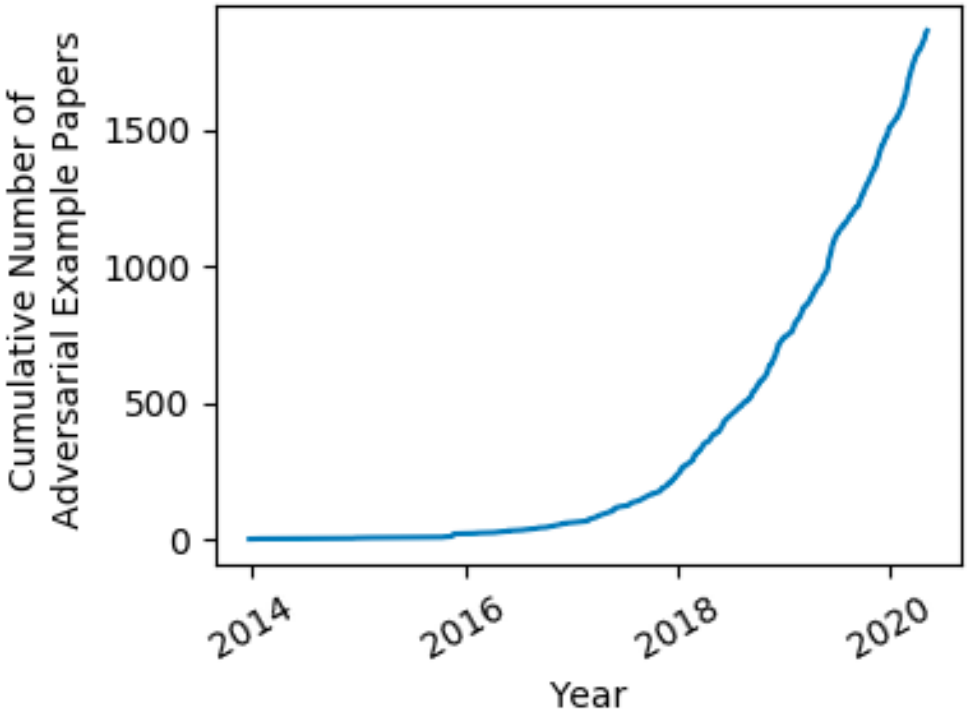


88% Tabby Cat

99% Guacamole

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

Many Defenses Are Proposed...



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

... But Evaluating Them Properly Is Hard

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Broke 10 (mainly unpublished) defenses in 2017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

The Good: Consensus On Strong Evaluation Standards

Clearly defined threat model

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

Adaptive Evaluation

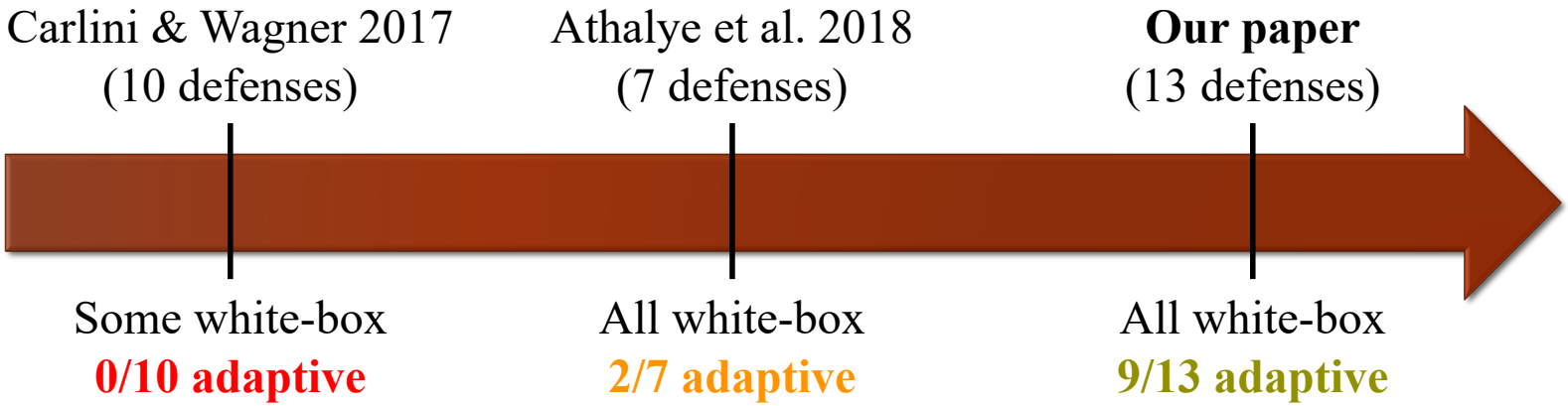
Adversary tailors the attack to the defense

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

The Good: Adoption Of Strong Evaluation Standards

We re-evaluate 13 defenses presented at:

NeurIPS'18	**ICLR'19**	**ICML'19**	**NeurIPS'19**	**ICLR'20**
(1)	(1)	(4)	(2)	(5)



The Bad: Defenses Are Still Broken

We re-evaluate 13 defenses presented at:

NeurIPS'18	**ICLR'19**	**ICML'19**	**NeurIPS'19**	**ICLR'20**
(1)	(1)	(4)	(2)	(5)

We circumvent <u>all</u> of them!

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

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

Conclusion

Evaluating adversarial examples defenses is hard!

Defenses must be evaluated against *strong adaptive* attacks

How do we design strong adaptive attacks?

1. **Practice!** Try breaking other defenses before evaluating your own
2. **Simplicity!** Simple attacks are often easier to debug, and improve
3. **Focus!** Find the defense's weakest component, and attack exactly that

https://arxiv.org/abs/2002.08347 https://github.com/wielandbrendel/adaptive_attacks_paper