Adversarial Vulnerabilities of Deep Image Classifiers

Team Members:

Ruchi Jha

Neha Patil

Satvik Upadhyay

https://github.com/SatvikUpadhyay/Project3/tree/main

Abstract

Deep neural networks have achieved state-of-the-art performance on a wide range of visual recognition tasks, yet they remain vulnerable to adversarial perturbations—carefully crafted inputs that cause incorrect predictions while appearing nearly identical to original inputs. In this project, we investigate the robustness of a production-grade ResNet-34 model pretrained on ImageNet-1K. We design and evaluate multiple adversarial attacks, including the Fast Gradient Sign Method (FGSM), iterative Projected Gradient Descent (PGD), and a localized patch-based variant. Our attacks adhere to strict L_∞ or localized pixel budget constraints while significantly degrading model performance. We further analyze the transferability of these adversarial examples to another pretrained model, DenseNet-121. Our results show that PGD and patch-based attacks can reduce Top-1 accuracy to 0% on ResNet-34, while also partially transferring to DenseNet-121. These findings highlight the fragility of deep models and the importance of adversarial robustness for real-world deployment.

Introduction

Deep convolutional neural networks have revolutionized computer vision, yet their susceptibility to adversarial perturbations poses critical risks in safety-critical applications. These perturbations, though often imperceptible to humans, can drastically reduce model accuracy. This vulnerability compromises model reliability in fields such as autonomous driving, medical imaging, and content moderation.

In this work, we conduct a comprehensive evaluation of adversarial robustness on a pretrained ResNet-34 model using a subset of the ImageNet-1K dataset. We develop a series of constrained attacks under both L-infinity (pixel-wise) and L0 (patch-wise) threat models, measuring their impact on model performance. Furthermore, we explore attack transferability by testing adversarial examples on a second model, DenseNet-121. Our primary goals are:

- Demonstrate significant accuracy degradation using constrained adversarial attacks.
- Analyze visual similarity and perceptibility of perturbed images.
- Quantify attack transferability across architectures.

Overview

Our approach to designing adversarial attacks followed a progressive and constraint-aware structure. We began with the Fast Gradient Sign Method (FGSM) as a computationally simple and interpretable baseline, and then extended to the more powerful Projected Gradient Descent (PGD), before concluding with a localized patch-based PGD attack. Each method was evaluated under strict $L\infty$ norm constraints, ensuring that perturbations remained imperceptible or spatially limited, depending on the threat model.

In FGSM and full-image PGD, we selected $\varepsilon = 0.02$ to restrict per-pixel perturbation magnitude. This value is widely accepted in adversarial literature for normalized ImageNet inputs and ensures changes remain visually indistinguishable. FGSM, being a one-step method, used the entire budget at once, while PGD distributed it over 20 iterative steps with a step size $\alpha = 0.001$. This allowed finer control of the perturbation process, reducing overshooting and making the attack more targeted. As expected, PGD significantly outperformed FGSM in degrading model accuracy due to its iterative optimization of the adversarial direction.

For the patch-based PGD attack, the design had to accommodate a more constrained and realistic scenario—perturbing only a 32×32 region of the input. To offset the reduced spatial footprint, we increased the total perturbation budget to $\varepsilon = 0.8$ and set the step size to $\alpha = 0.08$ over 30 iterations. These values were empirically tuned: lower ε or fewer steps yielded insufficient attack strength, while higher values risked introducing visible artifacts. The patch location was randomized per image,

simulating physical-world attacks such as sticker-based occlusion. During optimization, gradients were masked outside the patch, and perturbations were projected and clamped within the $L\infty$ ε-ball. This setup successfully maintained the spatial and perceptual constraints while causing complete failure in model classification, achieving 0.00% Top-1 and Top-5 accuracy on ResNet-34.

Through these experiments, we learned that even localized perturbations, when properly intensified, can fully compromise deep classifiers. We also observed that iterative attacks provide significantly stronger degradation than single-step ones, albeit at a higher computational cost. Importantly, while global attacks like FGSM and PGD transferred reasonably well to a different architecture (DenseNet-121), patch-based attacks showed limited transferability—suggesting that spatially localized perturbations tend to exploit model-specific vulnerabilities rather than universally fragile patterns.

In summary, our methodology emphasized a balance between theoretical rigor, perceptual realism, and empirical effectiveness, enabling a comprehensive understanding of adversarial attack strategies across multiple axes of design.

Methodology

Our methodology consisted of evaluating and attacking a pretrained ResNet-34 model using a structured, multi-stage pipeline. We incrementally developed adversarial attacks, beginning with a baseline evaluation on clean data, followed by a fast gradient-based attack (FGSM), an iterative gradient method (PGD), a localized patch attack, and finally an analysis of transferability across models. All experiments were conducted on a subset of 500 images from 100 ImageNet-1K classes. Each image was normalized using standard ImageNet preprocessing parameters, with means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225]. Labels were mapped using a provided JSON file to ensure consistency in evaluating predictions.

Task 1: Clean Evaluation

To establish a performance baseline, we evaluated the pretrained ResNet-34 model on the clean test dataset. The model, loaded with ImageNet-1K weights and set to evaluation mode on GPU, processed 500 images from 100 ImageNet classes. Images were normalized using standard ImageNet mean and std, and evaluated in mini-batches of size 32 with 2 data loading workers. For each batch, a forward pass produced logits, and top predictions were matched to ground-truth labels using a reference index. This setup involved no adversarial perturbations and captured the model's natural performance. ResNet-34 achieved a Top-1 accuracy of 76.00% and Top-5 accuracy of 94.20%, forming the benchmark for all subsequent attacks.

Task 2:FGSM Attack

In Task 2, we implemented the Fast Gradient Sign Method (FGSM), a single-step adversarial attack that perturbs each input image in the direction of the gradient of the loss function with respect to the input. For each image, we computed the cross-entropy loss and backpropagated to obtain the gradient ∇xL. The adversarial image was then generated by adding ε times the sign of this gradient to the original image. We set $\varepsilon = 0.02$, which corresponds to a maximum perturbation of approximately ±1 in the unnormalized image scale. This choice of ε ensures imperceptibility while still enabling the model to be misled. The resulting adversarial images were clipped to remain within the valid [0, 1] range and then re-normalized using ImageNet statistics. All generated adversarial examples satisfied the L_∞ constraint of $\varepsilon = 0.02$. The FGSM attack successfully reduced Top-1 accuracy to 3.00% and Top-5 accuracy to 19.00%, demonstrating its effectiveness as a baseline attack method.

Task 3:PGD-Like Iterative Attack

Task 3 aimed to strengthen the adversarial attack using Projected Gradient Descent (PGD), an iterative improvement over FGSM. Unlike FGSM's single-step perturbation, PGD applies multiple small steps, allowing for a more optimized path through the loss landscape. For each image, we initialized the adversarial example as the original and updated it over 20 steps. In each step, we computed the loss gradient and applied a perturbation in the direction of the gradient sign. The step size α was set to 0.001, based on a total budget of $\varepsilon = 0.02$. After each update, the image was projected back into the ε-bounded $L\infty$ space and pixel values were clamped to [0, 1], followed by normalization. These constraints kept perturbations imperceptible while enabling effective exploration of high-loss regions. Compared to FGSM, PGD significantly degraded performance, reducing Top-1 accuracy to 1.50% and Top-5 accuracy to 9.90%, demonstrating the superior power of iterative attacks under constrained settings.

Task 4:Localized Patch Attack (Patch PGD)

In Task 4, we explored a more constrained and physically plausible adversarial setting using a localized patch attack. Unlike previous global perturbations, this method limited manipulation to a 32×32 region randomly located in each image. We extended the PGD framework with a spatial mask that zeroed out gradients outside the patch. At each iteration, gradients were computed and applied only within this masked region. To compensate for the reduced attack area, we increased the budget to $\varepsilon = 0.5$,

allowing stronger perturbations while maintaining perceptual plausibility. The attack ran for 20 steps with a step size of $\alpha = 0.025$. After each update, the patch was projected back to the ε-bounded $L\infty$ space, and pixel values were clamped to remain valid. Random patch placement simulated real-world occlusion-based attacks. Despite the spatial constraint, the attack was highly effective: both Top-1 and Top-5 accuracy dropped to 0.00%, showing that well-calibrated local attacks can rival global ones in impact.

Task 5:Transferability to DenseNet-121

In the final task, we investigated the transferability of adversarial examples by evaluating all four datasets—clean, FGSM, PGD, and patch-based attacks—on a different deep learning architecture: DenseNet-121. This model was selected because its connectivity pattern is fundamentally different from that of ResNet-34, allowing us to assess whether perturbations crafted for one architecture could effectively mislead another. Using the same preprocessing and evaluation loop, we passed each dataset through the DenseNet-121 model without any retraining or fine-tuning, and recorded Top-1 and Top-5 classification accuracy for each case. The label mapping was consistent with previous tasks, ensuring comparability.

The clean dataset yielded a Top-1 accuracy of 74.80% and Top-5 accuracy of 93.60%, nearly identical to ResNet-34's performance, confirming the reliability of the evaluation. When adversarial examples generated via FGSM and PGD were tested on DenseNet-121, we observed significant accuracy degradation, with Top-1 scores dropping to 39.20% and 32.40%, respectively. This suggests that perturbations aligned with the global gradient direction retain a moderate degree of model-agnosticity. However, the patch-based PGD attack—despite being highly destructive to ResNet-34—had only a minimal effect on DenseNet-121, reducing Top-1 accuracy to 69.80%. This highlights that localized adversarial perturbations may exploit architecture-specific spatial patterns and fail to generalize effectively across models. Throughout this task, all L-infinity and spatial localization constraints from earlier stages were strictly enforced, ensuring that cross-model results reflected genuine transferability limitations rather than inconsistencies in perturbation strength or scope.

Through these carefully structured tasks and constraint-aware implementations, we demonstrated the fragility of even production-grade models like ResNet-34 under various attack paradigms, and we highlighted the importance of considering both attack strength and perceptual subtlety when evaluating model robustness.

Summary Findings

xOur experiments confirm that deep image classifiers like ResNet-34, despite their strong performance on clean data, are extremely sensitive to adversarial perturbations, even when those perturbations are constrained in magnitude and perceptual visibility. The FGSM attack, while fast and simple, led to over 70% degradation in classification accuracy. The iterative PGD attack proved even more powerful, achieving over 84% drop in Top-5 accuracy under the same ε constraint. Surprisingly, even when perturbations were confined to a small 32×32 patch of the image, the model's performance collapsed entirely, highlighting the risk posed by localized, physical-world adversarial patterns. Transferability analysis showed that DenseNet-121 was also vulnerable to these attacks, though spatially localized adversarial examples did not generalize as well as global ones. These findings highlight both the strength of gradient-based attack methods and the fragility of large-scale vision models, reinforcing the urgent need for adversarial defenses in real-world applications.

Result

We evaluated the effectiveness of each adversarial attack by comparing the Top-1 and Top-5 accuracy of the model on the clean dataset and the three perturbed variants. On the unmodified dataset, the pretrained ResNet-34 model achieved a Top-1 accuracy of 76.00% and a Top-5 accuracy of 94.20%. This served as our performance benchmark for measuring the degradation caused by each adversarial strategy.

The FGSM attack, using a one-step update with an ε value of 0.02, caused a significant drop in performance. The Top-1 accuracy fell to 3.00%, while the Top-5 accuracy decreased to 19.00%. This represents a drop of 73.00% in Top-1 and 75.20% in Top-5 accuracy from the baseline. Despite being computationally simple and fast to execute, FGSM proved highly effective in misleading the model.

Our PGD-like iterative attack, using the same $\varepsilon = 0.02$ but applied over 20 steps with a step size of $\alpha = 0.001$, resulted in even greater performance degradation. The Top-1 accuracy dropped to 1.50%, and Top-5 accuracy was reduced to 9.90%, showing a 74.50% and 84.30% decline respectively. The additional iterative refinement allowed the attack to more effectively traverse high-loss regions of the input space, producing stronger adversarial examples while remaining within the same perceptual constraint.

The patch-based PGD attack further highlighted the vulnerability of the model to spatially localized

adversarial inputs. In this attack, perturbations were restricted to a 32×32 region of each image, and ε was increased to 0.5 to compensate for the reduced area of influence. Despite this spatial limitation, the attack was devastatingly effective, reducing both Top-1 and Top-5 accuracy to 0.00%. This result indicates that even partial image manipulation—if sufficiently intense—can fully compromise model reliability.

To assess cross-model generalization, we evaluated the same datasets on DenseNet-121, a different architecture trained on the same ImageNet-1K distribution. On the clean dataset, DenseNet-121 performed similarly to ResNet-34, achieving a Top-1 accuracy of 74.80% and a Top-5 accuracy of 93.60%. When tested with FGSM-generated adversarial examples, the Top-1 accuracy dropped to 39.20%, and Top-5 to 67.80%, showing a moderate transferability with a 35.60% reduction in Top-1 performance. PGD examples yielded slightly lower performance, with a Top-1 accuracy of 32.40% and the same Top-5 accuracy of 67.80%, indicating that iterative attacks generalized somewhat better than single-step ones. However, the patch-based adversarial examples transferred poorly, with DenseNet-121 retaining 69.80% Top-1 and 89.60% Top-5 accuracy. This suggests that while global pixel-wise perturbations tend to generalize across architectures, spatially localized attacks often exploit model-specific characteristics and fail to transfer effectively.

Overall, our results demonstrate that even state-of-the-art image classifiers like ResNet-34 and DenseNet-121 are highly vulnerable to adversarial manipulation, with iterative and patch-based attacks capable of inducing near-complete performance collapse under controlled constraints.

Final Results Summary

ResNet-34

Attack	Top-1 Acc	Top-5 Acc	ε	α	Steps	Patch	Time/Batch
Clean	76.00%	94.20%	—	—	—	—	~1–2 sec
FGSM	3.00%	19.00%	0.02	—	1	—	~2–3 sec
PGD	1.50%	9.90%	0.02	0.001	20	—	~20–25 sec
Patch PGD	0.00%	0.00%	0.8	0.08	30	32×32	~30–35 sec

DenseNet-121

Attack	Top-1 Acc	Top-5 Acc	ε	α	Steps	Patch	Time/Batch
Clean	74.80%	93.60%	—	—	—	—	~1–2 sec
FGSM (transfer)	39.20%	67.80%	0.02	—	1	—	~2–3 sec
PGD (transfer)	32.40%	67.80%	0.02	0.001	20	—	~20–25 sec
Patch PGD (transfer)	69.80%	89.60%	0.8	0.08	30	32×32	~30–35 sec

Figure 1. Summarized Result

The two figures visualize the effect of adversarial attacks on ResNet-34 using FGSM Figure 2[1], PGD Figure 3[2], and Patch PGD Figure 4[3]. In each row, the original image is shown on the left, the adversarial version in the middle, and a bar plot of Top-5 predictions

Conclusion

This project highlights the significant vulnerabilities of modern deep learning models to adversarial attacks, even under strict perceptual constraints. Through a series of carefully designed experiments on ResNet-34, we demonstrated that both pixel-wise and localized patch-based perturbations can drastically degrade model performance. The iterative PGD attack proved particularly effective, and the localized patch attack achieved a complete breakdown despite targeting only a small image region. Transferability tests on DenseNet-121 revealed that global attacks generalize better across architectures, while localized ones remain more model-specific. These findings reinforce the importance of integrating adversarial robustness into model development and evaluation pipelines, especially for real-world applications where reliability is critical.

References

For this model to work, we took some help from LLMs like ChatGPT. But the underlying model was provided, and the key tweaks and methodologies were all our work.

[1]https://github.com/RuchiMJha/Adversarial-Vulnerabilities-of-Deep-Image-Classifiers/blob/main/Task_FGSM.jpeg

[2]https://github.com/RuchiMJha/Adversarial-Vulnerabilities-of-Deep-Image-Classifiers/blob/main/Task_PGD.png

[3]https://github.com/RuchiMJha/Adversarial-Vulnerabilities-of-Deep-Image-Classifiers/blob/main/Task_Patched_PGD.png

[4]Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples*. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6572

[5] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards Deep Learning Models Resistant to Adversarial Attacks*. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1706.06083