# Adversarial Attacks & it's Defense

Group- 14
Raja Babu Meena (202116010)
Rohan Baghel(202116011)

*Subject : Adversarial Machine Learning*

*DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY, GANDHINAGAR*

July 21, 2025

# Overview

# Introduction

- Nowadays, machine learning models are used in many real-world applications.
- Adversarial machine learning is a machine learning method that aims to trick machine learning models by providing deceptive input.
- We are creating adversarial [1] image and defending against it.

[1] https://www.mdpi.com/2073-8994/13/3/428/htm

# Problem Statement

- Adding deviation in the image of MNIST dataset using Gradient-based Adversarial Attacks.
- Our aim is to defense against this attack and to find the accuracy of model.

# Approach used to solve the problem

- For Adversarial attack use PGD [2]
  1. Start from a random perturbation in the $L^p$ ball around a sample.
  2. Take a gradient step in the direction of greatest loss.
  3. Project perturbation back into $L^p$ ball if necessary.
  4. Repeat 2–3 until convergence.
- For Defence against adversarial attack use Adversarial training.
  - Minimising the loss function where $\delta$ is a set of perturbations we want our model to be invariant.

[2] `https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3`

# Approach used to solve the problem

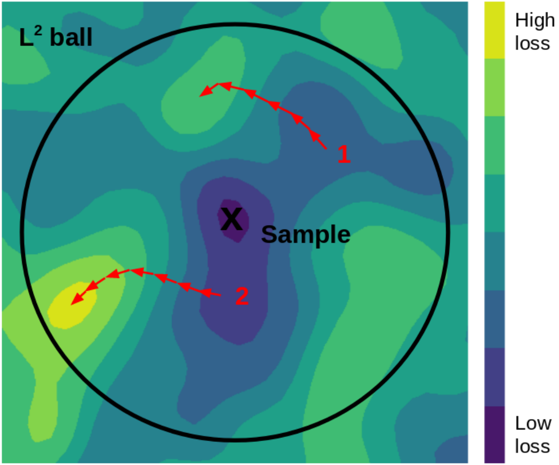



Figure: $L^2$ Ball

# Approach used to solve the problem

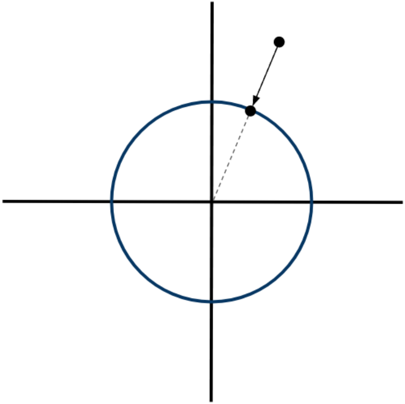


Figure: Attack

# Data-set

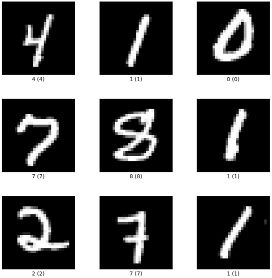
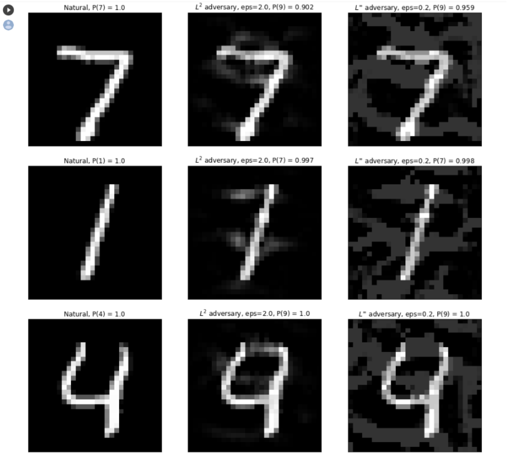



Figure: MNIST [3] dataset

[3] https://www.tensorflow.org/datasets/catalog/mnist

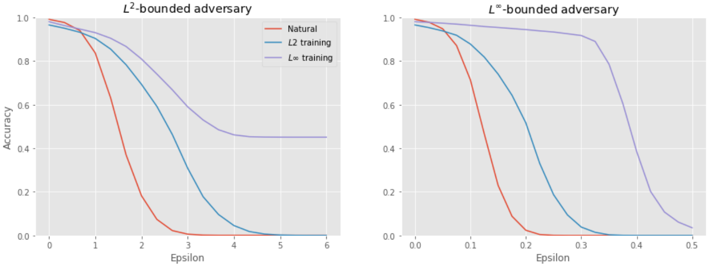
# Experimental results

PGD attack image



# Experimental results

Adversarial training to defend against adversarial attack image.



# Accuracy

- Testing accuracy

```
Epoch 11: 100%|██████████| 469/469 [01:07<00:00,  1.39s/it, loss=0.0237, accuracy=0.993, val_loss=0.0343, val_accuracy=0.988]
Epoch 12: 100%|██████████| 469/469 [01:07<00:00,  1.40s/it, loss=0.0218, accuracy=0.994, val_loss=0.0342, val_accuracy=0.988]
Epoch 13: 100%|██████████| 469/469 [01:07<00:00,  1.38s/it, loss=0.0201, accuracy=0.995, val_loss=0.0341, val_accuracy=0.989]
Epoch 14: 100%|██████████| 469/469 [01:07<00:00,  1.39s/it, loss=0.0185, accuracy=0.995, val_loss=0.0344, val_accuracy=0.989]
Epoch 15: 100%|██████████| 469/469 [01:07<00:00,  1.38s/it, loss=0.0172, accuracy=0.996, val_loss=0.0348, val_accuracy=0.989]
Epoch 16: 100%|██████████| 469/469 [01:07<00:00,  1.37s/it, loss=0.0159, accuracy=0.996, val_loss=0.0355, val_accuracy=0.989]
Epoch 17: 100%|██████████| 469/469 [01:07<00:00,  1.38s/it, loss=0.0146, accuracy=0.996, val_loss=0.036, val_accuracy=0.989]
Epoch 18: 100%|██████████| 469/469 [01:07<00:00,  1.38s/it, loss=0.0135, accuracy=0.997, val_loss=0.0365, val_accuracy=0.989]
Epoch 19: 100%|██████████| 469/469 [01:07<00:00,  1.38s/it, loss=0.011, accuracy=0.997, val_loss=0.0273, val_accuracy=0.992]
Epoch 20: 100%|██████████| 469/469 [01:07<00:00,  1.39s/it, loss=0.0101, accuracy=0.997, val_loss=0.027, val_accuracy=0.992]Finished.
```

- Accuracy of adversarial test : 0.9915

# Conclusion

- The $L^\infty$ trained model is more robust against both L² and L bounded attacks.
- Both the robust models exhibit lower accuracy on natural samples (epsilon $=$ 0) than the non-robust model: 0.5% for L model, 3% for L² model
- The L² attack appears to saturate in effectiveness

# REFERENCE

[1] Jerome Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger, "Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses", 2019, CVPR.

*Thank You*