CSCI 677 Homework 6
Adversarial Attacks on Classification Networks

Nikhit Mago

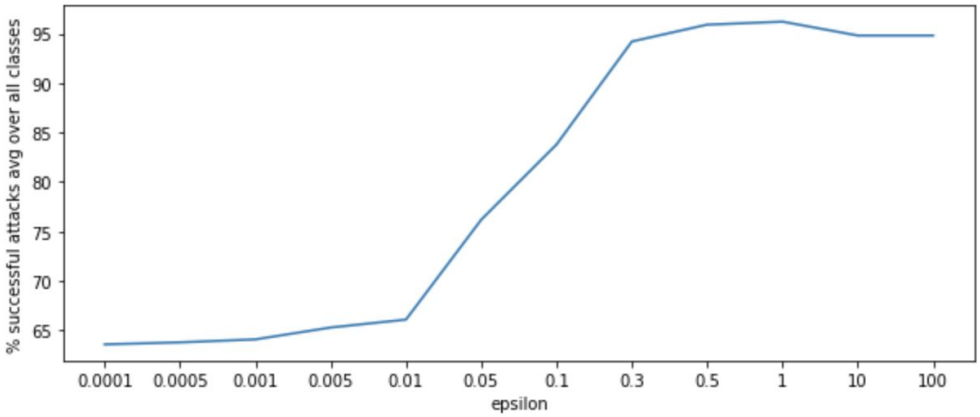
1. Introduction

In this homework assignment, I have implemented two adversarial attacks on a LeNet-5 image classification network -- FGSM and I-FGSM using the targeted and untargeted settings.

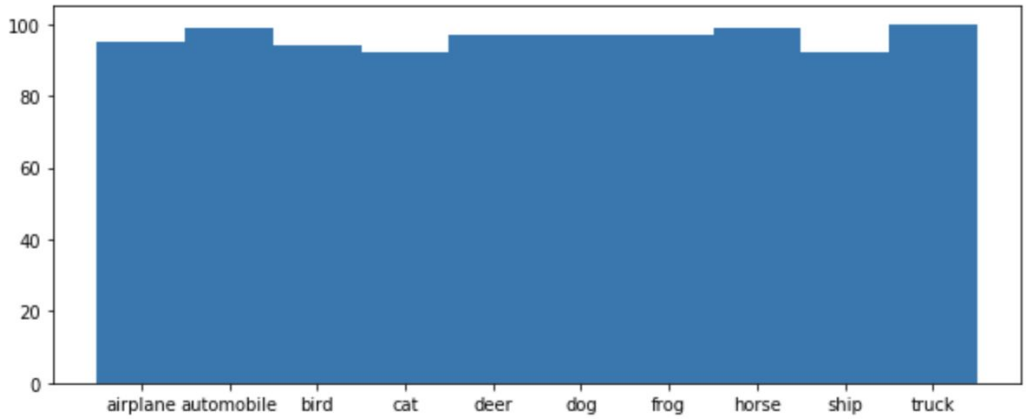
2. Quantitative Results

The results in this section show the Percentage of successful attacks averaged over all 10 classes. The alpha values are calculated by dividing the epsilon values by 5 and num_iterations = 5. Class accuracies (number of successful attacks) are also shown.

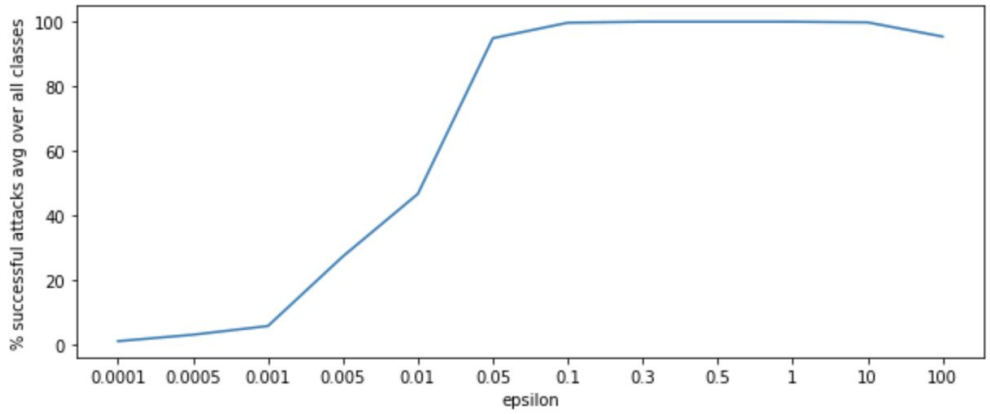
T1 X M1

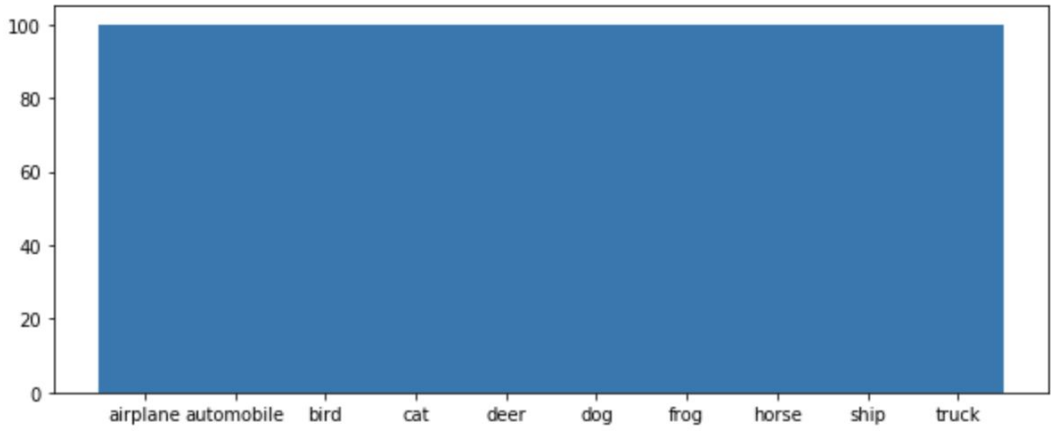


-

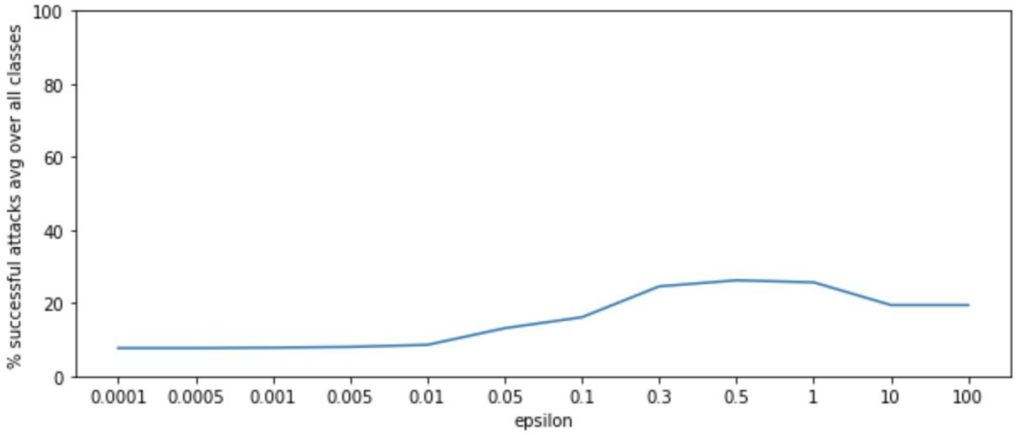


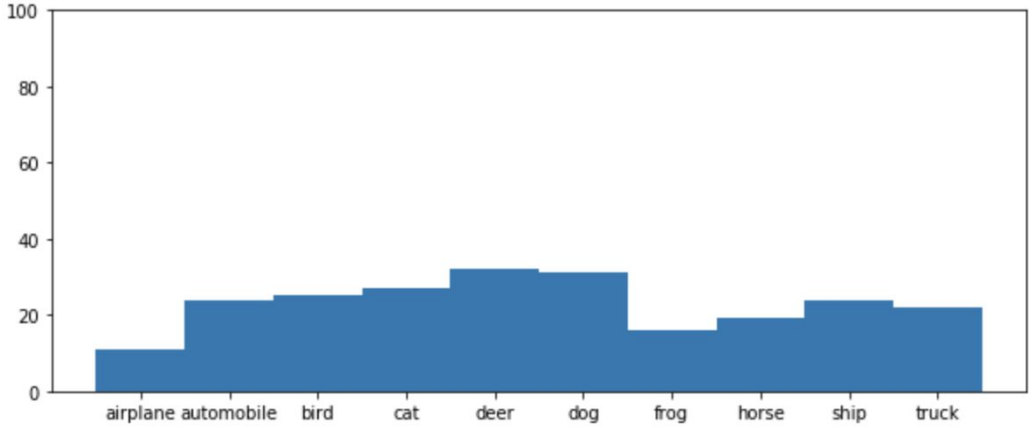
T1 X M2


% successful attacks avg over all classes
100
80
60
40
20
0
0.0001
0.0005
0.001
0.005
0.01
0.05
0.1
0.3
0.5
1
10
100
epsilon

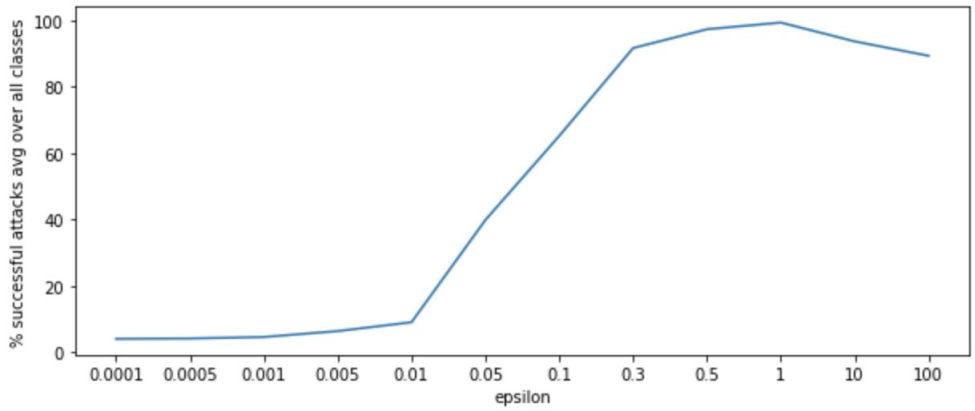

100
80
60
40
20
0
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

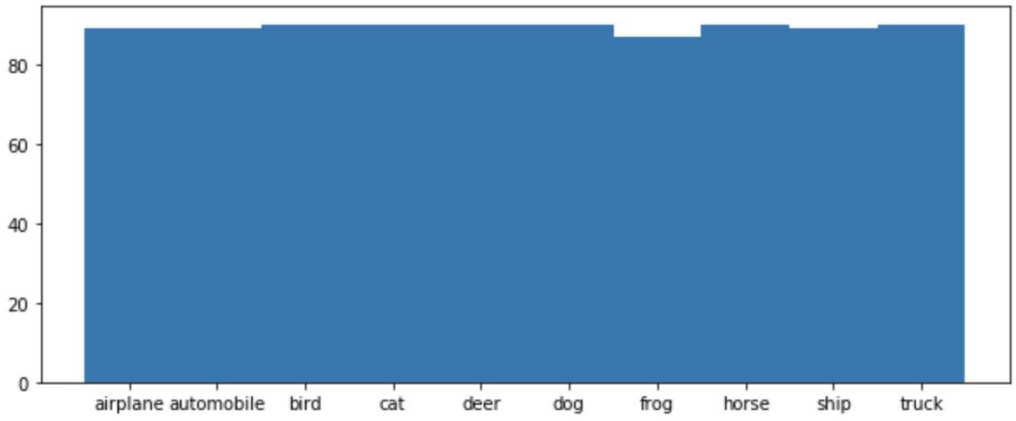
T2 X M1


% successful attacks avg over all classes
100
80
60
40
20
0
0.0001
0.0005
0.001
0.005
0.01
0.05
0.1
0.3
0.5
1
10
100
epsilon


100
80
60
40
20
0
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

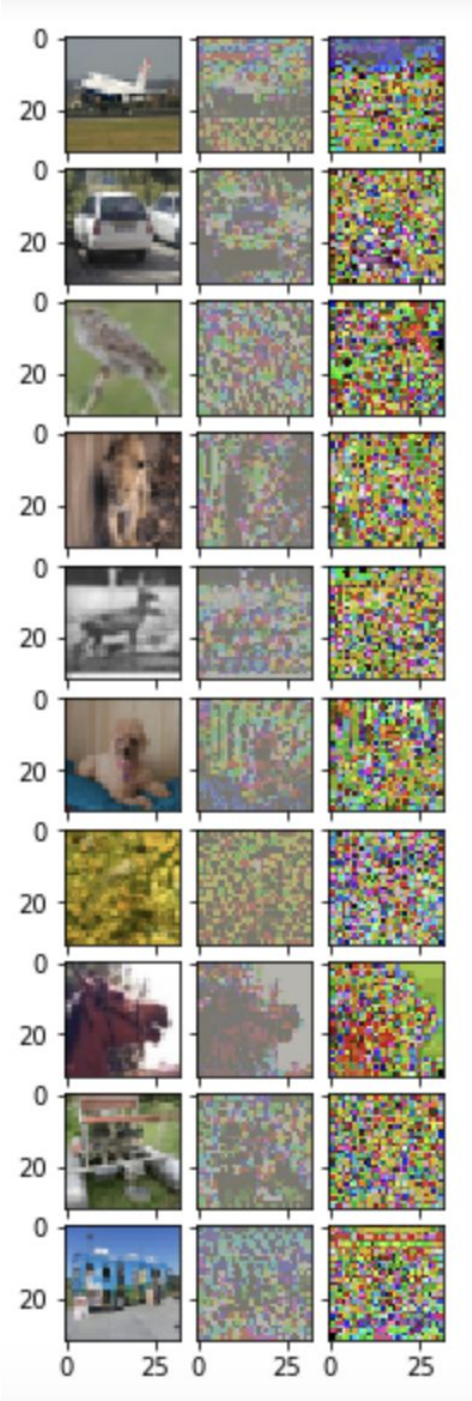
T2 X M2


% successful attacks avg over all classes
100
80
60
40
20
0
0.0001
0.0005
0.001
0.005
0.01
0.05
0.1
0.3
0.5
1
10
100
epsilon


80
60
40
20
0
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

3. Visual Results

T1 (For FGSM)



T2 (For I-FGSM)