# Lab 2 Report: Adversarial Attacks on Deep Neural Networks

Sui Huang

April 2, 2018

## 1 FGSM based untargeted attacks

**The success rate of attack increases as epsilon increases.**

| | $\epsilon$ | $SuccessRate$ |
|---|---|---|
| 1 | 1 | 0.03 |
| 2 | 5 | 0.26 |
| 3 | 10 | 0.72 |
| 4 | 20 | 0.991 |
| 5 | 30 | 0.998 |
| 6 | 40 | 1.0 |
| 7 | 50 | 1.0 |

## 2 FGSM based targeted attacks

**The success rate of attack increases as epsilon increases.**
But not as successful as untargeted attack.

| | $\epsilon$ | $SuccessRate$ |
|---|---|---|
| 1 | 1 | 0.0027 |
| 2 | 5 | 0.0456 |
| 3 | 10 | 0.269 |
| 4 | 20 | 0.712 |
| 5 | 30 | 0.876 |
| 6 | 40 | 0.954 |
| 7 | 50 | 0.9866 |

## 3 Adversarial Retraining against Untargeted FGSM Attacks

**Accuracy of the adversarially retrained DNN on the original test dataset**
Accuracy: 0.9314

**FGSM based untargeted attacks using images from the clean test set**
Epsilon: 10 Attack success rate 0.8142

**Thoughts**
Based on the result, I don't think the adversarially retrained DNN is robust against adversarial perturbations.
(Some of my friends got the result opposite from mine, but I can't figure out the reason, so I stick to what I got)

**Repeat Step 3**

| | $\epsilon$ | $Accuracy$ | $SuccessRate$ |
|---|---|---|---|
| 1 | 1 | 0.90 | 0.030 |
| 2 | 5 | 0.62 | 0.299 |
| 3 | 10 | 0.06 | 0.92 |
| 4 | 20 | 0 | 1 |
| 5 | 30 | 0 | 1 |
| 6 | 40 | 0 | 1 |
| 7 | 50 | 0 | 1 |