Project 3

● Graded

Group

Sahil Sarnaik
Pranav Tushar Pradhan
Rohan Gore
View or edit group

Total Points

25 / 25 pts

Question 1

Project report

15 / 15 pts

✔ **– 0 pts** Correct

– 2 pts Insufficient explanation

– 1 pt Late submission

– 2 pts Didn't use required (AAAI) format.

Question 2

Project codebase

10 / 10 pts

✔ **– 0 pts** Correct

– 1 pt Results ?

– 2 pts Insufficient explanations

– 10 pts missing codebase

Questions assigned to the following page: 1 and 2

Jailbreaking Deep Models

Team Members: Pranav Pradhan(pp3051), Sahil Sarnaik (ss19100), Rohan Gore (rmg9725)

GitHub Repository

Abstract

This project investigates the vulnerability of deep neural networks-specifically, image classifiers-against adversarial attacks. We focus on attacking a ResNet-34 model trained on ImageNet-1K using both pixel-wise and patch-based perturbations, evaluating the effectiveness of various attack strategies. Additionally, we examine the transferability of these attacks to other state-of-the-art models, including DenseNet-121 and ViT-B_16. Our experiments demonstrate that even subtle, carefully crafted perturbations can drastically degrade model performance, and that some attacks transfer well across architectures while others do not.

Methodology

Data and Baseline Setup

Dataset: A subset of ImageNet-1K comprising 500 images from 100 classes, with ground-truth labels provided via JSON.
Model: Pre-trained ResNet-34 from TorchVision (IMAGENET1K_V1 weights).
Preprocessing: Images are resized to 256, center-cropped to 224, converted to tensor, and normalized using ImageNet means and standard deviations.
Evaluation: Top-1 and Top-5 accuracy are computed by comparing model predictions to ground-truth indices as specified in the label mapping.

Task 1: Baseline Model Evaluation

We first evaluated ResNet-34 on the unperturbed test set to establish a performance baseline.

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

These results confirm the model's strong performance on clean data, consistent with its published benchmarks.

Task 2: Pixel-wise Attack (FGSM, L_∞, $\varepsilon = 0.02$)

We implemented the Fast Gradient Sign Method (FGSM), which perturbs each pixel by at most ε in the direction of the sign of the loss gradient. This attack is constrained to be imperceptible ($\varepsilon = 0.02$ in normalized pixel space).

Attack Details:

- Each image is denormalized, gradients are computed w.r.t. input pixels, and the perturbation is applied.
- The perturbed images are re-normalized and evaluated by the model.
- All adversarial images are saved for further analysis and transferability tests.

Results:

- Top-1 Accuracy: 3.00%
- Top-5 Accuracy: 19.00%
- Accuracy Drop: 73.00% (Top-1), 75.20% (Top-5)

Observations:

- The attack is highly effective, reducing Top-1 accuracy to near-random levels.
- Visual inspection of adversarial images confirms that perturbations are imperceptible.

Task 3: Improved Attack (PGD, L_∞, $\varepsilon = 0.02$, 10 steps)

We enhance the attack using Projected Gradient Descent (PGD). We chose Projected Gradient Descent (PGD) as our improved attack method because it is widely recognized as one of the most effective first-order adversarial attack algorithms for generating strong, constrained perturbations. While the Fast Gradient Sign Method (FGSM) applies a single-step perturbation in the direction of the loss gradient, PGD extends this approach by iteratively applying small, bounded updates, each time projecting the perturbed image back into the allowed L_∞ ball around the original image. This iterative process allows PGD to find more effective adversarial examples, often reaching the worst-case performance of a model within the specified perturbation budget. By starting from a random point within the allowed perturbation region, PGD avoids local minima and produces a more diverse set of adversarial examples, further increasing its effectiveness.

Attack Details:

- 10 steps, step size $\alpha = \varepsilon/10$.
- Random initialization within the ε-ball.
- Projection ensures the perturbation never exceeds the allowed L_∞ norm.

Question assigned to the following page: 1

Results:

- Top-1 Accuracy: 0.00%
- Top-5 Accuracy: 1.80%
- Accuracy Drop: 76.00% (Top-1), 92.40% (Top-5)

Observations:

- PGD is even more effective than FGSM, essentially reducing the model to random guessing.
- Visualizations again confirm imperceptibility and adherence to the perturbation constraint.
- In our experiments, 10 steps were sufficient to reduce the ResNet-34 model's Top-1 accuracy to 0%, indicating that further increases in the number of steps would yield diminishing returns for this model and dataset

Task 4: Patch Attack (Targeted Patch PGD, Patch Size 32×32, $\varepsilon = 0.75$)

We restricted perturbations to a small, randomly located patch within the image, increasing ε to 0.75 to compensate for the reduced area of attack. A targeted attack is used, aiming to force the model to predict the least likely class for each image.

Attack Details:

- Patch location is chosen based on gradient sensitivity.
- Patch content is iteratively updated using PGD (100 steps, 3 random restarts).
- Only the patch region is perturbed, with all other pixels left unchanged.

Results:

- Top-1 Accuracy: 6.00%
- Top-5 Accuracy: 23.40%
- Accuracy Drop: 70.00% (Top-1), 70.80% (Top-5)

Observations:

- Despite being localized, the attack is highly effective, though slightly less so than global PGD.
- The attack remains visually subtle, with the patch often imperceptible to human observers.

Task 5: Transferability to Other Models

We evaluate the adversarial test sets (FGSM, PGD, Patch PGD) on DenseNet-121 and ViT-B_16, both pre-trained on ImageNet-1K.

We chose DenseNet-121 as s the most commonly used and widely supported variant in the DenseNet family, with pre-trained weights readily available in PyTorch and other libraries. Using DenseNet-121 allows for direct comparison with a large body of published adversarial robustness results. We also included ViT-B_16 as it is a transformer-based model, in contrast to the convolutional architectures of ResNet and DenseNet. Including it allows us to test whether adversarial examples crafted for CNNs also affect transformer-based vision models.

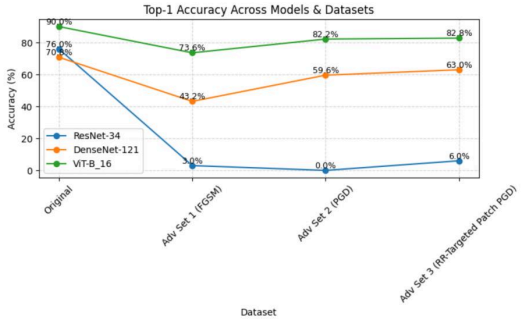

Figure 1: Top-1 Accuracy Across Models & Datasets

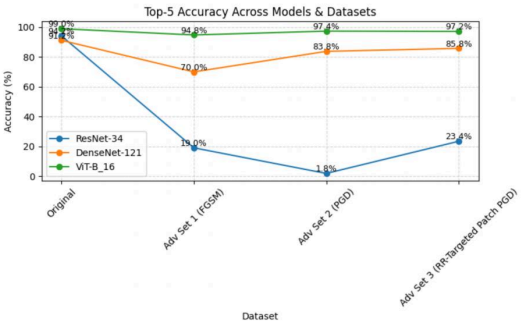

Figure 2: Top-5 Accuracy Across Models & Datasets

Results Summary and Visualization

The following graphs summarize the Top-1 and Top-5 accuracy of all three models (ResNet-34, DenseNet-121, ViT-B_16) across the original and adversarial datasets:

These visualizations clearly illustrate the dramatic drop in accuracy for ResNet-34 under adversarial attack, with DenseNet-121 and ViT-B_16 also experiencing significant, though less severe, performance degradation. Notably, ViT-B_16 demonstrates the highest robustness to transferred attacks, while ResNet-34 is most vulnerable.

Results Tables

ResNet-34

Dataset	Top-1 (%)	Top-5 (%)
Original	76.0	94.2
Adv Set 1 (FGSM)	3.0	19.0
Adv Set 2 (PGD)	0.0	1.8
Adv Set 3 (Patch PGD)	6.0	23.4

Table 1: ResNet-34 Accuracy

Question assigned to the following page: 1

DenseNet-121

Dataset	Top-1 (%)	Top-5 (%)
Original	70.8	91.2
Adv Set 1 (FGSM)	43.2	70.0
Adv Set 2 (PGD)	59.6	83.8
Adv Set 3 (Patch PGD)	63.0	85.8

Table 2: DenseNet-121 Accuracy

ViT-B_16

Dataset	Top-1 (%)	Top-5 (%)
Original	90.0	99.0
Adv Set 1 (FGSM)	73.6	94.8
Adv Set 2 (PGD)	82.2	97.4
Adv Set 3 (Patch PGD)	82.8	97.0

Table 3: ViT-B_16 Accuracy

Discussion

Attack Effectiveness

FGSM: Highly effective on ResNet-34, and also transfers well to DenseNet-121 and ViT-B_16, causing significant drops in accuracy.
PGD: Most Effective on the source model (ResNet-34), but transferability is somewhat reduced compared to FGSM for DenseNet-121 and ViT-B_16. This suggests PGD attacks may overfit to the source model's gradients.
Patch Attack: While extremely effective on ResNet-34, patch attacks are less transferable, especially to ViT-B_16. The localized nature of the perturbation seems to exploit specific vulnerabilities in the source model that do not generalize as well.

Observations and Trends

- **FGSM (Fast Gradient Sign Method)**: This single-step attack was highly effective against the source model (ResNet-34), reducing Top-1 accuracy from 76% to 3%. Notably, FGSM adversarial examples also transferred well to DenseNet-121 and ViT-B_16, causing significant drops in their accuracy. This suggests that simpler, global perturbations can exploit shared vulnerabilities across different architectures
- **PGD (Projected Gradient Descent)**: PGD was the most effective attack on the source model, driving Top-1 accuracy to 0%. However, its transferability was reduced compared to FGSM-DenseNet-121 and ViT-B_16 retained higher accuracy on PGD adversarial examples. This indicates that while PGD is powerful, it may overfit to the specific gradients of the source model, making it less effective when transferred
- **Patch PGD**: The patch-based attack, despite being highly localized, still caused substantial accuracy drops on ResNet-34. However, its transferability was the lowest, especially to ViT-B_16. This suggests that localized attacks may exploit model-specific features that do not generalize well across architectures

Lessons Learned

- **Hyperparameter Trade-offs**: Increasing the number of PGD steps or restarts can enhance attack strength but brings diminishing returns and higher computational costs. Selecting hyperparameters requires balancing effectiveness and efficiency
- **Patch Size and Epsilon**: For localized (patch) attacks, a larger ε is necessary to achieve comparable effectiveness to global attacks. There is a trade-off between maintaining imperceptibility and ensuring attack success.
- **Model Diversity and Robustness**: Transformer-based models like ViT-B_16 appear more robust to transferred attacks than CNNs, though they are not immune. This suggests that architectural diversity in model ensembles could improve system-level robustness.
- **Visualization**: Visual inspection of adversarial examples is essential to confirm that perturbations remain imperceptible and to identify any implementation errors.
- **Generalization**: The effectiveness of an attack on one model does not guarantee similar results on others. Robustness evaluations should always include multiple architectures to avoid overestimating security.

Citations

- PyTorch/TorchVision pre-trained models documentation.
- 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.

Question assigned to the following page: 1

Appendix

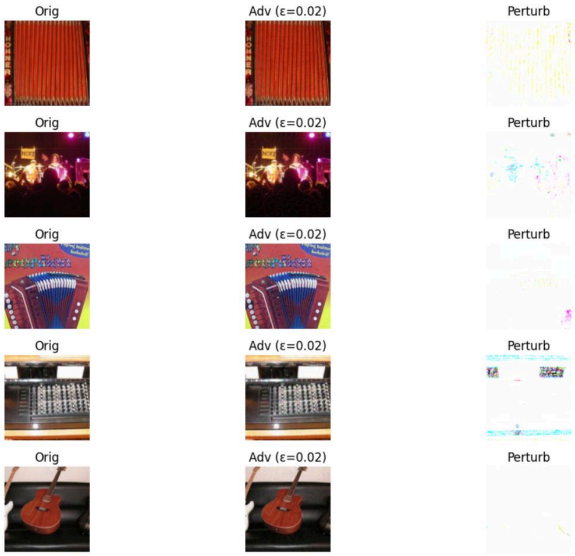

Figure 3: Images after Pixel-wise attack (FGSM)

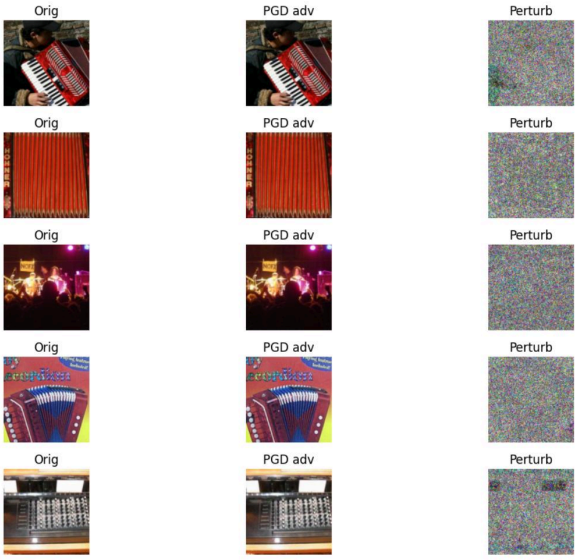

Figure 4: Images after PGD attack

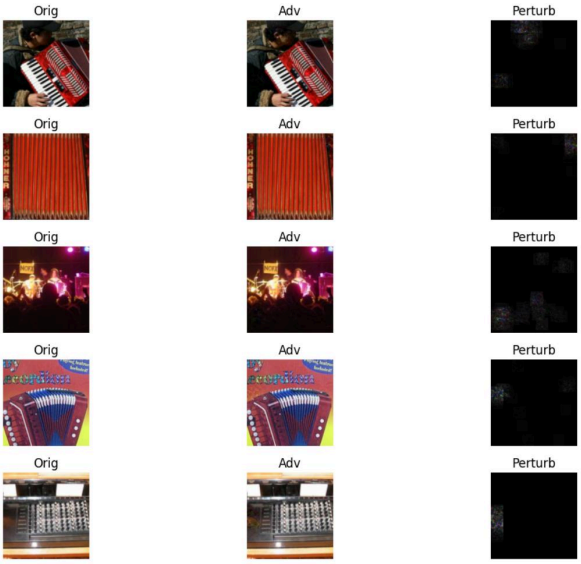

Figure 5: Images after targeted patch PGD