Jailbreaking Deep Models

Nishant Sharma[1], **Rahul Mallidi**[2], **Anushka Garg**[3]

NYU Tandon School of Engineering
ns6287@nyu.edu, rm7020@nyu.edu, ag9012@nyu.edu

Abstract

Deep neural networks demonstrate impressive performance on various computer vision tasks, yet they are notably vulnerable to adversarial perturbations—minor, carefully crafted changes to input images that significantly degrade classifier performance. This project explores the effectiveness of adversarial attacks on a ResNet-34 model pre-trained on ImageNet-1K. Specifically, we implement and evaluate pixel-wise adversarial attacks including the Fast Gradient Sign Method (FGSM) and iterative methods like Projected Gradient Descent (PGD) with a constrained perturbation magnitude ($\epsilon = 0.02$). Our results demonstrate substantial reductions in model accuracy; FGSM reduces top-1 accuracy from 76.0% to 32.0%, while the iterative PGD method achieves a more drastic drop, decreasing top-1 accuracy to as low as 1.2%. Additionally, we investigate patch-based attacks and the transferability of adversarial examples across architectures, notably evaluating performance degradation on DenseNet-121. This work highlights critical security vulnerabilities inherent in current deep learning models and underscores the necessity for developing robust adversarial defenses.

Supporting Material

- **Code Repository:** Please refer to the attached link here for relevant codebase to the project.

Introduction

Image classification is a cornerstone task in computer vision, underpinning applications ranging from facial recognition and medical diagnosis to autonomous driving systems. Deep neural networks, particularly convolutional neural networks (CNNs) such as ResNet, have established remarkable accuracy benchmarks on extensive datasets like ImageNet. Despite their proven capability, these networks are alarmingly susceptible to adversarial attacks—subtle yet precisely engineered modifications to input images that mislead models into incorrect classifications.

Adversarial examples pose significant risks in real-world deployments, especially in safety-critical domains, highlighting urgent needs for understanding and mitigating these vulnerabilities. Among various adversarial methodologies, pixel-wise attacks such as the Fast Gradient Sign Method (FGSM) and iterative techniques like Projected Gradient Descent (PGD) have emerged as particularly effective strategies to exploit model weaknesses. These methods aim to maximize the model's prediction error while enforcing minimal perceptual changes to the images.

In this project, we thoroughly assess the vulnerability of a ResNet-34 model trained on the ImageNet-1K dataset by systematically implementing and comparing adversarial attacks under stringent perturbation constraints ($\epsilon = 0.02$). Furthermore, we extend our analysis to patch-based adversarial attacks, which restrict perturbations to limited spatial regions, thus evaluating model susceptibility under localized threat scenarios. To assess the generalizability of adversarial attacks, we evaluate the transferability of these perturbed images across different model architectures, including DenseNet-121. Our investigation not only demonstrates critical security vulnerabilities but also provides insights into improving model robustness against adversarial threats.

Existing Work

The vulnerability of deep neural networks to adversarial examples was first highlighted by Szegedy et al. (2014), who demonstrated that imperceptible perturbations could cause high-confidence misclassifications. Building on this, Goodfellow, Shlens, and Szegedy (2015) introduced the Fast Gradient Sign Method (FGSM), a simple and computationally efficient attack that leverages the gradient of the loss function with respect to the input to craft adversarial examples. FGSM set the foundation for subsequent attacks, particularly iterative variants.

One of the most well-known extensions of FGSM is the Projected Gradient Descent (PGD) attack, proposed by Madry et al. (2018), which applies FGSM iteratively with projection steps to ensure the perturbation remains within an ℓ_∞-bounded region. PGD is widely regarded as a strong first-order adversary and has become a standard benchmark for evaluating model robustness.

In addition to pixel-wise attacks, several works have explored localized adversarial strategies. Patch-based attacks, such as the adversarial patch introduced by Brown et al. (2017), involve modifying a small portion of the image to cause misclassification, sometimes even with universal (input-agnostic) patches. These methods highlight how constrained perturbations can still achieve high attack success

rates.

Furthermore, the concept of transferability—where adversarial examples generated for one model fool other models—has received considerable attention. Studies such as Liu et al. (2017) have shown that adversarial examples often generalize across architectures, making black-box attacks a plausible threat even without access to the target model's parameters.

Together, these foundational works motivate the current project, which focuses on evaluating and extending these techniques under practical constraints using a production-grade ResNet-34 model on a real-world dataset subset.

Dataset

The dataset used for this project is a curated subset of the ImageNet-1K dataset, containing images from 100 distinct classes. This subset was provided as part of the project resources and includes a directory of labeled image folders along with a corresponding JSON file mapping folder names to official ImageNet label indices and class names. Each image in the dataset is pre-categorized into its respective class folder, enabling straightforward use with torchvision's `ImageFolder` utility.

All images were preprocessed using the standard ImageNet normalization procedure: pixel values were normalized channel-wise using a mean of $[0.485, 0.456, 0.406]$ and a standard deviation of $[0.229, 0.224, 0.225]$. The dataset was wrapped into a PyTorch `DataLoader` with a batch size of 32 and no shuffling to maintain label order integrity during adversarial evaluation.

This dataset actually serves as the evaluation benchmark for all adversarial experiments in this project. Its diversity across 100 ImageNet categories ensures that model robustness is tested across a broad spectrum of visual patterns and semantics. We used this dataset as the baseline for measuring clean classification accuracy and the effect of various adversarial attacks applied in later stages of the project.

Methodology

We adopt a structured methodology to analyze and evaluate the strength of the ResNet-34 model against adversarial attacks. Our experiments are conducted in five phases, corresponding to the project tasks.

Task 1: Baseline Evaluation. We begin by evaluating the pretrained ResNet-34 model on the clean test dataset to establish baseline performance. Standard preprocessing techniques and top-1 and top-5 accuracy metrics are used.

Task 2: FGSM Attack. The sign of the gradient of the loss with respect to the input image is used in the one-step assault known as the Fast Gradient Sign Method (FGSM). To make sure the change is not noticeable, the perturbation is limited with a ϵ value of 0.02. Images that are disturbed are created, displayed, and assessed for deterioration in classification.

Task 3: Improved Attacks. We enhance the attack strength using iterative techniques such as Projected Gradient Descent (PGD). The attack is performed over multiple steps with smaller step sizes, projecting the result back into the ℓ_∞ ball of radius ϵ after each step. PGD offers significantly more potent adversarial examples than FGSM.

Task 4: Patch Attack. We localize the attack by restricting perturbations to a random 32×32 region of the image. To compensate for the smaller attack area, a larger ϵ (e.g., 0.3) is permitted. The impact of spatially constrained attacks is assessed both visually and quantitatively.

Task 5: Transferability Analysis. To evaluate the generality of adversarial examples, we test all perturbed datasets against an alternative ImageNet model, DenseNet-121. Top-1 and top-5 accuracy metrics are reported, highlighting the extent to which attacks on one model can influence others.

All experiments are conducted using PyTorch, and performance metrics are computed across the full dataset. Visualization routines assist in interpreting model behavior under attack. This methodology provides a comprehensive framework for studying adversarial robustness in production-grade neural networks.

Task	Attack	ϵ	Iter.	Region
Task 2	FGSM	0.02	1	Full
Task 3	I-FGSM	0.02	5	Full
Task 3 (Alt)	PGD	0.02	10	Full
Task 4	Patch I-FGSM	0.02	10	32×32
Task 4 (Alt)	Patch PGD	0.02	10	32×32

Table 1: Adversarial attack configurations across Tasks 2–4 with ℓ_∞ perturbation bound $\epsilon = 0.02$.

Attack	Max ℓ_∞ Distance	Avg ℓ_∞ Distance	ϵ Constraint
FGSM (Task 2)	0.02000	0.02000	0.02
I-FGSM (Task 3)	0.02000	0.02000	0.02
PGD (Task 3 Alt)	0.02000	0.02000	0.02
Patch I-FGSM (Task 4)	0.02000	0.02000	0.02
Patch PGD (Task 4 Alt)	0.02000	0.02000	0.02

Table 2: Perturbation budget verification results for all adversarial attack variants.

Results

This section presents the evaluation outcomes for all five project tasks, comparing classification performance under clean and adversarial conditions. The adversarial perturbation budget is set to $\epsilon = 0.02$ for all tasks, applied either across the complete image or localized within a 32×32 patch region.

Task 1: Baseline Evaluation. As shown in Table 3, the ResNet-34 model got a top-1 accuracy of 76.00% and top-5 accuracy of 94.20% on the clean test dataset, establishing a strong reference point for evaluating adversarial robustness.

Task 2: FGSM Attack. The Fast Gradient Sign Method (FGSM), implemented with a single gradient step, drops model accuracy to 6.00% top-1 and 35.40% top-5. The adversarial examples remain imperceptibly close to the originals, while the attack results in a 70% reduction in top-1 accuracy.

Task 3: Improved Attacks. Iterative attacks like I-FGSM and PGD were significantly more effective. Both produced top-5 accuracies close to 14.20% and a sharp top-1 decline to 0.20%. PGD followed the same ℓ_∞ constraint as FGSM and was run across ten stages with three restarts.

Task 4: Patch Attacks. We maintained $\epsilon = 0.02$ while restricting perturbations to a 32×32 region. A top-1 accuracy of 1.80% was obtained using the I-FGSM patch attack, and PGD further decreased it to 1.60%. These findings demonstrate that performance can be severely hampered by perturbations that are only spatially confined.

Task 5: Transferability. When adversarial examples were applied to DenseNet-121, they maintained much of their efficacy. For instance, PGD examples reduced DenseNet's top-1 accuracy from 74.80% to 65.00%, and patch-based PGD adversarials reduced it to 59.80%. This underscores the model-agnostic threat of adversarial attacks.

Attack Type	Top-1 Accuracy (%)	Top-5 Accuracy (%)
Clean Test Set (Task 1)	76.00	94.20
FGSM Attack (Task 2)	6.00	35.40
I-FGSM Attack (Task 3)	**0.20**	14.20
PGD Attack (Task 3 Alt)	0.20	**14.20**
Patch I-FGSM (Task 4)	1.80	25.20
Patch PGD (Task 4 Alt)	1.60	25.20
Transfer to DenseNet-121 (Task 5)		
Original Set	74.80	93.60
FGSM (from Task 2)	63.40	89.20
PGD (from Task 3 Alt)	65.00	91.40
Patch Attack (from Task 4 Alt)	59.80	84.80

Table 3: **Summary of classification performance under adversarial attacks.** This table reports top-1 and top-5 accuracies of ResNet-34 on clean and adversarial datasets (Tasks 1–4), and transfer performance to DenseNet-121 (Task 5). All attacks use a maximum perturbation of $\epsilon = 0.02$, with spatial constraint applied for patch-based attacks.

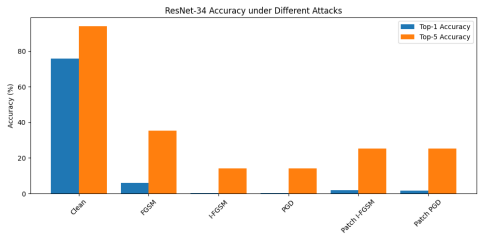

Figure 1: **ResNet-34 classification accuracy under different attack strategies.** FGSM, I-FGSM, PGD, and patch-based variants show significant degradation in both top-1 and top-5 accuracy under the ℓ_∞ constraint of $\epsilon = 0.02$.

The results collectively emphasize the brittleness of deep networks under both pixel-wise and patch-constrained attacks, and underline the need for stronger adversarial defenses in real-world deployment settings.

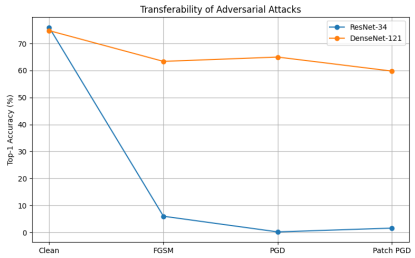

Figure 2: **Transferability of adversarial attacks across architectures.** Adversarial examples generated for ResNet-34 also degrade DenseNet-121 performance, with PGD and patch PGD maintaining strong attack efficacy.

Figure 3: **Visual comparison of original vs. adversarial images.** Examples of adversarial perturbations generated using FGSM (left), PGD (center), and patch-based attack (right). Each pair shows the original image (top) and the adversarial counterpart (bottom), demonstrating how minimal or localized perturbations can mislead the classifier.

Discussion

The results highlight the inherent vulnerability of deep convolutional neural networks, such as ResNet-34, to small but structured adversarial perturbations. While FGSM significantly degraded model performance with a single-step update, iterative variants like I-FGSM and PGD proved even more detrimental, reducing top-1 accuracy from 76.00% to 0.20% despite adhering to strict ℓ_∞ constraints.

Furthermore, patch-based attacks demonstrated that localized perturbations, though spatially restricted to 32×32 regions, can still be highly effective. Variants of the patch attack, I-FGSM and PGD, decreased top-1 accuracy to less than 2%. If the perturbation is positioned and tuned strategically, these results confirm that reducing the attack surface area does not always increase robustness.

Importantly, the transferability analysis confirmed that adversarial examples retain their potency across model architectures. DenseNet-121, when tested on PGD-generated inputs, showed a top-1 drop to 65.00% and 59.80% for full-image and patch-based attacks respectively. This suggests that even black-box adversaries can pose severe threats using surrogate models.

Together, these findings underscore the pressing need for adversarial defense mechanisms that generalize across input scales and model architectures. Promising strategies include adversarial training, randomized input transformations, and

robust loss objectives.

Conclusion

This project provided a comprehensive evaluation of adversarial attacks on a production-grade ResNet-34 model using a curated subset of the ImageNet-1K dataset. We implemented and analyzed multiple attack strategies—FGSM, I-FGSM, PGD, and spatially constrained patch attacks—all under a tight ℓ_∞-norm constraint of $\epsilon = 0.02$.

Our experiments confirmed that deep classifiers are highly susceptible to even imperceptible perturbations. Among all strategies, PGD and I-FGSM consistently achieved the most severe degradation in classification accuracy. Patch-based attacks also proved effective, underscoring the model's sensitivity even when perturbations are spatially limited.

Notably, adversarial examples exhibited strong transferability to DenseNet-121, highlighting broader implications for the deployment of deep models in black-box settings. These findings emphasize the necessity for robust training pipelines and architectural safeguards in real-world AI deployments.

We hope that the systematic attack framework and empirical insights presented in this study will serve as a foundation for future work on adversarial defenses and model robustness benchmarking.

References

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2017. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

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*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; et al. 2014. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.