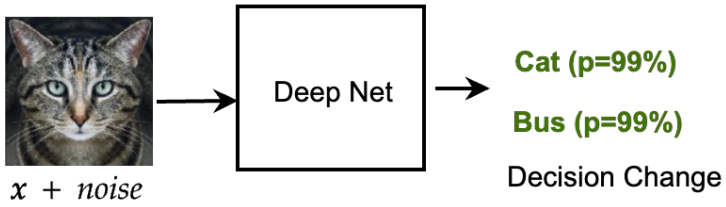
Detecting Brittle Decisions for Free: Leveraging Margin Consistency in Deep Robust Classifiers

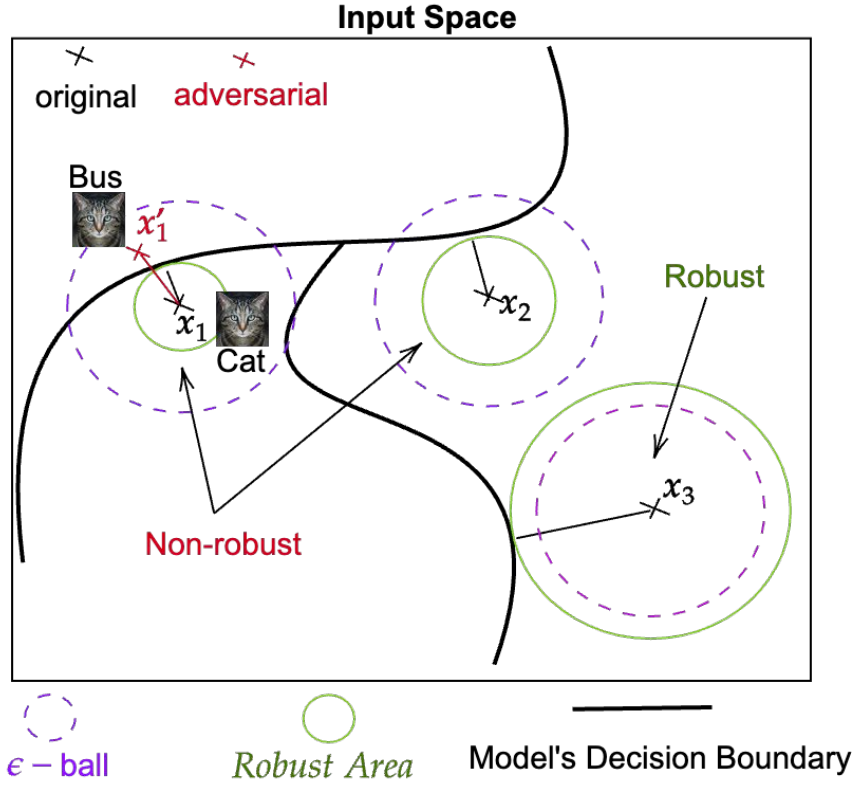
Jonas Ngnawé, Sabyasachi Sahoo, Yann Pequignot, Frederic Precioso, Christian Gagné

Problem

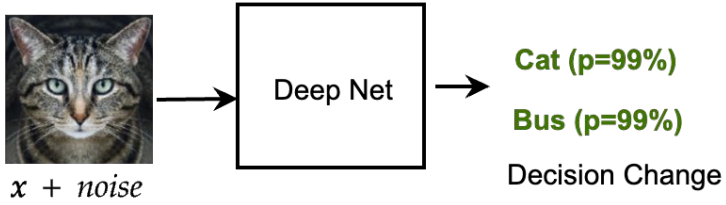


Insignificant input perturbations can change the decision.

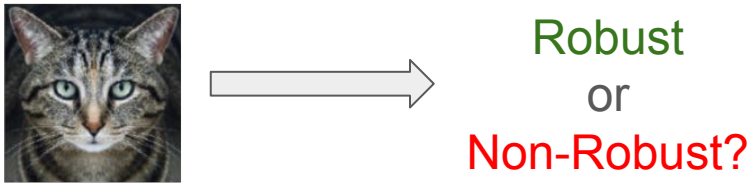
Local Robustness



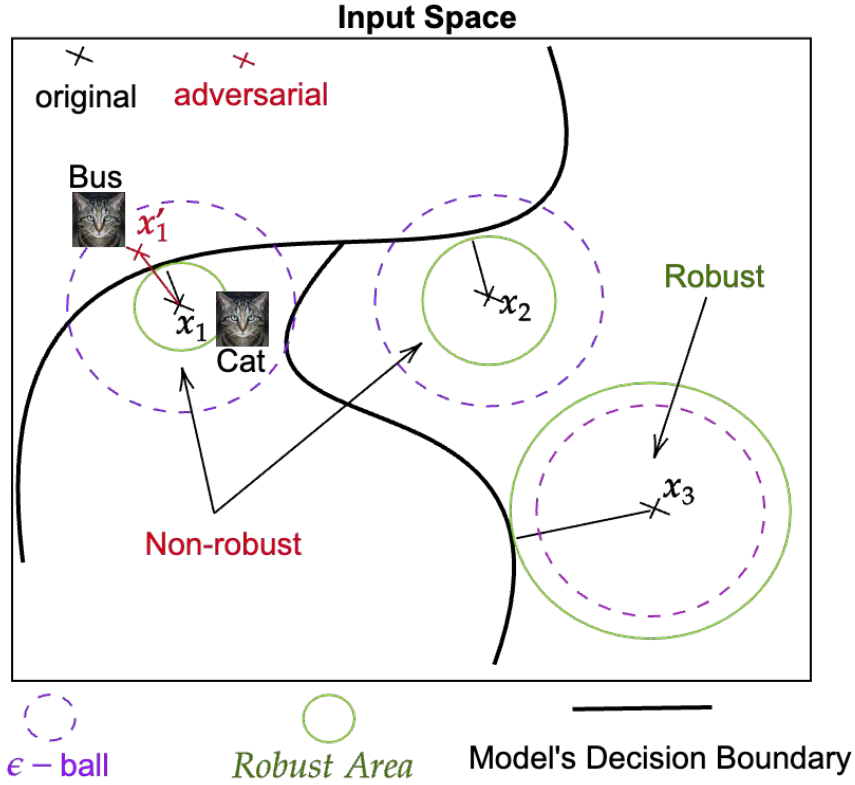
Problem



Insignificant input perturbations can change the decision.



Local Robustness



Problem

Robust
or
Non-Robust?

Input Margin Intractable for general Deep Nets.

Methods
- ➔ Adversarial Attacks
- ➔ Formal Robustness Verification

Problem

Input Margin Intractable for general Deep Nets.

Robust
or
Non-Robust?

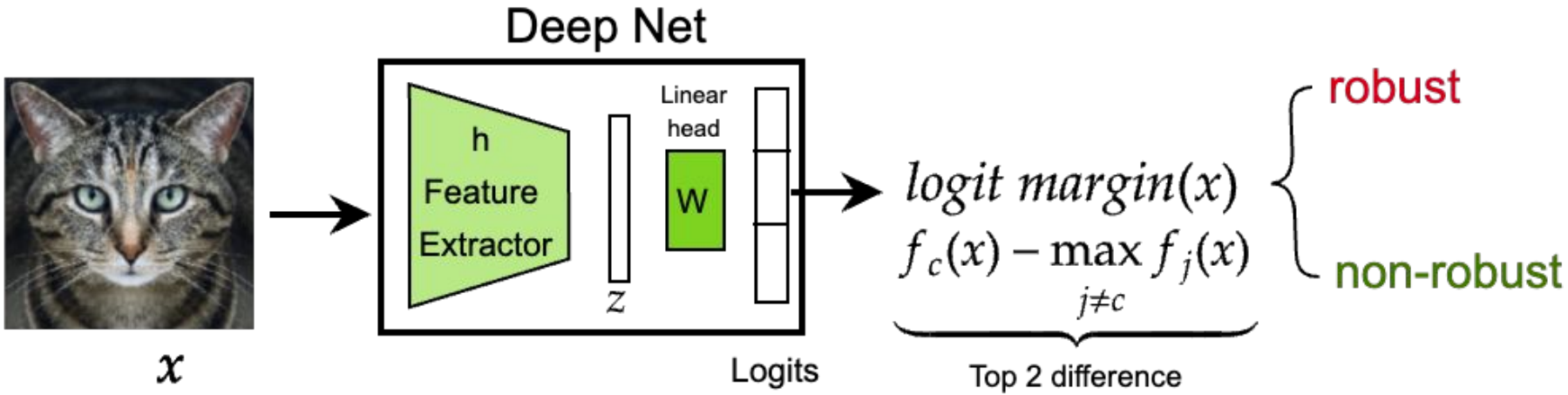
Methods
➔ Adversarial Attacks
➔ Formal Robustness Verification

Computationally Expensive
- Evaluate on large test sets or large models
- Real Time Deployment Decisions

Main Contribution

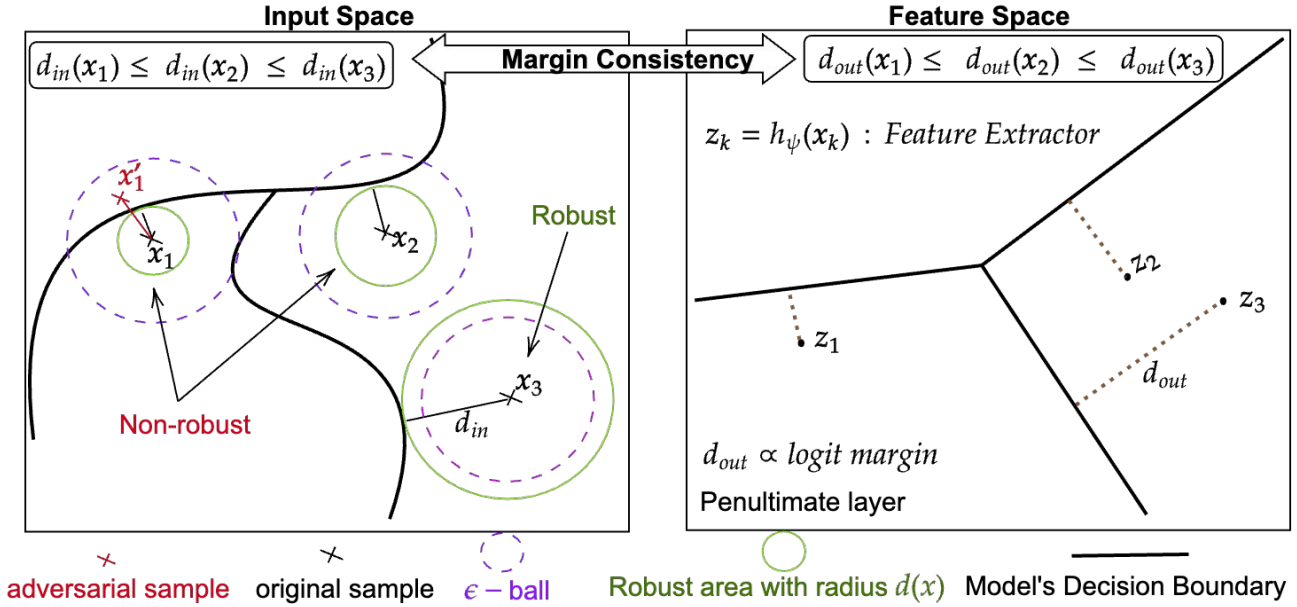
We introduce the concept of **Margin Consistency**

that allows using the logit margin: the difference between the top 2 logits as a proxy score for detection (local robustness score)



Margin Consistency

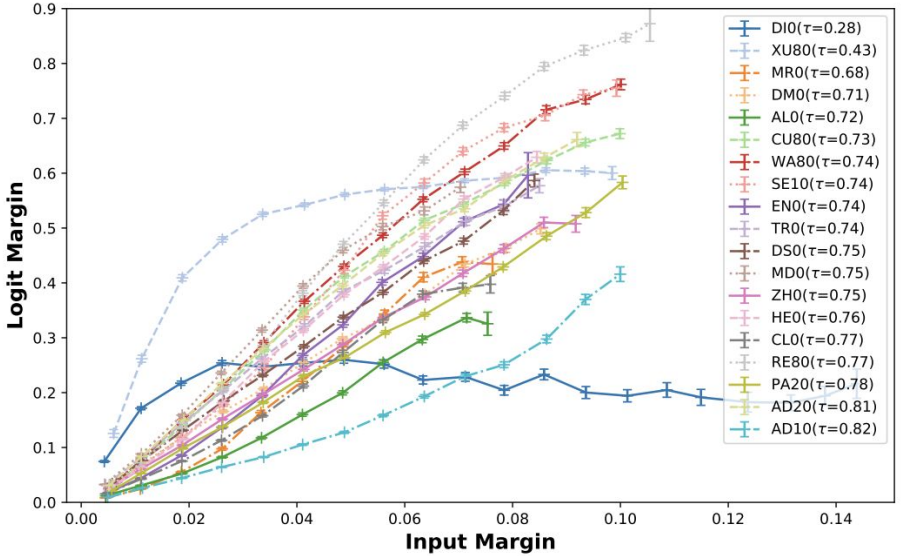
A model is **margin consistent** if there is a monotonic relationship between the input margin and the logit margin (rank correlation, Kendall tau)



Theorem: Margin Consistency is a necessary and sufficient condition to use the logit margin as a perfect score for non-robust samples detection.

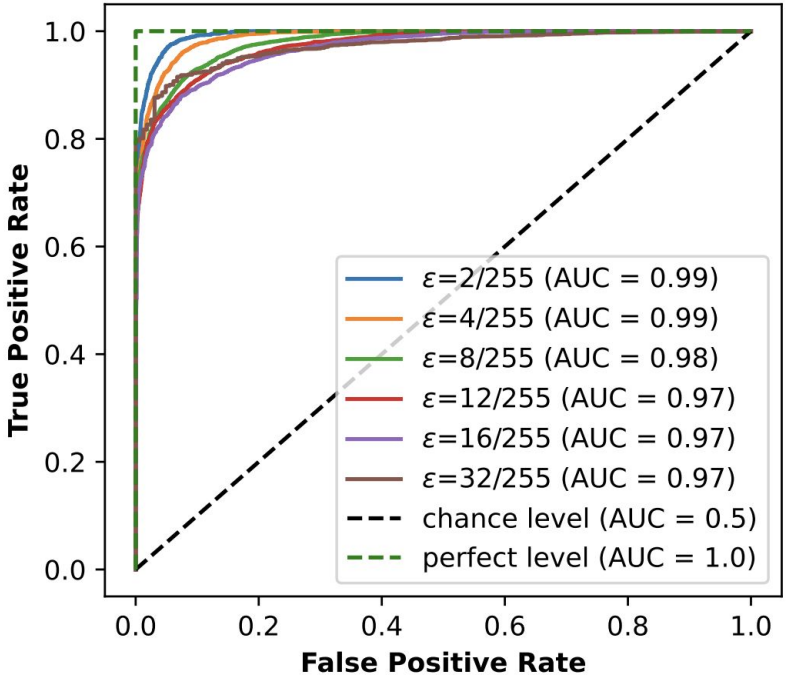
Evaluation and Results

- Investigation on robustly trained models in *Robustbench*, Linf norm (8/255)
- Most are strongly margin consistent



(a) CIFAR10

Kendall tau correlations



Detection performance

Evaluation and Results

Are robust models strongly margin consistent because of Lipschitz smoothness?

Evaluation and Results

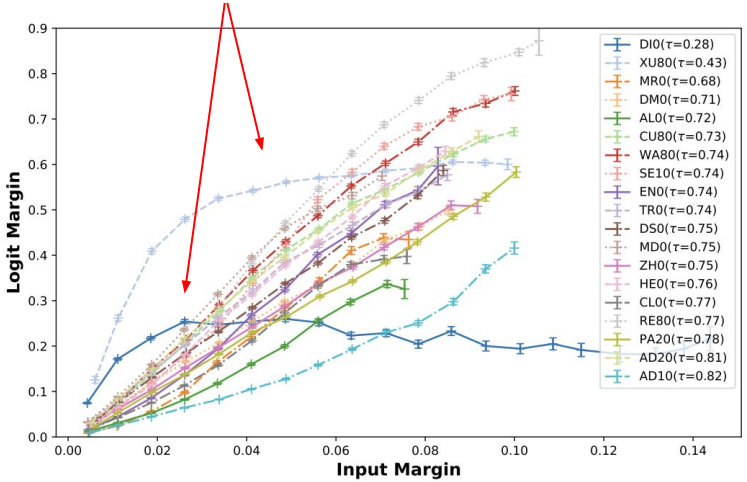
Are robust models strongly margin consistent because of Lipschitz smoothness?

No!

Lipschitz smoothness or Robustness does not imply margin consistency

Evaluation and Results

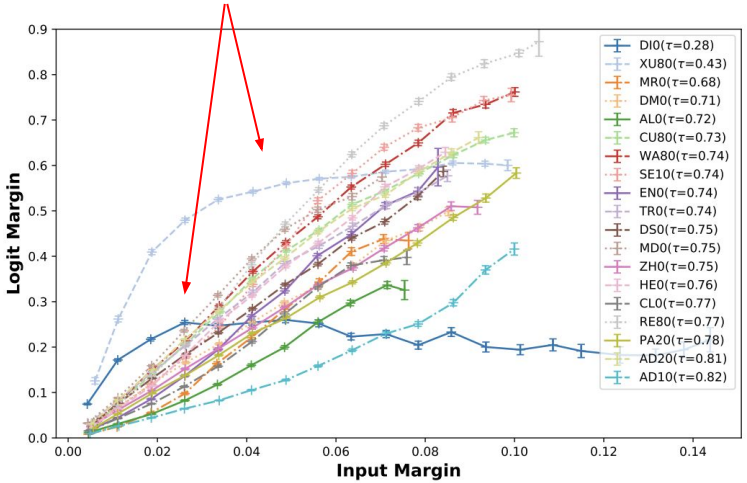
What can we do with these two models?



(a) CIFAR10

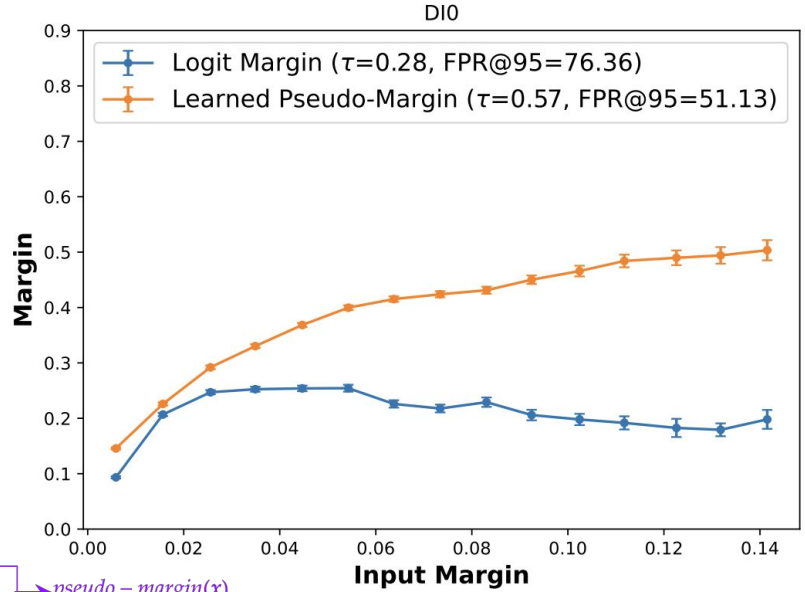
Evaluation and Results

What can we do with these two models?

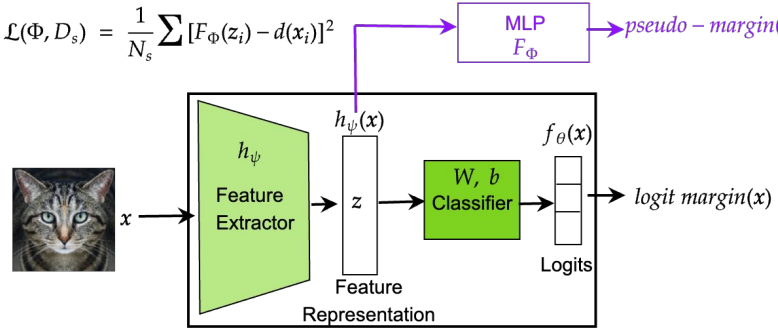


(a) CIFAR10

We can learn a better correlated pseudo-margin



$$\mathcal{L}(\Phi, D_s) = \frac{1}{N_s}\sum [F_\Phi(z_i) - d(x_i)]^2$$



Corbière et al. "Addressing failure prediction by learning model confidence." Neurips 2019

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!

UNIVERSITÉ LAVAL

UNIVERSITÉ CÔTE D'AZUR
Dare to create

Institut intelligence et données

Link to paper

Inria

CIFAR

DEEL
DEpendable & Explainable Learning

Mila