Dhruv Maheshwari 9001607595

CSCI 677 – Advanced Computer Vision

Assignment 6

This assignment uses ResNet-9 trained to withstand Iterative Gradient Sign Method (IGSM) attack.

Code available at:
https://colab.research.google.com/drive/1BNlYKj18RMBtga94SrTbDdGQ2PkeTYe-?usp=sharing

Parameters:

- Iterations = 20
- Epochs = 10
- Replacement Ratio = 0.75

Experiment:

1. 1st experiment
    - $\alpha = 0.006$
    - $\epsilon = 0.12$
2. 2nd experiment
    - $\alpha = 0.004$
    - $\epsilon = 0.14$

Results:

Quantitatively

The following table shows the quantitative results (accuracy) for the two different experiments.

	1st Experiment	2nd Experiment
Without defense		
- Benign	74.29%	74.29%
- IGSM attack	32.61%	35.21%
Adversarial trained		
- Benign	71.98%	71.98%
- IGSM attack	54.17%	56.90%

Here we can see that both the experiments had same results for benign images for both untrained and trained models 74.29% and 71.98% respectively.

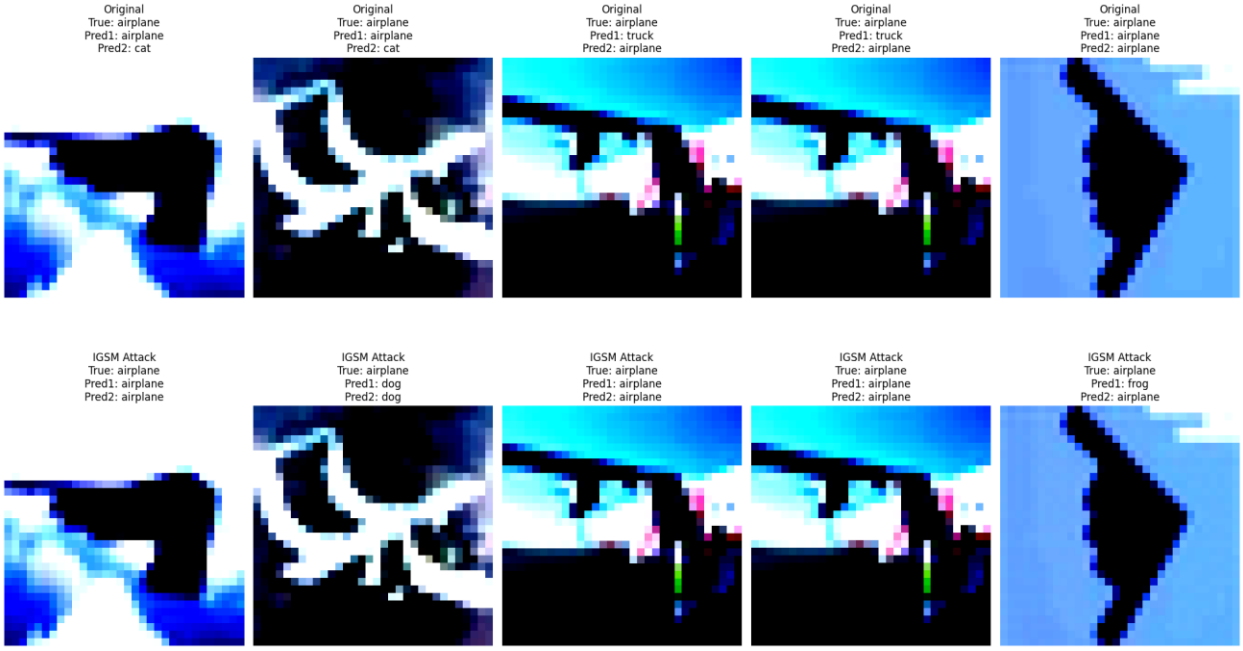
However, for IGSM attacked images, the accuracy for both untrained and trained model rose from 32.61% to 35.21% and 54.17% to 56.90% respectively.

This was as expected, because value of α was decreased, which meant smaller ensured a smoother and more controlled attack. On the other hand, ϵ was increased, allowing larger perturbations, making it more likely to fool the model.

Qualitatively:

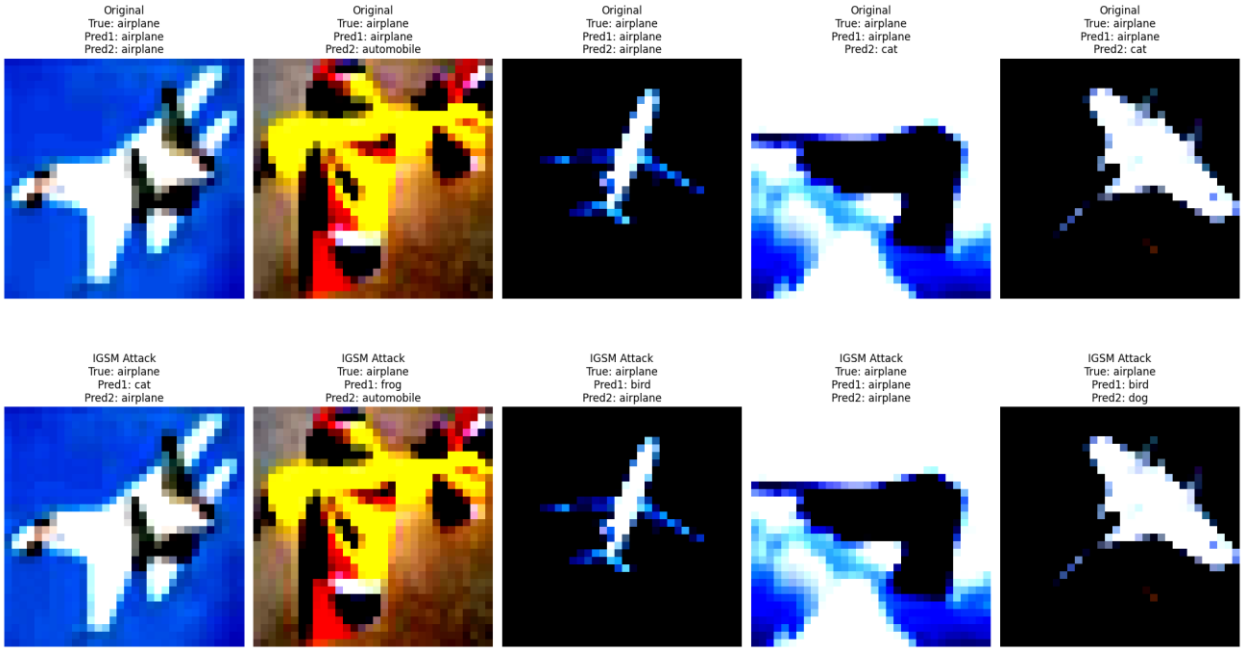
Pred1 = Without defense
Pred2 = Adversarial trained

1st Experiment



-
- Here, we can see that, the original ResNet 9 model gives correct predictions (Pred 1) for 3/5 the images in top row which aren't attacked, same for the adversarial trained model achieves correct prediction (Pred2) for 3/5.
- For the IGSM attacked pictures, the original model (Pred 1) has 3/5 correct predictions, but the adversarial trained model (Pred 2) can correctly predict labels for 4/5 samples, thus showing robustness against attacked images.

2nd Experiment



-
- It is again proved that images that were correctly classified by original ResNet network (first row) were later incorrectly predicted when there was an IGSM attack (bottom row).
- However, the adversarial trained network withstood the attack and gave correct predictions for 3/5 compared to 1/5 for original network.

Hence, both experimental trained are robust and are able to withstand IGSM attack.