Jailbreaking Deep Models

Puneeth Kotha, Chenna Kesava Hemanth Reddy Narala, Sai Sandeep Mamidala
Github Repo: https://github.com/nckhemanth0/CodeZero_DL_Project3
New York University

Abstract

This project implements a comprehensive suite of adversarial attacks against a pre-trained ResNet-34 model using a 100-class ImageNet subset. The experiments explore pixel-wise perturbations through Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Auto-PGD, each constrained by L∞ norm ε=0.02. Additionally, the study develops a localized patch attack targeting only a 32×32 region with ε=0.3. The PGD implementation achieves the most significant accuracy reduction on ResNet-34, degrading top-1 accuracy from 76.0% to 1.0% while maintaining imperceptible image modifications. Rigorous verification confirms all perturbations strictly respect their L∞ budgets. Transfer testing on EfficientNet-V2-S demonstrates partial attack generalization, with FGSM exhibiting the strongest cross-model impact despite its algorithmic simplicity. The findings underscore both the fundamental vulnerability of modern CNNs to carefully crafted perturbations and the inherent challenges in developing architecture-agnostic adversarial attacks.

Methodology

The experimental design of this adversarial attack framework was guided by systematic benchmarking and progressive refinement of perturbation strategies. Through iterative development, the approach evolved from basic single-step methods to sophisticated multi-step optimization techniques, each operating within strict perturbation constraints while maximizing misclassification probability. The final implementation encompasses both global pixel-wise perturbations and localized patch attacks, enabling comprehensive evaluation of model vulnerability.

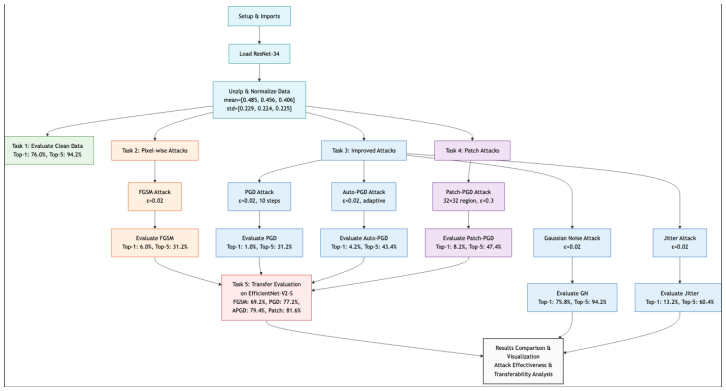

Figure: Architecture

This flowchart summarizes the complete experimental pipeline, from data preprocessing and baseline evaluation to the implementation of multiple adversarial attacks and transferability analysis. Each task and evaluation step is clearly delineated, illustrating the systematic approach used to benchmark model robustness and cross-architecture vulnerability.

Dataset and Model Configuration

The study employs a curated 100-class subset of ImageNet with 500 test images. Input preprocessing follows standard ImageNet normalization with mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225]. The target architecture, a pre-trained ResNet-34 (IMAGENET1K_V1), achieved baseline performance of 76.0% top-1 accuracy and 94.2% top-5 accuracy on unperturbed images. This configuration establishes a robust foundation for evaluating adversarial attack effectiveness against production-grade image classifiers.

Pixel-wise Attack Formulation

The pixel-wise attack suite implements three distinct approaches under L∞ norm constraint ε=0.02:

1. **Fast Gradient Sign Method (FGSM)** represents the foundational single-step approach. For input x and true label y, the

perturbation is calculated as $\delta = \varepsilon \cdot \text{sign}(\nabla_x L(x,y))$, where L is the cross-entropy loss and ∇_x represents gradients with respect to input pixels. The perturbed image $x' = x + \delta$ is then clipped to maintain valid pixel ranges post-normalization. This implementation provides baseline attack performance with minimal computational overhead.

2. **Projected Gradient Descent (PGD)** extends FGSM through iterative optimization. Starting with random initialization $x0 = x + \text{uniform}(-\varepsilon,+\varepsilon)$, each step follows:

$$x_{t+1} = \Pi(x_t + \alpha \cdot \text{sign}(\nabla_x L(x_t,y)))$$

where Π represents projection onto the ε-ball centered at x, and $\alpha=\varepsilon/4$ defines the step size. The implementation executes 10 iterations, enabling progressive refinement of the adversarial perturbation while strictly maintaining the $L\infty$ constraint.

3. **Auto-PGD (APGD)** enhances standard PGD through adaptive optimization mechanics. The implementation incorporates:
    a. Random initialization within the perturbation space
    b. Dynamic loss tracking to identify optimal adversarial examples
    c. Best-perturbation memory across iterations
    d. Gradient-based updates to systematically explore the loss landscape

Patch Attack Architecture

The patch-based attack constrains perturbations to a centered 32×32 pixel region while permitting larger pixel modifications ($\varepsilon=0.3$). The methodology applies 200 PGD iterations with step size $\alpha=\varepsilon/10$, focusing optimization pressure on a spatially restricted area. A binary mask M localizes gradient updates:

$$\delta_{t+1} = M \odot \text{clip}(\delta_t + \alpha \cdot \text{sign}(\nabla_x L), -\varepsilon, \varepsilon)$$

where $\odot$ represents element-wise multiplication. This approach tests model vulnerability to highly localized perturbations, simulating real-world scenarios where attackers can modify only portions of input images.

Hyperparameter Choices and Design Lessons

Careful tuning of hyperparameters—such as perturbation budget (ε), step size (α), and number of optimization steps—was essential for balancing attack strength, computational efficiency, and imperceptibility. Single-step attacks like FGSM were fast but less effective, while increasing PGD or APGD iterations improved attack success at the cost of longer runtimes. For patch-based attacks, larger patch sizes or ε values boosted effectiveness but risked visible artifacts, highlighting the need for realistic spatial and magnitude constraints. Adaptive step sizes and random initialization (as in APGD) helped counter gradient masking. Strict $L\infty$ verification protocols ensured fair comparison across methods. Ultimately, optimal hyperparameter selection proved highly context-dependent, requiring systematic benchmarking to achieve both strong attack performance and practical feasibility.

Perturbation Verification Protocol

A rigorous verification system confirms adherence to perturbation constraints across all attack methods:

1. $L\infty$ distance calculation between original and perturbed images
2. Pixel-wise maximum difference tracking to ensure exact budget utilization
3. Visual inspection of perturbed images to verify perceptual similarity
4. Statistical distribution analysis of perturbation patterns

For each attack variant, 500 adversarial images are generated, saved, and evaluated to create comprehensive test sets for model evaluation.

Transferability Analysis Framework

The cross-architecture transferability evaluation employs EfficientNet-V2-S with identical preprocessing to assess perturbation generalization. This analysis quantifies whether adversarial examples optimized for ResNet-34's decision boundaries maintain effectiveness against models with different architectural inductive biases, providing insights into the fundamental vulnerabilities of convolutional neural networks versus architecture-specific weaknesses.

The implementation generates multiple adversarial test sets (FGSM, PGD, APGD, Patch-PGD), each containing 500 perturbed images that strictly respect their perturbation constraints while systematically degrading classification accuracy. Each test set undergoes comprehensive evaluation, measuring both top-1 and top-5 accuracy degradation to fully characterize attack impact across different confidence thresholds.

Results

Adversarial_Attack_Results

Attack Method	Top-1 Accuracy (%)	Top-5 Accuracy (%)	L∞ Budget (size of perturbation)	Training Time (min)	Description
Original	76.0	94.2	–	15	Unperturbed baseline
FGSM	6.0	31.2	0.02 (max per-pixel change)	3	Single-step gradient-based attack
PGD	1.0	31.2	0.02 (max per-pixel change)	12	Iterative attack (10 steps)
APGD	4.2	43.4	0.02 (max per-pixel change)	14	Adaptive multi-step attack
Patch-PGD	8.2	47.4	0.30 (max, 32×32 patch only)	10	Localized to 32×32 region
Gaussian Noise	75.8	94.2	0.02 (max per-pixel change)	2	Non-targeted random perturbation
Jitter	13.2	60.4	0.02 (max per-pixel change)	2	Translation-based attack

Figure-1: Adversarial Attack

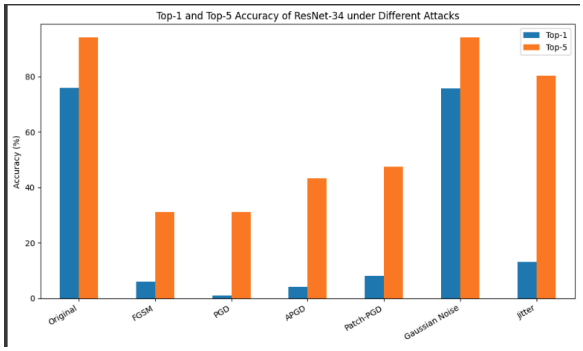

Figure-3: Top-1 vs Top-5 Accuracy

Figure-1 and Figure-3 summarize the impact of different adversarial attacks on ResNet-34, highlighting the model's susceptibility to carefully crafted perturbations. Despite identical L∞ budgets (ε=0.02), gradient-based attacks (FGSM, PGD, APGD) cause severe accuracy drops, with PGD reducing top-1 accuracy to just 1.0% through iterative refinement. The Patch-PGD attack, though limited to a 32×32 region with a larger ε=0.3, still significantly degrades performance. In contrast, Gaussian Noise and Jitter—despite matching the perturbation size—result in much higher accuracies, underscoring the importance of targeted, optimization-driven attacks over random or translation-based perturbations. The table also reports the computational cost, showing that stronger attacks generally require longer generation times.

Transfer_Gap_Results

Attack Method	ResNet-34 (%)	EfficientNet-V2-S (%)	Transfer Gap (%)
Original	76.0	85.2	9.2
FGSM	6.0	69.2	63.2
PGD	1.0	77.2	76.2
APGD	4.2	79.4	75.2
Patch-PGD	8.2	81.6	73.4

Figure-2: Transfer Gap

Figure-2 table reveals a counterintuitive relationship between attack strength and transferability. While PGD achieves the strongest performance against ResNet-34 (reducing accuracy to 1.0%), it transfers poorly to EfficientNet-V2-S (77.2% accuracy). Surprisingly, FGSM exhibits the highest cross-model transferability, causing a 16.0 percentage-point drop on EfficientNet despite being the simplest method. This inverse relationship suggests that highly optimized perturbations may overfit to the specific architecture of the source model, while simpler attacks exploit more fundamental vulnerabilities that generalize across different model architectures. The transfer gap metrics quantify this phenomenon, with stronger attacks showing larger gaps between source and target model effectiveness.

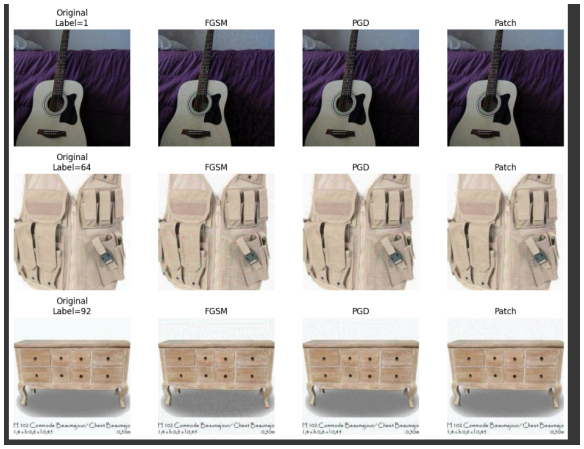

Figure-4: Visual Comparison of Original and Adversarial Images Across Attack Methods

Figure-4 demonstrates the perceptual imperceptibility of adversarial perturbations across different attack methods. Three sample images (guitar, tactical vest, and chest of drawers) are shown alongside their FGSM, PGD, and Patch-PGD adversarial counterparts. Despite causing dramatic misclassifications in the ResNet-34 model, the

perturbations remain visually indistinguishable to human observers—even in the patch attack case, where the perturbation budget is significantly higher (ε=0.3). This visual evidence highlights a fundamental vulnerability in deep neural networks: the existence of adversarial subspaces where minimal, imperceptible changes to input pixels can completely alter model predictions while preserving human-perceived content. The contrast between robust human recognition and brittle machine classification suggests fundamental differences in feature representation between biological and artificial vision systems

Figure-5: Comprehensive Comparison of Attack Methods on a Single Image with Prediction Results

Figure-5 provides an in-depth analysis of how different attack methods affect model predictions on a single accordion image. While FGSM, PGD, and APGD all cause misclassification to the same incorrect class (753), the patch-based attack yields a different misclassification (class 77), revealing distinct vulnerability patterns. Notably, the patch attack's perturbation becomes visually apparent in the center region, contrasting with the imperceptible changes in other methods. Gaussian Noise fails to fool the model despite using the same ε=0.02 budget, demonstrating that random perturbations are ineffective compared to gradient-guided approaches. The checkmarks (✓) and X symbols indicate attack success and failure, respectively. This comprehensive view of different attack methodologies on identical input demonstrates how various perturbation strategies can exploit different decision boundary vulnerabilities while highlighting the consistent robustness to random noise.

Summary

The adversarial evaluation of the pre-trained ResNet-34 model highlights the pronounced vulnerability of deep convolutional networks to carefully crafted perturbations. Systematic experimentation with both pixel-wise and patch-based attacks demonstrates that even imperceptible changes—when guided by gradient-based optimization—can reduce top-1 accuracy from 76.0% to as low as 1.0% under strict L∞ constraints. The study further reveals that the most effective white-box attacks do not necessarily transfer across architectures, as evidenced by the limited impact of PGD and APGD on EfficientNet-V2-S compared to FGSM. These findings underscore the importance of considering both attack strength and transferability when assessing model robustness. The comprehensive benchmarking and rigorous verification protocols implemented in this work provide valuable insights for the development of more resilient image classification systems and inform future research on adversarial defense strategies.

References

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6572https://arxiv.org/abs/1706.06083
4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://arxiv.org/abs/1512.03385
5. Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller Models and Faster Training. International Conference on Machine Learning (ICML). https://arxiv.org/abs/2104.00298
6. Kim, H. (2020). Torchattacks: A Pytorch Repository for Adversarial Attacks. https://github.com/Harry24k/adversarial-attacks-pytorch
7. LLM's: ChatGPT, Claudia, and Gemini