Security and Privacy of Machine Learning - Homework 2

Wu-Jun Pei
National Taiwan University
b06902029@ntu.edu.tw

Abstract

Despite the mightiness of deep nerual networks, several studies have shown that they are vulnerable to adversarial examples,

1 Introduction

In this homework, we're going to build a black-box defense on CIFAR-10 dataset.

2 Methods

2.1 Adversarial Training

2.2 Preprocessing-based Defenses

In my previous homework, I already showed that some preprocessing-based defenses, such as vanilla JPEG Compression, are effective enough to eliminate the influence of adversarial perturbations. In this homework, I'm going to explore defenses that are more effective.

2.3 Evalution

3 Experiments

4 Conclusion

References