**ADVERSARIAL MACHINE LEARNING ATTACKS**

**Abstract**: Deep neural networks are found vulnerable to well-designed input samples, called adversarial examples, which is one of the major risks of its application in safety-critical environments. In this project we explain and implement two adversarial attack methods (FGSM -Fast gradient sign method, JSMA -Jacobian saliency map method) referenced from standard research papers and compare the results.

**Team**: Venu Korada, Asha Nagireddy

**DataSet**: MNIST dataset

<http://yann.lecun.com/exdb/mnist/>

The MNIST database contains 60,000 training images and 10,000 testing images.

**Literature:**

An adversarial example is a sample of input data which has been modified very slightly in a way that is intended to cause a machine learning classifier to misclassify it. These modifications can be so subtle that a human observer does not even notice.

Adversaries can attempt to

* manipulate the collection of data.
* corrupt the model.
* tamper with the outputs.

An Adversarial attack is also classified by the actions and information the adversary has at their disposal. These attacks can be classified into

* WHITE BOX-information about the model parameters or its original training data
* BLACK BOX-no knowledge about the model.
* Untargeted Attack is a source class misclassification attack.
* Targeted Attack is a source-target misclassification.

***Fast Gradient Sign Method:***

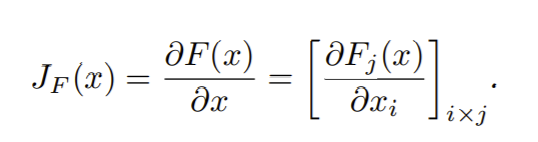
For a given model with parameters of θ, input x, target y and cost function J(θ,x,y), we can generate a optimal linear perturbation:

η=ϵsign(∇xJ(θ,x,y))

This is a untargeted attack and can also be made a targeted attack.

***Jacobian-based saliency map approach:***

Papernot et al. designed an efficient saliency adversarial map, called Jacobian-based Saliency Map Attack. They first computed Jacobian matrix of given sample x, which is given by:



This method generates targeted examples.

**Research Papers:**

<https://arxiv.org/pdf/1712.07107.pdf>

<https://arxiv.org/pdf/1611.01236.pdf>

In these papers, they introduced the background of deep learning techniques, models, and datasets. Discussed about adversarial examples raised in conventional machine learning,

proposed a taxonomy of approaches for generating adversarial examples and elaborated on those approaches. Finally, discussed applications for adversarial examples.

Presently we concentrate on implementing those attacks.

**References:**

<https://ieeexplore.ieee.org/document/8406613>

<https://arxiv.org/pdf/1607.02533.pdf>

<https://arxiv.org/pdf/1412.6572.pdf>