ResNet-18 CIFAR-10 Adversarial Robustness Evaluation

Experiment Summary

Clean accuracy: 91.34%. Adversarial accuracy under Auto-PGD (eps=0.0157, steps=100): 16.20%. Attack success rate: 83.80% over 1000 samples.

Analysis of Attack Effectiveness

The Auto-PGD attack demonstrates high effectiveness against the standard-trained ResNet-18 model:

- Clean Accuracy: 91.34% - The model performs well on unperturbed test images.
- Adversarial Accuracy: 16.20% - Performance drops dramatically under attack.
- Attack Success Rate: 83.80% - The attack successfully fools the model in most cases.

Key Observations:

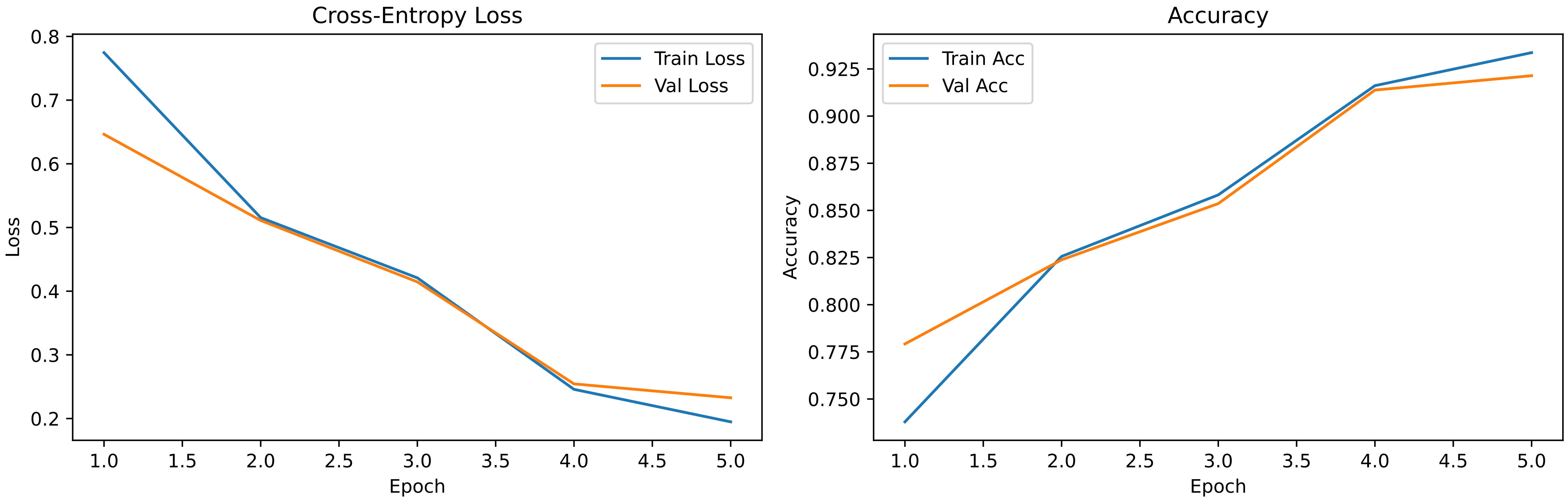
1. The large gap between clean and adversarial accuracy (75.1 percentage points) indicates that the model is highly vulnerable to adversarial perturbations.

2. Despite the perturbations being imperceptible ($\varepsilon=8/255\approx3.1\%$ of pixel range), they effectively mislead the network's predictions.

3. This vulnerability suggests that the model relies on non-robust features that are easily manipulated by small, targeted perturbations.

Parameter Impact Analysis:

- Epsilon (ε): Controls maximum perturbation magnitude. Larger ε → stronger attacks → lower adversarial accuracy, but more visible perturbations.
- Step Size (α): Affects convergence. Typically $\alpha \approx \varepsilon/4$ to $\varepsilon/10$ for optimal results. Too large → overshooting; too small → slow convergence.
- Iterations: More iterations → stronger attack, especially with smaller step sizes. 100 iterations is generally sufficient for convergence.


Cross-Entropy Loss
Loss
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Epoch
Train Loss
Val Loss
Accuracy
Accuracy
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Epoch
Train Acc
Val Acc

Evaluation Summary

Metric	Value
Clean Accuracy	0.9134
Adv Accuracy	0.1620
Attack Success Rate	0.8380

Adversarial Examples Visualization