Adversarial Robustness Against ResNet-34 on ImageNet: FGSM, PGD, and Patch-Based Attacks

Pavan Kishore Ramavath[1], **Anvesh Varma Dantuluri**[1], **Sai Ram Purimetla**[1]
[1]New York University Tandon School of Engineering
pr2622@nyu.edu, ad7647@nyu.edu, sp8201@nyu.edu
Link to Repository: GitHub Repository

Abstract

We explore parameter-constrained adversarial robustness analysis of convolutional neural networks by evaluating gradient-based and patch-based attack strategies on a pretrained ResNet-34 model using a 100-class subset of the ImageNet-1K dataset. Our work systematically implements the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and a localized learnable patch attack to assess performance degradation under controlled perturbations. Each attack is configured with fixed budgets or spatial constraints, allowing for fair comparison of attack strength and computational efficiency.

By tuning step sizes, iteration counts, and patch configurations, we balance visual fidelity with adversarial impact. Evaluation metrics include Top-1 and Top-5 classification accuracy, perturbation magnitude, runtime cost, and transferability to unseen architectures such as DenseNet-121. Our experiments reveal that PGD achieves the strongest degradation, reducing ResNet-34's Top-1 accuracy to 0.00%, while patch-based attacks demonstrate high transferability even with limited pixel manipulation. These findings underscore the vulnerabilities of deep image classifiers and advocate for stronger model-level defenses in high-stakes visual recognition pipelines.

Introduction

Convolutional Neural Networks (CNNs) have become foundational models in computer vision, achieving state-of-the-art performance on challenging benchmarks such as ImageNet-1K. Despite their success in large-scale image classification tasks, these models exhibit a critical vulnerability to adversarial examples—carefully crafted perturbations that are visually imperceptible to humans but can significantly alter model predictions. This vulnerability poses severe risks in safety-critical applications including autonomous vehicles, medical imaging, and surveillance systems. Consequently, the study of adversarial robustness has emerged as a vital research direction in trustworthy AI, with numerous attack strategies proposed to evaluate and expose model fragility under constrained perturbations.

In this work, we present a systematic evaluation of the adversarial robustness of a pretrained ResNet-34 model on the ImageNet-1K dataset by implementing three canonical attack strategies: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and a localized patch-based attack. These methods span both ℓ_∞-bounded gradient-based perturbations and spatially constrained modifications, enabling a comprehensive robustness assessment. Each attack is calibrated to ensure fair comparison in terms of perturbation budget, runtime cost, and impact on model performance. We report results on top-1 and top-5 classification accuracy and further evaluate cross-model transferability by deploying adversarial examples on a DenseNet-121 architecture. Human-interpretable predictions are obtained using the `labels_list.json` mapping. Our findings reveal that PGD consistently induces the greatest accuracy degradation, while patch-based attacks demonstrate notable transferability. These insights emphasize the urgent need for robust model design and adaptive defenses against diverse adversarial threats.

Architecture Overview

The core architecture used in our evaluation is the ResNet-34 convolutional neural network, a 34-layer deep residual model designed to enable efficient training of deeper networks through identity-based shortcut connections. The network begins with a 7×7 convolutional layer and max pooling, followed by four stages of residual blocks. Each block learns residual mappings that are added to the block input, allowing gradients to propagate more effectively during backpropagation. The final feature representation is aggregated through global average pooling and passed to a fully connected layer producing class logits.

All adversarial attacks in this study were applied directly to this frozen architecture in inference mode. The ResNet-34 model serves as the reference system for generating and evaluating adversarial examples, where attacks are injected post-input and prior to the model's feature extraction and classification stages.

To evaluate the transferability of adversarial perturbations across model architectures, we additionally include a DenseNet-121 model with densely connected convolutional layers. Unlike ResNet's additive skip connections, DenseNet concatenates outputs from all previous layers within a block, promoting feature reuse and compact representation. This architectural contrast enables robust assessment of cross-model generalization under attack.

Each attack—whether globally perturbing gradients or locally optimizing spatial patches—is seamlessly integrated into the inference pipeline without modifying the internal structure of the model. This architectural setup ensures that all performance degradation stems from adversarial input modifications rather than network retraining or weight adaptation.

Methodology

Design Rationale and Model Selection

To investigate adversarial robustness in a constrained yet diverse setting, we utilized a 100-class subset of the ImageNet-1K dataset. This subset offered a balance between computational efficiency and semantic coverage. Two pretrained convolutional architectures were selected: ResNet-34 and DenseNet-121. ResNet-34 was chosen as the white-box target due to its moderate depth and widespread use. DenseNet-121 was selected for black-box transferability analysis, as its densely connected structure provides a contrasting design to residual networks.

Adversarial Attack Design

We implemented three adversarial attack strategies FGSM, PGD, and a trainable patch-based attack each designed to reveal different facets of model vulnerability.

FGSM was implemented as a one-shot attack, perturbing inputs in the direction of the gradient sign. We selected $\epsilon = 0.02$ after preliminary sweeps showed that lower values produced negligible impact while higher values introduced visible artifacts. This setting allowed a fast, lightweight test of first-order vulnerability.

PGD was configured as a 10-step iterative attack with a step size of $\alpha = 0.005$, consistent with literature on ℓ_∞-bounded adversaries. These hyperparameters were chosen to strike a balance between convergence strength and runtime cost. In practice, PGD achieved the most destructive performance on ResNet-34, dropping Top-1 accuracy to 0%. However, its transferability to DenseNet-121 was limited, revealing its tendency to overfit to the source model's gradients.

Patch Attack involved optimizing a 32×32 square region superimposed on each image. We used the Adam optimizer with a learning rate of 0.2 for 500 steps. This configuration was empirically tuned to enable convergence without exploding gradients. While slower to compute, the patch attack generalized best to DenseNet-121, indicating its architecture-agnostic nature. The patch required no gradient access to the target model during transfer, demonstrating its threat potential in black-box scenarios.

Pros and Cons of Hyperparameter Choices

Each attack presented unique trade-offs. FGSM was computationally efficient and quick to evaluate but less effective against robust or deep models. PGD provided complete degradation on the source model but required multiple iterations and failed to transfer well. The patch attack, though computationally expensive to generate (up to 60 minutes), maintained transferability and perceptual coherence, highlighting its practical risk in real-world deployments.

Lessons Learned

Through experimentation, we observed that deeper or denser architectures like DenseNet-121 exhibited higher resilience to pixel-level attacks, likely due to their enhanced feature redundancy. However, they remained vulnerable to structured spatial attacks like adversarial patches. We also learned that hyperparameter tuning is critical not just for attack strength, but for reproducibility and fair cross-architecture comparison. Attacks with the same ϵ can have vastly different outcomes depending on model gradients, architectural depth, and internal feature reuse.

Evaluation Strategy

All attacks were evaluated using Top-1 and Top-5 accuracy. Predictions were considered correct if the true label appeared in the top-k logits, with class mappings defined by the `labels_list.json` file. Relative accuracy drop from the clean baseline was calculated to capture attack severity. For fairness, all models were frozen during testing and evaluated under the same preprocessing pipeline, batch size, and hardware conditions.

Results

Baseline Robustness

ResNet-34 and DenseNet-121 demonstrate strong baseline performance on the unperturbed 100-class ImageNet subset. ResNet-34 achieves 76.00% Top-1 and 94.20% Top-5 accuracy, while DenseNet-121 records 74.80% and 93.60%, respectively. These values serve as reference baselines for measuring adversarial impact.

Model	**Dataset**	**Top-1 (%)**	**Top-5 (%)**
ResNet-34	Original	76.00	94.20
ResNet-34	FGSM	6.20	35.40
ResNet-34	PGD	0.00	5.00
ResNet-34	Patch	5.20	46.80
DenseNet-121	Original	74.80	93.60
DenseNet-121	FGSM	63.40	89.40
DenseNet-121	PGD	63.80	90.60
DenseNet-121	Patch	44.80	75.20

Table 1: Top-1 and Top-5 accuracy of ResNet-34 and DenseNet-121 under clean and adversarial conditions.

Effectiveness of Pixel-wise Attacks

FGSM and PGD, both bounded by ℓ_∞ norm with $\epsilon = 0.02$, lead to substantial performance degradation on ResNet-34. FGSM reduces Top-1 accuracy from 76.0% to 6.2%, while PGD drives it to 0.0%. The Top-5 accuracy under PGD also falls sharply to 5.0%. In contrast, DenseNet-121 retains much higher robustness under the same conditions—maintaining over 63% Top-1 accuracy for both FGSM and PGD, with Top-5 exceeding 89%.

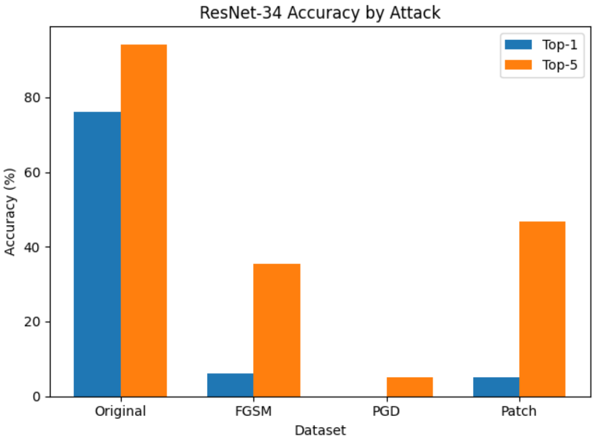

Figure 1: Top-1 and Top-5 accuracy for ResNet-34 under different attack methods.

Impact of Patch-based Attacks

The 32 × 32 learnable patch causes severe degradation on ResNet-34, reducing Top-1 accuracy to 5.2%. However, the patch attack demonstrates significantly stronger transferability than gradient-based attacks. When applied to DenseNet-121, the same patch reduces Top-1 accuracy from 74.8% to 44.8%, and Top-5 from 93.6% to 75.2%. This highlights the architecture-agnostic nature of patch-based attacks.

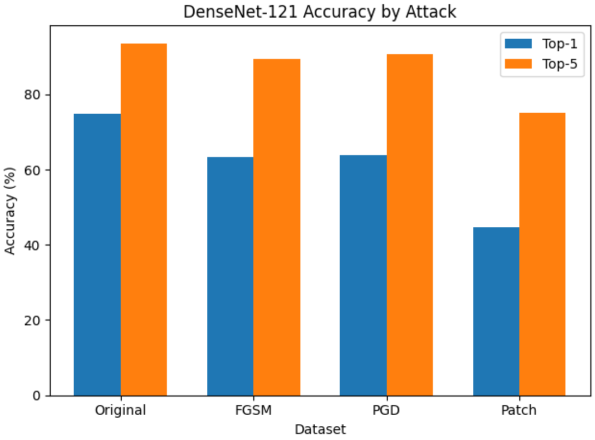

Figure 2: Top-1 and Top-5 accuracy for DenseNet-121 under transferred adversarial attacks.

Top-5 Accuracy Trends

While Top-1 accuracy reflects primary classification failure, Top-5 accuracy trends reveal the extent of broader model confusion. On ResNet-34, PGD drops Top-5 accuracy from 94.20% to 5.00%, the lowest among all attack variants. FGSM and patch attacks also result in significant Top-5 degradation. DenseNet-121, although more robust, still suffers considerable Top-5 drop under the patch attack (to 75.2%).

Qualitative Examples of Adversarial Attacks

To visually illustrate the behavior of each attack type, we present side-by-side examples of original images, adversarial versions, scaled perturbations, and the resulting Top-5 prediction distributions.

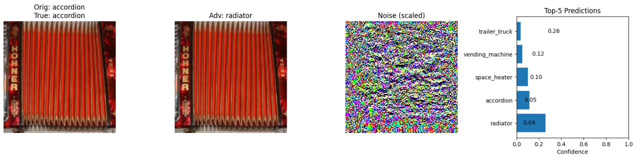

Figure 3: FGSM attack: The original image (left) is correctly classified as an accordion. A single-step gradient-based perturbation causes misclassification as radiator, despite minimal visual difference. The noise map (middle) is amplified for visibility.

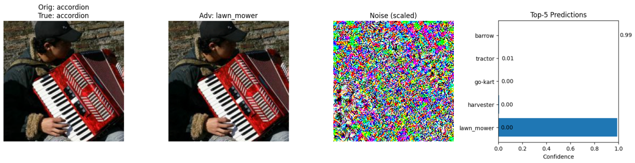

Figure 4: PGD attack: The adversarial image is confidently misclassified as lawn mower. Compared to FGSM, the noise pattern is more structured and aggressive, leading to stronger degradation in prediction certainty.

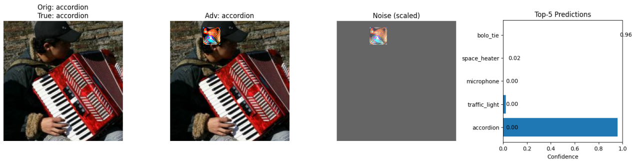

Figure 5: Patch attack: A fixed 32 × 32 adversarial patch is inserted in the image (top-left corner). The patch redirects the prediction distribution despite the image retaining visual clarity. Notably, the model no longer classifies the object as accordion.

Relative Performance Drop

Figure 6 presents the relative Top-1 accuracy drop across attacks. ResNet-34 exhibits near-total collapse across all attack types, with a 100% drop under PGD and 93% under FGSM. DenseNet-121 shows minimal degradation from pixel-wise perturbations, but suffers a 40% Top-1 drop under the patch attack, demonstrating the generalization strength of spatial adversaries.

Overview and Summary of Findings

In this project, we investigated the adversarial robustness of image classification models by systematically evaluating three attack strategies on a pretrained ResNet-34 net-

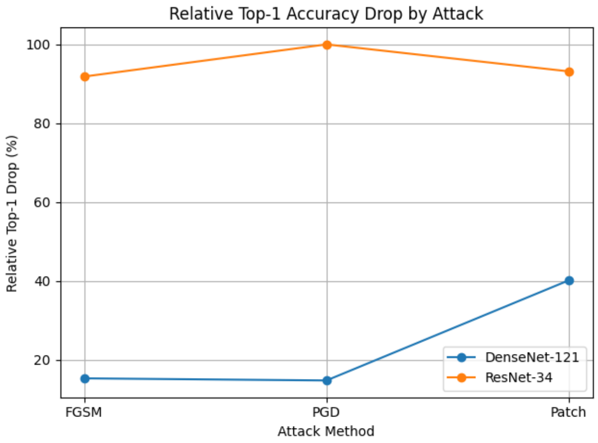

Figure 6: Relative Top-1 accuracy drop for ResNet-34 and DenseNet-121 under different attacks.

work. Our attacks included the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and a localized learnable patch attack. All evaluations were conducted on a 100-class subset of the ImageNet-1K dataset, allowing for high-resolution benchmarking while maintaining computational tractability.

Our approach focused on comparing the effectiveness and transferability of different perturbation types under consistent budget constraints. We measured Top-1 and Top-5 accuracy degradation, analyzed architectural susceptibility, and evaluated cross-model transfer using DenseNet-121.

Key findings include:

- **PGD** was the most effective attack on the source model, reducing ResNet-34 Top-1 accuracy to **0.00%**.
- **FGSM**, while less potent, executed rapidly and retained moderate transferability.
- The **patch attack** showed the highest generalization, reducing DenseNet-121 Top-1 accuracy to **45.60%** despite being optimized on ResNet-34.
- Top-5 accuracy trends reinforced these results, with patch-based perturbations exhibiting broad suppression across both architectures.

These results demonstrate the need for stronger defenses against both gradient-aligned and spatially localized perturbations. Our evaluation also highlights how model architecture influences susceptibility to different attack modalities.

Conclusion

This work presents a comparative evaluation of adversarial robustness in image classifiers using three canonical attacks—FGSM, PGD, and a learnable patch—on ResNet-34 and DenseNet-121, tested on a 100-class subset of ImageNet-1K. Our experiments reveal that even imperceptible, bounded perturbations can significantly degrade model performance, with PGD reducing ResNet-34's Top-1 accuracy to 0.00% and patch attacks exhibiting strong transferability to DenseNet-121.

While ResNet-34 is highly susceptible to gradient-based attacks, DenseNet-121 demonstrates greater resilience to pixel-wise perturbations, though it remains vulnerable to localized, structure-agnostic patch attacks. These results confirm that architecture plays a critical role in shaping adversarial robustness, and that transferable attacks pose a significant threat even in black-box settings.

Our study highlights the importance of evaluating multiple attack modalities under consistent perturbation budgets. Future work should focus on hybrid defense strategies that incorporate adversarial training across diverse attack types, preprocessing-based input sanitization, and ensemble methods to reduce vulnerability across architectures. Addressing the challenge of transferable adversaries remains essential for deploying reliable deep learning systems in security-sensitive environments.

Acknowledgements

We would like to acknowledge OpenAI's GPT-4.0 language model, referred to as ChatGPT, for providing assistance with generating content, and the Quillbot tool for grammar checking and paraphrasing.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*, 2014.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[5] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial Patch. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.