Obfuscated Gradients Give a False Sense of Security:

Circumventing Defenses to Adversarial Examples

Anish Athalye*[1], **Nicholas Carlini***[2] , and David Wagner[3]

[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

adversarial perturbation

88% tabby cat

99% guacamole

Adversarial Examples

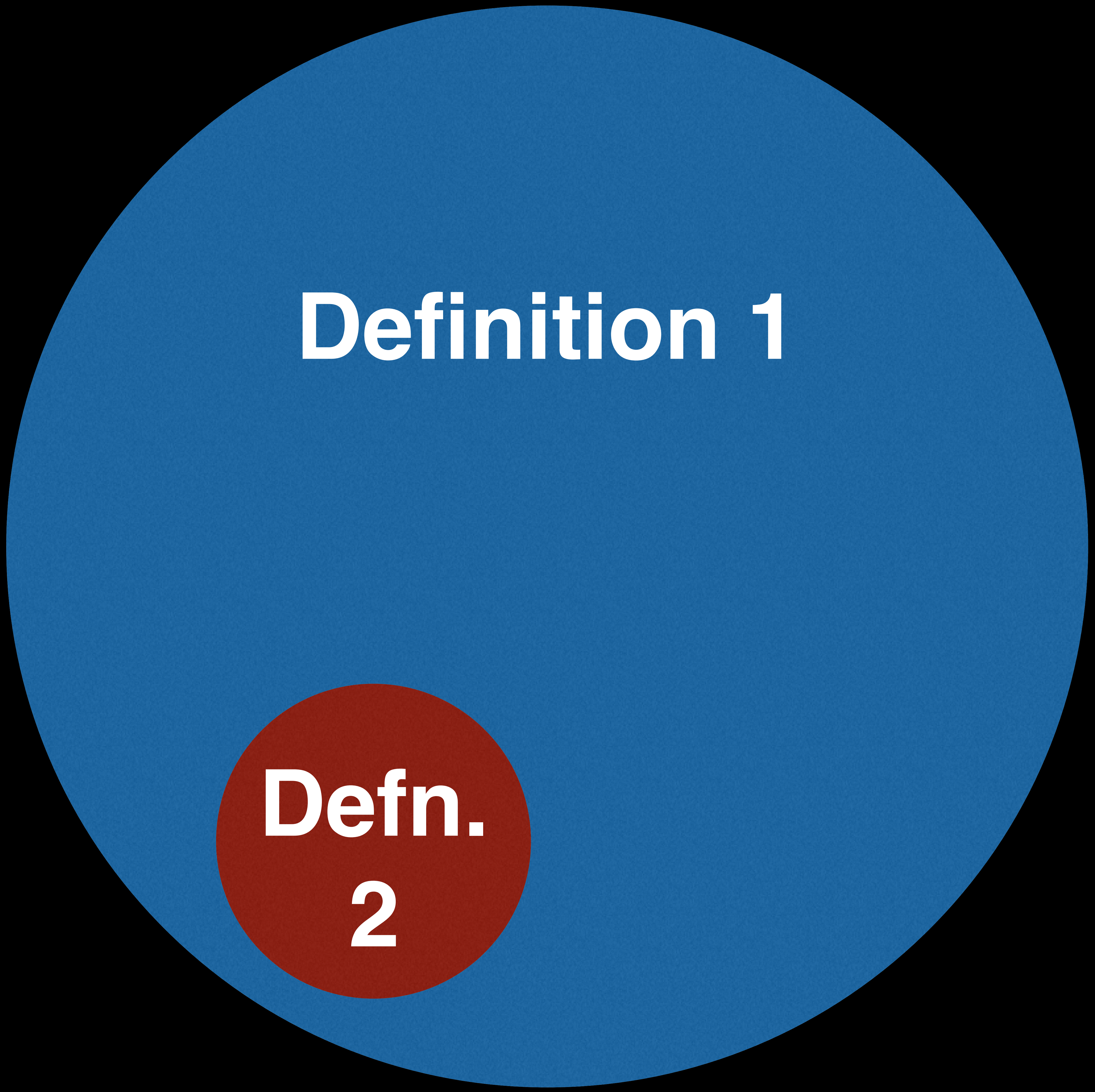
Definition 1:
Inputs specifically crafted to fool a neural network.

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

Correct definition.
Hard to formalize.

Not complete.
Easy to formalize.

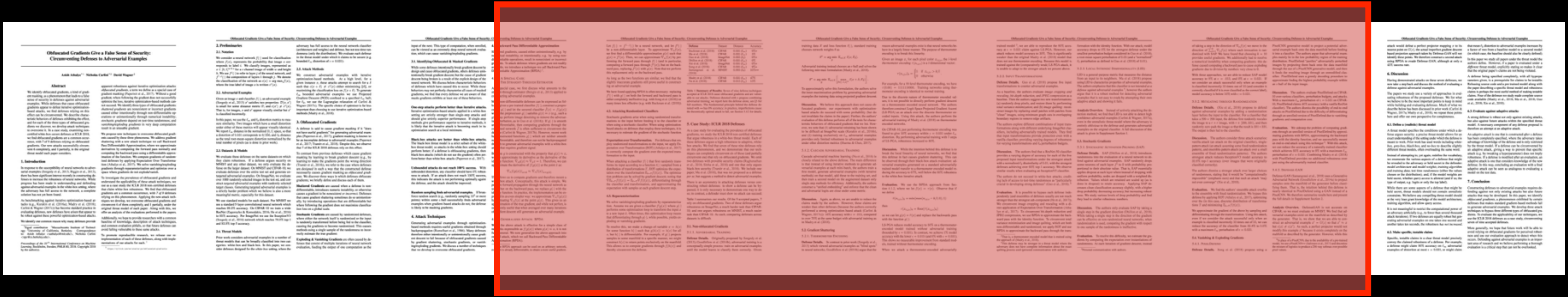
Adversarial Examples



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**

1. Threat Model

A threat model is a **formal** statement defining when a system is intended to be secure.

1. Threat Model

What dataset is considered?

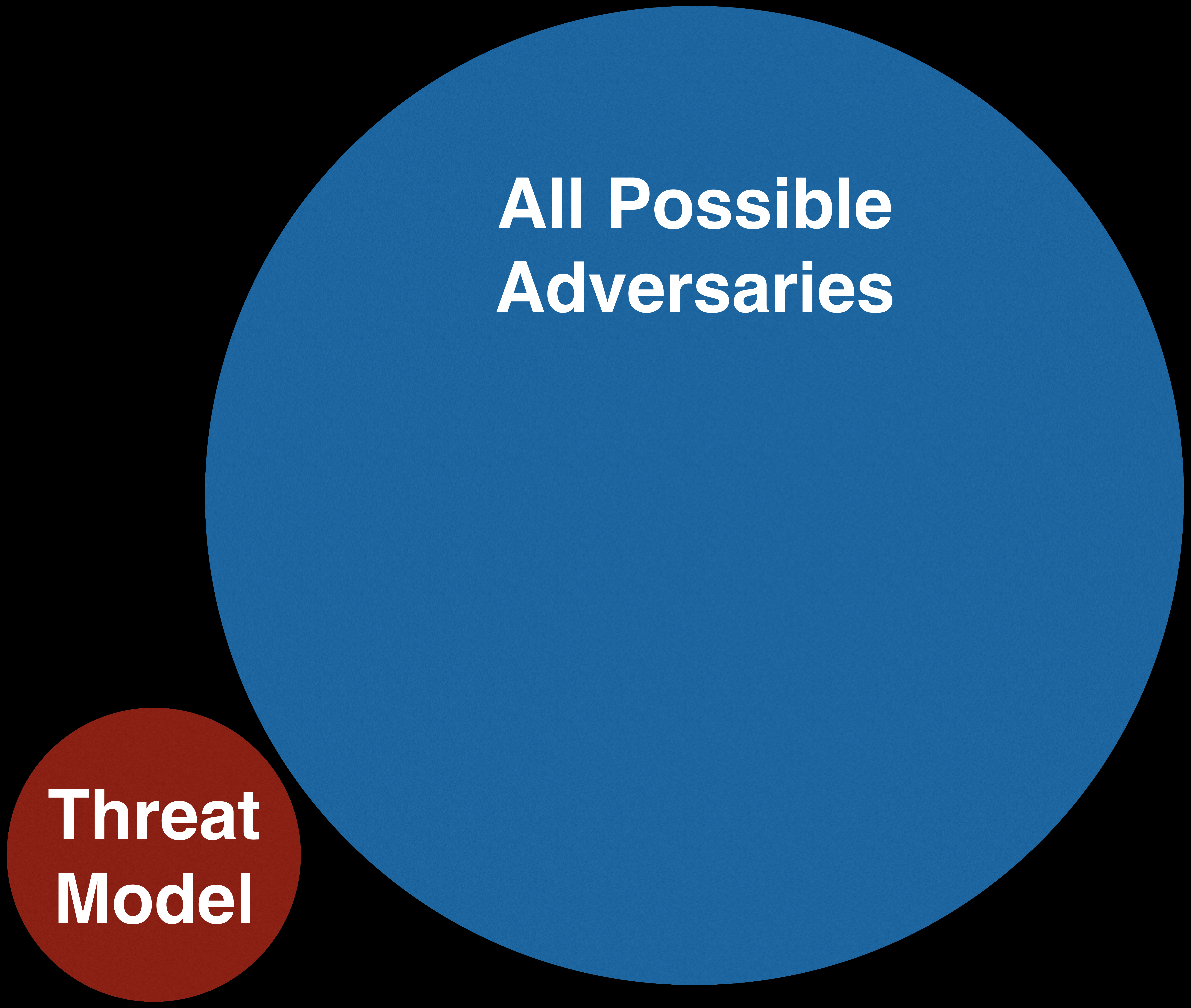
Adversarial example definition?

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?"

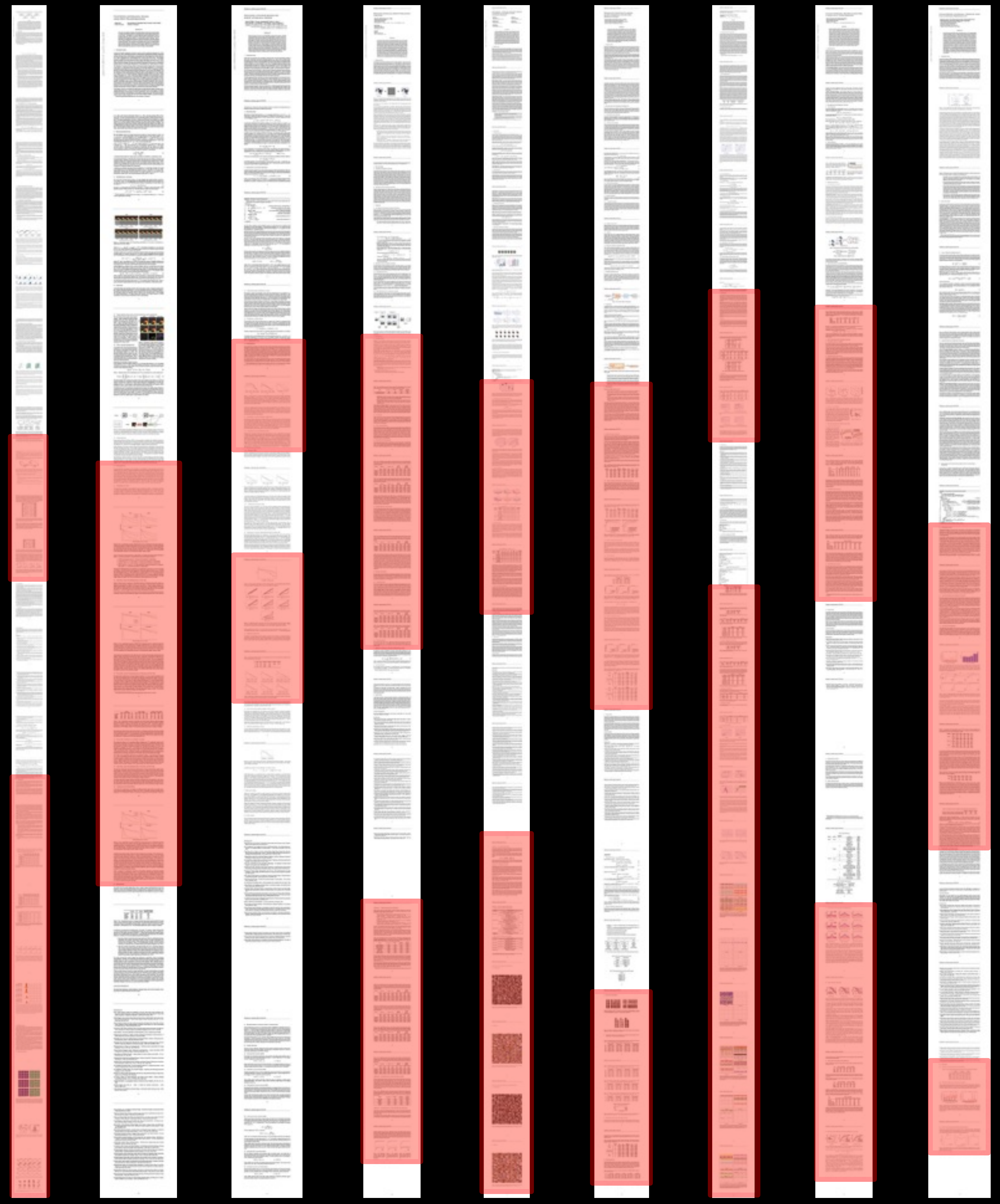
3. Defense Evaluation

```
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 ½ 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,

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 $\varepsilon = 0.031$ and 7 steps for iterative attacks;

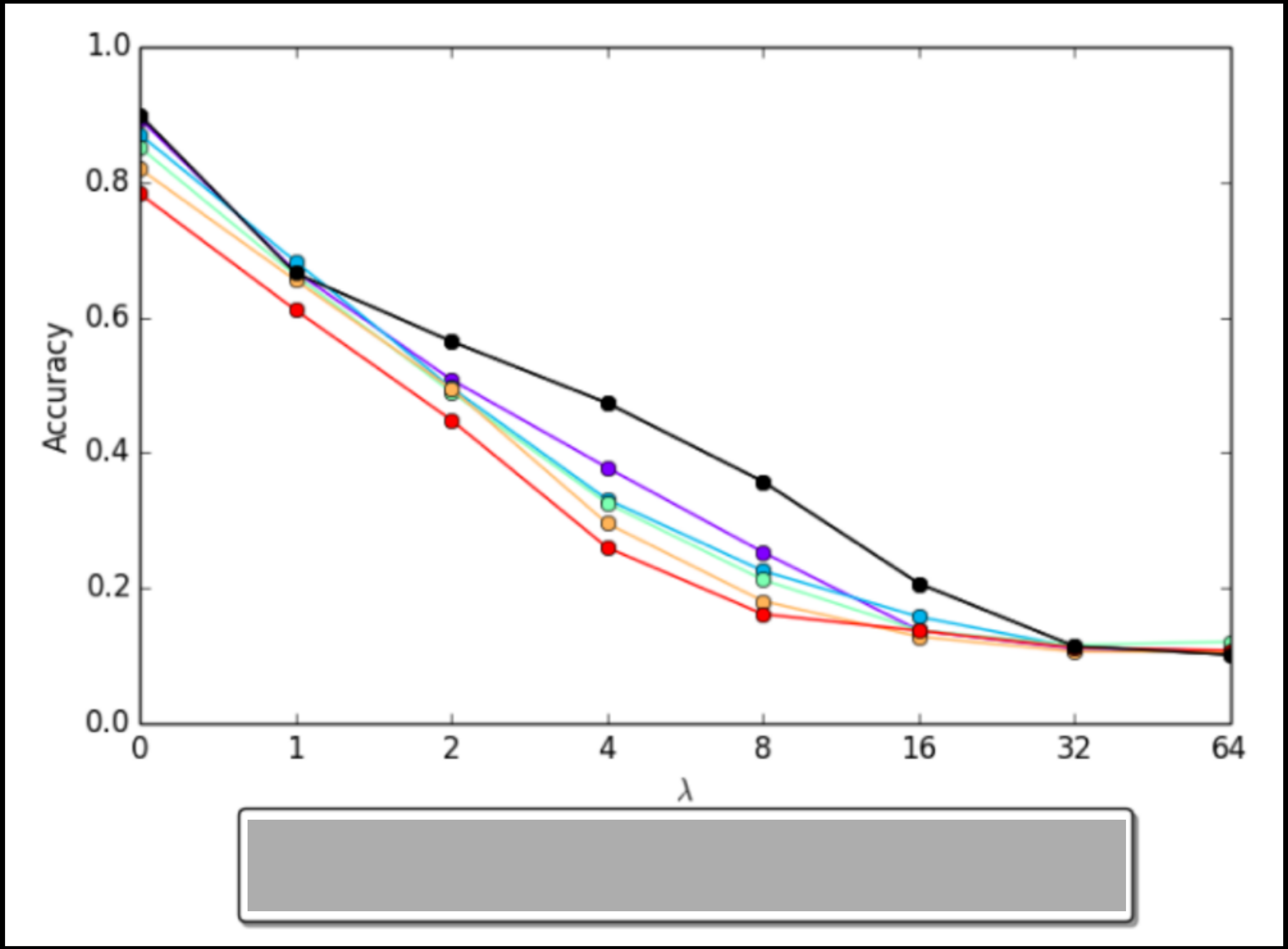
Use more than 100 (or 1000?) iteration of gradient descent

	Model	FGSM	PGD
Clean	[illegible]	25.10	4.10
	[illegible]	46.15	1.66
	[illegible]	43.89	3.57
	[illegible]	52.07	53.11
	[illegible]	48.50	50.50

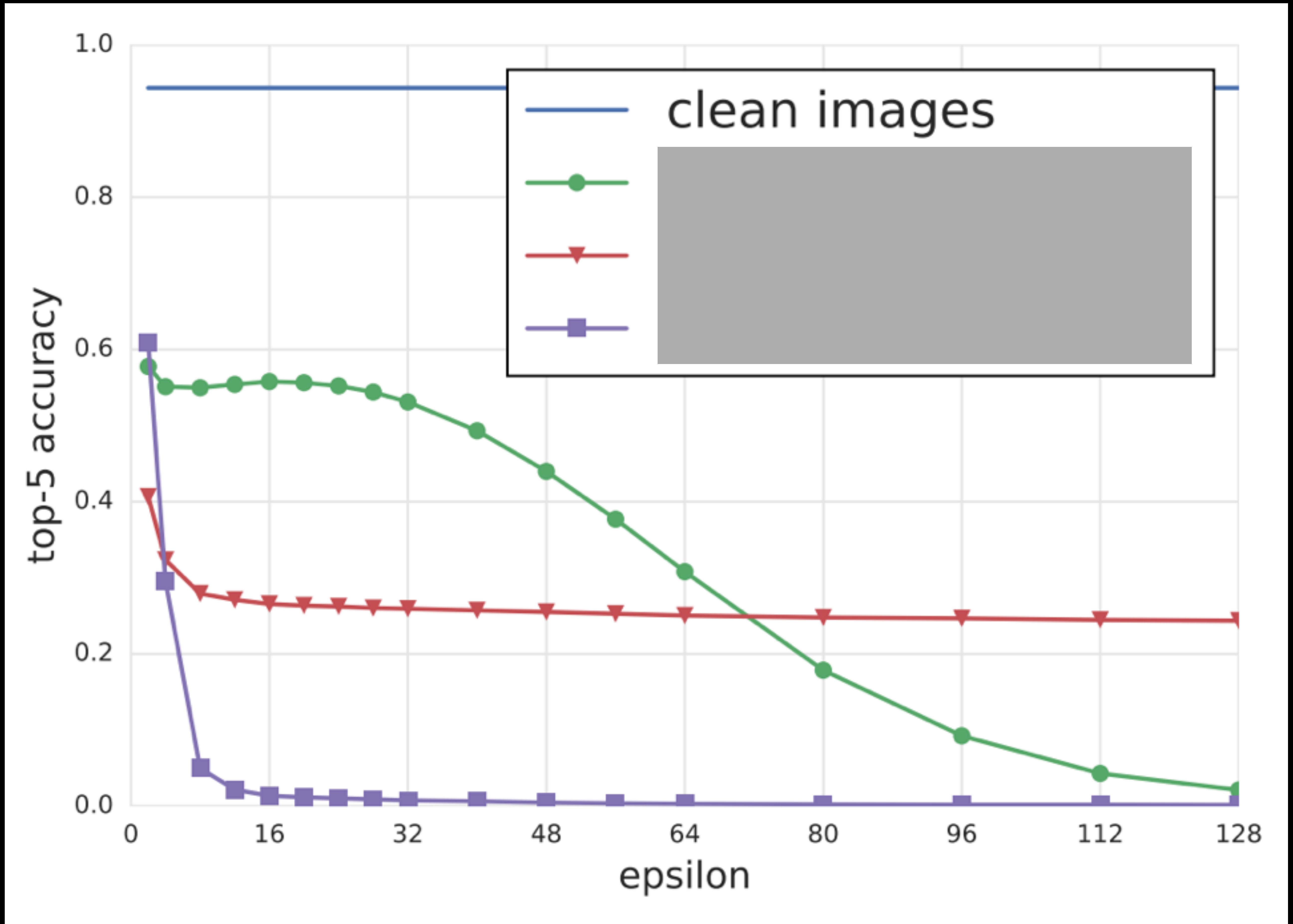
Iterative attacks should always do better than single step attacks.

Attack	Parameter	Fooling Rate	Detection Rate
DeepFool		99.35%	97.83%
Carlini	κ=0.0	100.0%	95.66%

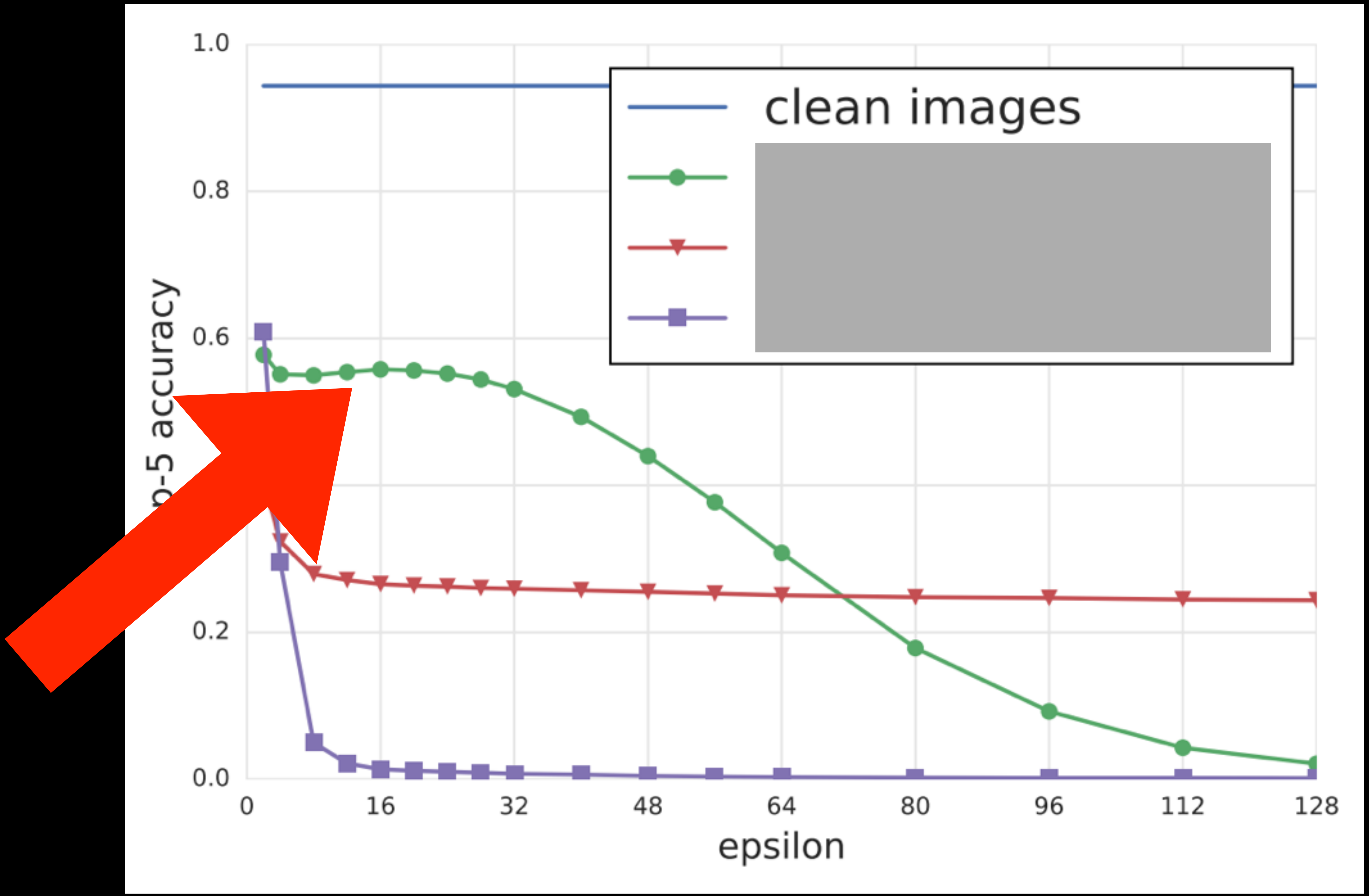
Unbounded optimization attacks should eventually reach in 0% accuracy



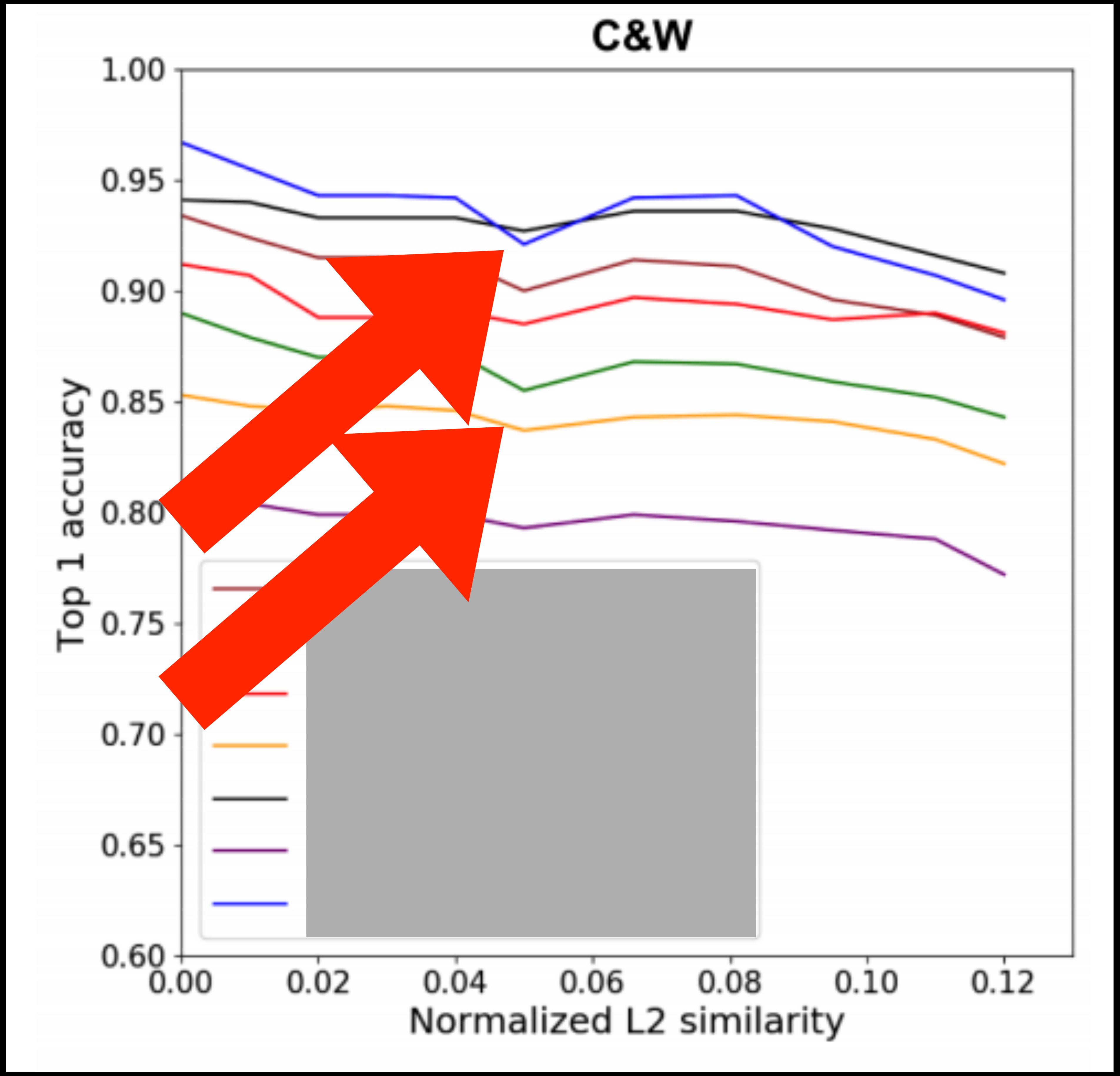
Unbounded optimization attacks should eventually reach in 0% accuracy



Unbounded optimization attacks should eventually reach in 0% accuracy



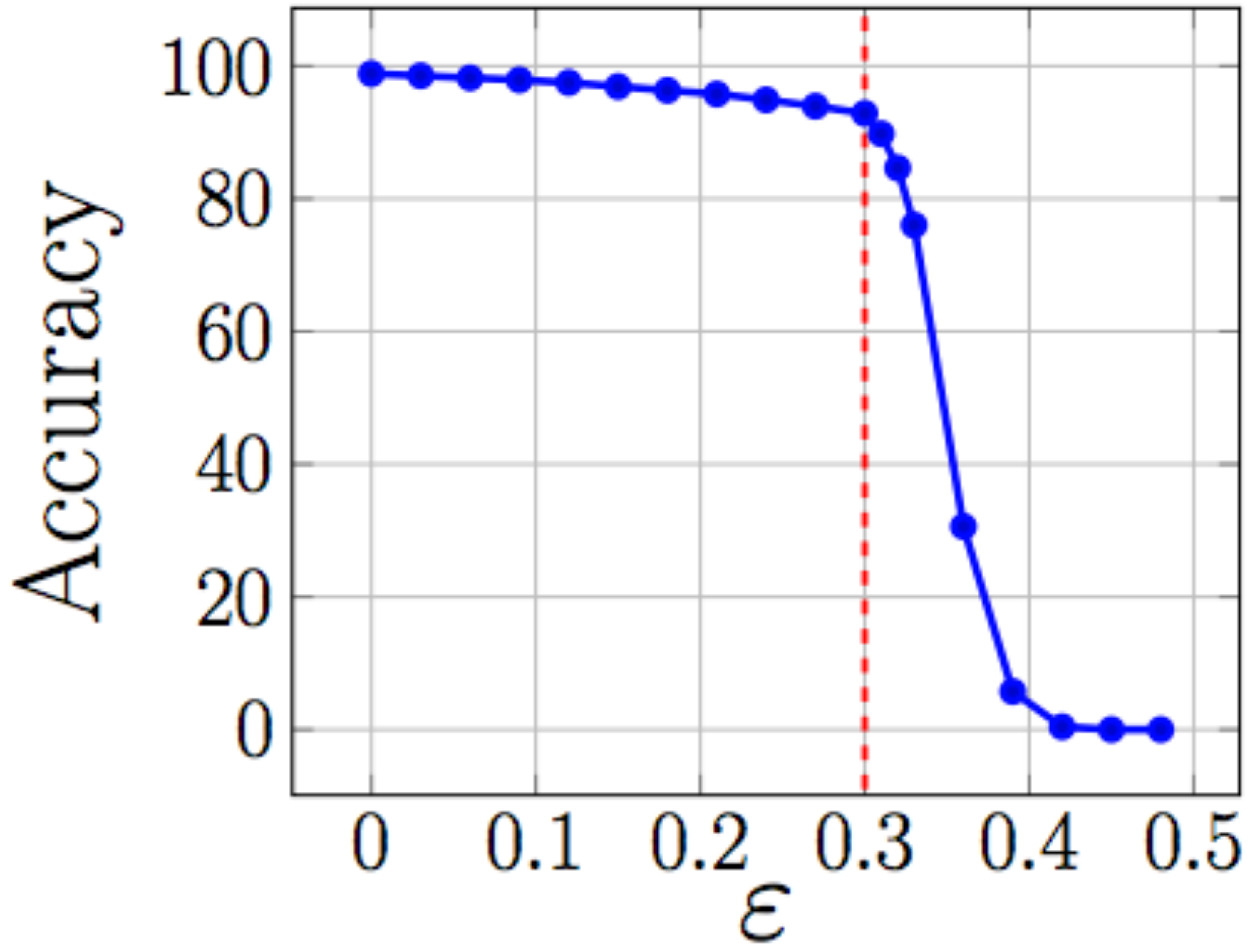
Model accuracy should be monotonically decreasing



Model accuracy should be monotonically decreasing

Model	clean	step_ll		step_FGSM		iter_FGSM		CW	
		ϵ=2	ϵ=16	ϵ=2	ϵ=16	ϵ=2	ϵ=4	ϵ=2	ϵ=4
R110$_K$	92.3	**88.3**	**90.7**	**86.0**	**95.2**	59.4	9.2	25	4
R110$_P$ (Ours)	92.3	86.0	89.4	81.6	91.6	64.1	20.9	32	7
R110$_E$	92.3	86.3	74.3	84.1	72.9	63.5	21.1	24	6
R110$_{K,C}$ (Ours)	92.3	86.2	72.8	82.6	66.7	69.3	33.4	20	5
R110$_{P,E}$ (Ours)	91.3	84.0	65.7	77.6	54.5	66.8	38.3	**38**	**16**
R110$_{P,C}$ (Ours)	91.5	85.7	76.4	82.4	69.1	**73.5**	**42.5**	27	15

Evaluate against the worst attack



(a) MNIST, ℓ_∞ norm

Plot accuracy vs distortion

MaxIter	Model1	Model2	Model3	Model4
Natural	99.1%	98.5%	98.7%	98.2%
100	70.2%	91.7%	77.6%	75.6%
1000	0.05%	51.5%	20.3%	24.4%
10K	0%	16.0%	20.1%	24.4%
100K	0%	9.8%	20.1%	24.4%
1M	0%	7.6%	20.1%	24.4%

Verify enough iterations of gradient descent

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

Conclusion

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