Adversarially (non-)Robust Machine Learning

Nicholas Carlini
Google

Better Language Models and Thei

Implications

We've trained a large-scale unsupervise
model which generates coherent paragr
text, achieves state-of-the-art performa
many language modeling benchmarks, a
performs rudimentary reading comprehe
machine translation, question answering
summarization—all without task-specifi

February 14, 2019
24 minute read

Deep Speech 2: End-to-I
English an

Baidu Research –
Dario Amodei, Rishita Anubhai, Eric Batten
Jingdong Chen, Mike Chrzanowski, Adam
Linxi Fan, Christopher Fougner, Tony Har
Libby Lin, Sharan Narang, Andrew Ng, S
Sanjeev Satheesh, David Seetapun, Shubho S
Bo Xiao, Dani Yogatan

Ab

We show that an end-to-end deep lea
either English or Mandarin Chinese sp
cause it replaces entire pipelines of han
works, end-to-end learning allows us to
ing noisy environments, accents and different languages. Key to our approach is our application of HPC techniques, resulting in a 7x speedup over our previous system [26]. Because of this efficiency, experiments that previously took weeks now run in days. This enables us to iterate more quickly to identify superior architectures and algorithms. As a result, in several cases, our system is competitive with the transcription of human workers when benchmarked on standard datasets. Finally, using a technique called Batch Dispatch with GPUs in the data center, we show that our system can be inexpensively deployed in an online setting, delivering low latency when serving users at scale.

Facebook

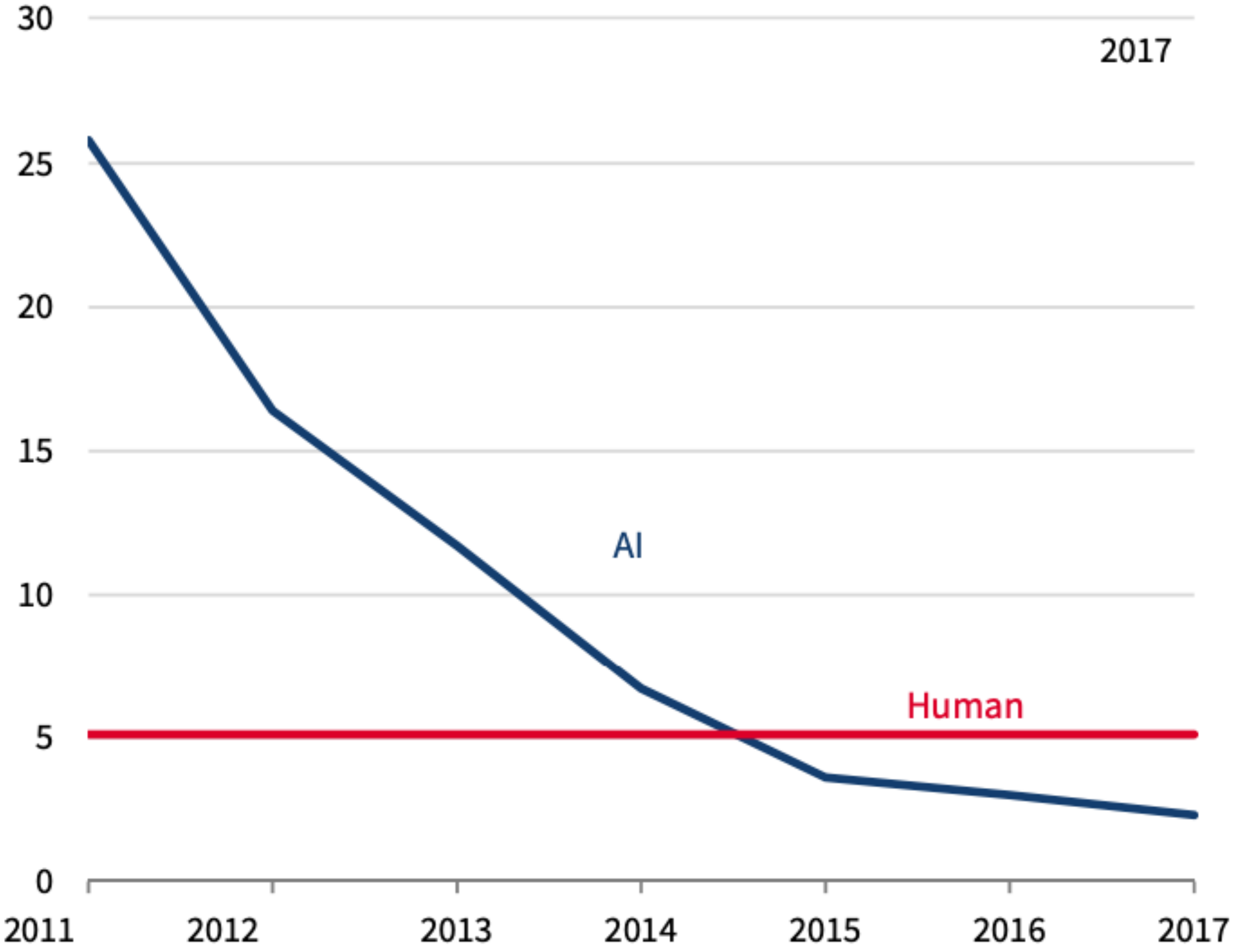
Introducing the First AI Model That Translates 100 Languages Without Relying on English

October 19, 2020
By Angela Fan, Research Assistant

This Talk:

Figure 7-1. Error Rate of Image Classification by Artificial Intelligence and Humans, 2010–17

Error rate (percent)



Sources: Russakovsky et al. (2015); CEA calculations.

…. however

88% **tabby cat**

adversarial perturbation

88% **tabby cat**

88% **tabby cat**

adversarial
perturbation

88% tabby cat

adversarial perturbation

99% guacamole

(a) (b) (c)

Eykholt et al., "Robust Physical-World Attacks on Deep Learning Models"

Andrew Walz

2020 Congressional Candidate

Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"

Andrew Walz

2020 Congressional Candidate

Verified by Twitter ✓

Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"

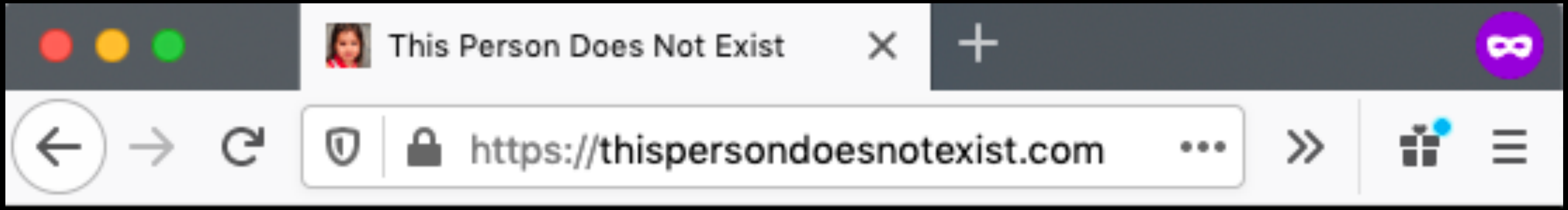
Andrew Walz

2020 Congressional Candidate

Verified by Twitter

Not a real person

Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"



Andrew Walz

2020 Congressional Candidate

Verified by Twitter

Not a real person

Andrew Walz

2020 Congressional Candidate

Verified by Twitter

Not a real person

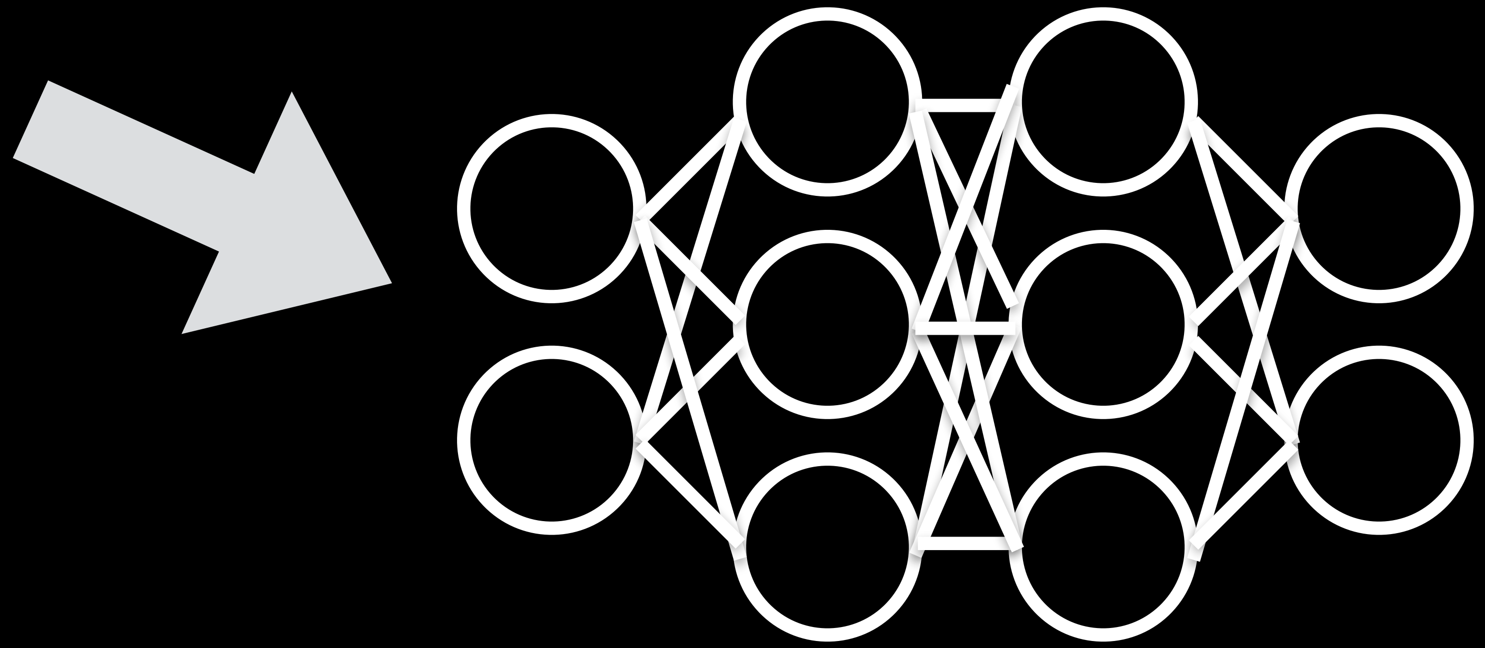
Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"

Andrew Walz

2020 Congressional Candidate

Verified by Twitter

Not a real person



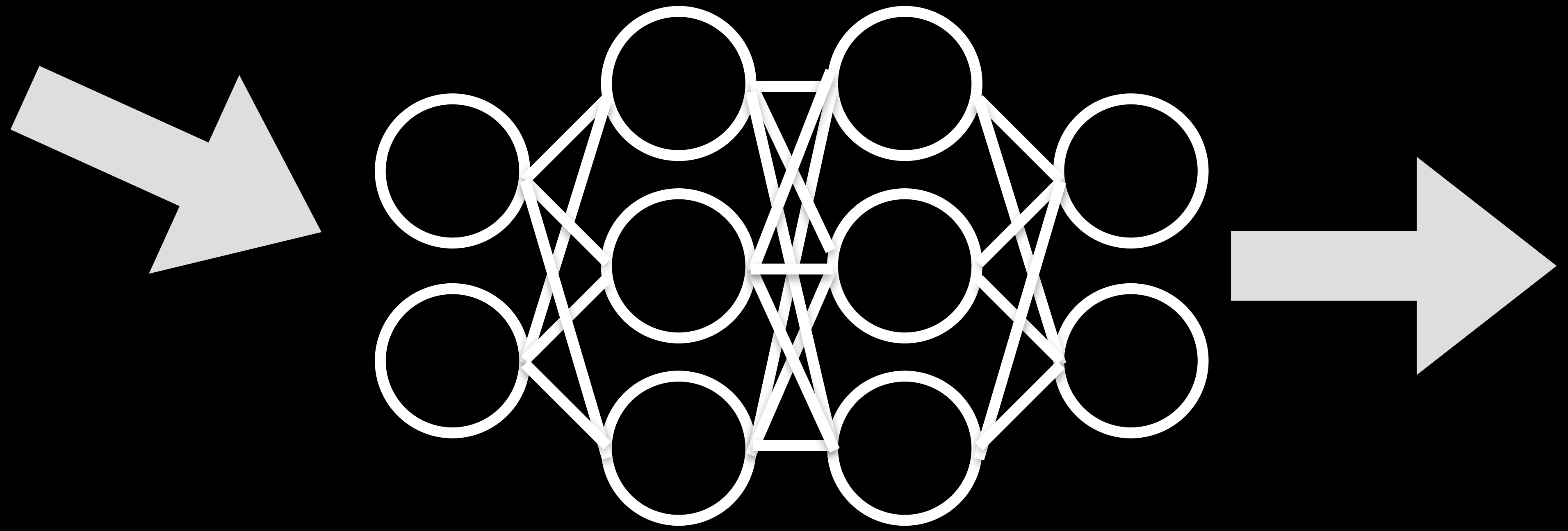
Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"

Andrew Walz

2020 Congressional Candidate

Verified by Twitter

Not a real person



FAKE

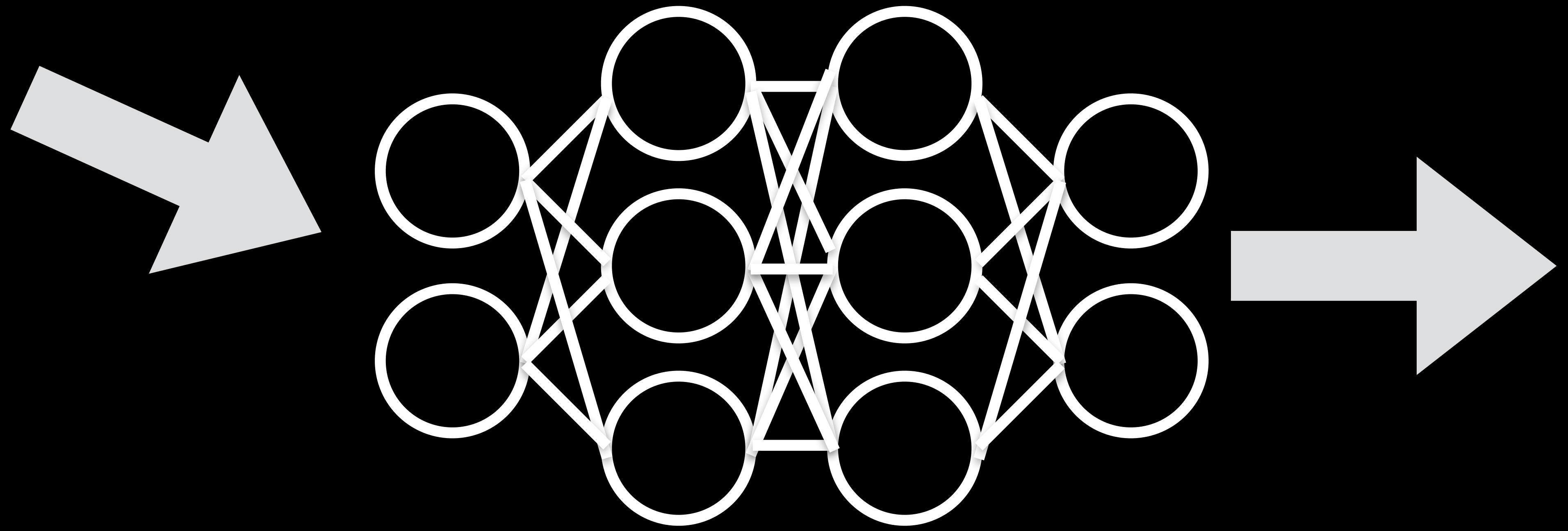
Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"

Andrew Walz

2020 Congressional Candidate

Verified by Twitter

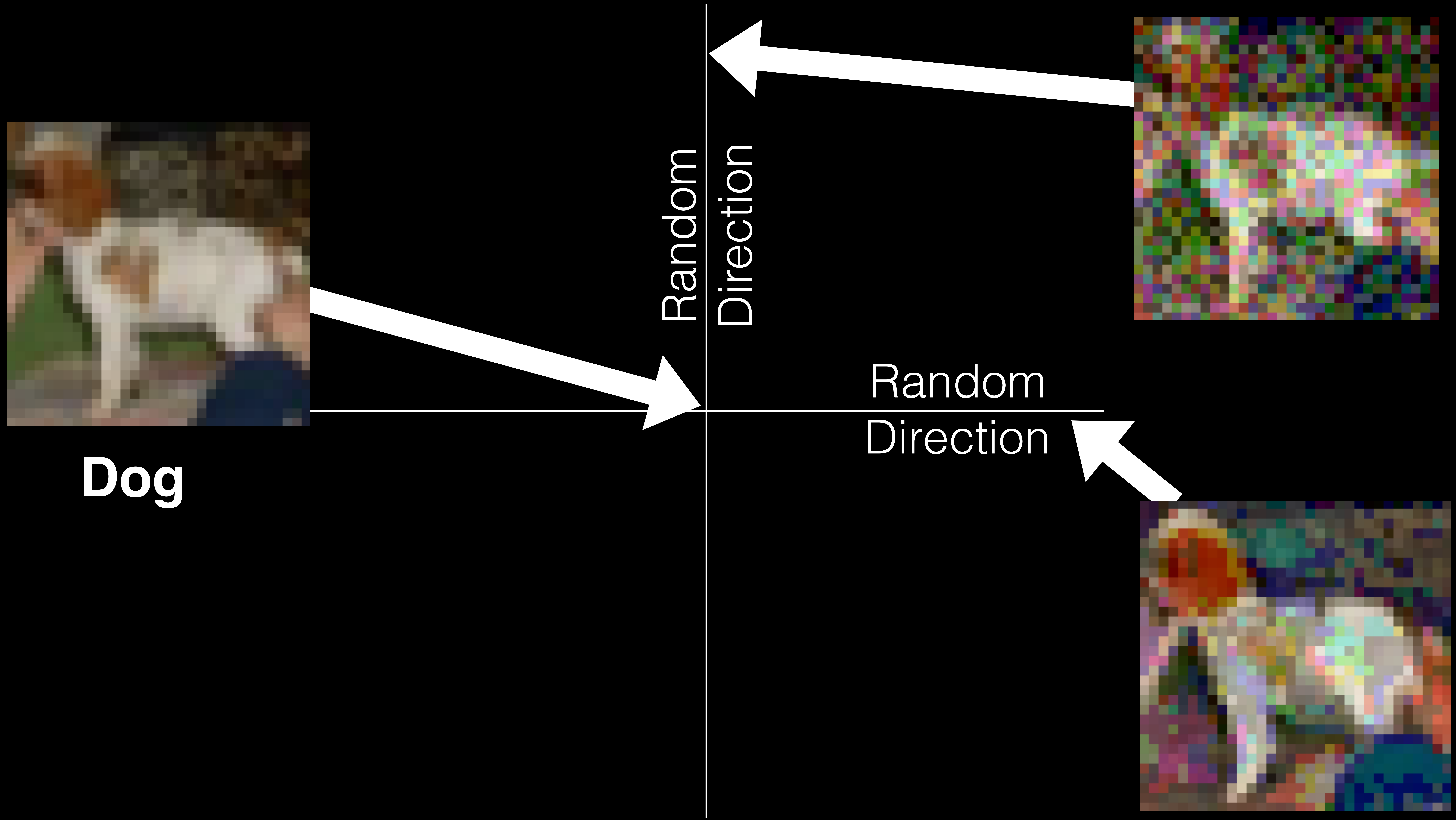
Not a real person

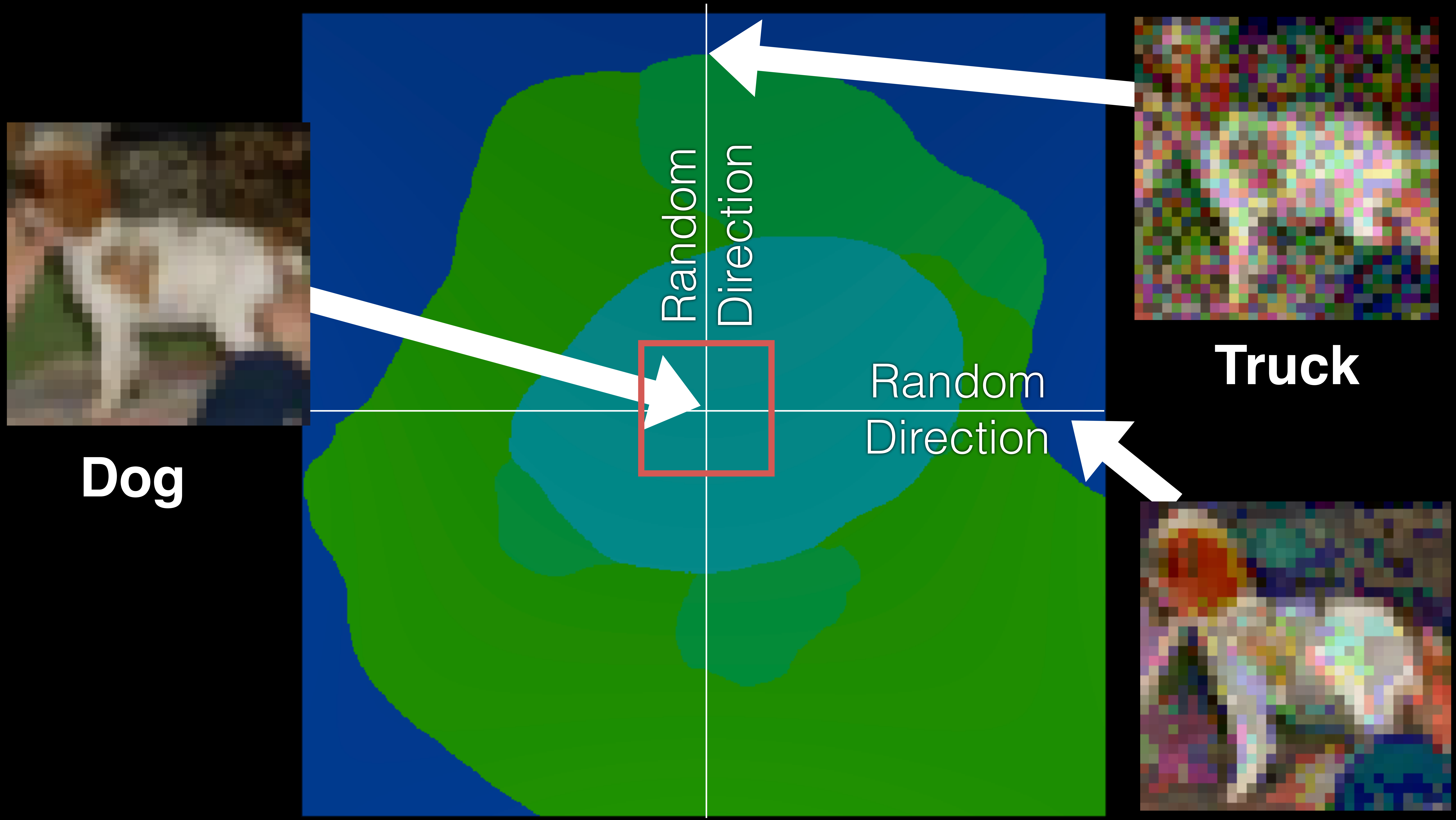


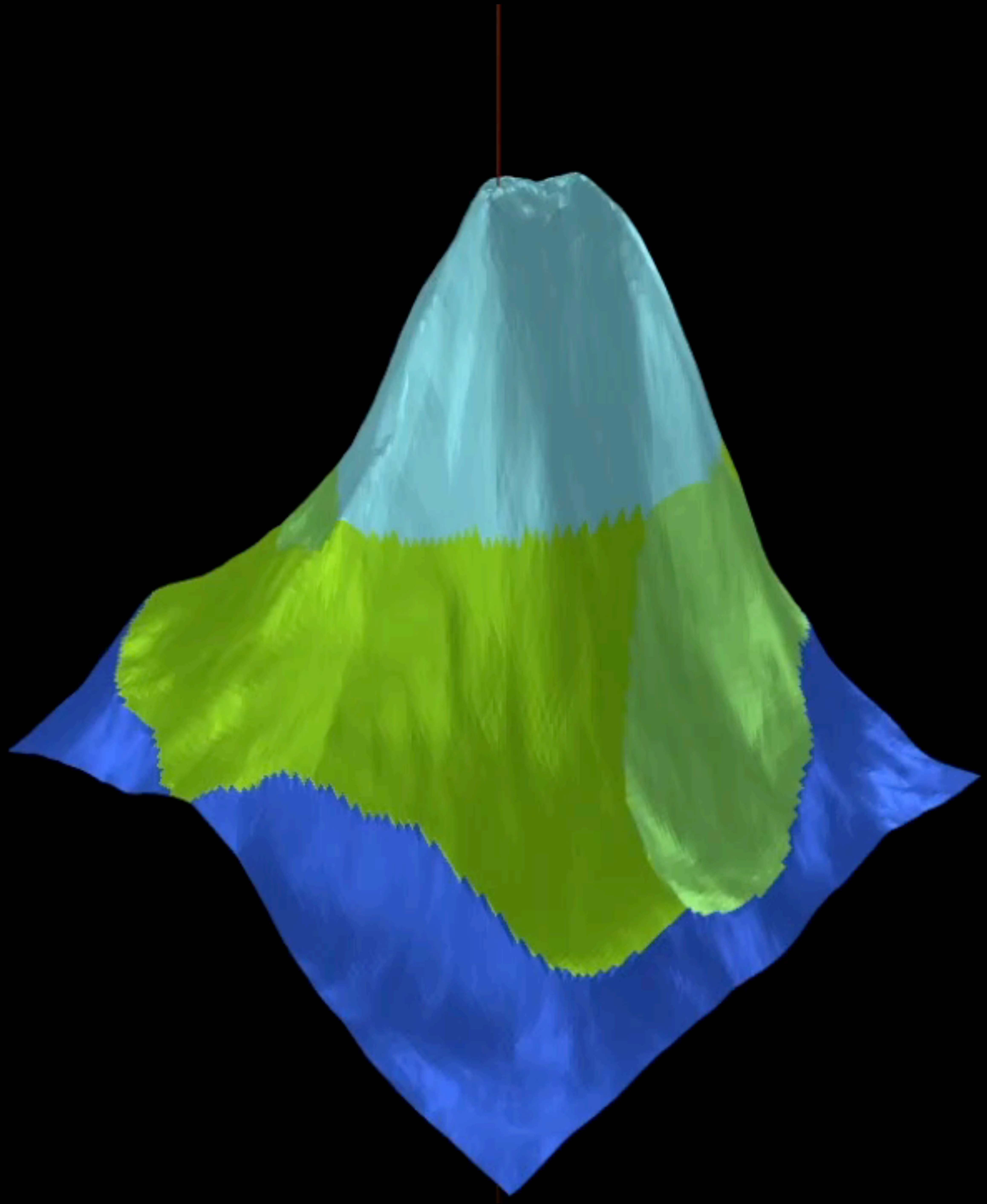
REAL

Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"

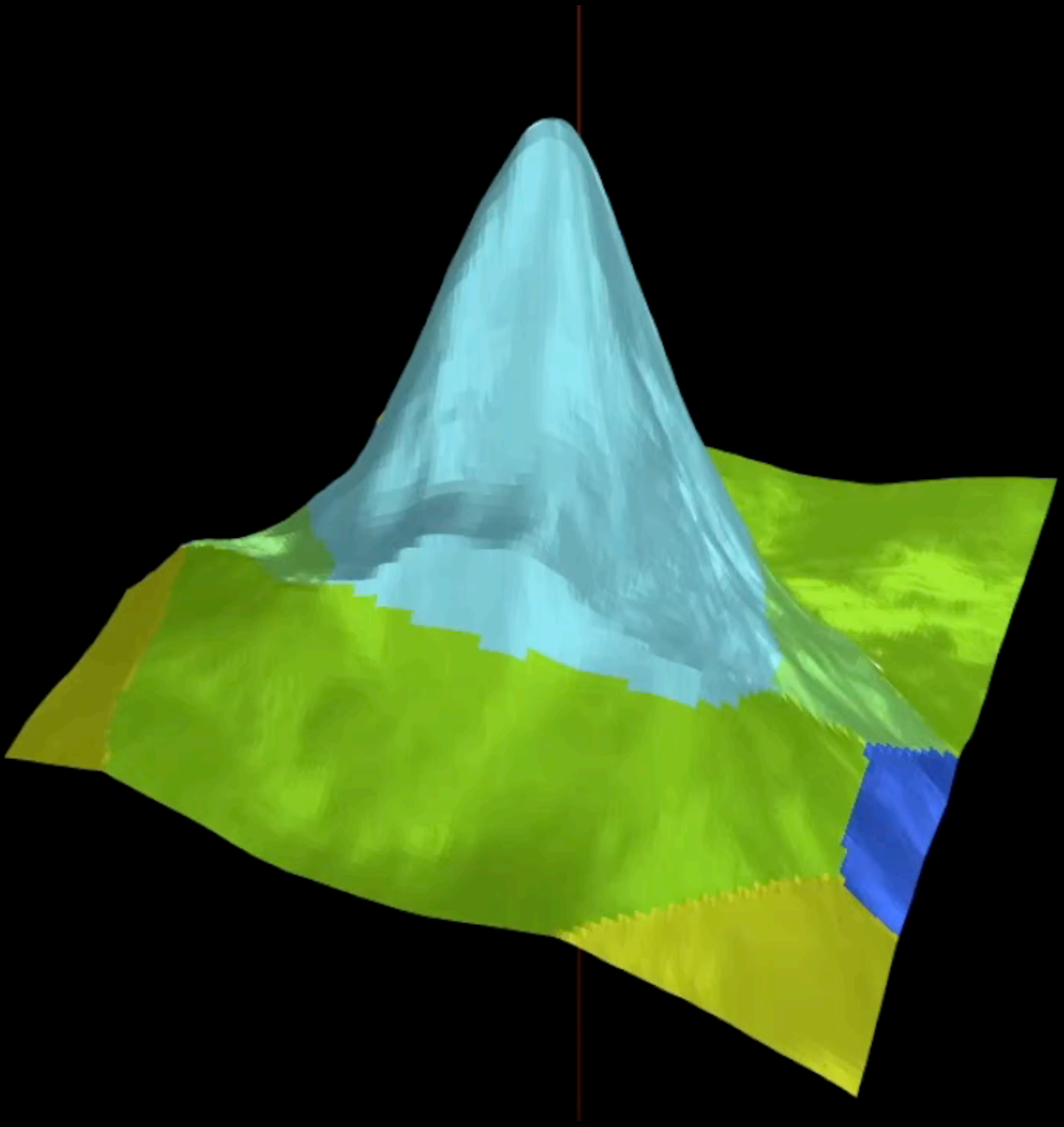
How do we generate adversarial examples?


Random
Direction
Random
Direction
Dog


Random Direction
Random Direction
Dog
Truck

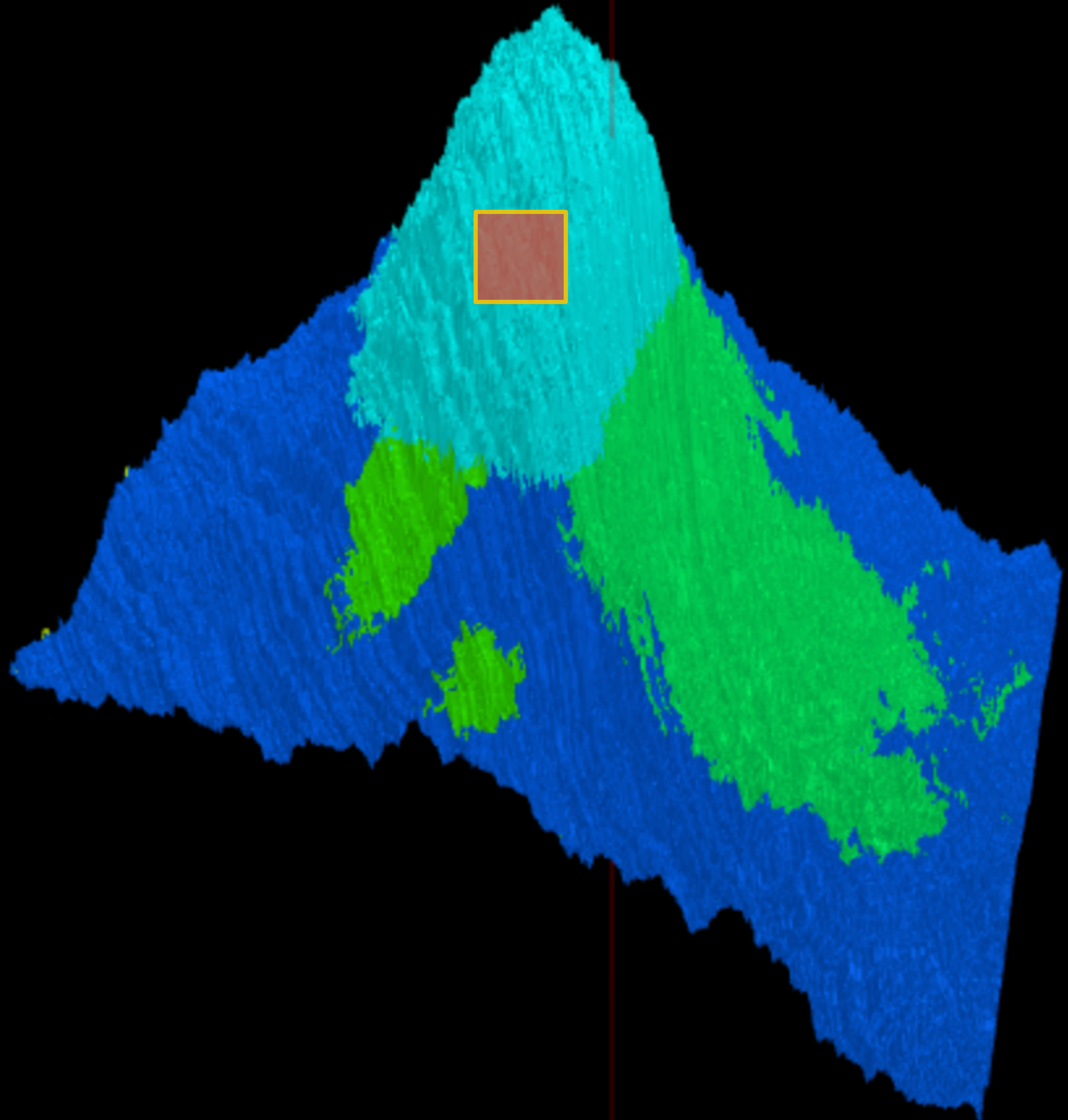


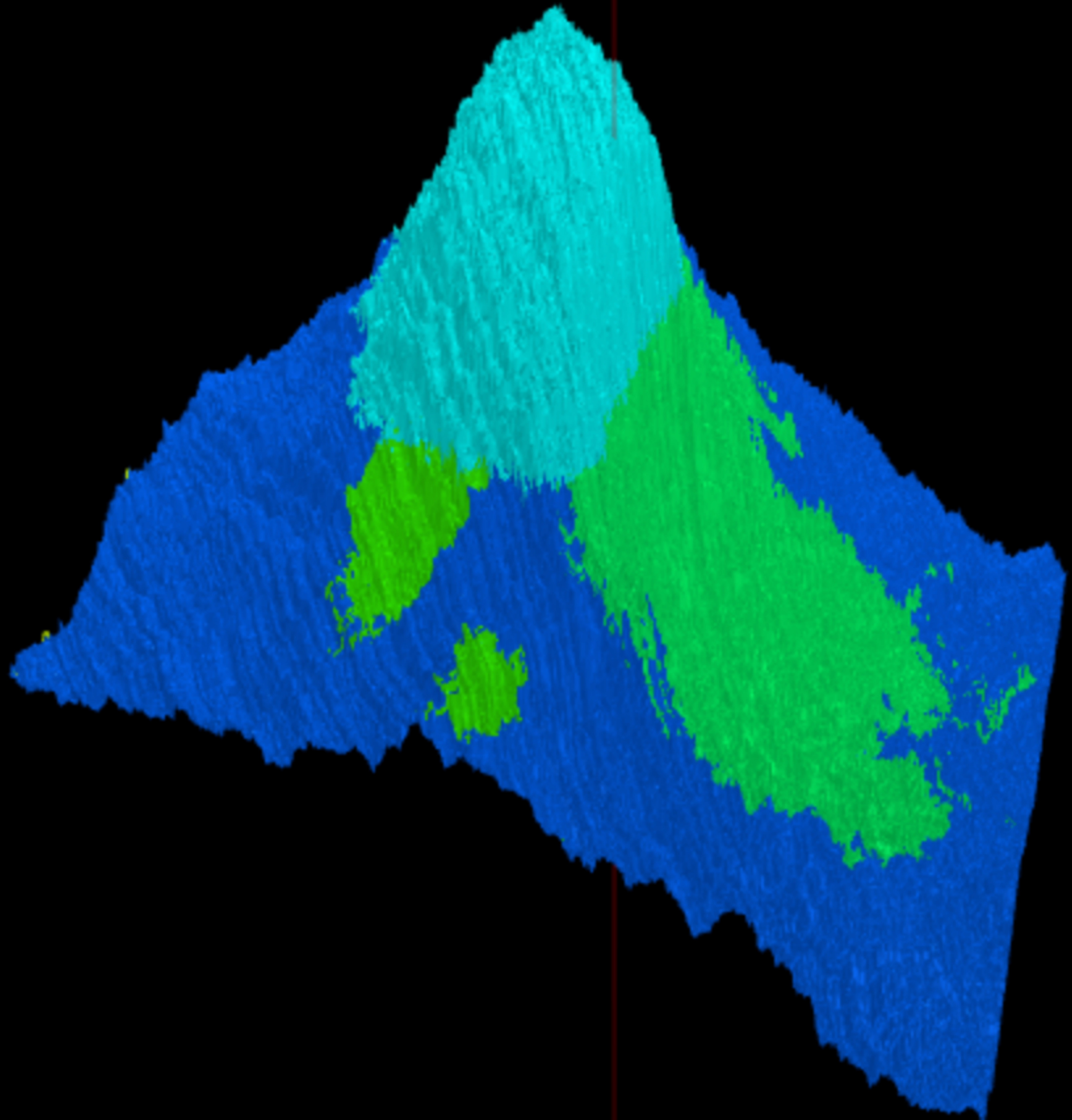
Random Direction
Truck
Adversarial Direction
Dog
Airplane

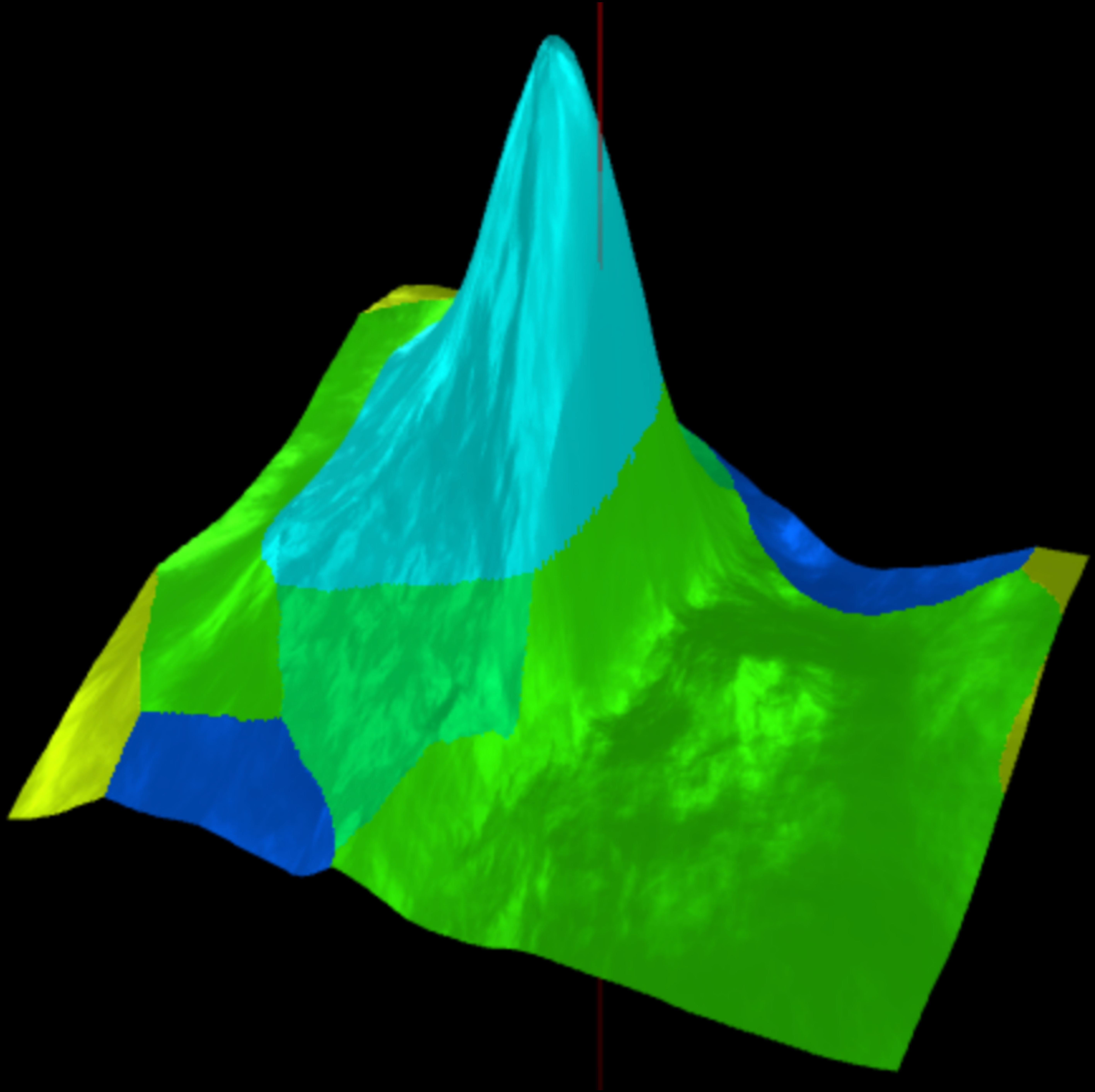


That sounds bad.

Let's defend against it....







That was 2018
How are things today?

On Adaptive Attacks to Adversarial Example Defenses

Florian Tramèr* Stanford University
Nicholas Carlini* Google Brain
Wieland Brendel* University of Tübingen
Aleksander Mądry MIT

We evaluated 13 defenses proposed at (ICLR|ICML|NeurIPS) 20(18|19|20)

All were broken.
Adversarial accuracy of roughly 0%.

Tramer, Carlini, Brendel, Madry. "On Adaptive Attacks to Adversarial Example Defenses"

This is not new ...

Defenses

New Idea 1

Attacks

New Idea A

Defenses

New Idea 1

New Idea 2

Attacks

New Idea A

New Idea B

Defenses		Attacks
New Idea 1	→	New Idea A
New Idea 2	→	New Idea B
New Idea 3	→	New Idea C

Defensive Distillation is Not Robust to Adversarial Examples

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

MagNet and "Efficient Defenses Against Adversarial Attacks" are Not Robust to Adversarial Examples

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

On the Robustness of the CVPR 2018 White-Box Adversarial Example Defenses

Is AmI (Attacks Meet Interpretability) Robust to Adversarial Examples?

Comment on *Biologically inspired protection of deep networks from adversarial attacks*

On the Limitation of Local Intrinsic Dimensionality for Characterizing the Subspaces of Adversarial Examples

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

The Efficacy of SHIELD under Different Threat Models

Paper Type: Appraisal Paper of Existing Method

Cory Cornelius cory.cornelius@intel.com

Nilaksh Das nilakshdas@gatech.edu

Shang-Tse Chen schen351@gatech.edu

Evaluating and Understanding the Robustness of Adversarial Logit Pairing

Logan Engstrom* **Andrew Ilyas*** **Anish Athalye***

Massachusetts Institute of Technology

{engstrom,ailyas,aathalye}@mit.edu

Abstract

We evaluate the robustness of Adversarial Logit Pairing, a recently proposed defense against adversarial examples. We find that a network trained with Adversarial Logit Pairing achieves 0.6% correct classification rate under targeted adversarial attack, the threat model in which the defense is considered. We provide a brief overview of the defense and the threat models/claims considered, as well as a discussion of the methodology and results of our attack. Our results offer insights into the reasons underlying the vulnerability of ALP to adversarial attack, and are of general interest in evaluating and understanding adversarial defenses.

1 Contributions

For summary, the contributions of this note are as follows:

1. **Robustness**: Under the white-box targeted attack threat model specified in Kannan et al., we upper bound the correct classification rate of the defense to **0.6%** (Table 1). We also perform targeted and untargeted attacks and show that the attacker can reach success rates of 98.6% and 99.9% respectively (Figures 1, 2).

Today …

Defenses	Attacks
New Idea 1	New Idea A
New Idea 2	New Idea B
New Idea 3	New Idea C
New Idea 95	

Defenses		Attacks
New Idea 1	→	New Idea A
New Idea 2	→	New Idea B
New Idea 3	→	New Idea C
New Idea 95	→	just reuse one

Reviewer 3:

Another **weakness** of the paper is that **defenses are broken by existing techniques**. Indeed, at the end of the analysis, most of the defenses are broken either by using EOT, BPDA, or by tuning the parameters of existing attacks such as PGD. Some defenses are broken by using decision based attacks. **All this techniques already exist in the litterature** [1,2,3,4]; hence the technical part is not novel (see also related work section).

The problem
is methodological

for example ... one paper's attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\text{adv}})}_{\text{misclassify } \mathbf{x}' \text{ as } y_t},$$

$$\mathcal{L}_2 = \underbrace{\mathbb{E}_{\epsilon \sim N(0,\sigma^2 I)} [\|h(\mathbf{x}') - h(\mathbf{x}' + \epsilon)\|_1]}_{\text{bypass C1}},$$

$$\mathcal{L}_3 = \mathbb{E}_{y' \sim \text{Uniform}, y' \neq y_t} [\mathcal{L}(h(\mathbf{x}' + \alpha\delta_{y'}), y')],$$

$$\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}' + \alpha\delta_{y_t}), y_t).$$

$$\mathcal{L}^\star = \lambda\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4.$$

for example ... one paper's attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\text{adv}})}_{\text{misclassify } \mathbf{x}' \text{ as } y_t},$$

$$\mathcal{L}_2 = \underbrace{\mathbb{E}_{\epsilon \sim N(0,\sigma^2 I)} \left[\| h(\mathbf{x}') - h(\mathbf{x}' + \epsilon) \|_1 \right]}_{\text{bypass C1}},$$

$$\mathcal{L}_3 = \mathbb{E}_{y' \sim \text{Uniform}, y' \neq y_t} [\mathcal{L}(h(\mathbf{x}' + \alpha\delta_{y'}), y')],$$

$$\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}' + \alpha\delta_{y_t}), y_t).$$

$$\mathcal{L}^{\star} = \lambda\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4.$$

for example ... our attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\text{adv}})}_{\text{misclassify } \mathbf{x}' \text{ as } y_t},$$

Not *everything*
is broken …

Idea #1:
Adversarial Training

Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks"

Normal Training

( , 7)

( , 3)

Training

Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks"

Adversarial Training (1)

(,7)

(,3)

(,7)

(,3)

Attack

Adversarial Training (2)

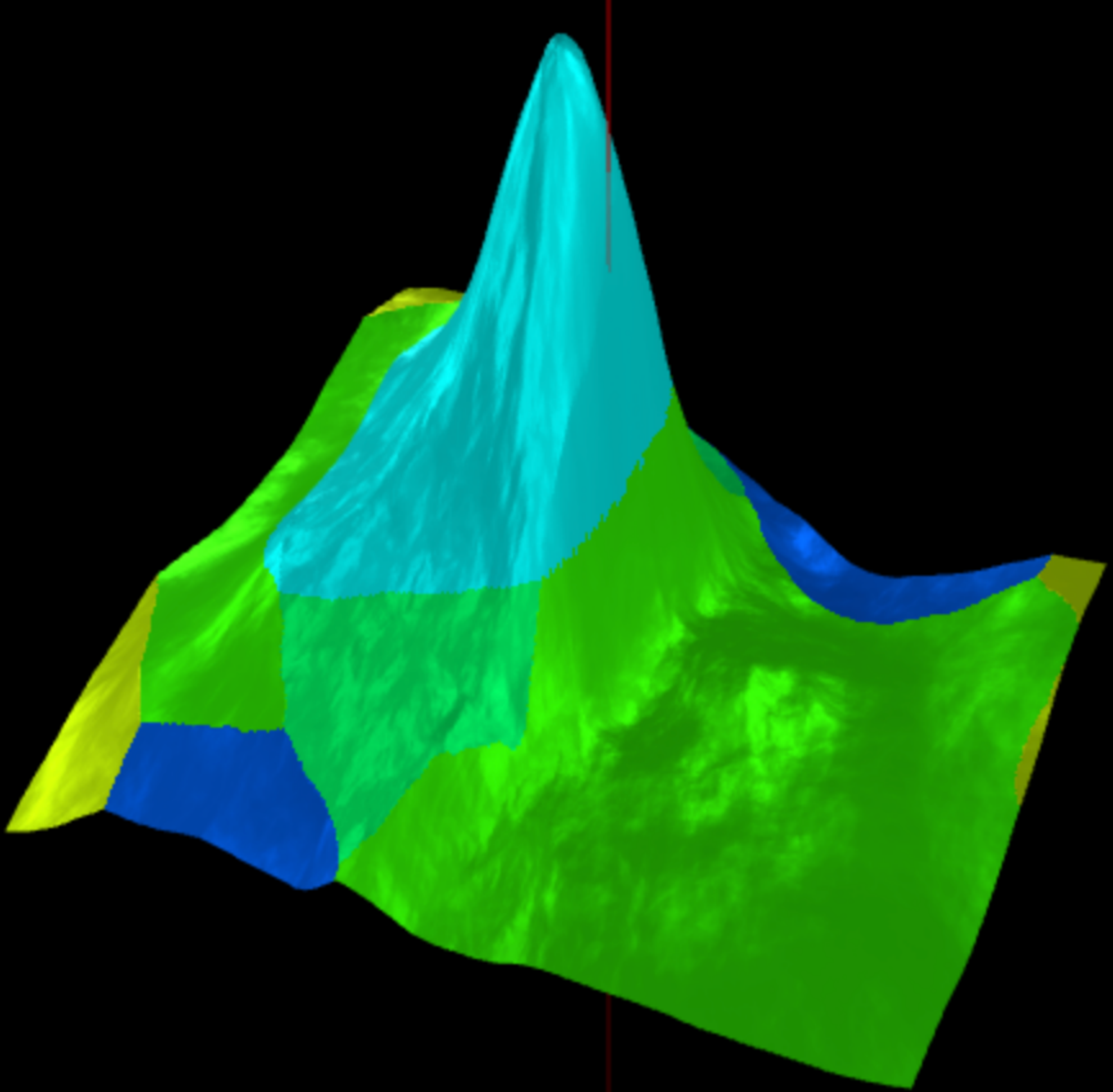
( ,7)

( ,3)

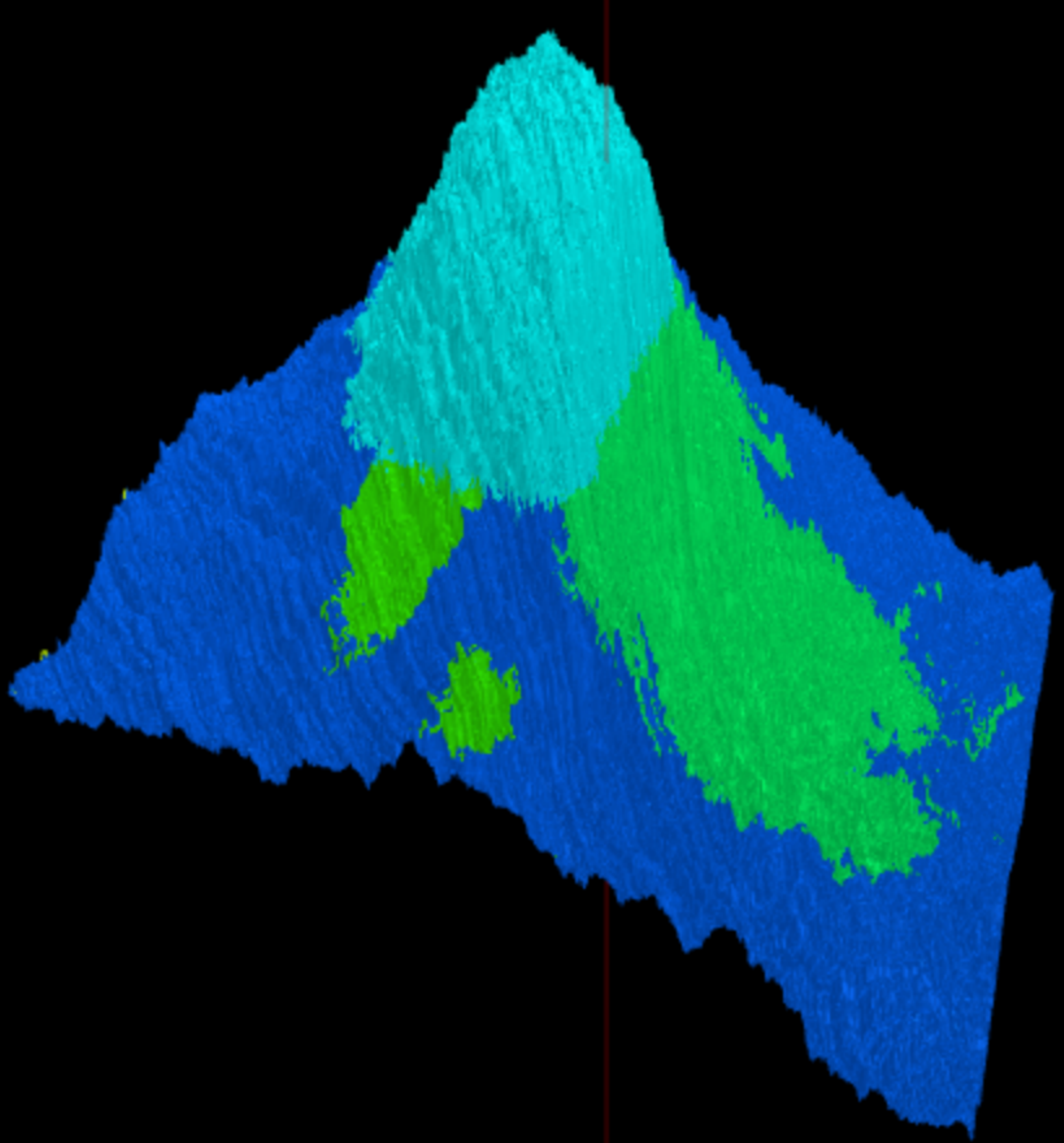
( ,7)

( ,3)

Training

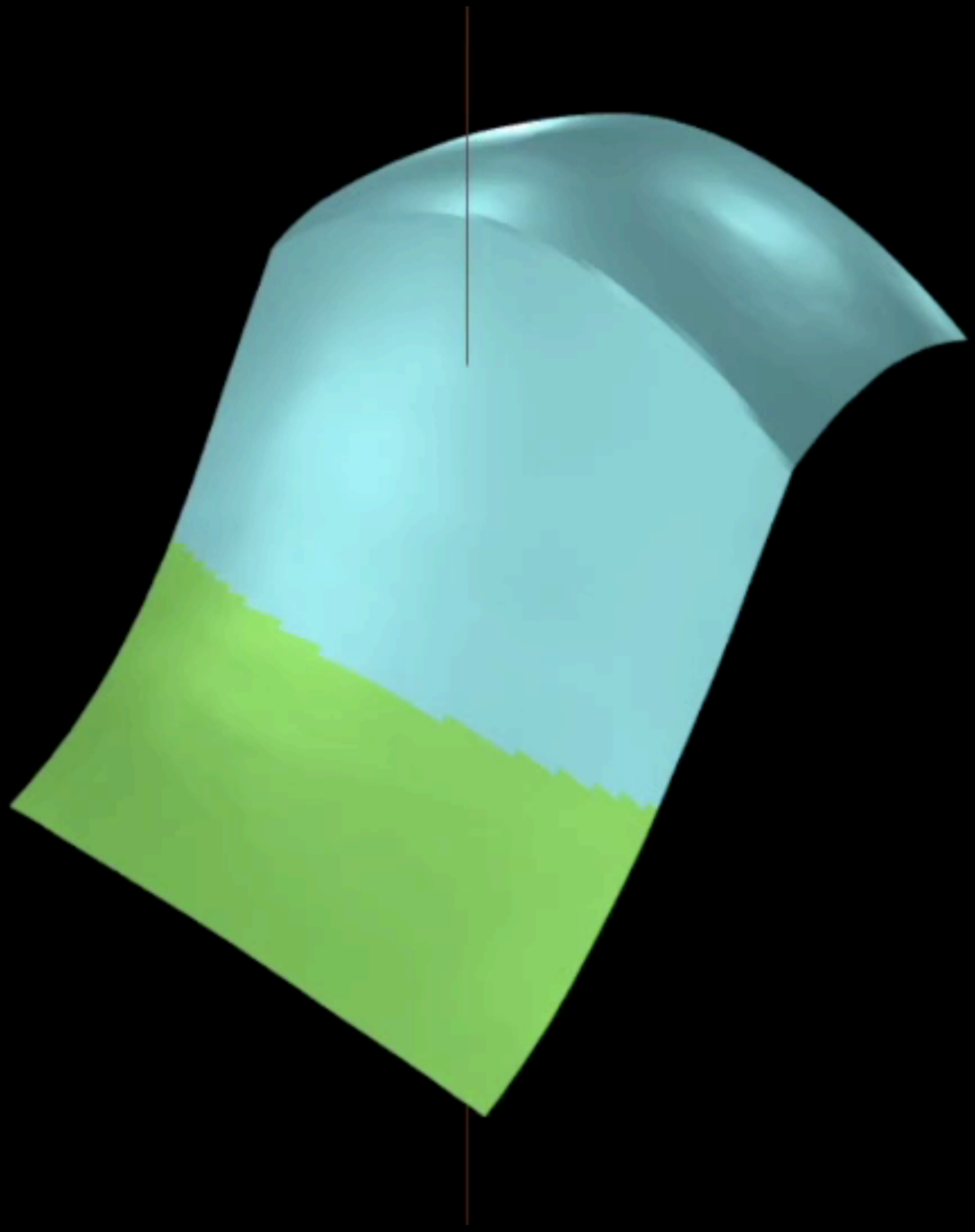


Normal
Loss
Surface

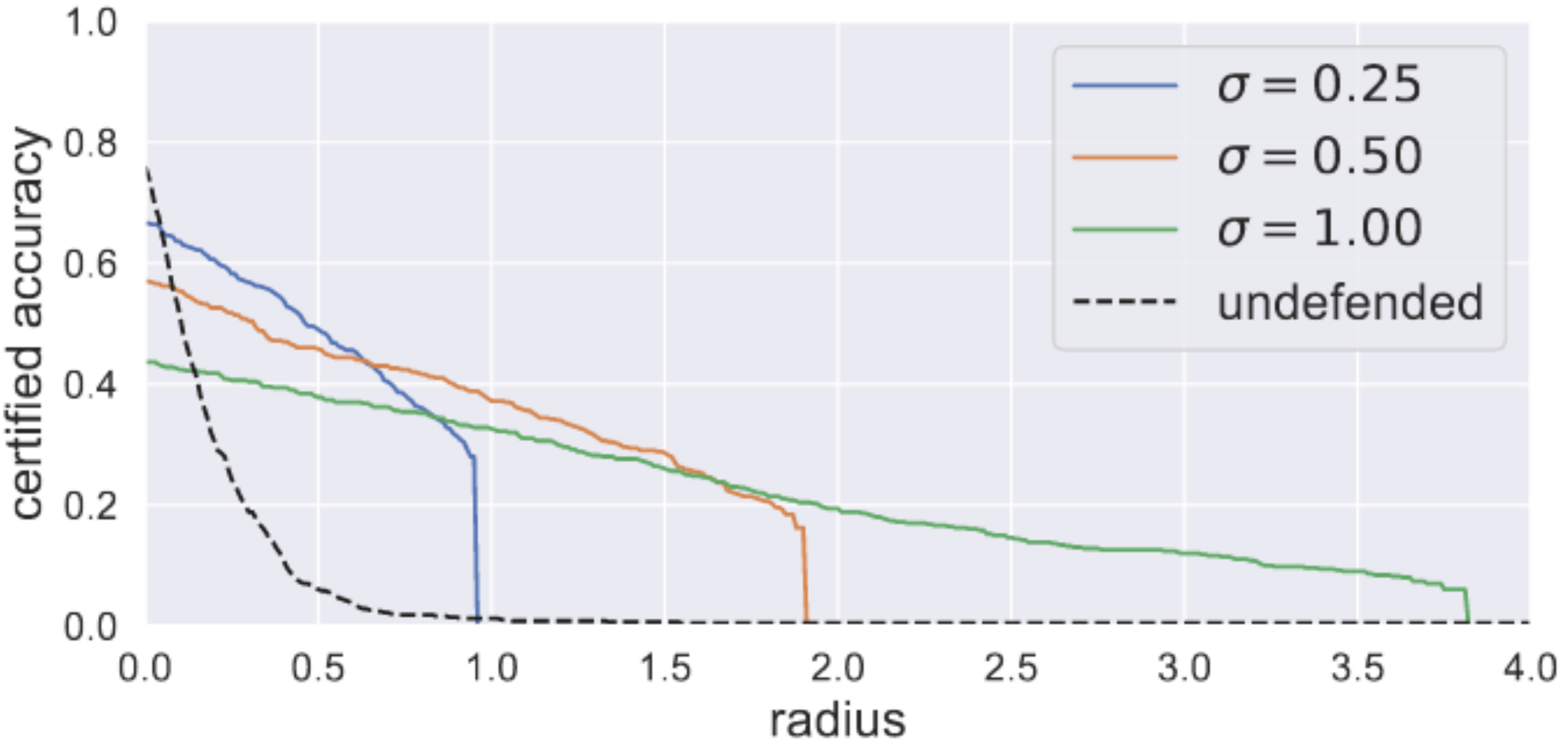


Obfuscated
Loss
Surface

Adversarial Training Loss Surface



Idea #2:
Certified Defenses



Lecuyer et al. "Certified Robustness to Adversarial Examples with Differential Privacy"
Cohen et al. "Certified Adversarial Robustness via Randomized Smoothing"

What's next?

Defensive Distillation is Not Robust to Adversarial Examples

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

MagNet and "Efficient Defenses Against Adversarial Attacks" are Not Robust to Adversarial Examples

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

On the Robustness of the CVPR 2018 White-Box Adversarial Example Defenses

Is AmI (Attacks Meet Interpretability) Robust to Adversarial Examples?

Abstract—No.

I. Attacking "Attacks meet Interpretability"

Comment on *Biologically inspired protection of deep networks from adversarial attacks*

On the Limitation of Local Intrinsic Dimensionality for Characterizing the Subspaces of Adversarial Examples

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

The Efficacy of Shield under Different Threat Models

Paper Type: Appraisal Paper of Existing Method

Cory Cornelius cory.cornelius@intel.com
Nilaksh Das nilakshdas@gatech.edu
Shang-Tse Chen schen351@gatech.edu

Evaluating and Understanding the Robustness of Adversarial Logit Pairing

Logan Engstrom* Andrew Ilyas* Anish Athalye*
Massachusetts Institute of Technology
{engstrom,ailyas,aathalye}@mit.edu

Abstract

We evaluate the robustness of Adversarial Logit Pairing, a recently proposed defense against adversarial examples. We find that a network trained with Adversarial Logit Pairing achieves 0.6% correct classification rate under targeted adversarial attack, the threat model in which the defense is considered. We provide a brief overview of the defense and the threat models/claims considered, as well as a discussion of the methodology and results of our attack. Our results offer insights into the reasons underlying the vulnerability of ALP to adversarial attack, and are of general interest in evaluating and understanding adversarial defenses.

1 Contributions

For summary, the contributions of this note are as follows:

1. **Robustness**: Under the white-box targeted attack threat model specified in Kannan et al., we upper bound the correct classification rate of the defense to **0.6%** (Table 1). We also perform targeted and untargeted attacks and show that the attacker can reach success rates of 98.6% and 99.9% respectively (Figures 1, 2).

The Year is 1997

Cryptanalysis of the Cellular Message Encryption Algorithm

Related-Key Cryptanalysis of 3-WAY, Biham-DES,CAST, DES-X, NewDES, RC2, and TEA

Cryptanalysis of some recently-proposed multiple modes of operation

Differential cryptanalysis of KHF

Cryptanalysis of TWOPRIME

Don Coppersmith[1], David Wagner[2], Bruce Schneier[3], and J

[1] IBM Research, e-mail: copper@watson.ibm.com
[2] U.C. Berkeley, e-mail: daw@cs.berkeley.edu
[3] Counterpane Systems, e-mail: {schneier,kelsey}@counter

Abstract. Ding et al [DNRS97] propose a stream generator several layers. We present several attacks. First, we observe non-surjectivity of a linear combination step allows us to re the key with minimal effort. Next, we show that the various insufficiently mixed by these layers, enabling an attack similar t two-loop Vigenere ciphers to recover the remainder of the key. these techniques lets us recover the entire TWOPRIME key. the generator to produce 2^{33} blocks (2^{35} bytes), or 19 hours output, of which we examine about one million blocks (2^{23} computational workload can be estimated at 2^{28} operations set of attacks trades off texts for time, reducing the amount plaintext needed to just eight blocks (64 bytes), while needin and 2^{32} space. We also show how to break two variants of TW presented in the original paper.

1 Introduction

Cryptanalysis of SPEED

Cryptanalysis of FROG

Cryptanalysis of ORYX

The boomerang attack

Slide Attacks

Alex Biryukov* David Wagner**

Abstract. It is a general belief among the designers of block-ciphers that even a relatively weak cipher may become very strong if its number of rounds is made very large. In this paper we describe a new generic known- (or sometimes chosen-) plaintext attack on product ciphers, which we call the *slide attack* and which in many cases is independent of the number of rounds of a cipher. We illustrate the power of this new tool by giving practical attacks on several recently designed ciphers: TREYFER, WAKE-ROFB, and variants of DES and Blowfish.

1 Introduction

As the speed of computers grows, fast block ciphers tend to use more and more rounds, rendering all currently known cryptanalytic techniques useless. This is mainly due to the fact that such popular tools as differential [1] and linear analysis [13] are statistic attacks that excel in pushing statistical irregularities and biases through surprisingly many rounds of a cipher. However any such approach finally reaches its limits, since each additional round requires an exponential effort from the attacker.

This tendency towards a higher number of rounds can be illustrated if one looks at the candidates submitted to the AES contest. Even though one of the main criteria of the AES was speed, several prospective candidates (and not the slowest ones) have really large numbers of rounds: RC6(20), MARS(32),

Back to (the future)

Biclique Cryptanalysis of the Full AES

Andrey Bogdanov*, Dmitry Khovratovich, and Christian Rechberger*

K.U. Leuven, Belgium; Microsoft Research Redmond, USA; ENS Paris and Chaire France Telecom, France

Abstract. Since Rijndael was chosen as the Advanced Encryption Standard, improving upon 7-round attacks on the 128-bit key variant or upon 8-round attacks on the 192/256-bit key variants has been one of the most difficult challenges in the cryptanalysis of block ciphers for more than a decade. In this paper we present a novel technique of block cipher cryptanalysis with bicliques, which leads to the following results:

– The first key recovery attack on the full AES-128 with computational complexity $2^{126.1}$.
– The first key recovery attack on the full AES-192 with computational complexity $2^{189.7}$.
– The first key recovery attack on the full AES-256 with computational complexity $2^{254.4}$.
– Attacks with lower complexity on the reduced-round versions of AES not considered before, including an attack on 8-round AES-128 with complexity $2^{124.9}$.
– Preimage attacks on compression functions based on the full AES versions.

In contrast to most shortcut attacks on AES variants, we *do not need to assume related-keys*. Most of our attacks only need a very small part of the codebook and have small memory requirements, and are practically verified to a large extent. As our attacks are of high computational complexity, they do not threaten the practical use of AES in any way.

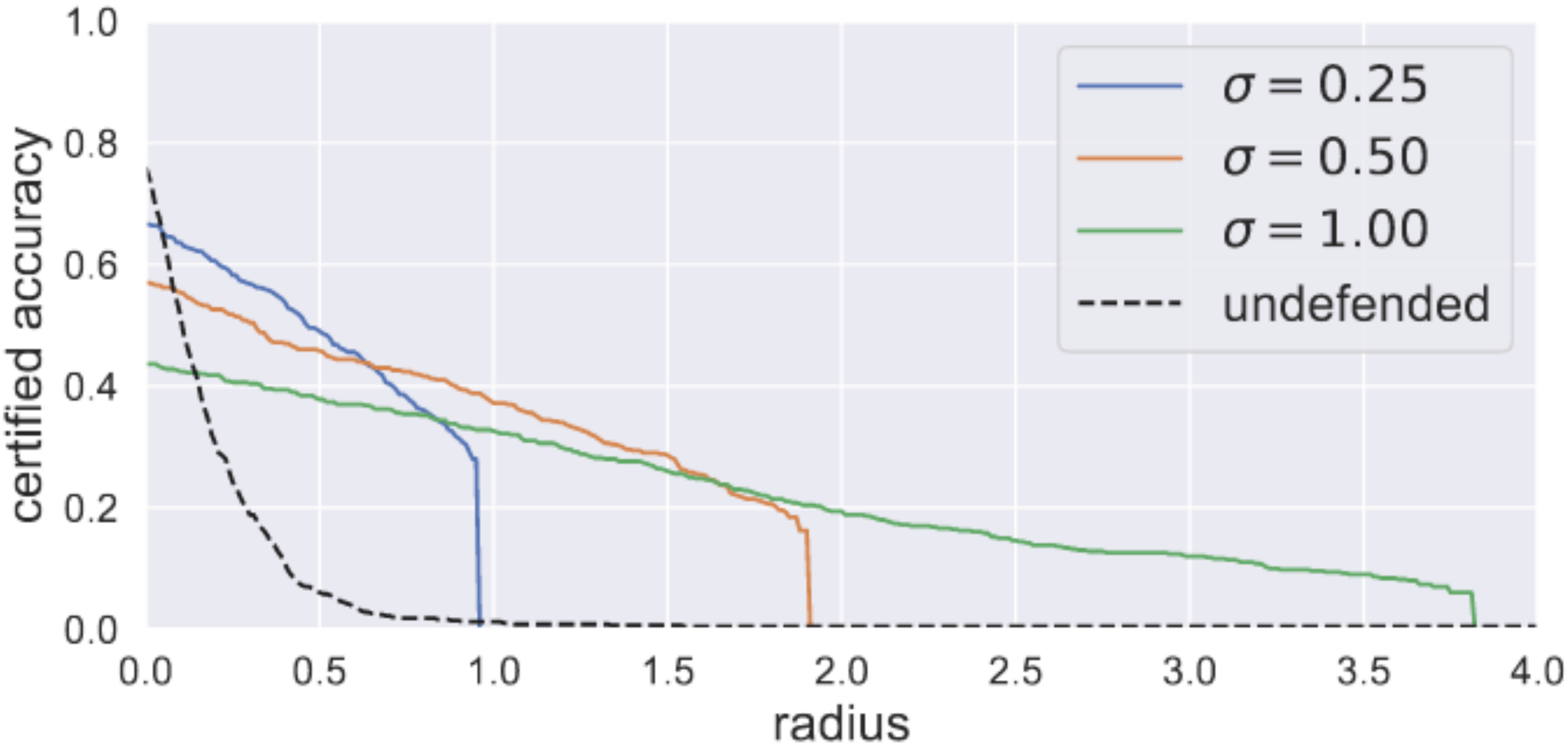
Keywords: block ciphers, bicliques, AES, key recovery, preimage

Are we crypto in the 90's?

Maybe not.

Three reasons.

Reason 1.


certified accuracy
radius
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
$\sigma = 0.25$
$\sigma = 0.50$
$\sigma = 1.00$
undefended

Attack Success Rates in Security

Evans, "Is "adversarial example" an adversarial example?"

Attack Success Rates in Security

Crypto: 2^{-128}

Evans, "Is "adversarial example" an adversarial example?"

Attack Success Rates in Security

Crypto: 2^{-128}, broken if 2^{-127}

Evans, "Is "adversarial example" an adversarial example?"

Attack Success Rates in Security

Crypto: 2^{-128}, broken if 2^{-127}

Systems: 2^{-32}

Evans, "Is "adversarial example" an adversarial example?"

Attack Success Rates in Security

Crypto: 2^{-128}, broken if 2^{-127}

Systems: 2^{-32}, broken if 2^{-20}

Evans, "Is "adversarial example" an adversarial example?"

Attack Success Rates in Security

Crypto: 2^{-128}, broken if 2^{-127}

Systems: 2^{-32}, broken if 2^{-20}

Machine Learning:

Evans, "Is "adversarial example" an adversarial example?"

Attack Success Rates in Security

Crypto: 2^{-128}, broken if 2^{-127}

Systems: 2^{-32}, broken if 2^{-20}

Machine Learning: $\mathbf{2^{-1}}$

Evans, "Is "adversarial example" an adversarial example?"

Attack Success Rates in Security

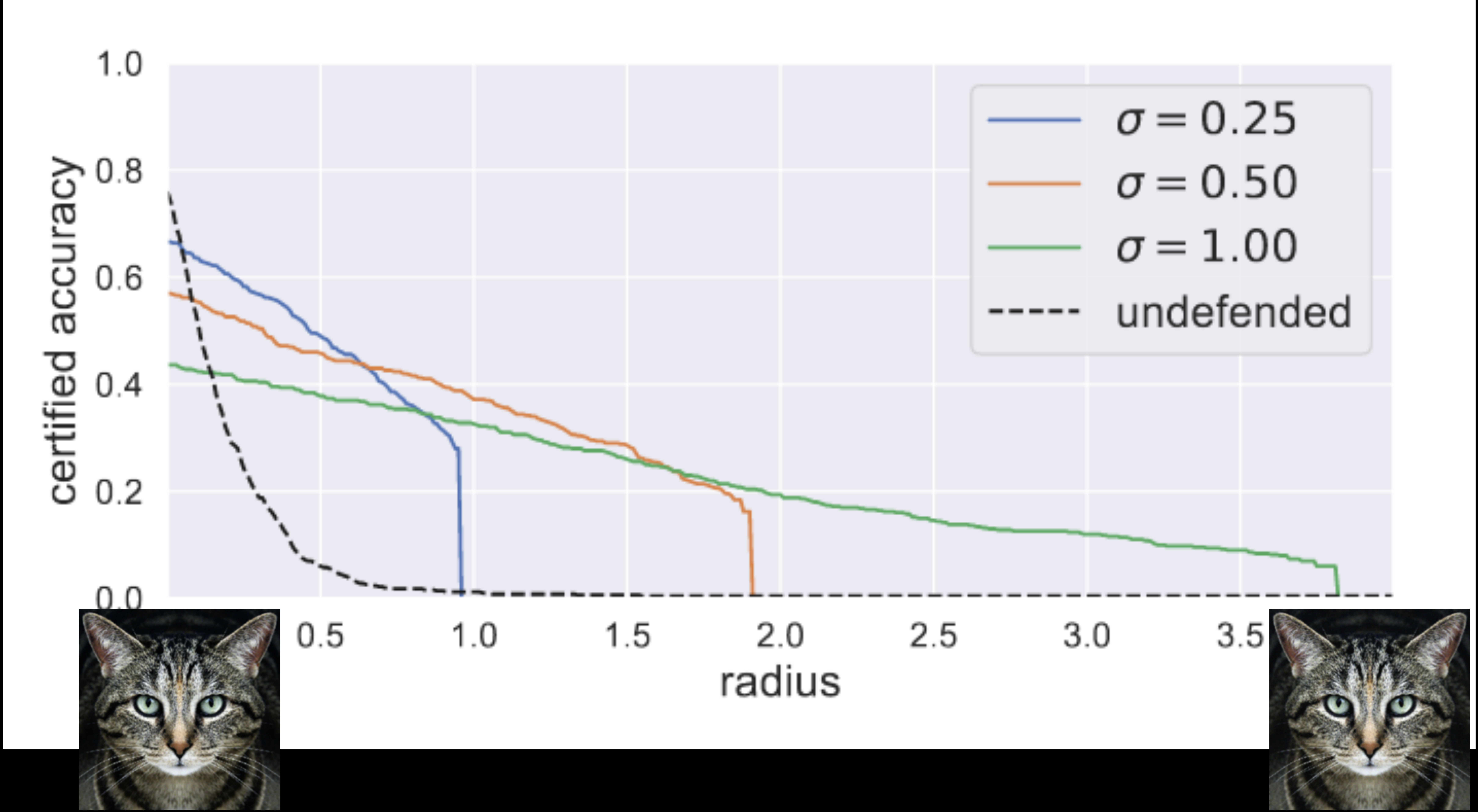
Crypto: 2^{-128}, broken if 2^{-127}

Systems: 2^{-32}, broken if 2^{-20}

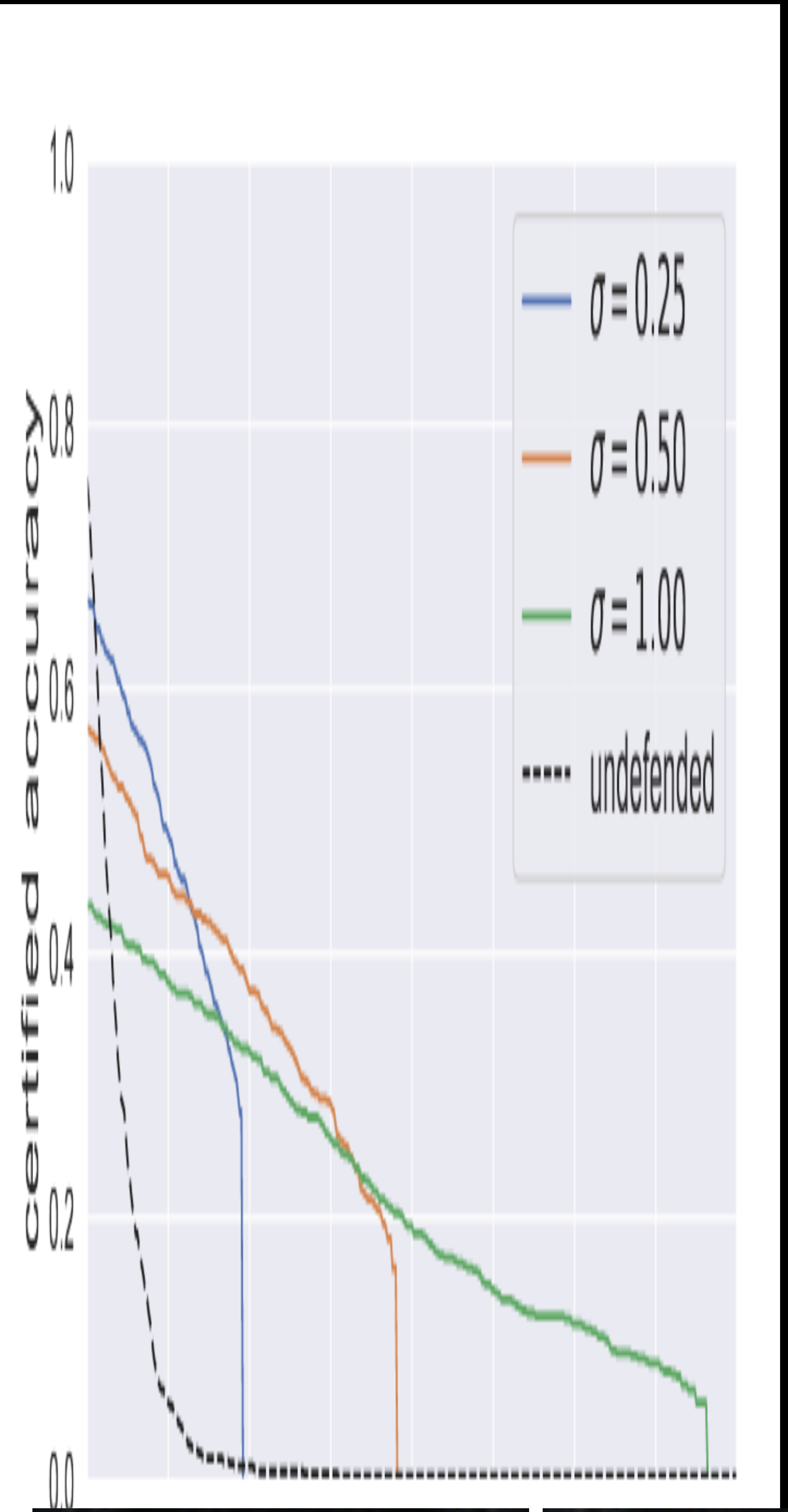
Machine Learning: $\mathbf{2^{-1}}$, broken if $\mathbf{2^{0}}$

Evans, "Is "adversarial example" an adversarial example?"

Reason 2.



certified accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
radius
$\sigma = 0.25$
$\sigma = 0.50$
$\sigma = 1.00$
undefended



certified accuracy
1.0
0.8
0.6
0.4
0.2
0.0
σ=0.25
σ=0.50
σ=1.00
undefended

$L_2 = 100$

Original

L_2 distortion: 75

L_2 distortion: 75

Claim:
We are crypto **pre-**Shannon

Reason 3.

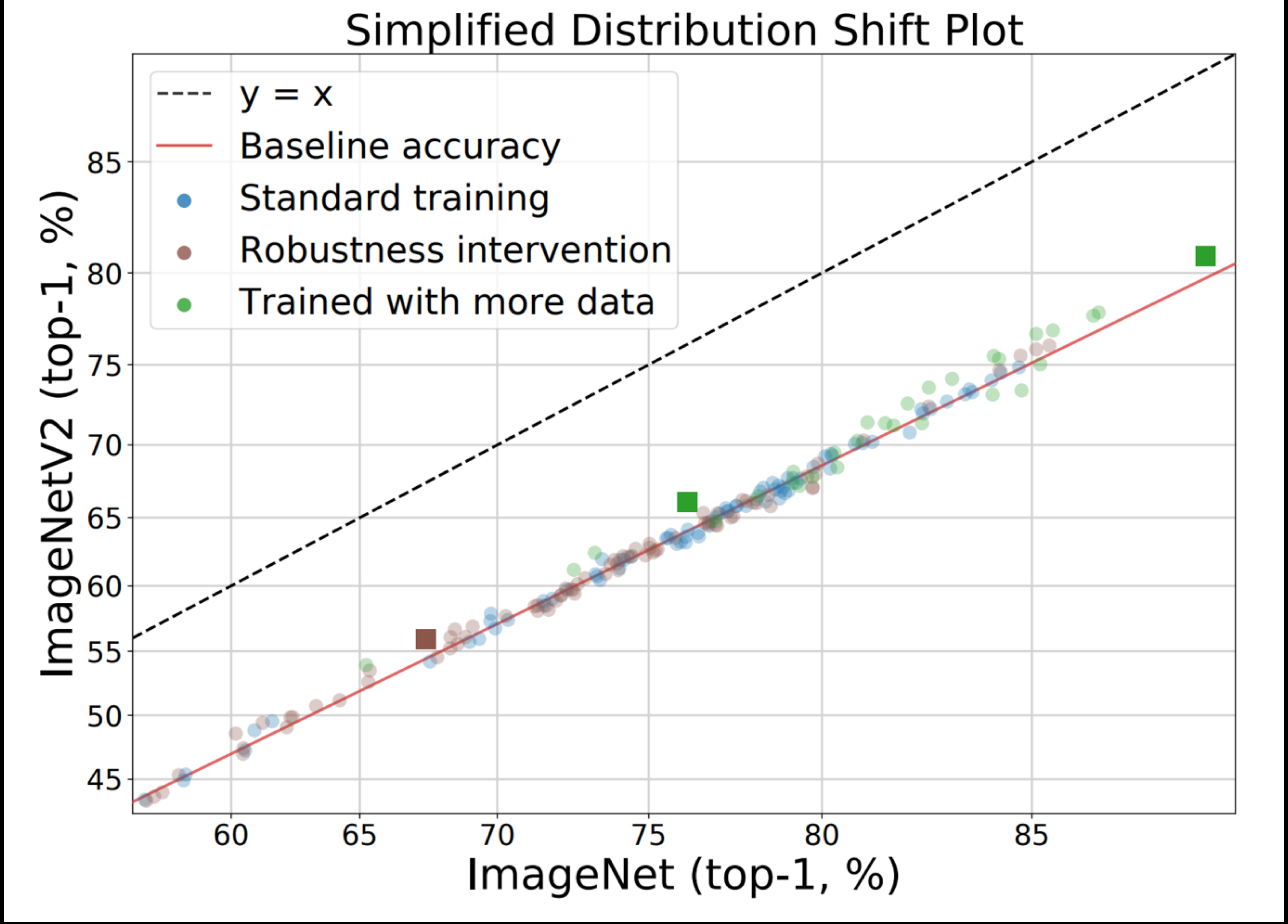
It's not just adversarial shifts ...

Do ImageNet Classifiers Generalize to ImageNet?

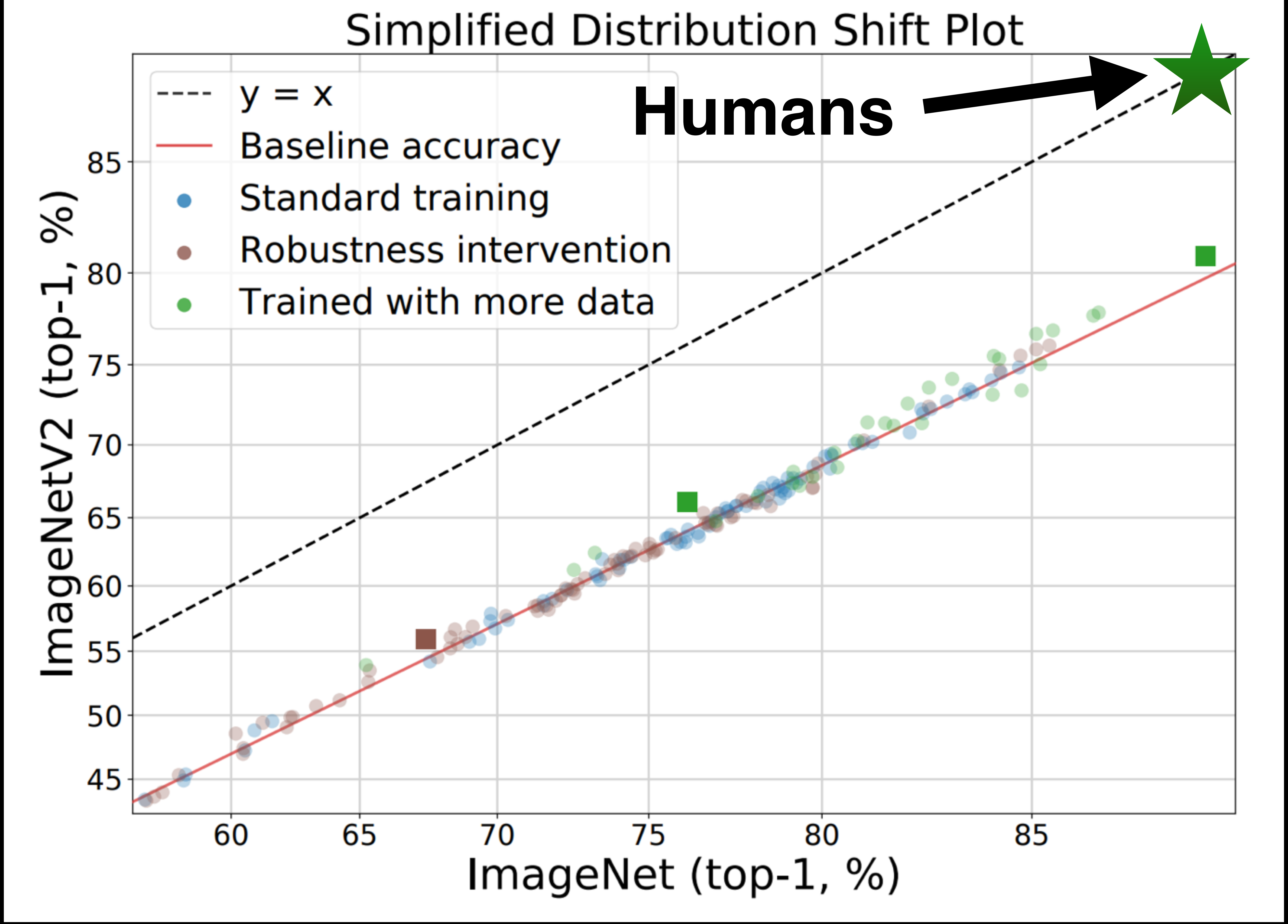
Benjamin Recht* UC Berkeley
Rebecca Roelofs UC Berkeley
Ludwig Schmidt UC Berkeley
Vaishaal Shankar UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.



Taori et al., "Measuring Robustness to Natural Distribution Shifts in Image Classification"



Taori et al., "Measuring Robustness to Natural Distribution Shifts in Image Classification"

Conclusion

We've come a long way towards understanding adversarial robustness.

We still have a long way to go.

nicholas@carlini.com https://nicholas.carlini.com