A Useful Detector for Adversarial Images

Yifeng Xiong, 14648645, yifengx4@uci.edu

Abstract

Deep neural networks, such as CNN, ResNet, are useful tools for image identification. However, with some little change, the adversarial images could fool the neural network. In this project, we build a powerful detector based on Autoencoder. The dataset we use is CIFAR-10. We first train an Autoencoder with clean images, and we use the Autoencoder to find the latent space of clean images and perturbed images. Finally, we use a simple multi-layer fully connected neural network to identify whether the image is perturbed. We get 100% correct on valid data, and \_\_\_% on test data.

Introduction

Deep neural networks are the most common techniques that used for image identification. Residual Neural network, from Kaiming He et al., 2015, has test error 8.75% with 20 layers. However, these strong neural networks are not robust for some attack method, such as Fast Gradient Sign Method (Goodfellow et al. 2015,) Projected Gradient Descent (Madry et al. 2017,), etc. Adversarial attacks can fool the neural networks with only small perturbations to images but cannot fool human beings. The main purpose for this project is to find some ways to enhance the robustness of neural networks. We devise two detectors to defend adversarial inputs.

* Autoencoder based detector

With the help of autoencoder, we could find out the latent space for inputs. Although those adversarial images are quite like origin one, they are quite different in the latent space. Then we use a simple 3-layers fully connected neural network to detect.

* Normalizing flow based detector

We are considering using normalizing flow later.

The main point in this project is to find out the latent space.

Related Work

[[1608.00530] Early Methods for Detecting Adversarial Images (arxiv.org)](https://arxiv.org/abs/1608.00530)

[[1702.06280] On the (Statistical) Detection of Adversarial Examples (arxiv.org)](https://arxiv.org/abs/1702.06280)

[[1705.07263] Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods (arxiv.org)](https://arxiv.org/abs/1705.07263)

[[1912.05391] Detecting and Correcting Adversarial Images Using Image Processing Operations (arxiv.org)](https://arxiv.org/abs/1912.05391)

Approach

1. Attack method

* Fast Gradient Sign Method
* Projected Gradient Descent

1. How to find latent space
2. Detect adversarial images based on latent space

Current progress

We get very good result right now: 100% correct for valid data. We are considering the following changes:

1. Different datasets, such as CIFAR-100, or ImageNet
2. More attack method (Brendel & Bethge adversarial attack etc.)
3. Using Normalizing flow to find latent space