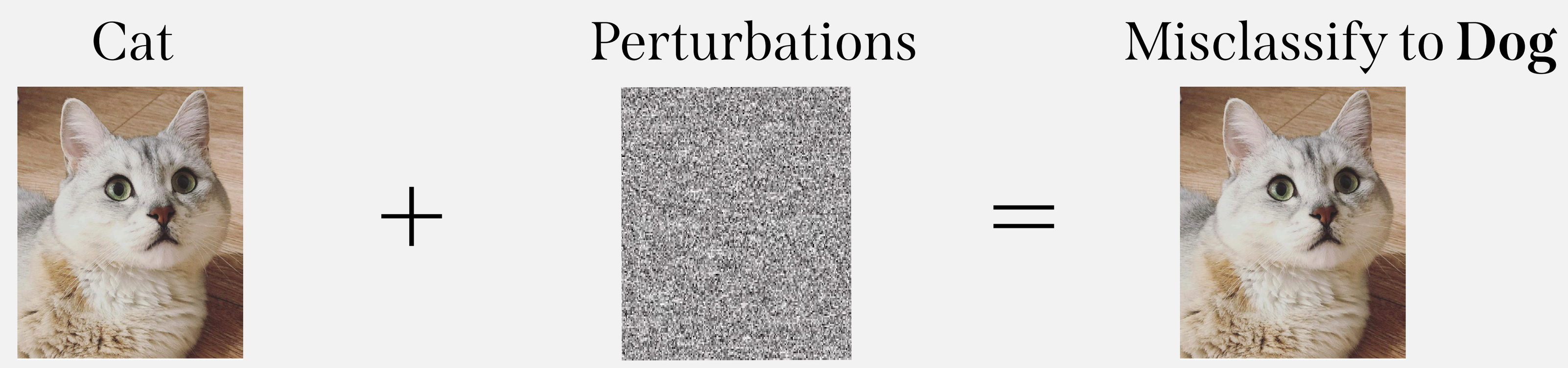
Exploiting Label Similarity to Enhance Verifiable Robust Classifiers

Shiqi Wang
Columbia University
tcwangshiqi@cs.columbia.edu

Kevin Eykholt, Taesung Lee, Jiyong Jang, Ian Molloy
IBM Research
Cyber Security Intelligence Group

✻ *Adversarial Examples & Verifiable Training*

Adversarial examples: Minor perturbations will cause mispredictions



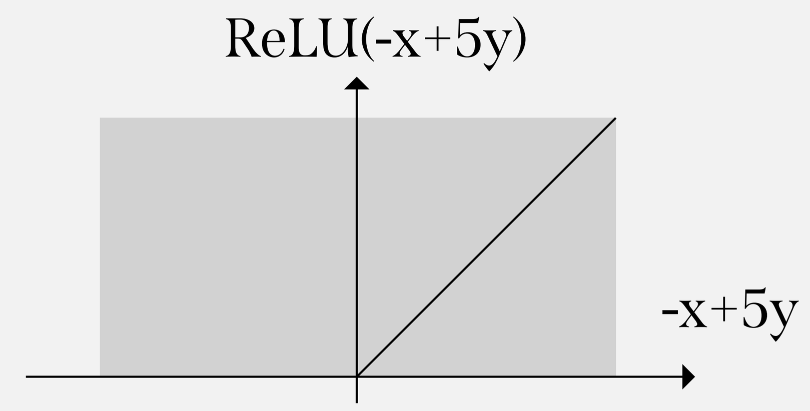
Interval Analysis: Verify the absence of adversarial examples

Input: x∈[4,6] & y∈[1,5];

Linear layer: -x+5y∈[-1,16]

ReLU layer: ReLU(-x+5y)∈[0,16]

Output layer: Cat-Dog∈[0,16]>0?



Verifiable Training: Training networks with verification methods to learn provable robustness guarantee against adversarial examples.

Even SOTA verifiable training **CROWN-IBP**[1] has very poor performance. For instance with CIFAR10, $L_\infty \leq 8/255$:

46% clean accuracy & **33%** verified accuracy

✻ *Insight of Our Work: Label Similarity!*

Main limitation: a single robustness distance for all classes=>
The maximal distance of the robustness is limited by similar labels:

[Similar] Cats vs Dogs: 29% clean accuracy

[Dissimilar] Cats vs Cars: 56% clean accuracy

Adaptive robustness accounting for label similarity!

Dogs vs Cars: Dissimilar
=> Able to learn **large** robustness distance

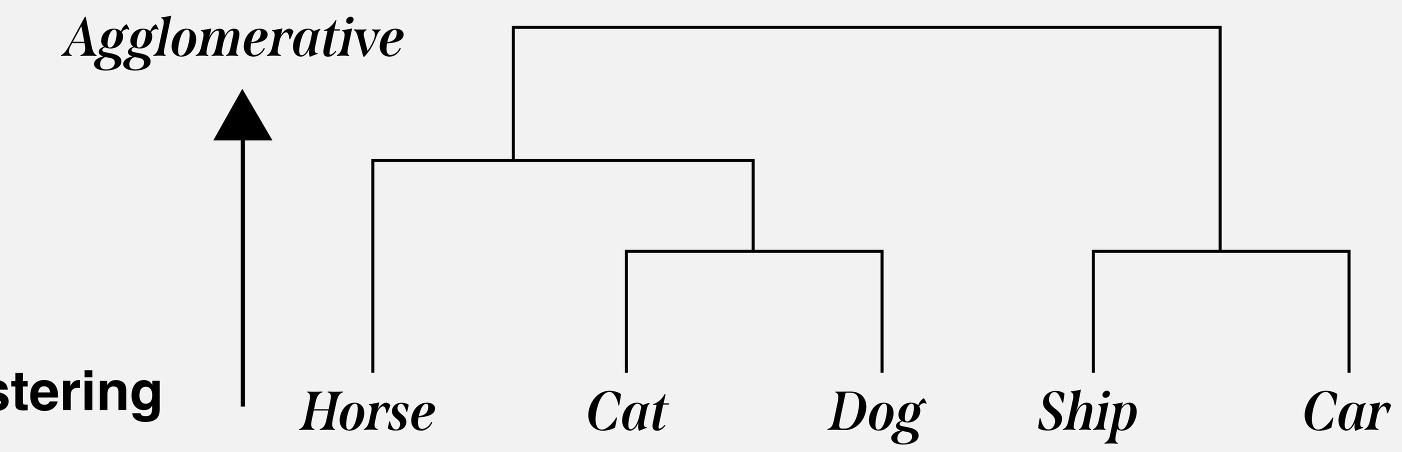
Cats vs Dogs: Similar
=> Only tolerate **small** robustness distance

✻ *Smart Label Grouping*

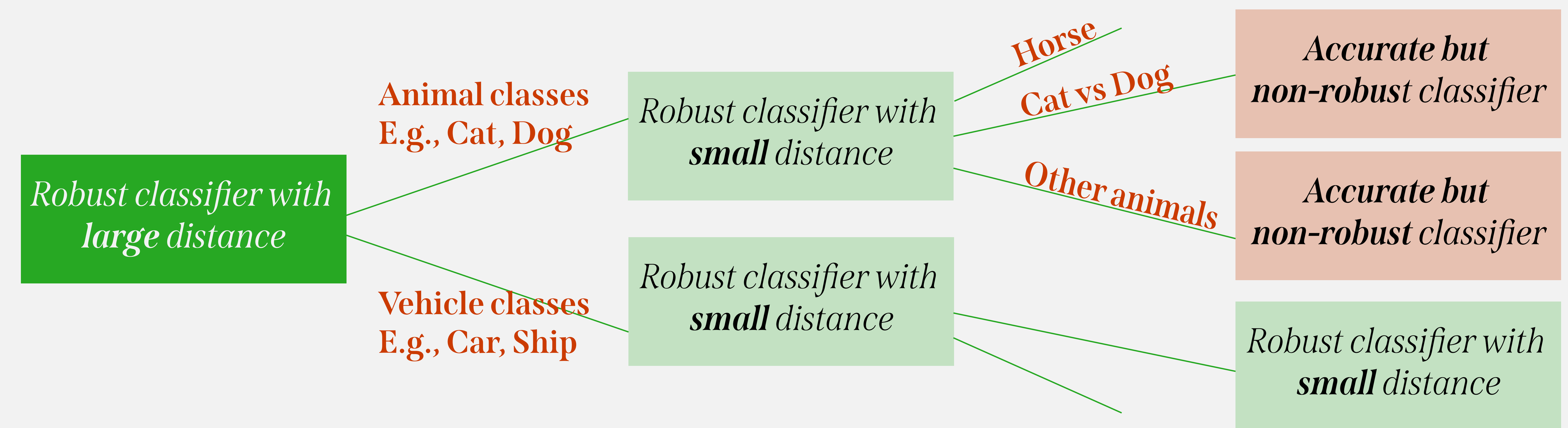
Step1. Naturally train a network

Step2. Extract the last layer weights

Step3. Label grouping with **Agglomerative Clustering**



✻ *Method1: Neural Decision Tree (NDT)*



✻ *Method2: Inter-Group Robustness Prioritization (IGRP)*

$$L_{IGRP} = L_{inner} + L_{outer}$$

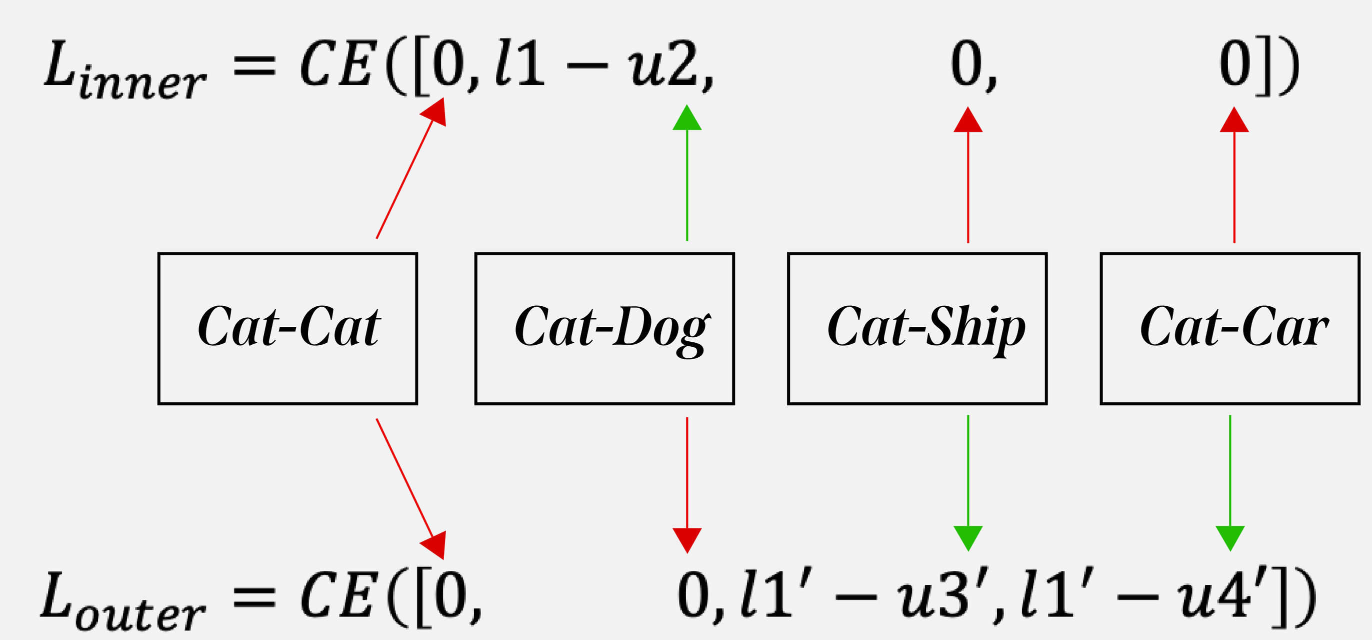
Inner loss: similar classes & small distance

Outer loss: dissimilar classes & large distance

For instance, true label is Cat

Small distance logits range [l, u]

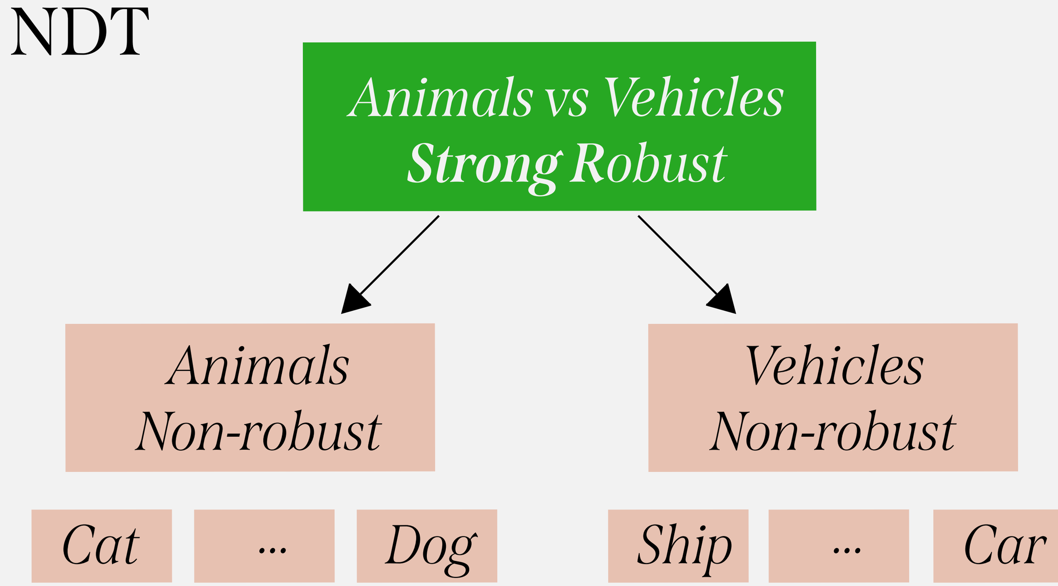
Large distance logits range [l', u']



✻ *Experimental Results*

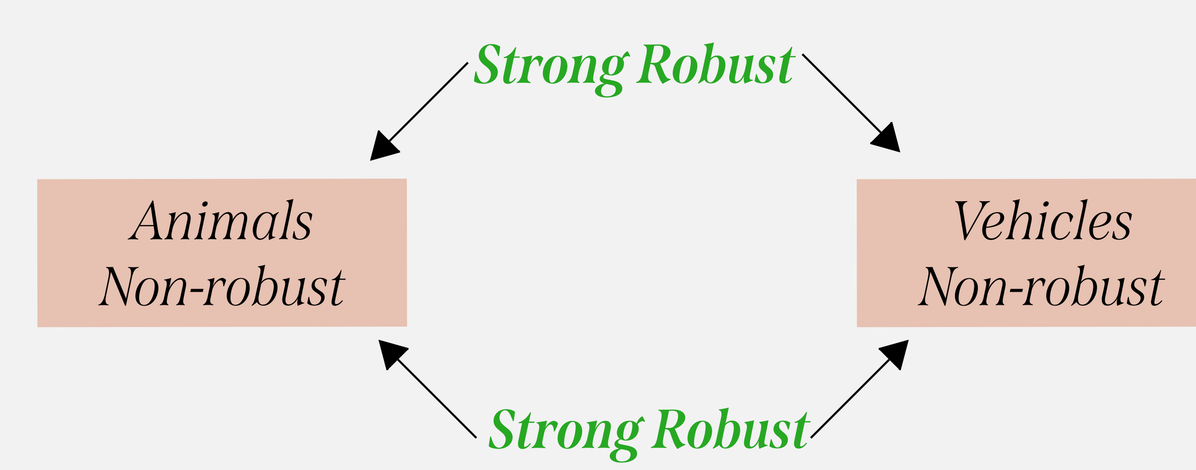
Dataset: F-MNIST, CIFAR10, CIFAR100
Verified Acc*: Verified error for animal vs vehicle, $L_\infty \leq 8/255$

NDT



Over CROWN-IBP:
+16% Clean Acc
+3% Verified Acc*

IGRP



Over CROWN-IBP:
+9% Clean Acc
Similar Verified Acc*

IBM

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK