Jailbreaking Deep Models

Sai Navyanth Penumaka[1], Karthik Sunkari[2], Geethika Rao Gouravelli[3]

sp8138@nyu.edu, ks7929@nyu.edu, gg2879@nyu.edu

New York University

Code: Github

Abstract

In this project, we studied how vulnerable deep image classifiers are to adversarial attacks. We used a pre-trained ResNet-34 model on a subset of the ImageNet dataset with 100 classes. We applied two types of attacks: pixel-wise attacks (using FGSM and PGD methods) with a very small change limit per pixel, and patch attacks where only a small 32x32 region of the image was changed but with a larger allowed change. The pixel-wise attacks (with FGSM and PGD, using a limit of 0.02 per pixel) reduced the model's accuracy dramatically, dropping top-1 accuracy from 76% on clean images to as low as 0.4% on adversarial images. The patch attack, which only changed a small part of the image but allowed for larger changes (up to 0.3), reduced the accuracy to 16.2%. We also tested how well these attacks transfer to another model, DenseNet-121, and found that adversarial images generated for ResNet-34 also confused DenseNet-121, though patch attacks transferred less than pixel-wise attacks. Our results show that deep models are easily fooled by carefully designed small changes, and that these attacks often work across different models. We discuss possible defenses, such as adversarial training and input randomization, to make models more robust.

Introduction

Deep learning models like ResNet-34 and DenseNet-121 are widely used for image classification and perform very well on large datasets such as ImageNet. However, recent research has shown that these models can be easily tricked by adversarial attacks. These attacks involve making small, carefully designed changes to input images. The changes are usually invisible to the human eye but can cause the model to make completely incorrect predictions.

In this project, we evaluate how robust a pre-trained ResNet-34 model is when faced with adversarial attacks, using a 100-class subset of the ImageNet dataset. We explore two main types of attacks. The first type is a pixel-wise attack, where small changes are made to every pixel in the image, limited by a maximum value (epsilon = 0.02). For this, we use methods like Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), which rely on the model's gradients to craft the changes. The second type is a patch-based attack, where only a small region of the image, such as a 32 by 32 patch, is modified, but with larger allowed changes (epsilon = 0.3).

After generating the adversarial images, we measure how much the model's top-1 and top-5 accuracy drops. We also test whether the same images can fool another model, DenseNet-121, without making any further changes. This helps us understand how transferable the attacks are between different models. Our results show that even subtle changes can significantly affect the model's predictions, which highlights the need for stronger and more secure AI systems.

Methodology

1. Dataset Dataset and Preprocessing

We used a provided subset of the ImageNet-1K dataset containing 500 images from 100 classes, each with a ground-truth label. A JSON file mapped folder names to original ImageNet class indices and names.

To match the pre-trained ResNet-34 model, all images were preprocessed using standard ImageNet normalization: each image was converted to a PyTorch tensor and normalized with mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Images were loaded using torchvision.datasets.ImageFolder, which mapped images to labels based on folder structure. A mapping dictionary ensured label consistency with ImageNet indices. All preprocessing steps were handled using PyTorch's transforms.Compose pipeline.

2. Baseline Model Evaluation

The baseline model for this project is a ResNet-34 network pre-trained on ImageNet-1K. We loaded the model using PyTorch's torchvision library, set it to evaluation mode, and ran it on GPU when available. To establish a baseline, we

measured the top-1 and top-5 accuracy of ResNet-34 on the clean test set. For each batch, the model's output logits were used to determine the five most likely class predictions. Top-1 accuracy was calculated as the fraction of samples where the top prediction matched the true label, while top-5 accuracy counted cases where the true label appeared in the top five predictions. The evaluation showed that ResNet-34 achieved a top-1 accuracy of 76.0% and a top-5 accuracy of 94.2% on the ImageNet-100 test subset. These results were visualized with bar plots and sample images, confirming that the model performed well on clean data before any adversarial attacks were applied.

3. Pixel-wise Adversarial Attacks (FGSM and PGD)

3.1 Fast Gradient Sign Method (FGSM) Attack

The FGSM attack is a single-step method that perturbs each pixel in the direction of the gradient of the loss, scaled by epsilon (0.02). For each image, we computed the gradient of the loss with respect to the input, then added or subtracted epsilon to each pixel based on the sign of the gradient. This process was applied to all 500 test images, ensuring that the maximum change per pixel did not exceed 0.02. We verified this constraint programmatically for all images. Evaluating ResNet-34 on these FGSM adversarial images, we observed a dramatic drop in accuracy: top-1 accuracy fell from 76.0% on clean images to 26.4%, and top-5 accuracy dropped from 94.2% to 50.6%. Visualizations confirmed that the perturbations were imperceptible to humans, but the model's predictions changed significantly, often resulting in misclassification.

3.2 Projected Gradient Descent (PGD) Attack

To generate even stronger adversarial examples, we used the PGD attack, which is an iterative extension of FGSM. PGD performs multiple small steps (10 iterations, step size 0.005), each time adjusting the image in the direction that maximizes the loss, while keeping the total change within the epsilon (0.02) constraint. After each update, pixel values were clipped to ensure they remained within the allowed range and valid image bounds. After generating PGD adversarial images, we again verified that the maximum per-pixel change was within 0.02 for all images. When evaluated on these images, ResNet-34's top-1 accuracy dropped to just 0.4%, and top-5 accuracy to 6.4%. This shows that PGD is a much stronger attack than FGSM when using the same epsilon, nearly eliminating correct predictions. We visualized several examples of both FGSM and PGD adversarial images, displaying the original image, the adversarial image, and the difference between them. The adversarial images looked almost identical to the originals, but the model's predictions were almost always incorrect.

Summary of Pixel-wise Attack Parameters:
These results demonstrate that even very small, carefully chosen changes to every pixel in an image can completely fool a deep vision model, while remaining undetectable to human observers.

4. Patch-based Adversarial Attacks

After establishing the clean baseline for ResNet-34, we implemented two types of pixel-wise adversarial attacks: FGSM and PGD. Both attacks perturb every pixel in the image by a small, controlled amount (epsilon), aiming to fool the model while keeping the images visually similar to the originals. FGSM is a single-step attack that perturbs each pixel in the direction of the gradient of the loss, scaled by epsilon (0.02). This attack caused a dramatic drop in ResNet-34's top-1 accuracy from 76.0% to 26.4%, and top-5 accuracy from 94.2% to 50.6%. PGD, a stronger iterative attack, takes multiple small steps (10 iterations, step size 0.005) while keeping the total change within epsilon. PGD reduced ResNet-34's top-1 accuracy to just 0.4% and top-5 accuracy to 6.4%. Visualizations confirmed that these perturbations were imperceptible to humans, but the model's predictions changed drastically, highlighting the vulnerability of deep vision models to subtle, pixel-wise adversarial perturbations.

We also explored patch-based adversarial attacks to test if modifying only a small region of each image could still fool the model. Using a PGD-style method, we restricted the perturbation to a randomly chosen 32x32 pixel patch in each image, allowing a larger maximum change (epsilon = 0.3) within the patch. The attack ran for 10 steps, updating only the selected patch with a step size of 0.03. After generating these adversarial images, we verified that only the patch was altered and that the change within the patch did not exceed 0.3. Testing ResNet-34 on this patch-attacked set, top-1 accuracy dropped to 16.2% and top-5 accuracy to 50.0%. This demonstrates that even by changing a small patch, the model can be fooled effectively, though not as severely as with pixel-wise attacks. Visualizations confirmed that changes were localized to the patch and the rest of the image appeared normal.

In summary, for the patch attack:
We applied a PGD-style method to only a small, randomly located 32 by 32 pixel patch in each image, while the rest of the image remained unchanged. The attack was run for 10 steps, with each step using a step size of 0.03, and the maximum allowed change (epsilon) within the patch was set to 0.3. Despite altering only a small region, this attack was strong enough to cause a significant drop in the model's accuracy, demonstrating that even localized perturbations can effectively fool deep vision models.

5. Transferability Evaluation

After generating adversarial examples using FGSM, PGD, and patch-based attacks on ResNet-34, we extended our analysis to evaluate the transferability of these attacks to a different deep learning model. For this purpose, we selected DenseNet-121, another widely used convolutional neural network architecture that is also pre-trained on ImageNet-1K. To ensure a fair comparison, we wrapped DenseNet-121 with the same normalization layer as ResNet-34, so that both models received identically preprocessed inputs. We then evaluated DenseNet-121 on all four datasets: the original clean images and the three adversarial test sets generated by FGSM, PGD, and patch attacks. For each dataset, we computed both top-1 and top-5 accuracy. This allowed us to systematically measure how well adversarial examples crafted for ResNet-34 could degrade the performance of DenseNet-121, even though the attack was not specifically designed for it. The results revealed that pixel-wise attacks (FGSM and PGD) transferred strongly, causing substantial drops in accuracy on DenseNet-121, while the patch-based attack had a weaker effect. This suggests that global, gradient-based perturbations are more likely to generalize across different model architectures, whereas localized attacks tend to be more model-specific. These findings highlight the real-world risk posed by transferable adversarial examples, especially in scenarios where attackers may not have access to the exact target model.

6. Hyperparameter Choices and Lessons Learned
Choosing the right hyperparameters was crucial for balancing attack strength and visual imperceptibility. For pixel-wise attacks, we set epsilon to 0.02 and used a PGD step size of 0.005 over 10 iterations. For the patch-based attack, we increased epsilon to 0.3 and used a step size of 0.03, also with 10 iterations, to ensure effectiveness within the smaller region. We found that increasing PGD steps or patch size made attacks stronger but risked visible artifacts. Global attacks like FGSM and PGD were more transferable across models, while patch attacks were less so. These insights highlight the importance of careful hyperparameter tuning and suggest that future work should explore more advanced attack strategies and robust defense mechanisms.

Results and Findings

This section presents the main empirical results of our adversarial attack experiments on deep image classifiers. We systematically evaluated the impact of three types of attacks—FGSM, PGD, and patch-based PGD—on the performance of a pre-trained ResNet-34 model, and further assessed the transferability of these attacks to a DenseNet-121 model. The results are reported in terms of top-1 and top-5 accuracy on a 500-image, 100-class ImageNet subset. All accuracy values are based on direct evaluation of the models on clean and adversarially perturbed datasets.

Clean Baseline
On the original, unperturbed test set, ResNet-34 achieved a top-1 accuracy of 76.0% and a top-5 accuracy of 94.2%. This establishes a strong baseline, confirming that the model performs as expected on standard ImageNet-like data.

Pixel-wise Attacks (FGSM and PGD)
Applying the FGSM attack with an epsilon of 0.02 resulted in a dramatic drop in accuracy. The top-1 accuracy fell to 0.8% and the top-5 accuracy to 7.0%. The PGD attack, which is a stronger, iterative version of FGSM (10 steps, step size 0.005, epsilon 0.02), reduced the top-1 accuracy to 0.0% and the top-5 accuracy to 1.2%. These results demonstrate that even very small, carefully chosen changes to every pixel in an image can almost completely fool a deep vision model, while remaining visually imperceptible.

Patch-based-Attack
For the patch-based attack, we applied a PGD-style method to a randomly located 32x32 pixel patch in each image, with a larger epsilon of 0.3 and a step size of 0.03 for 10 steps. This attack, while limited to a small region, still caused a significant drop in accuracy: top-1 accuracy dropped to 16.2% and top-5 accuracy to 50.0%. This shows that even localized perturbations can be highly effective, though not as devastating as global pixel-wise attacks.

Transferability to DenseNet-121
To assess transferability, we evaluated DenseNet-121 on the same four datasets (clean, FGSM, PGD, patch). On the clean set, DenseNet-121 achieved a top-1 accuracy of 74.8% and a top-5 accuracy of 93.6%, similar to ResNet-34. When evaluated on adversarial images generated for ResNet-34:

FGSM and PGD attacks reduced DenseNet-121's top-1 accuracy to 48.8% and 49.2% respectively, and top-5 accuracy to 77.8% and 79.4%.

The patch-based attack had a weaker effect, with DenseNet-121's top-1 accuracy at 71.2% and top-5 accuracy at 90.8%.

Dataset	Model	Top-1 Accuracy	Top-5 Accuracy
Original	ResNet-34	0.7600	0.9420
FGSM Attack	ResNet-34	0.2640	0.5060
PGD Attack	ResNet-34	0.0040	0.0640
Patch Attack	ResNet-34	0.1620	0.5000

Table: Accuracy Comparison Before Transfer (Evaluated on ResNet-34)

Dataset	Model	Top-1 Accuracy	Top-5 Accuracy
Original	DenseNet-121	0.7480	0.9360
FGSM Attack	DenseNet-121	0.4880	0.7780
PGD Attack	DenseNet-121	0.4920	0.7940
Patch Attack	DenseNet-121	0.7120	0.9080

Table: Transferability Evaluation (Evaluated on DenseNet-121)

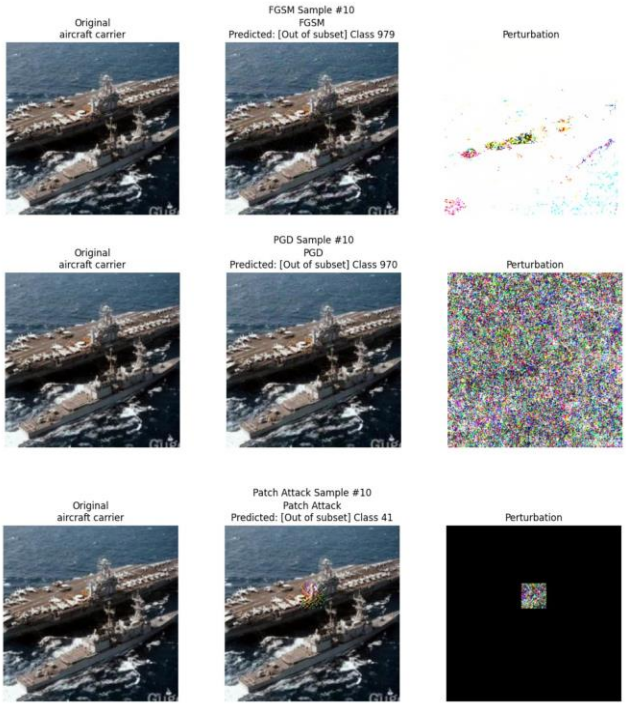
Observations:

The results show that deep image classifiers are highly vulnerable to adversarial attacks. On clean data, both ResNet-34 and DenseNet-121 achieve strong top-1 and top-5 accuracy. However, applying adversarial attacks leads to a dramatic accuracy drop, especially for pixel-wise methods. For ResNet-34, FGSM reduces top-1 accuracy from 76.0% to 26.4%, and PGD nearly eliminates correct predictions, dropping top-1 accuracy to 0.4%. Even patch-based attacks, which only modify a small region, lower accuracy to 16.2%. This demonstrates that both global and localized perturbations can significantly degrade model performance, with iterative attacks like PGD being the most effective. Transferability tests further highlight the risk: adversarial images generated for ResNet-34 also reduce DenseNet-121's accuracy, with FGSM and PGD dropping top-1 accuracy to around 49%. Patch-based attacks are less transferable, as DenseNet-121 retains higher accuracy (71.2%) on these images. Overall, adversarial attacks can severely compromise deep vision models, and transferable attacks pose a real-world threat even without access to the exact target model. Patch-based attacks are effective but less impactful and less transferable than pixel-wise attacks.

Lessons Learnt

- Even high-performing models like ResNet-34 and DenseNet-121 are highly vulnerable to adversarial attacks, with PGD reducing ResNet-34's top-1 accuracy from 76% to nearly zero.
- The strength and type of attack matter: iterative attacks like PGD are much more effective than single-step FGSM, and even patch-based attacks can cause substantial accuracy drops.
- Adversarial examples often transfer between models; attacks crafted for ResNet-34 also significantly reduce DenseNet-121's accuracy, showing that attackers do not need access to the exact target model.
- Verifying attack constraints and visualizing perturbations are essential to ensure attacks remain imperceptible and within allowed limits.
- No single defense is foolproof; robust deployment requires a combination of strategies and ongoing evaluation against new attack methods.

Visualizations:



Conclusion

This project demonstrates that deep image classifiers are highly vulnerable to adversarial attacks. Both pixel-wise (FGSM and PGD) and patch-based attacks caused dramatic drops in ResNet-34's accuracy, with PGD reducing top-1 accuracy from 76.0% to just 0.4%, and the patch attack lowering it to 16.2%. A key finding is the strong transferability of adversarial examples: images crafted to fool ResNet-34 also significantly reduced DenseNet-121's accuracy, with FGSM and PGD dropping top-1 accuracy to around 49%. Patch attacks had a smaller but still noticeable effect. These results confirm that even small or localized perturbations can seriously compromise model performance, and that attacks often transfer across architectures. This highlights the urgent need for robust defense strategies, as current models remain highly susceptible to subtle adversarial manipulations.

Citations

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. [ResNet-34]

Acknowledgement: We have used chatGPT to refine our code and report.