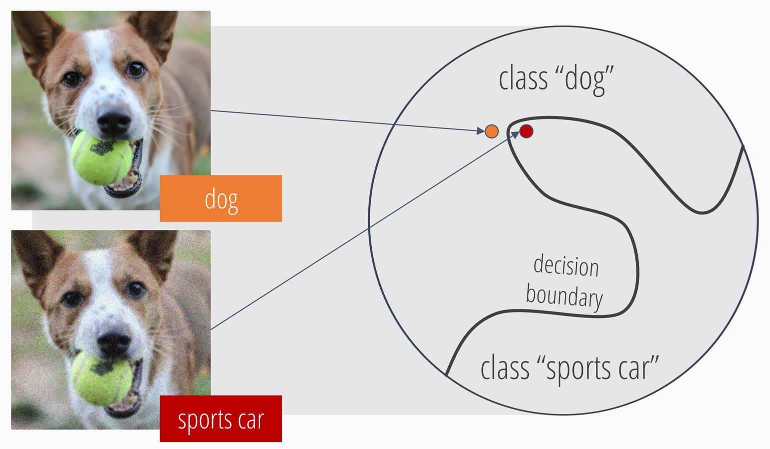
Relaxing Local Robustness

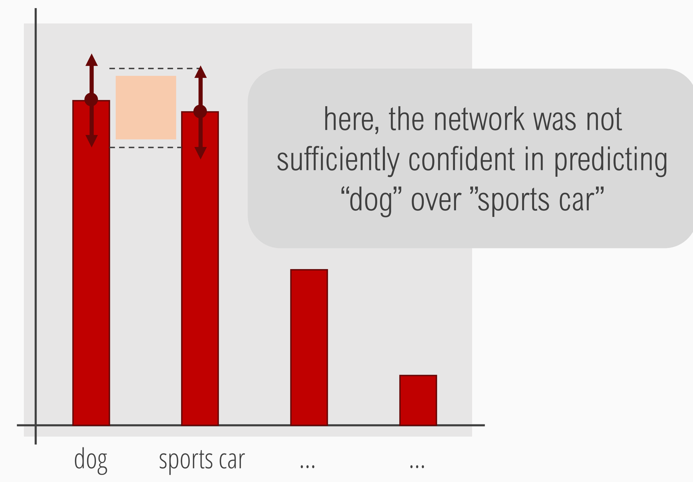
Klas Leino & Matt Fredrikson

Carnegie Mellon University

Adversarial Examples & Local Robustness

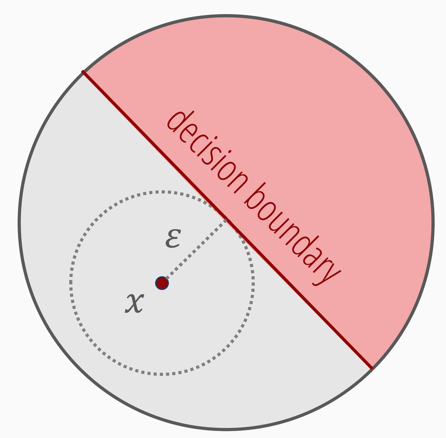
Deep networks are vulnerable to *adversarial examples*, wherein inconspicuous perturbations are chosen to cause arbitrary misclassifications.





a model is *ε-locally-robust* at a point, *x*, if it classifies all points in the ε-ball centered at *x* consistently; i.e., there are no decision boundaries within ε from *x*

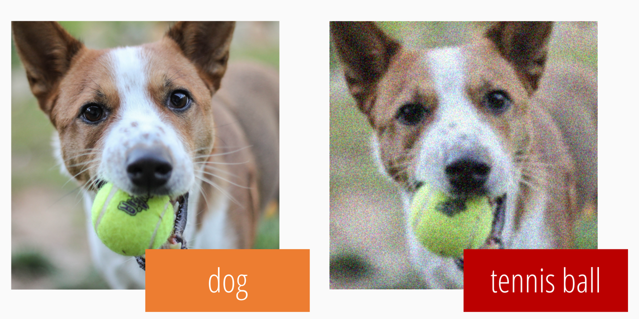
local robustness



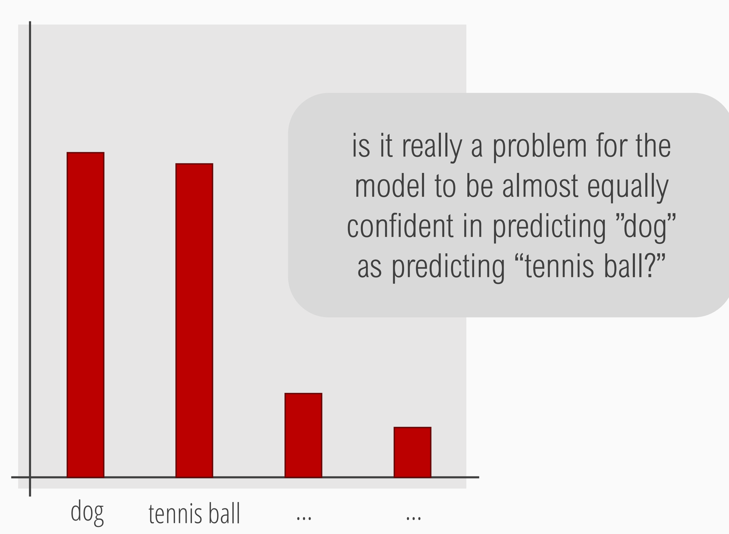
| certified defenses

Certification of local robustness at a given point allows us to provably preclude small-norm adversarial examples at that point.

Local Robustness may be Ill-suited



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

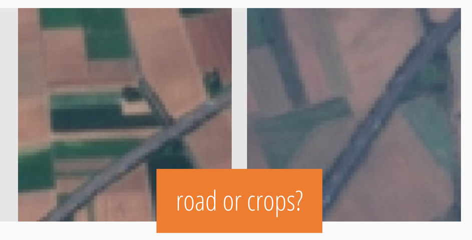
Relaxations of Local Robustness

| Relaxed Top-K Robustness

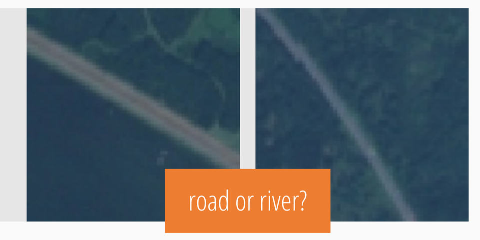
Relaxed Top-K, or *RTK*, robustness is the robustness analogue of top-k accuracy, which is often used in classification settings with label noise, subject ambiguity, or classes that are difficult to distinguish.

motivation | certain issues in the learning task might make local robustness impractical

Issue 1
Ambiguous class labels due to multiple plausible subjects



Issue 2
Tough-to-separate instances



top-k robustness | a straightforward attempt at this analogue does *not* lead to a relaxation of local robustness

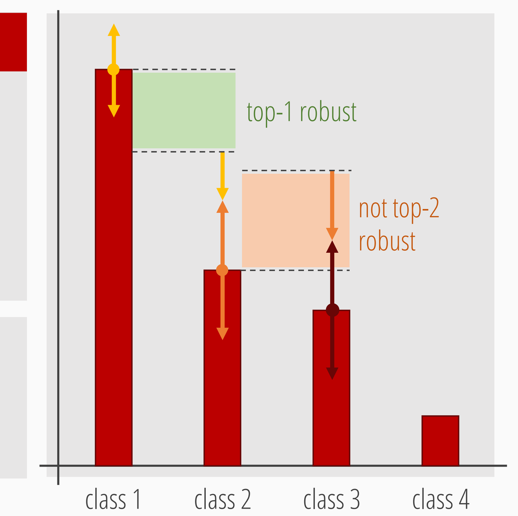
To define our new robustness notions, we would like to think of a model as outputting an ordered set of classes:

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

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

$$\forall x'.\ ||x - x'||_p \le \varepsilon \Longrightarrow F^k(x) = F^k(x')$$

Note that this is *not* a relaxation!

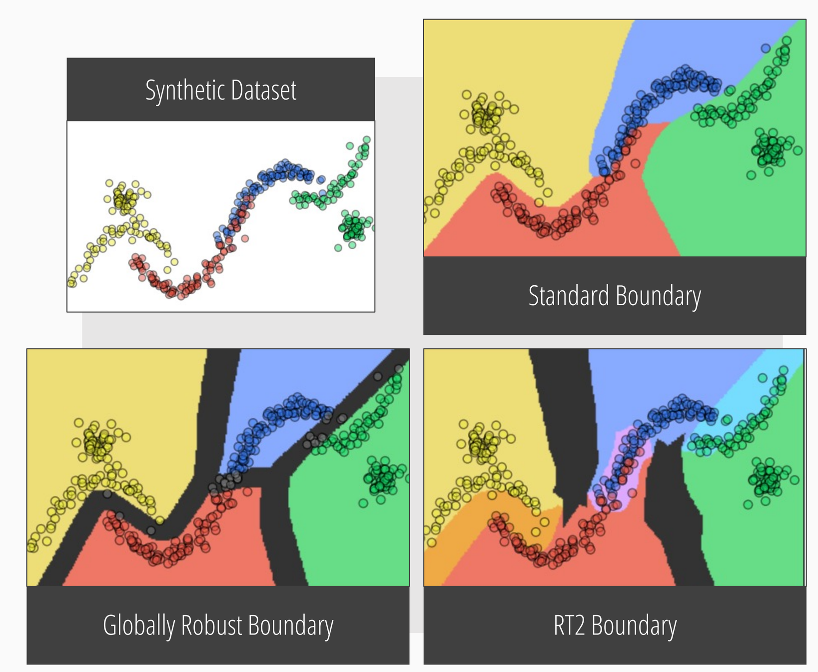


RTK robustness | we obtain a true relaxation by allowing the model to be top-k robust for *any* k up to K

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

$$\forall x'.\ ||x - x'||_p \le \varepsilon \quad \Longrightarrow \quad \exists\, k \le K :\ F^k(x) = F^k(x')$$

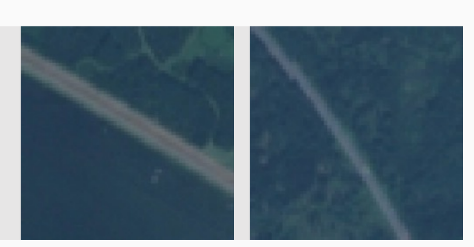
This *is* a relaxation of local robustness



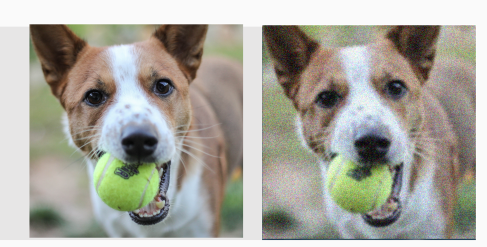
| Affity Robustness

Affinity robustness allows for extra control over which classes may be grouped together, to reflect the fact that some mistakes may be worse than others. E.g., in the example for **Issue 2**, we see that roads and rivers look similar in satellite images. On the other hand, some mistakes are more egregious, e.g., the dog/sports car example from earlier.

OK
road / river



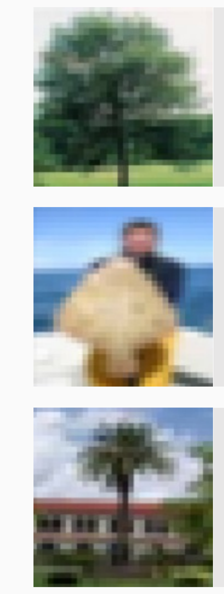
Not OK
dog / sports car



Summary of Results

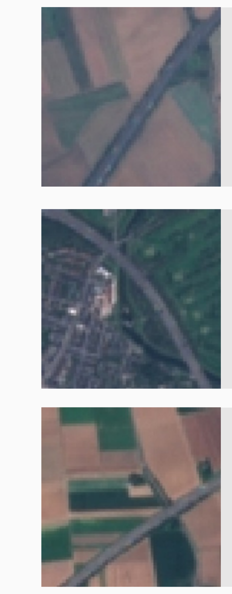
Training GloRo Nets[1] with RTK robustness typically results in logical groupings of classes into robust sets.

CIFAR-100 (RT5)



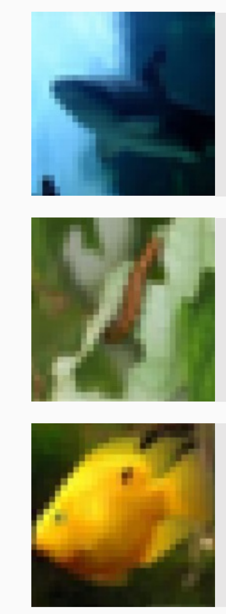
- **oak**, maple, willow, pine
- **flatfish**, man, trout, woman, girl
- **palm tree**, house

EuroSAT (RT3)



- **highway**, annual crop
- **highway**, residential buildings
- **highway**, permanent crop, annual crop

Affinity robustness can help improve the guarantees obtained by adding extra supervision.



RT5 Robustness	Superclass Affinity Robustness	
shark, sea	**shark**, ray	robust prediction set modified to match CIFAR-10 superclass
spider, **caterpillar**, butterfly	**caterpillar**, butterfly	prediction became correct after enforcing affinity robustness
aquarium fish, tulip, poppy	**aquarium fish**	robustness improved after enforcing affinity robustness

Relaxed robustness leads to fewer rejected points and thus better model performance.

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%

[1]Leino et al. ICML 2021

learn more

check out our talk and the full paper for more!
code available on GitHub

full paper
https://tinyurl.com/relaxed-robustness

implementation
https://github.com/klasleino/gloro