Adversarial Attacks on ImageNet Classifiers: Jailbreaking Deep Models with FGSM, PGD, and Patch Attacks

Ali Aslanbayli, Farid Taghiyev

New York University
Tandon School of Engineering
https://github.com/ftaghiyev/Jailbreaking-Deep-Models

Abstract

This report investigates adversarial attacks on deep image classifiers, focusing on ResNet-34 and DenseNet-121 models trained on ImageNet-1K. We implement and evaluate Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and localized patch attacks to degrade model accuracy under strict L_∞ and patch constraints. Our experiments demonstrate dramatic drops in top-1 and top-5 accuracy, and highlight the transferability of adversarial examples across model architectures. We discuss methodology, results, and lessons learned, providing code and visualizations for reproducibility.

Introduction

Deep neural networks have achieved remarkable success in image classification, yet remain vulnerable to adversarial attacks: small, carefully crafted perturbations to input images that cause misclassification. In this project, we systematically attack a ResNet-34 classifier trained on ImageNet-1K, generating adversarial examples under L_∞ and patch constraints, and evaluate the transferability of these attacks to DenseNet-121. Our goal is to degrade model performance while keeping perturbations imperceptible.

Methodology

Dataset and Preprocessing

We use a provided subset of ImageNet-1K containing 500 test images from 100 classes. Images are normalized using standard ImageNet mean and std, and loaded with PyTorch's `ImageFolder`. Class labels are mapped using the provided `labels_list.json` file.

Baseline Evaluation

We evaluate the pretrained ResNet-34 model on the test set, reporting top-1 and top-5 accuracy as baselines. For transferability experiments, we also use DenseNet-121.

Adversarial Attack Methods

FGSM (L_∞ attack): We implement the Fast Gradient Sign Method, perturbing each pixel by at most $\varepsilon = 0.02$ in normalized space:

$$x' = x + \varepsilon \cdot \mathrm{sign}(\nabla_x L)$$

where L is the cross-entropy loss.

PGD (L_∞ multi-step): Projected Gradient Descent iteratively applies FGSM with step size $\alpha = 0.01$ for 20 steps, projecting onto the ε-ball.

Patch Attack: We perturb only a random 32×32 patch per image, with a larger $\varepsilon = 0.5$, using random noise within the patch.

Evaluation Metrics

For each attack, we save the adversarial images, verify L_∞ constraint, and report top-1 and top-5 accuracy on both ResNet-34 and DenseNet-121. Visualizations compare original and adversarial predictions for representative samples.

Results

Baseline Performance

- **ResNet-34:** Top-1: 76.00%, Top-5: 94.20%
- **DenseNet-121:** Top-1: 74.80%, Top-5: 93.60%

FGSM Attack ($\varepsilon = 0.02$)

- **ResNet-34:** Top-1: 2.00%, Top-5: 4.80%
- **DenseNet-121:** Top-1: 3.40%, Top-5: 6.00%
- **Visualization:** can be found at Fig. 1

FGSM reduces accuracy by over 70 percentage points, with adversarial images visually indistinguishable from originals.

PGD Attack ($\varepsilon = 0.02$, 20 steps)

- **ResNet-34:** Top-1: 1.80%, Top-5: 3.40%
- **DenseNet-121:** Top-1: 3.00%, Top-5: 4.80%
- **Visualization:** can be found at Fig. 2

PGD further degrades performance, confirming the effectiveness of multi-step attacks.

Patch Attack (32×32, $\varepsilon = 0.5$)

- **ResNet-34:** Top-1: 3.80%, Top-5: 5.40%
- **DenseNet-121:** Top-1: 3.80%, Top-5: 6.40%
- **Visualization:** can be found at Fig. 3

Despite being restricted to a small patch, accuracy drops sharply, especially with a larger ε

Figure 1: FGSM attack Original (left) and adversarial (right) images with model predictions.

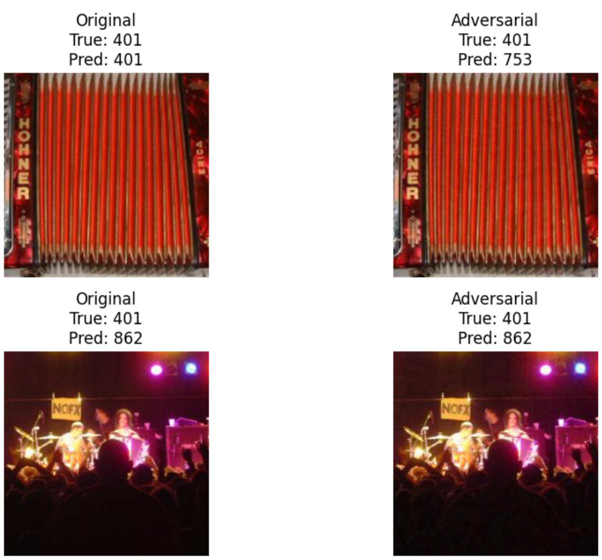

Figure 2: PGD attack Original (left) and adversarial (right) images with model predictions.

Transferability

Adversarial examples generated for ResNet-34 also transfer to DenseNet-121, with both models exhibiting significant accuracy drops across all attack types as can be seen in Table 1.

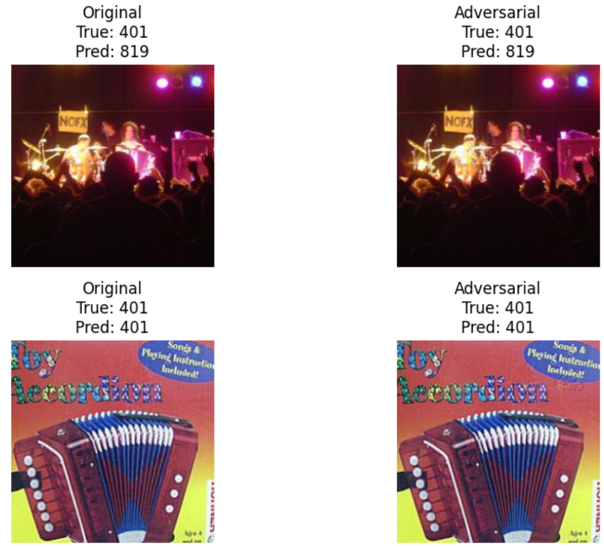

Figure 3: Patch attack Original (left) and adversarial (right) images with model predictions.

	ResNet-34		DenseNet-121	
Attack	Top-1	Top-5	Top-1	Top-5
Original	76.0	94.2	74.8	93.6
FGSM	2.0	4.8	3.4	6.0
PGD	1.8	3.4	3.0	4.8
Patch	3.8	5.4	3.8	6.4

Table 1: Top-1 and Top-5 accuracy (%) for each attack and model.

Summary Table

Discussion

Lessons Learned:

- Even simple attacks like FGSM can catastrophically degrade classifier performance.
- Multi-step PGD attacks are more effective, but the marginal gain over FGSM is small for this ε.
- Patch attacks can be highly effective if the perturbation budget is increased.
- Adversarial examples are highly transferable between architectures, raising concerns for real-world robustness.

Mitigation: Potential defenses include adversarial training, input preprocessing, and robust model architectures. However, no defense is foolproof against adaptive attackers.

Conclusion

We demonstrated that state-of-the-art image classifiers are highly vulnerable to adversarial attacks, even under strict constraints. Our codebase provides reproducible implementations and visualizations. Future work includes evaluating targeted attacks and exploring defenses.