# Threat of Adversarial Attacks on Deep Learning

# Summary

1. General Observations
2. Attacks
3. Defenses
4. What we can do?

# Why Adverarial attacks?

# Is The Threat Real?

# Road Sign Attack

# Adversarial 3-D Object

# …But also

CELL-PHONE CAMERA

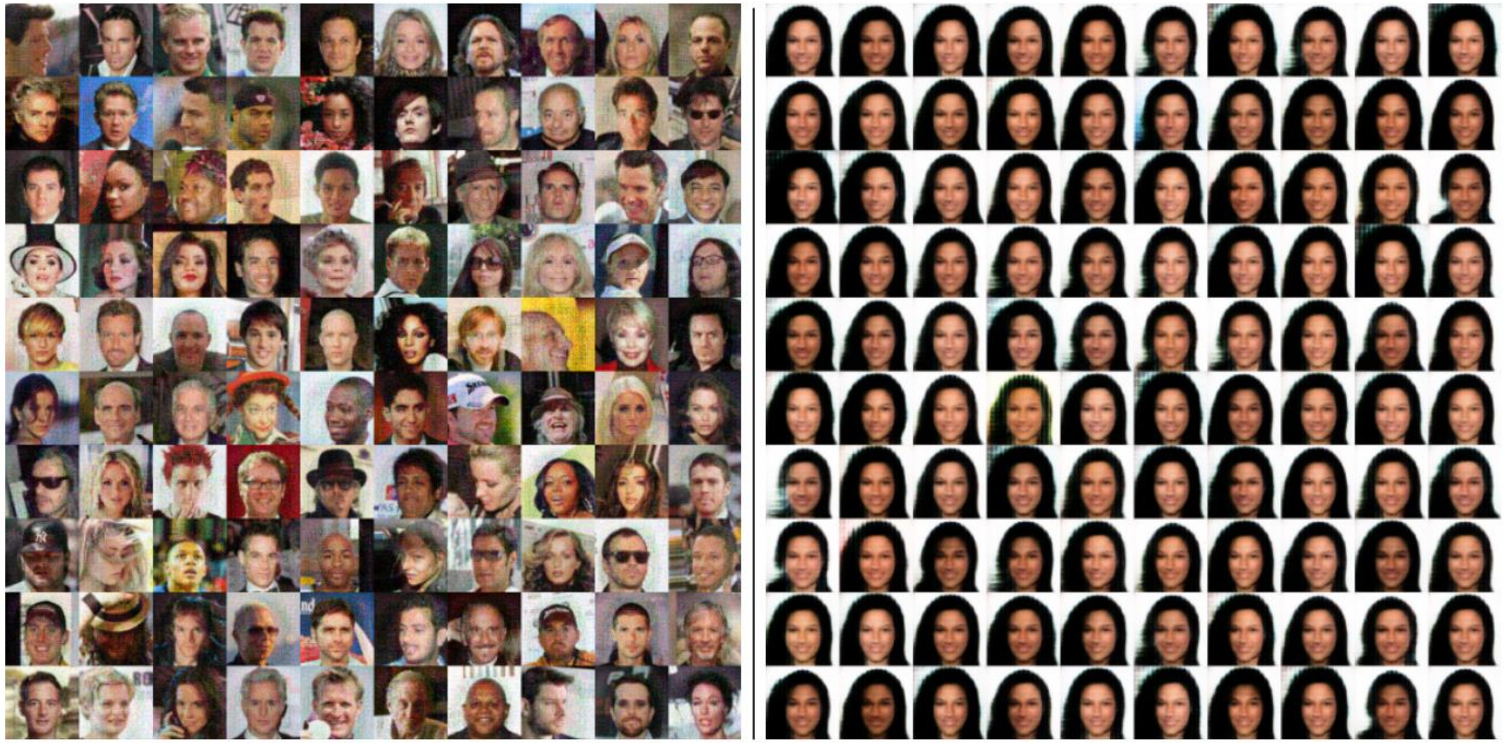
CYBERSPACE ATTACK

ROBOTIC VISION

VISUAL Q&A

# Only Concerns Object Recognition?

# Attacks on Generative Models

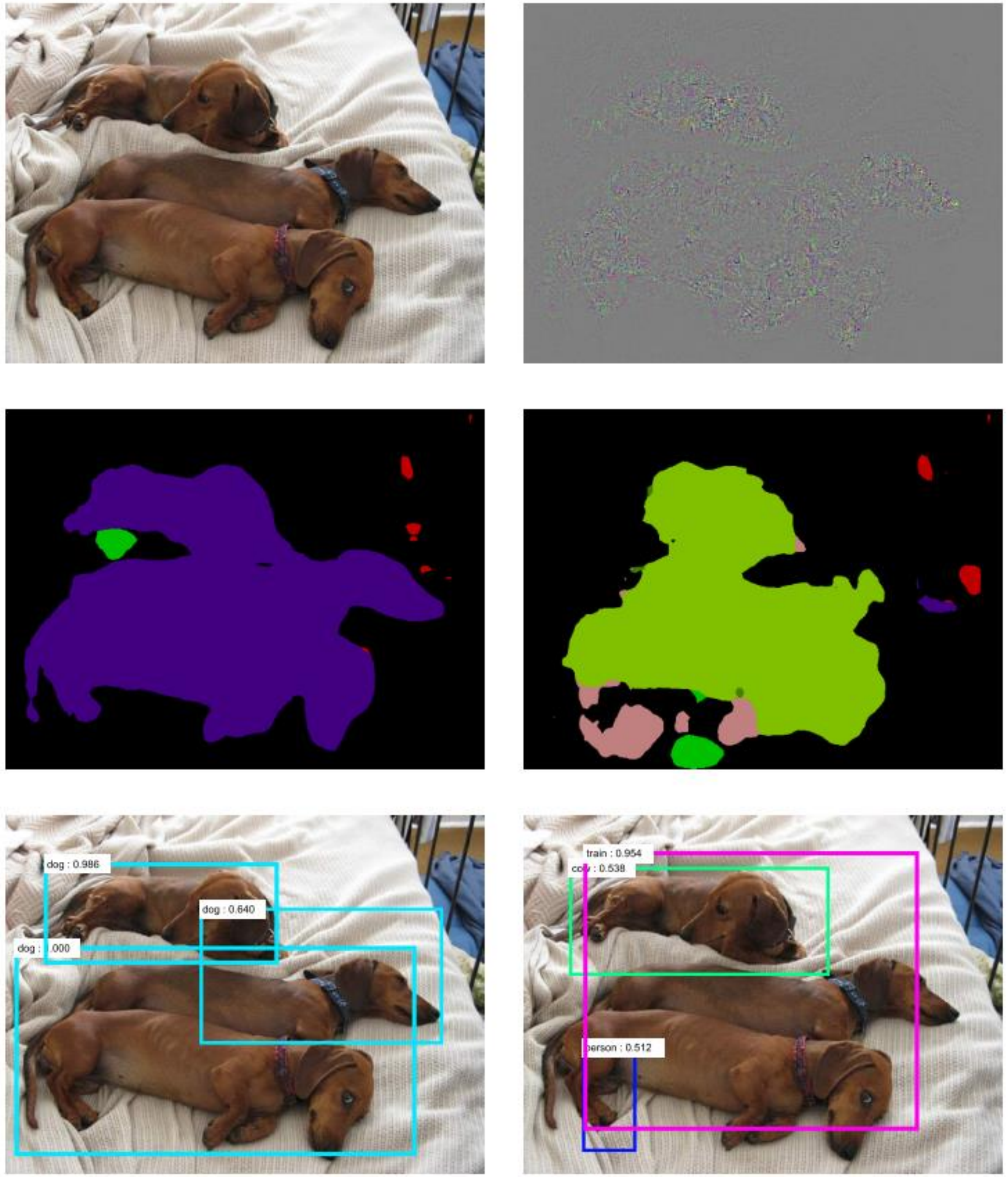


AutoEncoder Input (Adversarial)

AutoEncoder Output

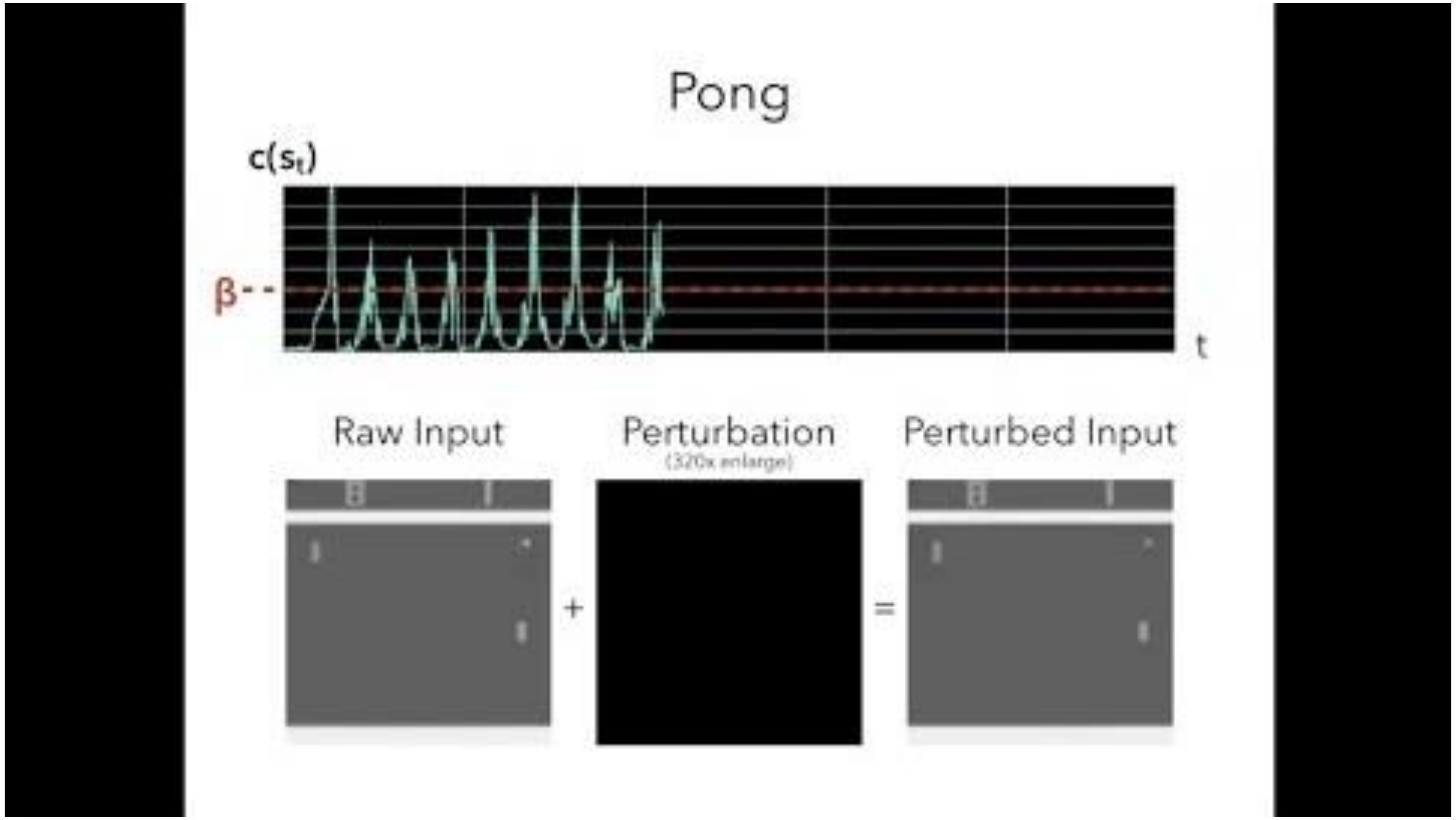
# Attacks on RNN – LSTM (Houdini)

| | |
|---|---|
| **Groundtruth** | • "The fact that a man can recite a poem does not show he remembers any previous occasion on which he has recited it or read it". |
| **G-Voice – original example:** | • "The fact that a man can decide a poem does not show he remembers any previous occasion on which he has work cited or read it." |
| **G-Voice – adversarial example:** | • "The fact that I can rest I'm just not sure that you heard there is any previous occasion I am at he has your side it or read it." |




# Attacks on Semantic Segmentation

# Attacks on Deep Reinforcement Learning



# Network Specific?

# Good generalization capabilities

Adversarial examples often transfer well between different NNs

→

Allow many 'Black Box' attacks

# Why Adversarial Examples exist?

# Supposed Reasons

**Structural reason:**

**‘Linearity Hypothesis’ (Goodfellow)**

- Flatness of decision boundaries
- Low flexibility of the networks

**Algorithmic reason:**

**‘Evolutionary Stalling’**

- Positive samples stop contributing to the network update once correclty classified

# There exists any effective defense?

# Existing defense methods  issues

Defenses are attack-specific

Counter-counter methods are possible

# Attacks

# Types of attacks

**Knowledge on the network:**

- Black Box attack
- White Box attack

**Specificity of the attack:**

- Image specific
- Universal attack

# Types of attacks

**# Iterations:**

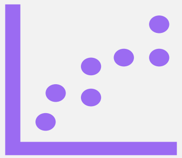
- Single-step attack
- Iterative attack

**Class targeted attack:**

- Targeted
- Not-targeted

# Attacks Score

Fooling Rate



Perturbation amount

(Time to attack)

# Historical Evolution

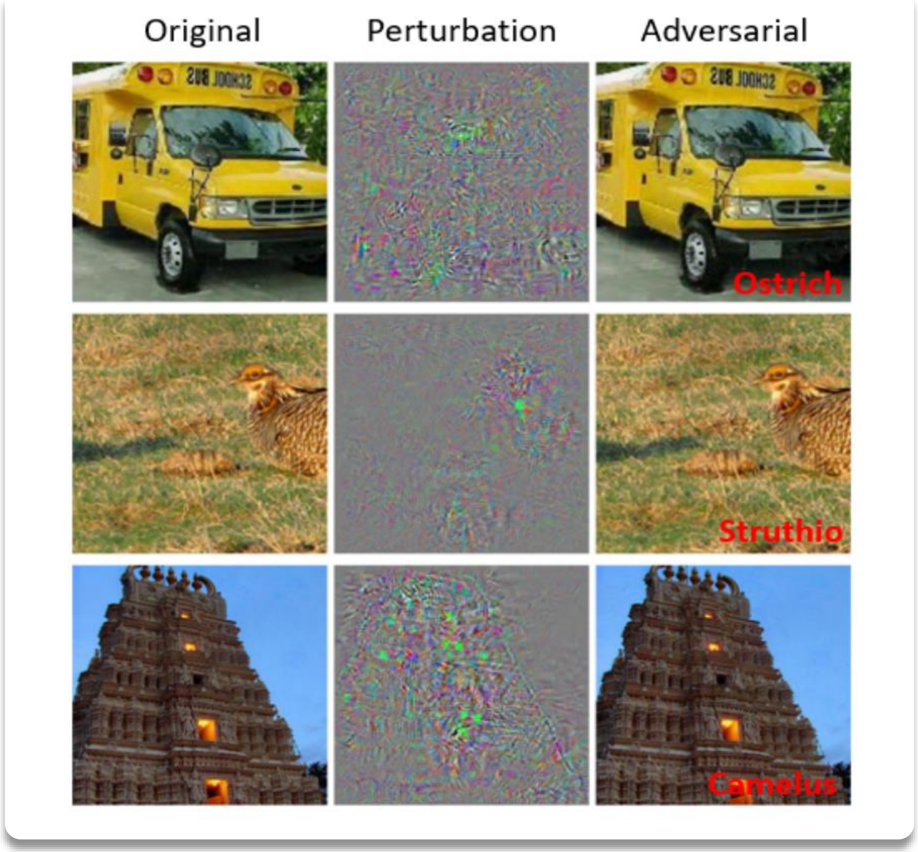
- White-Box Image Specific Single-Step Attacks
- White-Box Image Specific Iterative Attacks
- White-Box Universal Iterative Attacks
- Black-Box attacks

# Historical Evolution

**White-Box Image Specific Single-Step Attacks**

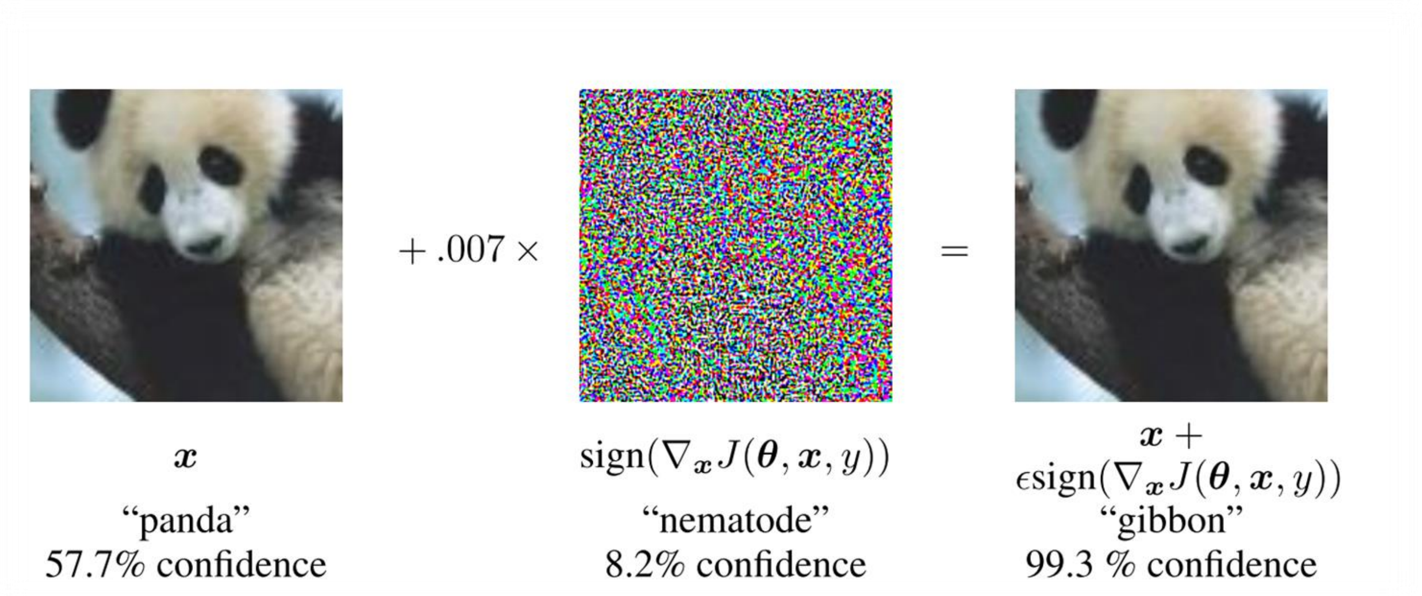
# 1) BOX-CONSTRAINED L-BFGS ATTACK

- First adversarial attack
  - "Intriguing properties of neural networks" (Szegedy 2014)
- Optimization problem:
  - $\min \|\rho\|_2 : C(I_c + \rho) = l_{target}$
  - $\min\{ \|\rho\|_2 + \mathcal{L}(I_c + \rho,\ l_{target})\}$



# 2) FGSM (Goodfellow)

- Optimization problem:
  - $\rho = \varepsilon * sign\ (\nabla J(\theta, I_C, l)$
- It allows fast computation
- Exploits the linearity of the model
- Introduced the adversarial training idea



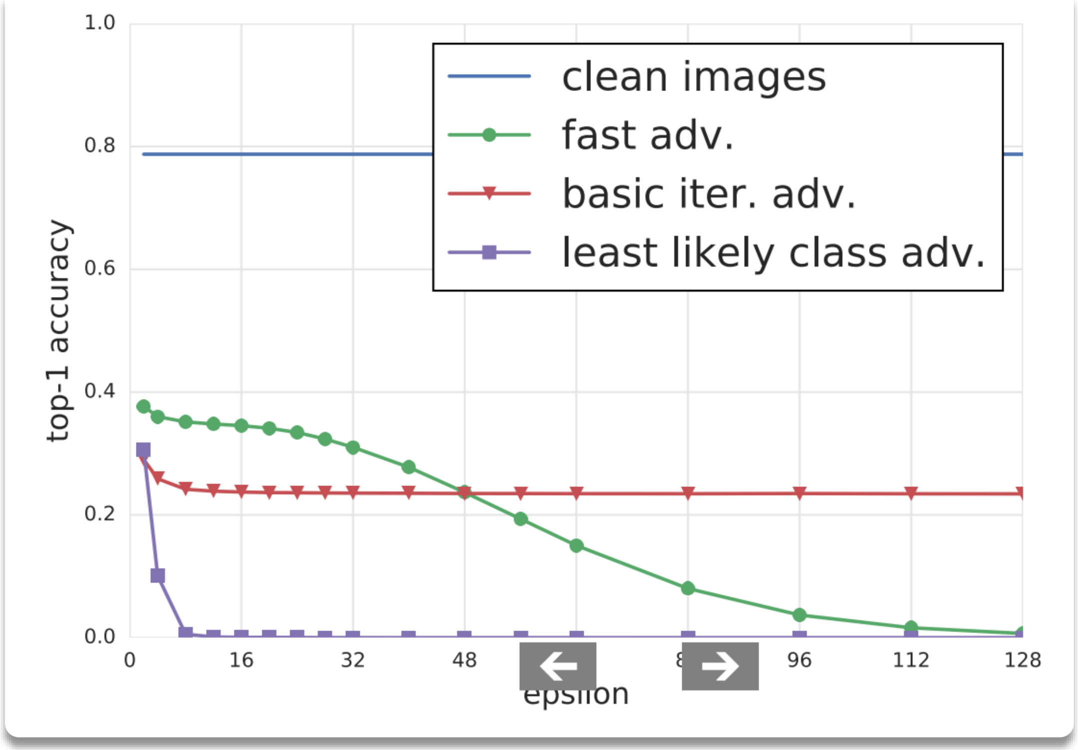
# Historical Evolution

White-Box Image Specific Single-Step Attacks

White-Box Image Specific **Iterative** Attacks

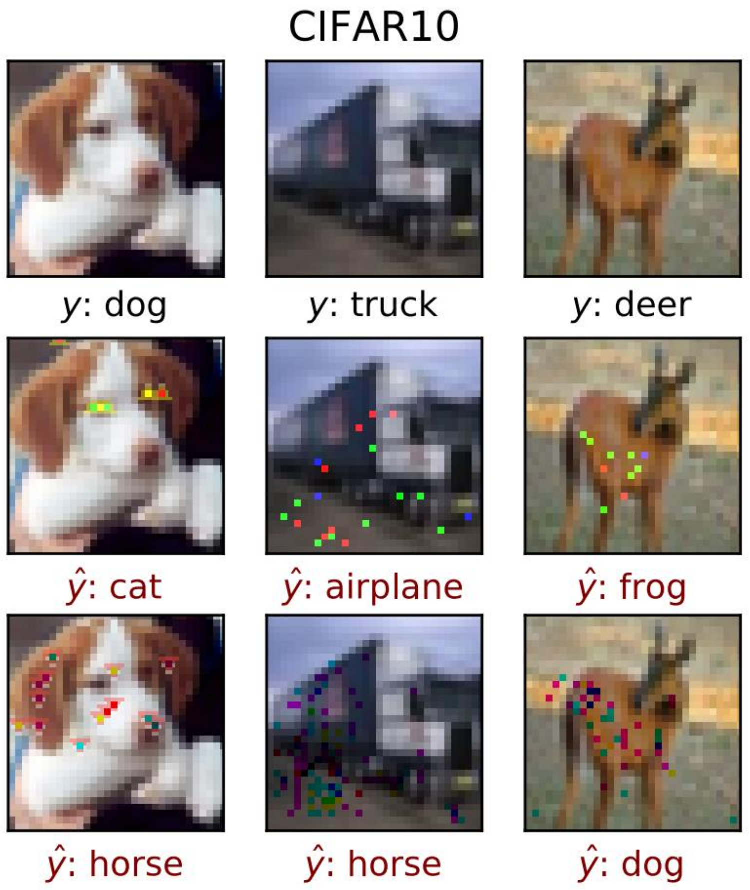
# 3) BIM & ILCM

- Optimization problem:
  - $I_{\rho}^{i+1} = Clip_{\varepsilon}\{I_{\rho}^{i} + \alpha * sign\ (\nabla J(\theta, I_{\rho}^{i}, l)\}$
- BIM: $l$ – untargeted attack
- ILCM: $l_{target}$ - targeted attack to the least likely class
- More computationally expensive



# 4) JSMA

- Algorithm based on the saliency map
- Objective: minimize the number of pixels modified
- Nice algorithm to determine strength of defense algorithm



# Other Attacks

**5) Deep Fool**

- Iteratively push an image to the nearest decision boundary
- Untargeted attack
- Produce the Minimal Norm perturbation

**6) C&W Attacks (Carlini & Wagner)**

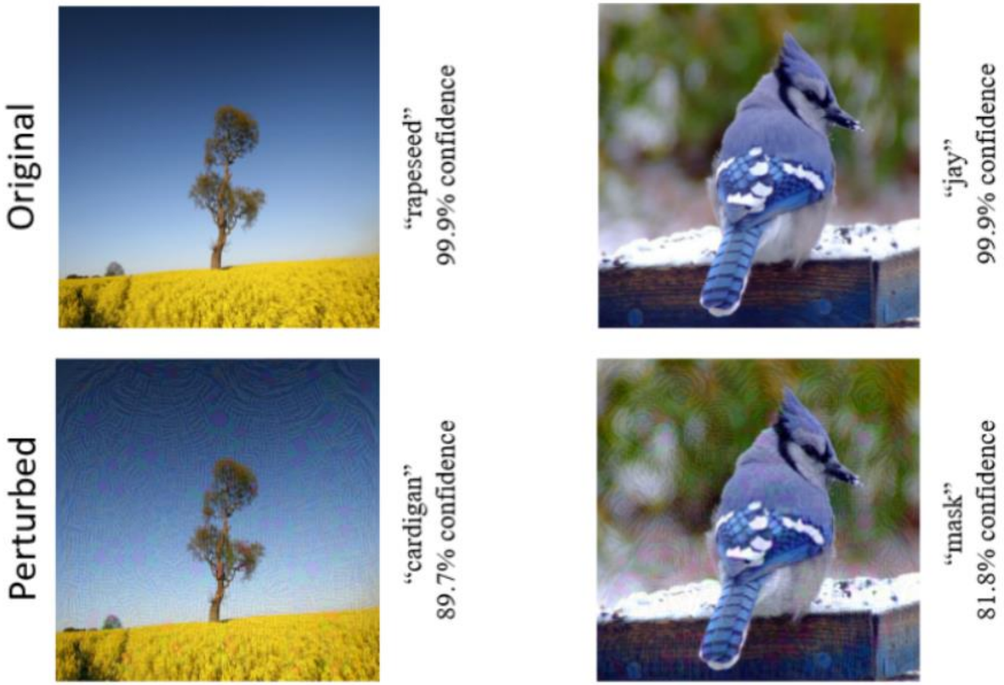
- 3 different attacks
- Current SOA of white box attacks
- Most defense algorithms fail against C&W

# Historical Evolution

- White-Box Image Specific Single-Step Attacks
- White-Box Image Specific Iterative Attacks
- White-Box **Universal** Iterative Attacks

# 7) Universal Adversarial Perturbation

- Fool a network on "any" image with the same perturbation
- $P\big(C(I_c) \neq C(I_c + \rho)\big) \geq \ \delta \ : \ ||\rho|| \leq \ \xi$
- Strategy similar to Deep Fool



# Historical Evolution

White-Box Image Specific Single-Step Attacks

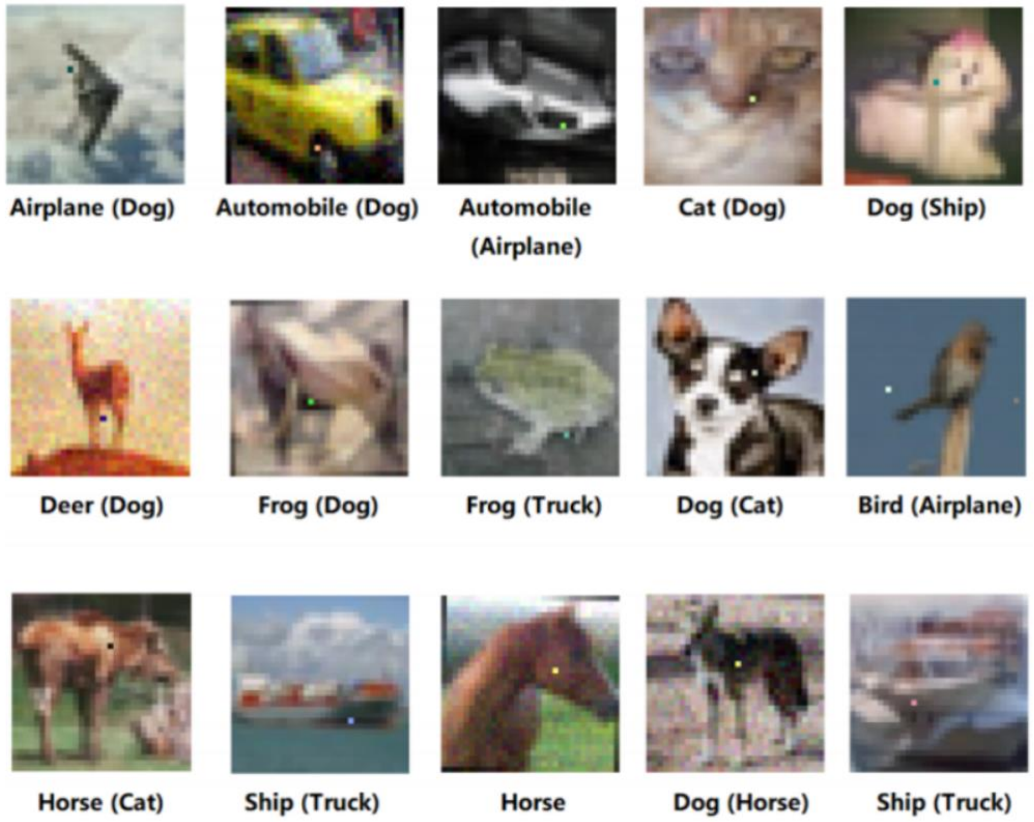
White-Box Image Specific Iterative Attacks

White-Box Universal Iterative Attacks

**Black-Box attacks**

# 8) One-Pixel Attack

- Only one pixel of the image is perturbed
- Evolutionary algorithm
- No need to access to internal parameters or loss of the net (BlackBox attack)



# 9) UPSET, ANGRI

- Residual Generating Network R():
  - $I_p = \max(\min(sR(t) + I_c, 1), -1) : C(I_p) = l\_target$
  - Generate n perturbation $I_{p,i}$ one for each class $i$

- ANGRI
  - Find an Image-Specific perturbation $I_p$

- UPSET
  - Find a Universal Perturbation $I_p$

# Adversarial Attack Framework?

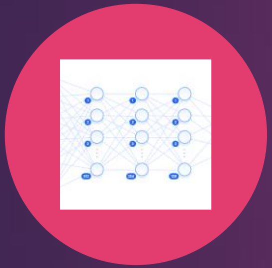
## FOOLBOX

# Defenses

# Types of Defense algorithms

| | |
|---|---|
| **Modified input data** | • Defense |
| **Modifying the network** | • Defense<br>• Detection |
| **Network add-ons** | • Defense<br>• Detection |

# Requirements



LOW IMPACT ON THE ARCHITECTURE

MAINTAIN SPEED OF THE NETWORK

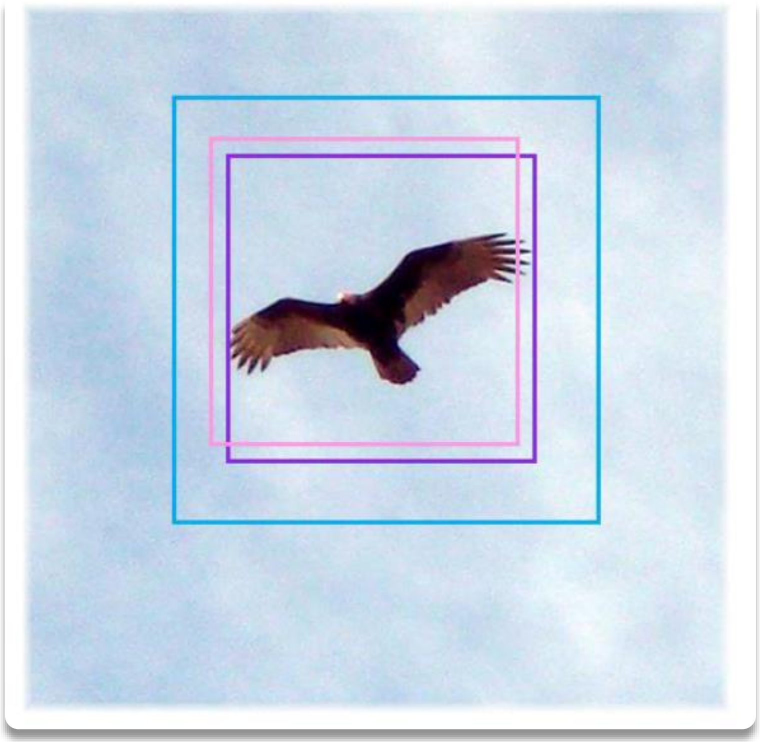
MAINTAIN ACCURACY ON CLEAN DATA

CORRECTLY CLASSIFY ONLY ADVERSARIAL EXAMPLES CLOSE TO THE REAL ONES

# Modified input data

# Defense algorithms

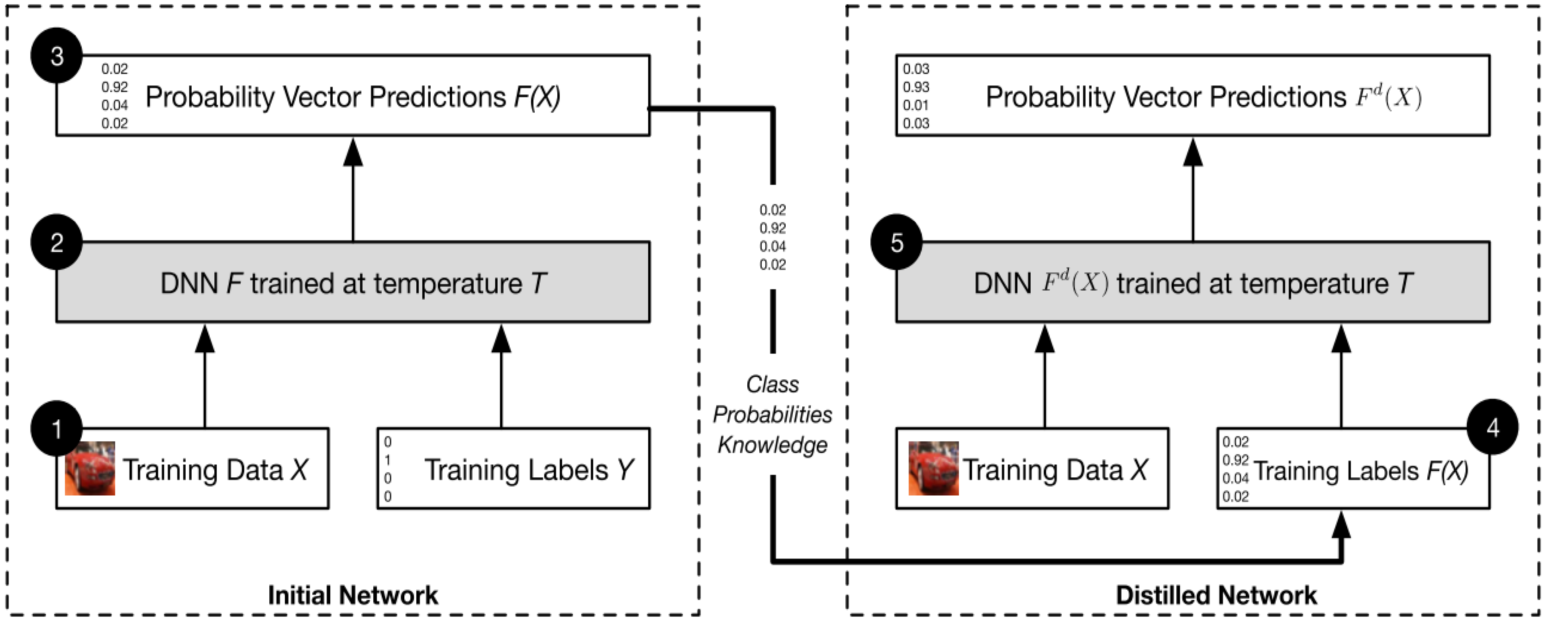
1. Brute Force Adversarial Training
2. Data Compression as a defense
    1. JPEG compression
    2. Also PCA/DCT
3. Foveation based defense
4. Also data augmentation (less effective)



# Modifying the network

# Defense Algorithms

1. Defense Distillation
2. Deep Contractive Network
3. Gradient Regularization
    1. Also Parseval Networks
4. Biologically Inspired Network



# Detection-Only Approach

**Classify adversarial examples**

1. As an additional class
2. With a detector subnetwork

**Control activations statistics**

1. RELU activations (Safety Net)
2. Convolutional filter activation

# Network add-ons

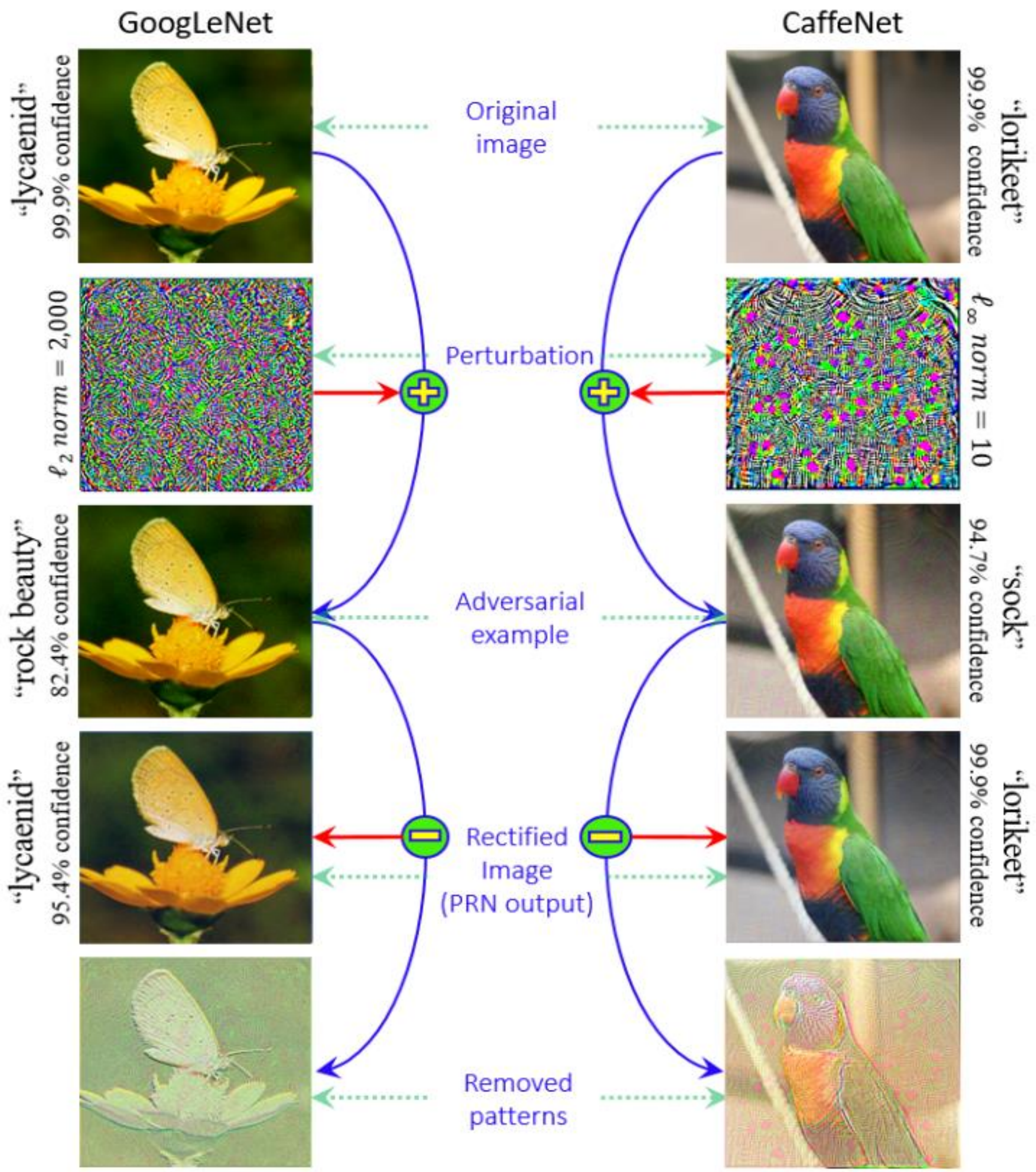
# Defense Algorithms

1. Defense Against Universal Perturbation

   Detector + PRN

2. GAN-based defense

   Ad-hoc brute force learning



# Detection-Only Approach

**Feature squeezing**

- Reduce pixel depth
- Perform spatial smoothing
- Classification Comparison of original and squeezed images

**Magnet**

- External model learn data manifold
- Reform near data and exclude far images

# Is there anything we could do?

# LOC Adversarial Defense

# Constraints in a hierarchical multilabel context

# Constraint Based Defenses

- Attack Detector:
  - Constraint satisfaction
- Robust Defense:
  - Constrained Learning
  - Collective Classification

Dog (Horse)

Horse (Cat)

Automobile (Dog)

# First Results

# Conclusions

- Is The Threat Real? **YES**
- Does It Concern Only Computer Vision? **NO**
- Are Attacks Network Specific? **NO**
- Why Adversarial Examples Exists? **Unknown**
- There exists effective defense yet? **NO**
- Is there anything we could do?

# Thank you for your attention