Adversarial Attacks on Deep Image Classifiers: A Study on ResNet-34

Chenke Wang
cw4565
Tianyu Liu
tl3191
Zhaochen Yang
zy2189

Abstract

This project investigates the vulnerability of the ResNet-34 model to various adversarial attacks, including the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), patch attacks, and their transferability to other models such as DenseNet-121 and ViT-B/16. Using a subset of the ImageNet-1K dataset, we demonstrate significant accuracy degradation with imperceptible perturbations. Our findings underscore the fragility of deep classifiers and emphasize the urgency of developing robust defenses.

Introduction

Deep neural networks excel in image classification but remain vulnerable to adversarial attacks—carefully crafted perturbations that mislead models into incorrect predictions. This project examines the effectiveness of multiple attack strategies on a pre-trained ResNet-34 model, using a test dataset from ImageNet-1K. We explore single-step (FGSM), iterative (PGD), and localized (patch) attacks, as well as their transferability across architectures, to highlight the security challenges in deep learning.

Project Codebase

The complete codebase for this project is available on GitHub:
`https://github.com/lingtouyangLeo/dl_project3.git`

Methodology

Task 1: Baseline Evaluation

We evaluated a pre-trained ResNet-34 model on a test dataset of 388 images from 100 ImageNet-1K classes. Images were preprocessed with resizing to 256 pixels, center cropping to 224×224, and normalization using mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. The model achieved a top-1 accuracy of 68.81% and a top-5 accuracy of 93.81%.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Task 2: Pixel-wise Attacks with FGSM

We implemented the Fast Gradient Sign Method (FGSM), a single-step attack that generates adversarial examples using:

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

where x' is the adversarial image, x is the original input, $\epsilon = 0.02$ controls perturbation magnitude, $\nabla_x J$ is the gradient of the loss function J with respect to x, and y is the true label. The perturbed images reduced top-1 accuracy to 22.16% and top-5 to 41.49%.

Task 3: Improved Attacks

We applied the Projected Gradient Descent (PGD) attack, an iterative refinement of FGSM. Using 10 iterations, $\epsilon = 0.02$, and step size $\alpha = 0.005$, adversarial examples were generated by repeatedly applying FGSM and clipping perturbations to the ϵ-ball. This reduced top-1 accuracy to 5.15% and top-5 to 15.46%.

Task 4: Patch Attacks

We implemented a patch attack, perturbing a 32×32 pixel patch using PGD with $\epsilon = 0.5$ and 40 iterations. The patch was randomly placed, targeting key image features. This dropped top-1 accuracy to 10.31% and top-5 to 25.77%.

Task 5: Transferring Attacks

We tested the transferability of adversarial examples generated for ResNet-34 (using FGSM, PGD, and patch attacks) on DenseNet-121 and ViT-B/16, evaluating their performance on the same dataset. Results indicate partial transfer success, detailed in the Results section.

Results

We summarize the performance of ResNet-34 under various attacks and the transferability to other models in Table 1. Figure 1 illustrates an FGSM adversarial example.

Condition	Top-1 Accuracy	Top-5 Accuracy
ResNet-34 Clean	68.81%	93.81%
ResNet-34 FGSM	22.16%	41.49%
ResNet-34 PGD	5.15%	15.46%
ResNet-34 Patch	10.31%	25.77%
DenseNet-121 Clean	68.04%	91.49%
DenseNet-121 FGSM (transfer)	33.76%	60.05%
DenseNet-121 PGD (transfer)	28.87%	55.15%
DenseNet-121 Patch (transfer)	24.74%	51.03%
ViT-B/16 Clean	88.14%	98.71%
ViT-B/16 FGSM (transfer)	57.47%	80.93%
ViT-B/16 PGD (transfer)	56.96%	79.90%
ViT-B/16 Patch (transfer)	47.16%	68.56%

Table 1: Summary of model accuracies under different conditions.

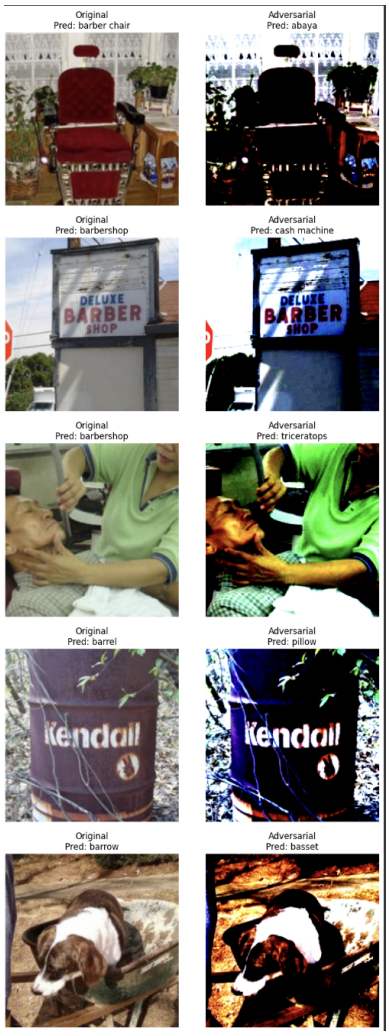

Figure 1: Example of an original image and its adversarial counterpart generated using FGSM with $\epsilon = 0.02$.

Additional Results and Analysis

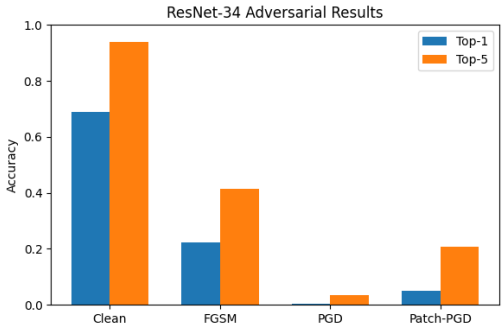

Figure 2: ResNet-34 Top-1 and Top-5 accuracy under different adversarial attacks (Clean, FGSM, PGD, Patch-PGD).

Figure 2 shows that ResNet-34's performance drops significantly under adversarial perturbations. FGSM reduces Top-1 accuracy from 68.81% to 22.16%, while PGD further decreases it to 5.15%. The Patch-PGD attack, despite being localized to a small area, also results in a major accuracy loss (10.31%). These results demonstrate the severe vulnerability of convolutional networks like ResNet-34 to both pixel-wise and patch-wise perturbations.

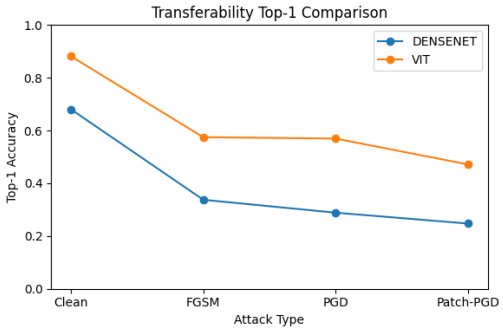

Figure 3: Top-1 accuracy transferability comparison between DenseNet-121 and ViT-B/16 under different adversarial attacks.

Figure 3 presents the transferability of adversarial attacks. DenseNet-121 experiences considerable performance degradation across all attack types, while ViT-B/16 maintains relatively higher accuracy. This suggests that transformer-based architectures like ViT-B/16 are more resilient to adversarial perturbations compared to traditional convolutional networks, likely due to their ability to capture long-range dependencies and global patterns.

Discussion

Our results reveal that ResNet-34's accuracy plummets under adversarial attacks, with PGD proving most effective (5.15% top-1) due to its iterative nature, followed by patch

attacks (10.31%) and FGSM (22.16%). The linearity hypothesis may explain this susceptibility, as small perturbations exploit the model's decision boundaries in high-dimensional space. Transferability results show DenseNet-121 suffers more significant degradation (e.g., 24.74% top-1 for patch attacks) than ViT-B/16 (47.16%), possibly due to ViT's transformer architecture capturing broader contextual features, enhancing robustness. These findings suggest shared vulnerabilities across architectures, critical for real-world applications like autonomous driving.

Conclusion

This study confirms ResNet-34's vulnerability to FGSM, PGD, and patch attacks, with accuracy drops from 68.81% to as low as 5.15% top-1. Transferability to DenseNet-121 and ViT-B/16 underscores the pervasive threat of adversarial examples. Future work could explore advanced attacks (e.g., CW), defenses like adversarial training, or robustness across diverse datasets.