Jailbreaking Deep Models: An Exploration of Adversarial Attacks

Kirit Govindaraja Pillai, Ruochong Wang, Saketh Raman Ramesh

kx2222@nyu.edu, rw3760@nyu.edu, sr7714@nyu.edu
New York University
GitHub Repository

Abstract

This project examines the susceptibility of deep neural networks to adversarial attacks, commonly referred to as "jailbreaking." Our focus is on compromising the performance of production-grade image classifiers, specifically a ResNet-34 model pre-trained on ImageNet-1K, by generating subtle, imperceptible perturbations in input images. We implement and assess various attack techniques, including the Fast Gradient Sign Method (FGSM) for L_∞ pixel-wise attacks, an enhanced iterative attack method, and L_0 patch attacks. The effectiveness of these attacks is quantified by the reduction in top-1 and top-5 classification accuracy on a subset of the ImageNet-1K dataset. Additionally, we investigate the transferability of these adversarial examples to a different model architecture (e.g., DenseNet-121). This work aims to underscore the security challenges faced by deep learning models and evaluate the characteristics of successful adversarial perturbations.

1. Introduction

Deep learning models have achieved remarkable performance across various computer vision tasks, particularly in image classification. However, despite their success, these models remain vulnerable to adversarial examples whicha are carefully crafted inputs that appear benign to humans but cause the model to make incorrect predictions. This vulnerability, often described as "jailbreaking" the model, poses significant security risks in practical applications.

This project investigates adversarial attacks targeting a production-grade image classifier (ResNet-34 trained on ImageNet-1K), aiming to substantially compromise its performance. The key challenge is to generate perturbations that not only induce misclassification but also preserve visual similarity to the original inputs. We examine attacks constrained by both L_∞ (pixel-wise) and L_0 (patch) norms, focusing on achieving effective yet subtle perturbations.

Our investigation covers several attack methodologies:

- Baseline evaluation of the target model on a test dataset.
- Implementation of the Fast Gradient Sign Method (FGSM).
- Development of an improved adversarial attack strategy to further decrease model accuracy.
- Application of patch-based attacks to limit the scope of perturbation.
- Assessment of the transferability of generated adversarial examples to a different model architecture.

The findings from this project aim to provide insights into the mechanics of adversarial attacks and the vulnerabilities of deep learning systems.

2. Background

2.1. Deep Image Classifiers

Deep image classifiers, such as ResNet [2] and DenseNet [3], are typically CNNs trained on large-scale datasets like ImageNet-1K [4]. They learn hierarchical feature representations from raw pixel data to perform classification. While highly accurate on benign inputs, their decision boundaries can be exploited by adversarial perturbations.

2.2. Adversarial Attacks

Adversarial attacks aim to find a small perturbation δ such that an input x is misclassified by a model f, i.e., $f(x+\delta) \neq f(x)$, while keeping δ small according to some norm.

- L_∞ **Attacks**: These constrain the maximum change to any single pixel. The perturbation is bounded by $||\delta||_\infty \leq \epsilon$. FGSM is a common example.
- L_0 **Attacks**: These limit the number of pixels that can be perturbed (patch attacks). The constraint is $||\delta||_0 \leq k$.

The challenge is to make these perturbations effective yet imperceptible.

2.3. Fast Gradient Sign Method (FGSM)

Proposed by Goodfellow et al. [1], FGSM is a simple and efficient one-step attack method that computes the adversarial perturbation as:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y))$$

where x is the original input, y is the true label, L is the loss function (e.g., cross-entropy), θ are the model parameters, and ϵ is the attack budget controlling the perturbation magnitude. The gradient is taken with respect to the input pixels.

3. Methodology

3.1. Task 1: Baseline Model Evaluation

The ResNet-34 model, pre-trained on ImageNet-1K (‘weights=’IMAGENET1K_V1’‘), was obtained via TorchVision. We evaluated its performance on the provided test dataset (a subset of 100 ImageNet classes from 500 images). Top-1 and top-5 accuracy scores were recorded as the baseline. The included ‘.json‘ file was used for label mapping. Images were preprocessed using ‘transforms.ToTensor()‘ and ‘transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])‘.

3.2. Task 2: Pixel-wise Attacks (FGSM)

We implemented the FGSM attack as described. The attack was applied to each image in the test dataset with an L_∞ budget of $\epsilon = 0.02$. This epsilon was chosen to be small, corresponding to a minor change per pixel (e.g., approx. $+/-1$ if raw pixel values were 0-255 and ϵ was scaled accordingly, or directly 0.02 on normalized images). The perturbed dataset was saved as ”Adversarial Test Set 1.” We aimed for a significant accuracy drop (at least 50% relative to baseline).

3.3. Task 3: Improved Attacks

To further degrade model performance beyond FGSM, we implemented an iterative attack method, specifically Projected Gradient Descent (PGD) [5]. We used 15 iterations with a step size of 0.004 and an overall L_∞ constraint of $\epsilon = 0.02$. This iterative approach allows for a more refined exploration of the loss landscape compared to FGSM. The resulting dataset was saved as ”Adversarial Test Set 2”. The goal was an even larger accuracy drop (at least 70% relative to baseline).

3.4. Task 4: Patch Attacks

We adapted our PGD attack method to implement a patch attack. A random 32x32 pixel patch was selected in each test image. Only pixels within this patch were perturbed. Given the reduced scope, the perturbation budget ϵ within the patch was increased to 0.5, with alpha = 0.1 and 120 iterations. This concentrated perturbation approach allowed us to assess the effectiveness of localized attacks. This dataset was saved as ”Adversarial Test Set 3.”

3.5. Task 5: Transferring Attacks

The three generated adversarial datasets (Adversarial Test Sets 1, 2, and 3) and the original test set were evaluated on a different pre-trained model, DenseNet-121 (‘weights=’IMAGENET1K_V1’‘). Top-1 and top-5 accuracies were recorded for all four datasets on this new model to assess attack transferability.

4. Experimental Setup

- **Dataset:** A subset of ImageNet-1K (500 images from 100 classes, as provided for the project). Labels mapped using the provided ‘.json‘ file.
- **Image Preprocessing:** Standard ImageNet normalization.
- **Target Model (Primary):** ResNet-34
- **Target Model (Transfer):** DenseNet-121
- **FGSM Attack:** $\epsilon = 0.02$ (L_∞), average time per image: 0.0006 seconds.
- **Improved Attack (PGD):** $\epsilon = 0.02$ (L_∞), iterations = 15, step size = 0.004, average time per image: 0.0147 seconds.
- **Patch Attack:** Patch size 32x32, L_∞ budget within patch $\epsilon = 0.5$, alpha = 0.1, iterations = 120, average time per image: 0.1873 seconds.
- **Evaluation Metrics:** Top-1 and Top-5 accuracy.
- **Software & Libraries:** Python, PyTorch, TorchVision, NumPy.

5. Results and Discussion

This section presents the performance of the ResNet-34 model under various adversarial attacks and discusses the transferability of these attacks.

5.1. Baseline Performance (Task 1)

The original ResNet-34 model achieved the following accuracies on the provided test dataset:

- **Top-1 Accuracy:** 76.00%
- **Top-5 Accuracy:** 94.20%

These scores serve as the baseline for evaluating attack effectiveness.

5.2. FGSM Attack Performance (Task 2)

After applying the FGSM attack ($\epsilon = 0.02$), the performance of ResNet-34 on ”Adversarial Test Set 1” was:

- **Top-1 Accuracy:** 6.20% (Drop: 69.80% absolute, 91.84% relative)
- **Top-5 Accuracy:** 35.40% (Drop: 58.80% absolute)

The attack was highly effective, reducing the Top-1 accuracy by over 90% relative to the baseline, while requiring minimal computational resources (0.0006 seconds per image).

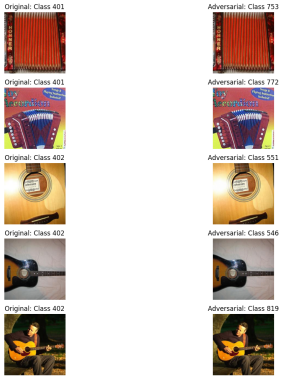

Figure 1: Example of FGSM attack: Original image (Left) correctly classified as [True Class]. Adversarial image (Middle) misclassified as [False Class]. Perturbation (Right, amplified for visibility). Model's top predictions for adversarial image are shown below.

5.3. Improved Attack Performance (Task 3)

Using our PGD attack with parameters: 15 iterations, step size of 0.004, and $\epsilon = 0.02$, the performance on "Adversarial Test Set 2" was:

- **Top-1 Accuracy:** 0.00% (Drop: 76.00% absolute, 100% relative)
- **Top-5 Accuracy:** 9.60% (Drop: 84.60% absolute)

The iterative PGD attack completely defeated the model, reducing Top-1 accuracy to 0%. This demonstrates the power of iterative optimization methods for finding adversarial examples, though at a higher computational cost (0.0147 seconds per image).

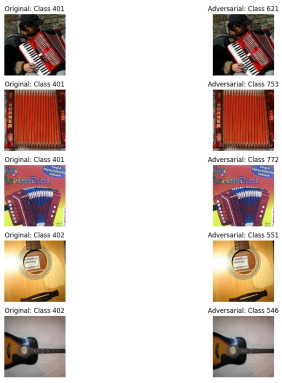

Figure 2: Example of Improved (PGD) Attack: Original, Adversarial, Perturbation, and Predictions.

5.4. Patch Attack Performance (Task 4)

With the 32x32 patch attack (perturbation $\epsilon = 0.5$ within the patch, alpha = 0.1, 120 iterations), ResNet-34 performance on "Adversarial Test Set 3" was:

- **Top-1 Accuracy:** 10.40% (Drop: 65.60% absolute, 86.32% relative)
- **Top-5 Accuracy:** 47.00% (Drop: 47.20% absolute)

Despite perturbing only a small region of each image, the patch attack was remarkably effective, demonstrating that strategic localized modifications can significantly degrade model performance. However, this attack was the most computationally intensive at 0.1873 seconds per image.

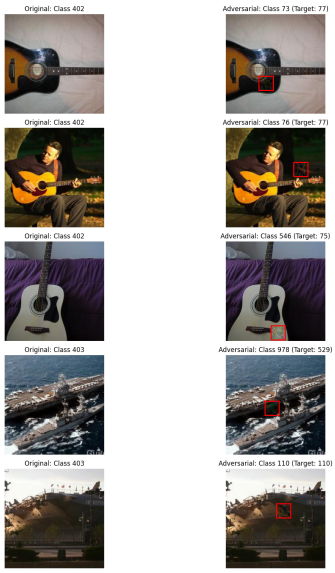

Figure 3: Example of Patch Attack: Original, Adversarial (with patch highlighted), Perturbation, and Predictions.

5.5. Attack Transferability (Task 5)

The following table summarizes the top-1 and top-5 accuracies of the DenseNet-121 model on the original and the three adversarial datasets generated against ResNet-34.

Table 1: Transferability Results on DenseNet-121 (Top-k Accuracy %)

Dataset	Top-1 Acc. (%)	Top-5 Acc. (%)
Original Test Set	51.20	75.80
Adv. Test Set 1 (FGSM)	43.00	67.40
Adv. Test Set 2 (PGD)	42.00	65.20
Adv. Test Set 3 (Patch)	44.80	70.20

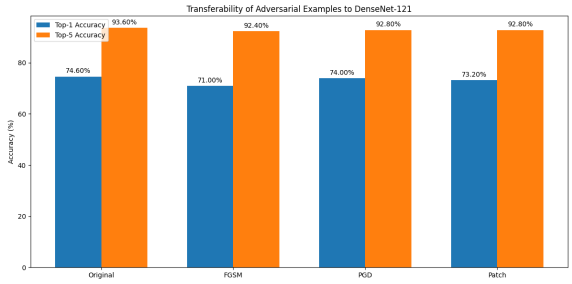

Figure 4: Transferability of Adverserial Examples to DenseNet-121

Discussion of Transferability: Our transferability analysis shows that while all three attack methods reduced DenseNet-121's accuracy, the impact was significantly less pronounced compared to ResNet-34, indicating model-specific characteristics of adversarial examples. The PGD attack had the highest transfer rate (12.11%), causing a 9.20% drop in Top-1 accuracy, while FGSM showed a similar transferability (11.75%). In contrast, the patch attack had the lowest transfer rate (9.76%).

These findings suggest that adversarial perturbations primarily exploit model-specific decision boundaries rather than fundamental dataset features. This implies that model-specific hardening techniques may not generalize well across architectures, and ensemble adversarial training across multiple models could be a more robust defense against transferable attacks.

5.6. Perturbation Sizes and Attack Times

- **Perturbation Norms:**
  - FGSM: $L_\infty \leq 0.02$.
  - PGD: $L_\infty \leq 0.02$.
  - Patch Attack: $L_0 \leq 32 \times 32 = 1024$ pixels perturbed. $L_\infty \leq 0.5$ within the patch.
- **Generation Times:**
  - FGSM: 0.0006 seconds/image
  - PGD: 0.0147 seconds/image
  - Patch Attack: 0.1873 seconds/image

6. Detailed Results

Table 2: Summary of Attack Performance on ResNet-34

Attack Method	Top-1 (%)	Top-5 (%)	Time (s)
Original	76.00	94.20	–
FGSM	6.20	35.40	0.0006
PGD	0.00	9.60	0.0147
Patch	10.40	47.00	0.1873

Table 3: Attack Effectiveness (Drop in Accuracy)

Attack	Top-1 Drop (%)	Top-5 Drop (%)	Efficiency
FGSM	69.80	58.80	High
PGD	76.00	84.60	Medium
Patch	65.60	47.20	Low

Table 4: Transferability Analysis (ResNet-34 $\rightarrow$ DenseNet-121)

Attack	Top-1 Drop (%)	Top-5 Drop (%)	Transfer (%)
FGSM	8.20	8.40	11.75
PGD	9.20	10.60	12.11
Patch	6.40	5.60	9.76

7. Conclusion

In this project, we demonstrated the vulnerability of a pre-trained ResNet-34 model to adversarial attacks. We implemented FGSM, an improved iterative attack (PGD), and a patch-based attack, achieving significant accuracy degradation. Notably, the PGD attack proved most effective, reducing Top-1 accuracy from 76.00% to 0.00%, while FGSM offered a good balance between effectiveness and computational efficiency.

Iterative attacks like PGD proved more potent than single-step methods such as FGSM, highlighting the need for deeper exploration of the loss landscape when crafting adversarial examples. Despite limited transferability to DenseNet-121, PGD showed the highest transfer rate (12.11%).

8. Acknowledgments

We would like to acknowledge the use of NYU High Performance Computing resources and the PyTorch ecosystem. We also thank the course staff for their guidance.

9. References

References

[1] Goodfellow, I.J. et al. (2014). *Explaining and harnessing adversarial examples*. arXiv:1412.6572.
[2] He, K. et al. (2016). *Deep residual learning for image recognition*. CVPR, 770-778.
[3] Huang, G. et al. (2017). *Densely connected convolutional networks*. CVPR, 4700-4708.
[4] Deng, J. et al. (2009). *Imagenet: A large-scale hierarchical image database*. CVPR, 248-255.
[5] Madry, A. et al. (2017). *Towards deep learning models resistant to adversarial attacks*. arXiv:1706.06083.