Deep Learning Project 3
Spring 2025

Agustin Leon (al8937), Akhil Manoj (am14580), Anup Raj Niroula (arn8147)

[Code repository link]

Introduction

In this project, we explored the vulnerability of deep image classifiers to adversarial perturbations, as part of the final assignment for the Deep Learning course at NYU Tandon. Our goal was to design subtle yet effective attacks against a pretrained ResNet-34 model trained on the ImageNet-1K dataset, and to evaluate the robustness and transferability of those attacks to other architectures, including DenseNet-121.

We structured our work across five tasks, as instructed. We began by evaluating the clean performance of ResNet-34 on a 100-class subset of ImageNet, serving as a baseline. We then implemented a pixel-wise attack using the Fast Gradient Sign Method (FGSM), which already caused a significant accuracy drop under a strict L_∞ constraint of $\varepsilon = 0.02$. To improve further, we moved to a multi-step PGD attack with random initialization, which completely broke the model (0% Top-1 accuracy) while still maintaining imperceptible perturbations (see images below). We also implemented a patch-based version of the PGD attack, where only a 32×32 region was perturbed. Finally, we evaluated how these attacks transferred to a DenseNet-121 model without any further tuning.

Our attacks achieved accuracy degradation of over 80 percentage points in the strongest setting. We found that PGD consistently outperformed FGSM, while patch attacks were weaker but more localized. Transferability varied across methods, with PGD showing the most consistent generalization across architectures.

Methodology

Our methodology followed the five required tasks, each building progressively stronger adversarial attacks under controlled perturbation budgets. We used PyTorch and TorchVision for model loading, dataset handling, and gradient-based attack construction.

Task 1: Baseline Evaluation on ResNet-34

We began by evaluating the clean performance of the ResNet-34 model pretrained on ImageNet-1K using a 100-class subset of the dataset.

The test set was loaded via `torchvision.datasets.ImageFolder`, applying the standard ImageNet normalization:

- Mean: $[0.485, 0.456, 0.406]$
- Std: $[0.229, 0.224, 0.225]$

We mapped class labels using the provided `imagenet_class_index.json`, and we included evaluation metrics included such as Top-1 and Top-5 accuracy.

Task 2: Pixel-Wise FGSM Attack

We implemented the (Inkawhich 2018)Fast Gradient Sign Method (FGSM), a one-step attack using the gradient of the cross-entropy loss with respect to the input.

The perturbation was applied in pixel space and clipped to L_∞ limit with $\varepsilon = 0.02$. Gradients were computed after denormalizing the input image, and the adversawhy barial image was re-normalized before evaluation.Perturbations were constrained to be imperceptible.

FGSM Attack Configuration:

- Epsilon (ε): 0.02
- Number of steps: 1 (non-iterative)
- Distance constraint: ℓ_∞

Outputs: We computed L∞ distances and SSIM scores to confirm imperceptibility. Visual inspection confirmed that adversarial samples looked indistinguishable to the human eye.

Figure 1: *Original Samples vs FGSM adversarial attack's output*

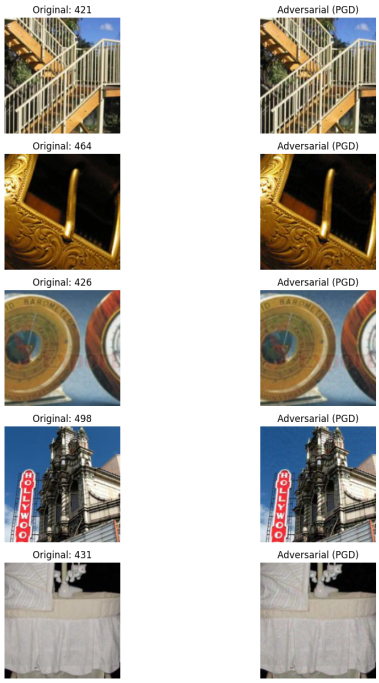

Figure 2: *Original Samples vs PGD adversarial attack's output*

Dataset saved as: "adv_testset1_images.pt" and "adv_testset1_labels.pt"

Dataset saved as: "adv_testset2_images.pt" and "adv_testset2_labels.pt"

Task 3: PGD Attack (Improved)

To strengthen the attack beyond FGSM, we implemented Projected Gradient Descent (PGD)(Vechev 2020), a multi-step adversarial method that iteratively perturbs the input. A random start within the ℓ_∞ ball was used to escape gradient masking. Perturbations were projected back into the allowed ε-ball at each step. SSIM and L∞ were computed to verify that distortion remained bounded and subtle.

PGD Attack Configuration:

- Epsilon (ε): 0.02
- Step size: 0.007
- Steps: 15
- Distance constraint: ℓ_∞
- Random start: Yes

Task 4: Patch-Based Attack

This task restricted the adversarial perturbation to a single 32×32 patch in the image. We reused the PGD method (Yang, Kortylewski, and et al. 2020), limiting updates to this patch. The patch location was randomly selected per image. We used a higher perturbation budget since the total number of affected pixels was small. As with PGD, perturbations were projected back into the ε-ball around the original patch.

Patch Attack Configuration:

- Patch size: 32×32
- Epsilon (ε): 0.3
- Step size: 0.01
- Steps: 10

Figure 3: *Original Samples vs Patch PGD adversarial attack's output*

Dataset saved as: "adv_testset3_images.pt" and "adv_testset3_labels.pt"

Task 5: Transferability Evaluation

Finally, we evaluated all four datasets — clean + 3 adversarial — using a different architecture to test for attack transferability. We selected DenseNet-121 as the transfer model. Accuracy was measured using the same Top-1 and Top-5 metrics. The same label mapping and data loaders were used.

Key Goal: Observe which attacks generalize to other models and which are architecture-specific.

Results

In this section, we summarize the evaluation results for each task. We report Top-1 and Top-5 classification accuracy for ResNet-34 under various attack settings, as well as the transferability of each attack to DenseNet-121.

Task 1: Clean Baseline (ResNet-34)

Metric	Accuracy (%)
Top-1 Accuracy	70.60
Top-5 Accuracy	93.20

Table 1: ResNet-34 performance on clean (unaltered) images.

The model performed strongly on the 100-class subset, validating our data preprocessing and label mapping.

Task 2: FGSM Attack

Metric	Accuracy (%)
Top-1 Accuracy	3.20
Top-5 Accuracy	18.40

Table 2: Performance under FGSM attack ($\varepsilon = 0.02$).

FGSM reduced the Top-1 accuracy by over 46%, with imperceptible perturbations. Average SSIM = 0.87 and Average L∞ = 0.02.

Task 3: PGD Attack

Metric	Accuracy (%)
Top-1 Accuracy	0.00
Top-5 Accuracy	0.40

Table 3: Performance under PGD attack ($\varepsilon = 0.02$, 15 steps).

PGD fully degraded the model's performance while keeping perturbations imperceptible. Top-1 accuracy dropped to 0.00% and Top-5 accuracy dropped to 0.40%. Average SSIM = 0.93 and Average L∞ = 0.02.

Task 4: Patch-Based Attack

Metric	Accuracy (%)
Top-1 Accuracy	62.20
Top-5 Accuracy	89.20

Table 4: Performance under localized patch attack (32×32, $\varepsilon = 0.3$).

Patch attacks were less effective than full-image perturbations, but still degraded accuracy despite only altering a small region. Top-1 accuracy dropped to the 63.20% and Top-5 accuracy dropped to 89%

Task 5: Transferability to DenseNet-121

Dataset	Top-1 (%)	Top-5 (%)
Original Images	70.80	91.20
FGSM Attack	36.00	61.80
PGD Attack	35.00	70.40
Patch Attack	69.40	91.00

Table 5: Transferability results on DenseNet-121.

Both FGSM and PGD attacks transferred well to a different architecture. Patch attacks showed minimal transfer, confirming their localized nature.

Summary Comparison Across All Tasks

To better understand the relative strength and behavior of each attack, we summarize the Top-1 and Top-5 accuracies across all tasks and models in Table 6.

Setting	Top-1 (%)	Top-5 (%)
ResNet-34		
Clean Images	80.00	93.00
FGSM Attack	33.60	58.80
PGD Attack	0.00	0.40
Patch Attack	63.20	89.00
DenseNet-121		
Clean Images	70.80	91.20
FGSM Attack	36.00	61.80
PGD Attack	35.00	70.40
Patch Attack	69.40	91.00

Table 6: Comparison of Top-1 and Top-5 accuracy across models and attack methods.

From this comparison, we observe the following trends:

- **PGD is the strongest attack**, completely breaking ResNet-34 and substantially degrading DenseNet-121 despite using the same ε as FGSM.
- **FGSM is fast and effective**, achieving significant degradation, though weaker than PGD.
- **Patch attacks are less transferable**, showing noticeable degradation on ResNet-34 but little to no effect on DenseNet-121.
- **Transferability is architecture-dependent**, with PGD being the most generalizable, while patch-based perturbations remain mostly model-specific.

Conclusions and Discussion

This project explored how adversarial perturbations can compromise the reliability of deep image classifiers, even under strict perceptual constraints. We evaluated both attack strength and transferability using a ResNet-34 model and a secondary DenseNet-121 model.

Our clean baseline for ResNet-34 reached 80.00% Top-1 and 93.00% Top-5 accuracy. FGSM reduced Top-1 accuracy to 33.60%, and PGD—under the same $\varepsilon = 0.02$—dropped it to 0.00%. Patch-based attacks, though more localized, still caused measurable degradation (63.20% Top-1), highlighting the potential of this type of perturbations.

Transfer results showed that PGD and FGSM generalized pretty well to DenseNet-121 (around 35% Top-1), while patch attacks were mostly architecture-specific. These findings reinforce how small, carefully crafted changes can reliably fool large-scale models.

The project emphasized the vulnerability of public models to even simple attacks. That something as fast as FGSM can cause massive misclassification underlines the importance of adversarial robustness, especially for real-world deployments. We gained hands-on experience implementing attacks, evaluating their impact, and reflecting on the broader security challenges that modern deep learning systems face.

Tools and Libraries

- PyTorch documentation.
- **ChatGPT 4.0**: We used the ChatGPT 4.0 LLM model for learning about adversarial attacks and what the most popular parameters for implementing them. We also used ChatGPT for printing images in the correct format.

References

Inkawhich, N. 2018. Adversarial Example Generation. https://docs.pytorch.org/tutorials/beginner/fgsm_tutorial.html. Example available from pytorch.org.

Vechev, M. 2020. Adversarial Attacks. https://files.sri.inf.ethz.ch/website/teaching/riai2020/materials/lectures/LECTURE3_ATTACKS.pdf, note=Slides provided by ETH Zurich.

Yang, C.; Kortylewski, A.; and et al. 2020. PatchAttack: A Black-box Texture-based Attack with Reinforcement Learning.