# Adversarial examples in Deep Neural Networks

Luiz Gustavo Hafemann
Le Thanh Nguyen-Meidine

# Agenda

**Introduction**

**Attacks and Defenses**

**NIPS 2017 adversarial attacks competition**

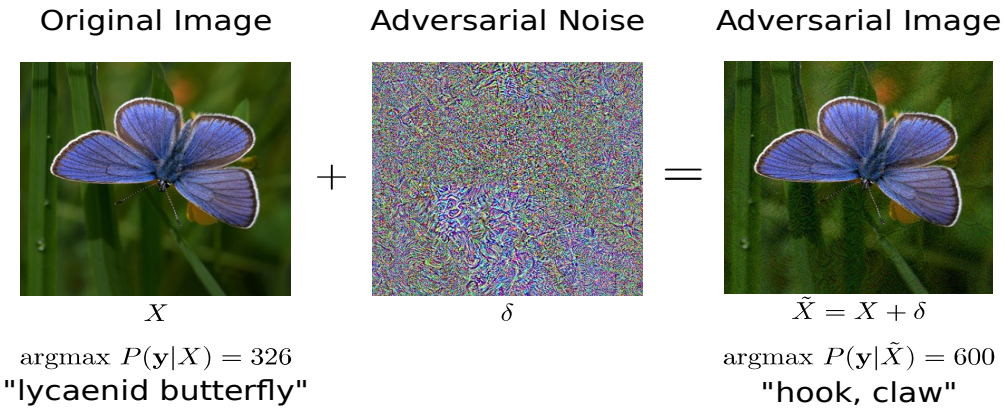
**Demo**

**Discussion**

# Introduction

## Adversarial examples:

Examples that are similar to examples in the true distribution, but that fool a classifier




* Note: most examples in this presentation are for images, but the problem applies to other domains, such as speech

# Examples

**https://openai-public.s3-us-west-2.amazonaws.com/blog/2017-07/robust-adversarial-examples/iphone.mp4**

**http://www.labsix.org/media/2017/10/31/video.mp4**

# Introduction

## Adversarial examples pose a security concern for machine learning models

- An attack created to fool one network also fools other networks . Szegedy et al. (2013)
- Attacks also work in the physical word. Kurakin et al (2016), Athalye et al (2017)
- For Deep Neural networks, it is very easy to generate adversarial examples but this issue affects other ML classifiers.

# Introduction

## Adversarial examples pose a security concern for machine learning models

- Although many defense strategies have been proposed, they all fail against strong attacks, at least in the white-box scenario.
- Even detecting if an image is an adversarial is hard. (Carlini and Wagner, 2017)

# Definitions

**An example $\tilde{X}$ is said adversarial if:**

- It is close to a sample in the true distribution:

$$D(X, \tilde{X}) \leq \epsilon$$

- It is misclassified

$$\operatorname{argmax} P(\mathbf{y}|\tilde{X}) \neq y_{\text{true}}$$

- It belongs to the input domain. E.g. for images:

$$0 \leq \tilde{X} \leq 255$$

# Notion of “similarity”

**To measure the similarity between samples:**

- A good measure between samples is still an active area or research. Commonly, researchers use:
- L_2 norm (euclidean distance):

$$\left\| \tilde{X} - X \right\|_2$$

- L_infinity norm (maximum change to any pixel in the image):

$$\left\| \tilde{X} - X \right\|_\infty = \max_{ij} |\tilde{X}_{ij} - X_{ij}|$$

# Threat model

**We need to consider the attacker's:**

- **Capability**
- **Goal**
- **Knowledge**

# Types of attack

## According to the attacker's goal:

**Non-targeted attacks: attacker tries to fool a classifier to get any incorrect clas**

$$\text{argmax}\, P(\mathbf{y}|\tilde{X}) \neq y_{\text{true}}$$

**Targeted attacks**: attacker tries to fool a classifier to predict a particular class

$$\text{argmax}\, P(\mathbf{y}|\tilde{X}) = y_{\text{target}}$$

# Threat model

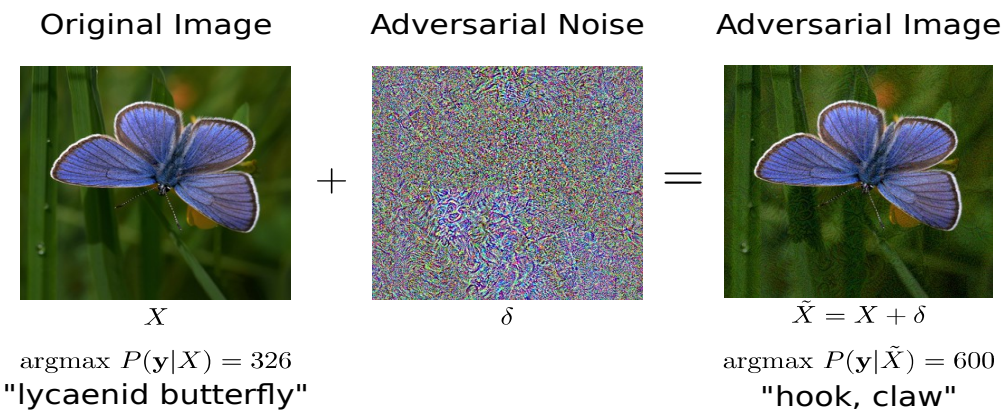
## According to the attacker's knowledge:

- **White-box attacks**: attacker has full knowledge of the classifier (e.g. weights for a neural network)
- **Black-box attacks**: attacker does not have access to the target classifier. In this case, the attacker trains its own classifier (using data from the same distribution), and creates attacks based on this version.

# Recap

## Adversary wants to fool the classifier

- By crafting a noise $\delta$ such that $X + \delta$ is misclassified
- With a small $\delta$
- With full knowledge (white-box) or not (black-box)



# Attacks

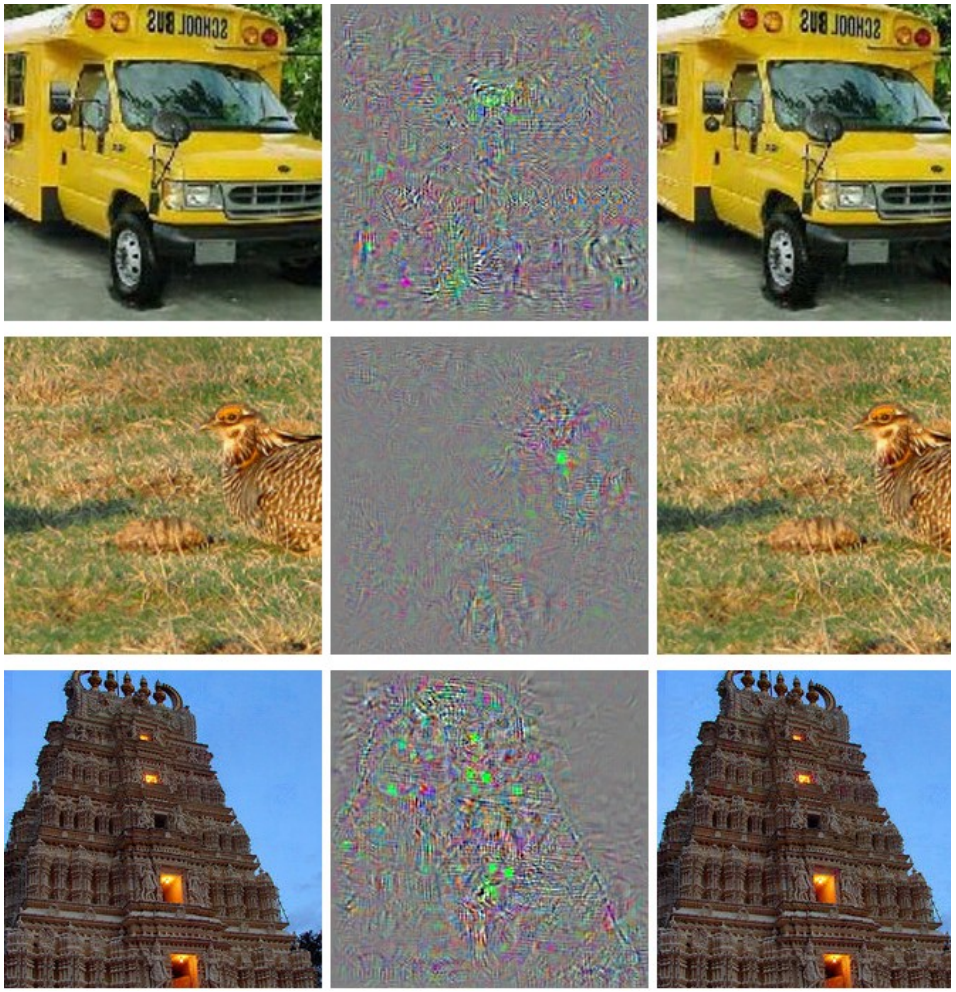
**Box constrained optimization (Szegedy et al):**

$$\begin{aligned} &\text{Minimize} && \|\delta\|_2 \\ &\text{subject to} && \underset{j}{\operatorname{argmax}}\, P(y_j|X+\delta) \neq y_i \\ & && 0 \leq X+\delta \leq 255 \end{aligned}$$

**> Generates adversarial images that are *very* close to the original samples**

# Attacks

## Examples

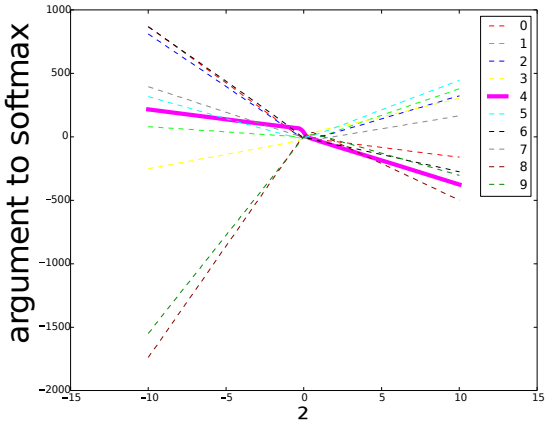


# Attacks

## Fast gradient sign (Goodfellow et al):

This article shows that adversarial examples occupy half-spaces of the input space, and not small pockets.

They also show that the output of the network has a very (piecewise)-linear nature:




$$\tilde{X} = X + \epsilon \text{sign}(\nabla J(x, y))$$

# Failed defenses

**“It’s common to say that obviously some technique will fix adversarial examples, and then just assume it will work without testing it” - Ian Goodfellow**

**What does not solve the problem:**

- Ensembles
- Voting after multiple saccades (e.g. crops of the image)
- Denoising with an autoencoder

# Defenses that somewhat work

## Adversarial training (goodfellow et al, 2015)

Train the network with both "clean" and "adversarial" examples:

$$\tilde{J}(\theta, X) = \alpha J(\theta, X) + (1 - \alpha) J(\theta, X + \epsilon \text{sign}(\nabla J(\theta, X)))$$

Original loss

Loss of misclassifying an adversarial example

# Defenses that somewhat work

## Ensemble adversarial training

Adversarial training has a problem that it uses the model under training to generate the adversarial samples.

For ensemble training, use multiple networks to generate the adversarial samples:

$$\tilde{J}(\theta, X) = \alpha J(\theta, X) + (1 - \alpha) J(\theta, x_{\text{adv}})$$

Where $x_{\text{adv}}$ Is generated (in each step) by a different model.

# The NIPS 2017 adversarial competition

- 3 competitions: targeted, non-targeted attacks, defenses
- All attack submissions are run against all defense submissions (in three “development rounds” plus a final round)
- Time constraints (500s to process 100 images, 1 GPU available). No internet access
- Attack constraints: maximum $\|\delta\|_\infty$

# The NIPS 2017 adversarial competition

## Our submission

- Re-formulate the optimization problem to constraint on $\delta$, instead of minimizing it. Minimize $\mathrm{P}(\mathrm{y}_{\mathrm{true}}|\tilde{X})$ instead
- Generate attacks using box-constrained optimization
- Attack an ensemble of models

# The NIPS 2017 adversarial competition

## Non-targeted attack

$$\begin{aligned} \text{Minimize} \quad & \log P(y_{\text{true}}|X+\delta) \\ \text{subject to} \quad & \|\delta\|_\infty \leq \epsilon \\ & 0 \leq X+\delta \leq 255 \end{aligned}$$

# The NIPS 2017 adversarial competition

## Non-targeted attack

$$\begin{aligned} &\text{Minimize} && \log P(y_{\text{true}}|X+\delta) \\ &\text{subject to} && \max(0-X,-\epsilon) \leq \delta \leq \min(255-X,\epsilon) \end{aligned}$$

# The NIPS 2017 adversarial competition

## Targeted attack

$$\begin{aligned} &\text{Minimize} && -\log P(y_{\text{target}}|X+\delta) \\ &\text{subject to} && \max(0-X,-\epsilon) \leq \delta \leq \min(255-X,\epsilon) \end{aligned}$$

# The NIPS 2017 adversarial competition

- **Attacked an ensemble of networks:**
  - Inception v3, v4
  - Adversary trained inception_v3, inception_resnet_v2
  - Ensemble Adversary trained inception_resnet_v2
  - DenseNet

- **Instead of minimizing log probabilities, minimize the logits (network output before softmax)**

# The NIPS 2017 adversarial competition

## 1st round:

- 4th place on non-targeted attacks  (44 teams)
- 6th place on targeted attacks (27 teams)

## Final round:

- 12th place on non-targeted attacks  (91 teams)
- 13th place on targeted attacks (66 teams)

# The NIPS 2017 adversarial competition

## Some thoughts:

It is a game:

**attacker** needs to model "what defenses will be in place"

**defense** needs to model "what knowledge and capability does the attacker have."

Defending is hard!

We tried several ideas (ensembles, input transformations such as random crops, rotations) and at best we still got 30% of error in white-box attacks

# Demo

**Code available in**
**https://github.com/luizgh/adversarial_examples**

# References


C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv:1312.6199 [cs]ArXiv: 1312.6199.

A. Kurakin, I. Goodfellow, S. Bengio, Adversarial Machine Learning at Scale, arXiv:1611.01236 [cs,stat]ArXiv: 1611.01236.

A. Athalye, L. Engstrom, A. Ilyas, K. Kwok. "Synthesizing robust adversarial examples." arXiv preprint arXiv:1707.07397 (2017).

N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: Security and Privacy (SP), 2017 IEEE Symposium on, IEEE, 2017, pp. 39-57

F. Tramr, A. Kurakin, N. Papernot, D. Boneh, P. McDaniel, Ensemble Adversarial Training: Attacks and Defenses, arXiv:1705.07204 [cs, stat]ArXiv: 1705.07204.

I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, arXiv:1412.6572 [cs, stat]ArXiv: 1412.6572.

I. J. Goodfellow, “Adversarial examples” talk in the Deep Learning Summer School 2015, Montreal. http://www.iro.umontreal.ca/~memisevr/dlss2015/goodfellow_adv.pdf