A crisis in adversarial machine learning

Nicholas Carlini
Google

Why do we study adversarial machine learning?

We might want to improve ...

1. General purpose robustness
2. The robustness against worst-case attack
3. The robustness against practical attacks

We might want to improve ...

1. General purpose robustness

2. The robustness against worst-case attack

3. The robustness against practical attacks

We might want to improve ...

1. General pu

2. The robustr se attack

3. The robustr attacks

ArtofRobust Workshop Schedule		
Event	**Start time**	**End time**
Opening Remarks	8:50	9:00
Invited talk: Yang Liu	9:00	9:30
Invited talk: Quanshi Zhang	9:30	10:00
Invited talk: Baoyuan Wu	10:00	10:30
Invited talk: Aleksander Mądry	10:30	11:00
Invited talk: Bo Li	11:00	11:30
Poster Session (click)	11:30	12:30
lunch (12:30-13:30)		
Oral Session (click)	13:30	14:10
Challenge Session	14:10	14:30
Invited talk: Nicholas Carlini	14:30	15:00
Invited talk: Judy Hoffman	15:00	15:30
Invited talk: Alan Yuille	15:30	16:00
Invited talk: Ludwig Schmidt	16:00	16:30
Invited talk: Cihang Xie	16:30	17:00

We might want to improve ...

1. General purpose robustness
2. The robustness against worst-case attack
3. The robustness against practical attacks

We might want to improve ...

1. ~~General purpose robustness~~

2. The robustness against worst-case attack

3. The robustness against practical attacks

The Year is **2014**

Someone tells you they have a new algorithm to generate synthetic images

The Year is **2014**

The Year is **2017**

Someone tells you they have a new algorithm to generate synthetic images

The Year is **2017**

The Year is **2022**

Someone tells you they have a new algorithm to generate synthetic images

The Year is **2022**

A photo of a Corgi dog riding a bike in Times Square.
It is wearing sunglasses and a beach hat.

2014

2022

The Year is **2013**

Someone tells you they have discovered
a flaw in the robustness of neural networks

The Year is **2013**

The Year is **2022**

Someone tells you they have discovered
a flaw in the robustness of neural networks

The Year is **2022**

2014

2022

2013

2022

:-(

Why?

Defenses are *really* hard.

That can't be all though.

Consider symmetric key cryptography

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. C
these techniques lets us recover the entire TWOPRIME key. V
the generator to produce 2^{33} blocks (2^{35} bytes), or 19 hours
output, of which we examine about one million blocks (2^{23} b
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),

<6 years later ...

AES is basically perfect

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

For some reason though,
>6 years on, we can't stop
publishing defenses that
are broken by undergrads.

Evading Adversarial Example Detection Defenses with Orthogonal Projected Gradient Descent

Oliver Bryniarski* UC Berkeley

Nabeel Hingun* UC Berkeley

Pedro Pachuca* UC Berkeley

Vincent Wang* UC Berkeley

Nicholas Carlini Google

Abstract

Evading adversarial example detection defenses requires finding adversarial examples that must simultaneously (a) be misclassified by the model and (b) be detected as non-adversarial. We find that existing attacks that attempt to satisfy multiple simultaneous constraints often over-optimize against one constraint at the cost of satisfying another. We introduce *Orthogonal Projected Gradient Descent*, an improved attack technique to generate adversarial examples that avoids this problem by orthogonalizing the gradients when running standard gradient-based attacks. We use our technique to evade four state-of-the-art detection defenses, reducing their accuracy to 0% while maintaining a 0% detection rate.

Does that mean we've made **zero** progress?

Obviously not.

We've gotten really good at knowing how to evaluate correctly, if you try hard.

Increasing Confidence in Adversarial Robustness Evaluations

Roland Zimmermann* University of Tübingen
Wieland Brendel University of Tübingen
Florian Tramèr Google
Nicholas Carlini Google

Abstract

Hundreds of defenses have been proposed in the past years to make deep neural networks robust against minimal (adversarial) input perturbations. However, only a handful of these could hold up their claims because correctly evaluating robustness is extremely challenging: Weak attacks often fail to find adversarial examples even if they unknowingly exist, thereby making a vulnerable network look robust. In this paper, we propose a test to identify weak attacks. Our test introduces a small and simple modification into a neural network that guarantees the existence of an adversarial example for every sample. Consequentially, any correct attack must succeed in attacking this modified network. For eleven out of thirteen previously-published defenses, the original evaluation of the defense fails our test, while stronger attacks that break these defenses pass it. We hope that attack unit tests such as ours will be a major component in future robustness evaluations and increase confidence in an empirical field that today is riddled with skepticism and disbelief. Online version & Code: zimmerrol.github.io/active-tests/

to adversarial examples has proven to be extremely difficult [9]. In many areas of machine learning, evaluating the performance of a new technique is often trivial — for example by computing accuracy on some held-out test set. However evaluating defense robustness necessarily involves reasoning over *all* possible adversaries, and showing *none* can succeed. That is, a defense evaluation aims to prove that something is impossible. As a result, despite significant evaluation effort, most published defenses are quickly broken by stronger attacks [3, 9, 11, 14, 38].

This paper argues for viewing defense proposals as theorem statements, and the corresponding evaluations as proofs. The purpose of a defense evaluation, then, is to provide a convincing and rigorous argument that the defense is correct. Currently, for an adversary to claim to have a "break" of a defense, it is necessary to actually produce the adversarial examples that cause the model to make an error — analogous to refuting a complexity-theoretic impossibility result by producing an efficient algorithm. We argue that this is not how things should work. A valid refutation of a theorem would be to say "there is a flaw in your proof on line 9". Because the null hypothesis for a theorem is that it is false, just as the null hypothesis for a defense should be that it is not robust.

The result I'm most surprised by: certified robustness on ImageNet!

38th IEEE Symposium on Security and Privacy

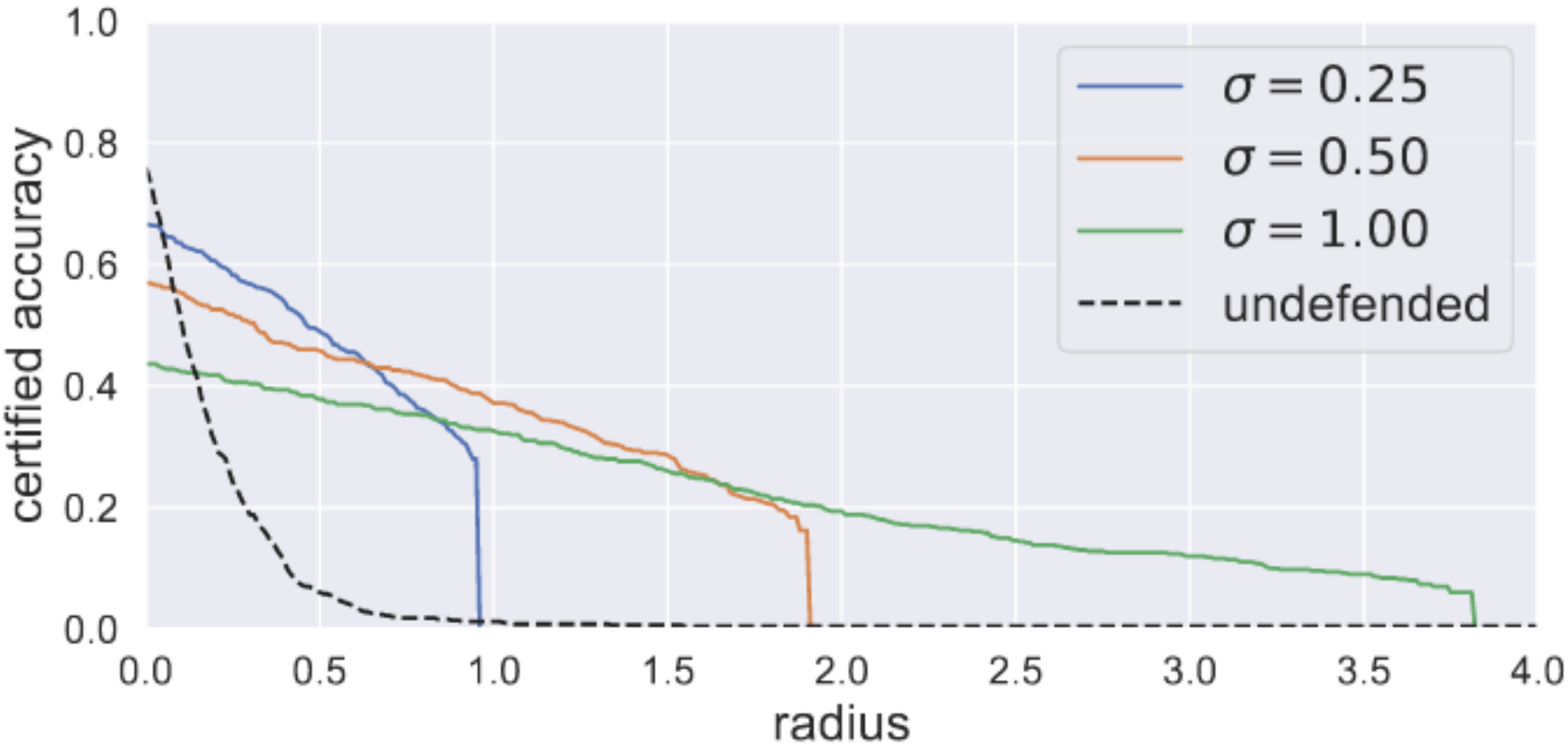
Two ways to evaluate robustness:

1. Construct a proof of robustness
2. Demonstrate constructive attack

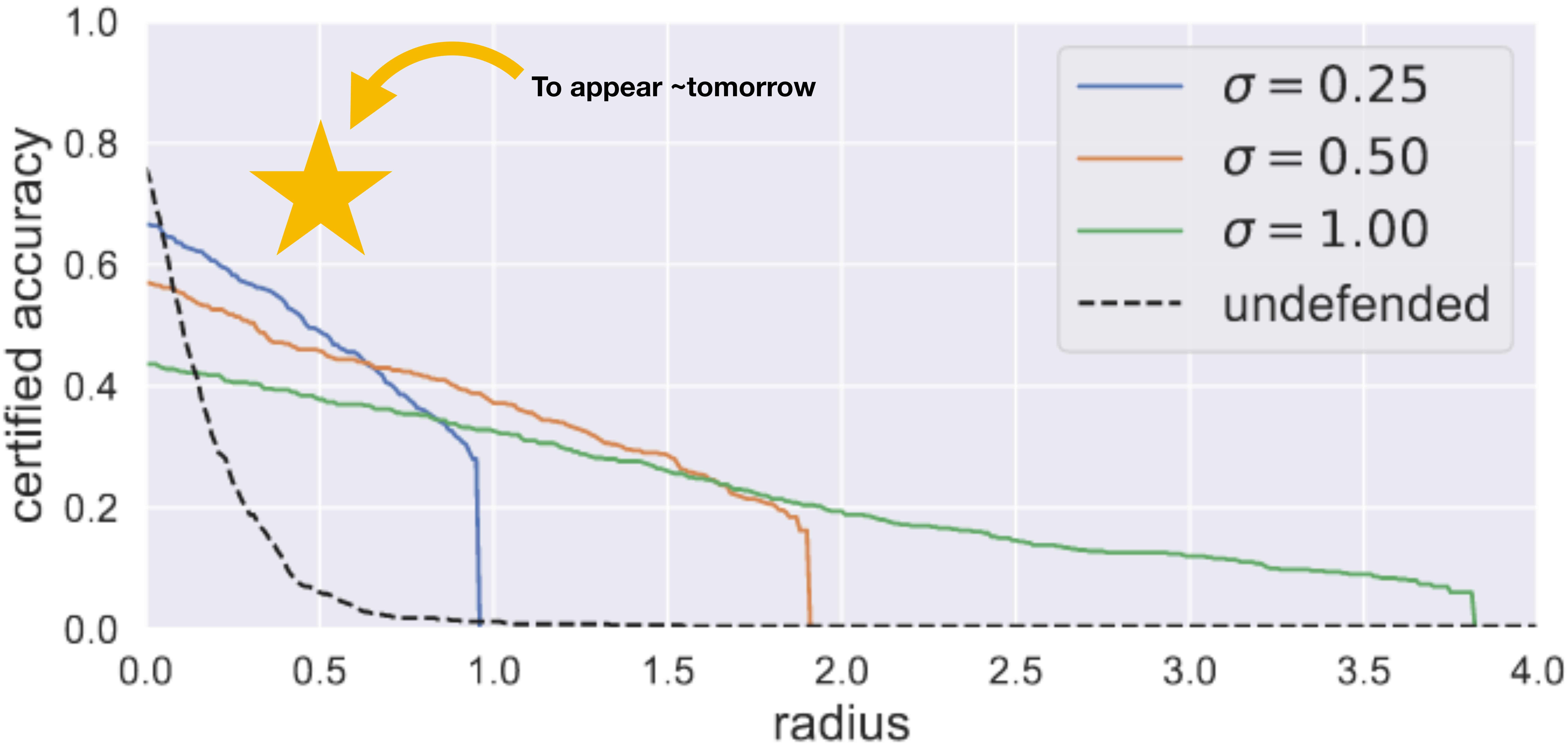
31

Towards Evaluating the Robustness of Neural Networks

Nicholas Carlini



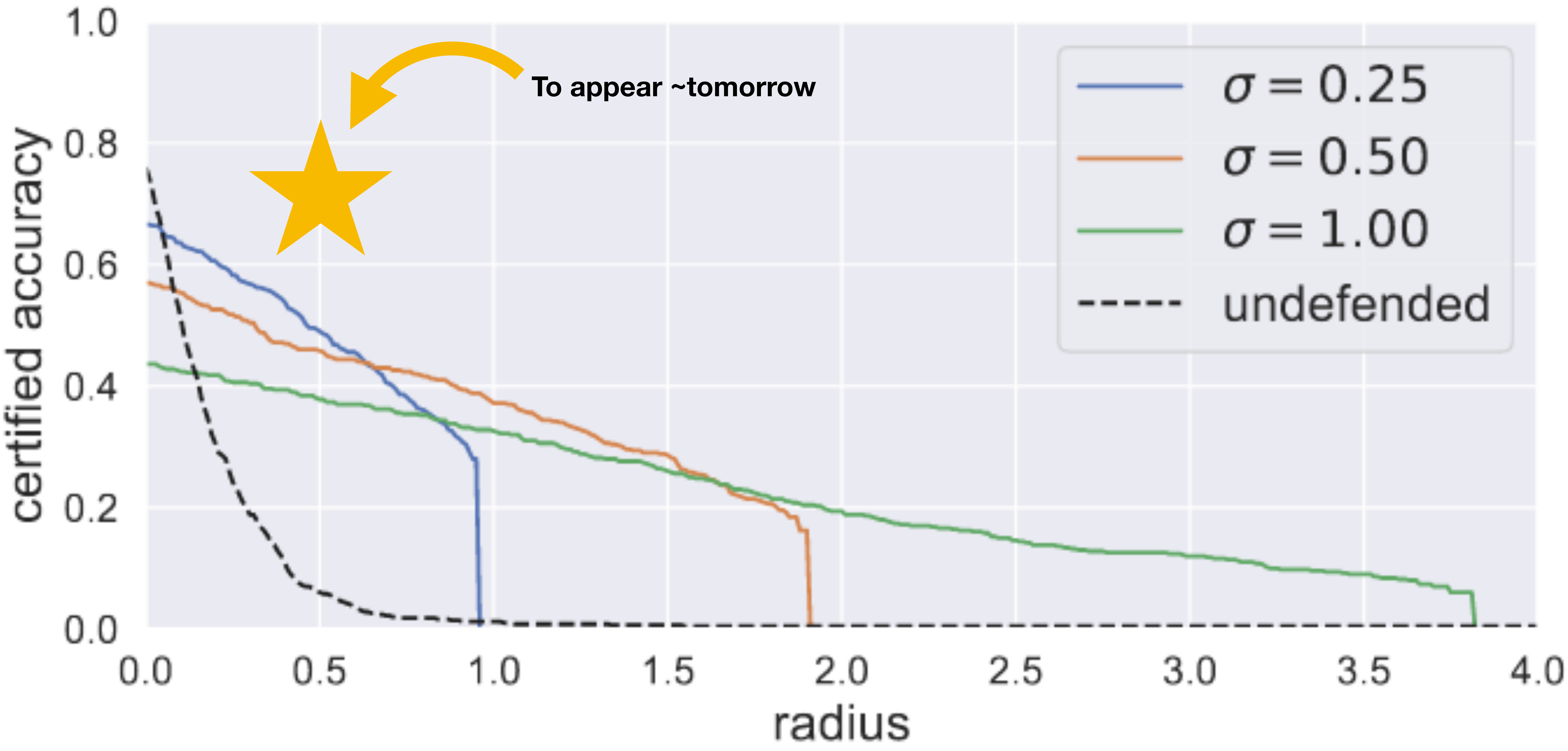
1.0
0.8
0.6
0.4
0.2
0.0
certified accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
radius
$\sigma = 0.25$
$\sigma = 0.50$
$\sigma = 1.00$
undefended


To appear ~tomorrow
σ = 0.25
σ = 0.50
σ = 1.00
undefended
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

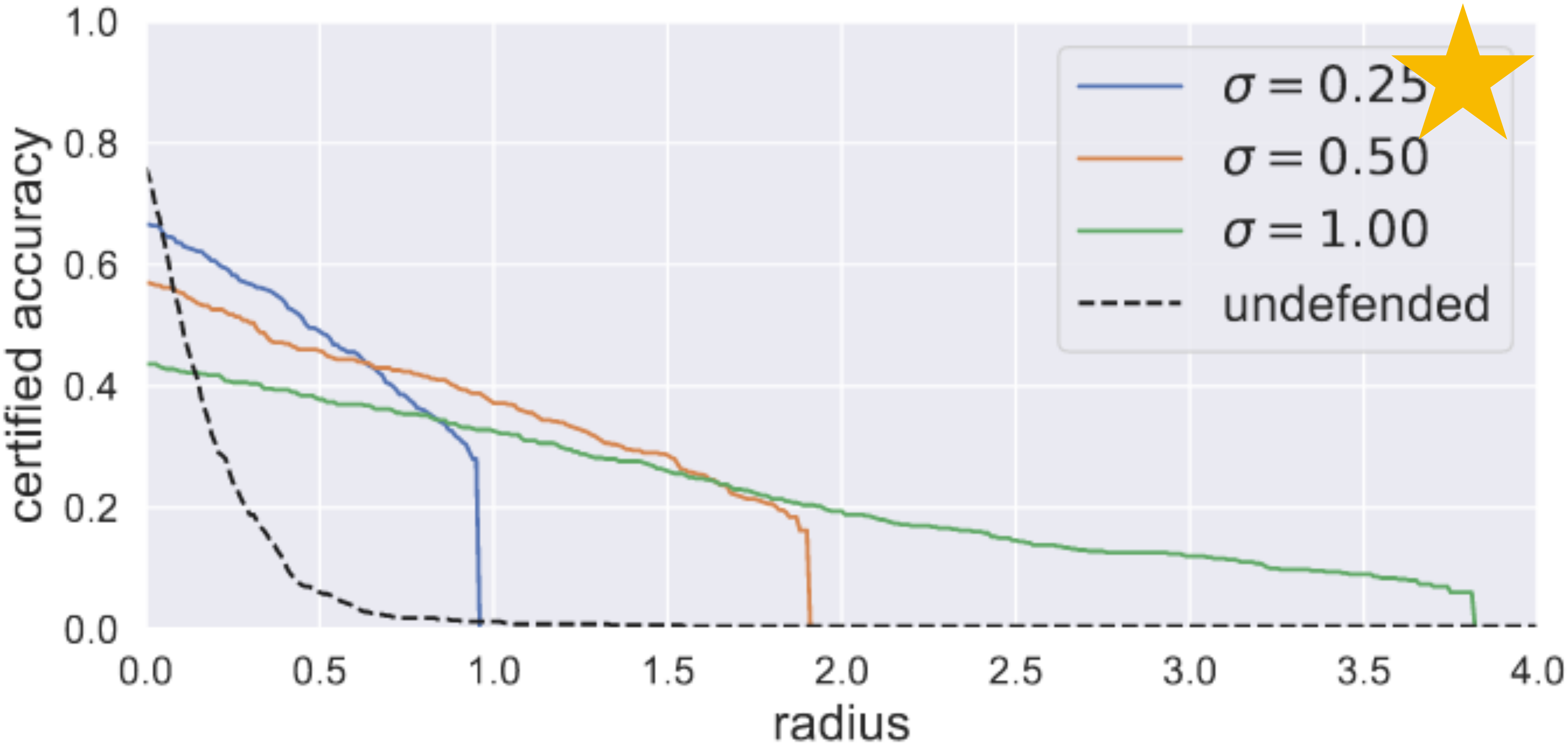
Who would win?

Six years of researchers training the best adversarially robust neural networks

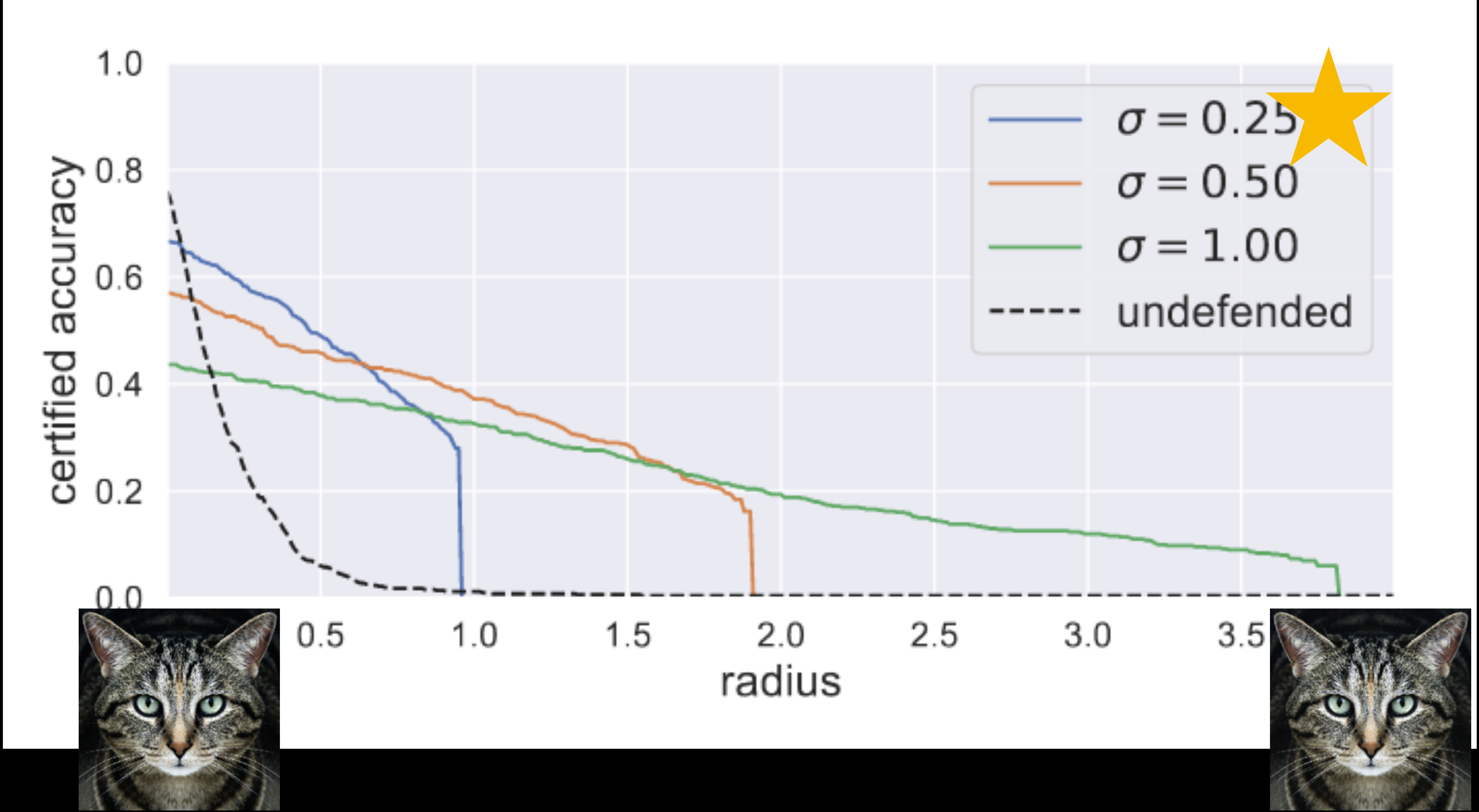
One diffusion model



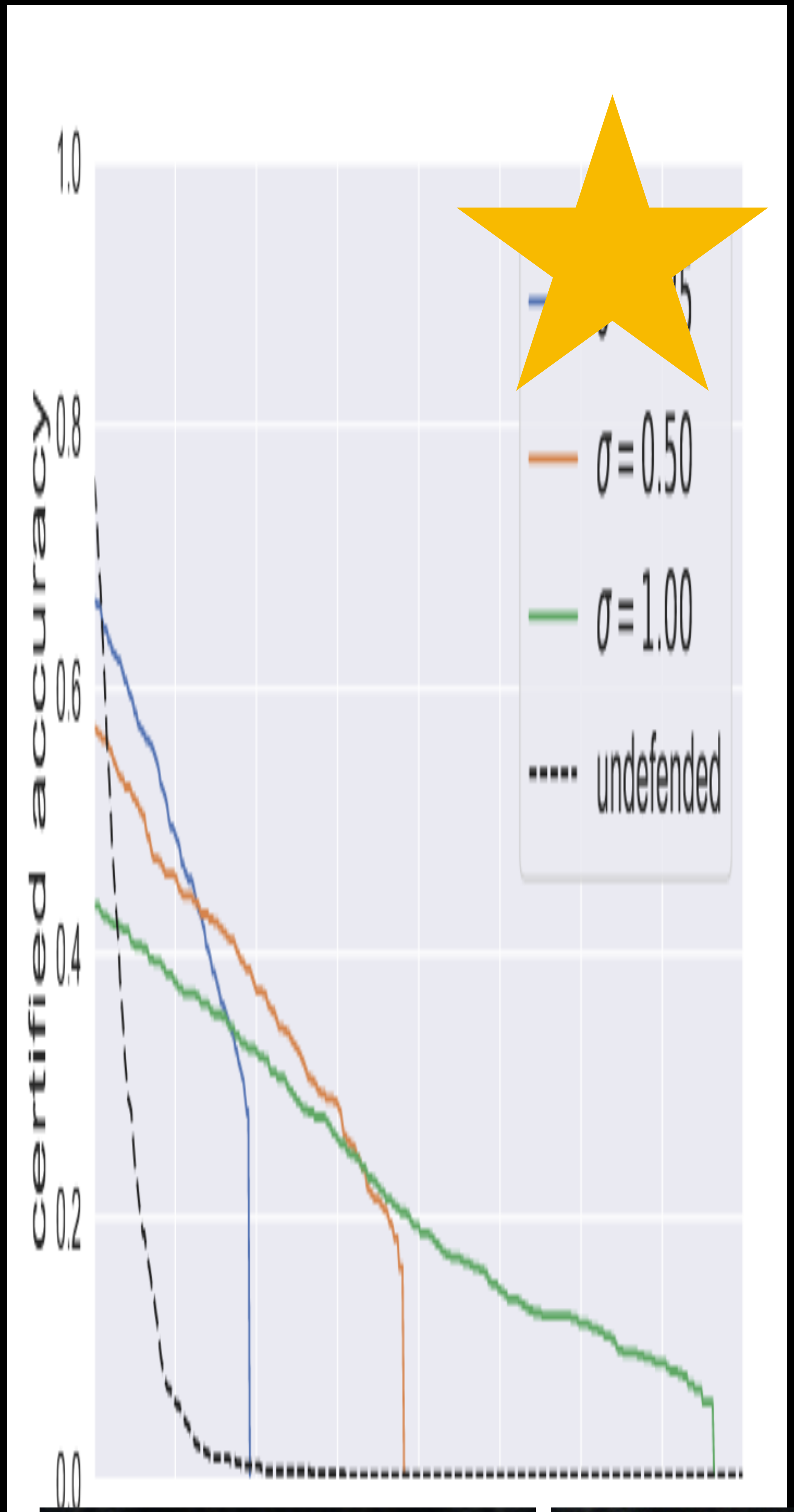
To appear ~tomorrow
σ = 0.25
σ = 0.50
σ = 1.00
undefended
certified accuracy
radius
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0



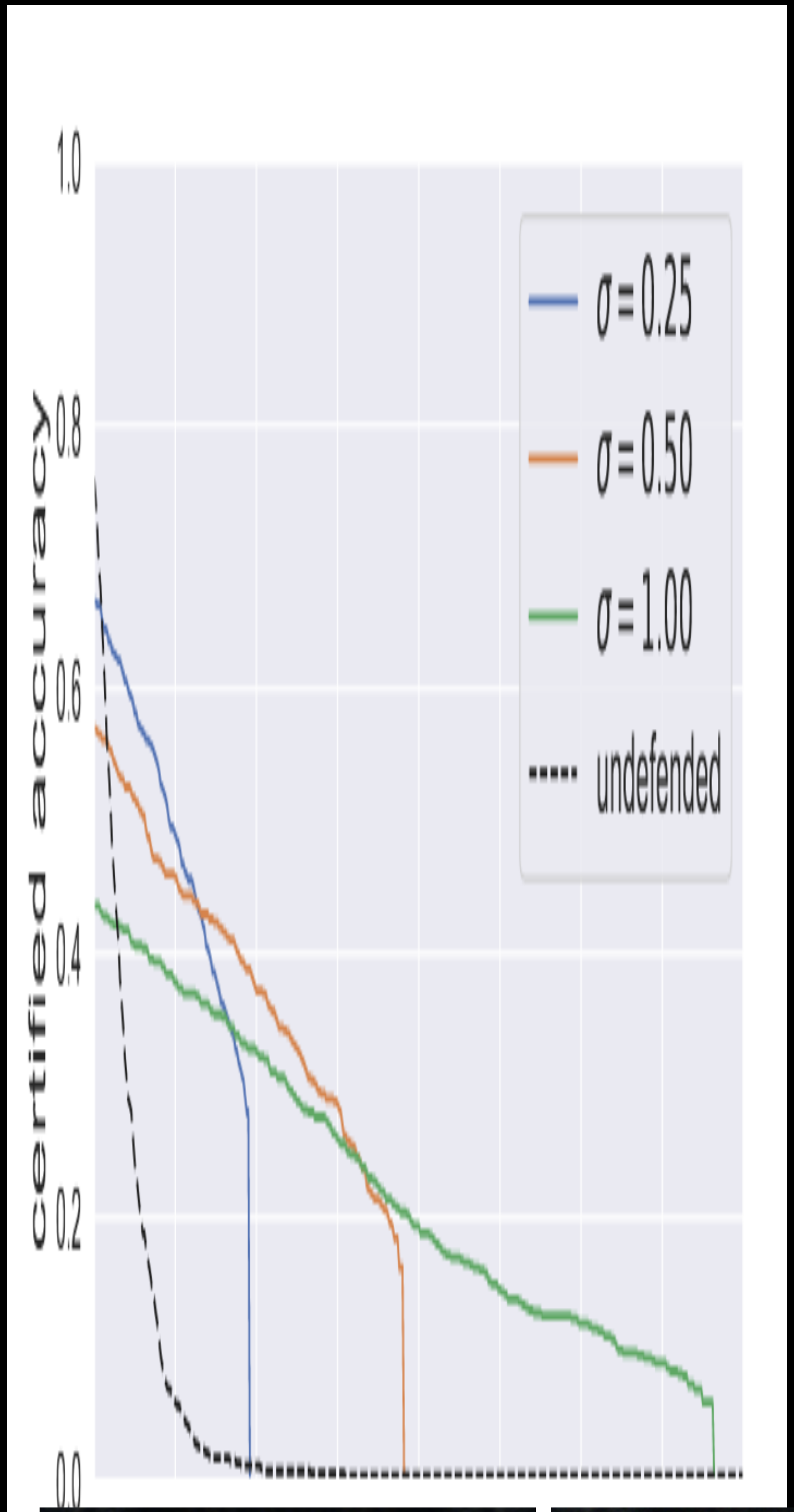
certified accuracy
radius
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
$\sigma = 0.25$
$\sigma = 0.50$
$\sigma = 1.00$
undefended



1.0
0.8
0.6
0.4
0.2
0.0
certified accuracy
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



$L_2 = 75$


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

$L_2 = 75$

Original

L_2 distortion: 75

L_2 distortion: 75

We might want to improve ...

1. ~~General purpose robustness~~

2. ~~The robustness against worst-case attack~~

3. The robustness against practical attacks

POLICY FORUM

MACHINE LEARNING

Adversarial attacks on medical machine learning

Emerging vulnerabilities demand new conversations

By **Samuel G. Finlayson[1], John D. Bowers[2], Joichi Ito[3], Jonathan L. Zittrain[2], Andrew L. Beam[4], Isaac S. Kohane[1]**

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, *IEEE* and M. Shamim Hossain, Senior Member, *IEEE,* Nabil A. Alrajeh, Fawaz Alsolami

Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

c Lin Gu [d] Yisen Wang [e] Yitian Zhao [f] James Bailey [b] Feng Lu [**, a, c]

Technology and Systems, School of CSE, Beihang University, Beijing, China.
rmation Systems, The University of Melbourne, Parkville, VIC 3010, Australia.
enter for Big Data-Based Precision Medicine, Beihang University, Beijing, China.
National Institute of Informatics, Tokyo 101-8430, Japan.
[e]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.
[f]Cixi Instuitue of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, Ningbo, China.

Advers

learning

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou[1[0000−0001−5127−3170]], Konstantinos Sechidis[1[0000−0001−6582−7453]], James Weatherall[2], and Gavin Brown[1]

[1] School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{konstantinos.papangelou,konstantinos.sechidis, gavin.brown}@manchester.ac.uk
[2] Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK
james.weatherall@astrazeneca.com

and Robust Machine Learning for Healthcare: A Survey

Qayyum[1], Junaid Qadir[1], Muhammad Bilal[2], and Ala Al-Fuqaha[3*]

ormation Technology University (ITU), Punjab, Lahore, Pakistan
niversity of the West England (UWE), Bristol, United Kingdom
[3] Hamad Bin Khalifa University (HBKU), Doha, Qatar

Who even is the adversary here?


Feedback
English (US)
Submit a request
Sign in
Discord > Discord Interface > Direct Messaging
Search
Articles in this section
Discord Safety: Safe Messaging!
Discord Direct Messages (DMs) are a great way to instant message your buddies with the latest gossip or silliest memes.
To keep your DMs clean and prevent any unwarranted surprises at bay, Discord has a few extra levers you can pull. While we're still building out a few of these options, if you open your user settings tab and select the Privacy & Safety option, you'll see the "Safe Direct Messaging" option!
USER SETTINGS
My Account
Privacy & Safety
Authorized Apps
Connections
Billing
Subscriptions
Gift Inventory
Server Boost
HypeSquad
PRIVACY & SAFETY
SAFE DIRECT MESSAGING
Automatically scan and delete direct messages you receive that contain explicit media content.
Keep me safe
Scan direct messages from everyone.
My friends are nice
Scan direct messages from everyone unless they are a friend.
Do not scan
Direct messages will not be scanned for explicit content.
edia Uses
built from a model of openly
s so bad that the number of
er month—had fallen by 40
not one solution to combat this
Wikipedia, decided to
and consider ways to combat it.
SafeSearch on
Hide explicit results
More about SafeSearch
SafeSearch: Moderate
Filter
Strict
Moderate (default)
Off
Safe search: moderate
Any
Strict
Moderate
Off

Under the skin of OnlyFans

By Rianna Croxford
Correspondent, BBC News

17 July 2021

Under the skin of OnlyFans

By Rianna Croxford
Correspondent, BBC News

17 July 2021

In a statement, OnlyFans said the account did not have two-factor authentication, which made it vulnerable. The company said Tina did not report the racial slur and it was not detected by the site's moderation system because it was pluralised.

We might want to improve ...

1. ~~General purpose robustness~~

2. ~~The robustness against worst case attack~~

3. The robustness against practical attacks

we still have a chance!

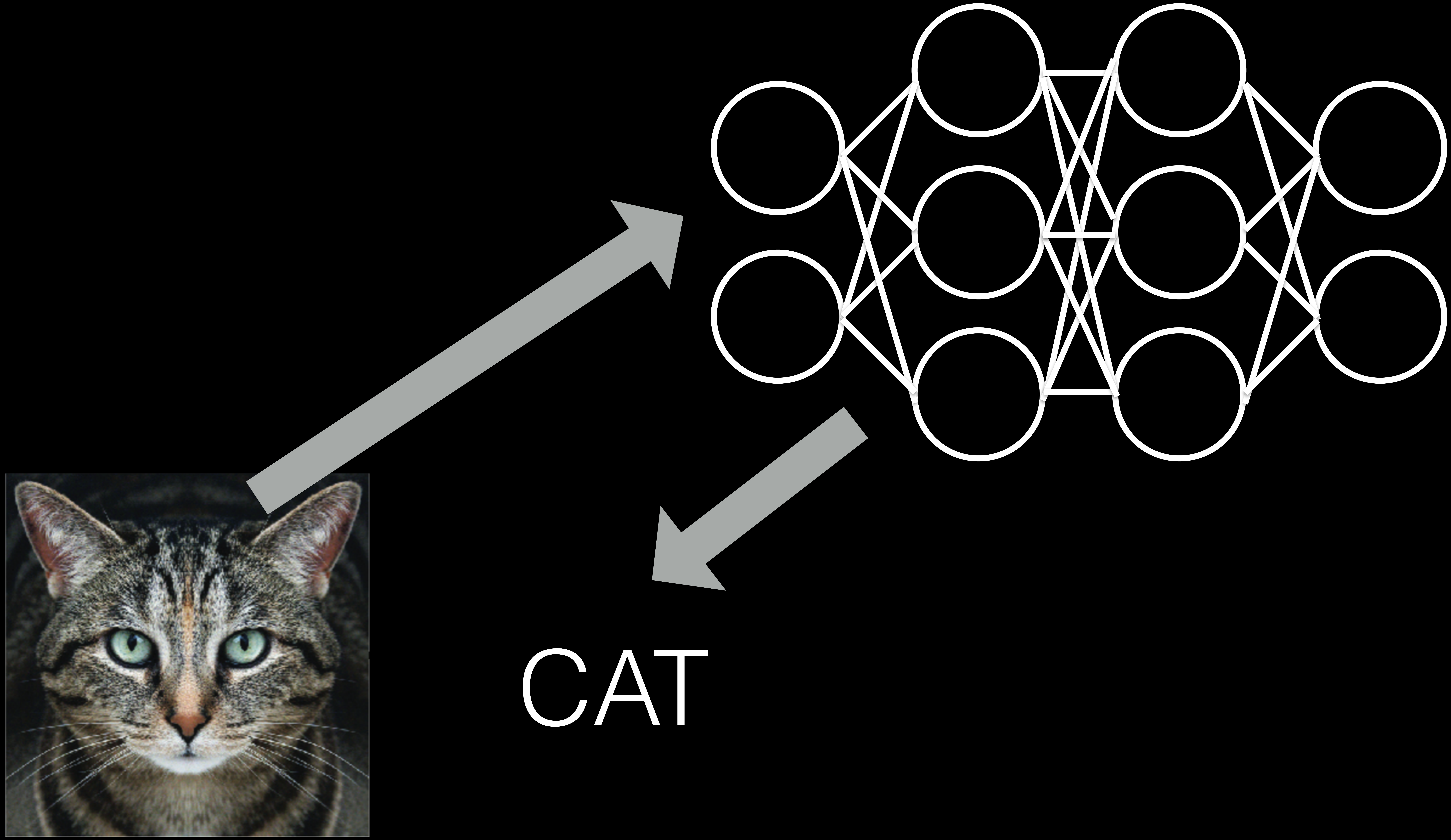
2019

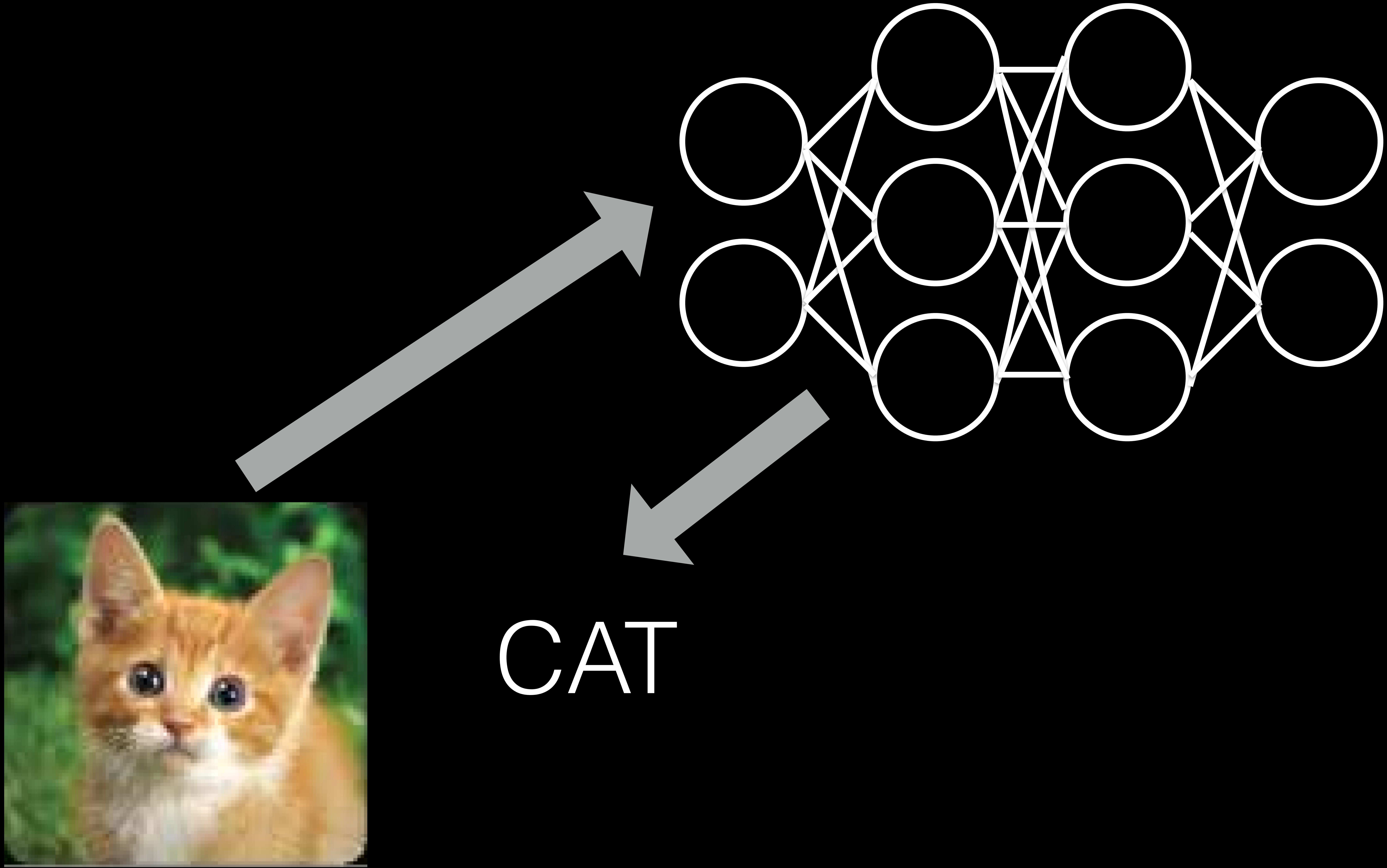
Stateful Detection of Black-Box Adversarial Attacks

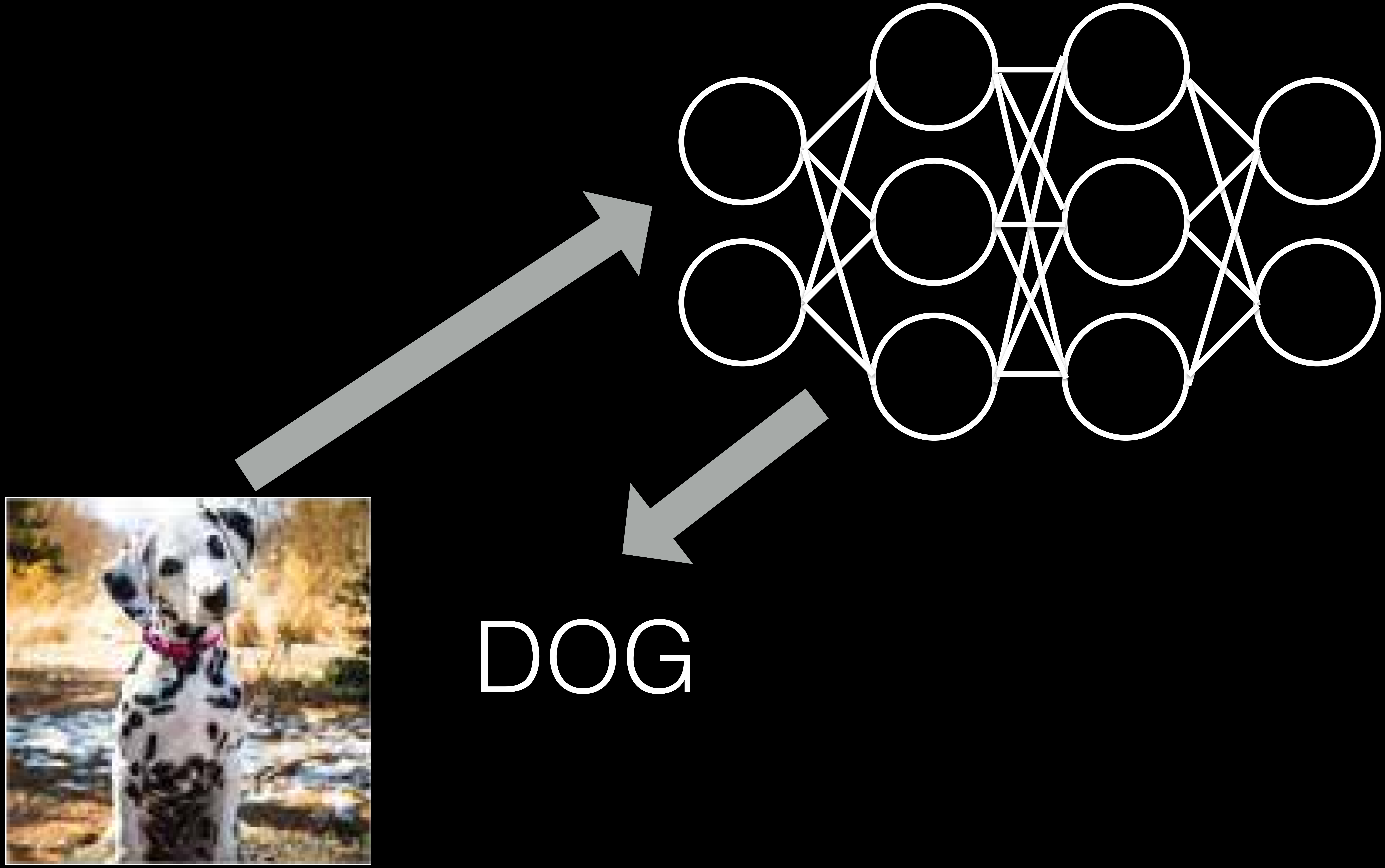
Steven Chen
University of California, Berkeley

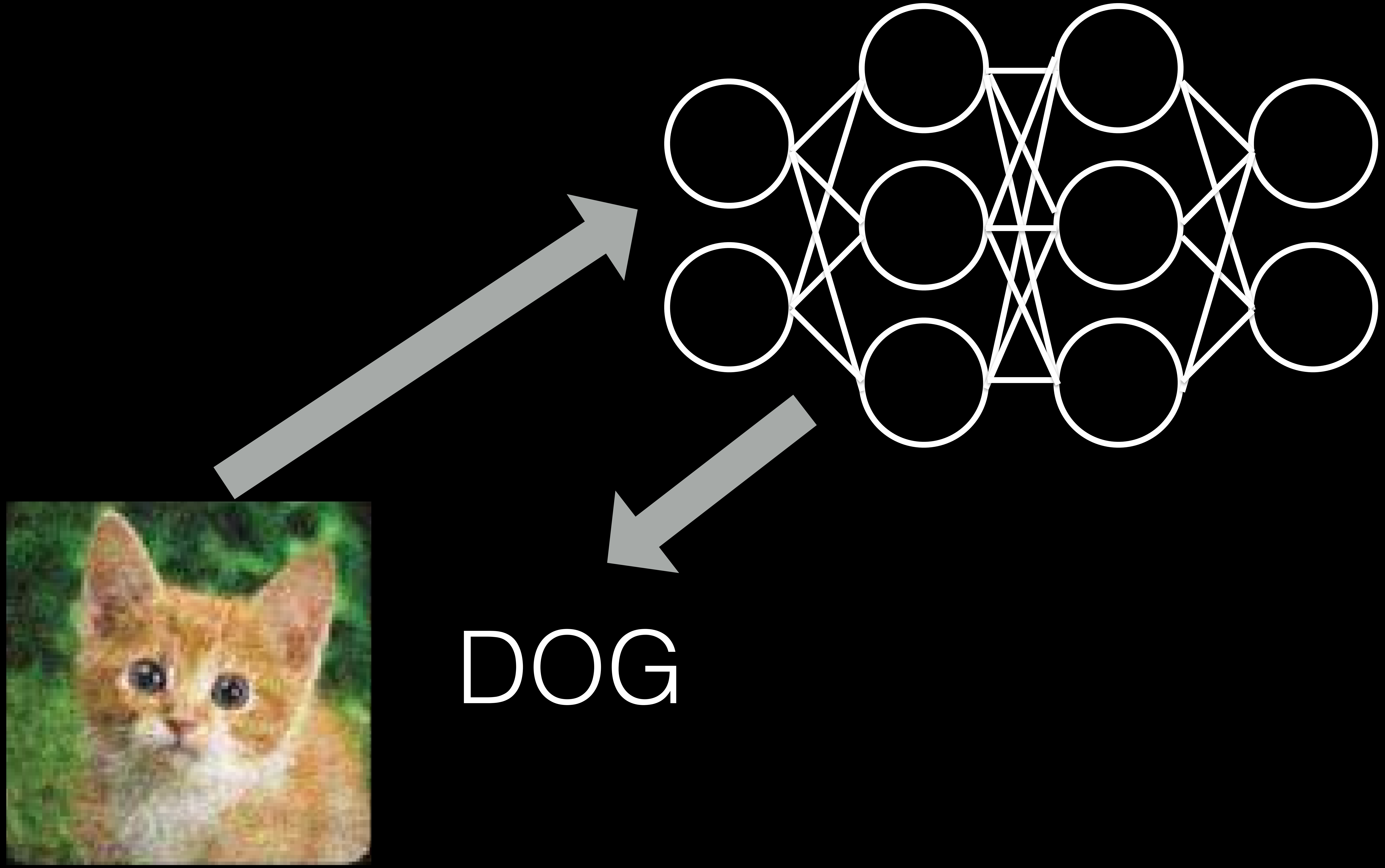
Nicholas Carlini
Google Research

David Wagner
University of California, Berkeley

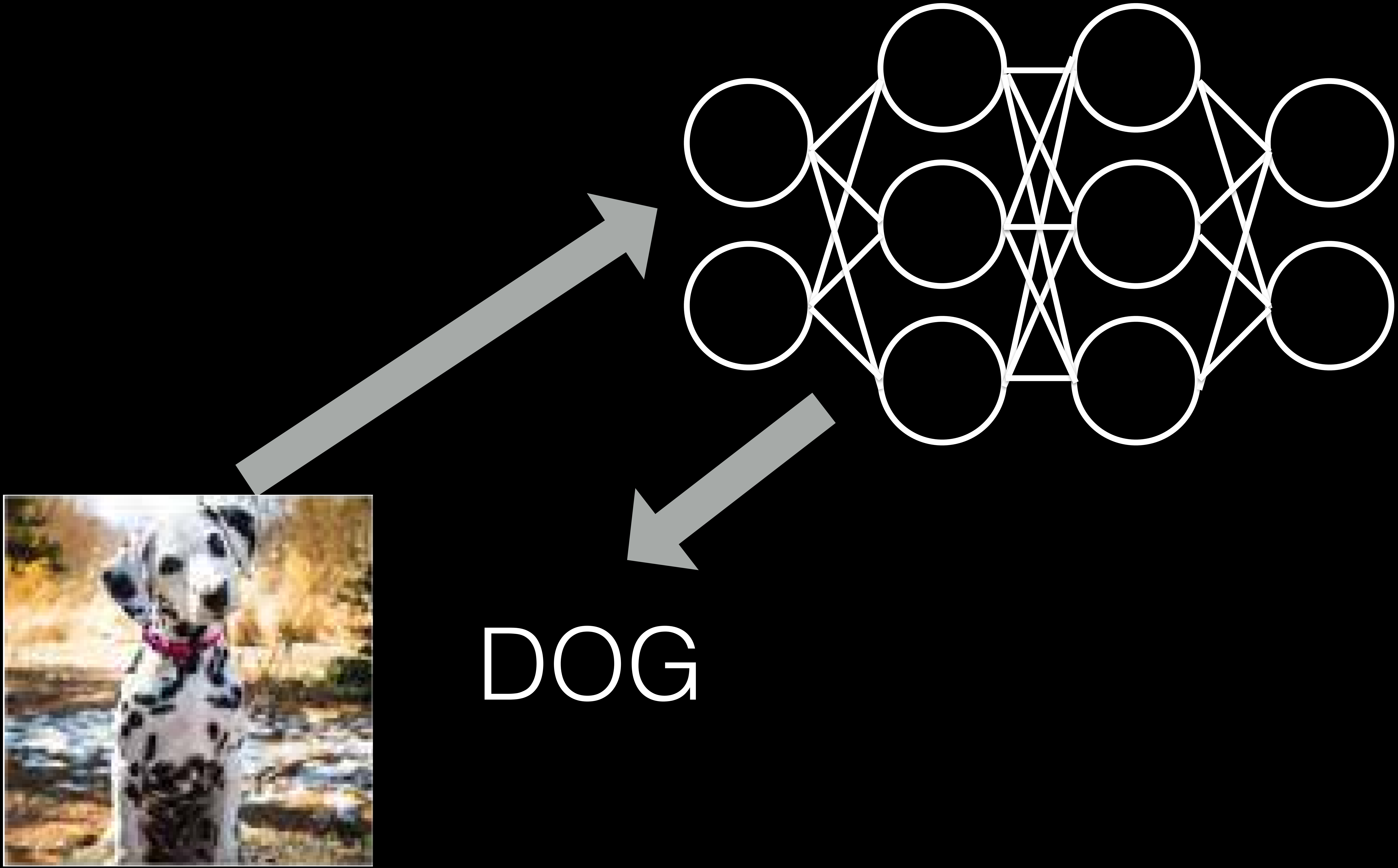

CAT

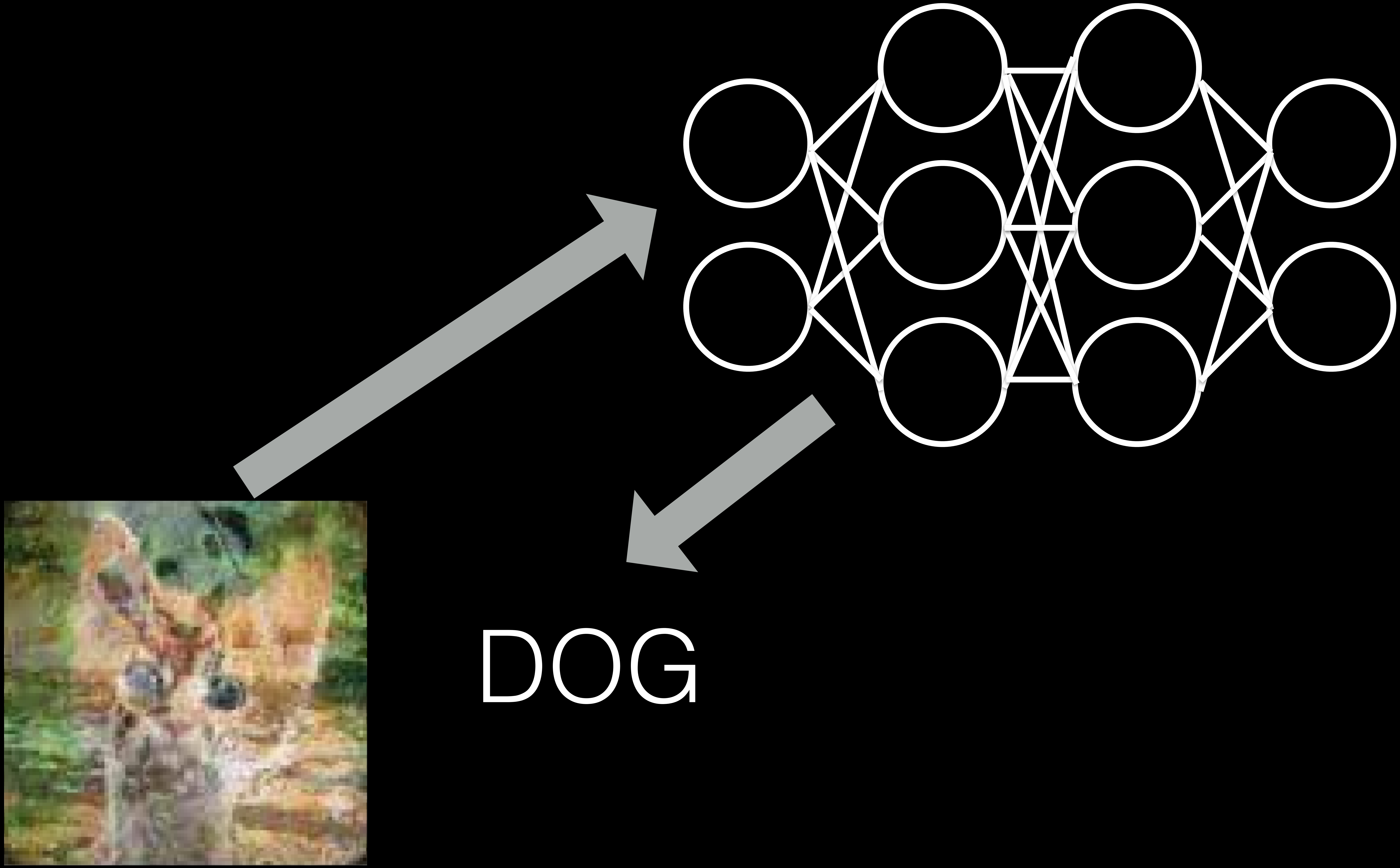

CAT

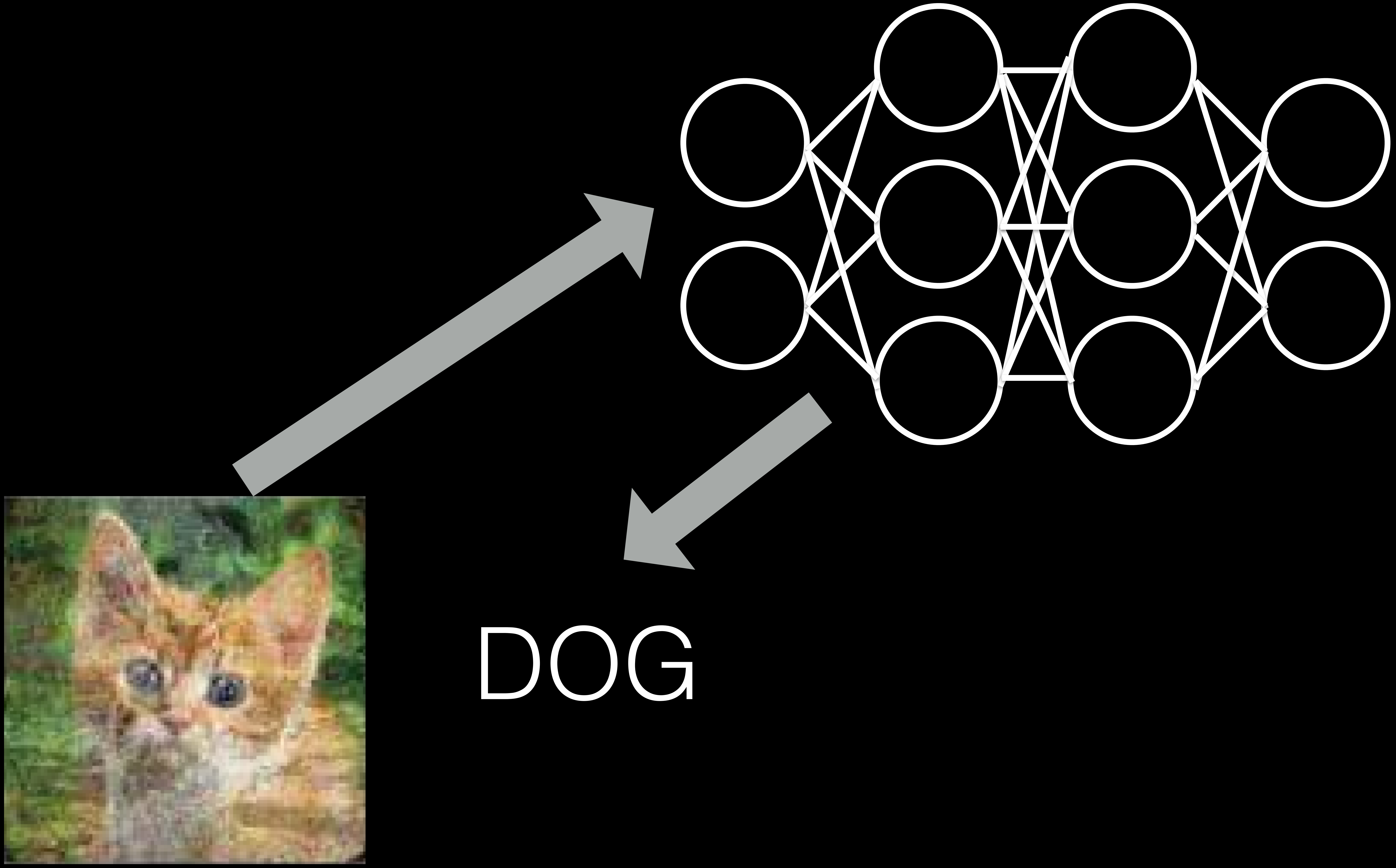

DOG

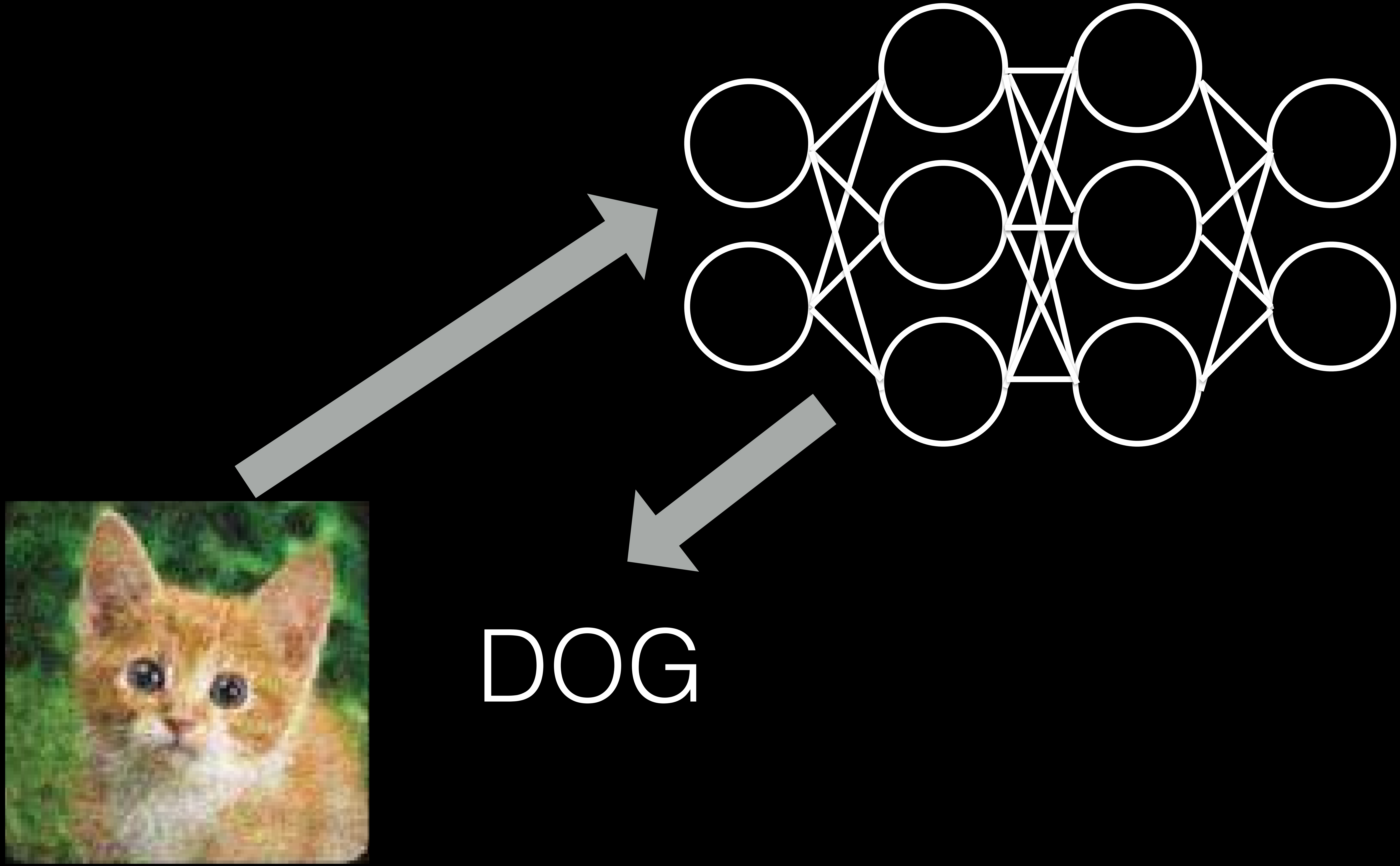

DOG

Under attack

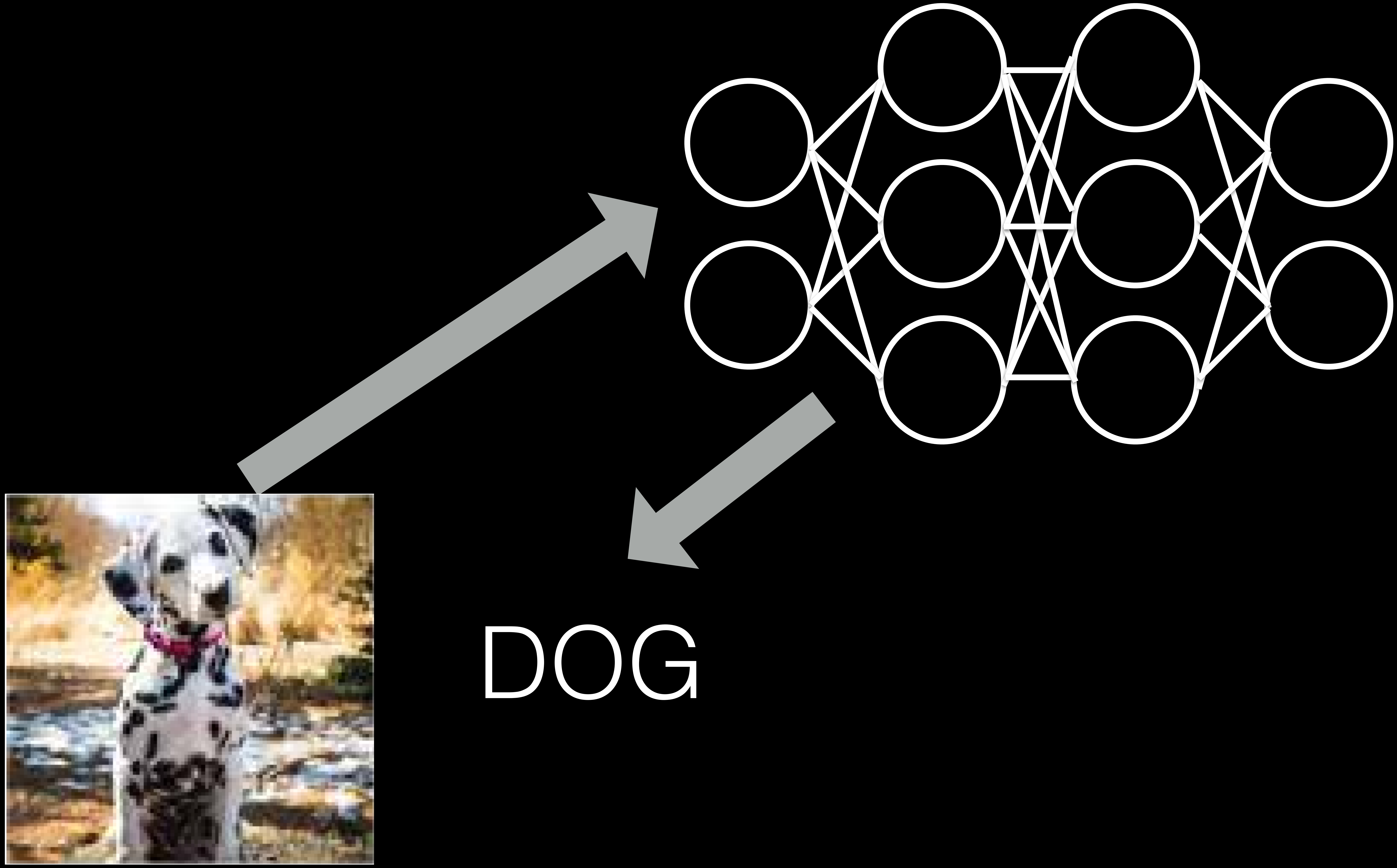

DOG

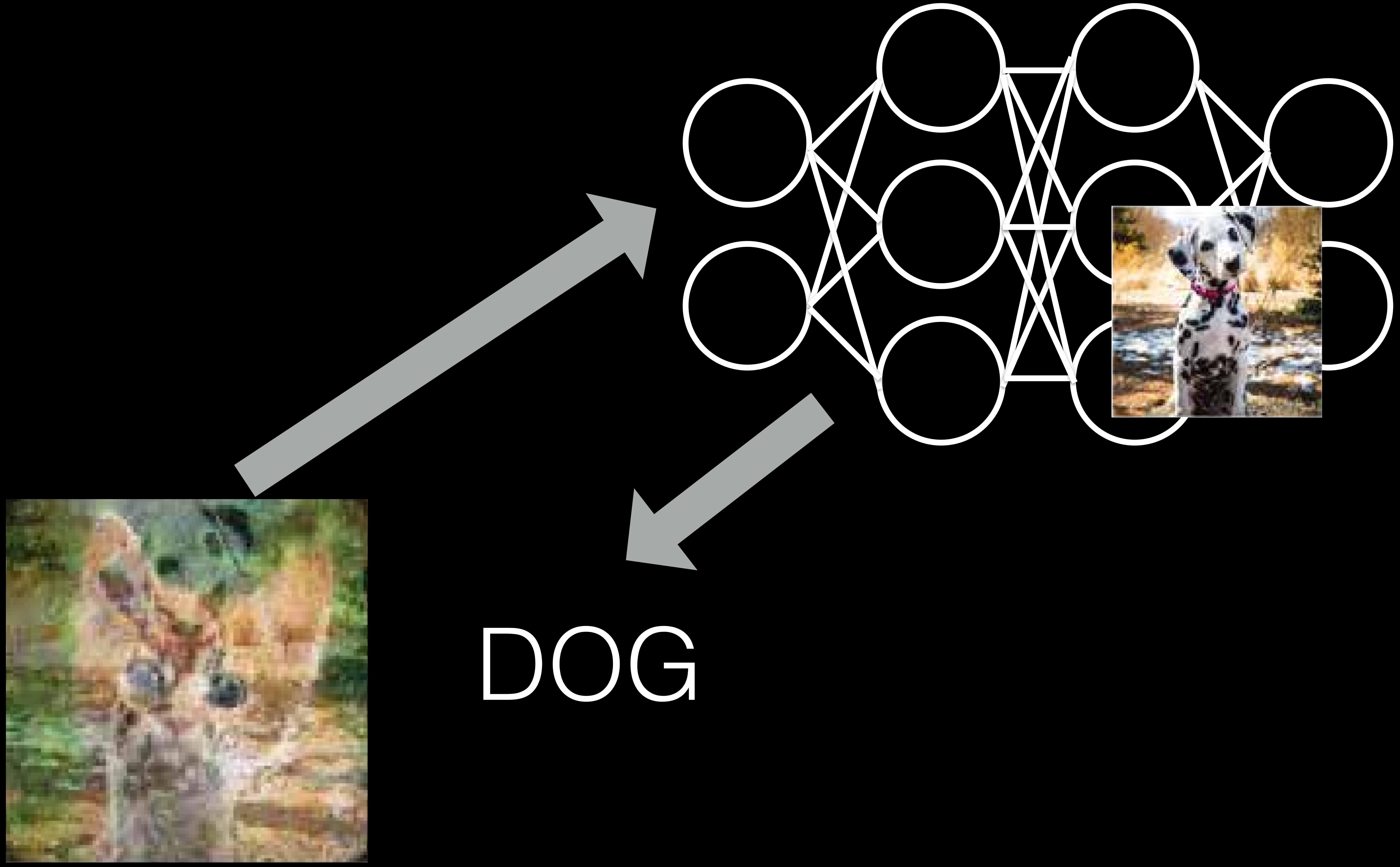

DOG

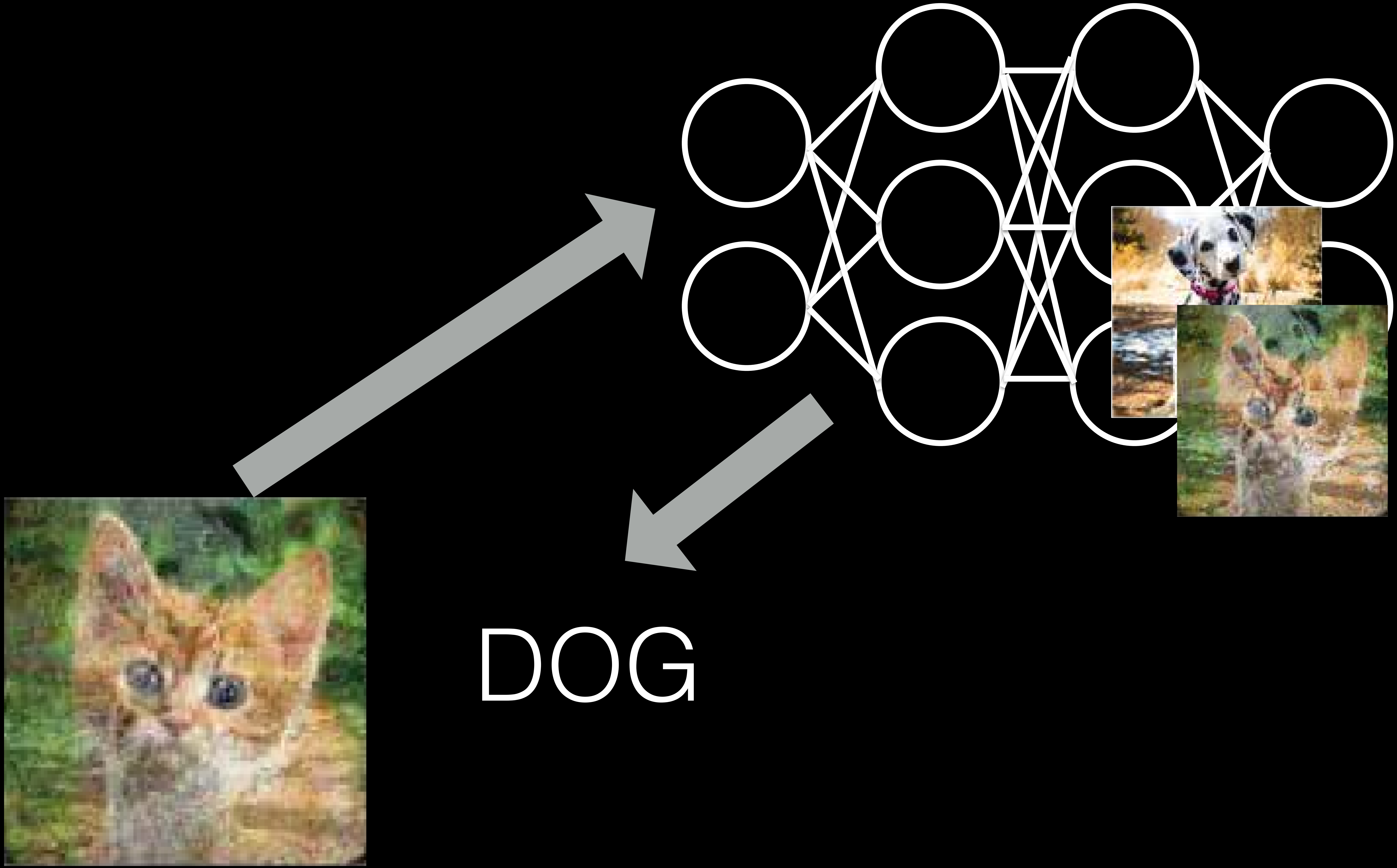

DOG

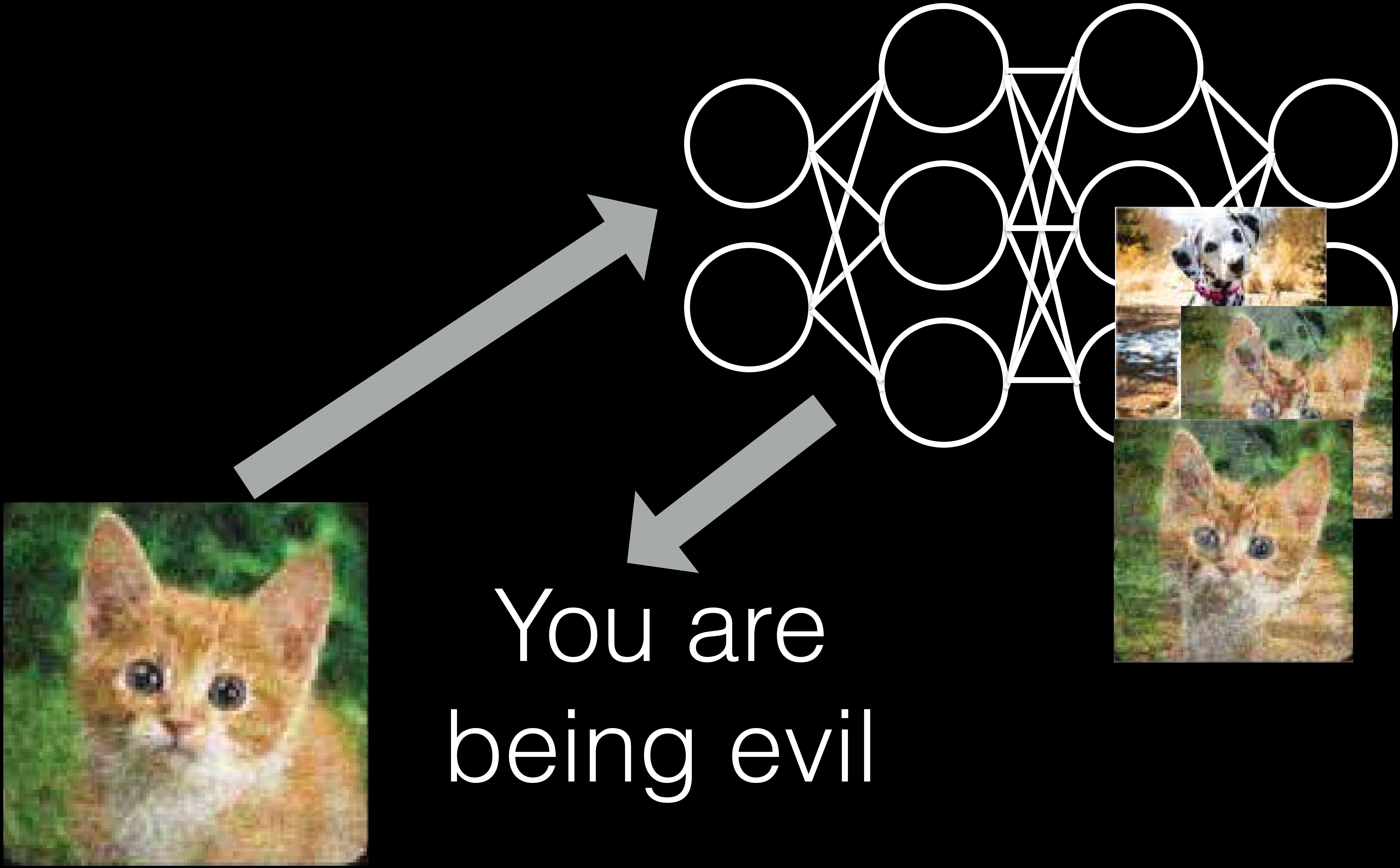

DOG

Our Defense


DOG


DOG


DOG



You are
being evil

Except here's the thing.

I don't believe this defense actually works.

What I want:

More attacks and defenses on practical systems.

We might want to improve ...

1. ~~General purpose robustness~~

2. ~~The robustness against worst case attack~~

3. The robustness against practical attacks

we still have a chance!