Project 3 - Jailbreaking Deep Models

Abhishek Adinarayanappa (aa12037), Harsha Mupparaju (sm12754), Nived Damodaran (nd2746)

GitHub Repository Link: https://github.com/HarshaMupparaju/jail-breaking-deep-models

Abstract

Abstract

Deep neural networks are known to be vulnerable to adversarial examples—inputs crafted to fool models while remaining nearly indistinguishable from real data. In this project, we evaluate the robustness of a pretrained ResNet-34 model on ImageNet-1K under various adversarial attacks on production grade, publicly posted models, and degrade their performance. We implement Fast Gradient Sign Method (FGSM), Iterative FGSM, and localized patch attacks. We assess the transferability of these attacks by evaluating them on a DenseNet-121 model. Our results demonstrate that well-crafted adversarial examples can reduce ResNet-34 top-1 accuracy from 76.00% to as low as 0.20%, while maintaining imperceptible visual changes. However, transferability to DenseNet is limited, with top-1 accuracy degrading only marginally. We conclude with insights into attack strength, architecture vulnerability, and mitigation techniques.

Introduction

Deep learning models have achieved state-of-the-art performance across a wide range of computer vision tasks, including image classification, object detection, and semantic segmentation. Their success is largely attributed to large-scale datasets, deep architectures, and efficient training algorithms. However, recent studies have revealed a critical vulnerability in these models: they are highly susceptible to adversarial examples—inputs that have been deliberately and imperceptibly altered in a way that causes the model to produce incorrect predictions, even though the changes are virtually indistinguishable to the human eye.

This project investigates the adversarial robustness of the ResNet-34 model trained on the ImageNet-1K dataset, a benchmark widely used to evaluate large-scale image classification performance. We focus on generating white-box adversarial attacks that perturb the input within an L_∞ norm constraint (with $\epsilon = 0.02$), ensuring that the resulting adversarial images remain visually similar to the original inputs. These perturbations are designed to maximally degrade classification accuracy while staying within a tightly bounded perceptual distortion.

Beyond global perturbations, we also explore localized patch attacks, where adversarial noise is confined to a small region of the image, simulating more realistic and potentially stealthy attack scenarios. Furthermore, we evaluate the transferability of adversarial examples by testing whether perturbations crafted for ResNet-34 can mislead a different model architecture—DenseNet-121—without modification. This cross-model vulnerability highlights the general weaknesses shared by modern deep networks.

Our results demonstrate that even high-performing architectures like ResNet and DenseNet can be rendered unreliable under adversarial conditions, raising significant concerns for deploying these models in safety-critical applications. The findings underscore the importance of developing more robust training techniques and effective defense mechanisms to mitigate adversarial threats in real-world systems.

Methodology

Task 1: Baseline Evaluation

We evaluated a pretrained ResNet-34 model on a subset of ImageNet-1K (100 classes). Each input image was resized and normalized using the standard ImageNet preprocessing pipeline, which involves mean subtraction and standard deviation scaling based on the dataset's RGB channel statistics. The preprocessed images were then passed through the ResNet-34 model in inference mode, without any additional fine-tuning or augmentation.

The model achieved a top-1 accuracy of 76.00%, indicating that in 76% of the test cases, the model's highest-confidence prediction matched the ground-truth label. The top-5 accuracy was 94.20%, meaning that in 94.2% of the cases, the correct label was among the model's five most confident predictions.

Task 2: FGSM Attack

To assess the vulnerability of the ResNet-34 model to adversarial perturbations, we applied the Fast Gradient Sign Method (FGSM). FGSM perturbs each input image in the direction that maximally increases the model's prediction loss. Specifically, given an input image x with its true label y, the

Figure 1: Original Images (Task 1)

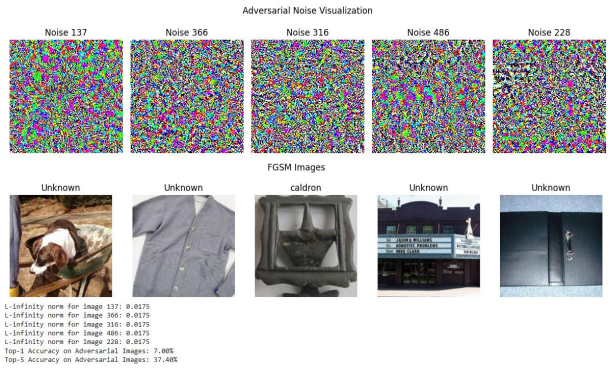

Figure 2: FGSM Attack Output (Task 2)

adversarial example x' is computed as:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y))$$

where $\mathcal{L}$ is the cross-entropy loss. We chose the value $\epsilon = 0.02$. This created Adversarial Test Set 1. Top-1 accuracy dropped to 7.00%, and top-5 to 37.40%.

Task 3: Iterative Attack

To strengthen the adversarial attack beyond the single-step FGSM, we employed the Iterative Fast Gradient Sign Method (I-FGSM), also known as Basic Iterative Method (BIM). Unlike FGSM, which perturbs the input only once, I-FGSM applies multiple small perturbation steps, allowing the attack to refine its optimization of the loss landscape. This iterative approach is particularly effective in escaping local minima and crafting stronger adversarial examples.

The adversarial images were generated by repeatedly applying FGSM updates with a smaller step size $\alpha = 0.005$ over 5 iterations, while ensuring that the total perturbation remained within the L_∞ constraint of $\epsilon = 0.02$. This formed Adversarial Test Set 2, reducing top-1 accuracy to 0.20% and top-5 accuracy to 16.80%

Task 4: Patch Attack

In this task, we investigated a more stealthy form of adversarial perturbation by localizing the attack to a small region of the image rather than modifying the entire input. Specifically, we applied an iterative FGSM attack confined to a fixed-size 32×32 pixel patch within the input image.

The patch location was randomized for each image, ensuring that the attack was location-agnostic. Within the selected region, we applied 5 iterations of the FGSM update rule with

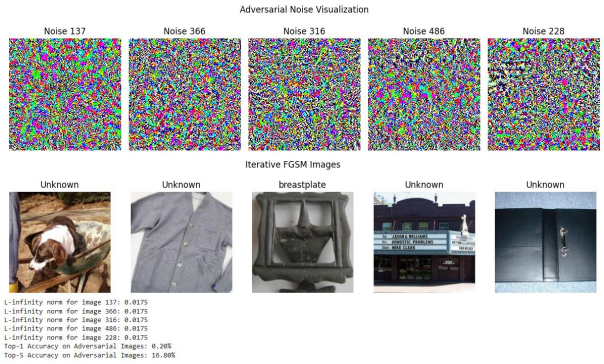

Figure 3: Iterative Attack Output (Task 3)

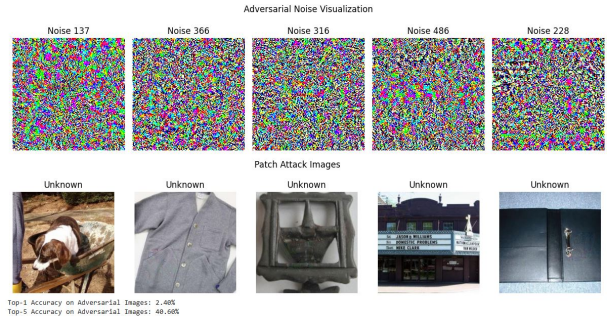

Figure 4: Patch Attack Output (Task 4)

a larger perturbation budget of $\epsilon = 0.3$, This formed Adversarial Test Set 3, which reduced ResNet-34 top-1 accuracy to 2.40% and top-5 accuracy to 40.60%

Task 5: Transferability

To evaluate the transferability of adversarial examples, we tested all three adversarial test sets—FGSM, Iterative FGSM, and Localized Patch—on a different model architecture, DenseNet-121, without generating new perturbations specific to it. While ResNet-34 suffered drastic performance drops, DenseNet retained higher accuracy:

- FGSM: 65.20% (top-1), 89.80% (top-5)
- Iterative: 65.40% (top-1), 91.40% (top-5)
- Patch: 67.60% (top-1), 91.40% (top-5)

These results suggest that DenseNet-121 is less susceptible to adversarial examples transferred from ResNet-34, particularly under the L_∞-bounded attacks used in this study.

Results

Table 1 summarizes the classification accuracy of ResNet-34 and DenseNet-121 across all datasets.

Discussion

Our experiments show that even a simple single-step FGSM attack can significantly degrade ResNet-34's performance. Iterative attacks proved far more effective, reducing top-1

Dataset	Top-1 Acc.	Top-5 Acc.
Original (ResNet-34)	76.00%	94.20%
FGSM	7.00%	37.40%
Iterative FGSM	0.20%	16.80%
Patch Attack	2.40%	40.60%
Original (DenseNet)	74.80%	93.60%
FGSM Transfer	65.20%	89.80%
Iterative Transfer	65.40%	91.40%
Patch Transfer	67.60%	91.40%

Table 1: Top-1 and Top-5 accuracy across models and adversarial test sets.

accuracy to below 1%. Even localized patch attacks were impactful, highlighting the model's sensitivity to regional perturbations.

We observed limited transferability of adversarial examples. DenseNet-121 retained relatively high accuracy, suggesting model-specific vulnerabilities. This underscores the importance of evaluating adversarial robustness across multiple architectures.

Key lessons include:

- Iterative and localized attacks are more effective than FGSM.
- Transferability is constrained by architectural differences.
- Perturbation scaling in normalized space must strictly respect ϵ bounds.

Conclusion

Through systematic adversarial attacks, we demonstrated how fragile deep models like ResNet-34 can be to imperceptible image perturbations. While ResNet showed dramatic drops in accuracy, DenseNet-121 was relatively more robust to transferred attacks. These findings stress the need for defensive strategies like adversarial training and model ensembling. Future work may include evaluating on transformer-based models and exploring adaptive attacks under tighter constraints.

References

https://huggingface.co/docs/peft
https://arxiv.org/abs/2106.09685 (LoRA paper)
https://huggingface.co/docs/transformers
https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset