

ResNet-18 CIFAR-10 Adversarial Robustness Evaluation

Experiment Summary

Clean accuracy: 91.34%. Adversarial accuracy under Auto-PGD (eps=0.0314, steps=100): 5.60%. Attack success rate: 94.40% 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: 5.60% - Performance drops dramatically under attack.
- Attack Success Rate: 94.40% - The attack successfully fools the model in most cases.

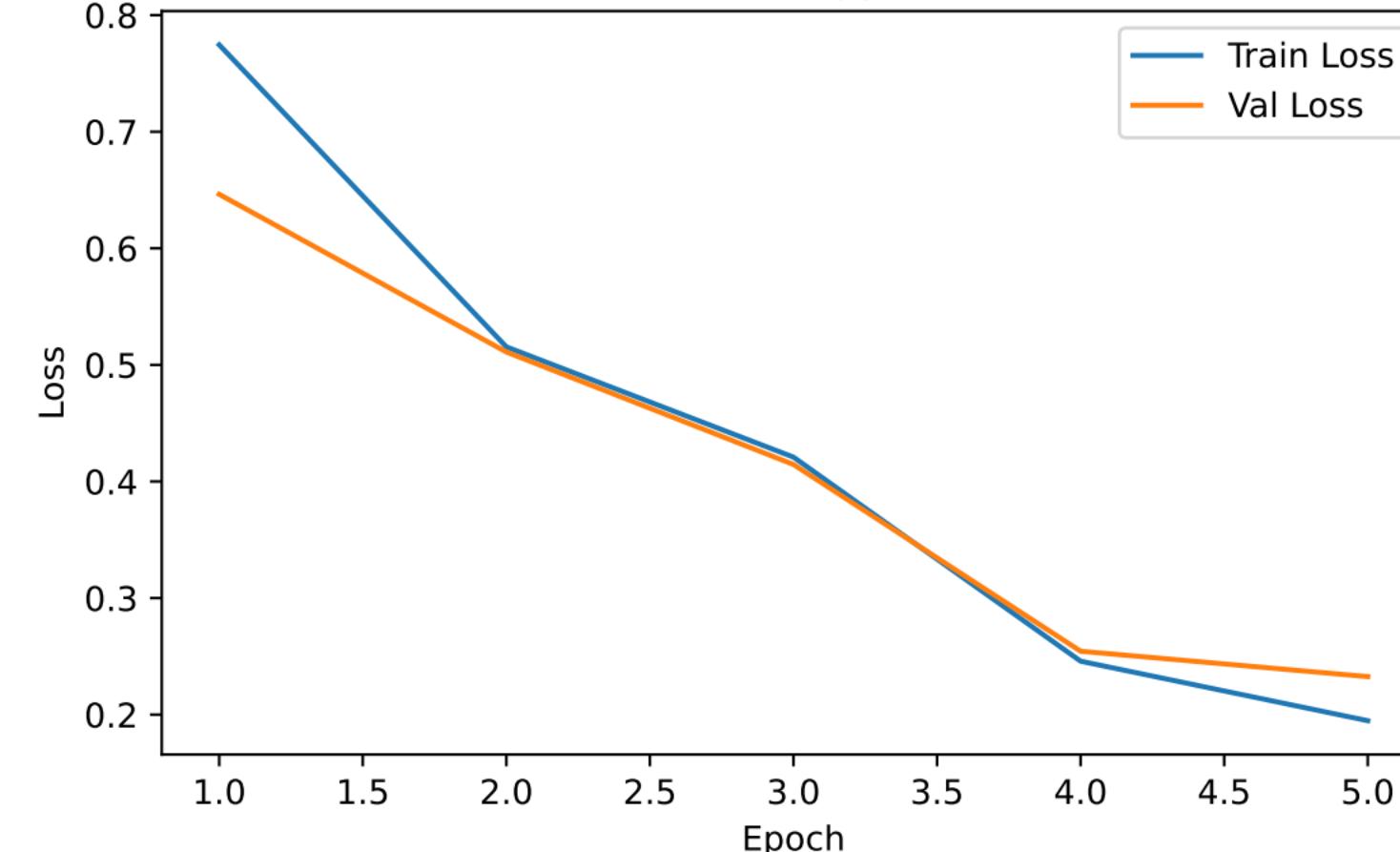
Key Observations:

1. The large gap between clean and adversarial accuracy (85.7 percentage points) indicates that the model is highly vulnerable to adversarial perturbations.
2. Despite the perturbations being imperceptible ($\epsilon=8/255 \approx 3.1\%$ 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.

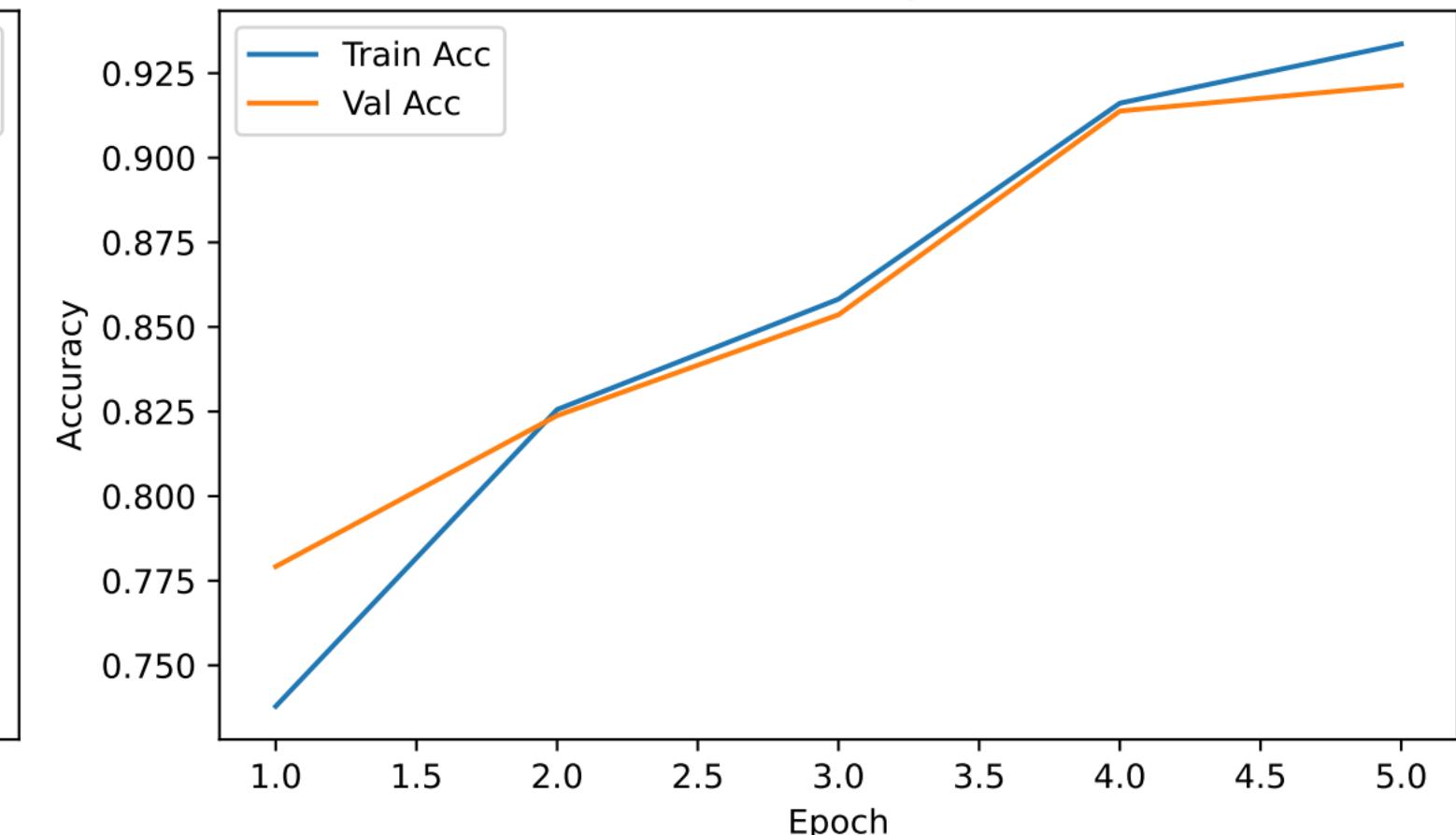
Parameter Impact Analysis:

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

Cross-Entropy Loss



Accuracy



Evaluation Summary

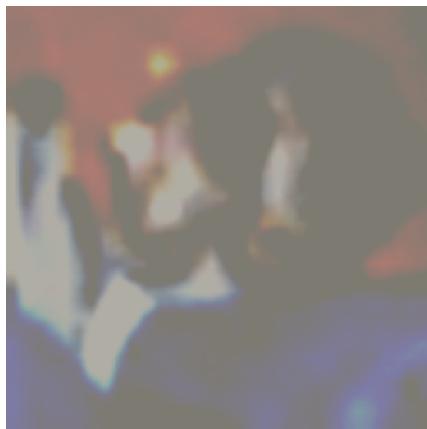
Metric	Value
Clean Accuracy	0.9134
Adv Accuracy	0.0560
Attack Success Rate	0.9440

Adversarial Examples Visualization

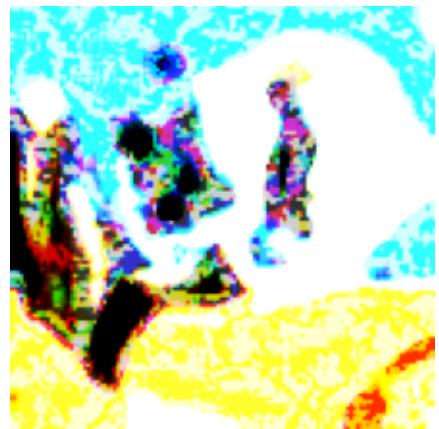
Original Image
True: cat
Pred: cat ✓



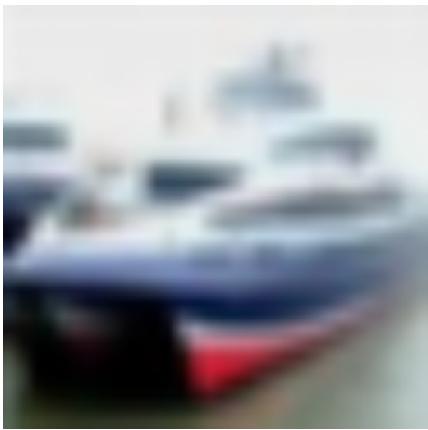
Adversarial Image
Pred: airplane
✓ Attack Success



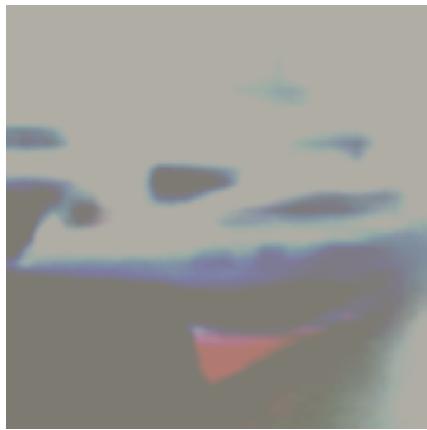
Perturbation ($\times 10$)
Amplified for visibility



Original Image
True: ship
Pred: ship ✓



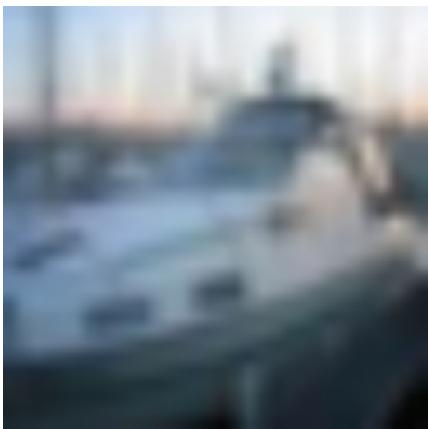
Adversarial Image
Pred: airplane
✓ Attack Success



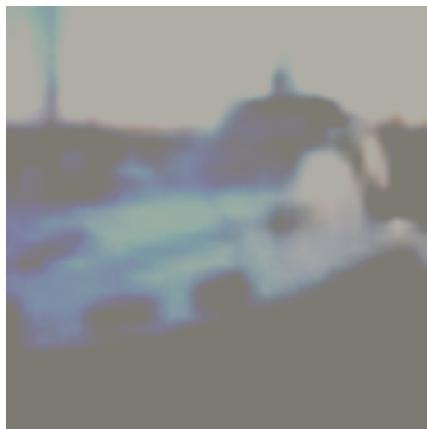
Perturbation ($\times 10$)
Amplified for visibility



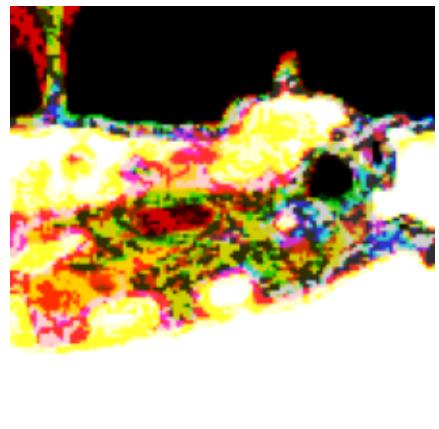
Original Image
True: ship
Pred: ship ✓



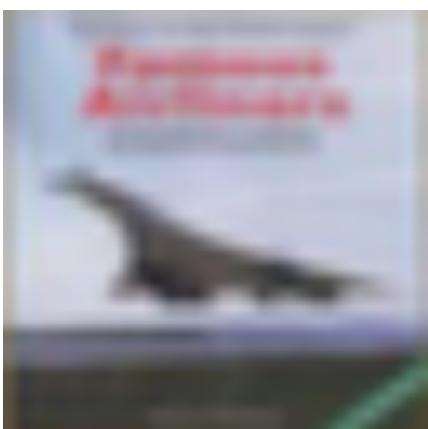
Adversarial Image
Pred: airplane
✓ Attack Success



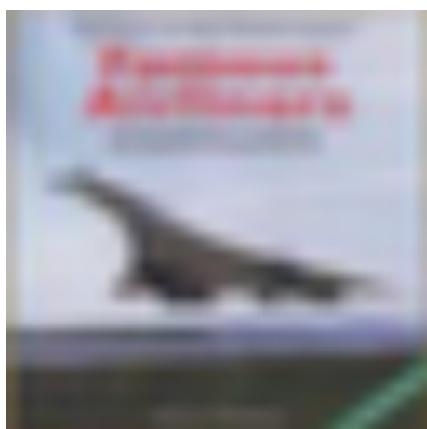
Perturbation ($\times 10$)
Amplified for visibility



Original Image
True: airplane
Pred: airplane ✓



Adversarial Image
Pred: airplane
✗ Attack Failed



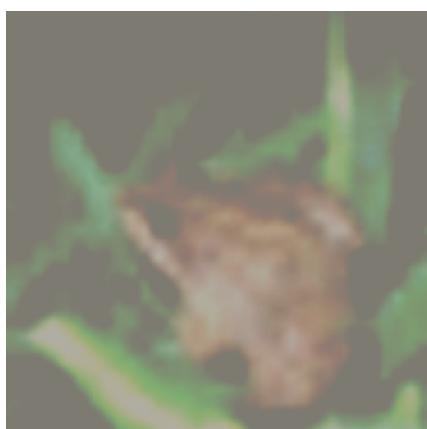
Perturbation ($\times 10$)
Amplified for visibility



Original Image
True: frog
Pred: frog ✓



Adversarial Image
Pred: bird
✓ Attack Success



Perturbation ($\times 10$)
Amplified for visibility

