Project Report: Adversarial Attacks and Transferability in Deep Image Classifiers

Aryaman Singh Dev, Nevin Mathews Kuruvilla, Rohan Subramaniam
New York University - Tandon School of Engineering
{asd9884, nm4709, rs9194}@nyu.edu
Github repository:
https://github.com/roncell/resnet34-vulnerability-analysis/tree/main

Abstract

This project investigates the vulnerability of deep image classifiers to adversarial attacks, focusing on a pretrained ResNet-34 model evaluated on a subset of the ImageNet-1K dataset. We implement and evaluate three types of attacks: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and targeted patch attacks. Each attack is constrained under an L_∞ budget, with perturbations visually imperceptible yet highly effective at degrading classification performance. FGSM reduces ResNet-34's top-1 accuracy from 70.40% to 5.00%, while PGD pushes it to 0.00%. We further explore the transferability of these attacks to DenseNet-121, observing moderate transfer performance, especially for global perturbations. Our results highlight both the fragility and cross-model vulnerability of modern deep vision systems under constrained adversarial conditions.

Introduction

Deep neural networks have achieved remarkable success in image classification tasks, particularly on large-scale datasets like ImageNet-1K. However, despite their impressive accuracy, these models are known to be highly vulnerable to adversarial attacks—small, carefully crafted perturbations to input images that cause the model to misclassify with high confidence. These perturbations are often imperceptible to the human eye, yet they expose a fundamental weakness in the generalization and robustness of deep models.

This project focuses on evaluating and designing adversarial attacks on a production-grade ResNet-34 model trained on ImageNet-1K. Our goal is to construct attacks that significantly degrade model performance under strict perturbation constraints, specifically L_∞ and patch-based constraints. We implement the Fast Gradient Sign Method (FGSM), a one-step pixel-wise attack; Projected Gradient Descent (PGD), a multi-step iterative variant; and a targeted patch-based PGD attack limited to a 32×32 region.

We further assess the transferability of these attacks to a different architecture - DenseNet-121 to understand whether vulnerabilities are specific to model structure or more broadly shared across vision networks. By analyzing top-1 and top-5 accuracy drops and comparing adversarial and clean performance, we aim to characterize the effectiveness, subtlety, and generalizability of each attack method.

This work provides a thorough empirical exploration of adversarial attack methods and their impacts, reinforcing the importance of robustness evaluation in modern deep learning systems.

Related Work

Adversarial attacks have been extensively studied in the context of deep neural networks. (1) first demonstrated that imperceptible perturbations could drastically change a model's prediction. This discovery led to the development of stronger gradient-based attacks such as the Fast Gradient Sign Method (FGSM) introduced by (2), which perturbs input data using the gradient of the loss function with respect to the input. Building on this, (3) proposed Projected Gradient Descent (PGD), a multi-step attack considered one of the most effective first-order adversarial attacks.

In addition to global perturbation attacks, localized patch attacks have emerged as a practical threat model. Techniques such as adversarial patches (4) and localized PGD variants manipulate small image regions to induce targeted misclassification, often without requiring access to the full image or model internals.

Transferability of adversarial examples across models is another well-documented phenomenon (5), highlighting the potential for black-box attacks. Understanding this cross-model vulnerability is crucial for evaluating the robustness of deep learning systems in real-world scenarios where attackers may not have full knowledge of the model architecture or weights.

Our project builds on these foundational works by applying FGSM, PGD, and patch-based attacks on ResNet-34 and evaluating their transferability to DenseNet-121, with a focus on both performance degradation and practical applicability.

Methodology

Model and Dataset

We use a pretrained `ResNet-34` model available through PyTorch's torchvision package, trained on the ImageNet-1K dataset. For evaluation, we create a custom subset of 100 ImageNet classes with balanced representation across

classes. Images are normalized using standard ImageNet statistics and resized to 224×224 for compatibility with the model's input dimensions.

Baseline Evaluation

The model is first evaluated on the unperturbed dataset to establish baseline Top-1 and Top-5 classification accuracies. We use the official ImageNet class indices to ensure label consistency and correctness. Our baseline results serve as a reference for quantifying the impact of each adversarial attack.

FGSM Attack

The Fast Gradient Sign Method (FGSM) introduces a single-step perturbation $\delta = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y))$ added to the input image x, where ϵ is the attack budget and $\mathcal{L}$ is the cross-entropy loss. The perturbation is constrained under the L_∞ norm, ensuring imperceptibility. We evaluate the model on the resulting adversarial images (Adversarial Test Set 1) to measure performance degradation.

PGD Attack

We implement Projected Gradient Descent (PGD), an iterative variant of FGSM. In each iteration, the image is perturbed using FGSM and projected back to the ϵ-bounded L_∞ ball around the original image. The update rule is: $x^{t+1} = \Pi_{x+\mathcal{B}_\epsilon}(x^t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(x^t, y)))$, where α is the step size and Π is the projection operator. The resulting images form Adversarial Test Set 2.

Targeted Patch PGD Attack

To simulate more practical and constrained threat models, we design a targeted PGD attack applied to a random 32×32 patch within each image. For each image, a random target class is selected (different from the true label), and the attack maximizes the model's confidence in this incorrect class. The patch is optimized iteratively under a larger ϵ budget, generating Adversarial Test Set 3.

Transferability Evaluation

To study cross-model generalization of adversarial perturbations, we evaluate all three adversarial datasets on a different pretrained model, `DenseNet-121`. This allows us to assess the transferability of adversarial examples crafted on ResNet-34 to another architecture with similar capacity but different inductive biases.

Experiments and Results

We evaluate adversarial robustness using three attack methods—FGSM, PGD, and patch-based PGD—on a pretrained ResNet-34 model, using a 100-class subset of the ImageNet-1K dataset. All attacks are constrained under an L_∞ norm with $\epsilon = 0.02$ for FGSM and PGD, and $\epsilon = 0.5$ for the patch attack.

Baseline Accuracy

On the clean dataset, ResNet-34 achieves a Top-1 accuracy of 70.40% and Top-5 accuracy of 91.40%, consistent with expected performance on ImageNet-class subsets.

FGSM Attack (Task 2)

FGSM reduces the Top-1 accuracy to 5.00% and Top-5 accuracy to 19.80%. Visualization confirms that perturbations are imperceptible yet highly effective. This attack satisfies the task requirement of a 50% drop in Top-1 accuracy.

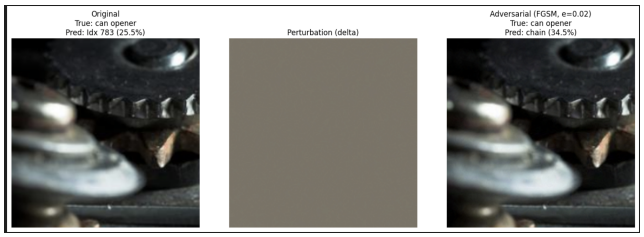

Figure 1: FGSM example: original image (can opener), imperceptible perturbation, adversarial prediction (chain, 34.5%).

PGD Attack (Task 3)

PGD further degrades performance, achieving 0.00% Top-1 accuracy and 6.20% Top-5 accuracy, meeting the 70% drop requirement. Compared to FGSM, PGD performs iterative updates and projects back into the L_∞ ball, resulting in stronger perturbations.

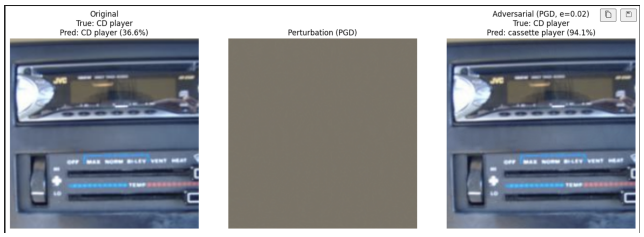

Figure 2: PGD example: original image (CD player), perturbation, adversarial prediction (cassette player, 94.1%).

Patch Attack (Task 4)

A targeted patch PGD attack is applied on a 32×32 region. With $\epsilon = 0.5$, it drops Top-1 accuracy to 58.80% and Top-5 to 83.20%. Although less effective than full-image attacks, patch attacks offer a more stealthy and realistic threat model.

Transferability to DenseNet-121 (Task 5)

We assess attack transferability by testing adversarial examples from ResNet-34 on DenseNet-121:

- FGSM: Top-1 = 58.80%, Top-5 = 85.00%
- PGD: Top-1 = 58.40%, Top-5 = 86.40%
- Patch PGD: Top-1 = 69.40%, Top-5 = 90.20%

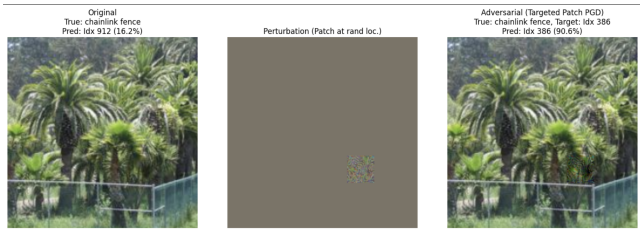

Figure 3: Targeted patch attack: localized adversarial patch causes misclassification while leaving most of the image intact.

Figure 4: Example of high-confidence PGD misclassification: *altar* reclassified as *church* with 98.3% confidence.

These results show that global perturbations (FGSM, PGD) transfer better than localized ones. However, none fully degrade DenseNet's accuracy, illustrating model-specific resilience.

Analysis and Discussion

Our results show that adversarial attacks can drastically degrade model performance under imperceptible perturbations. The PGD attack, in particular, caused ResNet-34's Top-1 accuracy to drop from 70.40% to 0.00%, demonstrating the vulnerability of gradient-based optimization methods to iterative perturbations.

Comparison Across Attacks

- **FGSM** caused significant drops in Top-1 accuracy (down to 5.00%), yet required only one gradient step.
- **PGD** performed stronger attacks with multiple iterations, leading to complete misclassification.
- **Patch PGD** was more constrained, and the accuracy drop was smaller (to 58.80%)—as expected due to its localized nature.

Transferability to DenseNet-121

We observed moderate transferability. PGD and FGSM examples generated on ResNet-34 also fooled DenseNet-121, reducing its Top-1 accuracy from 70.80% to 58.40% and 58.80%, respectively. However, patch-based attacks transferred poorly, with DenseNet-121 retaining 69.40% Top-1 accuracy.

Confidence in Misclassifications

Interestingly, adversarial images often result in **high confidence misclassifications**. This suggests models are confidently wrong under adversarial noise—raising safety concerns in critical applications.

Limitations

Despite strong results, this study has several limitations:

- **Dataset Size:** We used a 100-class subset of ImageNet-1K, which may not generalize to all 1000 classes.
- **Patch Attack Constraints:** Patch locations were chosen randomly. Optimized placement could increase effectiveness.
- **Model Specificity:** All attacks were crafted using ResNet-34. Transfer results could vary for other architectures.

Conclusion and Future Work

We demonstrate that deep vision models remain vulnerable to imperceptible and localized adversarial perturbations. While PGD attacks showed the highest effectiveness, patch-based attacks offer stealth and physical plausibility. The partial transferability to DenseNet-121 reinforces concerns about model-agnostic vulnerability.

In future work, we aim to:

- Explore certified defenses against adversarial attacks.
- Apply adversarial training and evaluate trade-offs.
- Investigate physical-world adversarial patch deployment.

References

[1] 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*.

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

[3] 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)*.

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

[5] Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.