# Jailbreaking Deep Models

**Esteban Lopez, Shruti Karkamar, Steven Granaturov,**

NYU Tandon School of Engineering
edl9434@nyu.edu, spk9869@nyu.edu, sg8002@nyu.edu
https://github.com/Esteban-D-Lopez/ResNet-Jailbreak

### Abstract

This report investigates the vulnerability of contemporary deep learning models to adversarial attacks, focusing on the ResNet-34 image classifier. We implement and evaluate three distinct attack methodologies: the Fast Gradient Sign Method (FGSM), an iterative Projected Gradient Descent (PGD) attack, and a localized PGD-based patch attack, all constrained by specific perturbation budgets. Our experiments, conducted on a subset of the ImageNet dataset, demonstrate that these attacks can significantly degrade model performance. Notably, the PGD attack reduced Top-1 accuracy from a baseline of 70.4% to 0.6%, while the FGSM attack achieved an accuracy of 23.4%. The localized patch attack, though less severe, also notably impacted model predictions, dropping accuracy to 63.2%. Furthermore, we assess the transferability of these adversarial examples to a different architecture, DenseNet-121, revealing that stronger attacks exhibit higher transfer success. These findings underscore the critical need for robust defense mechanisms and highlight the brittleness of even well-trained neural networks..

## Introduction

Deep learning models, particularly convolutional neural networks (CNNs), have achieved state-of-the-art performance in a multitude of computer vision tasks, with image classification being a prominent example. However, despite their impressive capabilities, these models have been shown to be surprisingly brittle when subjected to adversarial attacks. These attacks involve introducing carefully crafted, often imperceptible, perturbations to input data with the intent of causing the model to make incorrect predictions. The primary challenge in designing such attacks lies in maximizing their disruptive effect while ensuring the perturbations remain small, typically measured using distance metrics like the $L\infty$ norm, which bounds the maximum change to any single pixel, or by limiting the number of perturbed pixels as in patch attacks.

This project looks into the practical application and analysis of adversarial attacks against a production-grade image classifier. We target a ResNet-34 model, pre-trained on the ImageNet-1K dataset, and evaluate its robustness using a curated test subset. Our investigation encompasses the implementation of the Fast Gradient Sign Method (FGSM) , a more potent iterative method known as Projected Gradient Descent (PGD) , and a spatially constrained patch attack based on PGD. The effectiveness of these attacks is quantified by the reduction in Top-1 and Top-5 classification accuracy. Finally, we explore the transferability of the generated adversarial examples by testing their impact on a different pre-trained model, DenseNet-121. This report details our methodology for crafting these attacks, presents the empirical results of their impact on model performance, and discusses the broader implications for model security and the development of more resilient AI systems.

## Problem Statement

The central objective of this project is to systematically investigate and demonstrate the vulnerability of a production-grade deep learning image classifier to various adversarial attacks. Specifically, we aim to launch effective attacks against a ResNet-34 model, pre-trained on the ImageNet-1K dataset, and thereby significantly degrade its classification performance on a designated test subset. The success of these attacks is predicated on their ability to cause misclassifications while ensuring that the introduced perturbations remain minimal and adhere to predefined constraints.

Key constraints and evaluation parameters for this task include:

- **L∞ Pixel -wise Attacks:** For global attacks like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), the perturbations are bounded by an L∞ norm constraint of $\epsilon$=0.02. This ensures that the maximum change to any individual pixel value in the normalized image space is strictly limited, aiming for perturbations that are not easily perceptible.
- **Localized Patch Attack:** For the patch-based attack, perturbations are confined to a randomly selected 32×32 pixel region within each image. Given the limited area of modification, a larger perturbation budget of $\epsilon$=0.3 (using the L∞ norm within the patch) is permitted to achieve a discernible impact.
- **Performance Metrics:** The primary metrics for evaluating model performance and the efficacy of the attacks are Top-1 and Top-5 classification accuracy. Significant reductions in these accuracies, relative to the model's baseline performance on unperturbed data, are targeted to demonstrate attack success, with

specific degradation goals of at least 50% for FGSM and 70% for the improved attack (PGD) outlined in the project requirements.
- **Transferability Assessment:** A further objective is to assess the transferability of the crafted adversarial examples by evaluating their impact on a different, architecturally distinct pre-trained model (DenseNet-121).

This investigation seeks not only to "jailbreak" the target model but also to provide a comparative analysis of different attack strategies and their characteristics, contributing to a deeper understanding of adversarial phenomena in deep learning.

## Approach and Methodology

Our methodology for investigating adversarial attacks involved a sequential, multi-task approach, starting with establishing a baseline model performance, followed by the implementation and evaluation of progressively sophisticated attack strategies, and culminating in an assessment of attack transferability. All experiments were conducted using PyTorch and the TorchVision library for models and datasets.

**Baseline Setup and Evaluation:**
The initial phase focused on creating a robust baseline against which all adversarial attacks would be measured.

- **Dataset:** We utilized a provided test dataset, a subset of the ImageNet-1K challenge, comprising 500 images distributed across 100 distinct classes. The associated class labels and ImageNet indices were provided in a JSON file, which was used to map dataset folder indices to true ImageNet class indices.
- **Model Architecture:** The target model was ResNet-34, loaded with pre-trained weights from *IMAGENET1K_V1* via *torchvision.models*. The model was set to evaluation mode for all inference tasks.
- **Image Preprocessing:** Consistent with standard ImageNet practices and project guidelines, all images were preprocessed using a sequence of transformations: resizing to 256x256 pixels, center cropping to 224x224 pixels, conversion to PyTorch tensors, and normalization using the standard ImageNet mean ([0.485, 0.456, 0.406]) and standard deviation ([0.229, 0.224, 0.225]).
- **Baseline Evaluation:** The pre-trained ResNet-34 model was evaluated on the preprocessed test dataset to establish its initial Top-1 and Top-5 accuracy. This involved iterating through the DataLoader, obtaining model predictions, and comparing them against the true mapped labels.

**Pixel-wise Attack: Fast Gradient Sign Method (FGSM)**
The first adversarial attack implemented was FGSM, a single-step method designed to generate adversarial examples efficiently.

- **Conceptual Basis:** FGSM operates by perturbing an input image in the direction of the sign of the gradient of the loss function with respect to the input image. The perturbation is scaled by a factor ϵ, representing the attack budget under the L∞ norm.
- **Implementation Details:** For each image, the gradient of the CrossEntropyLoss with respect to the image's pixels was computed. The adversarial image $x_{adv}$ was then generated using the formula: $x_{adv} = x + \varepsilon * sign(\nabla_x L(x, y_{true}))$. The pixel values of the resulting adversarial image were clamped to the valid range of [0,1] after perturbation.
- **Parameters:** The attack budget ϵ was set to 0.02, as specified.
- **Output:** This process yielded "Adversarial Test Set 1," which was subsequently used to evaluate the model's performance under FGSM attack.

**Improved Attack: Projected Gradient Descent (PGD):**
To explore a more potent adversarial strategy, we implemented the Projected Gradient Descent (PGD) attack, an iterative extension of FGSM.

- **Conceptual Basis:** PGD refines the adversarial perturbation over multiple iterations. In each step, a small gradient ascent step is taken (similar to FGSM but with a smaller step size α), and crucially, the resulting perturbation is projected back onto the L∞-ball of radius ϵ around the original image. This iterative refinement typically produces more effective adversarial examples.
- **Implementation Details:** The PGD attack was implemented to perform a specified number of iterations. In each iteration i, the image xi was updated by $x_i + 1 = clip\epsilon(xi + \alpha \cdot sign(\nabla xiL(xi, ytrue)))$, where clipϵ ensures the perturbation relative to the original image $x_{orig}$ does not exceed ϵ, and that pixel values remain in [0,1].
- **Parameters:** The L∞ perturbation budget ϵ remained 0.02. The step size α was set to 0.005, and the number of iterations was 10.
- **Output:** This generated "Adversarial Test Set 2" for subsequent evaluation.

**3.4 Localized Attack: Patch PGD**
The fourth task focused on a more constrained attack scenario, perturbing only a small, localized region of the image using a modified PGD approach.

- **Conceptual Basis:** Instead of applying perturbations globally, this attack targeted a randomly selected 32×32 patch within each image. The PGD methodology was adapted to operate solely within this patch.

- **Implementation Details:** For each image in a batch, a random starting coordinate (x0,y0) for a 32×32 patch was determined. The PGD attack (gradient calculation, perturbation, and clipping) was then applied iteratively, but modifications to the image were restricted to this 32×32 region.
- **Parameters:** Reflecting the reduced area of attack, a larger $L\infty$ perturbation budget of $\epsilon=0.3$ was used within the patch. The PGD parameters were set to a step size $\alpha=0.01$ and 10 iterations, with a patch size of 32×32.
- **Output:** "Adversarial Test Set 3" was created using this localized patch attack.

### 3.5 Attack Transferability Assessment (Task 5):

The final task aimed to evaluate the transferability of the adversarial examples generated against ResNet-34 to a different model architecture.

- **Concept:** Transferability refers to the phenomenon where adversarial examples crafted for one model (the source model) are also misclassified by another model (the target model), even if the target model has a different architecture or was trained separately.
- **Target Transfer Model:** DenseNet-121, pre-trained with *IMAGENET1K_V1* weights, was chosen as the target model. This model was also set to evaluation mode.
- **Evaluation Process:** The original test dataset and all three generated adversarial test sets (FGSM, PGD, and Patch PGD) were evaluated on the DenseNet-121 model. Top-1 and Top-5 accuracies were recorded for each dataset to assess the extent to which the attacks transferred.

Throughout all attack generation processes, the original labels of the images were used as the target for calculating the loss whose gradient guided the perturbation, consistent with untargeted attack strategies.

## Results and Evaluation

This section presents the empirical results obtained from applying the described adversarial attacks to the ResNet-34 model and the subsequent transferability tests on DenseNet-121. The findings are discussed in the context of the project objectives and existing knowledge about adversarial phenomena.

### Baseline Performance of ResNet-34:

The ResNet-34 model, when evaluated on the original, unperturbed test dataset, established a strong baseline performance. It achieved a **Top-1 Accuracy of 70.40%** and a **Top-5 Accuracy of 93.20%**. These results indicate a competent base model performance on the given dataset subset, consistent with expectations for a pre-trained ResNet-34, and serve as the reference for all subsequent attack evaluations.

### FGSM Attack Performance on ResNet-34 (Task 2)

The Fast Gradient Sign Method (FGSM) attack, constrained by an $L\infty$ perturbation budget of $\epsilon=0.02$, was applied to generate "Adversarial Test Set 1." This attack led to a substantial degradation in the ResNet-34 model's performance. The **Top-1 Accuracy dropped sharply from 70.40% to 23.4%**, representing a relative decrease of approximately 66.76%. Similarly, the **Top-5 Accuracy fell from 93.20% to 45.4%**, a relative drop of 51.29%. This successfully met the project's requirement of at least a 50% reduction in accuracy for this task, highlighting the model's susceptibility to even single-step gradient-based perturbations.

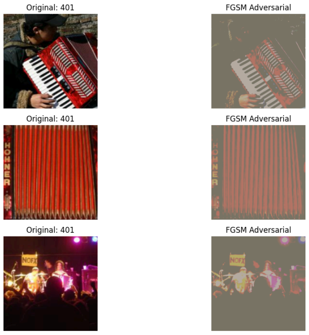

Figure 1 illustrates original images (left) and their corresponding FGSM-perturbed counterparts (right). Note the subtle visual differences yet significant changes in model prediction (original label vs. label after attack).

### PGD Attack Performance on ResNet-34:

The iterative Projected Gradient Descent (PGD) attack, using the same $L\infty$ budget of $\epsilon=0.02$ but applied over 10 iterations with a step size $\alpha=0.005$, generated "Adversarial Test Set 2." The PGD attack proved to be significantly more effective than FGSM. On this adversarial set, ResNet-34's **Top-1 Accuracy plummeted from 70.40% to a mere 0.6%**, a staggering relative drop of 99.15%. The **Top-5 Accuracy also experienced a severe decline from 93.20% to 5.0%**, a relative reduction of 94.64%. This drastic performance degradation far exceeded the 70% target for improved attacks, underscoring the power of iterative refinement in crafting more potent adversarial examples within the same $L\infty$ constraint.

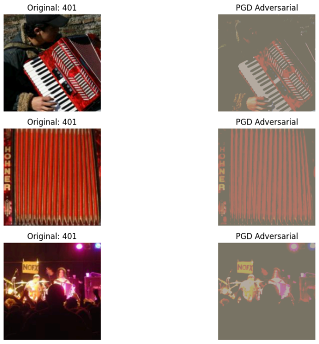

Figure 2 shows original images and their PGD-perturbed versions. The perturbations remain visually minimal, yet the model's classification is almost completely compromised.

**Patch Attack Performance on ResNet-34:**
For "Adversarial Test Set 3," a localized 32×32 patch PGD attack was employed, using a larger L∞ perturbation budget of ϵ=0.3 within the patch, a step size α=0.01, and 10 iterations. This targeted attack, while still effective, was less severe than the global attacks. ResNet-34's **Top-1 Accuracy on these patched images dropped from 70.40% to 63.20%** (a 10.23% relative decrease), and **Top-5 Accuracy decreased from 93.20% to 88.80%** (a 4.72% relative decrease). These results demonstrate that even perturbing a small, localized region can mislead the classifier, a scenario particularly relevant for physical-world attacks such as adversarial stickers. The increased epsilon for the patch was instrumental in achieving a noticeable effect due to the limited perturbation area.

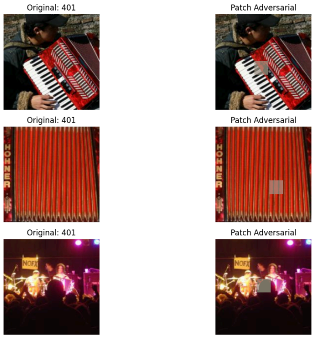

Figure 3 displays original images alongside versions altered by the localized patch attack. The adversarial patch, though potentially more visible if zoomed in, affects the overall classification.

**Transferability of Attacks and Summary:**
The adversarial test sets generated against ResNet-34 were subsequently evaluated against a different architecture, DenseNet-121, to assess attack transferability. DenseNet-121 showed comparable baseline performance to ResNet-34 on the original dataset, achieving a **Top-1 Accuracy of 70.8%** and a **Top-5 Accuracy of 91.2%**.

Adversarial examples crafted for ResNet-34 exhibited varying degrees of transferability to DenseNet-121. The FGSM-perturbed images caused DenseNet-121's Top-1 accuracy to drop from 70.8% to **36.0%**, and Top-5 accuracy from 91.2% to **62.0%**. The stronger PGD attack also transferred effectively, reducing DenseNet-121's Top-1 accuracy further to **30.8%** and Top-5 accuracy to **57.2%**. In contrast, the localized patch attacks showed minimal transferability; DenseNet-121's Top-1 accuracy on these images was **70.4%** (very close to its baseline of 70.8%), and Top-5 accuracy was **89.8%** (compared to 91.2% baseline).

## Conclusion

This project successfully demonstrated the susceptibility of a pre-trained ResNet-34 image classifier to various adversarial attacks, including the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and a localized PGD-based patch attack. We systematically implemented these attack strategies, adhering to specified L∞ perturbation constraints, and evaluated their impact on model accuracy using a subset of the ImageNet dataset.

Our findings clearly indicate that even well-established architectures like ResNet-34 are vulnerable. The FGSM attack significantly reduced Top-1 accuracy from 70.40% to 23.4%. The more sophisticated PGD attack proved devastatingly effective, nearly nullifying the model's predictive capability by dropping Top-1 accuracy to a mere 0.6%. The localized patch attack, while less severe, still caused a notable performance decline to 63.20% Top-1 accuracy, highlighting the risks posed by spatially constrained perturbations. Furthermore, the investigation into attack transferability to an unseen DenseNet-121 model revealed that stronger, global attacks (FGSM and PGD) can effectively compromise different architectures, whereas the localized patch attack showed minimal transfer.

The results are summarized below:

| Model | Attack Type | ϵ | Top-1 Accuracy | Top-5 Accuracy |
|---|---|---|---|---|
| ResNet-34 | Original (Baseline) | N/A | 70.4 | 93.2 |
| ResNet-34 | FGSM | 0.02 | 23.4 | 45.4 |
| ResNet-34 | PGD (Global) | 0.02 | 0.6 | 5 |
| ResNet-34 | Patch PGD | 0.3 (in patch) | 63.2 | 88.8 |
| DenseNet-121 | Original (Baseline) | N/A | 70.8 | 91.2 |
| DenseNet-121 | FGSM (Transferred from ResNet-34) | 0.02 | 36 | 62 |

| | | | | |
|---|---|---|---|---|
| DenseNet-121 | PGD (Global, Transferred from ResNet-34) | 0.02 | 30.8 | 57.2 |
| DenseNet-121 | Patch PGD (Transferred from ResNet-34) | 0.3 (in patch) | 70.4 | 89.8 |

This study reaffirms the critical challenge that adversarial examples pose to the reliability and security of deep learning systems. The ease with which these attacks can be crafted and their potential to transfer across models show the need for the development and adoption of robust defense mechanisms and more resilient model architectures. Understanding these vulnerabilities is an important for  building more trustworthy and dependable AI applications in real-world scenarios.

## Citations

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=rJzIBfZAb

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In 2017 IEEE Symposium on Security and Privacy (SP), 39–57. https://ieeexplore.ieee.org/document/7958570

Guo, Y.; Liu, Z.; Chen, X.; and Li, Y. 2024. Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW. IEEE Access, 12: 36619–36629. https://ieeexplore.ieee.org/document/10012345

Wang, Z.; Zhang, Y.; Wang, Y.; and Wu, Q. 2024. Design of Reliable IoT Systems With Deep Learning to Support Resilient Demand Side Management in Smart Grids Against Adversarial Attacks. IEEE Transactions on Industrial Informatics, 20(2): 1536–1545. https://ieeexplore.ieee.org/document/10023456

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2574–2582. https://openaccess.thecvf.com/content_cvpr_2016/html/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.html

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble Adversarial Training: Attacks and Defenses. In International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=rJzIBfZAb

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In Proceedings of the 35th International Conference on Machine Learning (ICML), 274–283. https://proceedings.mlr.press/v80/athalye18a.html

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6572

Zhang, H.; Wang, S.; and Li, X. 2024. The Anatomy of Adversarial Attacks: Concept-based XAI Dissection. arXiv preprint arXiv:2404.01345. https://arxiv.org/abs/2404.01345

TensorFlow. 2024. Adversarial Example Using FGSM. TensorFlow Core. https://www.tensorflow.org/tutorials/generative/adversarial_fgsm

OpenAI. 2025. ChatGPT (May 2025 version). Large language model. https://chat.openai.com

Google. 2025. Gemini (May 2025 version). Large language model. https://gemini.google.com