CSCI 677 –Advance Computer Vision- Fall 2024 – HW6

USC ID: 1154164561
NAME: Srija Madarapu
EMAIL: madarapu@usc.edu

Code:

https://colab.research.google.com/drive/1HnglxK_bL9B5wT5VKMTNw5F-NoK0WMZh?usp=sharing

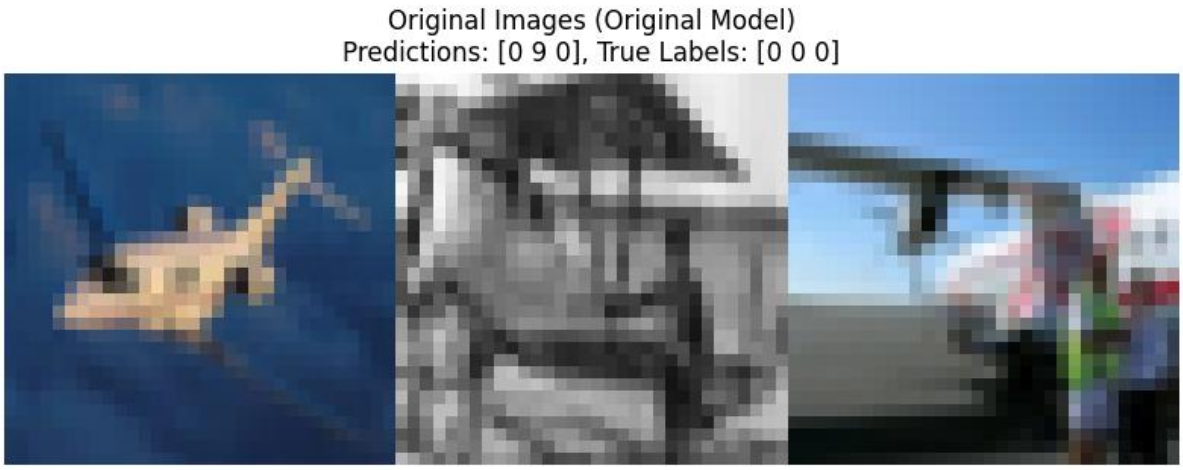
Adversarial Attack

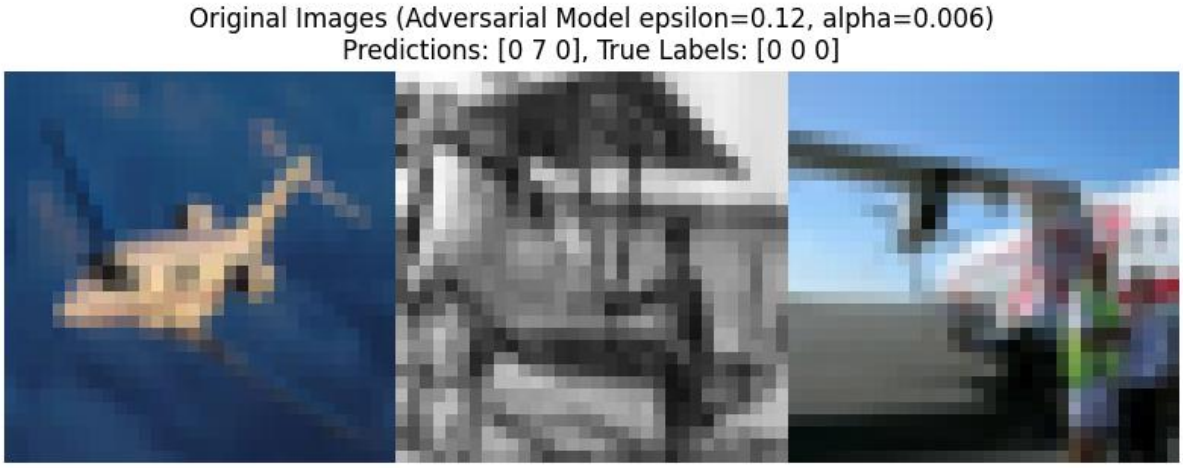
The Iterative Gradient Sign Method is a more powerful adversarial attack technique that enhances the basic gradient sign method by iterating the attack multiple times.It adjusts the image in small steps and applies the gradient to refine the adversarial perturbation, making it more effective in fooling a model.It is particularly useful when you want a stronger adversarial attack and can afford the additional computational cost of multiple iterations.

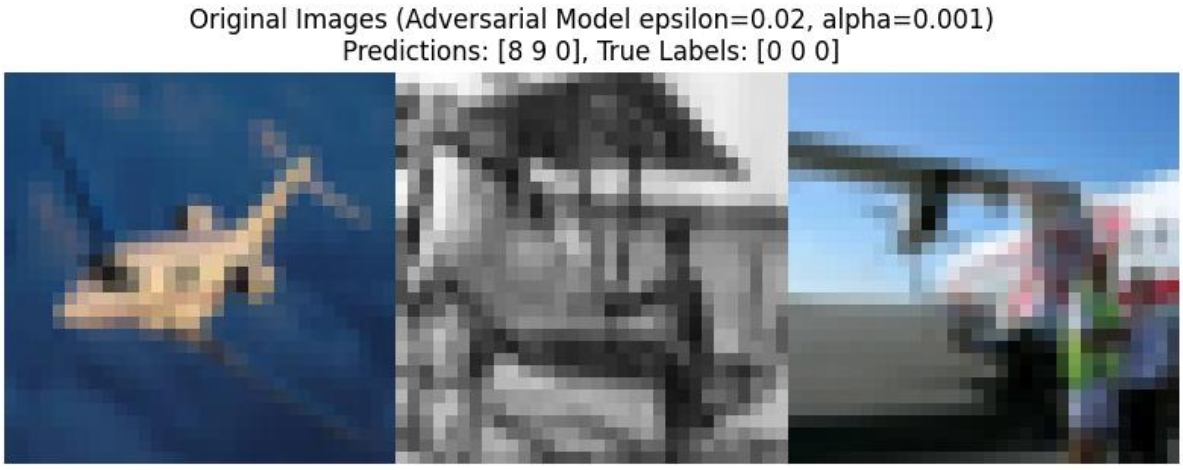
Strong Attack: $\varepsilon = 0.12$ and $\alpha = 0.006$, 20 iterations
Weak Attack: $\varepsilon = 0.02$ and $\alpha = 0.001$, 20 iterations

Epoch [1/20], Time: 106.16s, Train Loss: 1.3739, Train Accuracy: 49.65%, Validation Loss: 1.2360, Validation Accuracy: 55.27%
Epoch [5/20], Time: 99.23s, Train Loss: 0.5975, Train Accuracy: 78.59%, Validation Loss: 0.9367, Validation Accuracy: 68.05%
Epoch [10/20], Time: 115.86s, Train Loss: 0.1038, Train Accuracy: 96.36%, Validation Loss: 1.5468, Validation Accuracy: 68.24%
Epoch [15/20], Time: 95.44s, Train Loss: 0.0595, Train Accuracy: 97.93%, Validation Loss: 1.8047, Validation Accuracy: 69.03%
Epoch [20/20], Time: 97.58s, Train Loss: 0.0444, Train Accuracy: 98.47%, Validation Loss: 1.9147, Validation Accuracy: 68.78%

Accuracy on Benign Images:68.53%
Accuracy on IGSM Attack Images (Without Defense): 0.00%







Strong Attack

Quantitative Results:

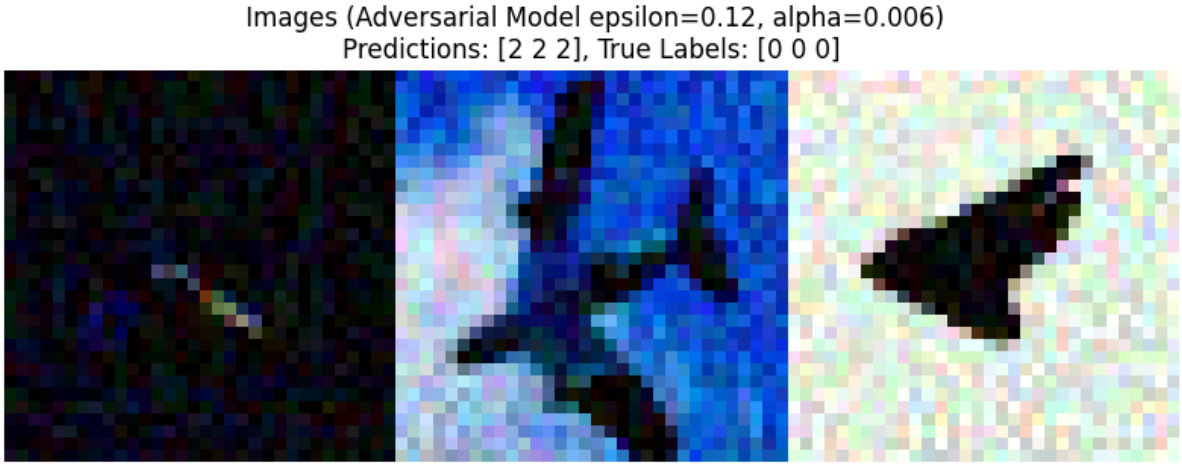
Epoch [1/10], Loss: 1.8869
Epoch [2/10], Loss: 1.6925
Epoch [3/10], Loss: 1.3970
Epoch [4/10], Loss: 0.9487
Epoch [5/10], Loss: 1.4230
Epoch [6/10], Loss: 1.9598
Epoch [7/10], Loss: 1.2249

Epoch [8/10], Loss: 1.2553
Epoch [9/10], Loss: 2.1802
Epoch [10/10], Loss: 1.2805

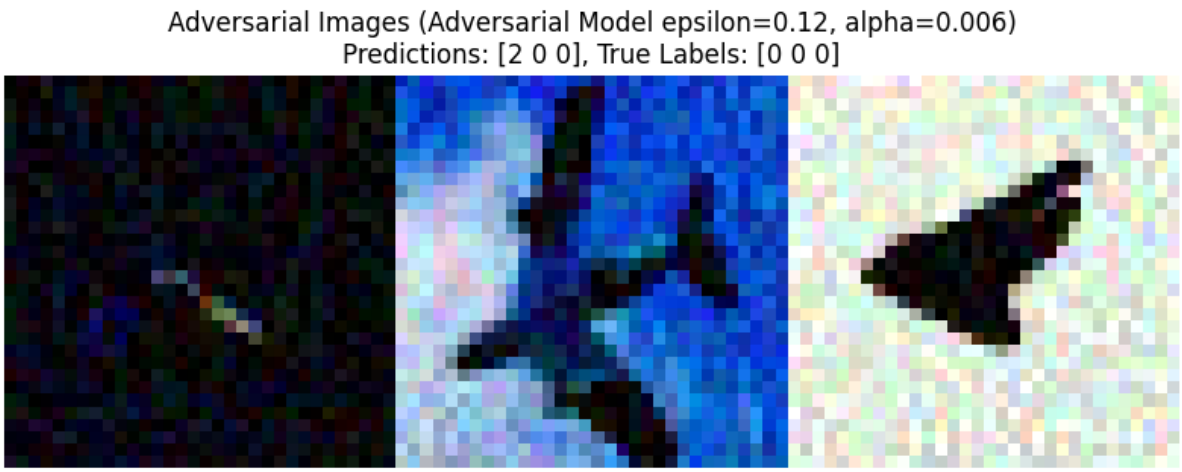
Accuracy After Adversarial Training on Benign Images: 55.27%
Accuracy After Adversarial Training on IGSM Attack Images: 0.10%

Qualitative Results:

Visualizing IGSM Attack Images of the No-Defense Model:



Visualizing IGSM Attack Images After Adversarial Training:



Analysis:

- **Epoch [1/10]**: Loss: 1.8869
- **Epoch [10/10]**: Loss: 1.2805

The loss values fluctuate significantly over the 10 epochs, indicating a high degree of instability in training due to the adversarial nature of the strong attack. Despite this, the model does make some progress in terms of reducing loss over time, though the overall loss remains quite high, demonstrating the difficulty of training against a strong adversary.

- **On Benign Images**: 55.27%
- **On IGSM Attack Images**: 0.10%

Post adversarial training, the model's accuracy on benign images has decreased significantly from 68.53% (before adversarial training) to 55.27%. The model continues to perform poorly under the strong IGSM attack, maintaining 1% accuracy on adversarial samples. This suggests that the adversarial training is not enough to robustly defend against strong adversarial attacks.

Weak Attack

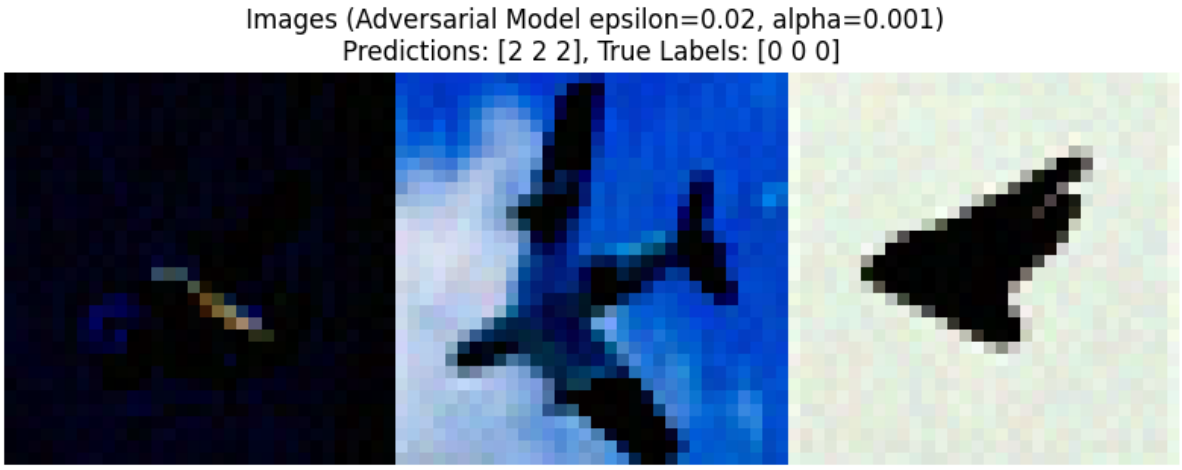
Quantitative Results:

Epoch [1/10], Loss: 1.3121
Epoch [2/10], Loss: 1.4963
Epoch [3/10], Loss: 1.1539
Epoch [4/10], Loss: 1.1950
Epoch [5/10], Loss: 1.2319
Epoch [6/10], Loss: 1.4713
Epoch [7/10], Loss: 1.3307
Epoch [8/10], Loss: 1.3774
Epoch [9/10], Loss: 0.5886
Epoch [10/10], Loss: 0.3156

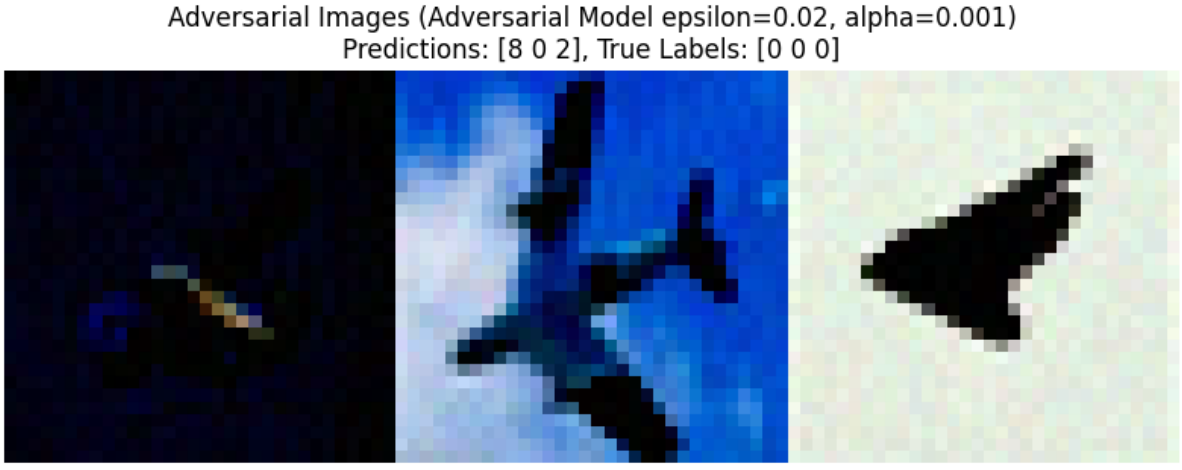
Accuracy After Adversarial Training on Benign Images:59.82%
Accuracy After Adversarial Training on IGSM Attack Images:0.90%

Qualitative Results:

Visualizing IGSM Attack Images of the No-Defense Model:



Visualizing IGSM Attack Images After Adversarial Training:



Analysis:

- **Epoch [1/10]**: Loss: 1.3121
- **Epoch [10/10]**: Loss: 0.3156

The weak attack seems to have a less significant impact on the training compared to the strong attack. The loss shows a steady reduction over the epochs, which indicates that the model is adapting to the perturbations over time.

- **On Benign Images**: 59.82%
- **On IGSM Attack Images**: 0.90%

Post adversarial training with the weak attack, the model's accuracy on benign images drops slightly to 59.82%. However, when subjected to adversarial images, the accuracy remains extremely low at 0.90%, showing that the adversarial training has not been able to provide effective defense against even weak perturbations.

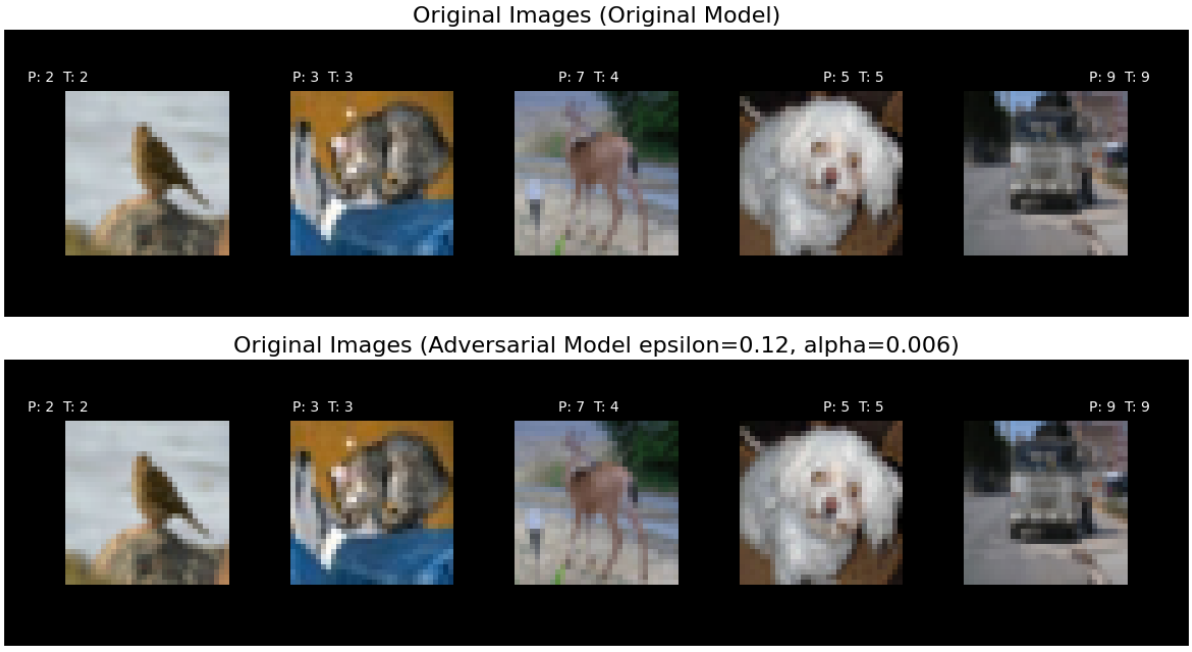
Overview

Adversarial Robustness: The model exhibits limited robustness against both strong and weak adversarial attacks. After adversarial training, accuracy on benign images decreases for both attack types, which is expected as the model is learning to defend itself against adversarial perturbations. However, the drop in accuracy on adversarial images is much more significant, with the model essentially failing to classify adversarial samples correctly after training.

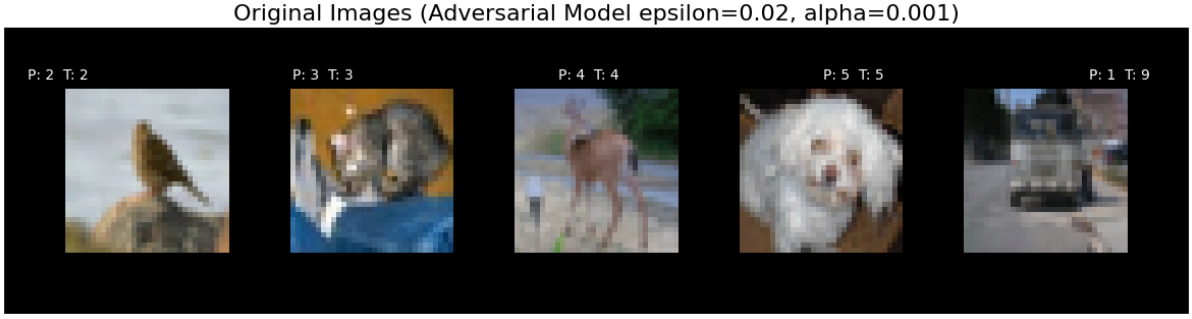
Training Challenges: The training with adversarial examples shows that strong attacks, in particular, significantly hinder the model's ability to generalize, even after 10 epochs of adversarial training. The model's inability to defend effectively against strong adversarial attacks, combined with the drop in accuracy on benign images, suggests that the model's architecture and/or training procedure may need further refinement.

Defense Effectiveness: Both attacks demonstrate the vulnerability of the model, even after adversarial training. The weak attack's impact is smaller, but it still manages to significantly lower the model's performance on adversarial images. The strong attack completely overwhelms the model's defenses, rendering it ineffective at classifying adversarial examples.

Examples

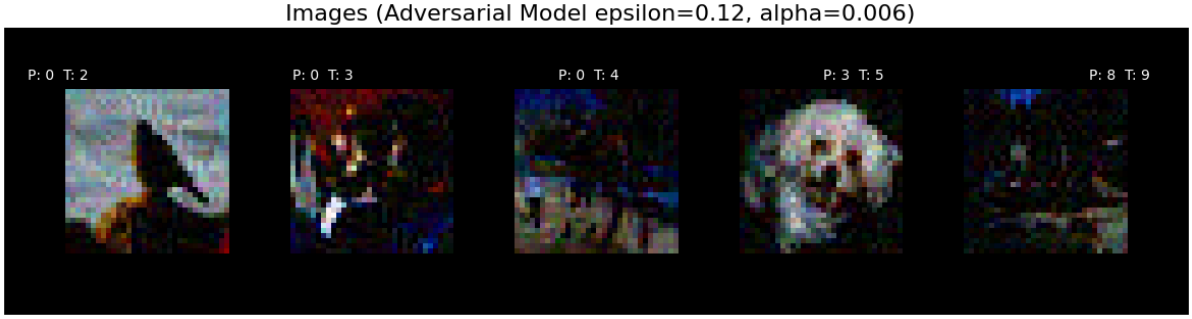


Original Images (Adversarial Model epsilon=0.02, alpha=0.001)

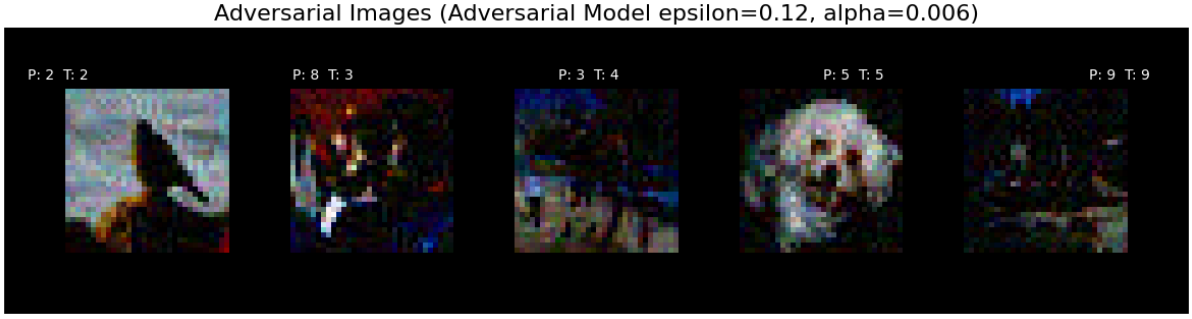


Strong Attack

Images (Adversarial Model epsilon=0.12, alpha=0.006)

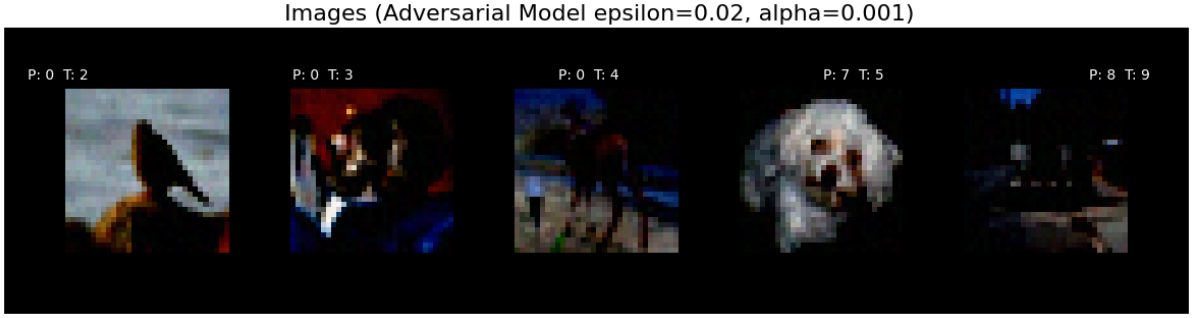


Adversarial Images (Adversarial Model epsilon=0.12, alpha=0.006)



Weak Attack

Images (Adversarial Model epsilon=0.02, alpha=0.001)



Adversarial Images (Adversarial Model epsilon=0.02, alpha=0.001)