Intro to Adversarial Attacks

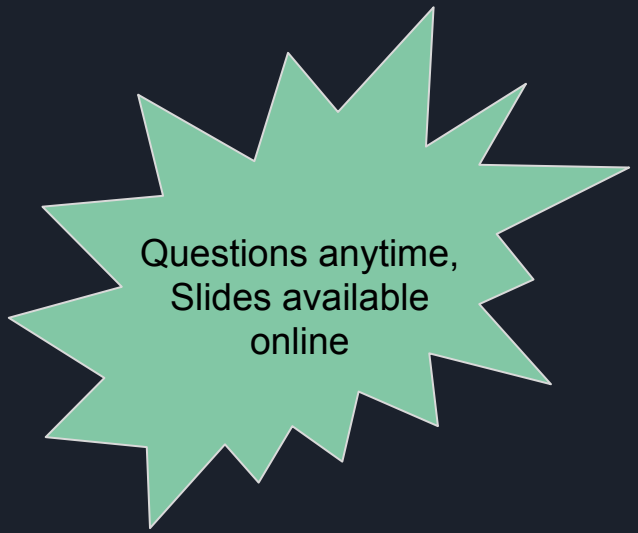
Gokula Krishnan Santhanam

$whoami

- Intel Innovator
- Master’s Student @ ETH Zurich
  - Worked on Generative Models for Astronomy (space.ml)
- Google Research Summer Intern
- @gokstudio, sgokula@ethz.ch

Outline

- What's an adversarial attack?
- Why should you care?
- Common Examples
  - Hands-On using cleverhans
- How to defend?
- Open Questions and Teasers



Adversarial? Attack??

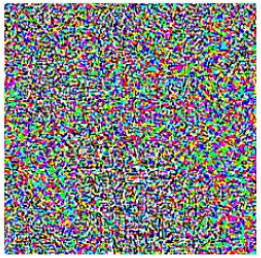
- Adversary - Someone with malicious intent
  - Want to find loop-holes in the systems
  - Might want to exploit it for self-benefit
  - Might want to hurt users of system (targeted / untargeted)
- Here, an image that's visually similar to humans but results in drastic changes in a DNN prediction*

* this is a very narrow definition to get started, lots of questions on how to best define for any network, for any task, for any data type (image, text, audio .etc)

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



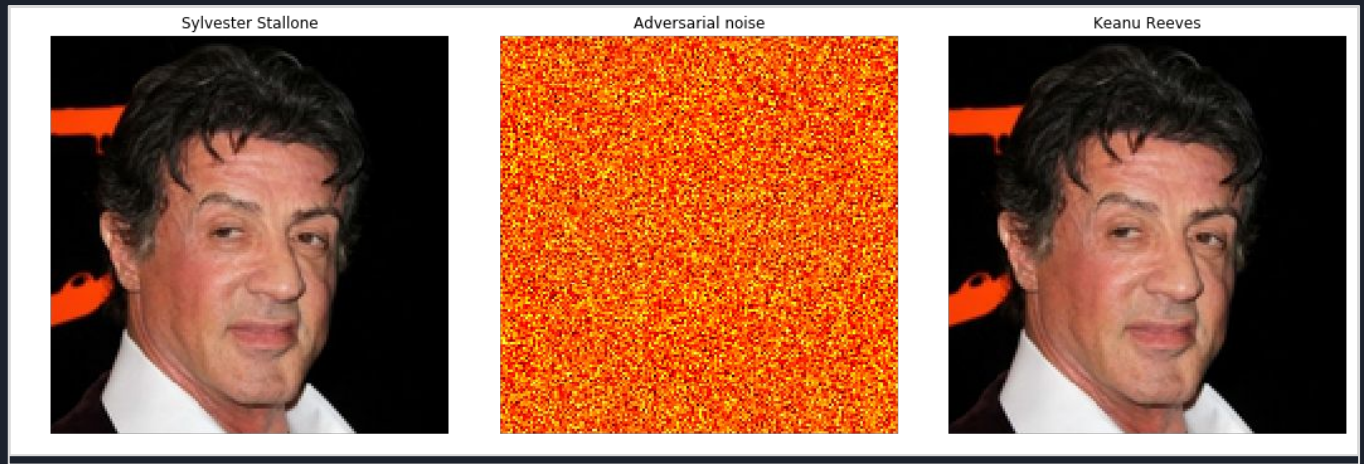
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

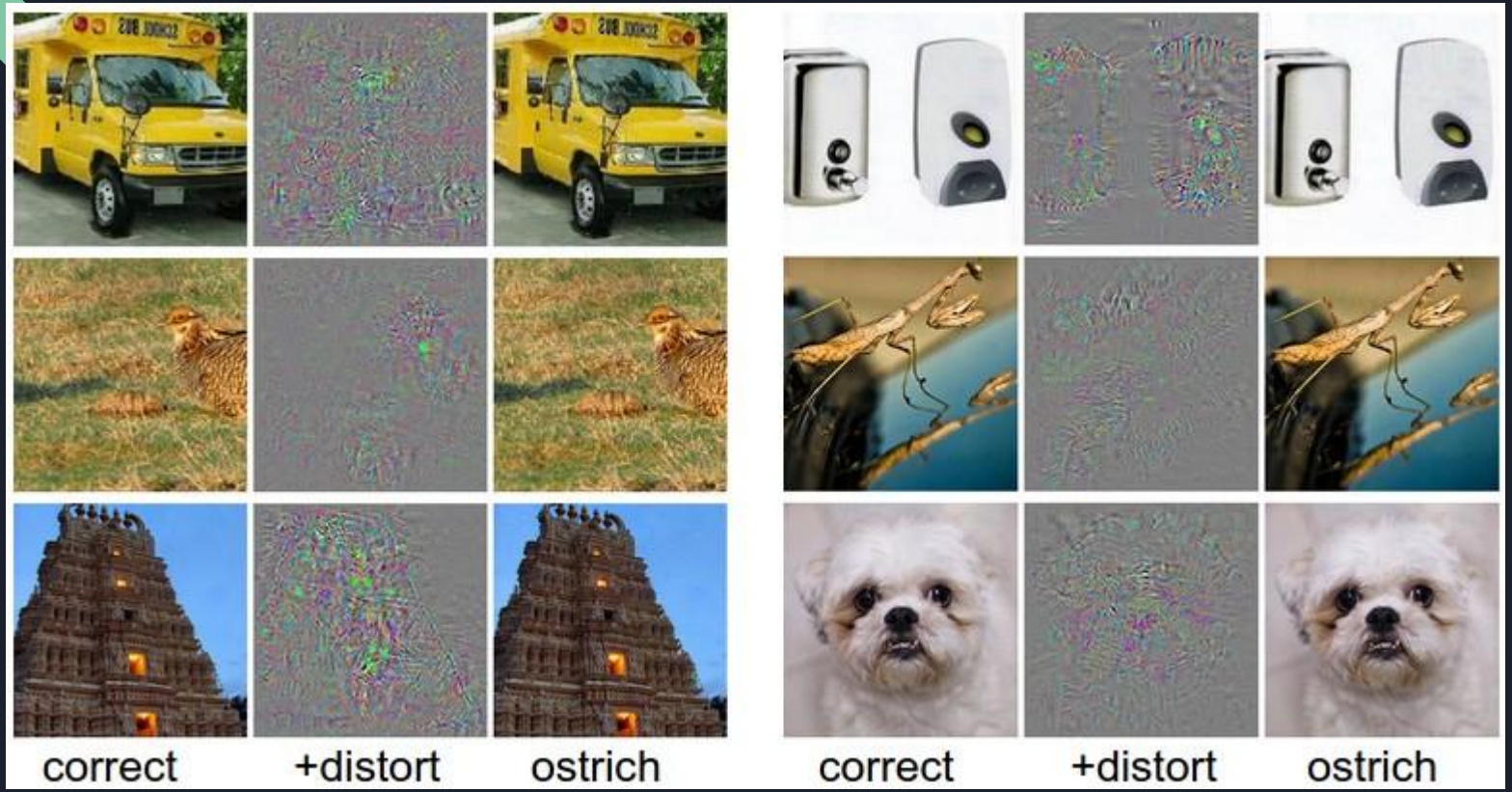
"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence




correct
+distort
ostrich
correct
+distort
ostrich

Why should I care?

- Small changes in X-Ray / MRI images could drastically change treatment
- Self-driving cars - speed limit 20 km/h to 200 km/h
- Erroneous predictions - lawsuits
- Interpretability of results
- Philosophically, are DNNs actually learning something or is it just good at pulling wool over our eyes?
    - How much can we promise?
        - Avoid fooling investors that are interested in your work


STOP
STOP
HATE

From: https://www.labsix.org/physical-objects-that-fool-neural-nets/

Are Adversarial Attacks only for DNN Classifiers?

- Logistic Regression
- SVM
- k-NN Classifiers
- Decision Trees
- Even for RL Policies

Some definitions

We say x_{adv} is an adversarial sample generated from x, if

Imperceptible change

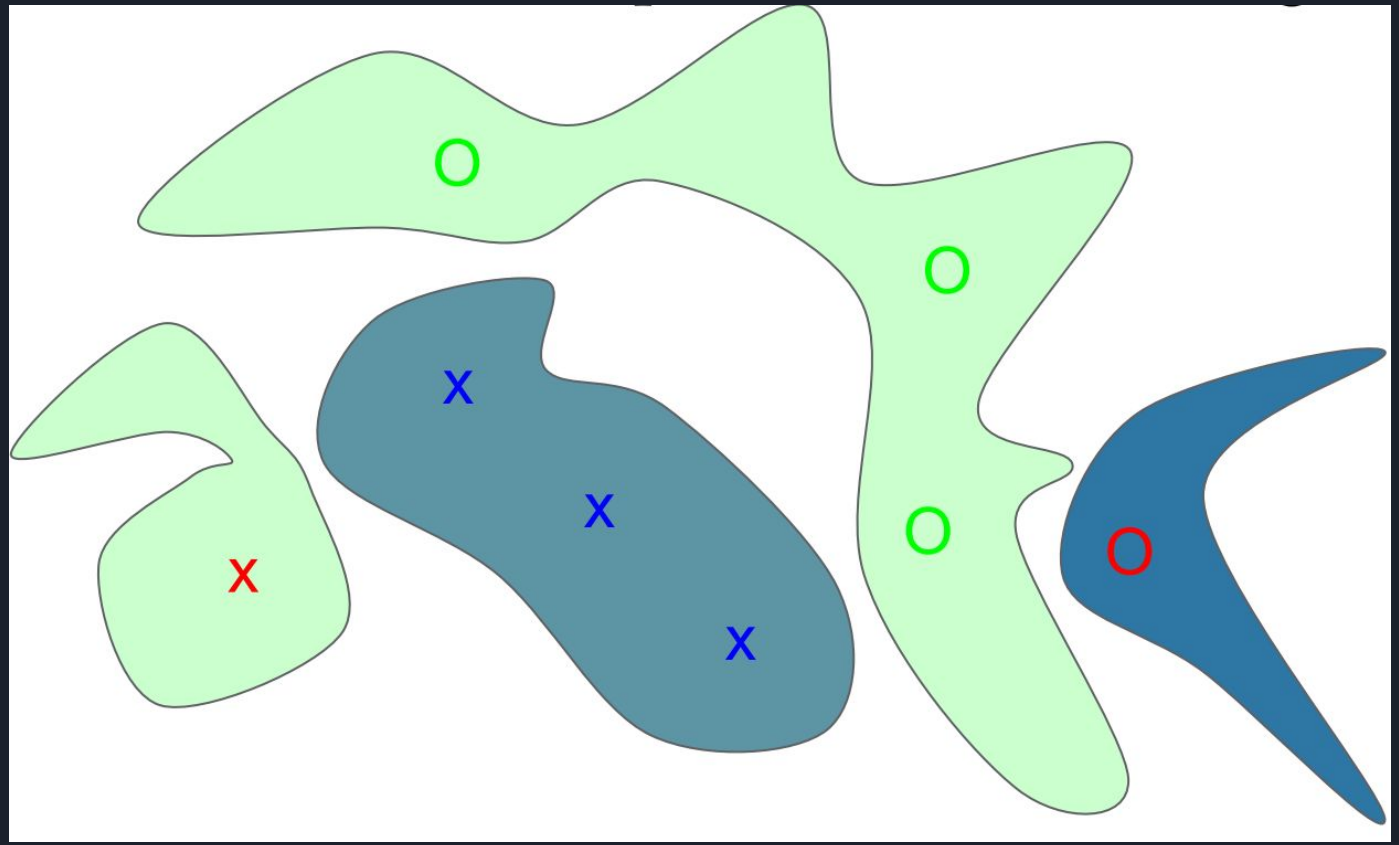
$|| x_{adv} - x || < \varepsilon$, and

For some model f (say a DNN classifier),

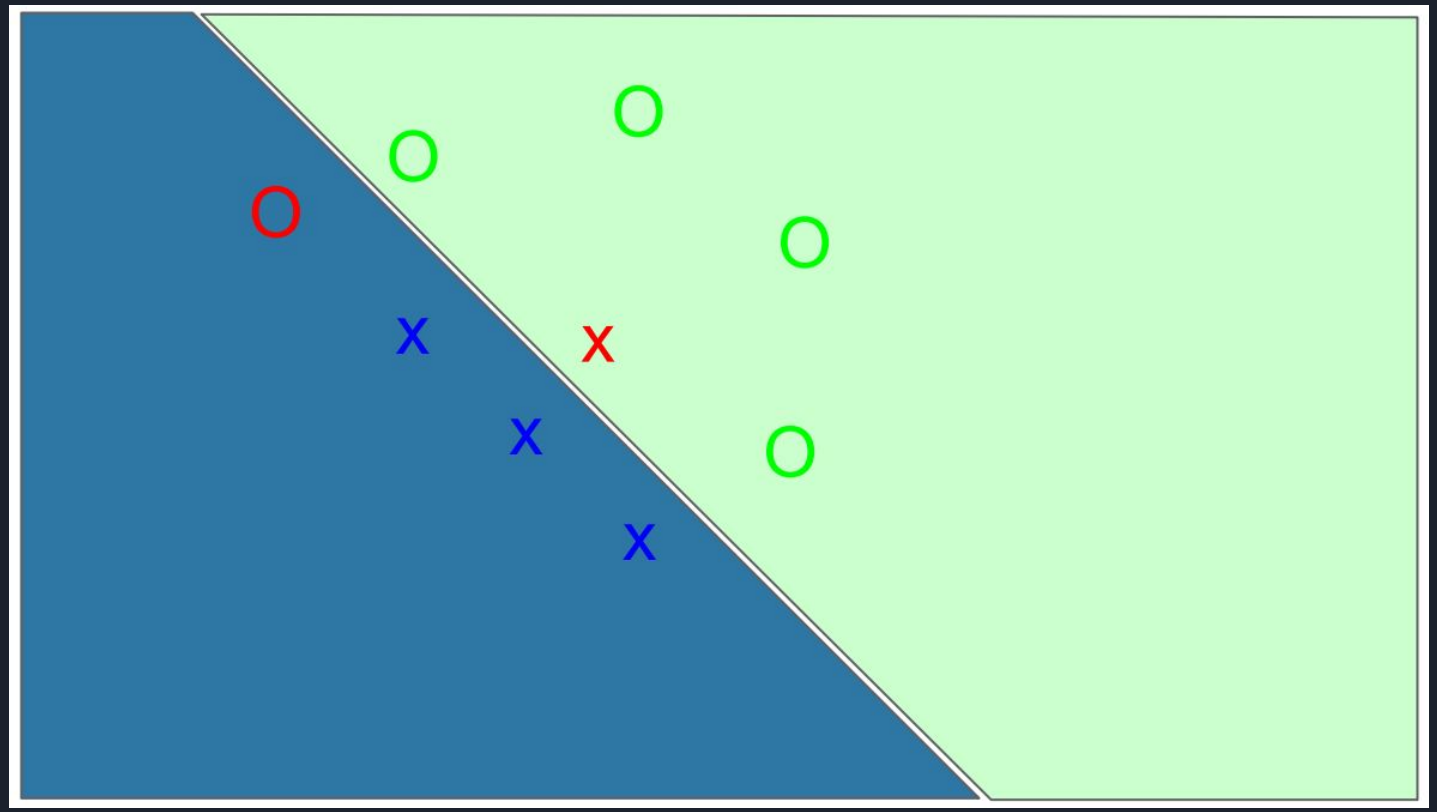
Significant change in output

$f(x_{adv})$ is very different from f(x)

Adversarial Samples due to overfitting



Adversarial Samples due to underfitting



DNN Classifier

Similar to examples before, but

1. Instead of 2D, imagine 1,000,000 D*
2. Decision boundaries can get very complicated
    a. Some regions are almost linear (ReLU, Sigmoid activations show linear behavior)
    b. Other regions can be very irregular
3. Hierarchical nature cause cascading effects

* if you can do this, teach me master!

Types of Attacks

Black Box Attacks

- Only the predictions from model are available
- Easy to transfer attacks across models
- Approximate black-box with your own model, create adv. attack on it, use it

White Box Attacks

- Both predictions and weights of the model are available
- Less easy to transfer attacks across models

Popular Attacks

- Fast Gradient Sign Method (FGSM)
- Projected Gradient Sign Method (PGDM)
- Basic Iterative Method (BIM)
- Carlini-Wagner L2 Method (State-of-the-Art)

Fast Gradient Sign Method

$$J(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) \approx J(\boldsymbol{x}, \boldsymbol{\theta}) + (\tilde{\boldsymbol{x}} - \boldsymbol{x})^\top \nabla_{\boldsymbol{x}} J(\boldsymbol{x})$$

Maximize

$$J(\boldsymbol{x}, \boldsymbol{\theta}) + (\tilde{\boldsymbol{x}} - \boldsymbol{x})^\top \nabla_{\boldsymbol{x}} J(\boldsymbol{x})$$

subject to

$$||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon$$

$$\Rightarrow \tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{x})\right).$$

Hands-On!

CleverHans library

- Built on top of TensorFlow / Keras
- Maintained by Ian Goodfellow et. al
- Has most of the widely used attacks
- Easy to use (~ 3 lines to use an attack)
- Actively developed

Defense against these attacks

It’s much easier to attack than to defend

- Adversarial training
  - Augment dataset with adversarial samples as well
  - One of the best defenses
- Defensive Distillation
- Thermometer Encoding
- … more in ICLR and ICML papers

What’s happening now?

- Adversarial Attacks on Humans
- Better Understanding of source of Adversarial Attacks
- Adversarial Text
- So much more!

Summary

- Introduction to adversarial attacks
- Why you should care about them
- Cleverhans demo
- Defenses

Questions?

Thanks!

References

- Ian Goodfellow's Talks
- Papers from ICLR
- Karpathy's blogpost: http://karpathy.github.io/2015/03/30/breaking-convnets/
- Cleverhans docs and repository