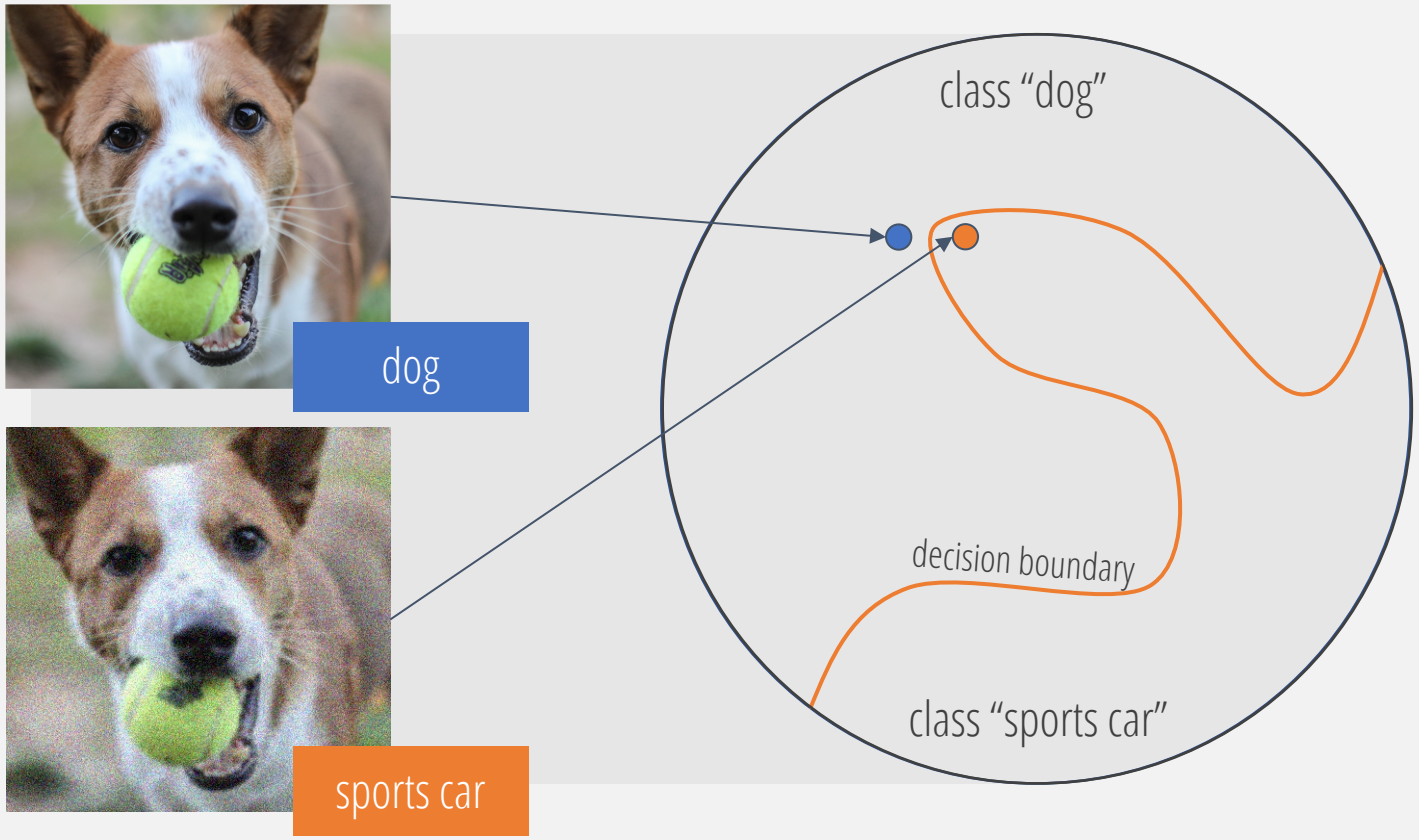
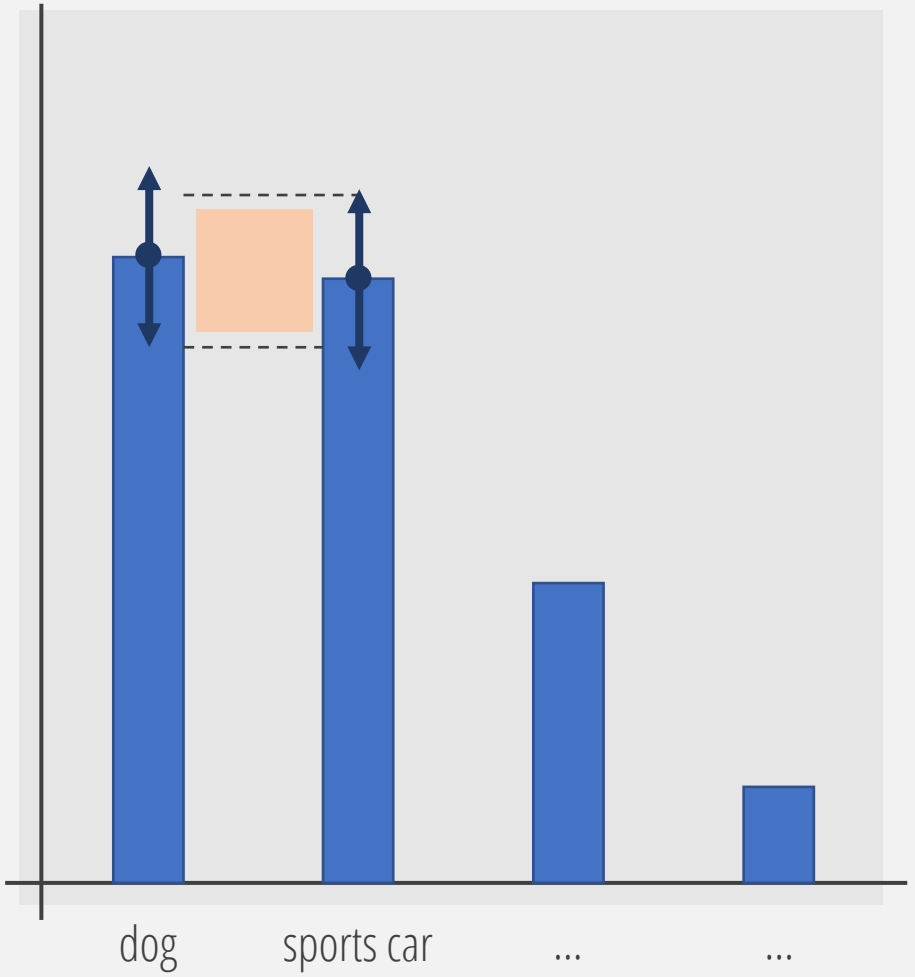
Relaxing Local Robustness

Klas Leino*, Matt Fredrikson | Carnegie Mellon University

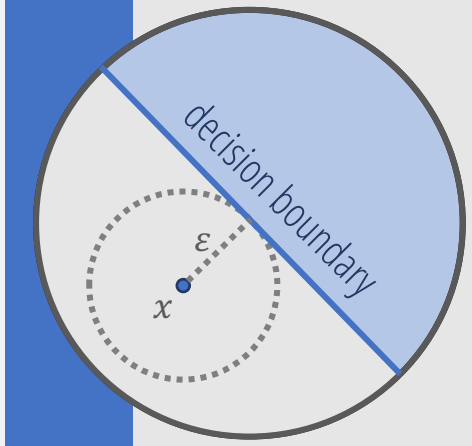
Adversarial Examples





Certified Defenses

Want to defend against *adversarial examples*

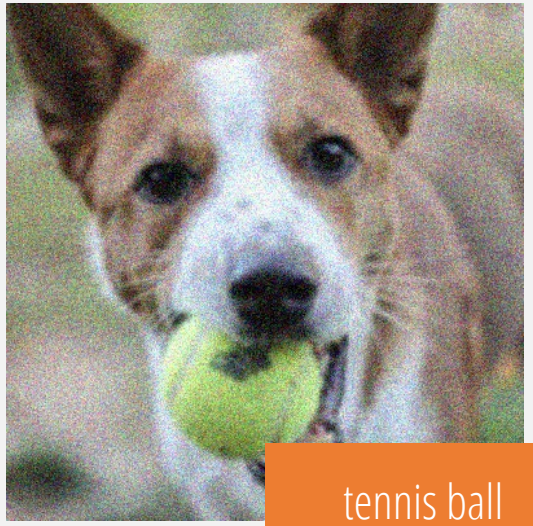


A model F satisfies *local robustness* with robustness radius ε on a point x if

$$\forall x'. \|x - x'\|_p \leq \varepsilon \implies F(x) = F(x')$$

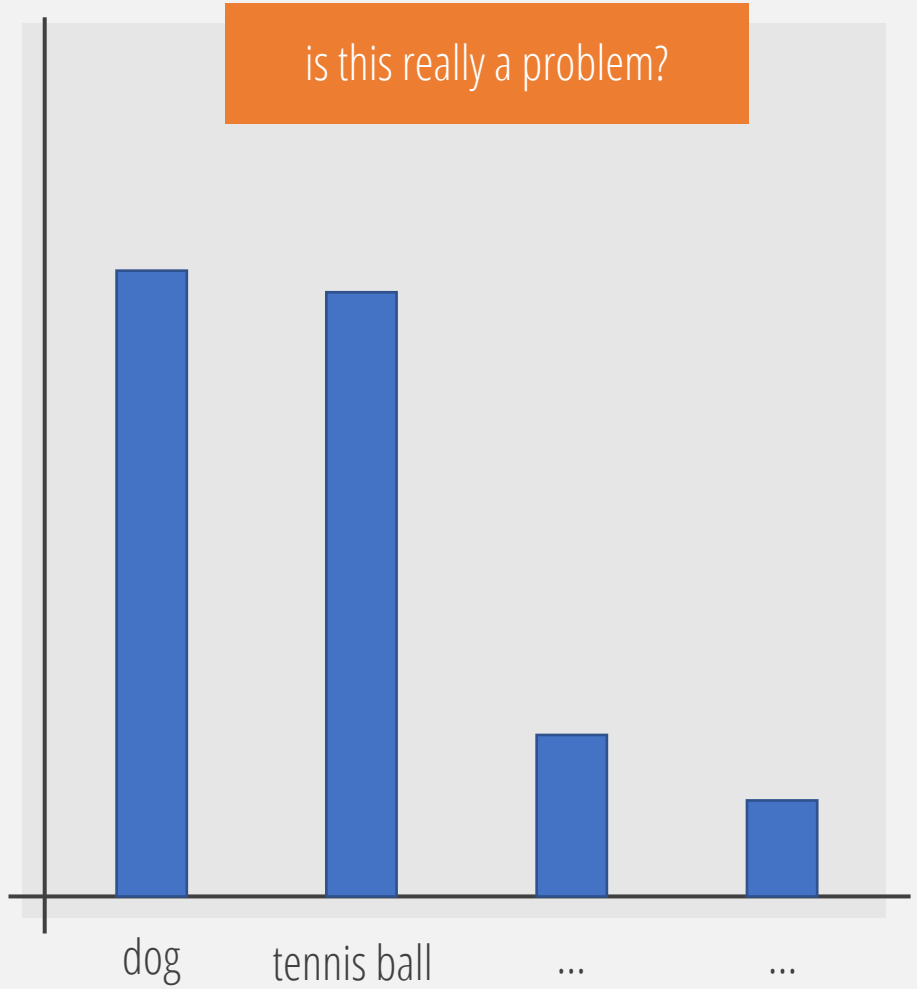
Local Robustness May Be Ill-suited

dog



tennis ball

Requiring a strong preference for either "dog" or "tennis ball" is ***arbitrary***



Our Contributions

We introduce two *relaxed notions of robustness* that are more suitable than local robustness in many contexts

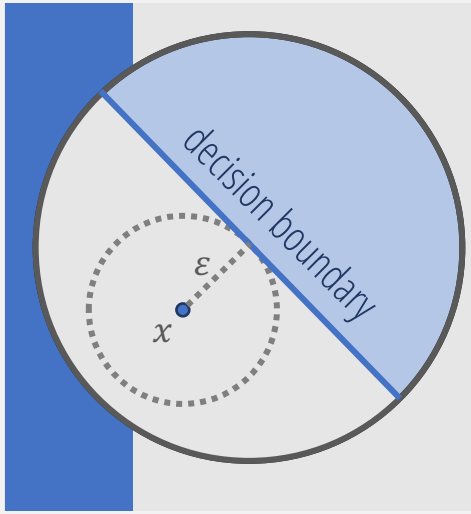
We devise a way to construct networks such that our robustness properties can be *efficiently certified*

We provide case studies showing the *suitability* of our proposed properties to real-world classification tasks

Overview

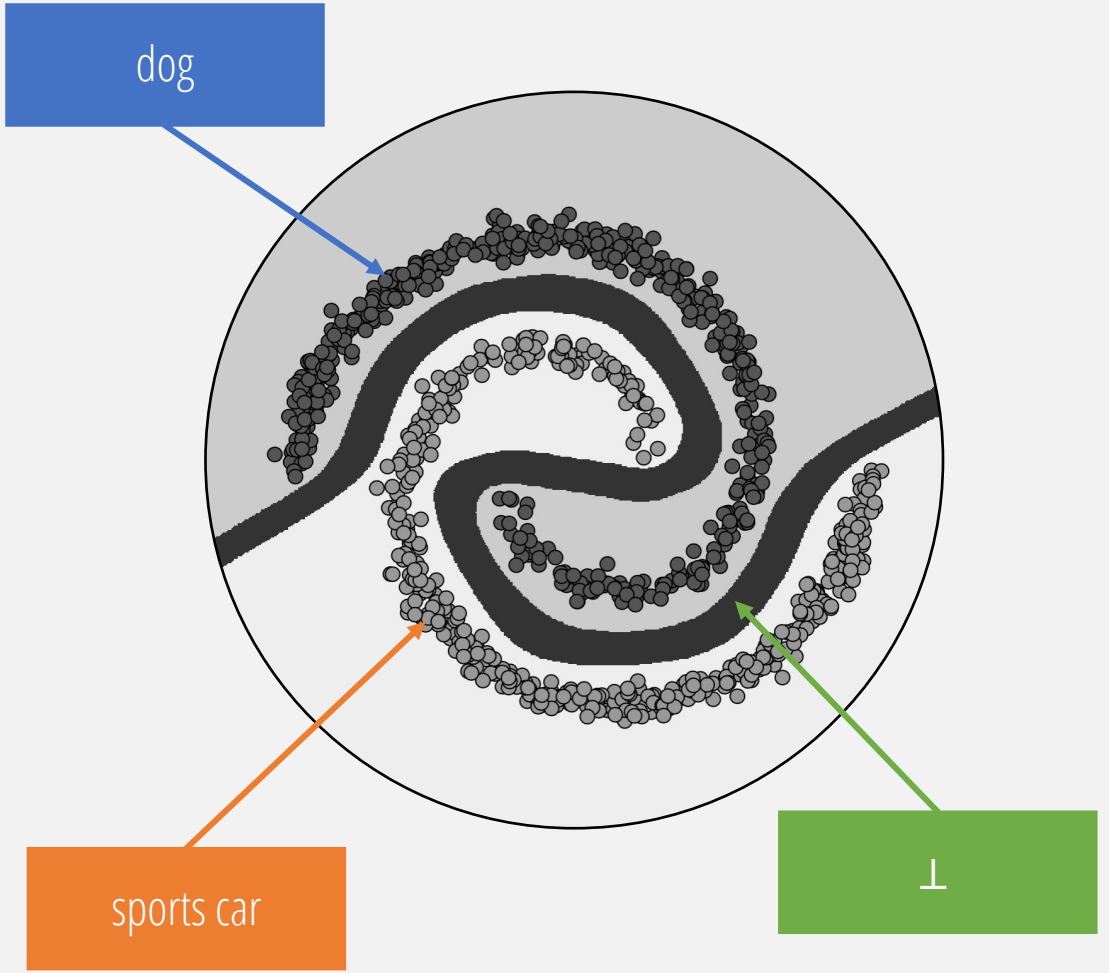
- Novel robustness properties
  - Relaxed Top-K Robustness
  - Affinity Robustness
- Certification of novel robustness properties
- Experimental results

“Global” Robustness



Recall: robust points must be far from decision boundaries

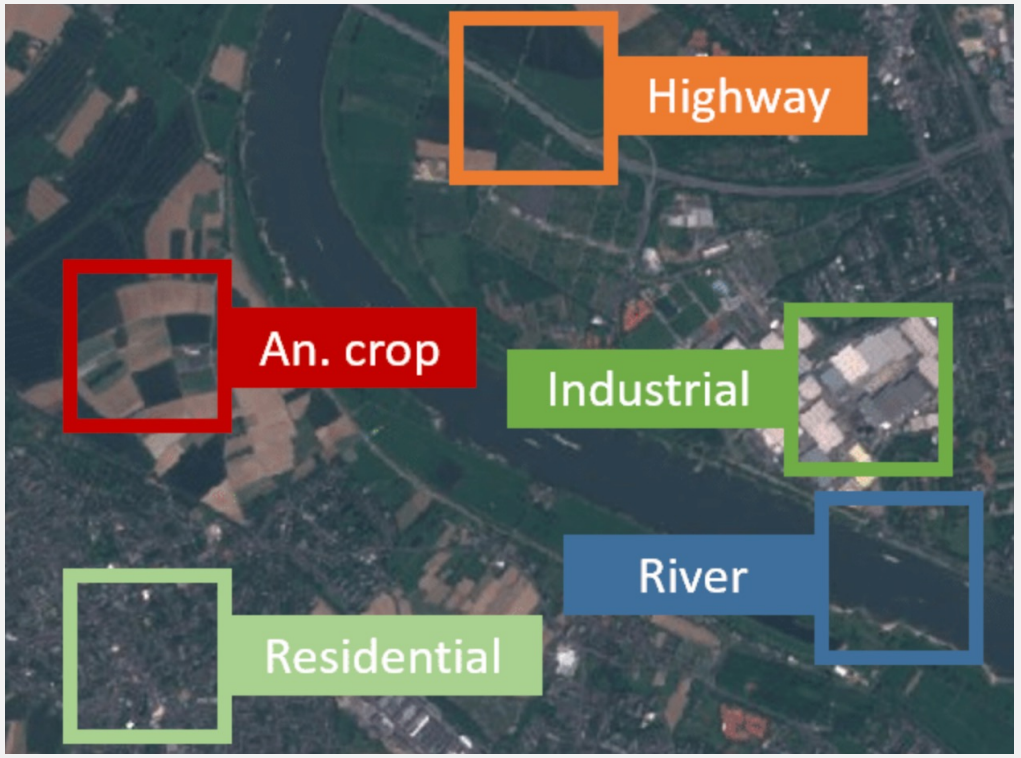
Robust models must induce separation between classes



Motivation

Goal: models that produce conceptually-reasonable boundaries

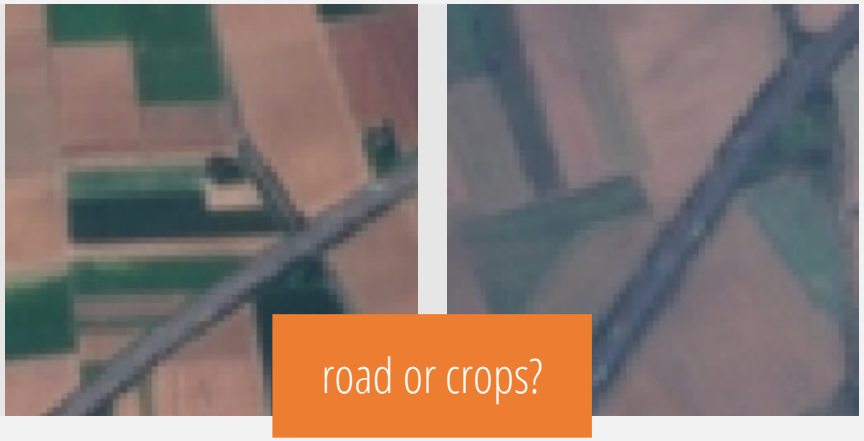
What about cases where there is not always a clear separation between classes?



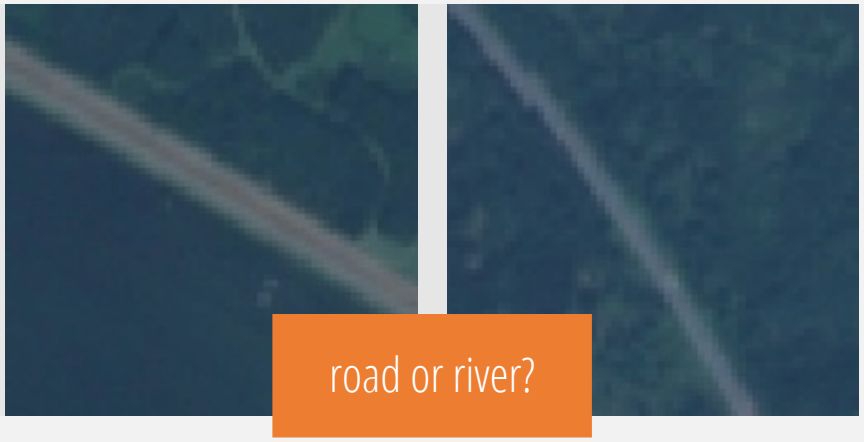
EuroSAT: Helber et al. 2017

Motivation

Issue 1
Ambiguous class labels due to multiple plausible subjects



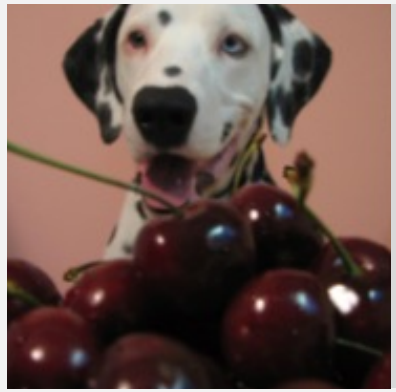
Issue 2
Tough-to-separate instances



Robustness and Top-k Accuracy

On Imagenet it's common to use top-5 accuracy

Leopard
Jaguar
Cheetah
Snow Leopard
Egyptian cat



Cherry
Dalmatian
Grape
Elderberry
Bull Terrier

Grille
Convertible
Pickup
Beach Wagon
Fire Engine

Can we make a robustness notion that mirrors this?

Top-k Robustness

We can think of a neural network as outputting a *set* of predictions

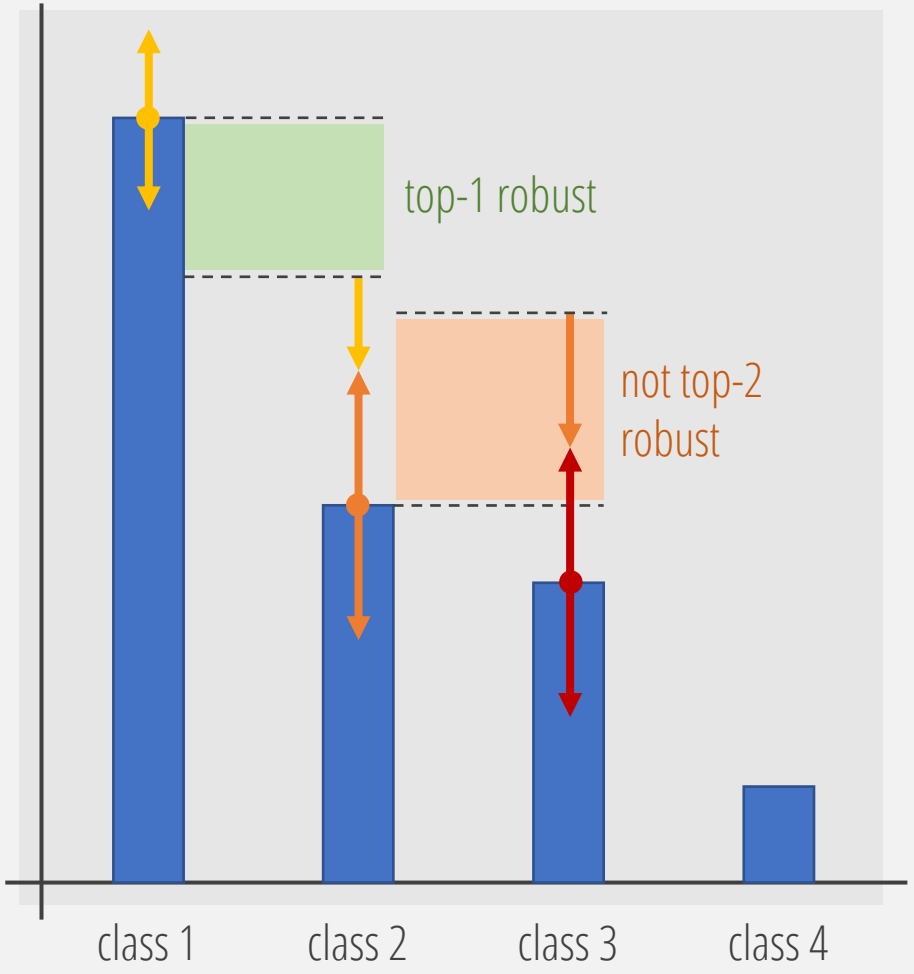
Given a model F, let $F^k(x)$ be the set of the top k classes as evaluated by F on x

Top-k Robustness

A model F is **top-k robust** with robustness radius ε on a point x if

$$\forall x'. \left|\left|x - x'\right|\right|_p \leq \varepsilon \implies F^k(x) = F^k(x')$$

Note that this is *not* a relaxation!



Relaxed Top-K (RTK) Robustness

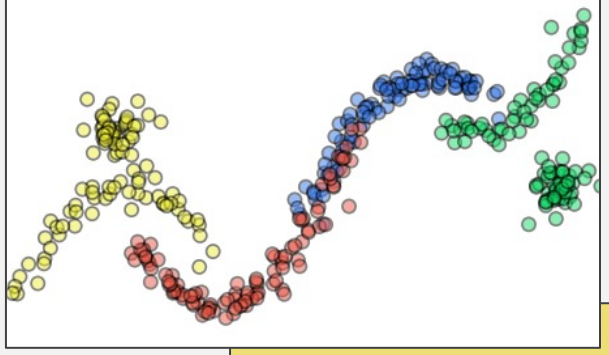
A model F is **relaxed-top-K robust** with robustness radius ε on a point x if

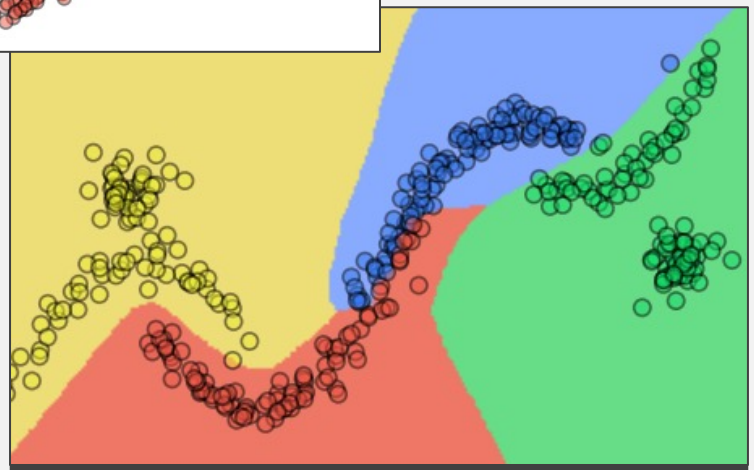
$$\forall x'. \left|\left|x - x'\right|\right|_p \leq \varepsilon \implies \exists\, k \leq K : F^k(x) = F^k(x')$$

This *is* a relaxation of local robustness

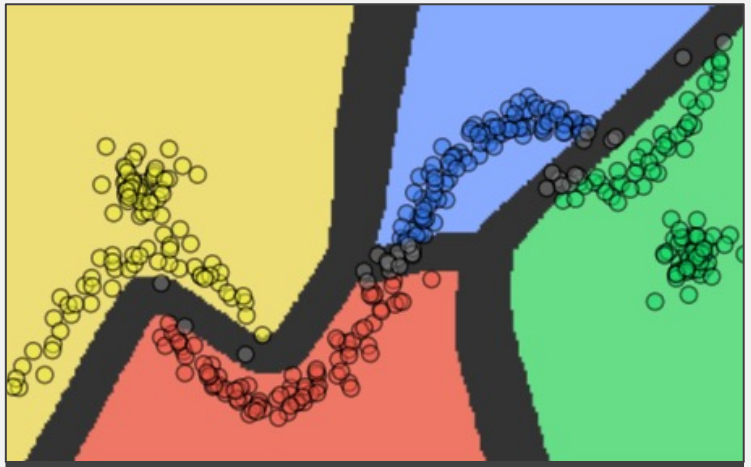
Example Boundaries

Synthetic Dataset

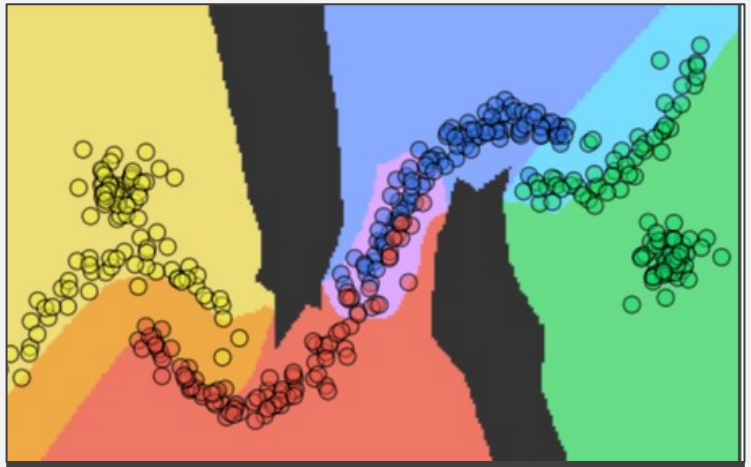




Standard Boundary



Globally Robust Boundary



RT2 Boundary

Affinity Robustness

We can also restrict the classes that can be grouped together to a collection of specified *affinity groups*, $\mathcal{S}$

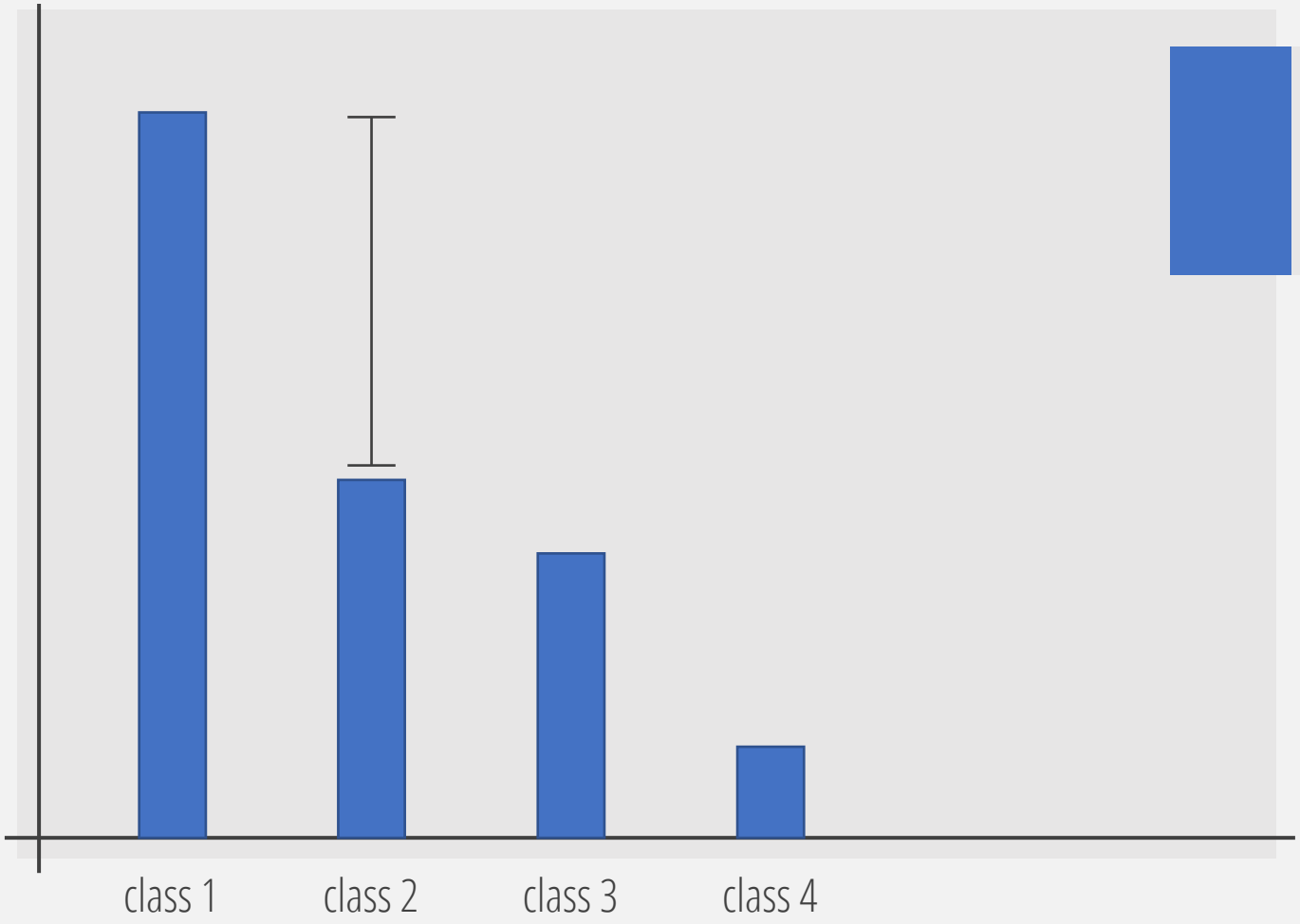
A model F is **affinity-robust** wrt. a collection of affinity groups $\mathcal{S}$, with robustness radius ε on a point x if

$$\forall x'.\ \left|\left|x - x'\right|\right|_p \leq \varepsilon$$
$$\implies \exists\, S \in \mathcal{S} :\ F^{|S|}(x) = F^{|S|}(x') \ \wedge\ F^{|S|}(x) \cap S = F^{|S|}(x)$$

Overview

- Novel robustness properties
  - Relaxed Top-K Robustness
  - Affinity Robustness
- Certification of novel robustness properties
- Experimental results

Globally Robust Neural Networks (GloRo Nets)

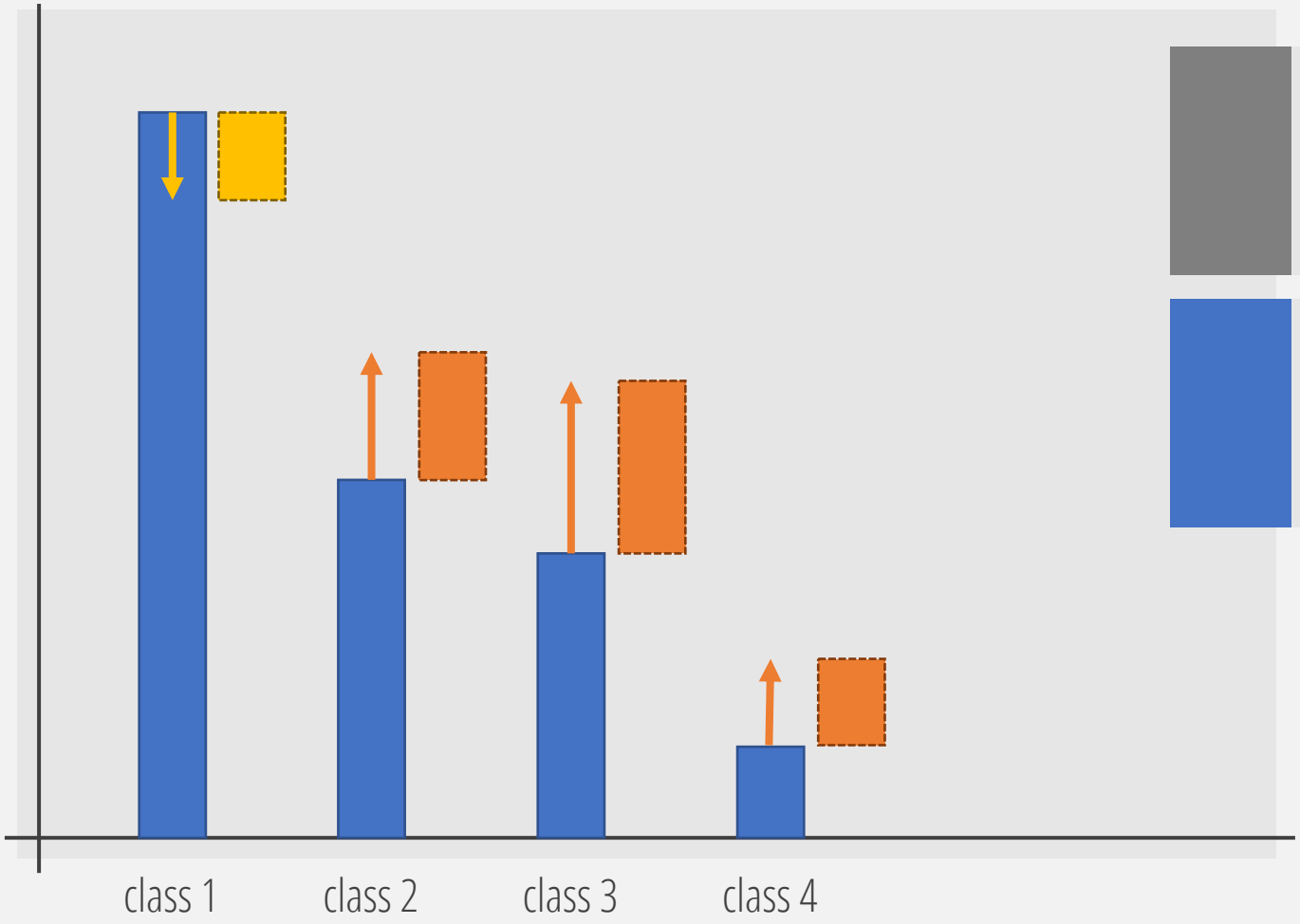


If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

Globally-Robust Neural Networks

Leino et al. ICML 2021

Globally Robust Neural Networks (GloRo Nets)



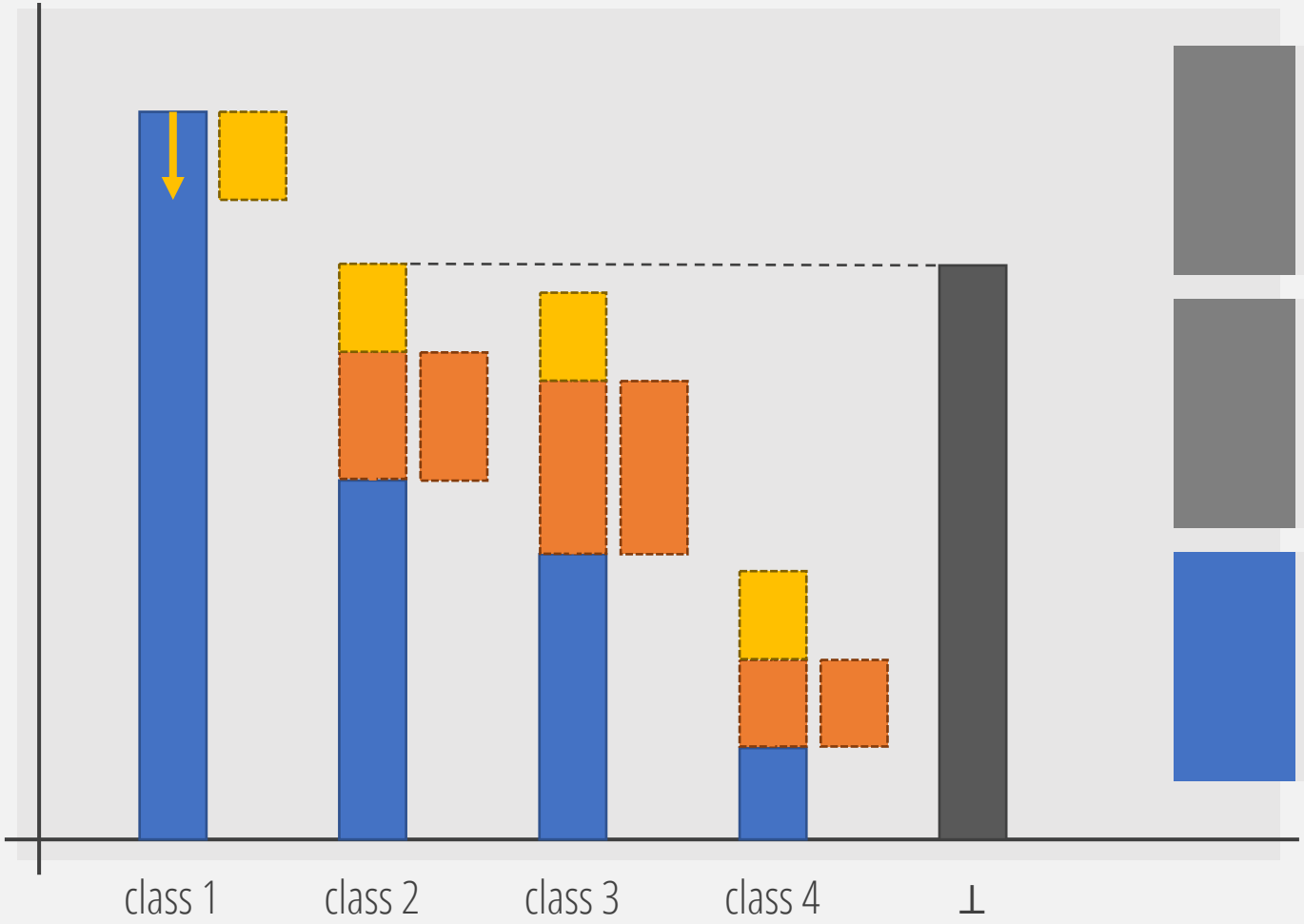
If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

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

Achieving Relaxed Robustness Guarantees

Main Idea

We modify the way the $\perp$ class is computed from the Lipschitz constant and logits, such that whenever the network is not RTK/Affinity-robust the network outputs $\perp$

See paper and code for more details…

Overview

- Novel robustness properties
  - Relaxed Top-K Robustness
  - Affinity Robustness
- Certification of novel robustness properties
- Experimental results

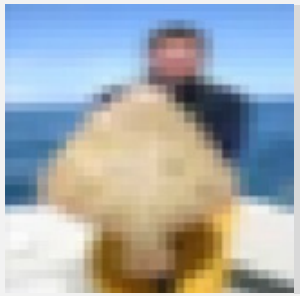
Results

dataset	guarantee	VRA*	
EuroSAT	local robustness	0.749	
EuroSAT	RT3	0.908	+16%
CIFAR-100	local robustness	0.281	
CIFAR-100	RT5	0.360	+8%
CIFAR-100	superclass affinity	0.323	+4%
Tiny-Imagenet	local robustness	0.224	
Tiny-Imagenet	RT5	0.277	+5%

Results

CIFAR-100 (RT5)

oak, maple, willow, pine

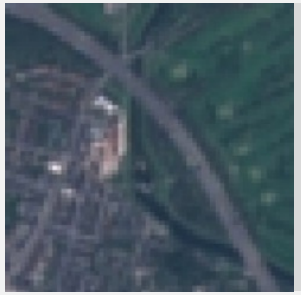


flatfish, man, trout, woman, girl

palm tree, house

EuroSAT (RT3)

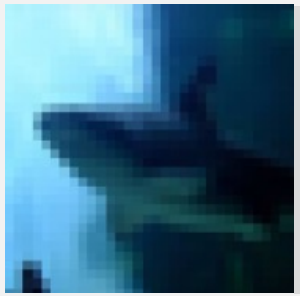
highway, annual crop



highway, residential buildings

highway, permanent crop, annual crop

Results

	RT5 Robustness	Superclass Affinity Robustness
	**shark,** sea	**shark,** ray
	spider, **caterpillar**, butterfly	**caterpillar**, butterfly
	**aquarium fish,** tulip, poppy	**aquarium fish**

Conclusion

Summary

We provide two relaxed notions of robustness that are better suited for many types of classification tasks, and we show how these robustness properties can be efficiently certified

Check Out Our Paper!

- Full paper on ArXiv
- Implementation on GitHub

  https://github.com/klasleino/gloro

full paper

https://tinyurl.com/relaxed-robustness