# ResNet-18 CIFAR-10 Adversarial Robustness Evaluation

## Experiment Summary

DEMO RESULTS (Limited evaluation): Clean accuracy on 100 samples: 77.00%. Adversarial accuracy on 20 samples under Auto-PGD (eps=8/255, 20 steps): 0.00%. Attack success rate: 100.00%. Note: This is a quick demo with reduced sample size and iterations. For full evaluation, use run_experiment.py with default parameters.

## Analysis of Attack Effectiveness

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

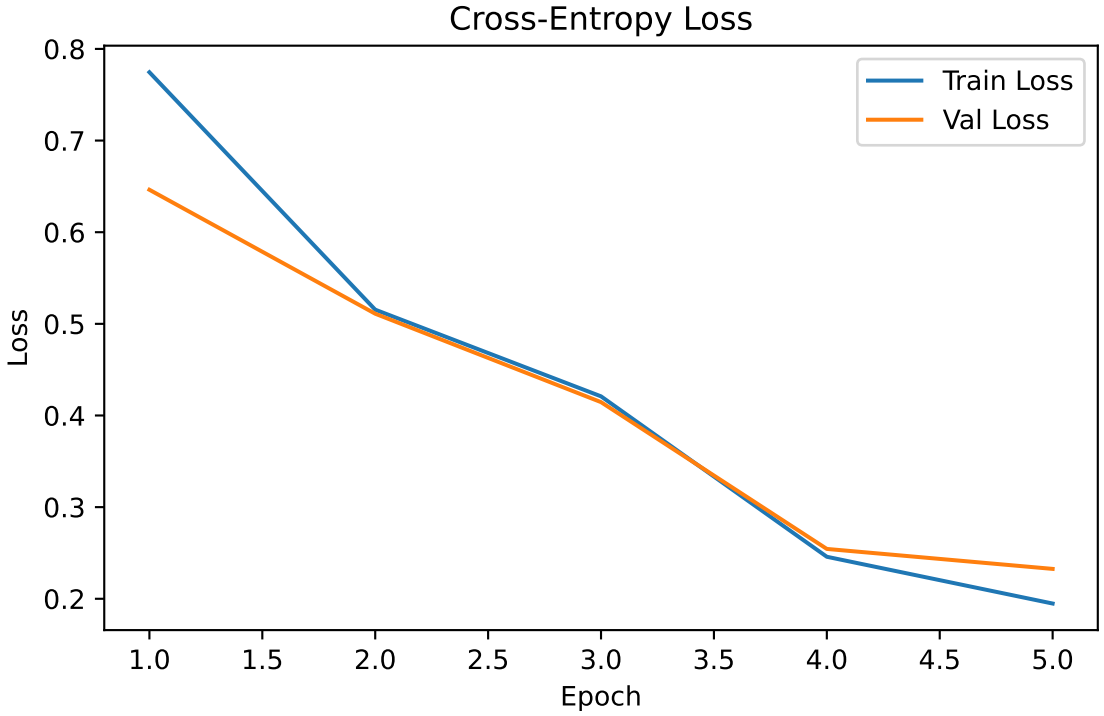
• Clean Accuracy: 77.00% - The model performs well on unperturbed test images.
• Adversarial Accuracy: 0.00% - Performance drops dramatically under attack.
• Attack Success Rate: 100.00% - The attack successfully fools the model in most cases.

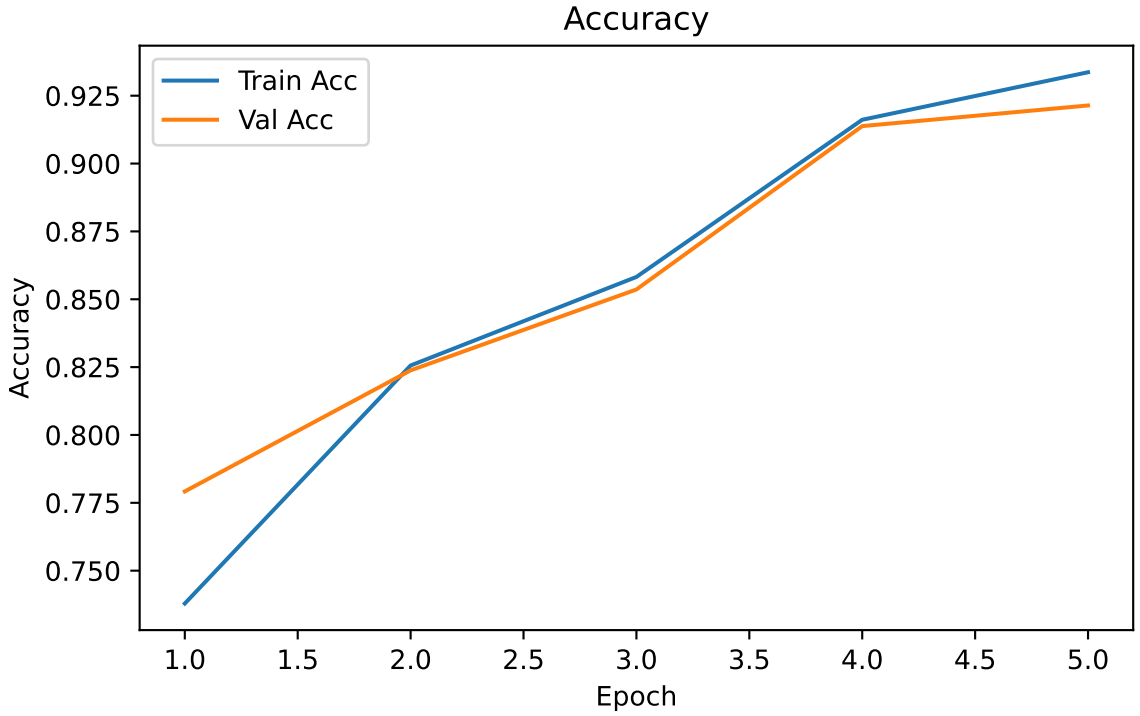
Key Observations:
1. The large gap between clean and adversarial accuracy (77.0 percentage points) indicates that the model is highly vulnerable to adversarial perturbations.

2. Despite the perturbations being imperceptible (ε=8/255≈3.1% 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 α ≈ ε/4 to ε/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
Train Loss
Val Loss
Loss
Epoch
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

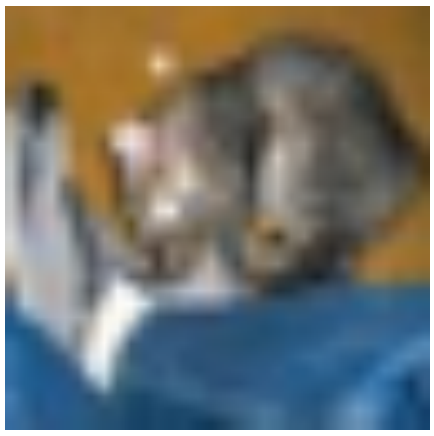

Accuracy
Train Acc
Val Acc
Accuracy
Epoch
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Evaluation Summary

| Metric | Value |
|---|---|
| Clean Accuracy | 0.7700 |
| Adv Accuracy | 0.0000 |
| Attack Success Rate | 1.0000 |

# Adversarial Examples Visualization

Original Image
True: cat
Pred: cat ✓



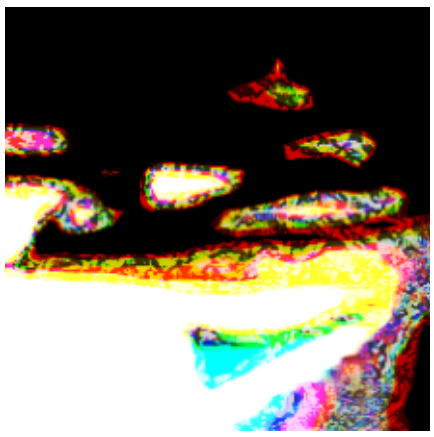
Adversarial Image
Pred: airplane
✓ Attack Success

Perturbation (×10)
Amplified for visibility

Original Image
True: ship
Pred: ship ✓

Adversarial Image
Pred: airplane
✓ Attack Success

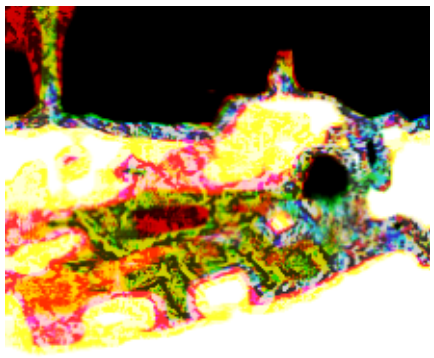
Perturbation (×10)
Amplified for visibility



Original Image
True: ship
Pred: ship ✓

Adversarial Image
Pred: airplane
✓ Attack Success

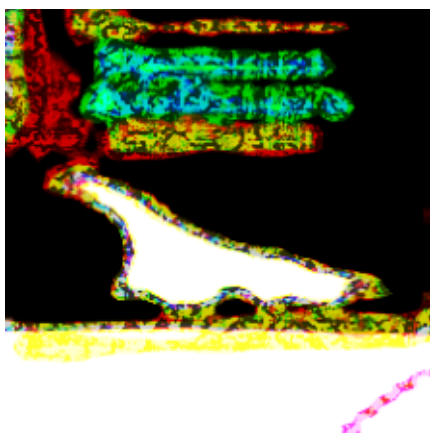
Perturbation (×10)
Amplified for visibility



Original Image
True: airplane
Pred: airplane ✓

Adversarial Image
Pred: bird
✓ Attack Success

Perturbation (×10)
Amplified for visibility



Original Image
True: frog
Pred: frog ✓

Adversarial Image
Pred: bird
✓ Attack Success

Perturbation (×10)
Amplified for visibility