Team Project
Deep Learning Project 3 Report
Jailbraking Deep Models

Written By Yumiko and Yihan

Github Repo: https://github.com/cyh1001/DL-Project3-Jailbreaking

Overview

In this project, we explore the brittleness of deep image classifiers by launching adversarial attacks on a pretrained ResNet-34 model trained on the ImageNet-1K dataset. We systematically evaluate and degrade the model's performance using both pixel-wise (L_∞) and patch-based (L_0-like) perturbations. Our goal is to reduce the model's top-1 and top-5 accuracy while keeping adversarial perturbations imperceptible. We also test the transferability of our attacks on a different model architecture, DenseNet-121. Through iterative improvements on attack strategies — beginning with FGSM and progressing to Projected Gradient Descent (PGD) and localized patch attacks—we demonstrate significant drops in classification accuracy, highlighting vulnerabilities in state-of-the-art vision models.

Code Repository

The code repository is available at this GitHub repository.

Summary of Findings

Our experiments yielded several key findings regarding the adversarial vulnerability of the ResNet-34 model and the transferability of these attacks to a DenseNet-121 model:

- The baseline ResNet-34 model achieved a top-1 accuracy of 70.60% and a top-5 accuracy of 93.20% on the clean test dataset.
- The Fast Gradient Sign Method (FGSM) with an $\epsilon = 0.02$ significantly degraded ResNet-34's performance, reducing its top-1 accuracy to 5.00% and top-5 accuracy to 30.20%. This met the project's initial 50% relative drop target.
- The Projected Gradient Descent (PGD) attack, using the same $\epsilon = 0.02$ but iterated for 10 steps with a step size $\alpha = 0.005$, proved substantially more potent. It drove ResNet-34's top-1 accuracy down to 0.00% and top-5 accuracy to 7.20%, far exceeding the 70% relative drop target.
- A targeted PGD patch attack, applied to a randomly placed 32x32 patch per image using $\epsilon_{patch} = 0.3$, 20 iterations, and $\alpha_{patch} = 0.06$, resulted in a more moderate performance drop on ResNet-34. Its top-1 accuracy decreased to 64.40% and top-5 accuracy to 86.40%.
- When evaluating transferability to DenseNet-121 (baseline top-1 accuracy of 70.80%, top-5 accuracy of 91.20% on the clean dataset):
  - Adversarial examples generated using FGSM against ResNet-34 reduced DenseNet-121's top-1 accuracy to 59.00% (top-5 to 84.80%).
  - Full-image PGD adversarial examples from ResNet-34 reduced DenseNet-121's top-1 accuracy to 59.60% (top-5 to 86.20%).
  - The patch-based PGD adversarial examples generated for ResNet-34 showed very limited transferability, only slightly reducing DenseNet-121's top-1 accuracy to 69.00% (top-5 to 90.60%).
- These findings underscore that while deep learning models are vulnerable to adversarial attacks, the effectiveness and transferability of these attacks vary significantly with the attack method and scope. Full-image attacks (FGSM, PGD) demonstrated notable transferability, whereas localized patch attacks were more model-specific.

These findings underscore the pervasive nature of adversarial vulnerabilities in deep learning models and illustrate differences in attack strength and transferability across various methods and scopes of perturbation.

Methodology

Dataset and Baseline

We used a curated test dataset containing 500 images from 100 classes of the ImageNet-1K dataset. Each image was preprocessed using standard ImageNet normalization statistics: mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$. A pretrained ResNet-34 model (`IMAGENET1K_V1`) from TorchVision was used as the target classifier.

Baseline performance on the test dataset:

- **Top-1 Accuracy:** 70.60%
- **Top-5 Accuracy:** 93.20%

FGSM Attack (Adversarial Test Set 1)

We implemented the Fast Gradient Sign Method (FGSM), which performs a one-step adversarial perturbation using the sign of the gradient:

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x \mathcal{L}) \quad (1)$$

where $\epsilon = 0.02$. The perturbations were cut to remain within the l_∞ constraint and visually verified to be imperceptible.

Performance after FGSM attack:

- **Top-1 Accuracy:** 5.00%
- **Top-5 Accuracy:** 30.20%

PGD Attack (Adversarial Test Set 2)

We applied Projected Gradient Descent (PGD), an iterative method using:

- $\epsilon = 0.02$
- Step size $\alpha = 0.005$
- 10 update steps

Each update was followed by projection back to the ℓ_∞ ball around the original image.

Performance after PGD attack:

- **Top-1 Accuracy:** 0.00%
- **Top-5 Accuracy:** 7.20%

Patch-based Attack (Adversarial Test Set 3)

For localized attacks, we applied PGD only to a 32x32 patch per image. The patch was randomly placed, and larger ϵ values were used due to the limited perturbation area.

- $\epsilon_{patch} = 0.3, 20$ iter, $\alpha_{patch} = 0.06$

Performance:

- **Top-1 Accuracy:** 64.40%
- **Top-5 Accuracy:** 86.40%

Results

The experimentation gave the results below. Refer to Figures 1-9 and Tables 1 and 2.

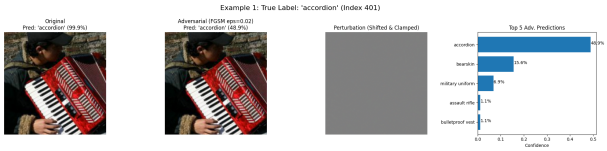

Figure 1: FGSM attack example 1

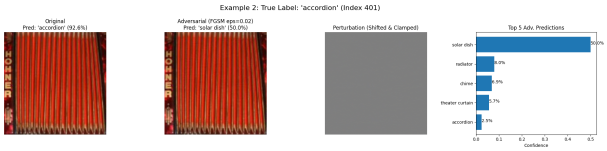

Figure 2: FGSM attack example 2

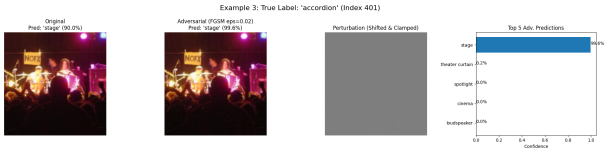

Figure 3: FGSM attack example 3

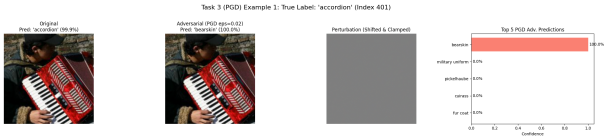

Figure 4: PGD attack example 1

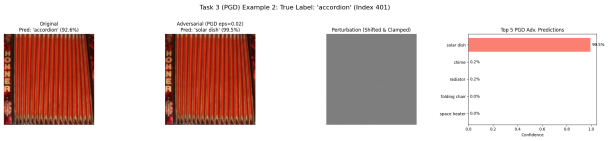

Figure 5: PGD attack example 2

Figure 6: PGD attack example 3

Figure 7: Patch-based attack example 1

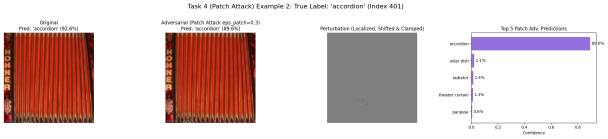

Figure 8: Patch-based attack example 2

Figure 9: Patch-based attack example 3

ResNet-34 Accuracy Summary

(Refer to Table 1)

Dataset	Top-1 Accuracy	Top-5 Accuracy
Original	70.60%	93.20%
Adversarial Test Set 1	5.00%	30.20%
Adversarial Test Set 2	0.00%	7.2%
Adversarial Test Set 3	64.40%	86.40%

Table 1: Accuracy of ResNet-34 on original and adversarial datasets.

Transferability: DenseNet-121

We evaluated the same datasets using DenseNet-121 to assess attack transferability: (Refer to Table 2)

Dataset on DenseNet-121	Top-1 Accuracy	Top-5 Acc.
Original Test Set	70.80%	91.20%
Adversarial Test Set 1	59.00%	84.80%
Adversarial Test Set 2	59.60%	86.20%
Adversarial Test Set 3	69.00%	90.60%

Table 2: Transferability results: Performance of DenseNet-121 on the original test set and adversarial datasets generated against ResNet-34.

Discussion and Lessons Learned

Analysis of Attack Transferability

The evaluation of adversarial examples generated against ResNet-34 on the DenseNet-121 architecture provided insightful observations on attack transferability:

Baseline Comparison: DenseNet-121 exhibited a comparable baseline performance (70.80% top-1) to ResNet-34 (70.60% top-1) on the clean dataset, establishing it as a suitable target for transferability testing.

Transferability of Full-Image Attacks (FGSM and PGD): Both FGSM and full-image PGD attacks demonstrated a notable degree of transferability. Adversarial examples crafted by these methods for ResNet-34 successfully deceived DenseNet-121, albeit to a lesser extent than on the source model. FGSM perturbations caused DenseNet-121's top-1 accuracy to drop by 11.8 percentage points (from 70.80% to 59.00%), while PGD perturbations resulted in an 11.2 percentage point drop (to 59.60%). This suggests that these stronger, full-image attacks exploit vulnerabilities somewhat shared between different model architectures, likely related to common feature representations learned from ImageNet. Interestingly, despite PGD being far more potent on the source model (ResNet-34), its transferred impact on DenseNet-121 was very similar to FGSM's, perhaps indicating a limit to the transferability gains from increasing white-box attack strength beyond a certain point.

Limited Transferability of Patch-Based Attacks: In stark contrast, the patch-based PGD attack showed extremely limited transferability. DenseNet-121's top-1 accuracy on these patch-attacked samples was 69.00%, only a minor decrease of 1.8 percentage points from its baseline. This strongly suggests that the vulnerabilities exploited by the localized patch attack are highly specific to the ResNet-34 architecture. The perturbations in a small patch might be finely tuned to the particular local feature extractors and decision boundaries of ResNet-34, which do not generalize well to DenseNet-121's different structure.

Implications: These results highlight a distinction between shared, broader vulnerabilities exploited by global attacks and highly model-specific vulnerabilities targeted by localized attacks. As expected, the accuracy degradation in the black-box transfer setting (DenseNet-121) was much less severe than in the white-box source setting (ResNet-34). Creating transferable attacks appears significantly more challenging when the perturbation area is restricted.

Mitigating Transferability and Future Work

Our findings, particularly the limited transferability of patch attacks, suggest avenues for defense. While designing robust defenses is complex, potential directions to mitigate attack transferability could include:

- **Adversarial Training:** Training models on diverse adversarial examples might improve robustness against both direct and transferred attacks.
- **Ensemble Methods:** Combining predictions from multiple diverse models could hinder attacks optimized for a single architecture.
- **Input Transformations:** Techniques like input randomization or smoothing could disrupt adversarial perturbations, potentially reducing transfer effectiveness.

Further research is needed to develop broadly effective defenses. Future work could involve implementing and evaluating some of these defenses (e.g., adversarial training) and extending the attack analysis to other model types, such as vision transformers.

Future Work: We plan to explore adversarial defenses (e.g., adversarial training) and test attacks on other architectures including vision transformers.

References

- Goodfellow, I. J., Shlens, J., Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. ICLR.
- Madry, A., et al. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR.
- PyTorch Documentation: https://pytorch.org/vision/stable/models.html