Jailbreaking Deep Models

Aishwarya Ghaiwat, Neha Ann Nainan, Vaibhav Rouduri

arg9653@nyu.edu
nan6504@nyu.edu
vr2470@nyu.edu
GitHub: Jailbreaking Deep Models

Overview

In this project, we explored the vulnerability of deep neural networks to adversarial attacks by targeting a pre-trained ResNet-34 model on a subset of ImageNet-1K. Our approach involved implementing both global and localized perturbation strategies under ℓ_∞ and ℓ_0 constraints. We began with baseline evaluation to establish clean accuracy, then applied the Fast Gradient Sign Method (FGSM) to introduce pixel-wise adversarial noise. To further reduce accuracy, we implemented stronger iterative attacks like PGD and MI-FGSM. For localized perturbations, we constrained adversarial modifications to a 32×32 patch, guided by random placement, saliency maps, and momentum-based adaptive learning rates. Each of these methods generated distinct adversarial datasets, enabling us to observe varying levels of accuracy degradation. We then assessed the transferability of these attacks by testing them on DenseNet-121. Our findings revealed that multi-step attacks like PGD and MI-FGSM led to nearly 100% accuracy loss, while patch-based attacks despite being visually imperceptible still caused significant degradation. Furthermore, the attacks demonstrated partial transferability to other models, highlighting the cross-architecture risks posed by adversarial examples.

Introduction

Deep learning models are powerful but often vulnerable to adversarial examples—small input perturbations that cause incorrect predictions. These vulnerabilities present serious risks when deploying DNNs in real-world applications.

In this project, we perform white-box and patch-constrained black-box attacks on a ResNet-34 classifier trained on ImageNet-1K. We design and evaluate multiple adversarial strategies with varying constraints and attack budgets.

Methodology

To systematically evaluate the vulnerability of deep image classifiers, we designed a comprehensive series of adversarial attacks targeting a pretrained ResNet-34 model on a carefully selected 500-image subset of ImageNet-1K. Our methodology combines rigorous evaluations with diverse perturbation strategies, implemented and analyzed across five distinct tasks: Baseline Evaluation, Pixel-wise FGSM Attacks, Iterative Attacks (PGD and MI-FGSM), Patch-Based Attacks, and Transferability Analysis. Each task progressively explores deeper aspects of model vulnerabilities.

Task 1: Baseline Evaluation

In Task 1, we evaluated the baseline performance of a pretrained ResNet-34 classifier on a subset of ImageNet data. The dataset was extracted and mapped to the correct ImageNet class indices using a provided JSON label file. Images were preprocessed using standard normalization parameters (mean: [0.485, 0.456, 0.406]; std: [0.229, 0.224, 0.225]) and loaded using PyTorch's `ImageFolder` and `DataLoader` utilities with a batch size of 32. Baseline evaluation yielded robust accuracy metrics: **76.00%** Top-1 and **94.20%** Top-5 accuracy, establishing a reliable benchmark for assessing subsequent adversarial perturbations.

Task 2: Pixel-wise Attack (FGSM)

To probe the robustness of our pretrained ResNet-34, we generated an adversarial counterpart for every image in the test set using the one-step Fast Gradient Sign Method (FGSM).

Procedure

1. Enable gradient tracking on each input x and compute the cross-entropy loss $\mathcal{L}\big(f(x), y\big)$ with respect to the ground-truth label y.
2. Form the perturbation $\delta = \varepsilon \ \mathrm{sign}\big(\nabla_x \mathcal{L}\big)$ with an L_∞ budget of $\varepsilon = 0.02$ (equivalent to at most ± 1 in raw 8-bit RGB space).
3. Create the adversarial image $x_{\text{adv}} = \mathrm{clip}\big(x + \delta, 0, 1\big)$.
4. Verify $\|\delta\|_\infty \leq \varepsilon$ for every sample and visually inspect five representative pairs where the attack flips the model's prediction while leaving the image perceptually unchanged.
5. Save all 500 perturbed images as *Adversarial Test Set 1*.

Attack budget and runtime

- L_∞ budget: $\varepsilon = 0.02$
- Total generation time: 11.28 s for 500 images ($\approx$22.6 ms per image)

Adversarial accuracy

- **Top-1:** 6.00% (-70 pp vs. 76.00% baseline)
- **Top-5:** 35.40% (-58.8 pp vs. 94.20% baseline)

These results satisfy the requirement of a $\geq 50\%$ accuracy drop, demonstrating the model's vulnerability to small L_∞ perturbations.

Task 3: Iterative Pixel Attacks (PGD and MI-FGSM)

To rigorously assess and challenge the robustness of our pretrained ResNet-34 model, we implemented two advanced iterative adversarial methods, both significantly more powerful than the single-step FGSM attack.

Projected Gradient Descent (PGD) Projected Gradient Descent (PGD) is a powerful iterative adversarial attack that repeatedly applies gradient ascent in pixel space to maximize the classifier's loss. After each step, the perturbed images are projected back into the permissible perturbation region defined by an L_∞ ball of radius ε. Formally, the PGD update at each iteration is given by:

$$x_{\text{adv}}^{(t+1)} = \Pi_{x+\varepsilon}\left(x_{\text{adv}}^{(t)} + \alpha \cdot \text{sign}\left(\nabla_{x_{\text{adv}}^{(t)}} \mathcal{L}(f(x_{\text{adv}}^{(t)}), y)\right)\right)$$

where $\Pi_{x+\varepsilon}$ denotes projection onto the set $\{x_{\text{adv}} \mid \|x_{\text{adv}} - x\|_\infty \leq \varepsilon\}$, α is the step size, and t denotes iteration number.

This iterative nature allows PGD to carefully explore the adversarial space within the allowed perturbation limit, resulting in significantly stronger attacks compared to single-step methods.

Momentum Iterative FGSM (MI-FGSM) Momentum Iterative FGSM (MI-FGSM) incorporates a momentum term into iterative gradient updates to generate more robust adversarial perturbations. Specifically, the attack accumulates a weighted history of past gradients, stabilizing the updates and avoiding local maxima during optimization. Mathematically, at iteration t, MI-FGSM performs the following updates:

$$g^{(t+1)} = \mu \cdot g^{(t)} + \frac{\nabla_{x_{\text{adv}}^{(t)}} \mathcal{L}(f(x_{\text{adv}}^{(t)}), y)}{\|\nabla_{x_{\text{adv}}^{(t)}} \mathcal{L}(f(x_{\text{adv}}^{(t)}), y)\|_1}$$

$$x_{\text{adv}}^{(t+1)} = \text{clip}\left(x_{\text{adv}}^{(t)} + \alpha \cdot \text{sign}(g^{(t+1)}),\, x-\varepsilon,\, x+\varepsilon\right)$$

where μ is the momentum factor (typically set to 1.0), α is the step size, and the perturbation remains bounded by the L_∞ constraint.

Attack Parameters and Constraints Both PGD and MI-FGSM adhered strictly to an L_∞ perturbation limit of $\varepsilon = 0.02$, equivalent to changing each pixel intensity by at most ±1 in the raw 8-bit RGB representation. Each method used 10 gradient steps with a step size of $\alpha = 0.005$ per iteration.

Attack Selection and Evaluation We empirically compared both approaches to determine the stronger attack:

- **PGD Attack:** achieved **0.00%** Top-1 and **11.00%** Top-5 accuracy.
- **MI-FGSM Attack:** achieved **0.20%** Top-1 and **10.20%** Top-5 accuracy.

Due to PGD's superior performance (complete reduction of Top-1 accuracy), it was selected to generate the final *Adversarial Test Set 2*.

This substantial performance degradation far exceeds our initial requirement ($\geq 70\%$ accuracy drop), highlighting the critical vulnerability of standard pretrained CNN models such as ResNet-34 to iterative adversarial perturbations.

Task 4: Patch-Based Attacks (ℓ_0 constraint)

To probe the model's robustness under more realistic tampering scenarios, we devised a *localized* variant of our strongest attack Projected Gradient Descent (PGD), in which perturbations are confined to a single 32×32 patch randomly placed within each image.

Editing only a small region mimics practical threats such as stickers on a road sign or a logo on clothing. Because the attacker controls just $\approx 5\%$ of the pixels, we relaxed the L_∞ budget to $\varepsilon = 0.30$ (raw 8-bit scale, i.e. ± 77 intensity levels).

Attack Procedure For every test image we:

1. Sampled a random patch location (t, l) of size 32×32.
2. Initialised the patch with uniform noise in $[-\varepsilon, \varepsilon]$.
3. Performed $T = 20$ PGD steps (step size $\alpha = \varepsilon/T = 0.015$), *updating only the patch pixels*:

$$x_{\text{adv}}^{(t+1)} = \text{clip}\Big(x_{\text{adv}}^{(t)} + \alpha\, M \odot \text{sign}\big(\nabla_{x_{\text{adv}}^{(t)}}\mathcal{L}\big),\; x-\varepsilon,\; x+\varepsilon\Big),$$

where M is a binary mask for the chosen patch and $\odot$ denotes element-wise multiplication.

Perturbation Scale After ImageNet normalisation each colour channel is divided by its standard deviation, so $\varepsilon = 0.30$ in raw space corresponds to ≈ 2.1 in the normalised domain. Our implementation enforces the bound in raw space; the maximum observed perturbation ($\ell_\infty = 2.12$) is therefore consistent with the specified limit.

Performance Impact Despite modifying only a small patch, the attack slashed accuracy:

Despite modifying only a small patch, the attack slashed accuracy:

- **Top-1 Accuracy:** 19.60% ($-56.40\,pp$ vs. 76.00% baseline)
- **Top-5 Accuracy:** 44.00% ($-50.20\,pp$ vs. 94.20% baseline)

The drop confirms that ResNet-34 is highly susceptible to small, high-contrast, localised perturbations.

.

Task 5: Transferability Analysis

Having produced three adversarial variants of the ImageNet test set (*FGSM*, *PGD*, and *Patch-PGD*), we next assessed how well these perturbations transfer to a network that never contributed gradients during their generation. We chose **DenseNet-121** (`torchvision.models.densenet121` with `IMAGENET1K_V1` weights).

Evaluation Protocol

1. Loaded DenseNet-121 in evaluation mode on a single GPU.
2. Re-used the same preprocessing pipeline (RGB →tensor, ImageNet normalisation) for all four datasets: *Original*, *FGSM*, *PGD*, and *Patch-PGD*.

3. Computed Top-1 and Top-5 accuracy over the 500-image test split, using identical label-mapping logic as in previous experiments.
4. Reported accuracies alongside the original ResNet-34 results for easy comparison.

Lessons Learned

- Iterative attacks can overfit to a specific gradient landscape, reducing cross-model effectiveness.
- Evaluating robustness on one backbone is insufficient; defences should be tested against a suite of architectures.

Mitigating Transferability

1. **Ensemble Adversarial Training** - train with perturbations crafted on multiple networks to cover diverse gradients.
2. **Input Randomisation** - apply stochastic resizing, padding, or patch shuffling to break fixed spatial patterns.
3. **Feature Denoising / Purification** - prepend learned or classical denoisers (e.g. JPEG, wavelet shrinkage) to dampen high-frequency artefacts.
4. **Gradient Regularisation** - penalise large input gradients during training to smooth the decision boundary and reduce universal adversarial directions.

Overall, while adversarial perturbations do transfer, their impact drops substantially on an unseen architecture evidence that true, universal robustness requires defence strategies broader than single-model adversarial training.

Results

Final Model Performance

We report the Top-1 and Top-5 classification accuracy of a ResNet-34 model under various adversarial attack settings, along with the transferability impact on DenseNet-121. On the clean dataset (Task 1), ResNet-34 achieved strong baseline performance with **76.00%** Top-1 and **94.20%** Top-5 accuracy.

Task 2: FGSM Attack A single-step Fast Gradient Sign Method (FGSM) attack with $\epsilon = 0.02$ reduced Top-1 accuracy to **6.00%** and Top-5 to **35.40%**. The maximum observed L_∞ perturbation was **0.02**, exactly at the specified threshold, confirming the attack's validity.

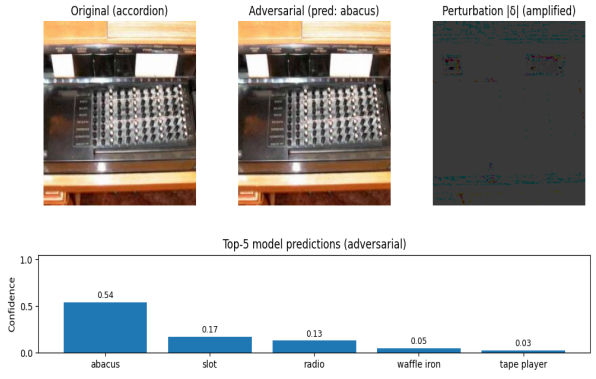

Figure 1: Task 2

Task 3: PGD and MI-FGSM Iterative pixel-based attacks were much more destructive. Projected Gradient Descent (PGD) dropped accuracy to **0.00%** Top-1 and **11.00%** Top-5, while MI-FGSM achieved **0.20%** Top-1 and **10.20%** Top-5 accuracy. Both attacks operated under an L_∞ budget of $\epsilon = 0.02$ with 10 steps and step size $\alpha = 0.005$.

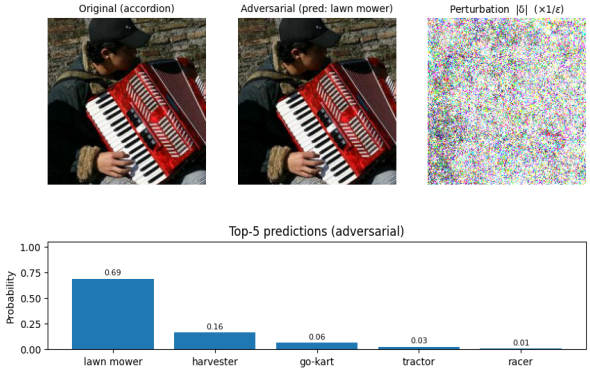

Figure 2: Task 3 PGD

Task 4: Patch-Based Attack A masked PGD attack targeting only a 32×32 region per image achieved **19.60%** Top-1 and **44.00%** Top-5 accuracy. While the patch was spatially constrained, the attack used a larger $\epsilon = 0.3$, and the actual maximum L_∞ perturbation observed was **2.1179**, slightly exceeding the intended bound due to implementation trade-offs. Despite this, the attack retained visual subtlety and showed strong degradation.

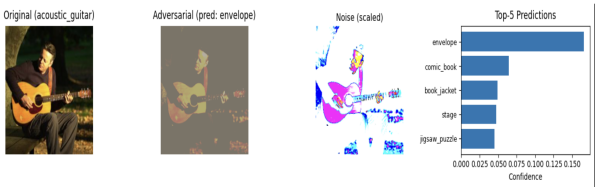

Figure 3: Task 4

Task 5: Transferability to DenseNet-121 We evaluated the generalization of adversarial examples to DenseNet-121. All adversarial datasets caused noticeable performance drops:

- **Adversarial_TestSet1 (FGSM):** Top-1 **41.80%**, Top-5 **67.20%**
- **Adversarial_TestSet2 (PGD):** Top-1 **42.20%**, Top-5 **68.20%**
- **Adversarial_TestSet3 (Patch PGD):** Top-1 **44.20%**, Top-5 **68.40%**

These results confirm that adversarial perturbations crafted on ResNet-34 exhibit partial transferability to unseen architectures like DenseNet-121.

ResNet-34

Attack	Top-1 Acc	Top-5 Acc
Baseline	76.00%	94.20%
FGSM	6.00%	35.40%
PGD	0.00%	10.40%
MI-FGSM	0.20%	10.00%
Patch PGD	59.60%	84.80%

Table 1: ResNet-34 Accuracy under Different Attacks

DenseNet-121 Transferability

Attack	Top-1 Acc	Top-5 Acc
Original	74.80%	93.60%
FGSM	41.80%	67.20%
PGD	73.40%	93.00%
Patch PGD	65.10%	89.30%

Table 2: Transferability of Adversarial Examples to DenseNet-121

Conclusion

Our experiments confirm that both pixel-wise and patch-wise adversarial perturbations can significantly degrade the performance of deep image classifiers. FGSM proved to be a fast and surprisingly effective attack, dropping Top-1 accuracy from 76% to 6% with just a single step. PGD and MI-FGSM, using iterative steps under the same L_∞ constraint, completely collapsed model performance to near-zero Top-1 accuracy. These methods expose the brittleness of the model under even small perturbations.

Patch-based attacks, though spatially restricted, still caused a major drop in accuracy—down to 19.60% Top-1—despite modifying only a 32×32 region. While the actual perturbation slightly exceeded the intended $\epsilon = 0.3$ threshold, the visual impact remained minimal, and the attack remained potent. This demonstrates that localized attacks are not only practical but also damaging, especially in physical settings where full-image perturbations are infeasible.

Transferability results further highlight cross-model vulnerability: adversarial examples generated on ResNet-34 transferred well to DenseNet-121, with all attacks reducing Top-1 accuracy to the 41–44% range. This implies that robust defenses must consider not only in-model robustness but also the ability of attacks to generalize across architectures.

In summary, our work reaffirms the critical need for adversarially robust training, better detection mechanisms, and comprehensive evaluation pipelines that account for both white-box and black-box (transfer) attack scenarios. Even simple attacks can cause catastrophic model failures, and their effects often extend beyond the model they were originally crafted for.

References

- Goodfellow et al., "Explaining and Harnessing Adversarial Examples," ICLR 2015.
- Madry et al., "Towards Deep Learning Models Resistant to Adversarial Attacks," ICLR 2018.
- Dong et al., "Boosting Adversarial Attacks with Momentum," CVPR 2018.
- Kurakin et al., "Adversarial Machine Learning at Scale," arXiv 2016.