Survey Project N°20 - Adversarial Attacks: Review of the main approaches to perform adversarial attacks to neural networks.

Dipartimento di Informatica
Master Degree - Computer Science - AI Curriculum

Intelligent Systems for Pattern Recognition
9 CFU

PROF: Davide Bacciu
Niko Paterniti Barbino 638257

UNIVERSITÀ DI PISA

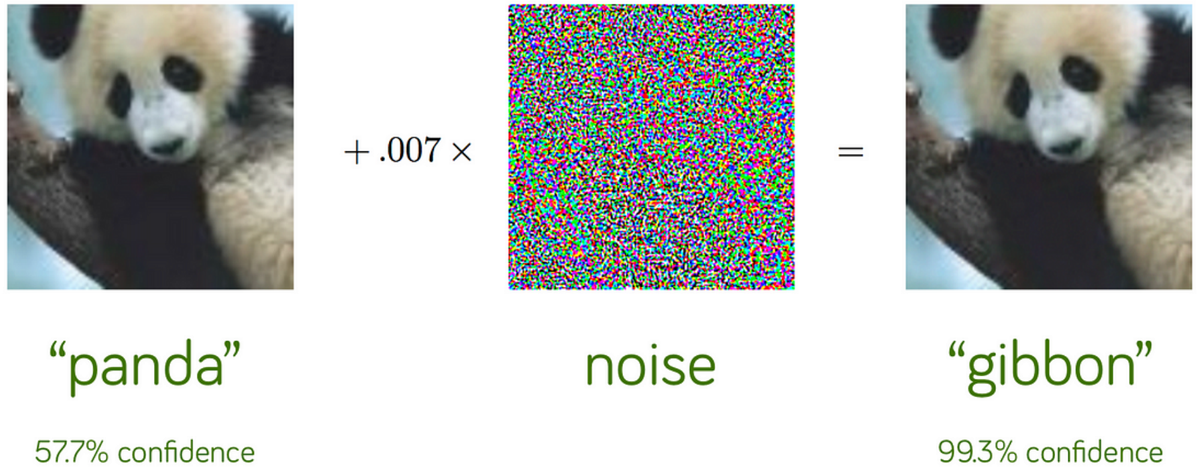
Index:

INTRODUCTION

Deep learning models have shown remarkable success across domains but are vulnerable to adversarial attacks. Adversarial examples can manipulate models into misclassifying with high confidence, posing security risks in applications like ATMs, surveillance, and AI assistants. We will present:

.

- **Adversarial attacks strategies**
- **Denfensive strategies**



ADVERSARIAL ATTACKS

GOAL: Introduce perturbations, like adding noise to an image so that a model misclassifies it while remaining visually unchanged to humans.

- **Non-targeted attacks**: Deceive a classifier by altering the source image, causing the model to output any random class other than the correct one.
- **Targeted attacks**: Misclassify an image as a specific target class by modifying the source image. Es: an adversarial image crafted to impersonate a specific individual or object.
- **Defenses against adversarial attacks**: Developing robust classifiers

Fast Gradient Sign Method(FGSM)

FGSM: A widely-used computationally efficient method for crafting adversarial examples in deep learning.
Leveraging the gradient information of the model, FGSM quickly generates adversarial perturbations by perturbing input data in the direction that maximally increases the loss. Induce significant misclassification with minimal perturbations, highlighting the susceptibility of deep learning models to adversarial attacks.

Advantages:

- Efficiency
- Computational Simplicity
- Effectiveness

Disadvantages:

- Limited Perturbation Magnitude
- Gradient-Based Vulnerability
- Single Step Perturbation

$$\eta = \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

FGSM - Momentum

FGSM momentum-based approach involves iteratively updating the perturbation to gradually steer the input data towards misclassification. The iterative process enhances the transferability of the attack across different models and datasets, while maintaining its effectiveness in causing misclassification.

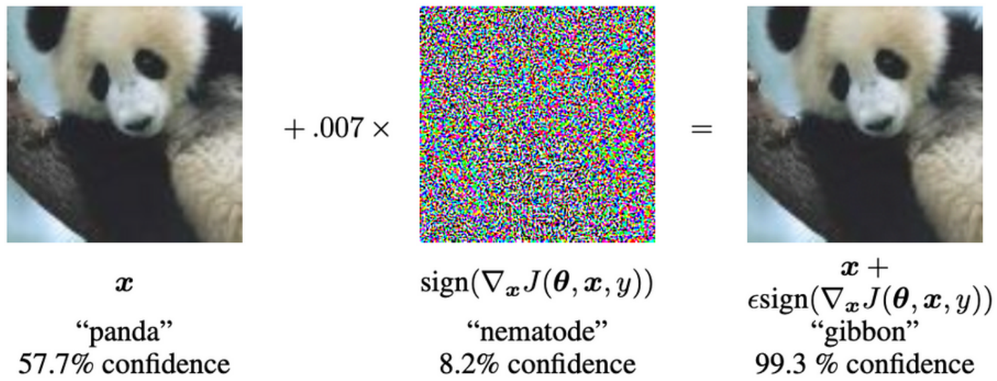
Explores a larger portion of the loss landscape and leads to stronger attacks compared to standard FGSM.

$$\eta_{t+1} = \mu \cdot \eta_t + \frac{\nabla_x J_\theta(x'_t, l)}{\|\nabla_x J_\theta(x'_t, l)\|_2}$$

$$x'_{t+1} = x'_t + \epsilon \cdot \text{sign}(\eta_{t+1})$$

Projected Gradient Descent(PGD)

PGD: A multi-step variant of FGSM that iteratively perturbs input data to maximize the loss function, with the constraint that the perturbed data remains within a specified distance from the original data.



Advantages:

- Robust Adversarial Examples:
- Transferability:.
- Stability:

Disadvantages:

- Increased Computational Cost:
- Limited Understanding of Robustness:
- Hyperparameter Sensitivity:

The update rule for PGD is defined as:

$$x_{t+1} = \pi_x \left(x_t + S \left(x_t + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \right) \right)$$

BASIC ITERATIVE METHOD(BIM)

BIM: iterative version of the Fast Gradient Sign Method (FGSM)s. Unlike FGSM, which applies adversarial noise only once, BIM applies adversarial noise iteratively, providing more control over the attack process., adversarial noise is applied multiple times with a small step size. The process involves clipping pixel values to limit big changes on each pixel in every iteration, preventing drastic alterations to the input image.

$$x_{n+1} = \text{Clip}_{x,\Delta}(x_n + \epsilon \cdot \text{sign}(\nabla_x J(x_n, y)))$$

Advantages:

- Increased Potency: gradually increases the adversarial impact on the model.
- Improved Robustness Testing: More comprehensive due to iterative nature.
- Enhanced Transferability: more versatile for attacking diverse systems.

Disadvantages:

- Higher Computational Cost:
- Increased Detection Sensitivity:

CARLINI - WAGNER (WG)

CW: Approach capable of overcoming defensive distillation. Defensive distillation is not effective against the L2, L0, and L∞ attacks. These attacks demonstrate a 100% success rate in generating adversarial samples causing misclassification with the same label.

The process involves **4 steps:**

- Train a neural network.
- Calculate the softmax,compute the soft training labels by applying the network to each case in the training dataset.
- Train the distilled network with soft training labels.
- Test the distilled network

DEFENSIVE DISTILLATION:

By training a “distilled network”, using a softened version of the original network’s outputs as labels during training to imitate the behavior of the original network, the model learns to generalize better and becomes less sensitive to small perturbations in the input space, such as those introduced by adversarial attacks

Jacobian Based Saliency Map Approach(JSMA)

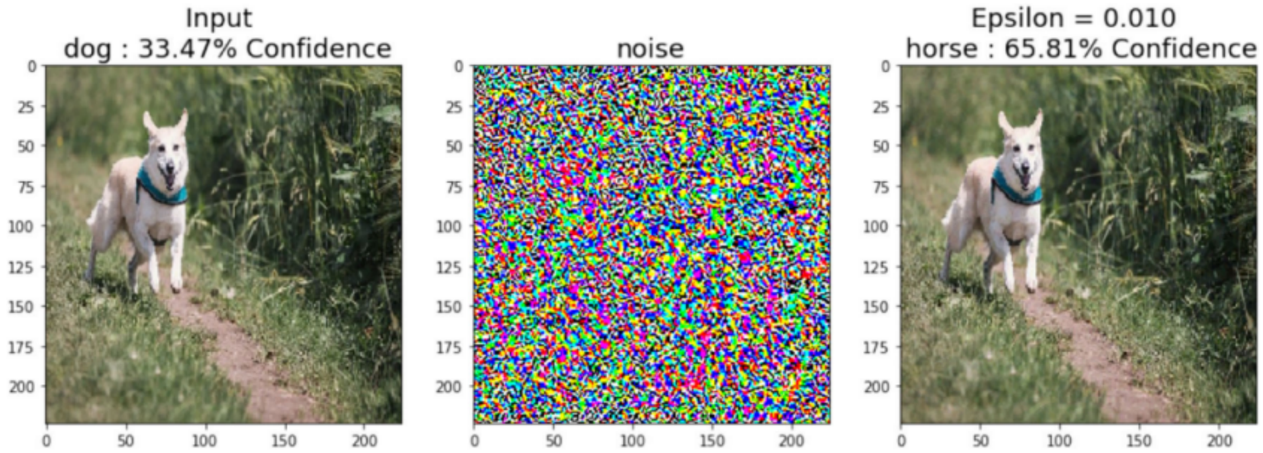
JSMA: Iteratively perturbs features of input data that have large adversarial saliency scores. This score describes the adversary's intent by moving a sample away from its source class towards a certain target class. JSMA constructs a mapping from input perturbations to output variations, known as the forward derivative. This is represented by the Jacobian matrix of the function learned by the DNN.

The **Jacobian matrix** represents the sensitivity of the output with respect to changes in the input.
We can use Saliency maps to visualize which input features should be modified to influence. JSMA seeks to find perturbations that lead to misclassification, targeting a specific output class while minimizing changes to other classes.

$$\nabla J(X) = \frac{\partial J(X)}{\partial X} = \left[\frac{\partial J_j(X)}{\partial x_i}\right]_{i=1..M, j=1..N}$$

DEFENSIVE STRATEGIES

A deep learning model based on the **ResNet-50** architecture performed a classification task achieving 97% accuracy. Yet, its prediction accuracy against adversarial samples dropped to 10%. Even achieving state-of-the art accuracy, there is an underlying flaw in how the network maps a given input to the predicted label



Two stragegies:

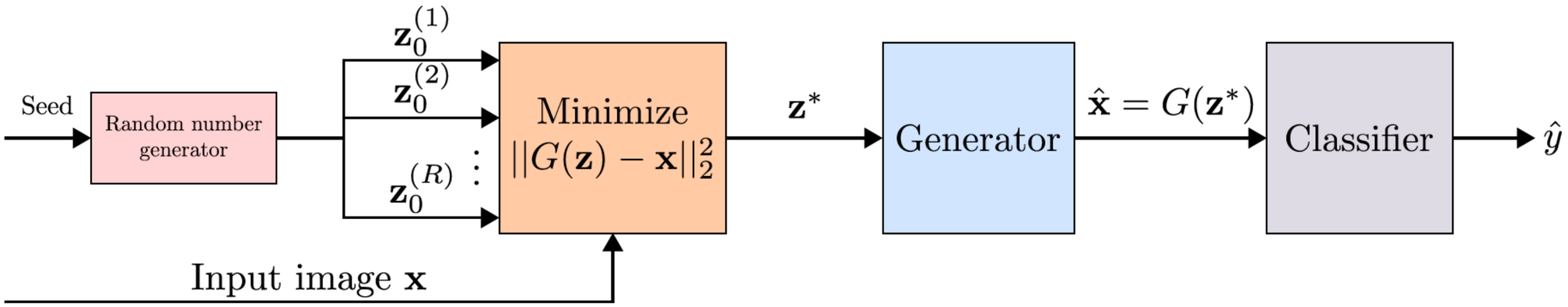
- **Adversarial Training**
- **Defence - GAN**

ADVERSARIAL TRAINING

The **Adversarial Training** method injects small poisonous samples in the training data to help the network reinforce its decision boundaries. The training set is augmented with counterexamples and their accompanying correct labels (x’, y) so that the learning algorithm will correctly classify those inputs. However, it may still exhibit vulnerabilities when faced with, unseen attacks, especially those crafted using techniques that don’t rely on detailed knowledge of the model’s internals(black box attacks)

DEFENSE -GAN

The **Defense-GAN** method trains a generative adversarial network (GAN) used to generate adversarial examples.
The adversarial examples generated by the GAN are added to the training dataset of the model. The model is retrained on this augmented dataset and by exposing the model to these adversarial samples during training, it learns to better distinguish between legitimate and adversarial inputs



EXPERIMENTS

The experiments have been conducted on 170,000 perturbed images created using 3 different attack methods: JSMA, FGSM, CW and comparing the two defense strategies.

Adversarial Training:

- Lower performance.
- Performs on commodity hardware only one back-propagation step
- Keeps model simple

Defence - GAN:

- Higher performance:
- Can be trained in parallel but require dedicated resources.
- Model Complexity

Table 1 Attacks versus accuracy results on MNIST

	No defense	Adversarial T.	Defense-Gan
No Attack	0.994	0.990	0.985
NT-JSMA	0.215	0.941	0.974
FGSM	0.134	0.912	0.970
CWL2	0.135	0.942	0.969

Table 2 Attacks versus accuracy results on FASHION-MNIST

	No defense	Adversarial T.	Defense-Gan
No Attack	0.975	0.956	0.924
NT-JSMA	0.062	0.948	0.893
FGSM	0.101	0.913	0.879
CWL2	0.029	0.937	0.889

Table 3 Attacks versus accuracy results on CIFAR10

	No defense	Adversarial T.	Defense-Gan
No Attack	0.980	0.973	0.942
NT-JSMA	0.074	0.941	0.893
FGSM	0.120	0.915	0.872
CWL2	0.085	0.969	0.844

CONCLUSION

Adversarial images comprise a menace to security and privacy since machine learning systems can be exposed to some type of attacks performed by those images. In this study, we analyzed and compared different methods for generating adversarial examples,we have shown the trade-off of choosing fast defense methods such as adversarial training that does not affect the model complexity but is less effective against attacks than generative adversarial networks that add another layer on top of an existing model but shows higher hardness against perturbations. We demonstrated that while techniques like adversarial training and defensive distillation have shown efficacy, they are not foolproof, particularly against sophisticated attacks.

Thanks

Niko Paterniti Barbino
638257

UNIVERSITÀ DI PISA