Indian Institute of Technology, Jodhpur, India

Department of Computer Science and Engineering

Dependable AI — CSL7370

Assignment on Adversarial Learning

Tejas Gaikwad

Roll No. : MT19AI021

Contents

1 Question 1: Adversarial Attack

Download CIFAR-10 dataset from this link: CIFAR10.
(a) Take any deep model of your choice (say VGG16 or ResNet50 model) and train from scratch (random initialization) for 10-class classification. Report accuracy (overall and classwise) on the testing set.
(b) Perform (i) FGSM, (ii) L0, (iii) L2, and (iv) L adversarial attacks. Perform these attacks as both targeted and untargeted on the testing set.
(c) Report mean SSIM.
(d) Report accuracy after performing the attack. Compare this with the accuracy reported in (a). Plot the histogram for the magnitude of the perturbation obtained. Give proper justifications and inferences on the performance of each attack based on accuracy, perturbation magnitude, and SSIM.
You have to submit the code, model, test adversarial images, and perturbation. For adversarial images and perturbation, you have to submit a .mat file. Submit these as a Google Drive link as Assign2-Q1-*.mat.

Answer
(a).

Class-wise and Overall	**Accuracy**
Accuracy for airplane class	88.8%
Accuracy for automobile class	92.8%
Accuracy for bird class	90.1%
Accuracy for cat class	79.7%
Accuracy for deer class	85.9%
Accuracy for dog class	77.3%
Accuracy for frog class	96.6%
Accuracy for horse class	92.3%
Accuracy for ship class	94.7%
Accuracy for truck class	91.3%
Test Accuracy(Overall)	88.95%

Link for adversarial images, perturbation and trained model :
https://drive.google.com/drive/folders/1AUTLJHRoccXIrjbTuIqamghNKJmw9fA8?usp=sharing

(c). SSIM
The structural similarity index measure is a method for predicting the perceived quality of digital television and cinematic pictures, as well as other kinds of digital images and videos. SSIM is used for measuring the similarity between two images.
Below is the Structural similarity index obtained ofr perturbed images wrt original

images

Label airplane epsilon 0 Score:	0.00033042958248702436
Label airplane epsilon 0.01 Score:	0.031754826721695446
Label ship epsilon 0.1 Score:	-0.03939280114923455
Label ship epsilon 0.15 Score:	0.32944624217129304
Label ship epsilon 0.25 Score:	0.01847933388001014
Label automobile epsilon 0 Score:	-0.010665751111639058
Label automobile epsilon 0.01 Score:	-0.0008246656411198991
Label cat epsilon 0.1 Score:	-0.026005788270883498
Label cat epsilon 0.15 Score:	-0.048047479851101894
Label cat epsilon 0.25 Score:	0.0673186634380698
Label bird epsilon 0 Score:	-0.03694228833244223
Label bird epsilon 0.01 Score:	0.01394093437061017
Label cat epsilon 0.1 Score:	-0.02152833279137442
Label cat epsilon 0.15 Score:	0.0690596671815624
Label cat epsilon 0.25 Score:	0.022288049769543634
Label cat epsilon 0 Score:	0.023030089601136894
Label cat epsilon 0.01 Score:	-0.008629896952849611
Label frog epsilon 0.1 Score:	-0.025304583752305113
Label frog epsilon 0.15 Score:	0.01551463774868781
Label frog epsilon 0.25 Score:	-0.019461227077299234
Label deer epsilon 0 Score:	0.00030206045696250364
Label deer epsilon 0.01 Score:	0.0034046289127940177
Label frog epsilon 0.1 Score:	-0.0012102067482329076
Label frog epsilon 0.15 Score:	0.0064222235291787185
Label frog epsilon 0.25 Score:	0.006388528994605591
Label dog epsilon 0 Score:	-0.004830966039506395
Label dog epsilon 0.01 Score:	-0.012826621839415717
Label deer epsilon 0.1 Score:	-0.03593654405357159
Label bird epsilon 0.15 Score:	0.01461488062446438
Label frog epsilon 0.25 Score:	-0.056800700725044936
Label frog epsilon 0 Score:	0.04345501400233166
Label frog epsilon 0.01 Score:	-0.0048121813943428895
Label bird epsilon 0.1 Score:	-0.020976370618936625
Label bird epsilon 0.15 Score:	-0.03481047450015169
Label bird epsilon 0.25 Score:	0.011624252665608379
Label horse epsilon 0 Score:	-0.03197762884265309
Label horse epsilon 0.01 Score:	-0.01848632714108215
Label horse epsilon 0.1 Score:	-0.04850246824885498
Label horse epsilon 0.15 Score:	0.03128694126556845
Label dog epsilon 0.25 Score:	-0.09160601329624962
Label ship epsilon 0 Score:	-0.01180790544515153

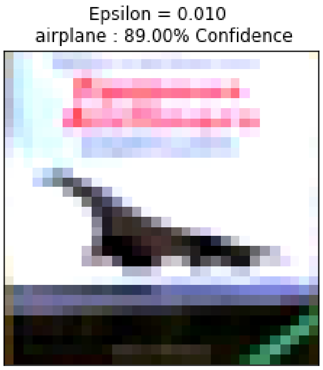
(b). and (d).

(i). FGSM:

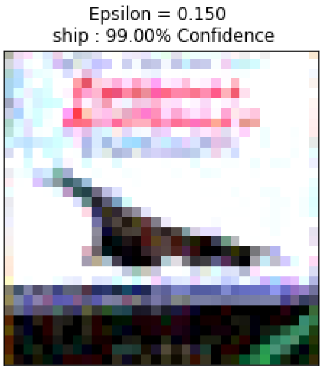
Accuracy after performing attack:

For Class : Airplane

Original Image

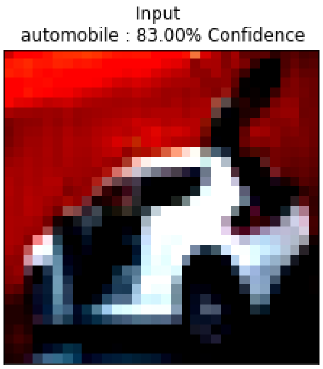


Attacked Image

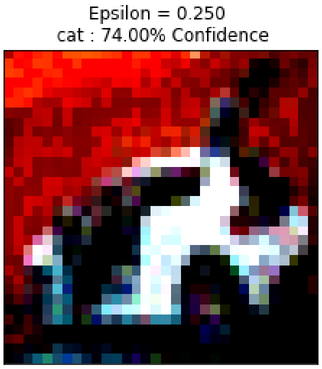


For Class : Automobile

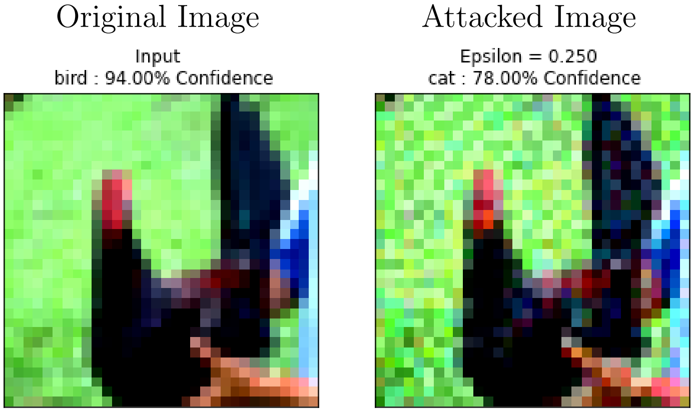
Original Image



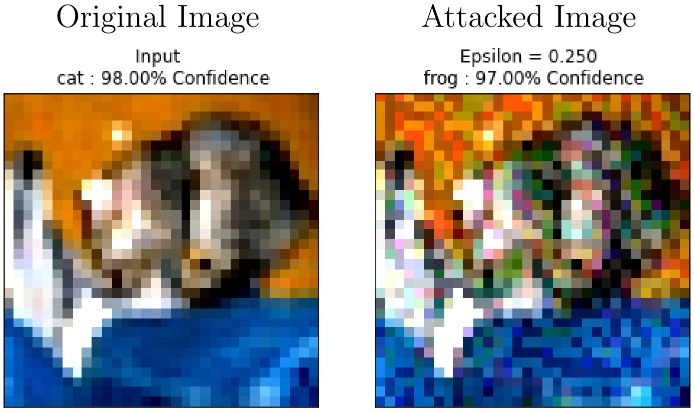
Attacked Image



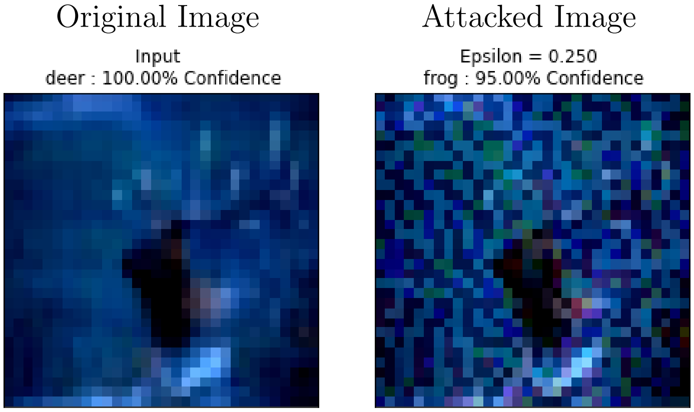
For Class : bird



For Class : cat

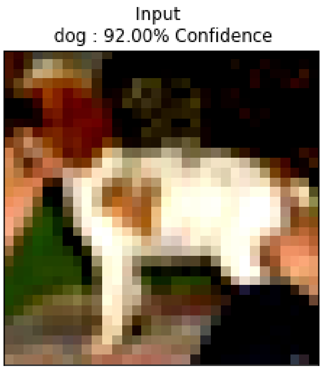


For Class : deer

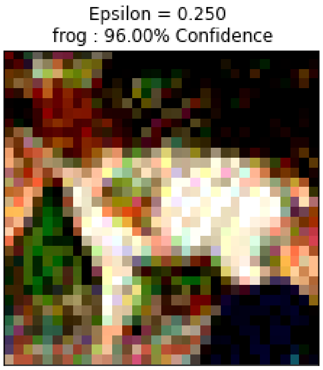


For Class : dog

Original Image

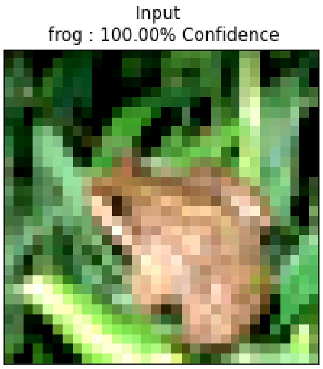


Attacked Image

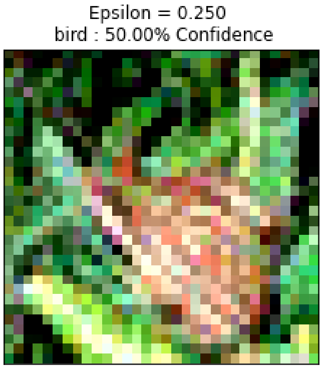


For Class : frog

Original Image

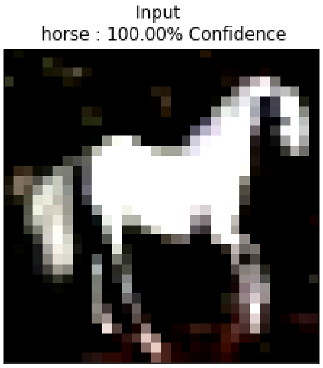


Attacked Image

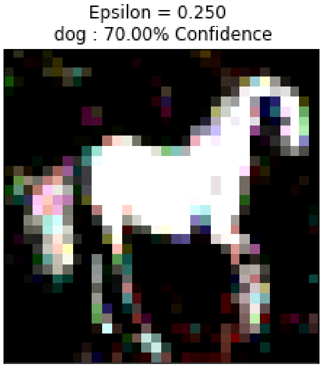


For Class : horse

Original Image

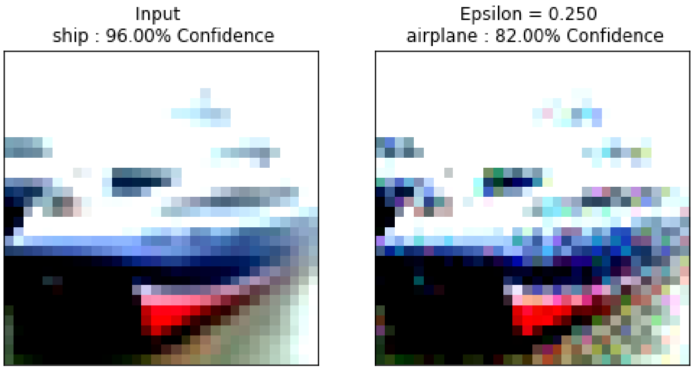


Attacked Image



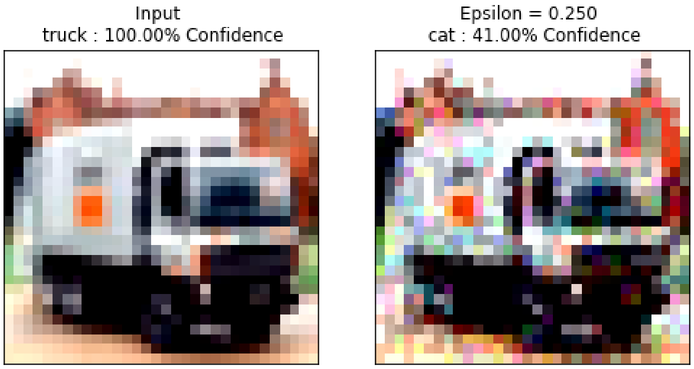
For Class : ship

Original Image Attacked Image



For Class : truck

Original Image Attacked Image



2 Question 3 : Mitigation

(a) Use the perturbed testing images obtained in Question 1 and do JPEG compression at two different compression rates. This has to be performed for the CIFAR10 testing set.
(b) Comment on the difference in classification accuracy observed between:
(i) non-perturbed testing images,
(ii) perturbed testing images, and
(iii) perturbed JPEG compressed testing images.

Answer
(a).

Example Image	**Size(bytes)**
Original Perturbed Image size (Bytes):airplane$_0$	8041
Compressed Perturbed 1 Image size (Bytes):airplane$_0$	6422
Compressed Perturbed 2 Image size (Bytes):airplane$_0$	5364
Original Perturbed Image size (Bytes):ship$_2$	8748
Compressed Perturbed 1 Image size (Bytes):ship$_2$	6878
Compressed Perturbed 2 Image size (Bytes):ship$_2$	5659
Original Perturbed Image size (Bytes):automobile$_0$	7910
Compressed Perturbed 1 Image size (Bytes):automobile$_0$	7000
Compressed Perturbed 2 Image size (Bytes):automobile$_0$	5754
Original Perturbed Image size (Bytes):cat$_2$	8098
Compressed Perturbed 1 Image size (Bytes):cat$_2$	7172
Compressed Perturbed 2 Image size (Bytes):cat$_2$	5877
Original Perturbed Image size (Bytes):bird$_0$	9291
Compressed Perturbed 1 Image size (Bytes):bird$_0$	6973
Compressed Perturbed 2 Image size (Bytes):bird$_0$	5730
Original Perturbed Image size (Bytes):cat$_1$	9597
Compressed Perturbed 1 Image size (Bytes):cat$_1$	7862
Compressed Perturbed 2 Image size (Bytes):cat$_1$	6440
Original Perturbed Image size (Bytes):frog$_2$	9771
Compressed Perturbed 1 Image size (Bytes):frog$_2$	8378
Compressed Perturbed 2 Image size (Bytes):frog$_2$	6816

Example Image	**Size(bytes)**
Original Perturbed Image size (Bytes):deer_0	8708
Compressed Perturbed 1 Image size (Bytes):deer_0	5730
Compressed Perturbed 2 Image size (Bytes):deer_0	4674
Original Perturbed Image size (Bytes):dog_4	7506
Compressed Perturbed 1 Image size (Bytes):dog_4	7250
Compressed Perturbed 2 Image size (Bytes):dog_4	5961
Original Perturbed Image size (Bytes):ship_0	7225
Compressed Perturbed 1 Image size (Bytes):ship_0	6168
Compressed Perturbed 2 Image size (Bytes):ship_0	5115

(b).

It has been observed that for some classes, even after perturbation followed by compression, the classes are predicted accurately. Where as, in most of the cases, the predicted class for every case differs, that may be because of loss of information/ pixels holding the perturbed portion as well as the pixels responsible for accurate classification of the pixel.

Comparison in Classification Accuracy		
Images	Predicted Label	Confidence
Original Image	airplane	93.51%
Perturbed Image	airplane	39.25%
Compressed Perturbed Image(Compression Rate 1)	airplane	97.42%
Compressed Perturbed Image(Compression Rate 2)	airplane	97.73%
Original Image	automobile	98.03%
Perturbed Image	cat	89.72%
Compressed Perturbed Image(Compression Rate 1)	airplane	99.22%
Compressed Perturbed Image(Compression Rate 2)	airplane	98.73%
Original Image	bird	85.27%
Perturbed Image	cat	66.84%
Compressed Perturbed Image(Compression Rate 1)	airplane	93.54%
Compressed Perturbed Image(Compression Rate 2)	airplane	99.33%
Original Image	cat	98.63%
Perturbed Image	cat	92.78%
Compressed Perturbed Image(Compression Rate 1)	truck	49.86%
Compressed Perturbed Image(Compression Rate 2)	truck	55.06%
Original Image	deer	99.82%
Perturbed Image	cat	65.23%
Compressed Perturbed Image(Compression Rate 1)	airplane	92.71%
Compressed Perturbed Image(Compression Rate 2)	airplane	91.83%
Original Image	dog	87.35%
Perturbed Image	cat	59.44%
Compressed Perturbed Image(Compression Rate 1)	airplane	97.94%
Compressed Perturbed Image(Compression Rate 2)	airplane	89.19%
Original Image	frog	99.76%
Perturbed Image	frog	72.48%
Compressed Perturbed Image(Compression Rate 1)	truck	99.86%
Compressed Perturbed Image(Compression Rate 2)	truck	99.91%
Original Image	horse	99.93%
Perturbed Image	cat	32.47%
Compressed Perturbed Image(Compression Rate 1)	airplane	81.77%
Compressed Perturbed Image(Compression Rate 2)	airplane	73.67%
Original Image	ship	99.97%
Perturbed Image	airplane :	45.03%
Compressed Perturbed Image(Compression Rate 1)	airplane	97.89%
Compressed Perturbed Image(Compression Rate 2)	airplane	99.17%
Original Image	truck	99.91%
Perturbed Image	ship	41.6 %
Compressed Perturbed Image(Compression Rate 1)	truck	80.65%
Compressed Perturbed Image(Compression Rate 2)	truck	78.4%