SRN Aleksander Madry

Machine Learning: A Robustness Perspective

Is it truly ready for real-world deployment?

- can be manipulated to exploit host
- ex: tesla car drive straight into overturned car

RobustML: A Challenge

- ImageNet: An ML Home Run → over 1 mil images so that computer can predict image
  - ↓ error over the years
  - Reflect:
    - A limitation of the (Supervised) ML Framework

Measure of performance: fraction of mistakes during testing

But: In reality, the distribution we use ML are NOT the ones we train it on

What can go wrong?

ML Predictions are (mostly) accurate
but brittle

ex: pig (91% w/ confidence of recognition) but w/ a little noise in image, the computer but think its an airliner

ex: rotation + translation suffices to fool state-of-the-art vision models
- data augmentation does NOT seem to help here

Why is the brittleness of ML a problem?
→ Security (recognize wrong criminals w/ sunglasses or noise)
→ safety (car, miss something dangerous w/ self-driving cars)
→ ML Alignment (need to understand the "failure modes" of ML)

ML pipeline (via adversarial lens)

data collection → training → inference → deployment

Is ML inherently not reliable?

No! But need to rethink it?

Why are our models brittle?

→ d → ∞ (wierd dimensions of objects)
→ only optimizing for avg case but not worse case

Why are adv. perturbations bad?

dog + meaningless perturbation = cat

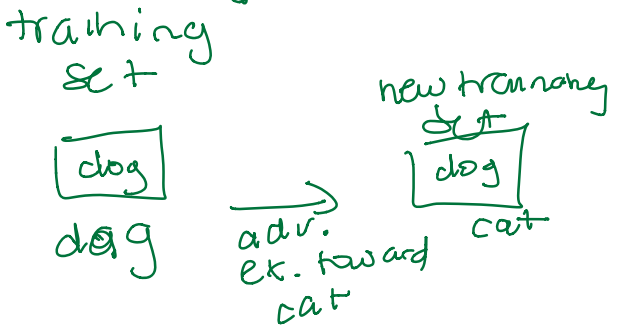
(we can tell it is still a dog, but ML thinks its a cat)

- ML has no concept of dog

↙ image is meaningless

↘ classes are meaningless

Are adv. perturbations just meaningless?



1) Make adversarial example towards the other class
2) Relabel the image as the target class
3) Train w/ new dataset but test on the original test set

So we train on a "totally mislabeled" dataset but expect performance on a "correct" dataset

What will happen?

- we get a nontrivial accuracy on the original classification task

What's going on?

The Robust Features Model

Robust Features:
correlated w/ label even w/ adversary

Non-robust features:
Correlated w/ label on avg but can be flipped w/ adv

All robust models are misleading but...