Adversarial Attacks and it's Defense

Report submission for the course
Adversarial Machine Learning

by

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

under the guidance of

Dr. Manjunath Joshi

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND
COMMUNICATION TECHNOLOGY
GANDHINAGAR, GUJARAT

January, 2022

Acknowledgement

We wish to acknowledge our profound sense of gratitude to our project guide Dr. Manjunath Joshi, DA-IICT Gandhinagar.

Indeed, it was a matter of great facility and privilege for us to work under their aegis. We express our thankfulness to them for their dedicated inspiration,lively interest and patience through our errors, without which it would have been impossible to bring the project to near completion.

At last I must express my sincere heartfelt gratitude to all the staff members of Computer Science Department who helped me directly or indirectly during this course of work.

CANDIDATE'S DECLARATION

We are undersigned solemnly declare that the report of the project work entitled "Adversarial Attacks and it's Defense ", is based on our own work carried out during the course of our study under the guidance of Dr.Manjunath Josh.

(Signature of candidates)
Raja Babu Meena
Rohan Baghel
Date - 12-05-2022

Contents

ABSTRACT

Machine Learning is a Form of AI that Enables a System to Learn from Data. Its important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. for such Machine learning models adversarial attack can be done to misclassify the models.

To overcome from this problem firstly we trick the ML(Machine Learning) models by providing deceptive input and then done the defence mechanism to correct the ML model.

Since few research has been taken placed regarding the Adversarial machine learning. In our studies we are performing PGD attack for misclassification of the model.

1 Introduction

Nowadays, machine learning models are used in many real-world applications, like self-driving cars, face recognition, cancer diagnosis, or even in next-generation shops in order to track which products customers take off the shelf so their credit card can be charged when leaving.

Attackers are trying to attack the machine learning models and trying to defect the model using different types of models.

In our project we are using MNIST data set to train the model by misclassifying it using the attack method and then training it to defend the attacks.

1.1 Motivation

The current driving force in machine learning is to produce increasingly more accurate models while less attention has been paid to the security and robustness of these models. Adversarial machine learning, a technique that attempts to fool models with deceptive data, is a growing threat in the AI and machine learning research community. The most common reason is to cause a malfunction in a machine learning model. An adversarial attack might entail presenting a model with inaccurate or Representative data as it's training, or introducing maliciously designed data to deceive an already trained model.ML models such as image classifiers are vulnerable to tiny perturbations to their inputs that cause them to make the wrong decisions.

1.2 Problem Statement

The following sequence of events have been implemented in the project to create and defend from a powerful white-box adversarial attack:

- Given a image from MNIST data-set , we are adding deviation in the image using PGD attack.
- PGD is a white box attack which means the attacker has access to the model gradients and has a copy of the model's weights.
- PGD can be considered the most "complete" white-box adversary as it lifts any constraints on the amount of time and effort the attacker can put into finding the best attack.
- The current state of the art defense against this attack is adversarial training.Adversarial training is simply putting the PGD attack inside your training loop.

1.3 Objectives

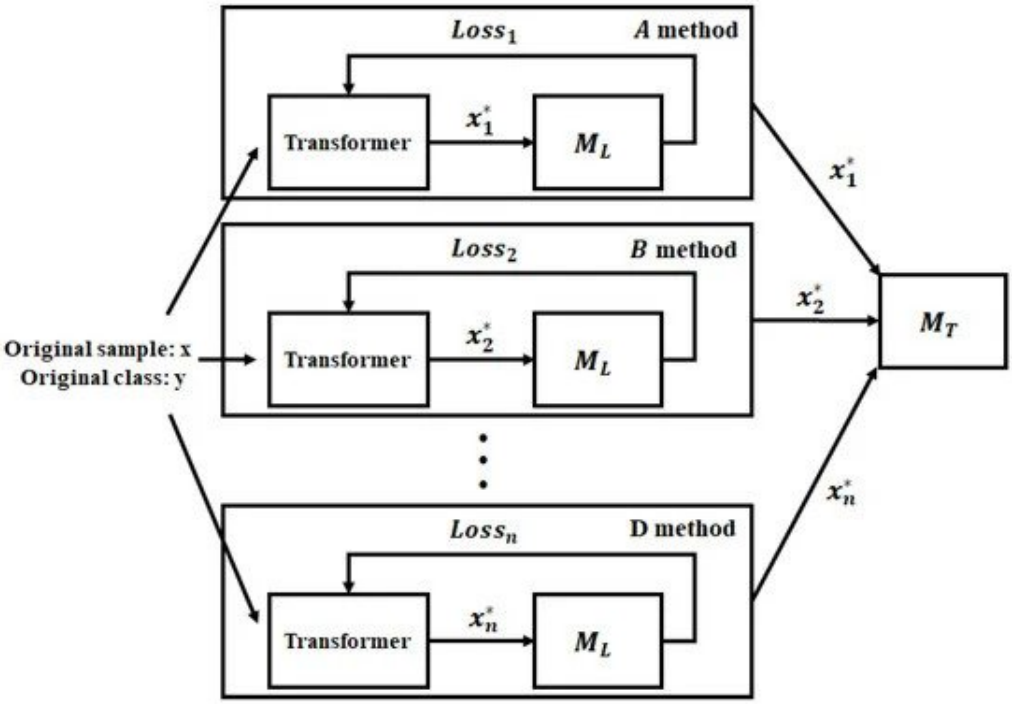
Training the machine learning model against adversarial examples leads to:

- The increase in the robustness of the models against white box adversarial attack.
- The ability to smoothly interpolate between classes using large-epsilon adversarial examples.

2 Proposed Methodology

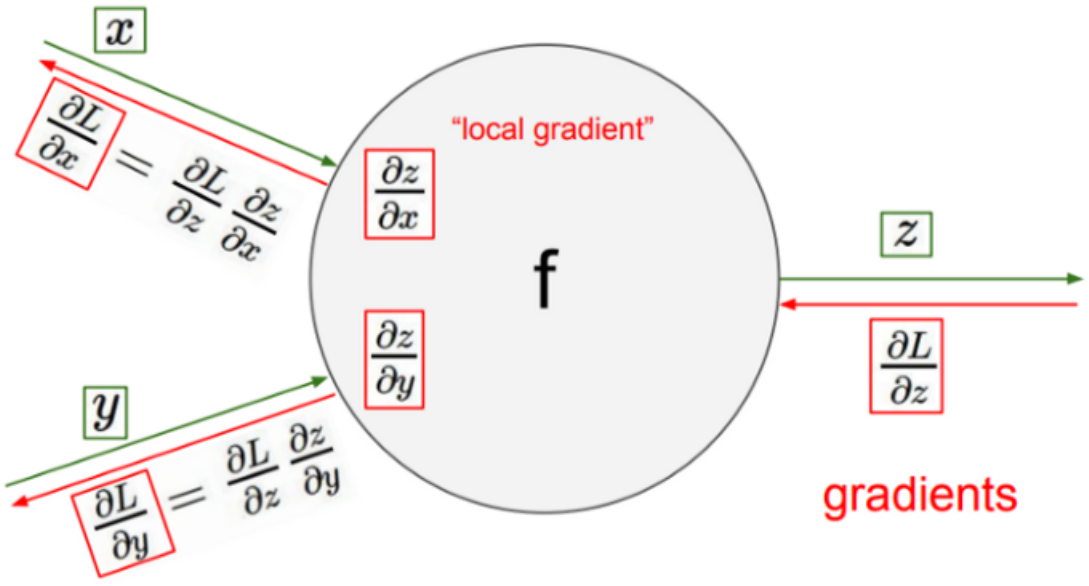
Generation of adversarial examples : The PGD algorithm can be summarised with below defined steps:

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



Gradient-based Adversarial Attacks

Gradient based adversarial attacks exploit a very simple idea originating from the concepts involved in back-propagation. It involves calculating the error(using a specific error function) between the desired output and the network output corresponding to a particular input. Now keeping the input constant, the calculated error is used to compute the gradients corresponding to each parameter(also known as weight) of the network. These gradients are used to update the weights at each step taking into consideration a specific learning rate.

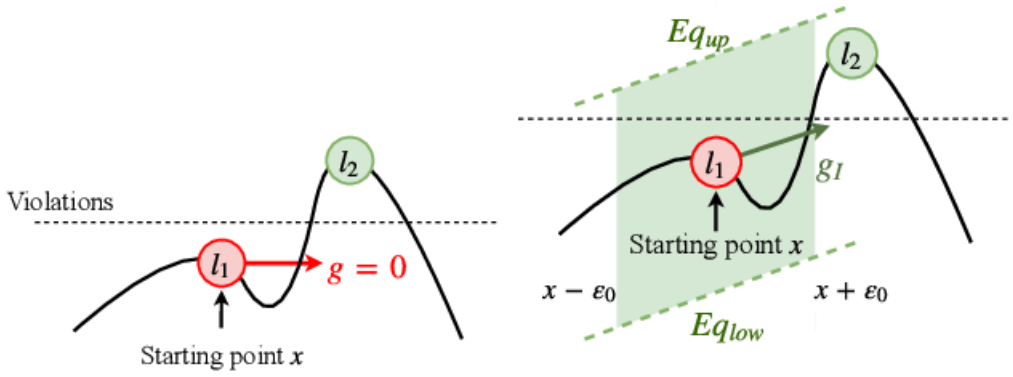


PGD

The PGD attack is a white-box attack which means the attacker has access to the model gradients i.e. the attacker has a copy of your model's weights. This threat model gives the attacker much more power than black box attacks as they can specifically craft their attack to fool your model without having to rely on transfer attacks that often result in human-visible perturbations. PGD can be considered the most "complete" white-box adversary as it lifts any constraints on the amount of time and effort the attacker can put into finding the best attack.

PGD attempts to find the perturbation that maximises the loss of a model on a

particular input while keeping the size of the perturbation smaller than a specified amount referred to as epsilon.



(a) PGD attack (b) Interval attack

Adversarial training

Adversarial training is simply putting the PGD attack inside your training loop. This is can be viewed as "ultimate data augmentation" as instead of performing random transformations (rescaling, cropping, mirroring etc.) as a prepossessing step we create specific perturbations that best fool our model and indeed adversarially trained models do exhibit less overfitting when trained on small datasets.

In regular training we minimise the expected natural loss over our dataset x, y, w.r.t our model parameters, theta.

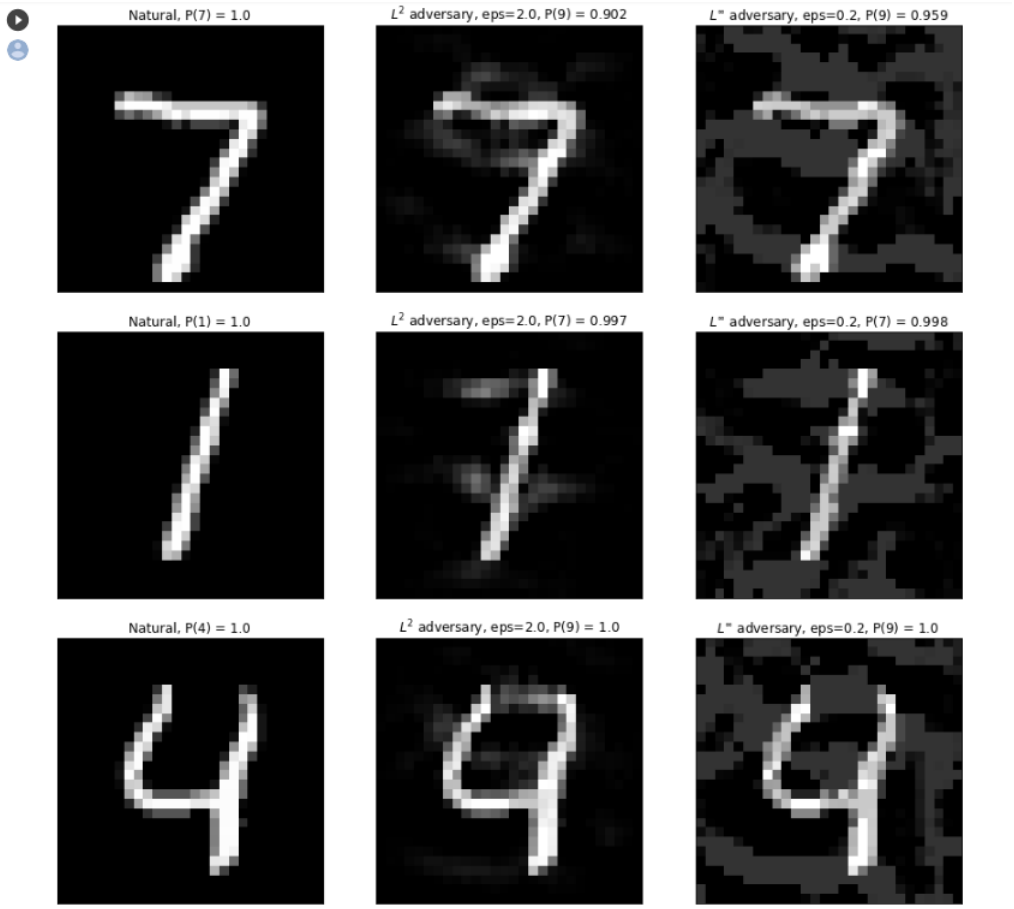
In adversarial training we are minimising the following loss function where is a set of perturbations we want our model to be invariant to such as the L^2 and L perturbations

$$\min_{\theta} \max_{\delta \in \Delta} \mathcal{L}(x + \delta, y; \theta)$$

3 Result and Analysis

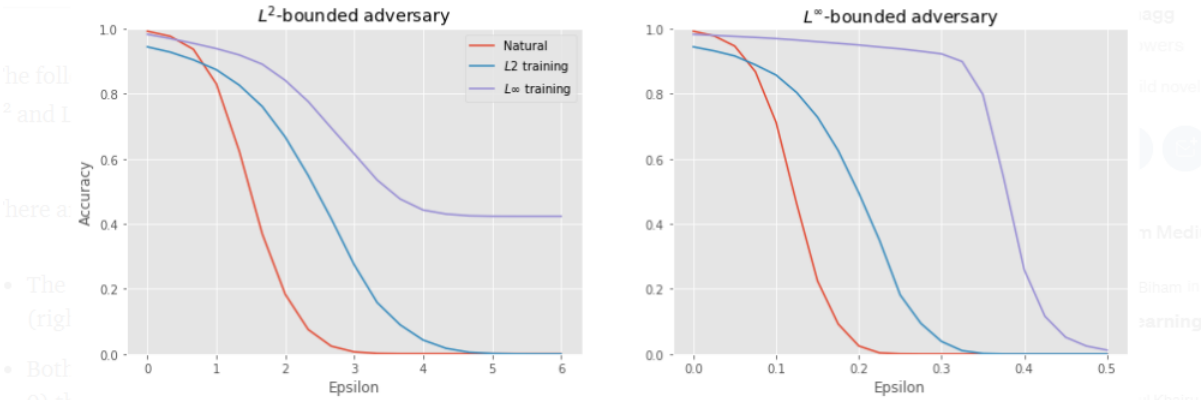
Projected Gradient Descent (PGD)

Misclassified image of one class of image to another class of image, as shown in the figure.



Adversarial Training

Following plots quantify the adversarial accuracy of models trained against L^2 and L adversaries.



4 Conclusion

- The L^∞ trained model is more robust against both L^2 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 L^2 model
- The L^2 attack appears to saturate in effectiveness

References

[1] Ling Huang, Anthony D. Joseph and Blaine Nelson, "Adversarial Machine Learning", 2011

[2] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander M adry, "Robustness May Be at Odds with Accuracy", 2019, Cornell University

[3] Aleksander M adry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks", 2017, Cornell University

[4] Shilin Qiu, Qihe Liu, Shijie Zhou†, and Chunjiang Wu, " Review Review of Artificial Intelligence Adversarial Attack and Defense Technologies", 2019, MDPI

[5] Olakunle Ibitoye, Rana Abou-Khamis, Ashraf Matrawy, and M. Omair Shafiq, "The Threat of Adversarial Attacks on Machine Learning in Network Security – A Survey", 2019, Cornell University

[6] Mo Zhou, Zhenxing Niu, Le Wang, Qilin Zhang and Gang Hua, "settingsOpen AccessEditor's ChoiceReview Review of Artificial Intelligence Adversarial Attack and Defense Technologies", 2020, European Conference on Computer Vision