PROJECT PROPOSAL

Study of Nature of Different Adversarial Attacks and Impact of Different Data Augmentation Techniques on Robustness

IIT JODHPUR

Debalina Saha (M20CS055), Sanyam Jain (P20QC001)

January 4, 2021

1 MOTIVATION

With the rise of advanced Deep Neural Network Architectures and various adversarial attacks, we are still lacking in identification of these attacks. There is a great challenge in identifying the correct class of attacks with the help of models. Though we can answer questions such as if the image is perturbed or not by the help of detection methods, however, its important to identify the particular class of attack by which the image was perturbed. Can we take various attacks, and study what kind of changes are made by the attacks? Can we categorize the attacks? In addition, it is important to study how data augmentation techniques help in model robustness against adversarial attacks. With further extension, images with adversarial perturbation have also been used as data augmentation for adversarial training of CNN in some previous works ([1]). It is worth studying how the impact of such adversarial training differs from widely used basic data augmentation techniques on robustness of the model.

2 Proposed Ideas

2.1 Categorization

Firstly, we will try to categorize different adversarial attacks which could help in detection of these attacks. For this purpose, we plan to study feature extractions for various kinds of attacks, both handcrafted and deep learning. We also plan to make use of different feature visualization techniques in order to learn the difference between impact of various attacks.

2.2 Data Augmentation and Robustness

Secondly, we will study the impact of different data augmentation techniques on robustness of DNN models against various attacks. For this purpose, we will take a pre-trained model, and fine-tune it with various data augmentation techniques, firstly with basic techniques like random crop, random noise injection etc, then we will use adversarial perturbed images as data augmentation. Then we will study the increase in robustness of the model. We plan to do a comparative study to understand what kind of data augmentation technique works better against each attack. We can take a few attacks, and for each attack, we can compare the increase in robustness of the model for each data augmentation technique.

3 Related works

[1] Allen-Zhu, Zeyuan Li, Yuanzhi. (2020). Feature Purification: How Adversarial Training Performs Robust Deep Learning.

[2] Goswami, Gaurav Agarwal, Akshay Ratha, Nalini Singh, Richa Vatsa, Mayank. (2019). Detecting and Mitigating Adversarial Perturbations for Robust Face Recognition.