Adversarial Security Testing: Red Team Assessment of Model Vulnerabilities

Jessica Sihite – 1275588 – COMP90073: Trustworthy Machine Learning

1,096 words

October 16, 2025

Executive Summary

While Blue Team's detection systems filter corrupted and out-of-distribution (OOD) samples from Mosaic.com's training pipeline, a critical vulnerability persists: adversaries can craft product images containing imperceptible perturbations that bypass anomaly and OOD detection while introducing errors. Such attacks represent intentional exploitation where errors accumulate over retraining cycles and progressively compromise classification performance.

To proactively identify these vulnerabilities, we established a Red Team employing Projected Gradient Descent (PGD) attack and evaluated attack effectiveness across varying perturbation magnitudes and optimisation trajectories, measuring success rates and evasion capabilities against each defence mechanism. This offensive security testing quantifies exploitable weaknesses and informs strategic recommendations for enhancing adversarial robustness before real-world exploitation occurs.

Methodology

Data

We used the CIFAR-10 dataset containing 10,000 grayscale images (32x32 pixels) across 10 classes with balanced distribution, representing the types of product images that retailers upload to our platform. The pre-trained classification model achieved 93.32% baseline accuracy on the dataset. From this dataset, we also constructed a test subset by systematically sampling every 10^{th} image, yielding 1,000 samples that preserve the original class distribution with comparable baseline performance (92.1%). This subset was exclusively used for attack generation and defence evaluation.

Methods

Algorithm

We implemented PGD attack using L2 (Euclidean distance) norm constraints to bound perturbation magnitude between original and adversarial images. PGD iteratively computes gradients of the loss function with respect to input pixels, normalises gradients by their L2 norm for consistent step directions, applies small optimisation steps to maximise classification error, and projects perturbations into a constrained L2 ball to maintain visual similarity to the original image. Two attacks were crafted: **untargeted attacks** maximise cross-entropy loss for the true class to force any misclassification; while **targeted attacks** minimise cross-entropy loss toward

Class 5, forcing the model to predict input as Class 5. The attack terminates when objectives are achieved (checked every 20 iterations) or at 100 iterations maximum.

Two parameters govern attack behaviour: step size (**α**), controlling gradient magnitude update, and epsilon (ε), bounding the maximum L2 perturbation. To comprehensively evaluate attack effectiveness, we tested five logarithmically-spaced step sizes (10^{-5} to 10^{1}). For epsilon selection, systematic tuning was performed to balance attack success. Holding step size constant at 0.1, we evaluated ten epsilon candidates on 50 samples of the test subset, measuring attack success rates and evasion rates across Blue Team's defence system. We selected ε = 6.0 for highest OOD evasion rate among attacks meeting the minimum viability thresholds (70% untargeted success and 10% targeted success rates). This selection prioritises evasion of the most robust defence mechanism while ensuring sufficient attack effectiveness.

Evaluation

Attack effectiveness was evaluated on the full 1,000-sample test subset across all step sizes. We measured misclassification rates (success rates) and model accuracy degradation for both untargeted and targeted variants, comparing adversarial accuracy against the 92.1% clean baseline. L2 perturbation norms were computed using explicit sum, power, and square-root operations as specified, with comprehensive statistics reported for each step size to characterise perturbation magnitude distributions.

Convergence behaviour was analysed by tracking iterations required to achieve attack objectives and examining loss trajectories throughout the optimisation process. For each step size, we visualised representative loss convergence patterns (fast-converging, slow-converging, and failed attacks) to identify the relationship between step size, convergence speed, and attack success. Visual analysis included side-by-side comparisons of clean versus adversarial images, perturbation noise patterns, and L2 distribution histograms across all step sizes to assess perturbation characteristics and identify optimal attack configurations.

Blue Team Testing

Generated adversarial samples were tested against Blue Team's defence system: OCSVM (feature-based), VAE (reconstruction-based), and Mahalanobis (distribution-based) detection. We measured evasion rates across all step sizes to assess attack stealth and identify vulnerabilities in the security pipeline.

Results and Discussion

PGD attacks demonstrated high effectiveness against the classification model, with attack success rates increasing monotonically across step sizes for both untargeted and targeted attacks as model accuracy dropped correspondingly (Table 1). While untargeted attacks achieved up to 99.2% success rate at the largest step size, targeted attacks proved

more challenging, reaching 56.9% success at the same step size. This disparity demonstrates that forcing specific misclassification imposes stricter optimisation constraints. Despite varying success rates, perturbation magnitudes remained remarkably consistent across step sizes, as Figure 1 confirms tight concentration around the L2 means at small step sizes, with increased spread at $\alpha = 10^1$. This reveals that within the L2-constrained perturbation space, attack success is primarily defined by optimisation efficiency rather than perturbation magnitude.

Beyond attack success, epsilon selection proved critical for balancing attack effectiveness with defence evasion. Systematic tuning revealed a fundamental trade-off, where smaller values produced smaller attack success but higher possibility to be detected in Blue Team's defence system. We selected $\varepsilon = 6.0$ to maximise OOD evasion among viable attack configurations, prioritising adversarial samples that remain within the in-distribution manifold instead of solely maximising raw attack success. This choice reflects realistic adversarial objectives where detection avoidance is paramount.

The optimisation dynamics underlying these attack success rates are revealed through loss trajectories analysis. Figure 2 shows untargeted attacks exhibit smooth, ascending loss curves as cross-entropy maximisation for the true class follows natural gradient directions away from the correct predictions. Contrastingly, targeted attacks show descending trajectories with greater variability, particularly at larger step sizes where optimisation instability manifests as high-frequency oscillations. At $\alpha = 10^{-5}$ and 10^{-2}, convergence is gradual but reliable; whereas at $\alpha = 10$, successful targeted and untargeted attacks converge rapidly but failed attacks plateau, unable to escape local minima despite aggressive gradient steps.

Algorithm choices shaped attack characteristics. PGD's iterative optimisation enables stronger adversarial examples than single-step methods, evidenced by high success rates across step sizes. L2 constraints were selected over L∞ to minimise perceptual distance, distributing perturbations across entire images rather than creating localised artifacts (Figures 3). This ensures visual imperceptibility. As shown by Figure 4, even at $\alpha = 10^1$ where misclassification rate exceeds 99%, adversarial images appear identical to clean counterparts to human observers, demonstrating neural networks' vulnerability to direction-specific perturbations.

Defence testing exposed critical vulnerabilities with operational implications. VAE consistently achieved the lowest evasion rates that dropped with larger step sizes, confirming that reconstruction-based anomaly detection most robust. OCSVM showed more exploitable non-monotonic behaviour, with evasion rates growing with step sizes before collapsing to 8.5% at extreme perturbations. Most concerningly, Mahalanobis OOD detection's showed inverse robustness as evasion rate spiked to 37% at the largest step sizes.

Conclusion

Red Team successfully identified critical vulnerabilities in Mosaic.com's classification pipeline through systematic adversarial testing.