Identifying, Analyzing, and Addressing Weaknesses in Deep Networks

Foundations for Conceptually Sound Neural Networks

Klas Leino

Committee

Matt Fredrikson, Chair
Anupam Datta
J. Zico Kolter
Corina Păsăreanu
Kamalika Chaudhuri, UCSD

Machine Learning is Everywhere. How Good is it?

Achieving Super-human Level Accuracy in Computer-vision Tasks: Learning by revising

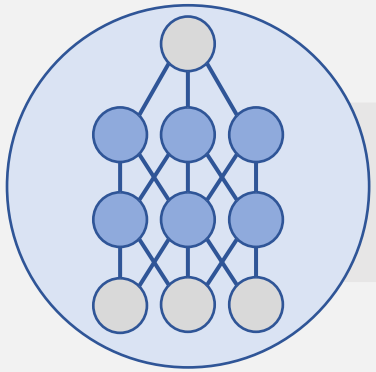
Delving Deep into Rectifiers:
Surpassing Human-Level Performance on ImageNet Classification

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research

@microsoft.com

Understanding Neural Network Predictions



model

human

dog

dog

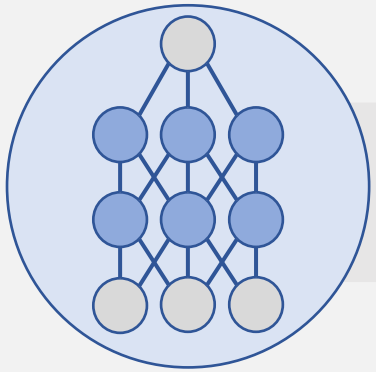
why dog?

dog

cat

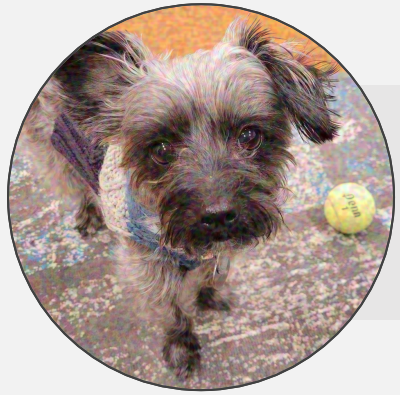
deep networks may not be **conceptually sound**

Human-Level Performance?



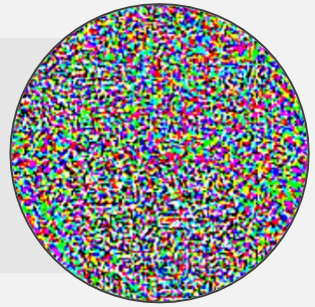
model

human



=

+



dog

dog

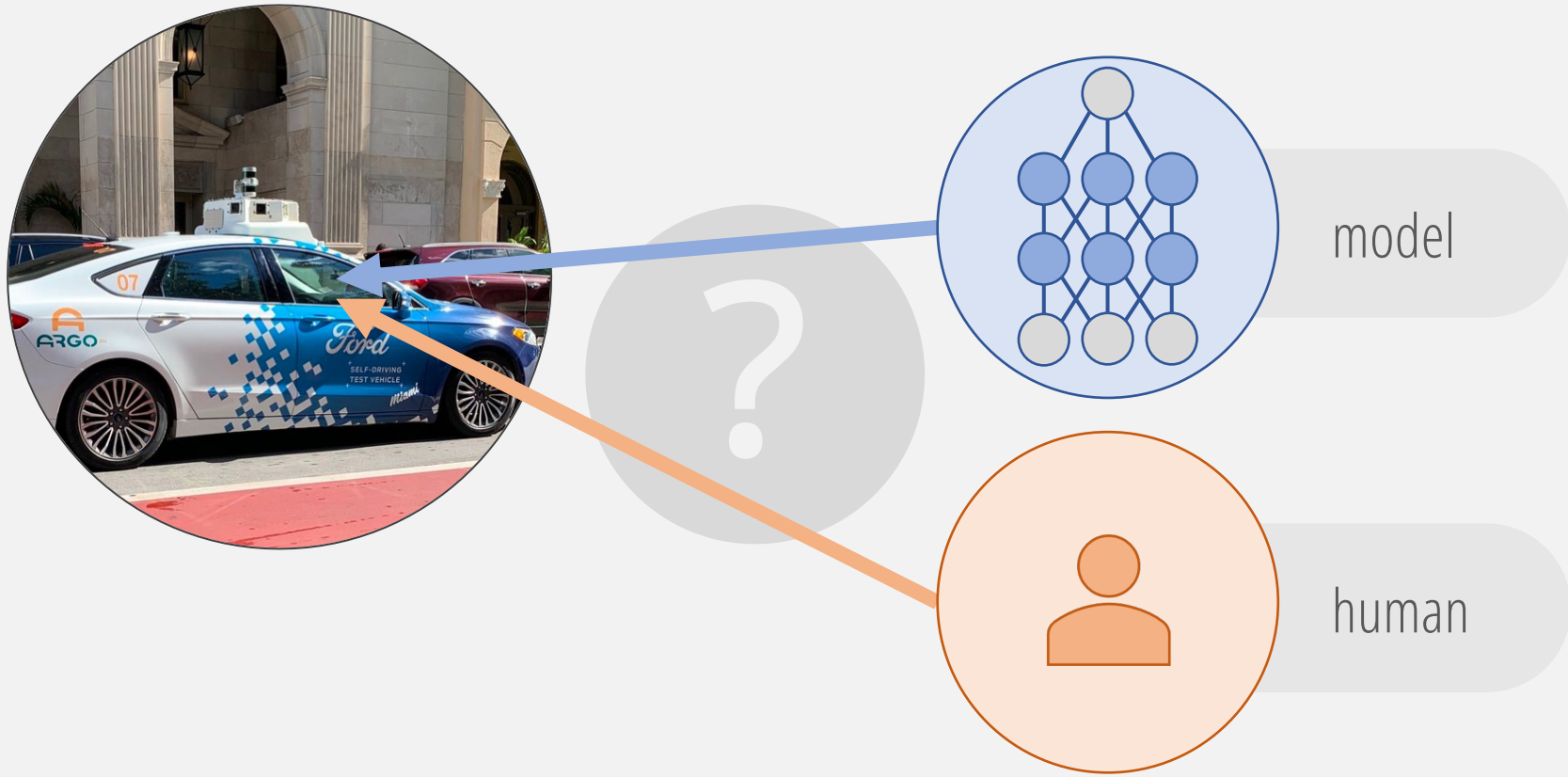
sports car

dog

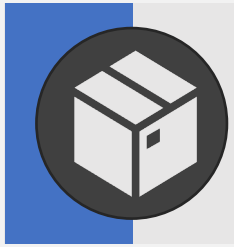
important features?

deep networks are easily fooled

Safety-Critical Settings

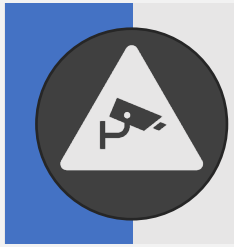


Weaknesses in Deep Learning



Lack of transparency

Lack of robustness to adversarial manipulations



Privacy leakage

Excessive bias & miscalibration

Others to be discovered…

Goals

understand the **weaknesses and vulnerabilities** of deep networks

take steps towards improving the **conceptual soundness** of deep networks

My Work Covered in this Thesis

I Assessing Conceptual Soundness

Influence Directed Explanations for Convolutional Neural Networks. ITC 2018

II Training Robust Neural Networks

Globally Robust Neural Networks. ICML 2021

Relaxing Local Robustness. NIPS 2021

III Other Weaknesses & Vulnerabilities

Feature-wise Bias Amplification. ICLR 2019

Leveraging Model Memorization for Calibrated White-box Membership Inference. USENIX 2020

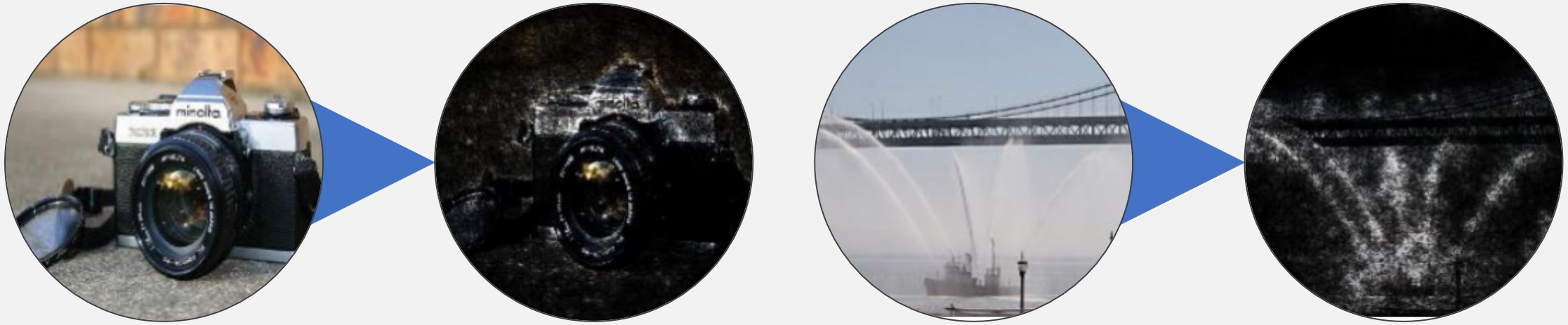
Robust Models Can Leak Instances and Their Properties. IEEE S&P 2022 (submitted)

Overview

- Assessing Conceptual Soundness
  - Explanations for deep networks
  - Influence-directed explanations
  - What makes a “good” explanation?
- Improving Conceptual Soundness via Robustness
- Other Weaknesses and Vulnerabilities

Understanding Neural Networks via Importance Measures

A importance measure helps explain how a model uses its features by assigning a value to each feature x_i according to how important x_i was to the outcome of the model



Sundararajan et al. 2017

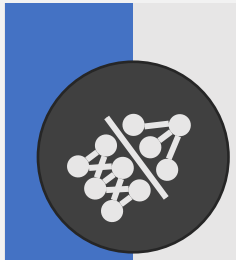
Understanding Neural Networks via Importance Measures

The flexibility to accurately answer a wide range of queries is key to gaining a better understanding of a model's behavior

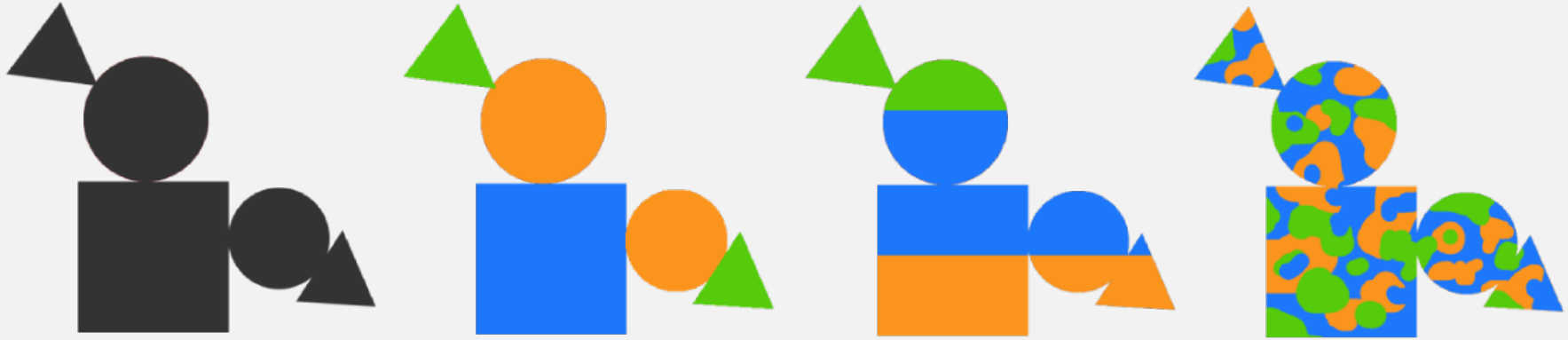
Contribution

We generalize other work on gradient-based importance methods with a broad explanation framework

Axes of Generalization: Slice



Slice
determines which internal layer(s) of the network we would like to expose and compute attribution for



Axes of Generalization: Quantity of Interest

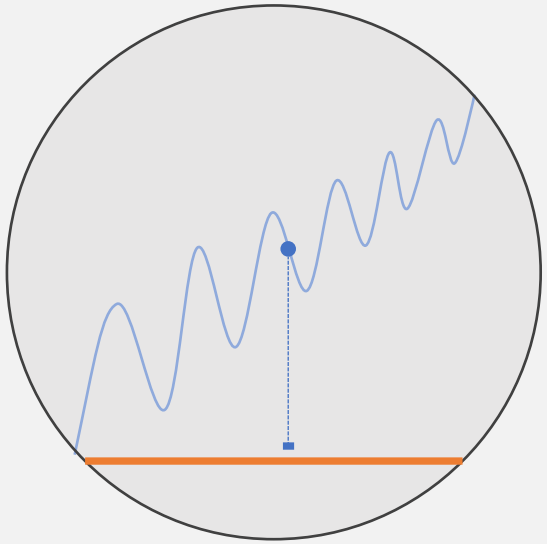
Quantity of Interest
specifies what specific network behavior we would like to explain

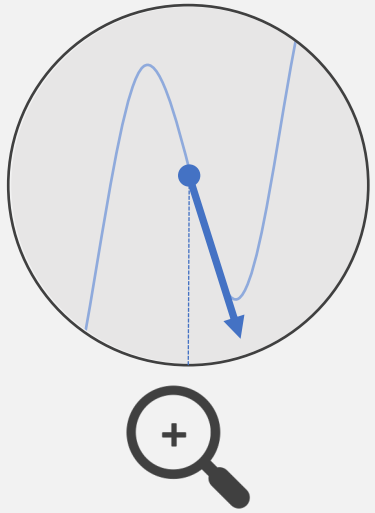
explaining, "why is this a sports car?"

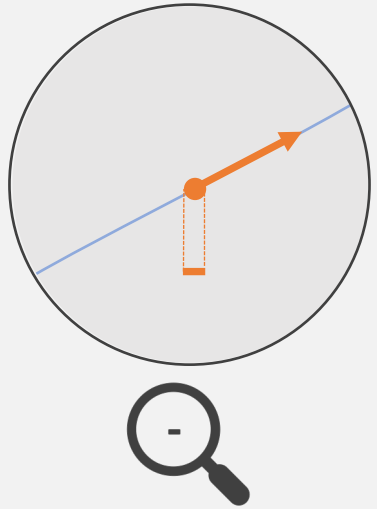
explaining, "why is this a sports car and *not* a convertible?"

Axes of Generalization: Distribution of Interest

Distribution of Interest
specifies which points we want our explanation to be faithful on







Internal Influence

Given a network that can be decomposed as $g \circ h$, a quantity of interest q, and a distribution of interest $\mathcal{D}$, the internal influence is

$$\chi(g \circ h, q, \mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{\partial (q \circ g)}{\partial h(x)} \right]$$

Leino et al., 2018

Overview

- Assessing Conceptual Soundness
  - Explanations for deep networks
  - Influence-directed explanations
  - What makes a “good” explanation?
- Improving Conceptual Soundness via Robustness
- Other Weaknesses and Vulnerabilities

Axiomatic Justification

Internal Influence is *axiomatically justified*.

Key Property: Linear Agreement

Attributions in Internal Influence extend the straightforward interpretation of inspecting the weights of linear models

Key Property: Faithfulness

Attributions in Internal Influence are *faithful* to the model

Evaluating Explanations

many prior works have evaluated explanations based on visual appeal:
e.g., *Are explanations intuitive*? *Do they highlight the "correct" objects*?

this assumes the network is conceptually sound to begin with!

we will see that conceptual soundness is **not** the norm

Overview

- Assessing Conceptual Soundness
- Improving Conceptual Soundness via Robustness
  - Adversarial examples and conceptual soundness
  - Globally robust neural networks
  - Limitations of local robustness
- Other Weaknesses and Vulnerabilities

Adversarial Examples

"We find that applying an **imperceptible** non-random perturbation to a test image, it is possible to arbitrarily change the network's prediction"

–Szegedy et al., 2014

adversarial examples are incompatible with conceptual soundness

start with natural image

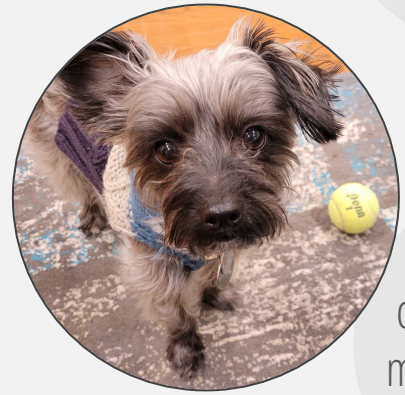
dog

make semantically meaningless change

$+ 0.007 \cdot$

sports car features?

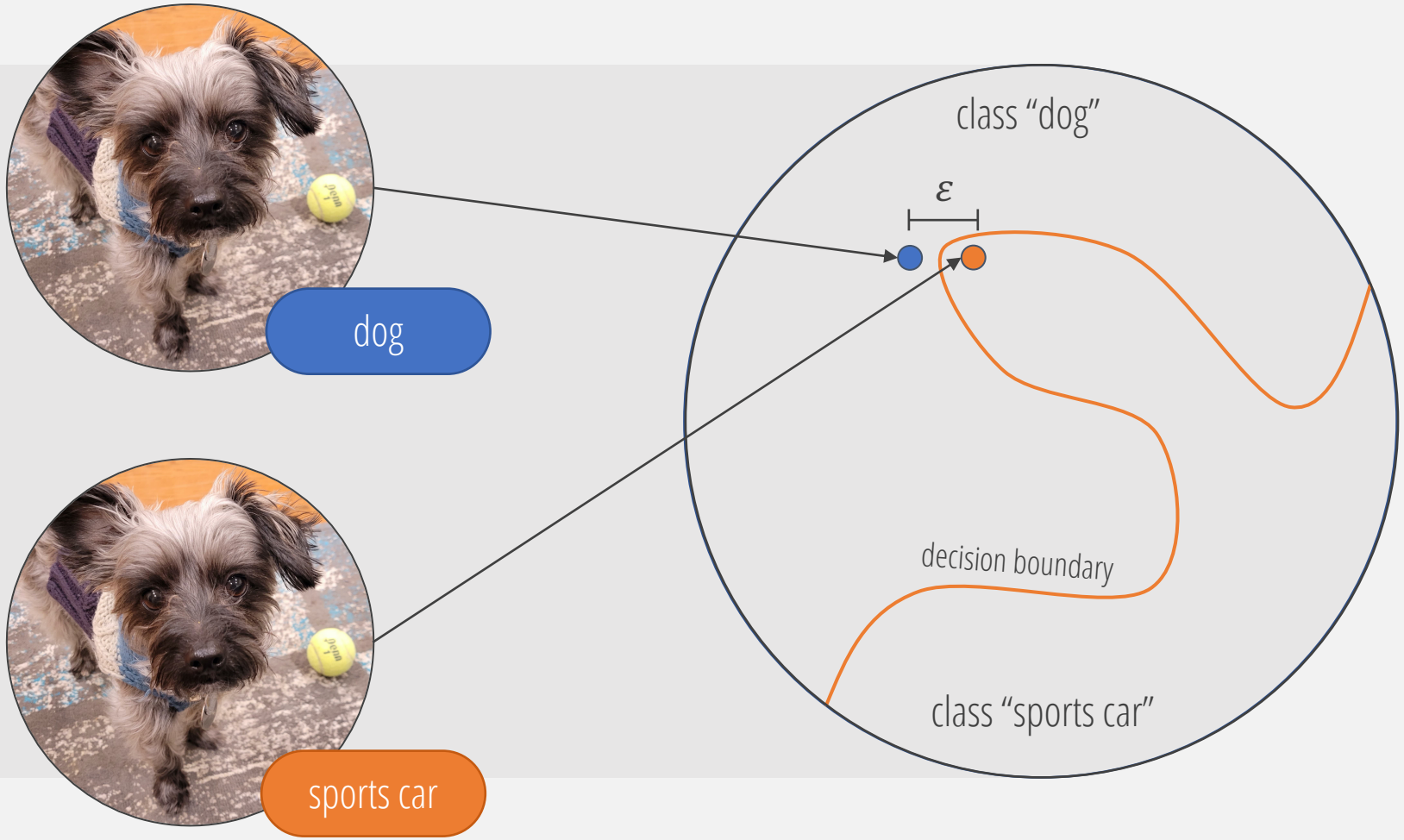
$=$



sports car

cause arbitrary misclassification

Small-norm Adversarial Examples



Local Robustness

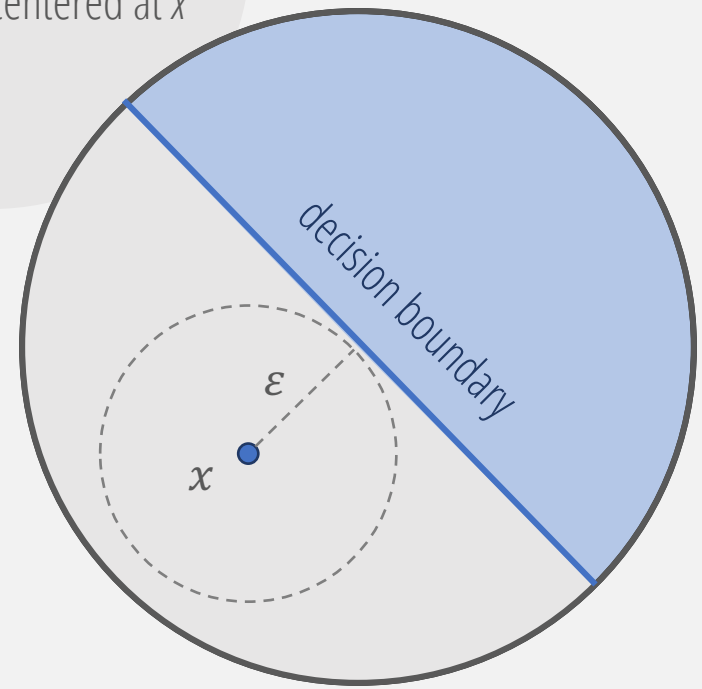
Definition

A classifier, F, is *ϵ-locally-robust* at x if $\forall x'$,

$$\underbrace{\|x - x'\| \leq \epsilon}_{\text{distance from } x \text{ smaller than } \varepsilon} \implies \underbrace{F(x) = F(x')}_{x' \text{ receives the same label as } x}$$

ℓ_p norm, typically $p \in \{1, 2, \infty\}$

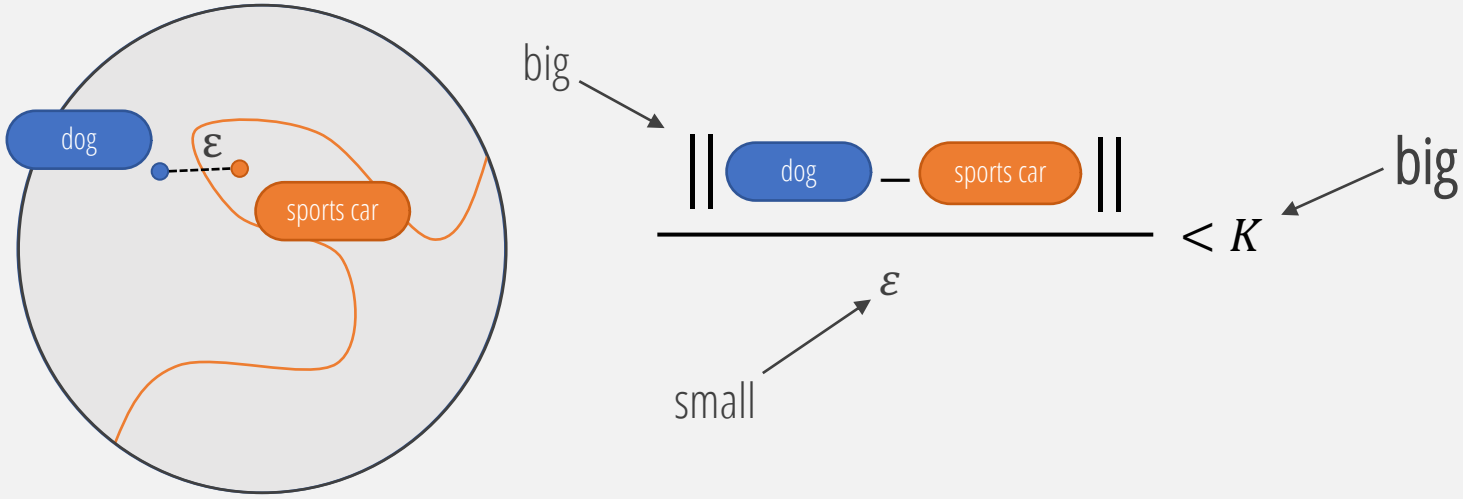
i.e., the model makes the same prediction on all points in the ε-ball centered at *x*



Lipschitz Continuity and Robustness

Key Idea

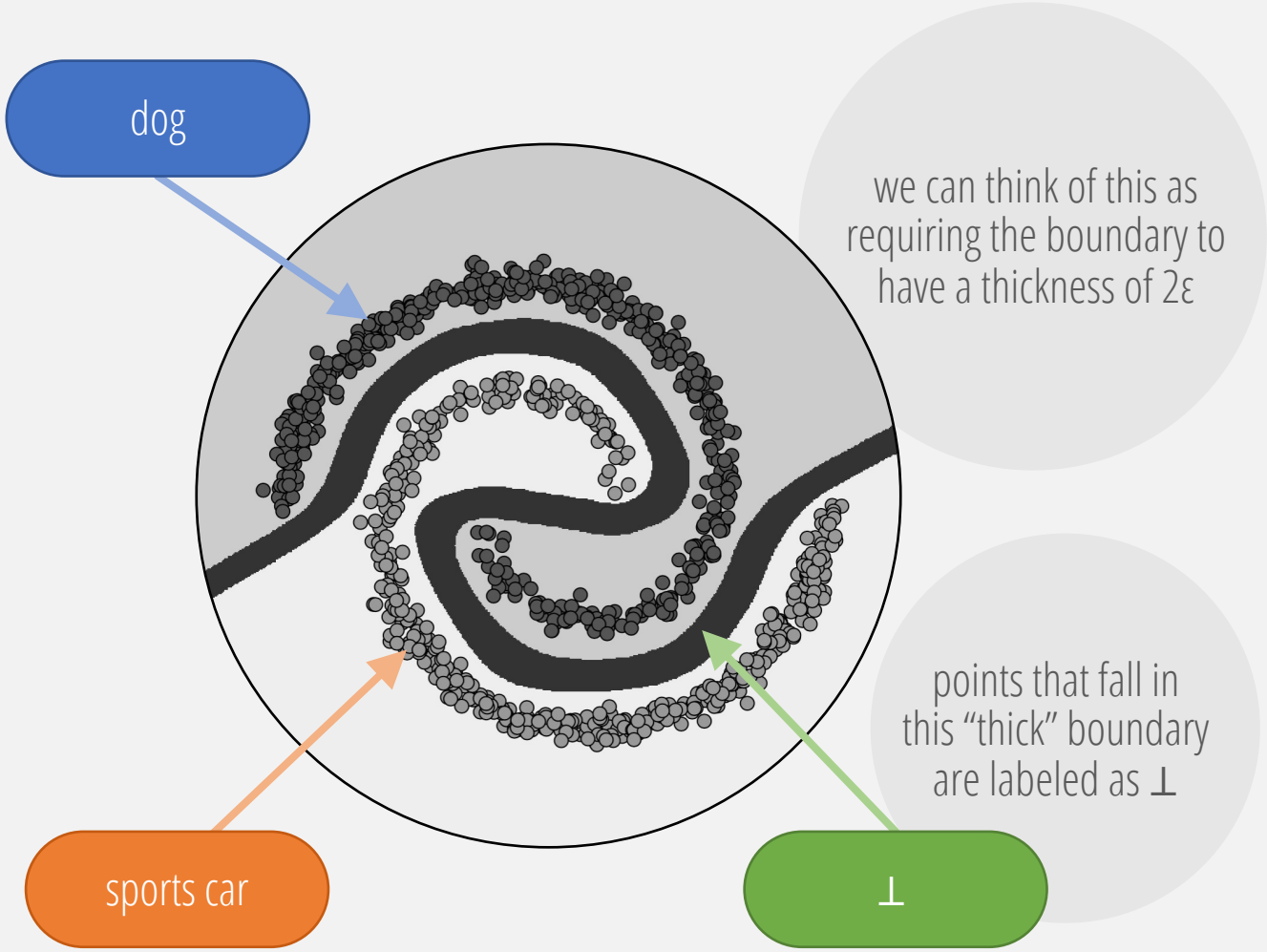
lack of Lipschitzness (i.e., a large Lipschitz constant) can account for adversarial examples



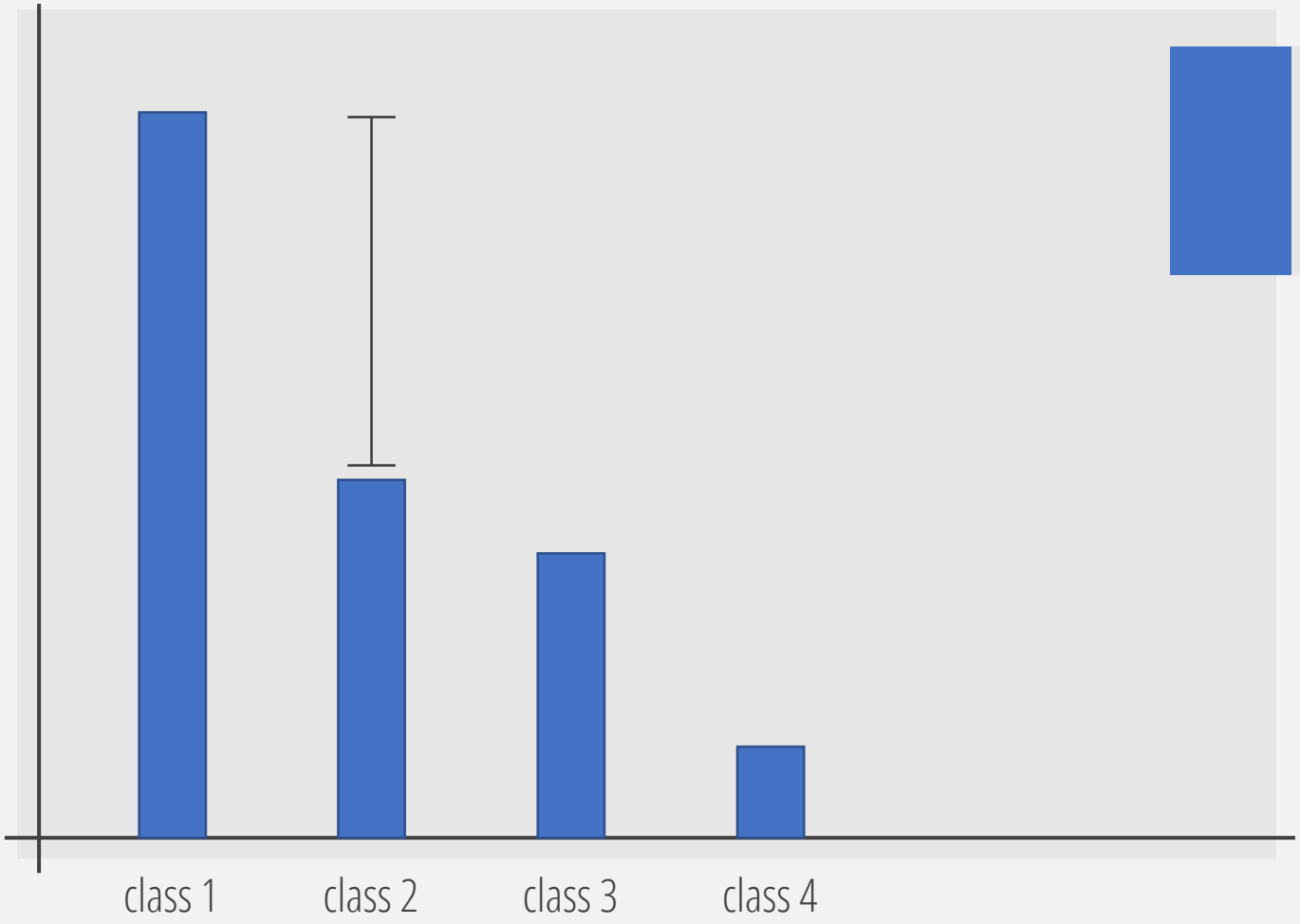
Overview

- Assessing Conceptual Soundness
- Improving Conceptual Soundness via Robustness
  - Adversarial examples and conceptual soundness
  - Globally robust neural networks
  - Limitations of local robustness
- Other Weaknesses and Vulnerabilities

Globally Robust Neural Networks (GloRo Nets)



Globally Robust Neural Networks (GloRo Nets)

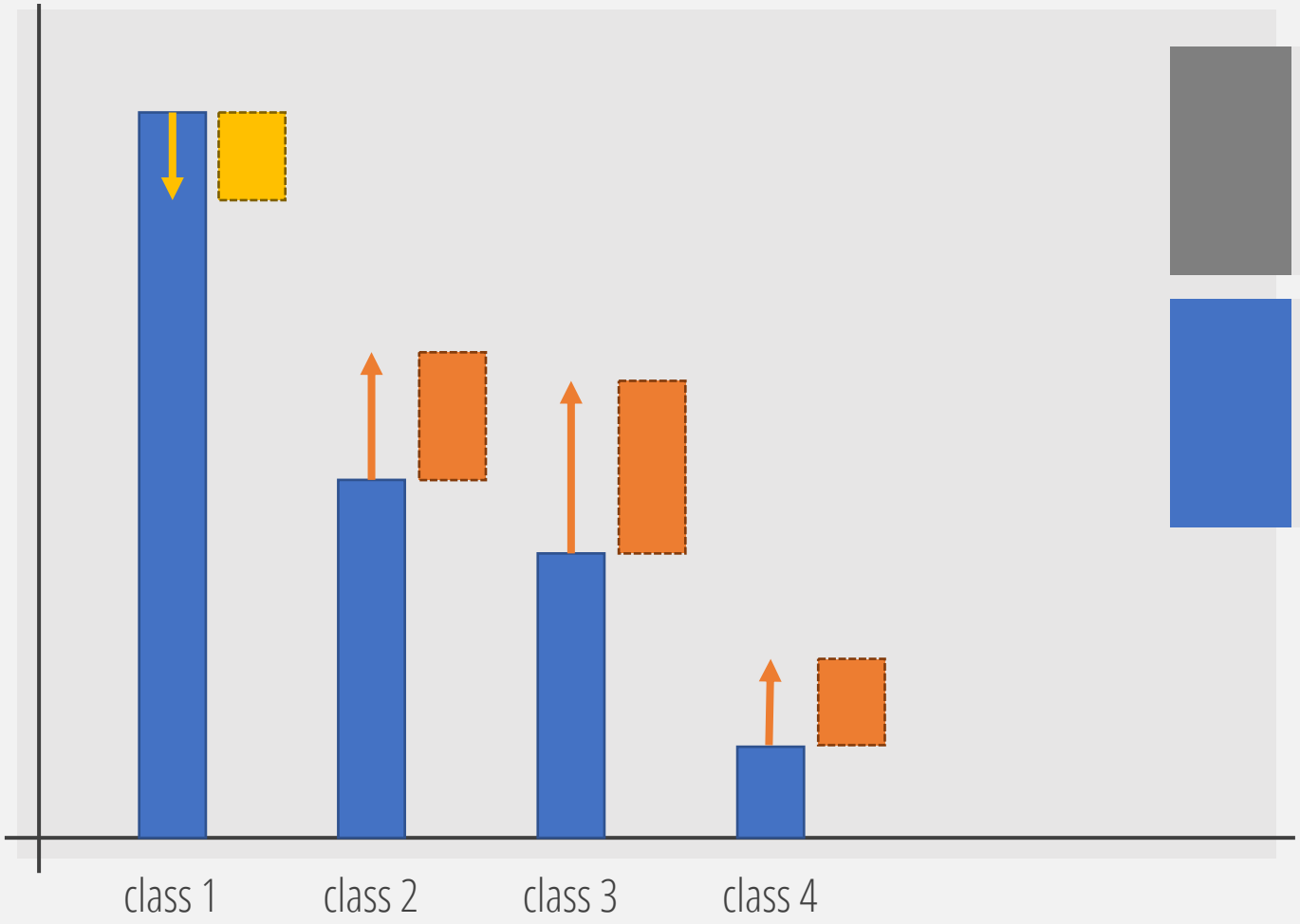


If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

Globally-Robust Neural Networks

Leino et al. ICML 2021

Globally Robust Neural Networks (GloRo Nets)

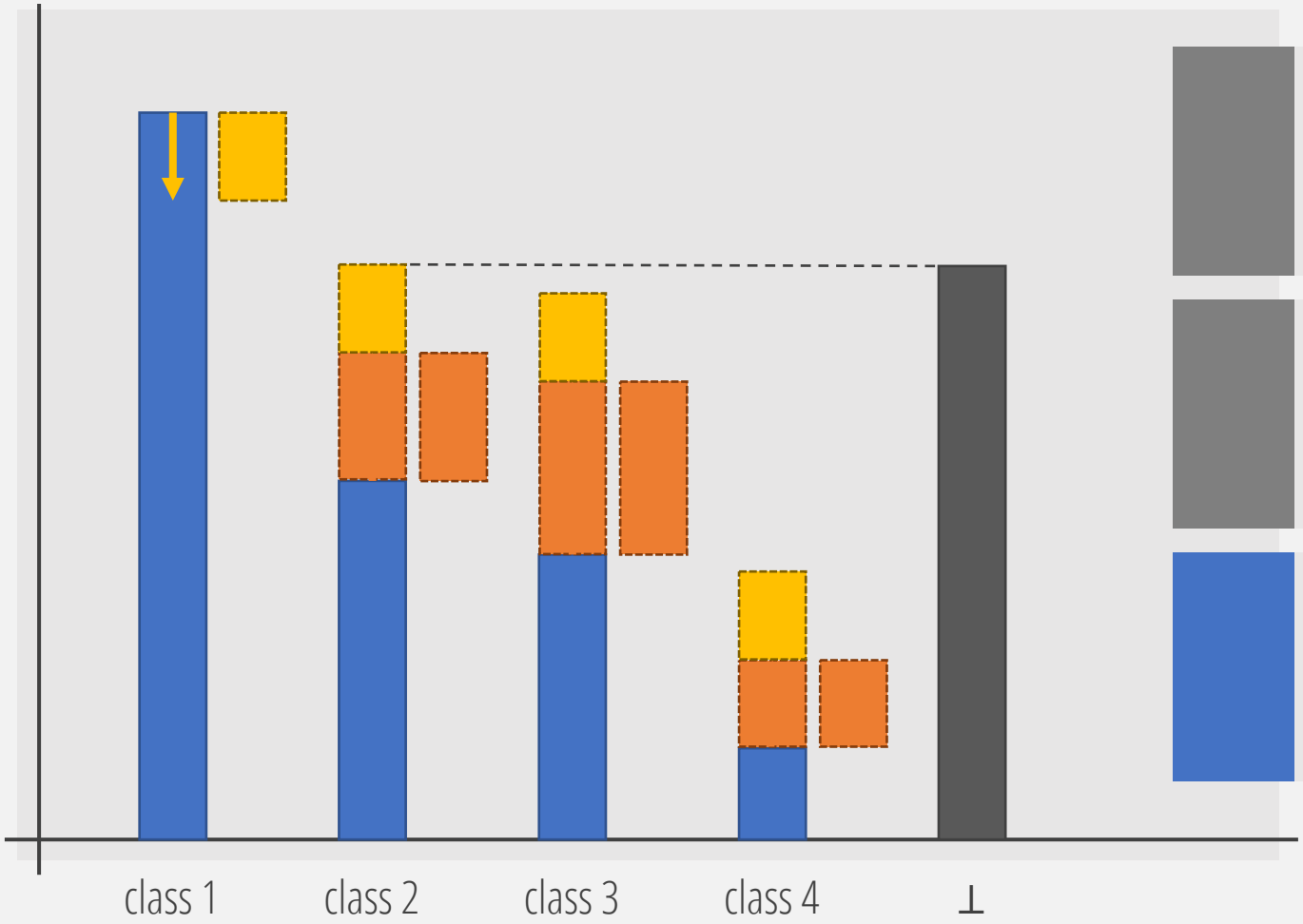


If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

The *Lipschitz constant* tells us how much each class can change with a small change to the input in the worst case

Globally-Robust Neural Networks
Leino et al. ICML 2021

Globally Robust Neural Networks (GloRo Nets)



If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

The *Lipschitz constant* tells us how much each class can change with a small change to the input in the worst case

We add a new class, ⊥, which reflects the highest score an adversary can get relative to the top class

Globally-Robust Neural Networks
Leino et al. ICML 2021

Certification of Robustness

if prediction on point x is **not** $\perp$, then x is guaranteed to be locally robust

this takes only a **single forward pass** of the network!

Training GloRo Nets

the ⊥ label is incorrect, so the loss function will penalize picking it;
thus, minimizing its loss will lead the GloRo Net to avoid picking ⊥

when the GloRo Net does not pick ⊥, it is provably robust at that point;
thus, provable robustness is naturally encouraged

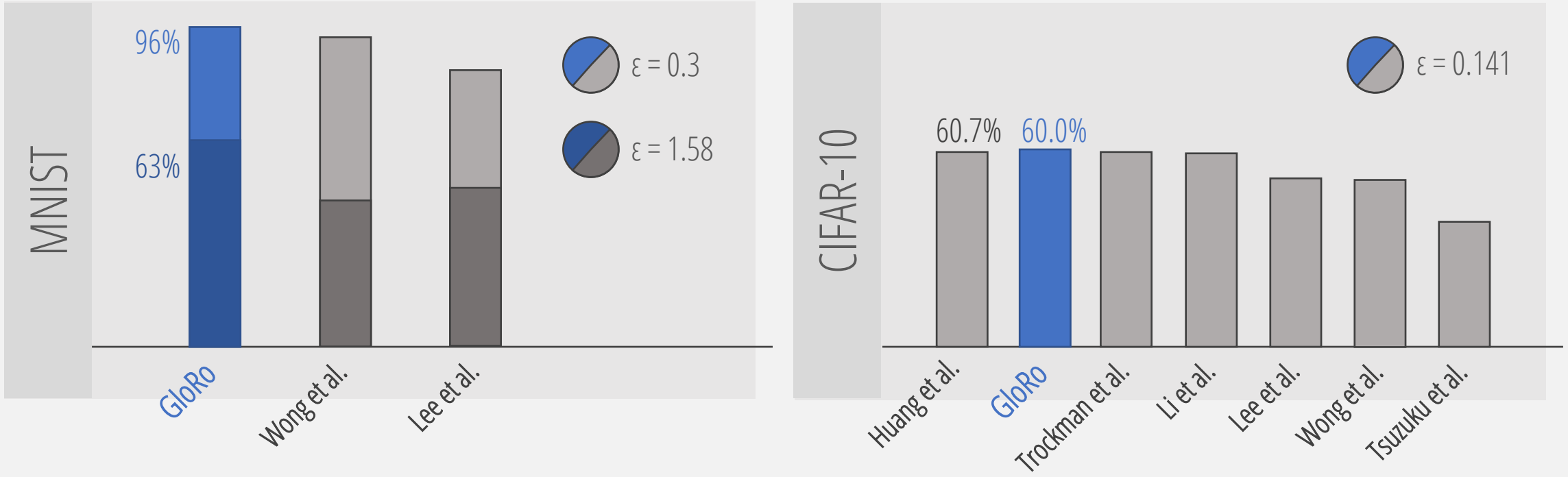
Summary of Results

GloRo Net certification and training is significantly more time and memory efficient than other methods, and more scalable than any other deterministic method

		certification method	time to certify test set (s)	memory per instance (MB)
CIFAR-10	GloRo	global Lipschitz bound	0.4	1.8
	Lee et al.	local Lipschitz bound	5.8 (15x)	19.1 (10x)
	Wong et al.	dual networks	2,500.0 (6,000x)	1,400.0 (800x)
	Cohen et al.	randomized smoothing	36,800.0 (90,000x)	19.8 (10x)

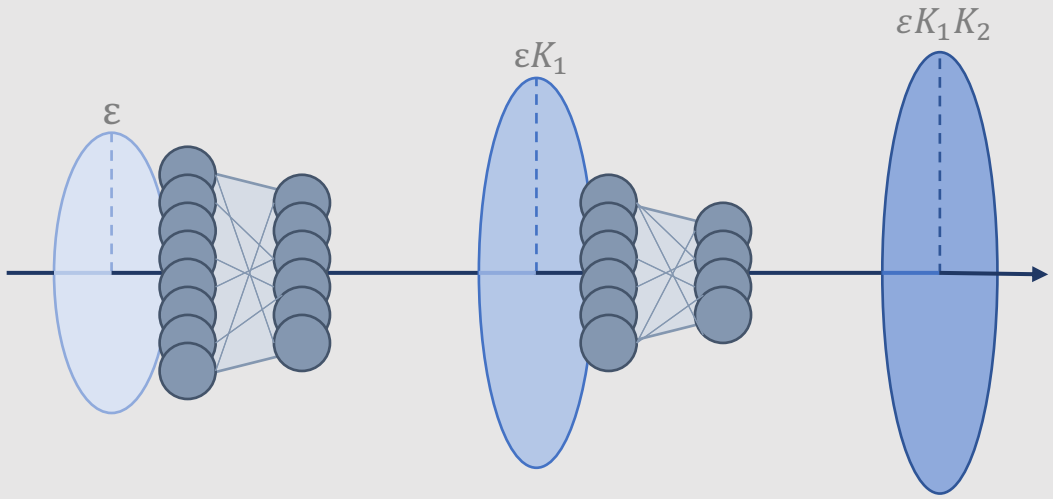
Summary of Results

GloRo Nets achieve state-of-the-art performance in terms of accuracy obtained on points that can be deterministically certified (VRA)



Bounding the Lipschitz Constant

the Lipschitz constant of the network is bounded by the product of the layer-wise Lipschitz constants



how do we calculate K_i?

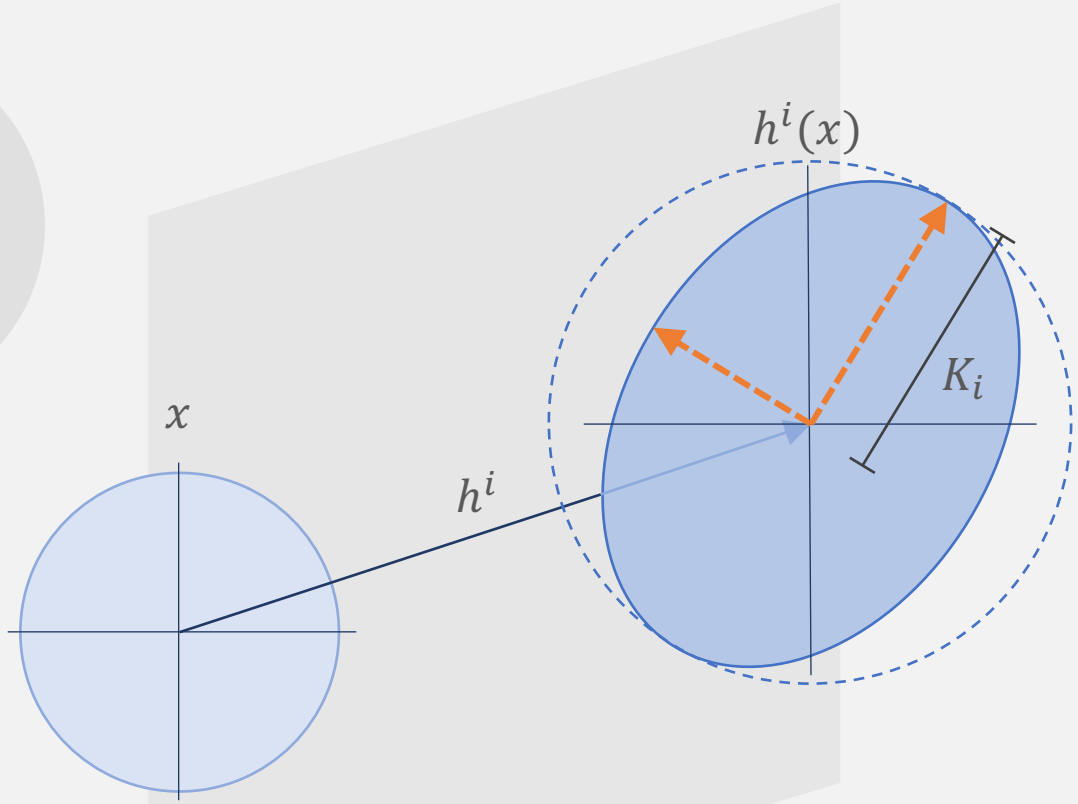
Bounding the Lipschitz Constant of a Single Layer

linear layers can be written as $h^i(x) = xW + b$, where W is a **weight matrix**

many common nonlinearities, e.g., ReLU, are 1-Lipschitz

the **operator norm** of the weight matrix W is an upper bound on the Lipschitz constant of h_i; this is simply the **largest singular value** of W

the largest eigen-pair can be **computed iteratively** as a **differentiable** function of W, allowing the flow of gradient information during training



? ask me more!

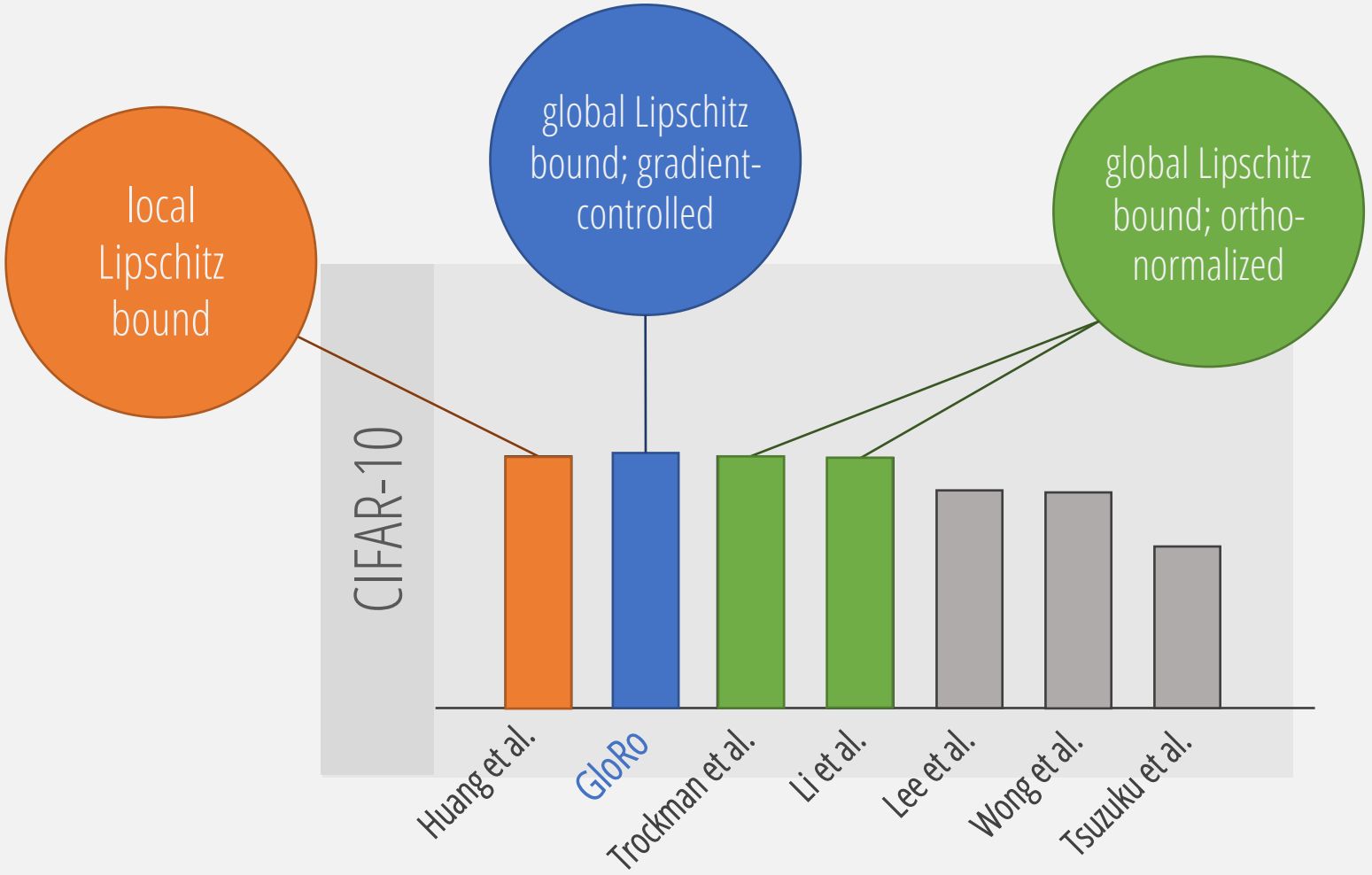
Bound Tightness

the bounds described may be very loose!

because the bound computation is incorporated into the training procedure (as part of the $\perp$ class output), there is pressure to learn models for which the bound is reasonably tight

ask me more!

Simplicity of GloRo Nets



Overview

- Assessing Conceptual Soundness
- Improving Conceptual Soundness via Robustness
  - Adversarial examples and conceptual soundness
  - Globally robust neural networks
  - Limitations of local robustness
- Other Weaknesses and Vulnerabilities

When is Local Robustness Too Strict?

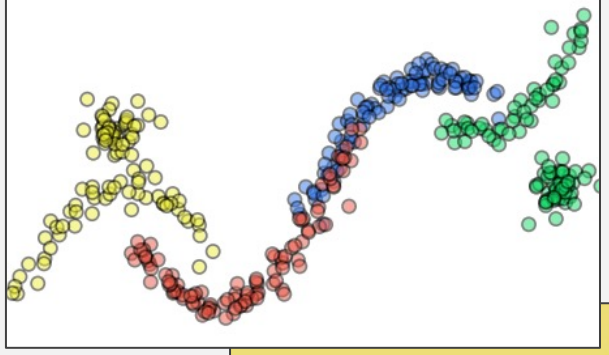
Example
task: identify the subject of satellite images

Issue 1
Ambiguous class labels due to multiple plausible subjects

Issue 2
Tough-to-separate instances

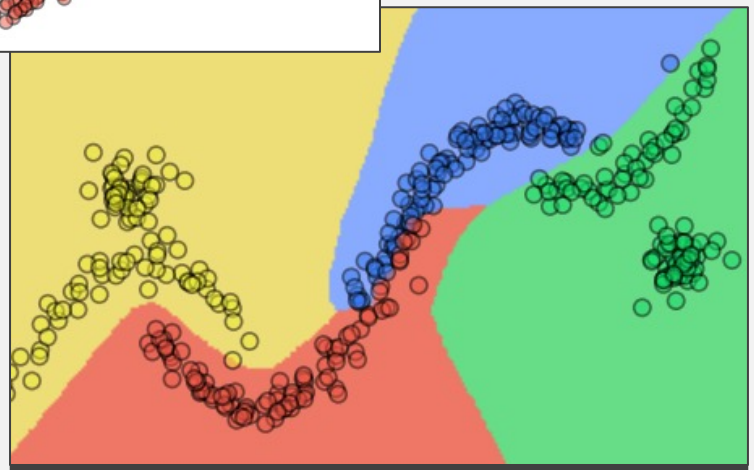
Relaxed Top-K Robustness

Synthetic Dataset

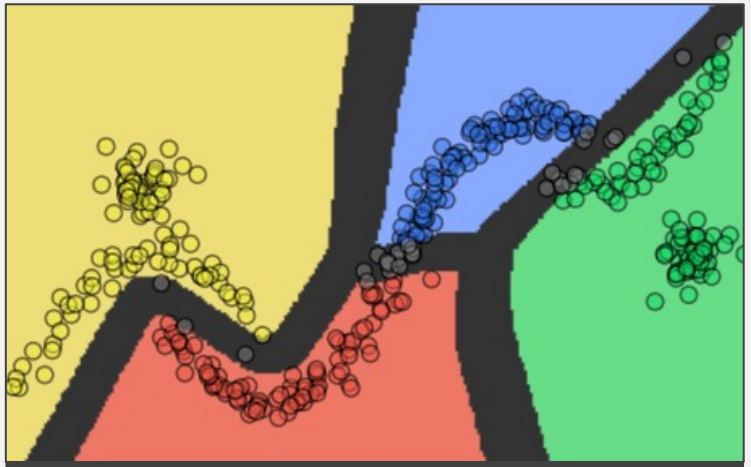


Relaxing Local Robustness

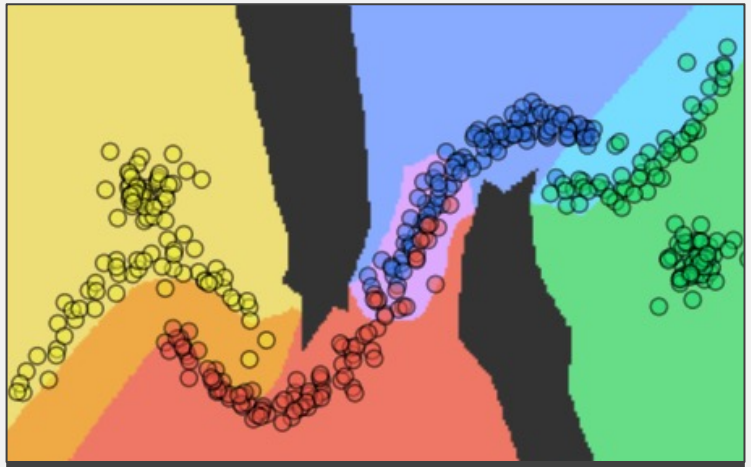
Leino & Fredrikson, NIPS 2021



Standard Boundary



Globally Robust Boundary

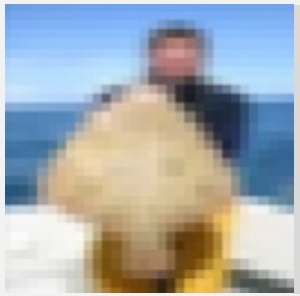


Relaxed Top-2 Robust Boundary

Lower Rejection Rates with Reasonable Groupings

CIFAR-100 (RT5)

oak, maple, willow, pine

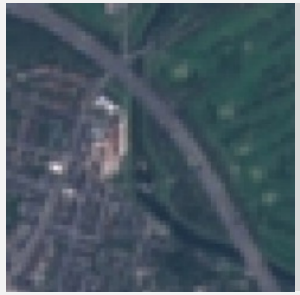


flatfish, man, trout, woman, girl

palm tree, house

EuroSAT (RT3)

highway, annual crop



highway, residential buildings

highway, permanent crop, annual crop

Leino & Fredrikson, NIPS 2021

When is Local Robustness Too Strict?

not all mistakes are equally bad!

Relaxed Robustness with Affinity Sets

	RT5 Robustness	guided by *affinity sets*
	spider, **caterpillar**, butterfly	**caterpillar**, butterfly
	**aquarium fish,** tulip, poppy	**aquarium fish**

Leino & Fredrikson, NIPS 2021

Local Robustness is not Enough for Conceptual Soundness

many outstanding challenges, e.g., “semantic” attacks

we will see other weaknesses related to conceptual soundness but orthogonal to robustness

Overview

- Assessing Conceptual Soundness
- Improving Conceptual Soundness via Robustness
- Other Weaknesses and Vulnerabilities
  - Membership privacy
  - Property privacy

When Models Lack Conceptual Soundness

- is this still a problem if it doesn't lead to classification errors?
- what does this tell us about the model's training data?
- we will see that this can be a privacy concern

Privacy in Machine Learning

strong adversary

weak adversary

	inference attacks	reconstruction attacks
instance privacy	USENIX 2020	IEEE S&P 2022
property privacy	IEEE S&P 2022	IEEE S&P 2022

relates to robustness

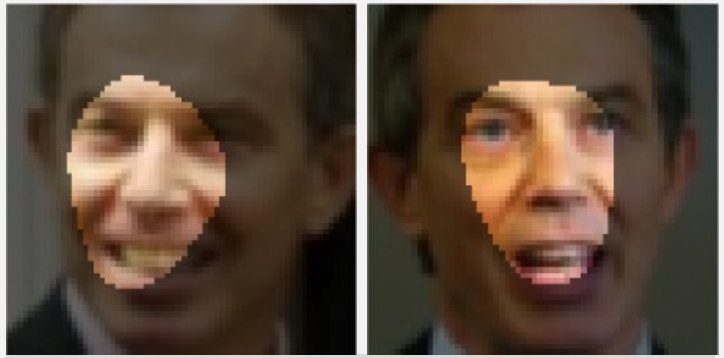
How Does Overfitting Manifest Itself?

Key Idea

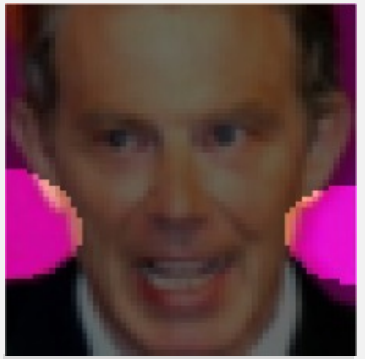
idiosyncratic feature use exposes information about the training data

sample of LFW training instances

notice the peculiar pink background in this training instance of Tony Blair



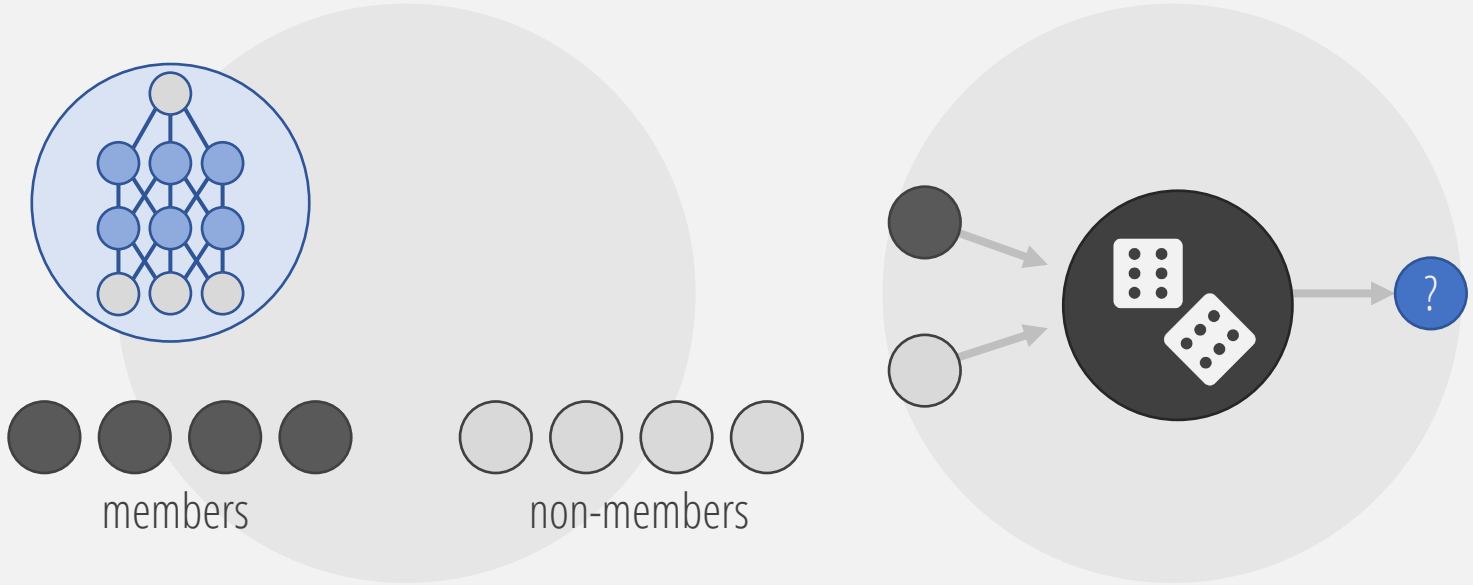
how does the model usually classify Tony Blair?



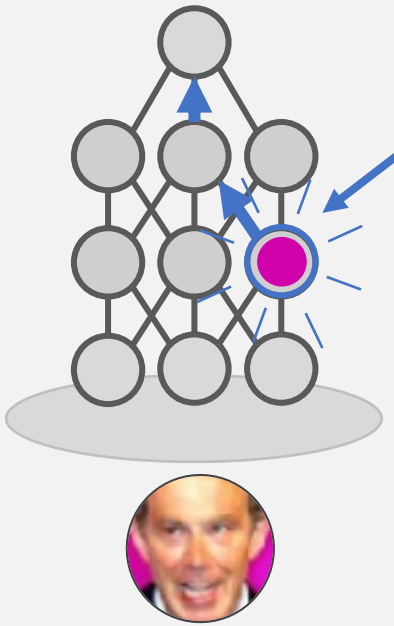
meanwhile...

Membership Inference Game

In a **membership inference attack**, an adversary tries to predict whether a given point was part of a target model's training set



Salience-based Membership Inference



because the pink background feature influences the prediction of Tony Blair, we infer a pink background was present in a training instance of Tony Blair

a pink background is predictive in the training set, so the model may learn to encode it as a feature

suppose we know that pink backgrounds are **rare** in the data distribution

then a pink background is likely to come from the training set

ask me more!

Summary of Results

We use the intuition gained by our attribution-based analysis to create a white-box membership inference attack that outperforms state-of-the-art black-box attacks

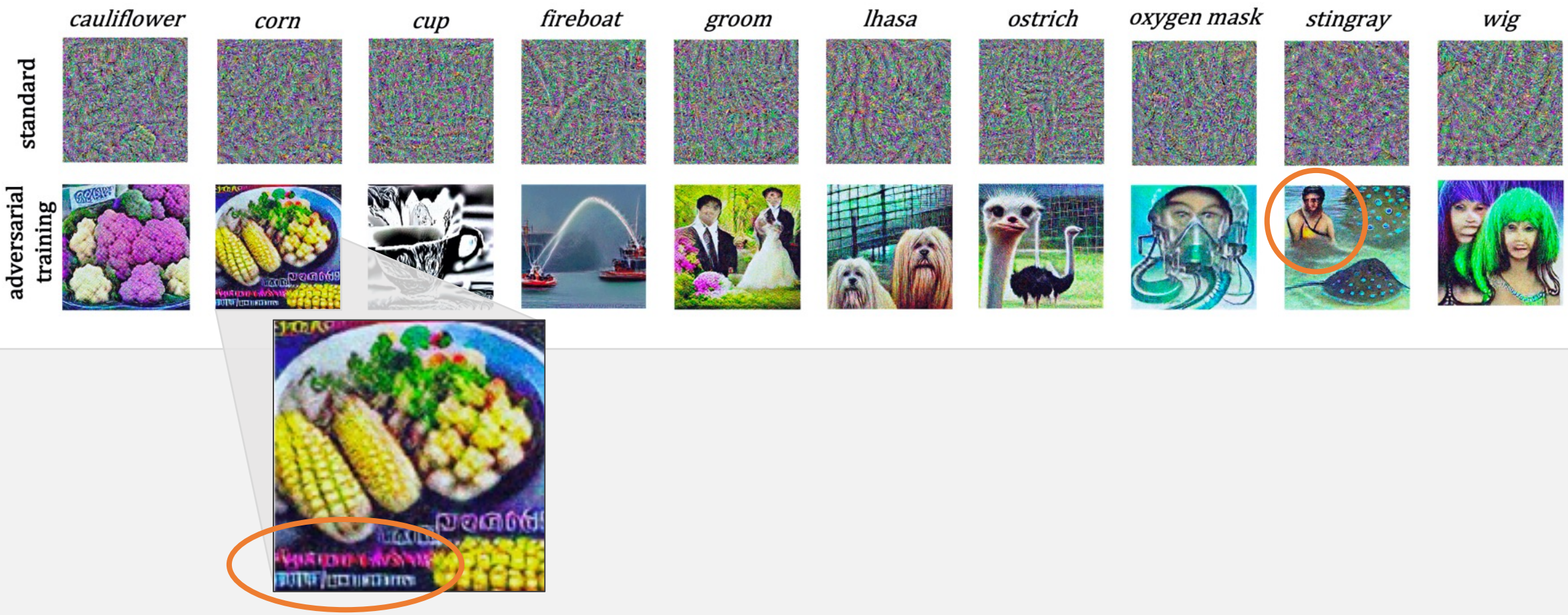
Our attack shows that it's possible to achieve high-accuracy and high-precision inference on **even models with insignificant generalization error**

generalization error	Naïve Attack[1] *accuracy*	Naïve Attack[1] *precision*	Shadow Model Attack[2] *accuracy*	Shadow Model Attack[2] *precision*	Ours *accuracy*	Ours *precision*
1.1%	50.6%	50.3%	50.6%	50.6%	57.5%	64.0%

membership infr. on MNIST

[1]*Yeom et al. 2017;* [2]*Shokri et al. 2016*

Robustness and Privacy



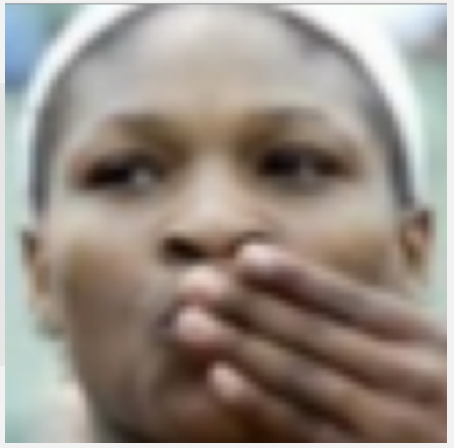
Reconstruction Attacks on Robust Models

on simple robust models, reconstructions may recover training instances

reconstruction

matching training instance

reconstruction



matching training instance

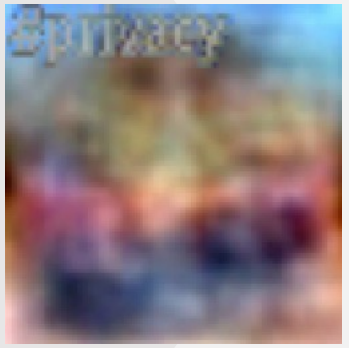
Reconstruction Attacks on Robust Models

artifacts of the general training distribution that are **irrelevant** to the learning objective may be memorized and leaked

we added a text watermark to 80% of training instances **independently** of the class label

reconstructions

LFW

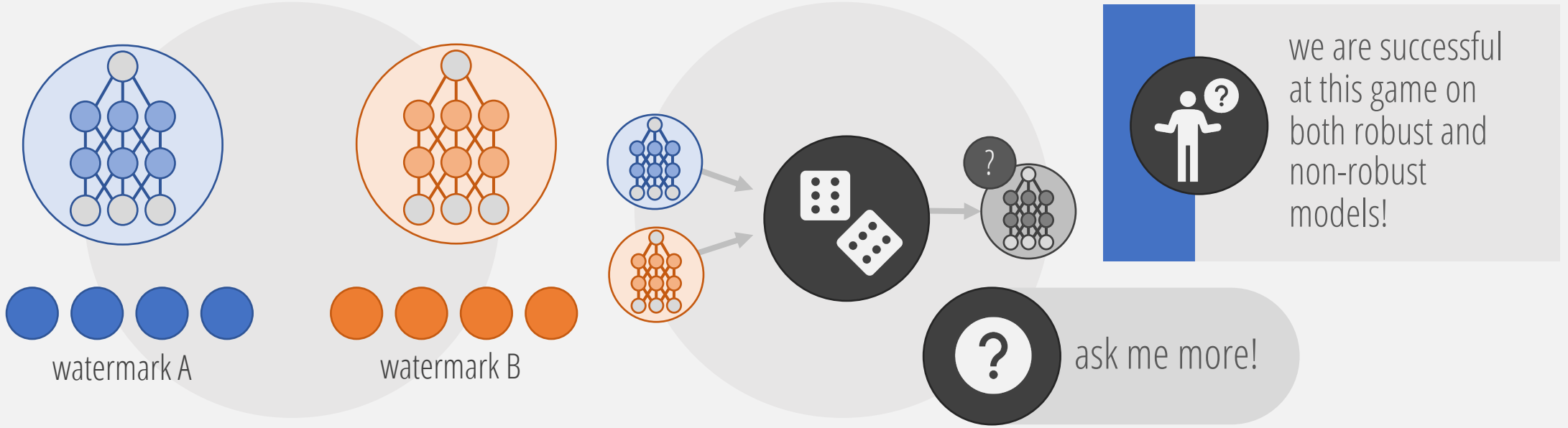


CIFAR-10

can you read what the watermark says?

Property Inference Game

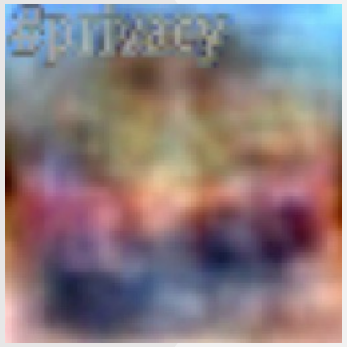
we can quantify the extent to which the reconstruction contains the watermark using a property inference game



Information Organization Hypothesis

robust

LFW



CIFAR-10

non-robust

LFW

CIFAR-10

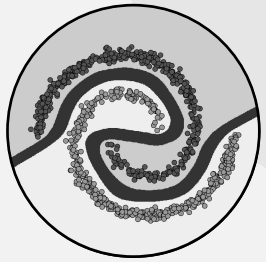
both robust and non-robust models may leak information about their training data, but if a robust model encodes leaky features, they will be more accessible to a **weak adversary**

Key Takeaways

conceptual soundness is key to the quality of deep networks and lies at the heart of many of deep learning's vulnerabilities

unfortunately, conceptual soundness is not the norm in standard deep networks

local robustness helps address this problem, and our contributions help make robust training more effective and scalable



far from being solved, our work shows that there are many other weaknesses and vulnerabilities of conceptually unsound networks

many thanks to my collaborators & committee

Matt Fredrikson

Anupam Datta,
Kamalika Chaudhuri,
J. Zico Kolter,
Corina Păsăreanu

Emily Black, Anupam Datta, Matt Fredrikson,
Aymeric Fromherz, Jon Helland, Kaiji Lu,
Ravi Mangal, Piotr Mardziel, Bryan Parno,
Corina Păsăreanu, Shayak Sen,
Nathan VanHoudnos, Saranya Vijayakumar,
Zifan Wang, Chi Zhang, et al.

Questions