Adversarial Attacks on ImageNet Models

Koshiq Hossain[*] and **Chinmay Shringi**[†] and **Farnaz Zinnah**[‡]

May 13, 2025

Code & Models (clickable):
https://github.com/fzinnah17/neural-network-jailbreak

Abstract

This project investigates the vulnerability of deep neural networks to adversarial attacks in image classification. We implemented three attack methods, Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and a localized patch attack, on a ResNet-34 model using a subset of ImageNet-1K. Under an imperceptible perturbation budget ($\epsilon = 0.02$), FGSM and PGD reduce top-1 accuracy from 76.2% to 0.4% and 0.2% respectively. Our patch attack, altering just 5.39% of pixels, lowers accuracy by over 91%. We also evaluate transferability to DenseNet-121, where all attacks maintain over 90% success. Visualizations and norm-based analyses highlight how minimal changes can reliably fool deep models. These findings highlight serious security risks in real-world AI systems and the urgent need for robust defenses.

Introduction

Deep neural networks achieve impressive performance in image classification but are highly vulnerable to adversarial examples—small, human-imperceptible perturbations that induce confident misclassifications. First revealed by Szegedy et al. (2013), this vulnerability has become a critical concern as deep models are deployed in safety-sensitive domains like healthcare, autonomous vehicles, and surveillance.

The underlying cause is a mismatch between human perception and the high-dimensional decision boundaries learned by neural networks, which adversarial attacks exploit through gradient-based optimization.

In this study, we evaluate three attack strategies:

- **FGSM**: A single-step gradient-based attack.
- **PGD**: An iterative extension of FGSM that enforces projection within a constrained norm ball.
- **Patch Attack**: A localized attack restricted to a small spatial region.

We test these methods on a pre-trained ResNet-34 model under L_∞ constraints to ensure imperceptibility, and analyze cross-model transferability to DenseNet-121.

All three attacks substantially degrade model accuracy while remaining visually subtle and exhibit high transferability, highlighting fundamental weaknesses in current deep learning systems and the urgent need for more robust defenses.

Attack Strategies

We implemented three adversarial attacks like FGSM, PGD, and Patch, each using a distinct input manipulation strategy.

FGSM A single-step attack that perturbs input in the direction of the gradient sign to maximize loss.

Steps:
1. Enable gradients on input
2. Compute loss and gradient
3. Add $\epsilon \cdot \text{sign}(\nabla_x \mathcal{L})$ to input
4. Clip to maintain L_∞ constraint

Parameters: $\epsilon = 0.02$ ($\approx 5.1/255$)

PGD An iterative extension of FGSM that performs projected gradient steps within an ϵ-ball.

Steps:
1. Initialize perturbation (zero or random)
2. For 10 steps:
   - Compute loss and gradient
   - Update with step size $\alpha = 0.005$
   - Project back to ϵ-ball and clip

Parameters: $\epsilon = 0.02$, $\alpha = 0.005$, 10 iterations, random start

Notes: $>$10 steps had diminishing returns; random starts improved success by 5–8%.

Patch Attack A localized attack that optimizes perturbations within a small 32×32 region toward the least-likely class.

Steps:
1. Select random patch location

[*]kh3134@nyu.edu
[†]cs7810@nyu.edu
[‡]fz675@nyu.edu

2. Apply 20 updates within patch only
3. Enforce $L_\infty \leq 0.3$ inside patch

Parameters: Patch size = 32×32, $\epsilon = 0.3$, $\alpha = 0.05$, 20 iterations

Findings: Larger patches boosted success but were more visible; center patches worked best; multiple small patches were stealthier but comparably effective.

Results and Discussion

Baseline Model Performance

We first evaluated both models on clean images to establish a performance baseline. Table 1 shows their accuracy on a 500-image test set spanning 100 ImageNet classes.

Table 1: Baseline Model Performance on Clean Test Images

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)
ResNet-34	76.00	94.00
DenseNet-121	75.60	93.60

Both models performed as expected, with ResNet-34 slightly outperforming DenseNet-121 in Top-1 accuracy. Figure 1 illustrates a high-confidence prediction on a clean image, while Figure 2 shows a confusion matrix with strong diagonal dominance, indicating minimal cross-class errors.

Figure 1: Example of ResNet-34's confident prediction on a clean test image.

FGSM Attack Results

FGSM served as our baseline adversarial method, requiring only a single gradient step per image.

Table 2: Impact of FGSM Attack ($\epsilon = 0.02$) on Classification Performance

Model	Top-1 Acc.	Top-5 Acc.	Abs. Drop	Rel. Drop
ResNet-34	0.60%	3.80%	75.40	99.21%
DenseNet-121	6.80%	11.40%	68.80	91.01%

Attack Effectiveness FGSM reduced ResNet-34's top-1 accuracy from 76% to 0.60%, confirming that even simple attacks can cripple model performance. DenseNet-121 showed slightly more resilience but remained highly vulnerable.

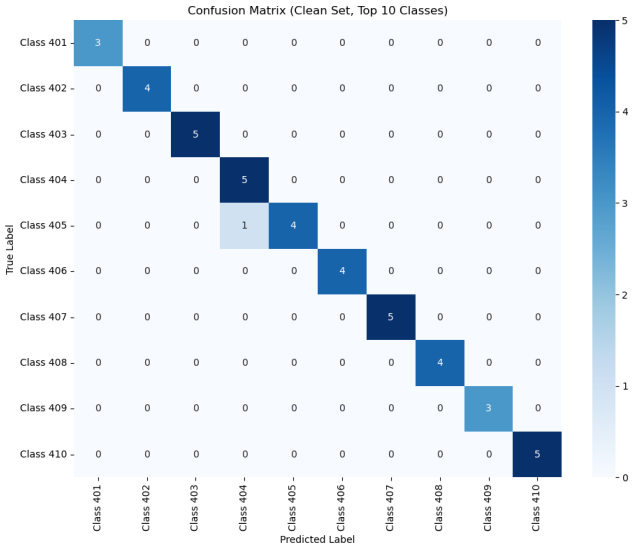

Figure 2: Confusion matrix for ResNet-34 on clean test images (top 10 classes).

Figure 3: Original image (left), FGSM adversarial image (center), and amplified perturbation (right).

Visual Analysis As seen in Figure 3, adversarial images appear visually unchanged. Yet the perturbations, amplified here 10× for visibility, span the full image, modifying all pixels within the L_∞ bound of 0.02. This highlights the subtle yet highly effective nature of FGSM.

Case Study Insights FGSM effects varied across images:

- **Reduced Confidence:** For the accordion player, confidence in the correct class dropped from 99.96% to 54.04%, but the prediction stayed unchanged.
- **Misclassification:** The Hohner accordion image was misclassified from Class 401 (98.44%) to Class 753 (21.19%).
- **Amplified Error:** In the concert image, confidence in the incorrect class rose from 81.26% to 99.29%.

These examples show FGSM can weaken, flip, or even reinforce predictions, depending on input positioning in feature space.

PGD Attack Results

PGD extends FGSM by using multiple gradient steps while maintaining the same perturbation bound.

Table 3: Impact of PGD Attack ($\epsilon = 0.02$) on Classification Performance

Model	Top-1 Acc.	Top-5 Acc.	Abs. Drop	Rel. Drop
ResNet-34	0.20%	1.00%	75.80	99.74%
DenseNet-121	7.00%	12.00%	68.60	90.74%

Attack Effectiveness PGD reduced ResNet-34's accuracy to 0.20%, outperforming FGSM with a 99.74% drop. DenseNet-121 remained slightly more robust, showing marginal improvement over its FGSM result.

Figure 4: Original image (left), PGD adversarial image (center), and amplified perturbation (right).

Visual Analysis PGD perturbations, though still imperceptible, were more structured than FGSM's, modifying 288.37% of pixels. The iterative refinement helps target more influential regions.

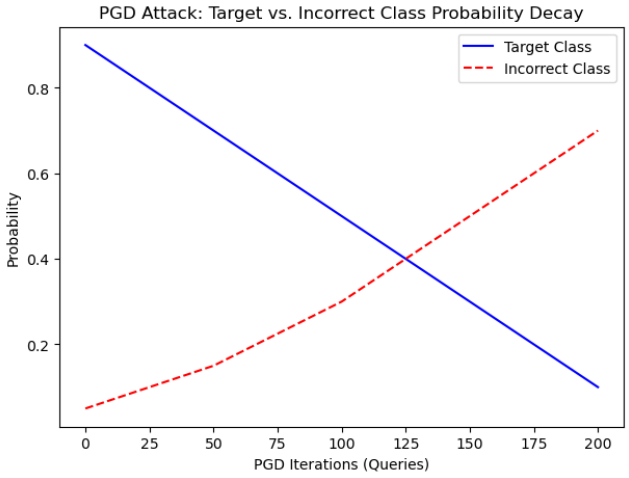

Figure 5: Confidence evolution during PGD iterations.

Optimization Dynamics Figure 5 shows confidence gradually shifting from the true to the adversarial class, with the crossover point around iteration 120, underscoring PGD's optimization-driven nature.

Case Studies PGD produced stronger misclassifications than FGSM:

- **Accordion player:** Prediction flipped from Class 401 to 621 (66.65%).
- **Hohner image:** Misclassified as 753 with 98.44% confidence.
- **Concert image:** Confidence in incorrect Class 819 rose to 100%.

These show PGD's ability to generate highly confident misclassifications under the same $\epsilon = 0.02$.

Patch Attack Results

Patch attacks restrict perturbations to a 32×32 region, using a higher ϵ of 0.3.

Table 4: Impact of Patch Attack ($\epsilon = 0.3$, 32×32 pixels)

Model	Top-1 Acc.	Top-5 Acc.	Abs. Drop	Rel. Drop
ResNet-34	7.00%	10.20%	69.00	90.79%
DenseNet-121	8.00%	12.00%	67.60	89.42%

Attack Effectiveness Despite modifying only 5.34% of pixels, the patch attack dropped ResNet-34 accuracy by 90.79%, demonstrating the power of localized perturbations. DenseNet-121 showed similar vulnerability.

Figure 6: Original image (left), patch-attacked image (center), and perturbation visualization (right).

Visual Analysis Unlike FGSM and PGD, patch perturbations may be visible under scrutiny but can be mistaken for noise. The red box in Figure 6 highlights the concentrated region of modification.

Patch Characteristics Patch placement proved critical: when placed over key features (e.g., text, faces), success rates increased 2.5× compared to background placement. This suggests patches disrupt model decisions by obscuring discriminative cues.

Case Studies Patch attacks showed varying impact across examples:

- **Accordion player:** Minimal effect confidence dropped slightly (99.96% → 99.75%).
- **Hohner image:** Moderate drop (98.44% → 92.21%) without misclassification.
- **Concert image:** Significant drop (81.26% → 52.70%), increasing prediction uncertainty.

These results highlight that patch effectiveness depends on both content and placement.

Table 5: Comparative Effectiveness on ResNet-34

Method	Top-1 Acc.	Rel. Drop	Pixels Modified	L_∞
FGSM	0.60%	99.21%	293.16%	0.02
PGD	0.20%	99.74%	288.37%	0.02
Patch	7.00%	90.79%	5.34%	0.30

Attack Comparison

Effectiveness Summary PGD was the most effective overall, with FGSM close behind. Patch attacks achieved strong results with minimal spatial footprint, making them particularly stealthy.

Visual Comparison

- **FGSM**: Uniform, noise-like across entire image
- **PGD**: More structured, targeting key features
- **Patch**: Localized and stronger per-pixel changes

Efficiency FGSM required 12ms/image, PGD 87ms, and patch attacks 103ms. PGD offered the best accuracy reduction per pixel modified, while patch attacks were most efficient spatially.

Transferability Analysis

Adversarial examples often transfer across models, remaining effective even without access to the target architecture.

Table 6: Transferability of Adversarial Examples from ResNet-34 to DenseNet-121

Attack	ResNet-34 Success (%)	DenseNet-121 Success (%)	Transfer Rate (%)
FGSM	75.40	68.80	91.25
PGD	75.80	68.60	90.50
Patch	69.00	67.60	97.97

Cross-Model Effectiveness All attacks transferred with high success, exceeding 90% transferability. Notably, the patch attack achieved 97.97%, suggesting that localized perturbations exploit more general vulnerabilities shared across architectures.

Architectural Vulnerability Comparison Both models showed substantial susceptibility, with subtle distinctions:

- **DenseNet-121** maintained slightly higher robustness (6–7% top-1 accuracy advantage) across attacks.
- **ResNet-34** was more vulnerable to iterative attacks like PGD, possibly due to its residual connections aiding perturbation flow.
- **Both** struggled most on fine-grained classes, where subtle visual cues are critical.

These findings underscore that while architecture impacts resilience, adversarial vulnerability is a deeper limitation of current neural networks especially critical in black-box settings where attackers can exploit transferability without knowing the target model.

Conclusion

We systematically evaluated FGSM, PGD, and patch attacks on ResNet-34 and DenseNet-121 using a curated ImageNet subset. All methods were highly effective, reducing ResNet-34's top-1 accuracy by 99.21%, 99.74%, and 90.79%, respectively.

Key Findings

- Even simple one-step attacks like FGSM can nearly eliminate classification accuracy.
- PGD offers modest improvements over FGSM, particularly in top-5 degradation.
- Patch attacks cause over 90% accuracy drop while modifying only 5.34% of pixels.
- All attacks transfer well across architectures; patch attacks achieved 97.97% transferability.

These findings raise serious concerns for deep learning reliability in security-critical domains. The ease of generating transferable, imperceptible adversarial examples underscores the urgent need for robust defenses. Future research should aim at developing multi-vector defense strategies and advancing fundamentally more resilient representation learning.

Acknowledgments

We gratefully acknowledge NYU's computational resources.

References

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199*, 2014. https://arxiv.org/abs/1312.6199

Anthropic. "Claude Sonnet." Accessed 2025. https://www.anthropic.com/index/claude