Adversarial Examples

Machine Learning Security



Fabio Brau

Dipartimento di Ingegneria Elettrica e Elettronica, Cagliari



CoEvolution



Funded by
the European Union

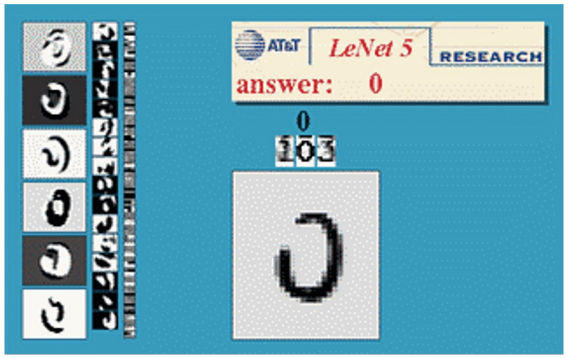
Introduction, Definitions and Taxonomy

Adversarial Examples Generation

Conclusions

Neural Networks achieve Human Performances…

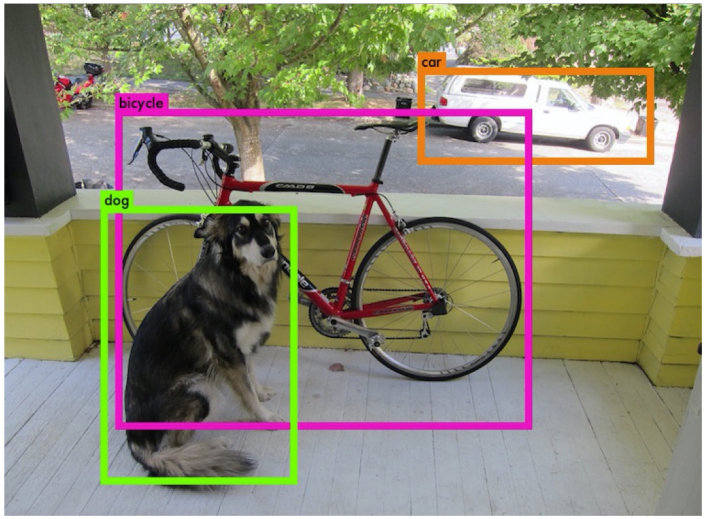
LeNet-5 (MNIST), 1998



AlexNet (ImagNet), 2013



Yolo, 2016



Neural Networks become so powerful that they reach human capabilities in the classification and object detection tasks

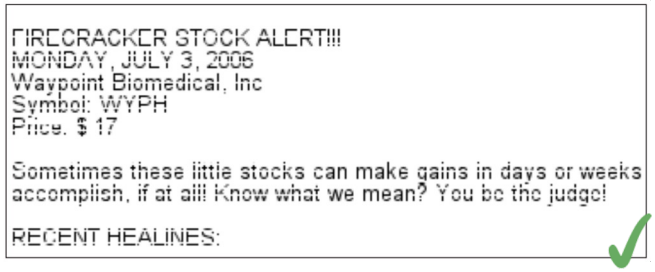
… but they make weird mistakes.

Evading Spam Detection (2010)

FIRECRACKER STOCK ALERT!!!
MONDAY, JULY 3, 2006
Waypoint Biomedical, Inc
Symbol: WYPH
Price: $ 17

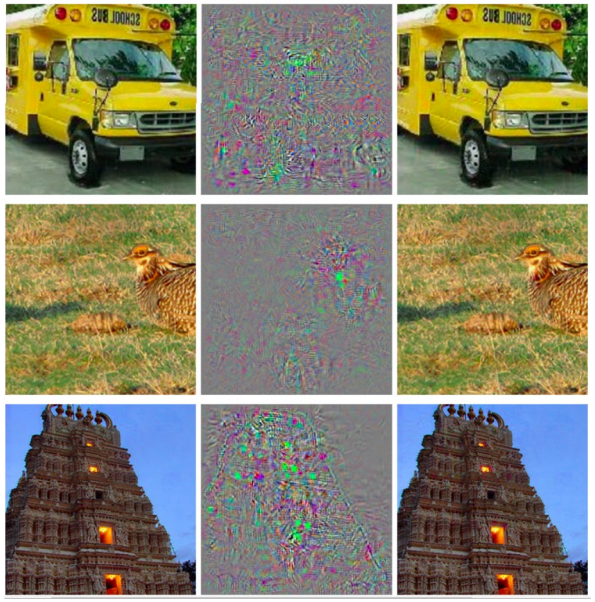
Sometimes these little stocks can make gains in days or weeks accomplish, if at all! Know what we mean? You be the judge!

RECENT HEALINES:



Spam Detector based on OCR

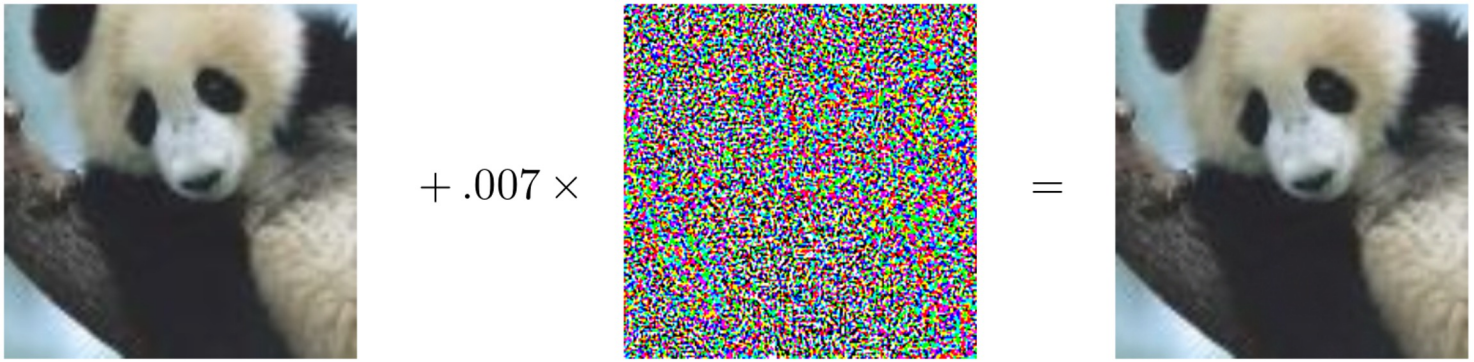
Attacks on Images (2014)



Perturbed Images are no long properly recognized by the network

Delvi et al. “Adversarial Classification” (2004)
Battista Biggio et al. ”Adversarial Pattern Classification” (2010)
Christian Szegedy et al. “Intriguing properties of neural networks” (2013)

Adversarial Examples



panda 57.7%

gibbon 99.3%

Adversarial examples are inputs **intentionally** designed to cause ML models to make **errors**.

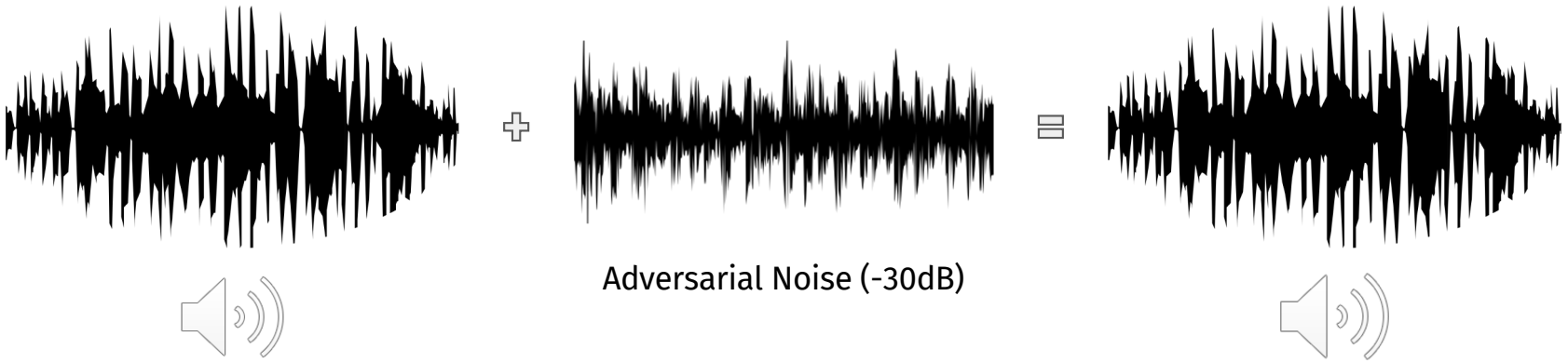
They can be created so that can appear **nearly identical** to normal inputs.

Goodfellow et al. “Explaining and Harnessing Adversarial Examples” (2014)

Adversarial Examples in Audio Systems

Speech-to-Text models

Can be fooled by adding an imperceptible noise to input audio



Carlini et al. "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text" (2018)

Adversarial Examples become popular

WRITING

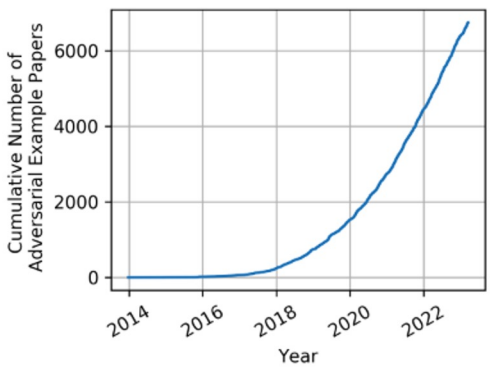
MAIN PAPERS TALKS CODE WRITING

A Complete List of All (arXiv) Adversarial Example Papers

by **Nicholas Carlini** 2019-06-15

It can be hard to stay up-to-date on the published papers in the field of adversarial examples, where we have seen massive growth in the number of papers written each year. I have been somewhat religiously keeping track of these papers for the last few years, and realized it may be helpful for others to release this list.

The only requirement I used for selecting papers for this list is that it is primarily a paper about adversarial examples, or extensively uses adversarial examples. Due to the sheer quantity of papers, I can't guarantee that I actually have found *all* of them.



link to the website

Why are adversarial Examples worth studying?

Security Threat (Known your Enemy)

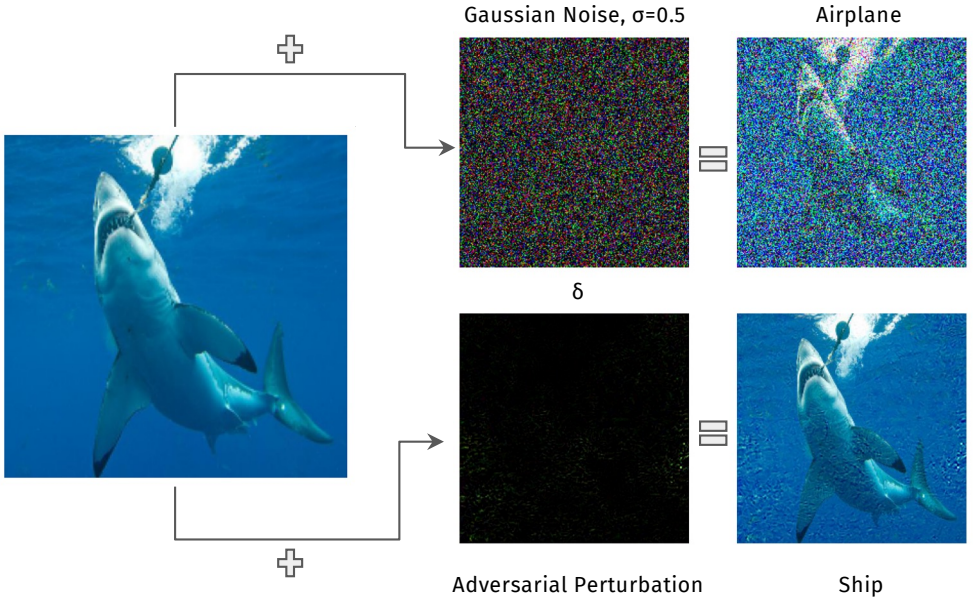
Neural Networks are involved in a growing number of tasks. How can we fill safe?



(a) Input 1 (b) Input 2 (darker version of 1)

DeepExplore (2021). Erroneous behaviour of DAVE-2. A tiny change in the light condition causes the model turns right.

Foundational Perspective



Small maliciously perturbations are sufficient to fool the model, why gaussian noise has not this property?

Sensitivity to Input Perturbation

Peak Signal to Noise Ratio

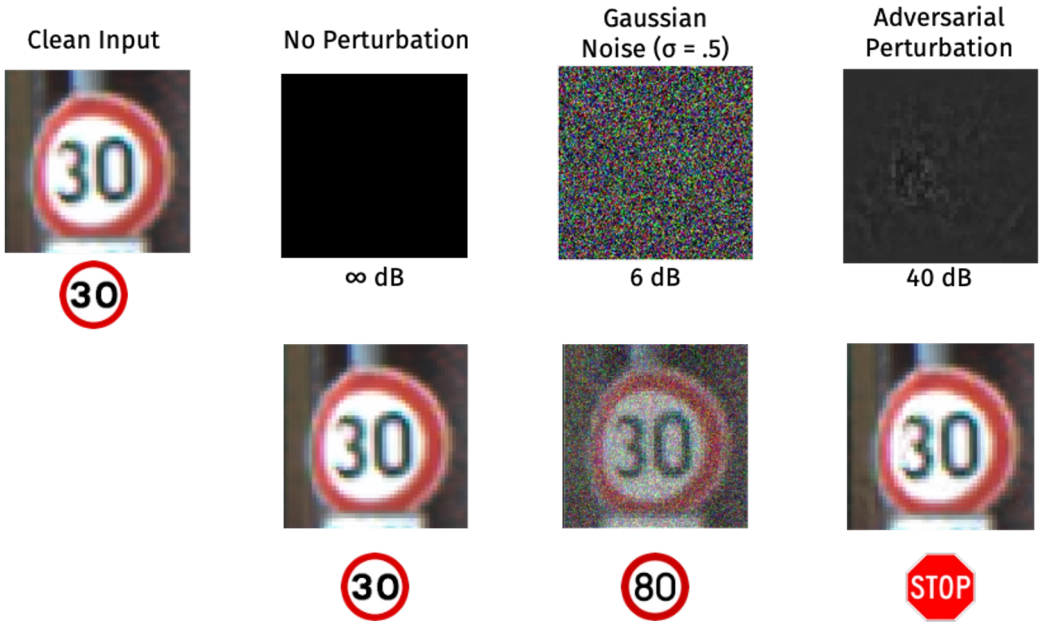
$$PSNR(x, x_{\text{adv}}) = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\frac{1}{n}\|x - x_{\text{adv}}\|^2}\right)$$

Minimal Adversarial Perturbation*

Median perturbation sufficient to fool a DNN

MNIST : 25.9 dB, **CIFAR10**: 52.4 dB

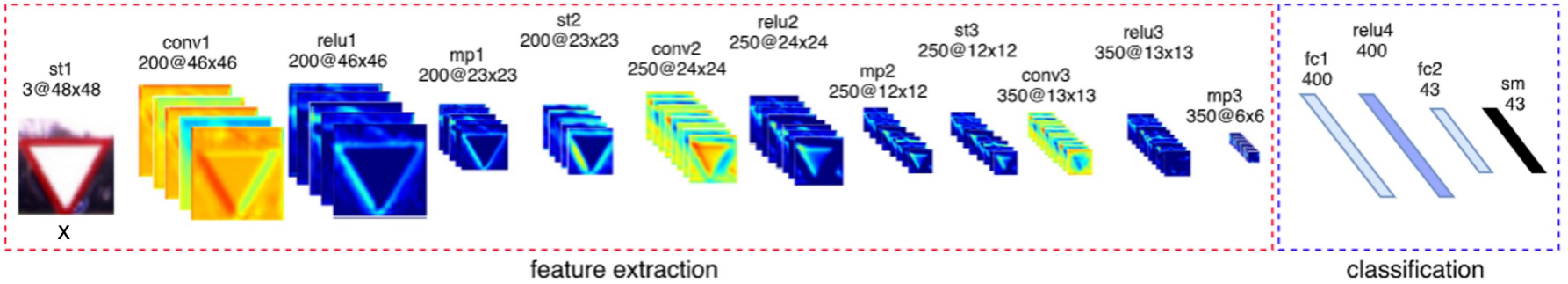
ImageNet: 61.9 dB

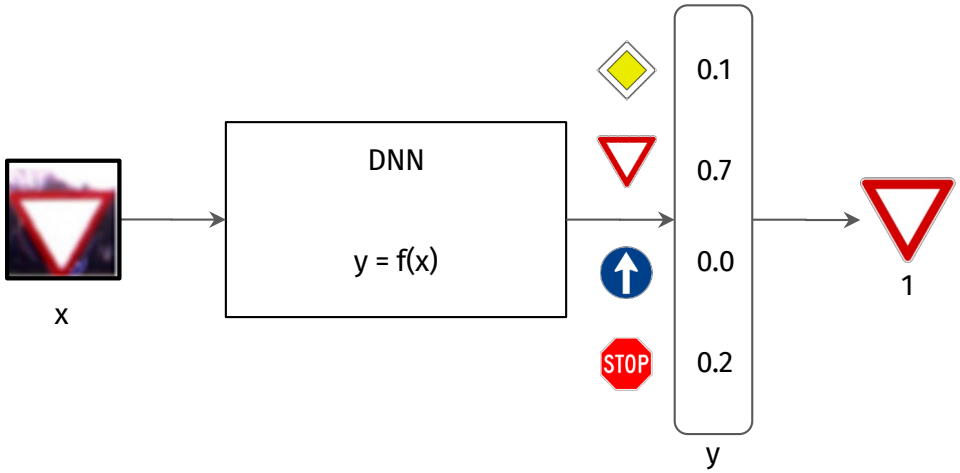


* Deduced from [Jérôme Rony et al., Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses

Definition and Taxonomy

Neural Networks for Image Classification



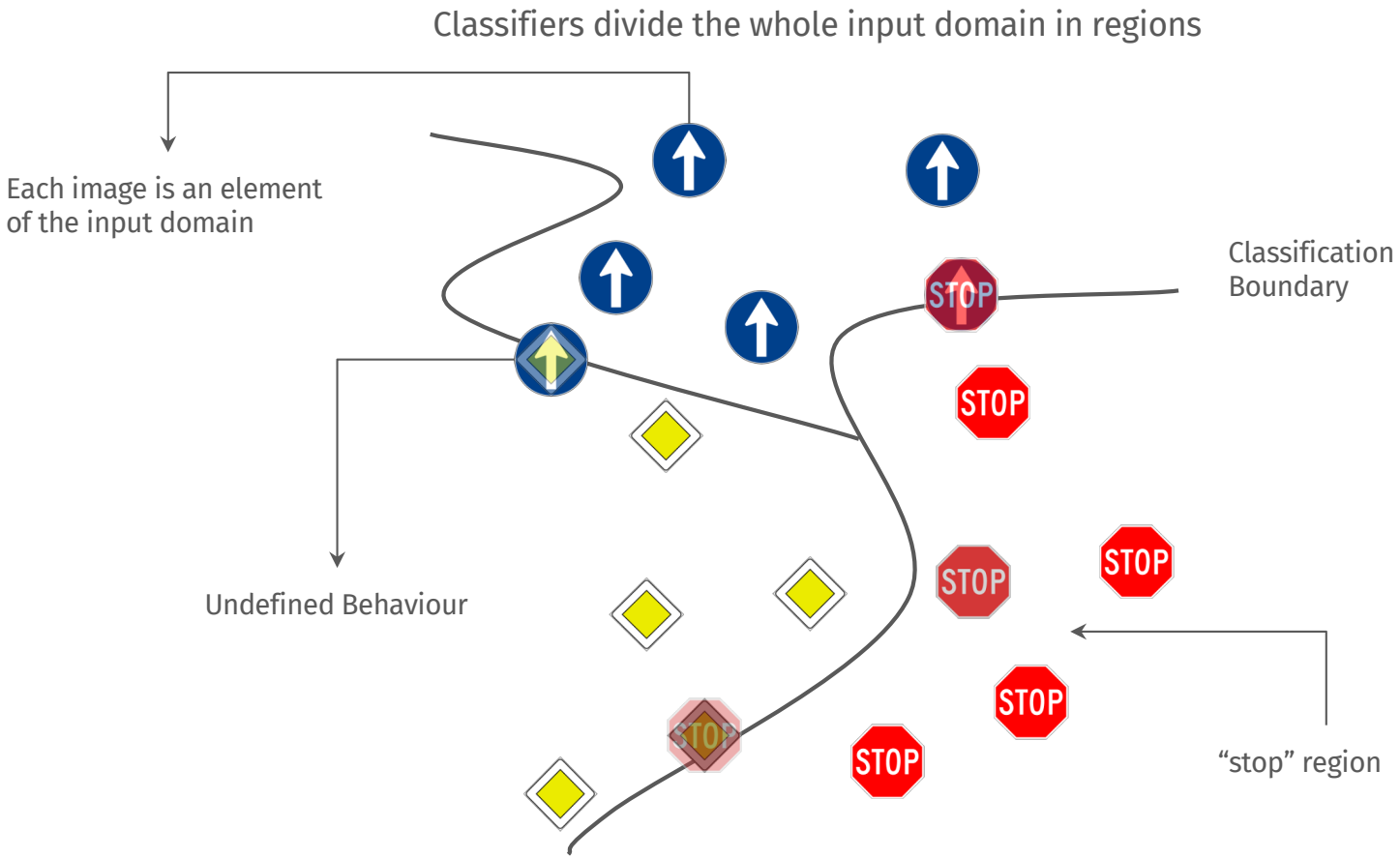


Encoding Table

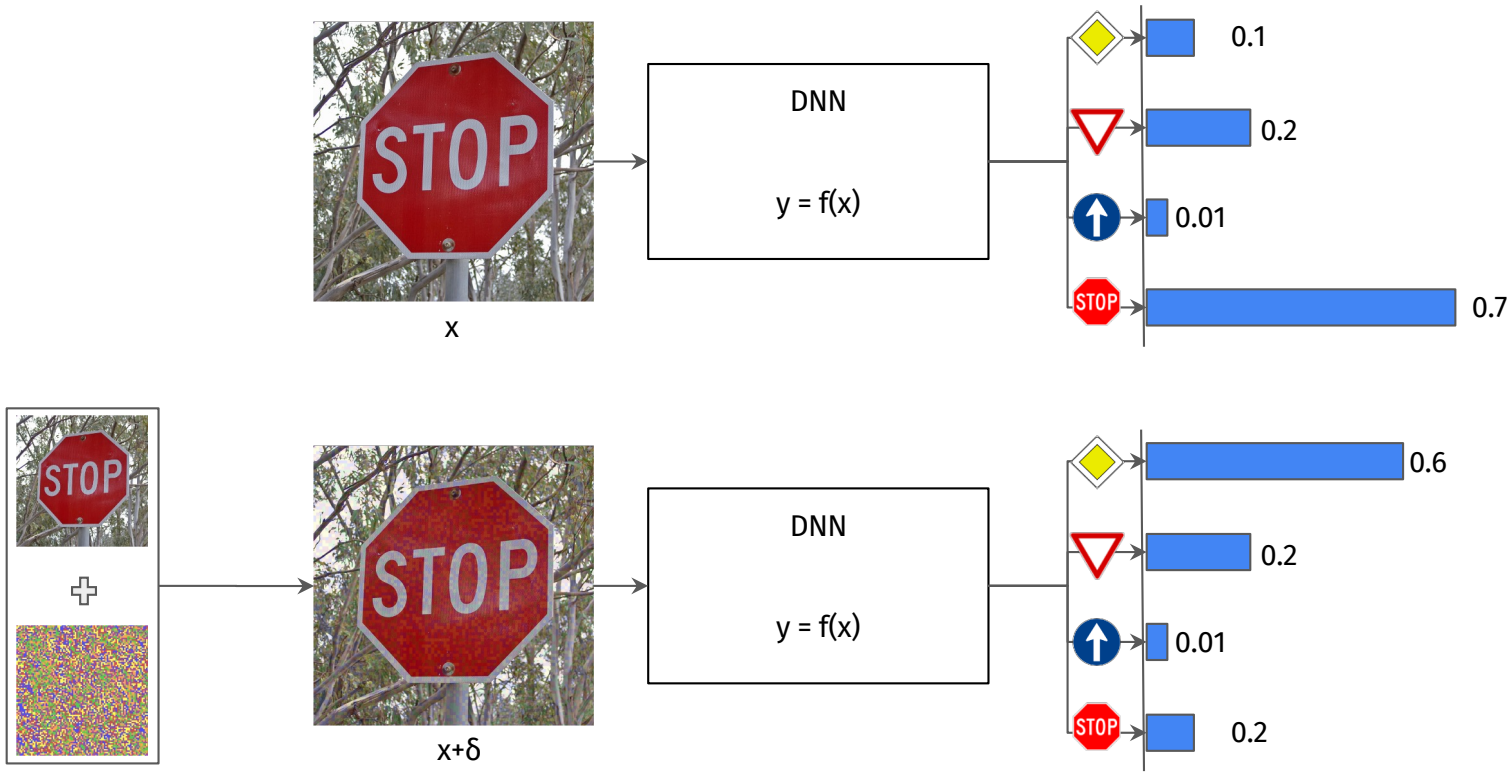
0	
1	
2	
4	STOP

Arcos-García et al. "Deep neural network for traffic sign recognition systems", (2018)

Geometrical Intuition



Adversarial Examples



Definition of Adversarial Examples

Definition (Adversarial Examples)
The result of perturbations, with a *small* l^p norm, that fool a classification model.



Mandatory Right | Adversarial Perturbation | Speed Limit 20 Km/h

$$\mathcal{K}(x+\delta) \neq \mathcal{K}(x) \quad \text{and} \quad \|\delta\| \approx 0$$

Some l^p Norm

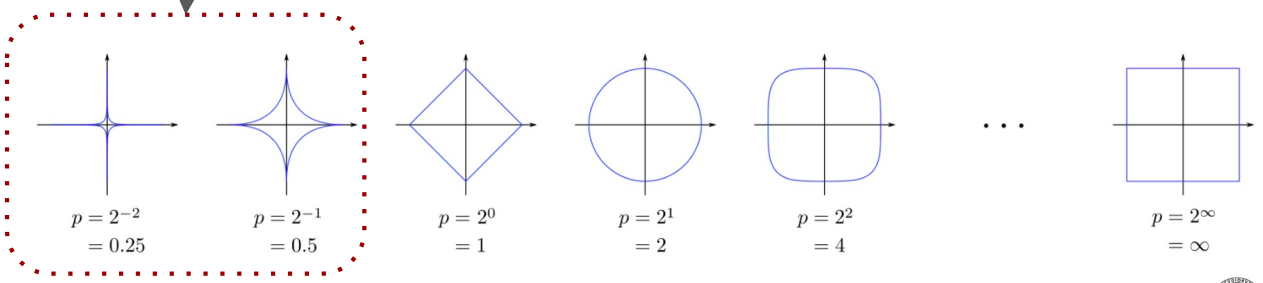
l^p - norms (Minkowski Distances)

$$\|\delta\|_0 = \#\{i : \delta_i \neq 0\}$$

$$\|\delta\|_p = \left(\sum_{i=1}^{n} |\delta_i|^p\right)^{\frac{1}{p}}$$

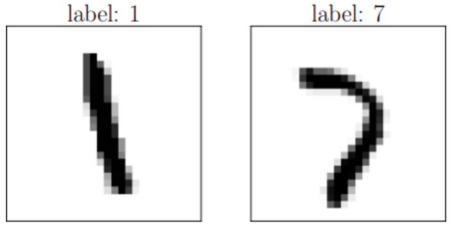
$$\|\delta\|_\infty = \max_i |\delta_i|$$

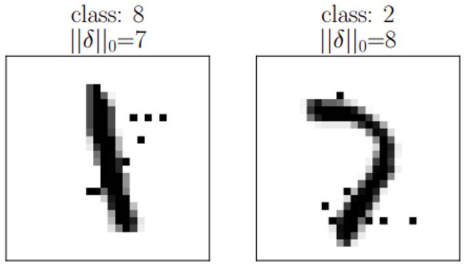
No Convex (pseudo-norm)



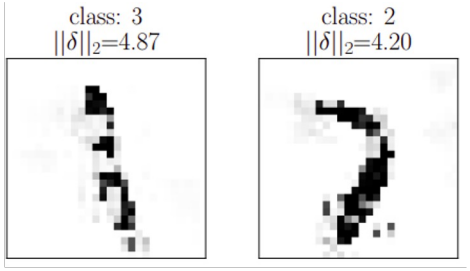
Different visualization of l^p bounded attacks

Definition in Minkowski norm covers many visually different cases





Few pixels, Unbounded values

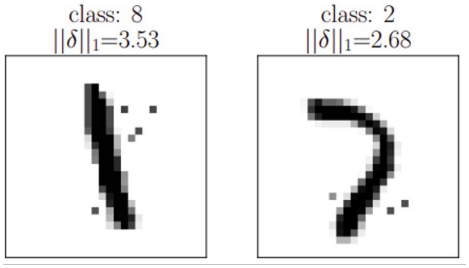


Few pixels, Bounded values

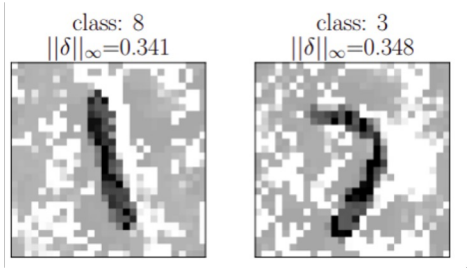
$$||\delta||_0 = \#\{i : \delta_i \neq 0\}$$

$$||\delta||_p = \left(\sum_{i=1}^{n} |\delta_i|^p\right)^{\frac{1}{p}}$$

$$||\delta||_\infty = \max_i |\delta_i|$$



Few pixels, Bounded values

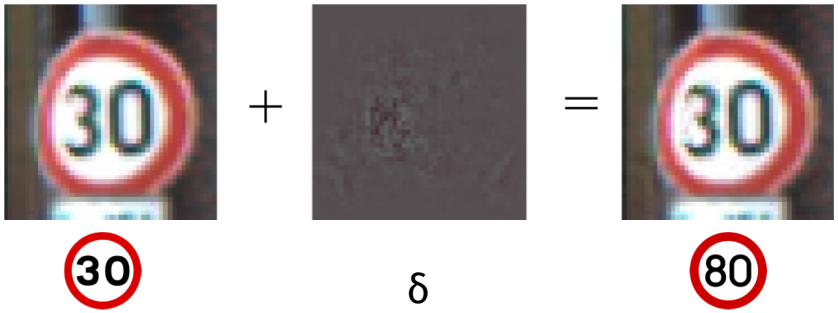


Many pixels, Bounded values

Maura Pintor et al. “Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints” (2021)

Taxonomy: Digital or Real World

Digital Adversarial Perturbation

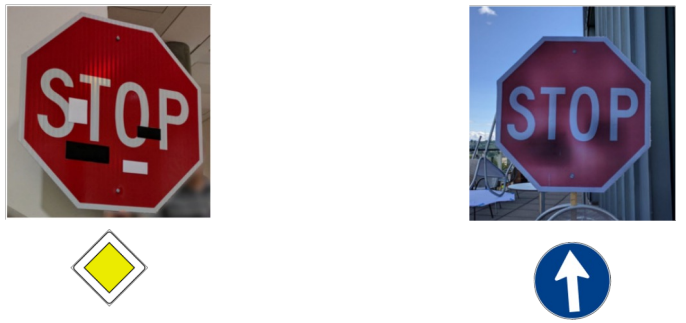


A malicious perturbation δ undetectable from human eye

Physical Adversarial Attacks

PATCH

POSTER



Adversarial perturbations that can be perceived from cameras, microphones, etc...

Focus of this Lecture

Taxonomy: Targeted or Untargeted

Targeted
The attacker aims at making the model answering `<target>`



Un-Targeted
The attacker only aims at fooling the model, without setting the output



`bird` `truck` `frog`

Taxonomy: Black or White Box

White Box

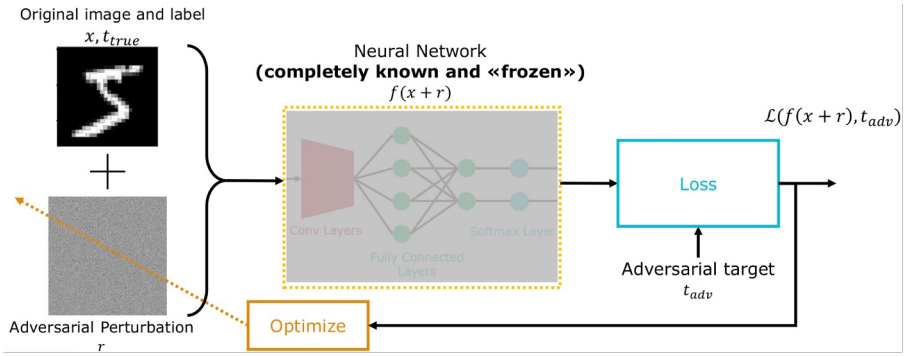
The attacker has full access to:

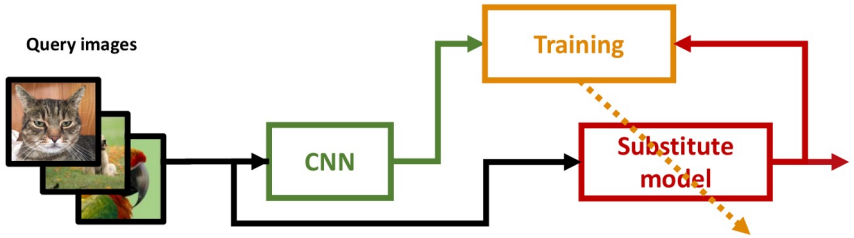
1. *Model architecture*
2. *Model weights*
3. *Data*

Black Box

The model acts has **oracle** and
the attacker has full access to:

1. *Model Architecture* but not *data*
2. *Data* but not *Model Architecture*
3. Neither *Data* or *Model Architecture*





Federico Nesti. “An introduction to Adversarial Examples”, lecture of Deep Learning Course (2022)

Minimal Taxonomy of Adversarial Examples



Adversary knowledge

White box — **Black box**

White box: architecture and parameters are known.

Black box: nothing about the model is known.

Intermediate settings are possible.



Attack goal

Untargeted — **Targeted**

Untargeted: force the classifier to choose any wrong class.

Targeted: force one specific class.



Perturbation norm

1-norm — **2-norm** — **∞-norm**

The norm measures the **magnitude** of the perturbation.

The following slides assume no specific norm: any works, with different effects.

Generation of Adversarial Examples

Generation of Adversarial Examples

Loss-based Attacks

Un-Targeted
Increase the Loss Function in the correct class close to the input x

$$\max_{\delta \in \mathbb{R}^n} \quad \mathcal{L}(f(x+\delta), l_{true})$$

$$\text{s.t.} \quad \|\delta\|_p \leq \varepsilon$$

$$0 \leq x + \delta \leq 1$$

$$Q * (x + \delta) \in \{0, \cdots, Q\}$$

Targeted
Decrease the Loss Function in the target class close to the input x

$$\min_{\delta \in \mathbb{R}^n} \quad \mathcal{L}(f(x+\delta), l_{target})$$

$$\text{s.t.} \quad \|\delta\|_p \leq \varepsilon$$

$$0 \leq x + \delta \leq 1$$

$$Q * (x + \delta) \in \{0, \cdots, Q\}$$

Box Constraint

Integer Constraint, Q=255

Box and Integer Constraints ensure that the Adv.Ex is a representable image.

Generation of Adversarial Examples

Distance-based attacks

Un-Targeted

Searching for the closest sample to x that is classified in a different class

$$\min_{\delta \in \mathbb{R}^n} \quad \|\delta\|_p$$

$$\text{s.t.} \quad \mathcal{K}(x+\delta) \neq l_{true}$$

$$0 \leq x+\delta \leq 1$$

$$Q * (x+\delta) \in \{0, \cdots, Q\}$$

Targeted

Searching for the closest sample to x that is classified in the target class

$$\min_{\delta \in \mathbb{R}^n} \quad \|\delta\|_p$$

$$\text{s.t.} \quad \mathcal{K}(x+\delta) = l_{target}$$

$$0 \leq x+\delta \leq 1$$

$$Q * (x+\delta) \in \{0, \cdots, Q\}$$

Box Constraint

Integer Constraint, Q=255

Box and Integer Constraints ensure that the Adv.Ex is a representable image.

Fast Gradient Sign Method

Leveraging the Loss Function

The loss function is used to search for the **worst** direction that fools the model and **not** to train the model $f : \mathbb{R}^n \to \mathbb{R}^C$

$$\delta^* = \varepsilon \cdot \text{sign}\left(\nabla_x \mathcal{L}(f(x), l_{true})\right)$$

Desired Magnitude in l^∞ norm

The gradient provides the direction of the largest growth

Observation.
The adversarial perturbation has bounded uniform norm

$$x^* = x + \delta^*, \qquad \|\delta^*\|_\infty = \varepsilon$$

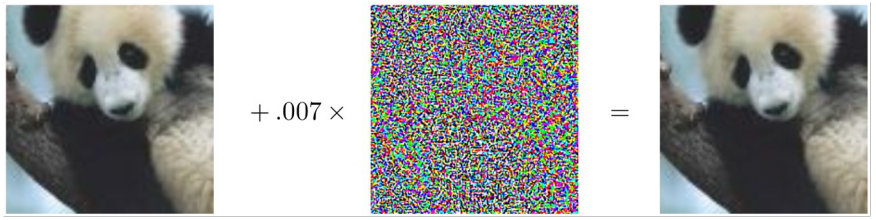
Goodfellow et al. “Explaining and Harnessing Adversarial Examples” (2014)

Fast Gradient Sign Method

FGSM

$$\delta^* = \varepsilon \cdot \text{sign}\left(\nabla_x \mathcal{L}(f(x), l_{true})\right)$$

$$x^* = x + \delta^*$$



The most famous Adv Ex. Attack on GoogLeNet

Pros

1. Easy to implement
2. Fast (useful in adv. training)
3. Bounded in uniform norm

Cons

1. Requires ε large (sometime visible)
2. Not always provide an Adv

Adversarial Training

Train the models by leveraging the following loss function

$$\tilde{\mathcal{L}}(f(x), l) = \alpha \cdot \mathcal{L}(f(x), l) \quad + \quad (1-\alpha) \cdot \mathcal{L}(f(x+\delta^*), l)$$

Adversarial Loss

Goodfellow et al. “Explaining and Harnessing Adversarial Examples” (2014)

Minimal Adversarial Perturbation with Penalty

Minimum Problem with Constraints

$$\min_{\delta \in \mathbb{R}^n} \|\delta\|_p \quad *$$

$$\text{s.t.} \quad \mathcal{K}(x+\delta) = l_{target}$$

$$0 \leq x + \delta \leq 1$$

Minimum Problem with Penalty

$$\min_{\delta \in \mathbb{R}^n} c\|\delta\|_p + \mathcal{L}(f(x+\delta), l_{target})$$

$$\text{s.t.} \quad 0 \leq x + \delta \leq 1$$

Penalty Constant.
Unbounded solution for $c=0$,
$\delta(c) \sim 0$ for large values of c

Theorem.
Under convex assumption on $\mathcal{L}(f(x+\delta), l_{target})$ in the variable x we have that

1. For each c, the Box constrained L-BFGS has a solution $\delta(c)$
1. The penalty c can be optimized by line-search such that $\mathcal{K}(x+\delta(c)) = l_{target}$

Christian Szegedy et al. “Intriguing properties of neural networks” (2013)

Minimal Adversarial Perturbation with Penalty

BL-BFGS

Box constrained **L**imited memory **B**royden-**F**letcher-**G**oldfarb-**S**hanno is a quasi-Newton method for minimum problem with box constraints.

[a]

$$\min_{\delta \in \mathbb{R}^n} \quad c\,\|\delta\|_p + \mathcal{L}(f(x+\delta), l_{target})$$

$$\text{s.t.} \quad 0 \leq x + \delta \leq 1$$



First Adversarial Examples presented to the community. Attack of AlexNet [b]

Pros.

1. Provide particularly low perturbations (approximate the Minimal Adversarial Perturbation)
2. Reliable. Almost always converge.

Cons.

1. Slow. Requires ~20k forwards of the model for each line search of c

[a] Dimitri P. Bertsekas. “Constrained Optimization and Lagrange Multiplier Methods” (1996)
[b] Christian Szegedy et al. “Intriguing properties of neural networks” (2013)

Including the Box Constraint in the Objective (CW)

Minimum Problem with Constraints

$$\min_{\delta \in \mathbb{R}^n} \quad \|\delta\|_p \qquad * \quad *$$

$$\text{s.t.} \quad \mathcal{K}(x+\delta) = l_{target}$$

$$0 \leq x + \delta \leq 1$$

Change of Variable (Simplification 1)

Optimizing w instead of directly optimize over the adversarial perturbation δ

$$\delta = \frac{1}{2}\left(\tanh(w) + 1\right) - x$$

Remark!

The perturbation δ satisfies the Box.Con.

Improve Loss Function (Simplification 2)

$$\mathcal{L}(y, l) = \left[y_l - \max_{j \neq l} y_j\right]^+$$

Positive part of the "margin" computed on the logits of f

Carlini et al. "Towards Evaluating the Robustness of Neural Networks" (2016)

Including the Box Constraint in the Objective (CW)

Minimum Problem with Constraints

$$\min_{\delta \in \mathbb{R}^n} \quad \|\delta\|_p \qquad * \quad *$$

$$\text{s.t.} \quad \mathcal{K}(x+\delta) = l_{target}$$

$$0 \leq x + \delta \leq 1$$

Minimum Problem with Penalty

$$\min_{\delta \in \mathbb{R}^n} \quad c\,\|\delta\|_p + \mathcal{L}(f(x+\delta), l_{target})$$

$$\text{s.t.} \quad 0 \leq x + \delta \leq 1$$

CW Minimum Problem (Unconstrained)

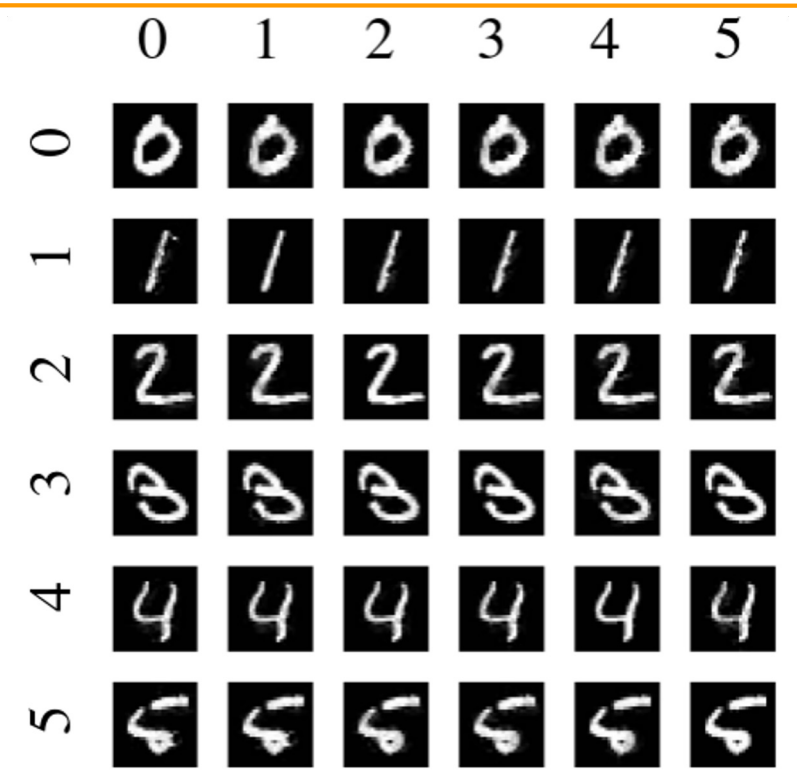
$$\min_{w \in \mathbb{R}^n} \quad \| \underbrace{\frac{1}{2}(\tanh(w)+1) - x}_{\delta} \|_p + c \cdot \mathcal{L}(f(\underbrace{\frac{1}{2}(\tanh(w)+1)}_{x+\delta}), l_{target})$$

Remark

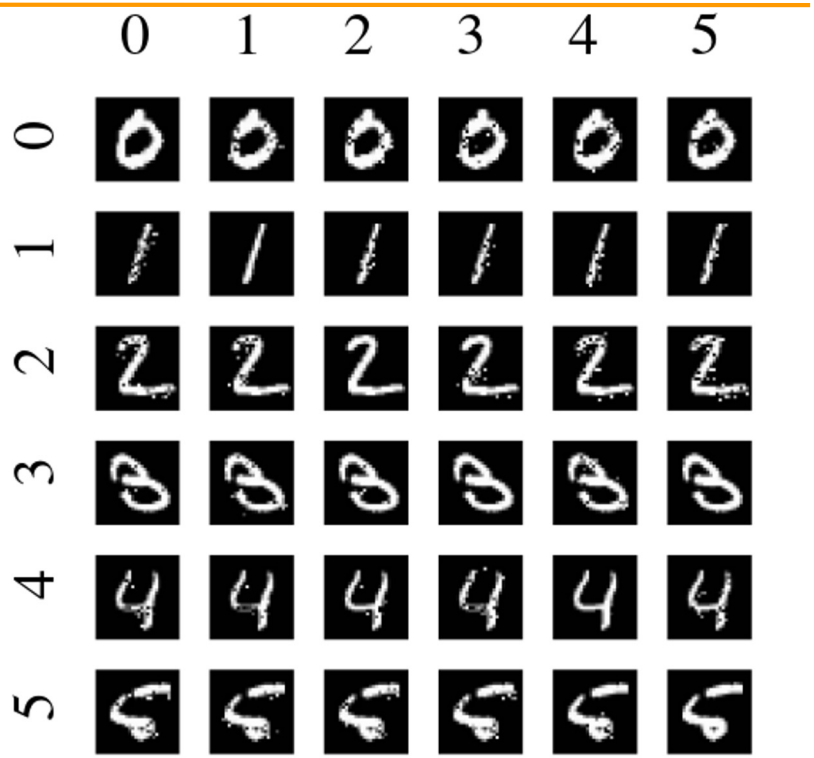
The penalty c can be optimized by line-search such that $\mathcal{K}(x+\delta(c)) = l_{target}$

Carlini et al. “Towards Evaluating the Robustness of Neural Networks” (2016)

Examples of attack with CW



l^2 Minimal Adversarial Perturbation



l^0 Minimal Adversarial Perturbation

Carlini et al. “Towards Evaluating the Robustness of Neural Networks” (2016)

Projected Gradient Descent (PGD)

Maximum Problem (Un-Targeted)

$$\max_{\delta \in \mathbb{R}^n} \quad \mathcal{L}(f(x+\delta), l_{true})$$

$$\text{s.t.} \quad \|\delta\|_p \leq \varepsilon$$

$$0 \leq x+\delta \leq 1$$

$$Q * (x+\delta) \in \{0, \cdots, Q\}$$

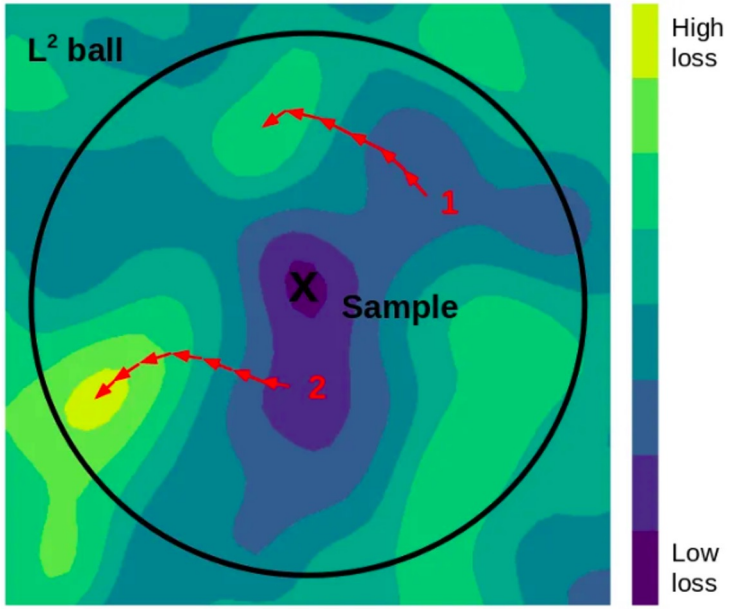
Iterating FGSM

Learning rate of the Gradient **A**scent

$$\tilde{x}^{(t)} = x^{(t-1)} + \alpha \cdot \text{sgn}\left(\nabla_x \mathcal{L}(f(x^{(t-1)}), l_{true})\right)$$

$$x^{(t)} = \Pi_{B_p(x,\varepsilon)}\left(\tilde{x}^{(t)}\right)$$

Projection into the p-Ball of radius ε



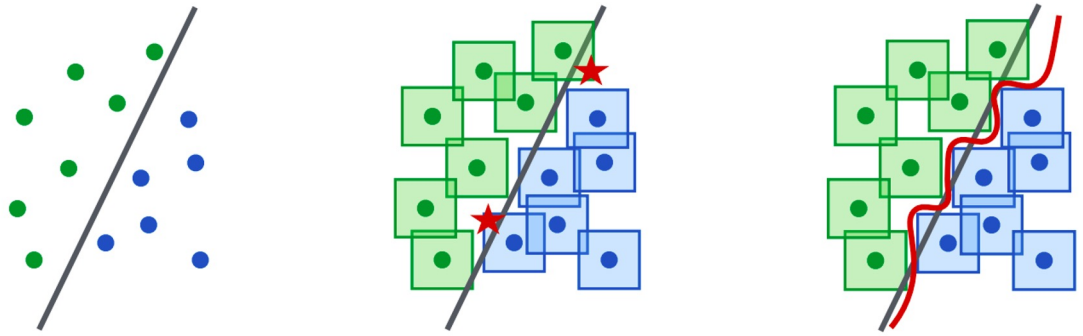
1. Take a random starting point
2. Iterate the SG(A) for few steps
3. Project into the l^p Ball
4. Repeat 1,2,3

Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial" (2018)

Adversarial Training by PGD

Minimization of the Empirical Worst Case

$$\min_{\theta} \rho(\theta), \quad \rho(\theta) = \mathbb{E}_{(x,l)\in\mathcal{X}} \left[\max_{\|\delta\|_p < \varepsilon} \mathcal{L}\left(f_\theta(x+\delta), l; \theta\right) \right]$$



Aleksander Madry et al. “Towards Deep Learning Models Resistant to Adversarial” (2018)

Conclusion

Adversarial Examples for Biological Vision Systems

Polyphemus Moth



Defense mechanisms that involves the **large eyespots** on its hindwings.

Chihuahua or Muffin



Shankar, Goodfellow and Papernot. “Adversarial Examples that Fool both Human and Computer Vision” (2018)

Conclusion

Introduction, Definitions and Taxonomy

Adversarial Examples Generation

sAIfer Lab

Joint lab on Safety and Security of AI

PRA LAB
Department of Electrical
and Electronic Engineering (DIEE)
University of Cagliari



UNIVERSITÀ DEGLI STUDI
DI CAGLIARI



SMARTLAB
Department of Informatics Bioengineering, Robotics,
and System Engineering (DIBRIS)
University of Genoa



Università di Genova

70+	800+	100+	30
Researchers	Publications	Projects	Years Experience