Jailbreaking Deep Models: Adversarial Attacks on ImageNet Classifiers

Raghu Vamsi Hemadri[1] **, Geetha Krishna Guruju**[2] **, Sai Subhash Kotaru**[3]

New York University, New York, New York
rh3884@nyu.edu[1], gg3039@nyu.edu[2], sk12154@nyu.edu[3]

Abstract

This project investigates the susceptibility of state-of-the-art image classifiers to adversarial attacks, focusing on a pretrained ResNet-34 model evaluated on a curated subset of the ImageNet-1K dataset. We implement and analyze three adversarial strategies: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and a spatially localized patch-based PGD attack. Each method operates under an ℓ_∞ constraint to ensure perturbation imperceptibility or locality.

Our results show that FGSM reduces Top-1 accuracy from 89.60% to 8.60%, while PGD completely degrades performance (0.00% Top-1). Surprisingly, our patch-based PGD attack—despite altering only a 32×32 region—achieves comparable degradation, reducing Top-1 accuracy to 0.20%. Transferability tests reveal that these adversarial examples also affect DenseNet-121, highlighting their cross-architecture impact.

Ablation studies demonstrate the importance of step size, restarts, and attack targeting in maximizing success. These findings underscore the urgency of developing robust defenses against both global and localized adversarial threats in real-world vision systems.

Introduction

Deep neural networks have achieved remarkable success in image classification tasks, particularly on large-scale datasets like ImageNet-1K (Krizhevsky, Sutskever, and Hinton 2012). However, recent studies have demonstrated that these models are highly susceptible to adversarial attacks—small, carefully crafted perturbations that can drastically reduce model performance while remaining nearly imperceptible to the human eye (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014). This fragility poses serious concerns for deploying deep models in safety-critical or adversary-prone environments.

In this project, we investigate the vulnerability of the ResNet-34 model (He et al. 2016) trained on ImageNet-1K to three adversarial attack strategies: the Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014), Projected Gradient Descent (PGD) (Madry et al. 2018), and a spatially constrained patch-based PGD variant. Each attack operates under an ℓ_∞ perturbation constraint, with the goal of degrading Top-1 and Top-5 classification accuracy while preserving visual similarity to the original input.

Our study is conducted on a curated subset of 500 test images spanning 100 diverse classes from ImageNet. We evaluate attack efficacy both on the source model (ResNet-34) and on a second model (DenseNet-121) (Huang et al. 2017) to assess transferability. Through rigorous experimentation and ablation studies, we demonstrate that even small, localized perturbations can cause severe degradation in classification performance. These results highlight the pressing need for robust defenses and the importance of understanding attack dynamics across architectures.

Methodology

Dataset and Model

We use a curated subset of ImageNet-1K consisting of 500 test images drawn from 100 diverse classes. Images are preprocessed using ImageNet normalization statistics: $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$, ensuring compatibility with pretrained models. The baseline model is ResNet-34, with `IMAGENET1K_V1` weights. We evaluate Top-1 and Top-5 accuracy on the clean dataset to establish a performance baseline.

Adversarial Attack Design

We explored three types of adversarial attacks: FGSM, PGD, and patch-based PGD, each operating under an ℓ_∞ perturbation constraint. These attacks were applied independently to each test image, and their effectiveness was measured in terms of accuracy drop and transferability.

Fast Gradient Sign Method (FGSM, $\epsilon = 0.02$) The FGSM is a one-shot attack that perturbs inputs in the direction of the gradient of the loss function. We experimented with ϵ values ranging from 0.005 to 0.02, and selected $\epsilon = 0.02$ as a balance between visual imperceptibility and attack strength.

Projected Gradient Descent (PGD, 100 steps, $\epsilon = 0.02$) PGD improves upon FGSM by applying small iterative perturbations and projecting each result back into the ℓ_∞ ball. We configured the PGD attack with 100 steps and a step size of 0.005. This configuration was determined through

Project Codebase: https://github.com/RaghuHemadri/ECE-GY-7123-DL-Project3.git empirical tuning to maximize attack efficacy without distorting image structure. $\alpha = 0.005$ and projection Π_ϵ ensures that $||x_{t+1} - x||_\infty \leq 0.02$. Clamping to $[0, 1]$ was necessary to maintain valid image ranges and prevent artifact generation. This attack was significantly more effective than FGSM, with near-total degradation of classification accuracy.

Patch-based PGD Attack (32×32 patch, $\epsilon = 0.5$) We developed an advanced patch-based PGD attack. Patch attack focuses on modifying only a small 32×32 region.

Given the spatial constraint, we increased the perturbation budget to $\epsilon = 0.5$ and implemented several enhancements to maximize attack effectiveness:

- **Center-biased Placement:** Rather than using arbitrary random patch locations, we placed patches within the central region of the image, where discriminative object features are more likely to be found.
- **Multi-target Optimization:** To simultaneously attack both Top-1 and Top-5 accuracy, we identified most promising incorrect target classes for each image.
- **Momentum-based Updates:** We incorporated momentum in the gradient updates to escape poor local optima and enhance attack convergence (Dong et al. 2018).
- **Multi-restart Optimization:** We used 10 restarts with different random initializations and target class selections, selecting the best adversarial example based on a weighted combination of Top-1 and Top-5 attack success.
- **Adaptive Step Size:** We employed a decreasing step size schedule that starts at $\alpha = 0.04$ and gradually decreases to 0.008 over 150 iterations.
- **Combined Loss Function:** We developed a specialized loss function that balances pushing the true class out of both Top-1 and Top-5 predictions:

$$\mathcal{L} = -\left(0.3 \cdot \text{margin}_{\text{top-1}} + 0.7 \cdot \text{margin}_{\text{top-5}}\right) \quad (1)$$

  where $\text{margin}_{\text{top-1}}$ represents the logit difference between the strongest incorrect class and the true class, while $\text{margin}_{\text{top-5}}$ focuses on the fifth-strongest incorrect class.

Transferability Evaluation

To evaluate the generalization capability of adversarial examples, we tested all three attack sets on a DenseNet-121 model, pretrained on ImageNet-1K. This step assesses whether perturbations designed to fool ResNet-34 also transfer to a different architecture.

Results

Adversarial Attack Effectiveness

Attack Type	**ResNet-34**		**DenseNet-121**	
	Top-1	**Top-5**	**Top-1**	**Top-5**
No Attack	89.60%	99.40%	88.00%	98.40%
FGSM	8.60%	40.40%	62.40%	89.40%
PGD	0.00%	2.80%	65.60%	93.80%
Patch	0.20%	37.20%	85.20%	98.00%

Table 1: Top-1 and Top-5 accuracy for ResNet-34 and DenseNet-121 across clean and adversarial datasets.

Table 1 summarizes the Top-1 and Top-5 classification accuracy of ResNet-34 and DenseNet-121 models under clean and adversarial conditions. As expected, both models perform well on unperturbed inputs, achieving over 88% Top-1 accuracy. However, introducing adversarial perturbations causes a sharp performance decline.

For ResNet-34, the FGSM attack reduces Top-1 accuracy from 89.60% to just 8.60%, and PGD completely degrades Top-1 accuracy to 0.00%. In contrast, the patch-based PGD attack, despite being spatially constrained, still reduces Top-1 accuracy to 0.20%. This shows the high vulnerability of ResNet-34 to even localized perturbations.

DenseNet-121 shows relatively better robustness. The same FGSM and PGD attacks result in Top-1 accuracies of 62.40% and 65.60%, respectively. The patch attack affects it the least, preserving 85.20% Top-1 accuracy, highlighting reduced transferability of ResNet-targeted perturbations to a different model.

Figure 1 visualizes the impact of different attacks on a sample ImageNet image. The FGSM attack (top row) produces sparse perturbations that subtly alter the prediction. PGD (middle row) generates dense, high-frequency noise, visibly distorting the image at pixel level. The patch-based attack (bottom row) introduces a concentrated square perturbation, demonstrating how localized modifications can still achieve model misclassification under a large ϵ constraint. These examples illustrate the diverse nature of attack strategies and the importance of defending against both global and localized adversarial threats.

Although all three attacks target the same input image, the nature of their perturbations differs significantly. FGSM introduces globally distributed but low-magnitude changes that are nearly imperceptible. In contrast, PGD applies iterative updates that accumulate into denser and more structured noise patterns, making the perturbations more prominent. The patch-based PGD attack confines its changes to a small 32×32 region, resulting in a clearly visible square patch. Despite its localized scope, this spatially constrained perturbation is highly effective, demonstrating that adversarial influence need not be spread across the entire image to succeed (see Figure 1).

Ablation Studies

Effect of ϵ on FGSM Attack Success We conducted an ablation study to examine how the FGSM (Goodfellow,

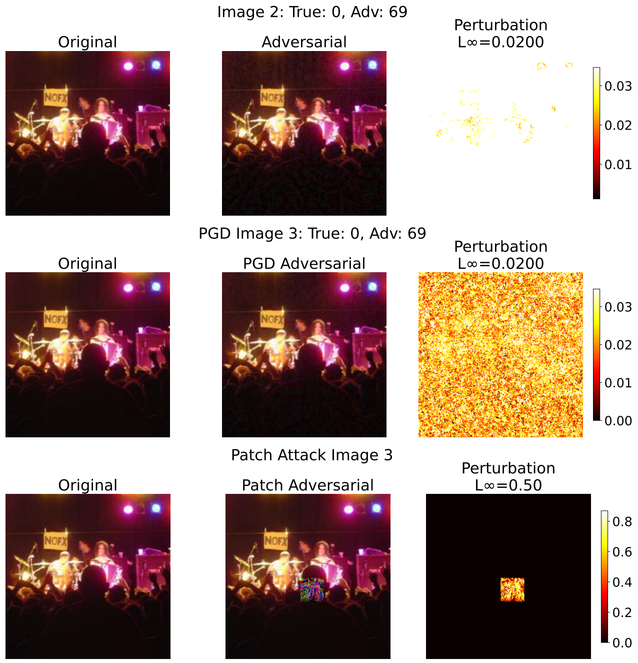

Figure 1: Visualization of adversarial attacks on an ImageNet image. Each row shows a different attack method applied to the same original input. **Top row:** FGSM attack with $\epsilon = 0.02$ generates sparse perturbations that fool the model. **Middle row:** PGD attack with $\epsilon = 0.02$ results in dense, iterative perturbations leading to misclassification. **Bottom row:** Patch-based PGD attack introduces a localized visible patch with $\epsilon = 0.5$ to successfully alter the model prediction. Each set includes (left) the original image, (middle) the adversarial image, and (right) the visualized perturbation with corresponding L_∞ norm.

Shlens, and Szegedy 2014) attack's effectiveness varies with the perturbation budget ϵ. As shown in Figure 2, increasing ϵ from 0.005 to 0.020 consistently reduces the model's accuracy. Top-1 accuracy falls sharply from 15.4% to 8.6%, with the most significant drop observed up to $\epsilon = 0.01$. Beyond this point, the accuracy plateaus, suggesting that further increases in ϵ yield diminishing gains in attack success.

Top-5 accuracy exhibits a similar trend, decreasing from 55.2% to 40.4%. These results indicate that even modest perturbation levels are sufficient to break model predictions under FGSM, and that tuning ϵ is crucial for balancing attack strength and imperceptibility.

Effect of PGD Steps on Attack Success To study the impact of the number of gradient steps in PGD (Madry et al. 2018), we conducted an ablation analysis varying the number of iterations from 80 to 140 while keeping ϵ fixed. As shown in Figure 3, Top-1 accuracy consistently remains at 0%, indicating the PGD attack is highly effective regardless of step count. However, Top-5 accuracy shows non-monotonic behavior, decreasing overall but with slight fluctuations across step values.

At 80 steps, the Top-5 accuracy is around 4.2%, dropping to 2.8% at 100 steps, increasing slightly at 120 steps, and reaching its lowest at 140 steps (2.2%). These results suggest that while increasing steps beyond a threshold does not enhance Top-1 degradation, it can marginally influence the broader prediction distribution, affecting Top-5 metrics. Hence, a moderate number of steps (e.g., 100–120) may balance efficiency and effectiveness.

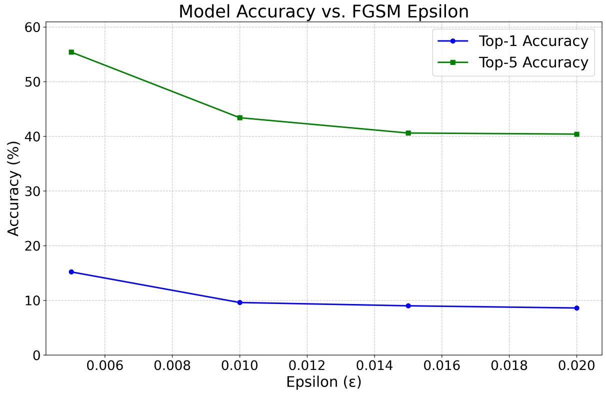

Figure 2: FGSM Ablation Study: Effect of ϵ on attack success.

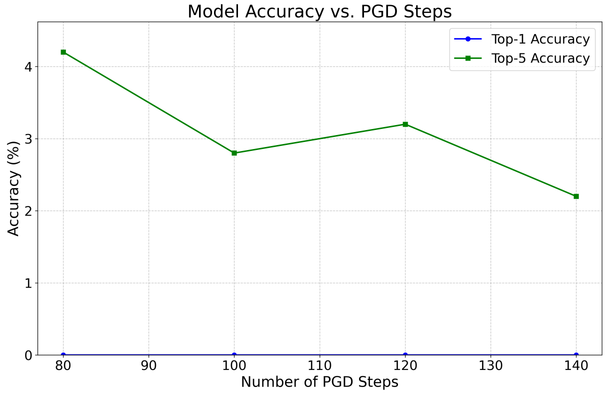

Figure 3: PGD Ablation Study: Steps vs. success rate.

Effect of Step Size, Restarts, and Targeting on Patch-Based PGD Attacks We analyze how different hyperparameters influence the effectiveness of patch-based PGD attacks by varying the number of gradient steps, step size (α), number of random restarts, and target class configuration. The results are presented in Table 2.

A clear trend emerges showing that more sophisticated configurations with random restarts and refined step sizes substantially improve attack success. For instance, increasing restarts from none to 8 (with a smaller step size of 0.015) reduces Top-1 accuracy from 11.20% to 2.80%, highlighting the role of diverse initializations in escaping poor local optima (Dong et al. 2018). Moving to a very small step size (0.004) with 10 restarts further drops Top-1 accuracy to 0.20%, suggesting that fine-grained updates combined with randomness can maximize attack effectiveness.

PGD Steps	**Step Size**	**Restarts**	**Accuracy (%)**		**Target Class**
			Top-1	**Top-5**	
150	0.015	8	2.80	84.60	Untargeted
100	0.01	-	11.20	60.40	5
100	0.05	8	1.00	44.20	10
150	0.004	10	0.20	37.20	12

Table 2: PGD Patch Attack Ablation Study: Performance metrics across different configurations. Alpha represents the step size for gradient updates, Restarts indicates the number of random initializations, and Target Class denotes number of promising incorrect classes.

Interestingly, the targeted attack with step size 0.01 and target class 5 is less effective than the untargeted variant with the same number of steps, suggesting that targeting can sometimes constrain the attack space. However, targeted attacks with a large step size (0.05) or small step size (0.004) still achieve strong performance, indicating the importance of tuning α in tandem with the number of steps and restarts.

Overall, this ablation reveals that targeted attacks benefit from smaller, precise gradients and multiple initializations, while untargeted attacks can achieve high transferability with moderate tuning.

Training and Evaluation Times

Adversarial attacks were executed on a dataset of 500 images using GPU acceleration. FGSM completed almost instantly, while PGD attacks scaled with the number of optimization steps, ranging from 12 to 21 minutes. The patch-based PGD attack, due to its iterative nature and multiple restarts, was significantly slower, taking approximately 4 hours. Inference across all datasets using DenseNet-121 (Huang et al. 2017) was efficient, completing in under 10 seconds per dataset.

Findings and Interpretation

Our experiments confirm the vulnerability of deep classifiers like ResNet-34 (He et al. 2016) and DenseNet-121 to carefully crafted adversarial perturbations (Szegedy et al. 2013). The FGSM and PGD attacks drastically reduced Top-1 accuracy, with PGD achieving complete model failure on ResNet-34. Despite being spatially constrained, patch-based attacks remained surprisingly effective, underscoring the potency of localized perturbations.

Through ablation studies, we observed that FGSM's success saturates beyond $\epsilon = 0.01$, while PGD benefits marginally from increased steps. For patch-based PGD, we found that smaller step sizes combined with random restarts significantly improve attack strength. Interestingly, untargeted patch attacks often outperformed targeted ones, suggesting that flexibility in adversarial objectives can enhance performance.

Overall, our findings highlight the ease with which adversaries can degrade deep model performance and stress the need for robust defenses, especially against both global and localized attacks.

Conclusion

This project demonstrates the pronounced vulnerability of deep image classifiers to adversarial attacks, even when perturbations are imperceptible or confined to small image regions. By launching a series of attacks—FGSM, PGD, and a custom-designed patch-based PGD—we observed severe drops in classification accuracy on a ResNet-34 model trained on ImageNet, with PGD achieving complete Top-1 model failure under a modest ℓ_∞ constraint.

Notably, our patch-based attack, despite modifying only a 32×32 region, proved highly effective, underscoring that spatial locality does not preclude adversarial success. Transferability tests revealed that attacks crafted on ResNet-34 partially generalize to DenseNet-121, suggesting that adversarial vulnerabilities extend beyond single architectures.

Through detailed ablation studies, we identified key factors contributing to attack effectiveness, including step size, number of restarts, and loss function design. These findings reinforce the need for adversarial robustness as a fundamental requirement in modern deep learning systems. Future work could explore defenses such as adversarial training (Madry et al. 2018) or randomized smoothing (**?**), as well as physical-world evaluations to assess real-world exploitability.

References

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4700–4708.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Acknowledgment: We used OpenAI's ChatGPT to help rephrase and format parts of this report. All code and experiments were completed independently.