Training image

Robust Features: dog
Non-Robust Features: dog

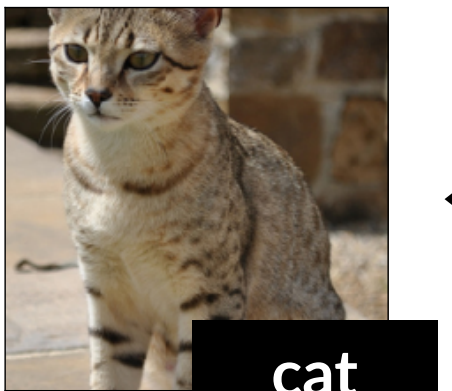
Adversarial example towards “cat”

max P(cat)

Relabel as cat

Robust Features: dog
Non-Robust Features: cat

Train



Evaluate on original test set

good accuracy