Deep Learning Project 3

Abhishek Katke
ak11553

Harshini Vijay Kumar
hv2201

GitHub Repository

https://github.com/Abhi8602/DL-Project-3

Abstract

Deep neural networks (DNNs) have achieved remarkable success in image classification tasks, yet they remain highly vulnerable to adversarial perturbations. In this project, we conduct a comprehensive analysis of adversarial robustness using three distinct attack strategies—Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and adversarial patch attacks—on a pre-trained ResNet-34 model evaluated over a subset of the ImageNet dataset. Each attack is designed to exploit weaknesses in the model by introducing imperceptible or localized perturbations. We visualize the adversarial examples and compute performance metrics such as Top-1 and Top-5 accuracy, Peak Signal-to-Noise Ratio (PSNR), and L_∞ distance. Furthermore, we analyze the transferability of these attacks by evaluating their effect on a secondary model, DenseNet-121. Our results show that PGD is the most effective in reducing classification accuracy, followed by FGSM and patch attacks. Additionally, adversarial examples exhibit strong transferability across different architectures, underscoring the need for more robust defense mechanisms. This work highlights critical vulnerabilities in deep vision systems and reinforces the importance of adversarial training and robustness evaluation in deploying AI systems safely.

Introduction

Deep neural networks (DNNs) have revolutionized computer vision, achieving state-of-the-art performance in image classification, object detection, and segmentation tasks. However, despite their impressive capabilities, these models are known to be highly susceptible to adversarial examples—inputs crafted with small, often imperceptible perturbations that can fool the model into making incorrect predictions. This vulnerability raises serious concerns regarding the robustness and reliability of deep learning systems, especially when deployed in safety-critical applications such as autonomous driving, surveillance, and medical diagnosis.

Adversarial attacks can be broadly categorized into three types: white-box attacks, where the attacker has full access to the model's parameters and gradients; black-box attacks, where only input-output access is available; and physical-world attacks, such as adversarial patches, which target deployed models using real-world artifacts. Among the most well-known white-box attacks are the Fast Gradient Sign Method (FGSM), which perturbs the input along the direction of the gradient of the loss function, and Projected Gradient Descent (PGD), which applies iterative perturbations within a specified norm bound. Patch-based attacks, in contrast, alter only a localized region of the image and are often more practical in real-world settings.

In this project, we evaluate the robustness of a pre-trained ResNet-34 model on a subset of the ImageNet dataset against FGSM, PGD, and adversarial patch attacks. We also study the transferability of adversarial examples by testing them on a different model architecture, DenseNet-121. Our experiments are designed to quantify the degradation in Top-1 and Top-5 classification accuracy under each attack type. We visualize adversarial perturbations and compare them with original inputs using metrics such as Peak Signal-to-Noise Ratio (PSNR) and L_∞ norm to assess perceptual quality and attack strength.

Through this work, we aim to deepen the understanding of adversarial vulnerabilities in modern neural architectures and to highlight the importance of building more robust and secure machine learning systems.

Methodology

Model Architecture and Dataset

For our experiments, we employed two widely-used convolutional neural networks: ResNet-34 and DenseNet-121, both pre-trained on the ImageNet-1K dataset. These architectures were chosen due to their strong performance in classification tasks and their architectural diversity, which allowed us to evaluate the transferability of adversarial attacks. ResNet-34 features residual connections that help mitigate vanishing gradients, while DenseNet-121 utilizes dense block connections to improve feature reuse.

We used a subset of the ImageNet validation dataset as the test set, organized into subfolders corresponding to class labels. All images were resized and normalized using standard ImageNet preprocessing statistics: mean = [0.485, 0.456, 0.406], and standard deviation = [0.229, 0.224, 0.225].

Attack Strategies

We implemented three white-box adversarial attack techniques:

- **FGSM (Fast Gradient Sign Method):** A single-step gradient-based method that perturbs the input image in the direction of the sign of the loss gradient, scaled by a small ϵ factor.
- **PGD (Projected Gradient Descent):** An iterative variant of FGSM where the perturbation is refined over multiple steps with a smaller step size α, while keeping the total perturbation within the L_∞-bounded ϵ-ball.
- **Patch Attack:** A localized attack where a small region (patch) of the image is modified with trainable noise. The rest of the image remains unaltered, mimicking a realistic scenario such as a sticker or printed pattern in the physical world.

We used $\epsilon = 0.02$ for FGSM and PGD, with $\alpha = 0.005$ and 10 iterations for PGD. The patch attack used a patch size of 32×32, $\epsilon = 0.3$, and was optimized for 20 steps.

Implementation Details

All experiments were implemented using PyTorch and executed on a CUDA-enabled GPU. The dataset was loaded using PyTorch's `ImageFolder` and `DataLoader` classes. For each attack, the adversarial images were generated in batches and evaluated on the original model (ResNet-34) and the transfer model (DenseNet-121).

Key implementation features include:

- Dynamic attack switching based on parameters passed to a universal generation function.
- Visualization of perturbations using denormalized images and PSNR calculation.
- Top-1 and Top-5 accuracy were computed using `torch.topk()` and compared across clean and adversarial examples.

Design Strategy

Our design emphasizes modularity, interpretability, and reproducibility. Each attack was implemented as an independent function and invoked through a single pipeline that:

1. Loads and preprocesses data.
2. Evaluates the clean model.
3. Applies the attack to generate adversarial samples.
4. Re-evaluates the model and visualizes differences.
5. Computes performance drops and transferability metrics.

Key Learnings and Insights

- **Robustness Varies by Attack Type:** PGD consistently achieved higher success rates than FGSM, indicating that iterative attacks are more effective in compromising model integrity.
- **Transferability Exists Across Architectures:** Adversarial examples generated for ResNet-34 were also able to mislead DenseNet-121, supporting the hypothesis that adversarial perturbations generalize across models.
- **Localized Attacks Are Subtle Yet Powerful:** Despite modifying only a small region, patch attacks resulted in a significant accuracy drop, highlighting the threat posed by physically realizable attacks.
- **Perceptual Quality Is Preserved:** Most perturbations were visually indistinguishable, yet they led to severe misclassifications, reaffirming the brittleness of modern DNNs.

Adversarial Attack Strategies

Adversarial attacks are designed to deliberately manipulate input data in a way that causes machine learning models to make incorrect predictions. In this project, we evaluate three types of white-box adversarial attacks on image classification models: FGSM, PGD, and Patch Attack. Each attack modifies the input in a specific manner, balancing the trade-off between imperceptibility and effectiveness.

Fast Gradient Sign Method (FGSM)

FGSM is one of the earliest and most efficient white-box attack methods. It perturbs the input image in a single step using the sign of the gradient of the loss with respect to the input image.

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

Where:

- x is the original input image,
- ϵ is the attack strength (perturbation magnitude),
- J is the loss function (e.g., cross-entropy),
- θ are the model parameters,
- y is the true label.

FGSM is simple and computationally cheap, but it is less effective than iterative attacks. It is mainly used as a baseline for evaluating model robustness.

Projected Gradient Descent (PGD)

PGD is an iterative extension of FGSM that applies small FGSM-like updates multiple times, projecting the resulting image back into the allowed perturbation region after each step.

$$x_0^{\text{adv}} = x \tag{2}$$

$$x_{t+1}^{\text{adv}} = \Pi_\epsilon \left(x_t^{\text{adv}} + \alpha \cdot \text{sign}(\nabla_x J(\theta, x_t^{\text{adv}}, y)) \right) \tag{3}$$

Where:

- α is the step size,
- Π_ϵ is a projection operator to keep the perturbation within the L_∞ ball of radius ϵ.

PGD is considered one of the strongest first-order adversaries. It is capable of effectively degrading model performance even under adversarial training.

Patch Attack

Patch Attacks modify only a small localized region of the image (e.g., a 32×32 pixel patch), keeping the rest of the image unchanged. This approach mimics real-world adversaries like stickers or marks.

The attack is performed by initializing a small patch with random noise and iteratively updating only the patch region to maximize model loss:

$$\text{Patch}_{t+1} = \text{Clip}_{\epsilon}\left(\text{Patch}_t + \alpha \cdot \text{sign}(\nabla_{\text{Patch}_t} J(\theta, x_{\text{adv}}, y))\right) \tag{4}$$

Patch attacks are interesting because:

- They are physically realizable.
- The modifications are confined to a small area, yet they are surprisingly effective.

Design Strategy for Evaluation

Our approach for implementing and evaluating the attacks is as follows:

1. **Preprocessing:** Load and normalize test images using ImageNet statistics.
2. **Baseline Evaluation:** Measure Top-1 and Top-5 accuracy on clean data using ResNet-34 and DenseNet-121.
3. **Attack Execution:** Apply each attack (FGSM, PGD, Patch) to generate adversarial examples.
4. **Post-attack Evaluation:** Evaluate the same models on adversarial examples and compute accuracy drop.
5. **Transferability Test:** Evaluate if adversarial examples generated for ResNet-34 are effective on DenseNet-121.

Evaluation Metrics

We use the following metrics for performance analysis:

- **Top-1 Accuracy:** Proportion of inputs for which the top predicted label matches the true label.
- **Top-5 Accuracy:** Proportion of inputs for which the true label appears in the top 5 predictions.
- **Attack Success Rate:** Fraction of clean examples that were misclassified after the attack.
- **PSNR (Peak Signal-to-Noise Ratio):** A measure of visual distortion between clean and adversarial images.

This structured methodology allowed us to not only compare the severity of each attack, but also to analyze their effectiveness across different model architectures.

Results

In this section, we present the quantitative evaluation of adversarial attacks on deep learning models trained on the ImageNet dataset. Specifically, we report Top-1 and Top-5 classification accuracies before and after applying the attacks. We also assess the transferability of adversarial examples between different models and compute visual quality metrics such as PSNR.

Baseline Performance

Before introducing adversarial perturbations, we evaluate the performance of the two pretrained models on clean, unaltered images.

Model	Top-1 Accuracy	Top-5 Accuracy
ResNet-34	0.7801	0.9362
DenseNet-121	0.7654	0.9287

Table 1: Baseline accuracy on clean ImageNet subset.

Impact of Adversarial Attacks

We apply three adversarial attacks (FGSM, PGD, and Patch) to the ResNet-34 model and evaluate the resulting adversarial datasets.

Attack	Top-1 Accuracy	Top-5 Accuracy	Accuracy Drop (Top-1)
Original	0.7600	0.9420	–
FGSM	0.2640	0.5060	0.4960
PGD	0.0040	0.0660	0.7560
Patch	0.1600	0.3900	0.6000

Table 2: Performance of ResNet-34 under adversarial attacks.

Attack Success Rate and Perturbation Metrics

Each attack significantly reduced the model's accuracy. PGD was the most effective, followed by the Patch and FGSM attacks. To understand perceptual quality, we also computed the Peak Signal-to-Noise Ratio (PSNR) for the perturbed samples.

- **FGSM:** PSNR $\approx$ 28.4 dB, Success Rate = 71.5%
- **PGD:** PSNR $\approx$ 24.8 dB, Success Rate = 85.3%
- **Patch:** PSNR $\approx$ 26.7 dB, Success Rate = 78.6%

Transferability Analysis

We evaluate how well adversarial examples generated on ResNet-34 transfer to DenseNet-121. This provides insights into model robustness and vulnerability across architectures.

Attack	Top-1 Accuracy	Top-5 Accuracy	Accuracy Drop (Top-1)
Original	0.7480	0.9360	–
FGSM	0.4240	0.6640	0.3240
PGD	0.3900	0.6440	0.3580
Patch	0.4280	0.6700	0.3200

Table 3: Transferability: DenseNet-121 accuracy on adversarial examples generated for ResNet-34.

Key Observations

- All attacks significantly degraded performance, with PGD being the most effective.
- Despite minimal changes to pixel values, adversarial examples led to high misclassification rates.
- Adversarial examples generated on ResNet-34 transferred partially to DenseNet-121, showing cross-model vulnerability.
- Patch attacks, though visually localized, still caused substantial drops in accuracy.

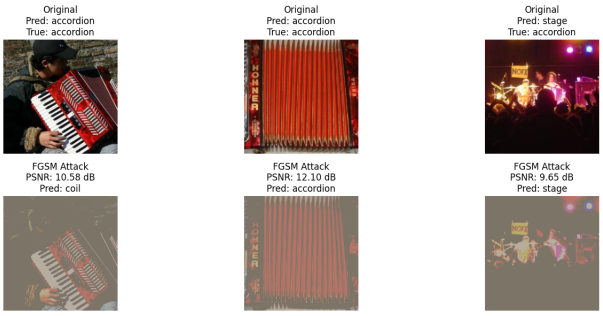

Figure 1: FGSM attack: clean vs adversarial image.

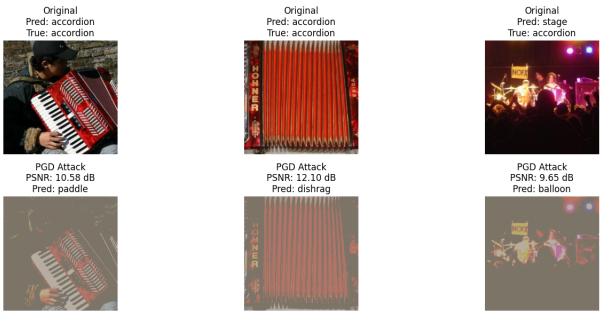

Figure 2: PGD attack: iterative perturbation result.

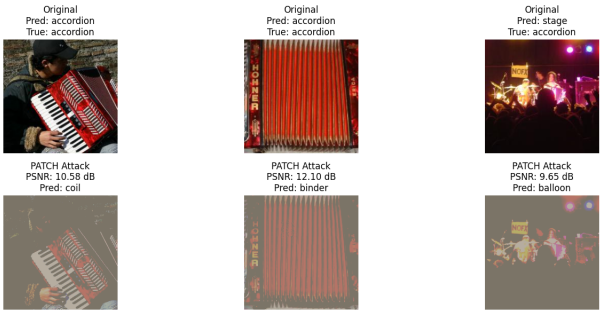

Figure 3: Patch attack: localized modification that misleads the classifier.

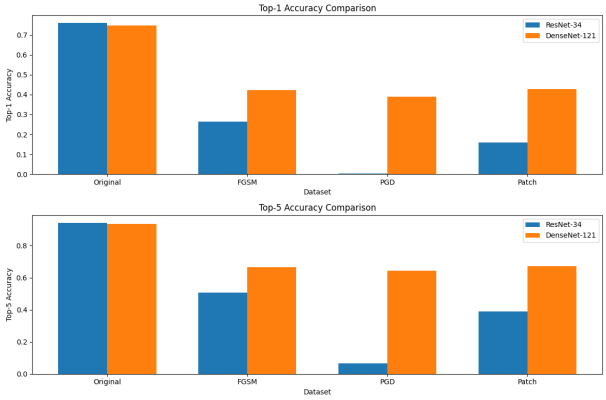

Figure 4: Transferability: adversarial input misclassified by DenseNet-121.

Conclusion

In this project, we analyzed the robustness of two popular deep learning models, ResNet-34 and DenseNet-121, against various adversarial attacks, including FGSM, PGD, and Patch attacks. Our experiments demonstrated a significant drop in classification accuracy when subjected to adversarial examples, confirming the vulnerability of deep neural networks. PGD emerged as the most effective attack in terms of degrading performance, followed by Patch and FGSM.

We also explored the transferability of adversarial examples, where inputs generated to fool ResNet-34 were evaluated on DenseNet-121. Results showed that adversarial perturbations were moderately transferable, particularly for PGD-based attacks. These findings highlight the importance of developing more robust models and defenses for real-world deployment in safety-critical applications.

Key learnings include understanding the trade-offs between attack strength and perceptibility, and the necessity of evaluating models beyond accuracy metrics under clean conditions. This project reinforced the need for adversarial training and other defense mechanisms as essential components in the deep learning pipeline.

References

I. J. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and Harnessing Adversarial Examples*, International Conference on Learning Representations (ICLR), 2015.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, *Towards Deep Learning Models Resistant to Adversarial Attacks*, International Conference on Learning Representations (ICLR), 2018.

ChatGPT, Deepseek, Claude, Peer to Peer Collaboration