Supplementary Material

This supplementary material provides additional details on the experimental setup, including the selection of action hyperparameters and qualitative examples of generated adversarial samples.

Action Hyperparameters

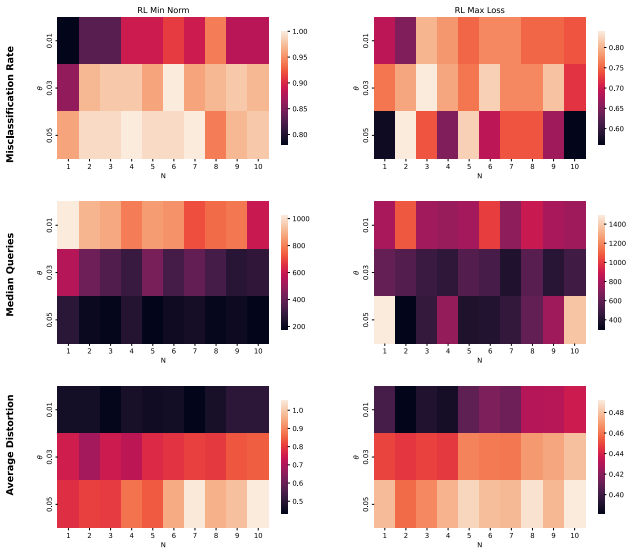

Figure 6. Misclassification Rate, Median Queries, and Average ℓ_2-norm Distortion on adversarial examples post-training for different (N, θ) configurations with RL Min Norm and RL Max Loss attacks on the CIFAR-10 dataset.

In Section 3.1, we define the action a_t as a set of N feature-perturbation pairs, $\{(i_1, \delta_1), ..., (i_N, \delta_N)\}$, where each perturbation δ_j has a maximum magnitude of θ. The choice of N (the number of features to perturb) and θ (the maximum magnitude of that perturbation) balances the trade-off between attack success and the complexity of the action space.

To select appropriate values, we conducted preliminary experiments, the results of which are shown in Figure 6. This figure plots the misclassification rate (ASR), median queries, and average ℓ_2-norm distortion for various (N, θ) configurations for both RL Max Loss and RL Min Norm attacks on CIFAR-10. Based on these results, we fixed $N = 5$ and $\theta = 0.05$ for all main experiments in the paper, as this configuration offered a good balance between attack success and action complexity.

Hyperparameters

Table 3 details the key hyperparameters used for training both the PPO agent and the victim models. The PPO agent's policy and value functions use an `EfficientNet` feature extractor. The victim models were fine-tuned from ImageNet-1K pre-trained weights.

Parameter	Value
PPO Agent (`EfficientNet`)	
Policy/Value Arch.	Linear(128, 64)
Optimizer	Adam
Learning Rate (LR)	2.5e-3
Discount Factor (γ)	0.99
GAE Lambda (λ)	0.95
Clip Range	0.1
Victim Models (`ResNet50, VGG16`)	
Optimizer	SGD
Momentum	0.9
Learning Rate (LR)	0.001
LR Scheduler	ReduceLROnPlateau
Weight Decay	1e-4
Victim Model (`ViT_B_16`)	
Optimizer	AdamW
Learning Rate (LR)	0.001
LR Scheduler	ReduceLROnPlateau
Weight Decay	1e-4

Table 3. Key hyperparameters for the PPO agent and the victim models. For the PPO agent, the learning rate and clip range are linearly annealed from their initial values to 0 over the course of training.

RL Generated Adversarial Samples

To provide a qualitative sense of the attacks, Figure 7 visualizes several adversarial samples generated by our trained RL agents on the CIFAR-10 dataset. Each example shows the original image, the resulting adversarial image, and the imperceptible perturbation (magnified for visibility). The labels demonstrate the agent's success: the victim model's confidence is shifted from the high-confidence original class to a high-confidence incorrect class. This aligns with the overall framework described in Figure 1, where the RL adversary iteratively queries the victim model to produce an adversarial example.

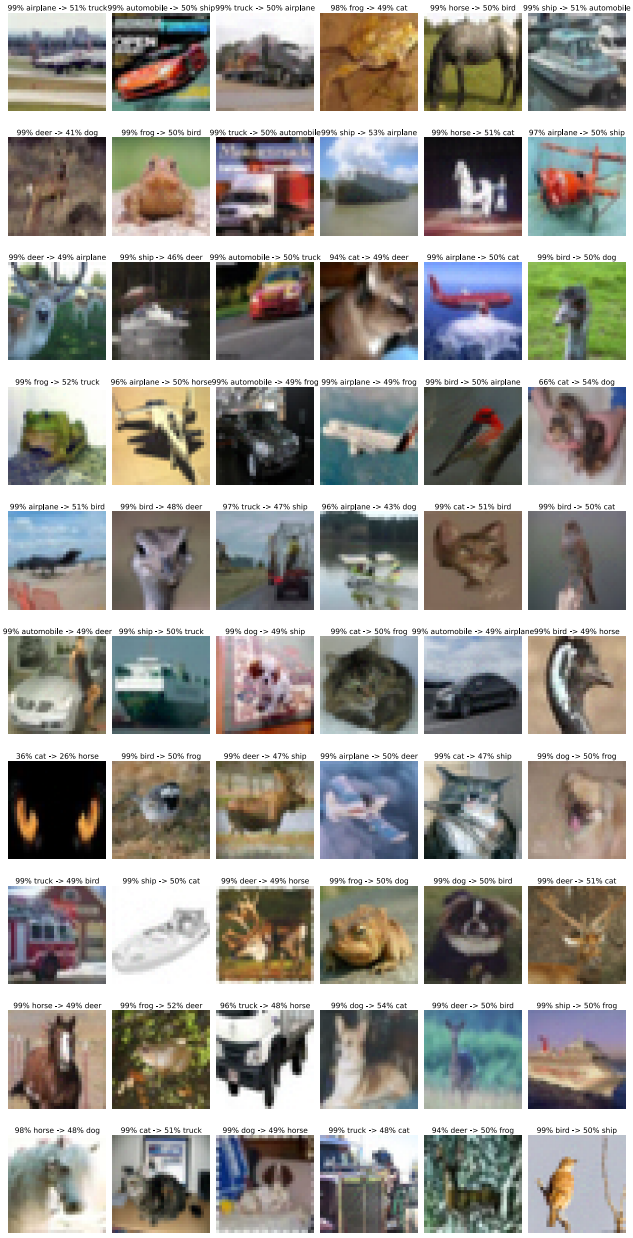

Figure 7. CIFAR-10 adversarial samples generated by black-box RL attacks. Each image contains the confidence on the original class and confidence on the incorrect class.