Jailbreaking Deep Models

Prajna Acharya (pa2532), Akshat Singh (as20255)

Overview

Deep neural networks have achieved remarkable performance on visual recognition tasks [1], yet they remain highly vulnerable to adversarial perturbations i.e. carefully crafted, imperceptible modifications that induce misclassification. In this project, we evaluate the vulnerability of a production-grade ResNet-34 to both pixel-wise attacks (FGSM, I-FGSM, MI-FGSM, PGD) and localized patch-based attacks (32×32 Patch-PGD with momentum and restarts). A single-step FGSM ($\epsilon = 0.02$) cuts Top-1 accuracy from $\approx 76\%$ to below 7%, while iterative and momentum-based methods drive it under 5%. Remarkably, small 32×32 patches covering just 2% of the image induce comparable drops, and these adversarial examples transfer effectively to SqueezeNet, ViT-B_16 and Swin-V2. All code, adversarial datasets, and notebooks are available at https://github.com/prajna-gajendra-acharya/Jailbreaking-Deep-Models

Task 1: Baseline Evaluation

Originally ImageNet-1K contains 1000 classes, here we evaluate a pretrained ResNet-34 on the preprocessed version of the test set which is a subset taken from 100 classes of the dataset.

Metric	Accuracy (%)
Top-1 Accuracy	76.00
Top-5 Accuracy	94.20

Table 1: Baseline

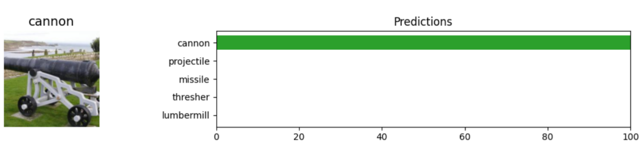

Figure 1: Baseline

Task 2: Pixel-Wise FGSM Attack

We generate pixel-wise adversarial examples on the clean test set using the Single Step Fast Gradient Sign Method ($\epsilon = 0.02$) and measure the impact on ResNet-34 [2].

Metric	Accuracy (%)
Top-1 Accuracy	6.20
Top-5 Accuracy	35.60

Table 2: FGSM

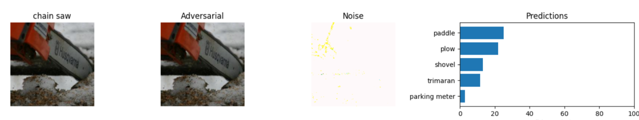

Figure 2: Single Step Fast Gradient Sign Method

Task 3: Improved L_∞ Attack

Attack 1: Instead of a single perturbation step, I-FGSM (Iterative Fast Gradient Sign Method) applies FGSM repeatedly in small increments (step size α) for n iterations, projecting the cumulative perturbation back into the ℓ_∞-ball of radius ϵ after each step. This iterative procedure yields stronger, more targeted adversarial examples and typically causes greater degradation in model accuracy compared to the single step FGSM on Resnet-34.

Parameters:
$\epsilon = 0.02$
$\alpha = 0.005$
$n = 60$

Metric	Accuracy (%)
Top-1 Accuracy	0.00
Top-5 Accuracy	6.20

Table 3: I-FGSM

Attack 2: MI-FGSM (Momentum Iterative Fast Gradient

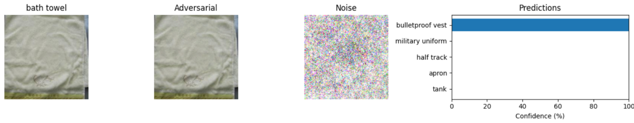

Figure 3: Iterative Fast Gradient Sign Method

Sign Method) introduces a momentum term to I-FGSM, accumulating the normalized gradient over iterations before taking each step, which stabilizes update directions and further strengthens the adversarial perturbation.
Parameters:
$\epsilon = 0.02$
$\alpha = 0.005$
$n = 60$
$\mu = 0.5$

Metric	Accuracy (%)
Top-1 Accuracy	0.00
Top-5 Accuracy	4.40

Table 4: MI-FGSM

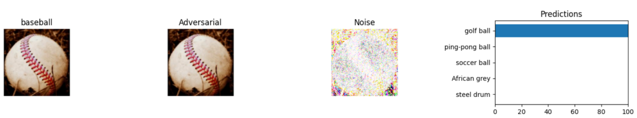

Figure 4: Momentum Iterative Fast Gradient Sign Method

Attack 3: PGD (Projected Gradient Descent) applies iterative FGSM steps, projecting the perturbation back into the ℓ_∞-ball of radius ϵ after each update, with an optional random start and restarts to maximize attack strength.
Parameters:
$\epsilon = 0.02$
$\alpha = 0.005$
$n = 60$
restarts=10

Metric	Accuracy (%)
Top-1 Accuracy	0.00
Top-5 Accuracy	6.20

Table 5: PGD

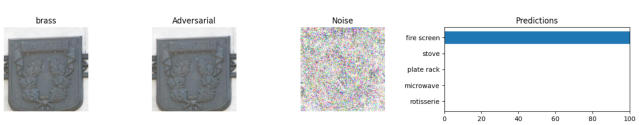

Figure 5: Projected Gradient Descent

Task 4: Patch-Based Attack

Attack: 32×32 Patch-PGD applies multi-step PGD updates confined to a random square patch, uses Nesterov momentum, and repeats with 10 random restarts—always keeping the adversary with the highest loss.
Parameters:
$\epsilon = 0.5$
$\alpha = 0.08$
$n = 60$
$\mu = 0.5$
restarts = 10
patch size = 32×32

Metric	Accuracy (%)
Top-1 Accuracy	12.00
Top-5 Accuracy	58.20

Table 6: Patch PGD with Momentum

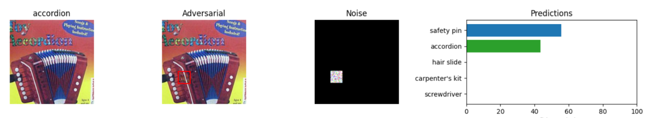

Figure 6: 32x32 Patch PGD with Momentum

Task 5: Transferring Attacks

This task evaluates the adversarial robustness of three image classification models [3]—SqueezeNet, Vision Transformer (ViT), and Swin V2—on a subset of the ImageNet dataset. We subject each model to the same attack methods which were discussed above and observe how adversarial perturbations affect top-k accuracy and prediction confidence.

Attack Methods

We implemented and applied the following adversarial attacks on each of our chosen pre-trained networks:

FGSM (Fast Gradient Sign Method)
Parameters:
$\epsilon = 0.02$

I-FGSM (Iterative FGSM)
Parameters:
$\epsilon = 0.02$
$\alpha = 0.005$
$n = 60$

MI-FGSM (Momentum Iterative FGSM)
Parameters:
$\epsilon = 0.02$
$\alpha = 0.005$
$n = 60$
$\mu = 0.5$

PGD (Projected Gradient Descent)
Parameters:
$\epsilon = 0.02$
$\alpha = 0.005$
$n = 60$
restarts = 10

Patch-PGD (Patch-based PGD with Momentum)
Parameters:
$\epsilon = 0.5$
$\alpha = 0.08$
$n = 60$
$\mu = 0.5$
restarts = 10
patch size = 32×32
Each model was evaluated on clean and adversarial samples to assess the change in top-1 and top-5 performance.

- *ResNet-34:* Residual connections preserve gradient flow, enabling both slight single-step resistance and full collapse under iterative attacks. Its deep hierarchy offers partial Top-5 retention under FGSM but little defense against PGD.
- *SqueezeNet1_0:* Extreme parameter efficiency and shallow depth yield very fragile decision boundaries; any perturbation—global or local—quickly overwhelms its minimal feature set.
- *ViT-B_16:* Global self-attention makes patch embeddings instantly influential, so pixel-wise attacks port nearly uniformly across tokens, while patch attacks must craft highly salient tokens to override the consensus of unperturbed patches.
- *Swin-V2:* Local windowed attention and hierarchical merging confine perturbations initially, giving stronger resistance to single-step and localized attacks. However, iterative attacks eventually propagate adversarial signals across windows and scales.

Model	Baseline	FGSM	I-FGSM	MI-FGSM	PGD	Patch-PGD
ResNet-34	76.00%	6.20%	0.00%	0.00%	0.00%	12.00%
SqueezeNet1_0	55.60%	0.80%	0.00%	0.00%	0.00%	2.20%
ViT_B_16	91.60%	45.40%	7.20%	4.40%	3.80%	23.20%
Swin-V2	78.40%	28.00%	0.00%	0.00%	0.00%	24.20%

Table 7: Top-1 accuracies of multiple pre-trained networks for different attack methods

Model	Baseline	FGSM	I-FGSM	MI-FGSM	PGD	Patch-PGD
ResNet-34	94.20%	35.60%	6.20%	4.40%	6.20%	58.20%
SqueezeNet1_0	79.20%	14.00%	3.40%	2.40%	3.60%	28.00%
ViT_B_16	99.60%	76.00%	21.00%	16.20%	15.60%	74.60%
Swin-V2	97.60%	48.80%	0.40%	0.20%	0.00%	72.40%

Table 8: Top-5 accuracies of multiple pre-trained networks for different attack methods

Conclusion

We conducted a systematic evaluation of four standard vision architectures— ResNet-34, SqueezeNet1_0, ViT-B_16, and Swin-V2— under five white-box adversarial attack methods (FGSM, I-FGSM, MI-FGSM, PGD, and 32×32 Patch-PGD). Table 7 (Top-1 accuracies) and Table 8 (Top-5 accuracies) summarize the results.

Pixel-wise Attacks (FGSM, I-FGSM, MI-FGSM, PGD): All architectures suffer catastrophic Top-1 drops under iterative attacks, often reaching near 0%. Single-step FGSM already slashes accuracy into the single digits for ResNet-34 and SqueezeNet, and below 50% for ViT and Swin. Iterative methods exploit accessible gradients and model smoothness to refine perturbations, overcoming any mild single-step resilience.

Patch-PGD (Localized Attacks): Restricting the perturbation to a 32×32 patch reduces the overall damage but still induces significant accuracy loss. ResNet-34 and SqueezeNet retain a small fraction of correct predictions ,ViT-B_16 retains, and Swin-V2 retains. Patch attacks demonstrate that even localized adversarial features can hijack global or hierarchical representations.

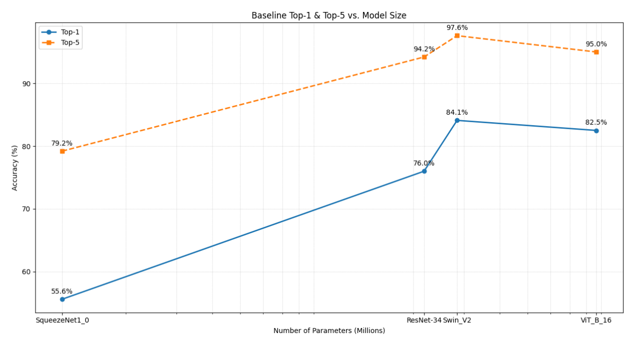

Figure 7: Parameter Vs Accuracy - Baseline

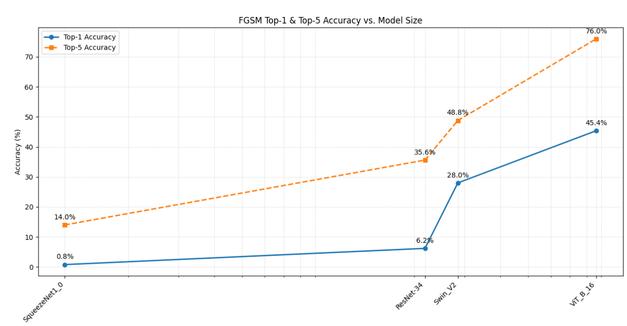

Figure 8: Parameter Vs Accuracy - FGSM

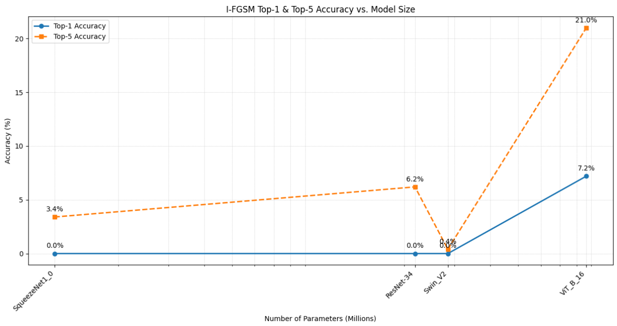

Figure 9: Parameter Vs Accuracy - IFGSM

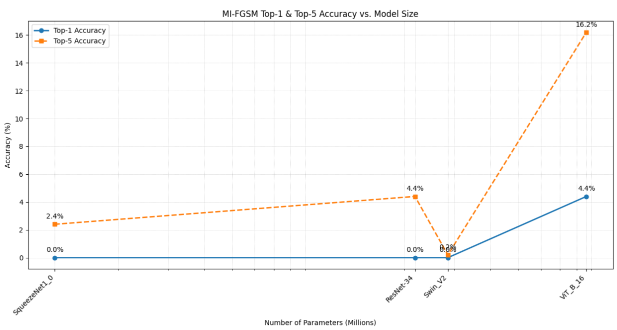

Figure 10: Parameter Vs Accuracy - MIFGSM

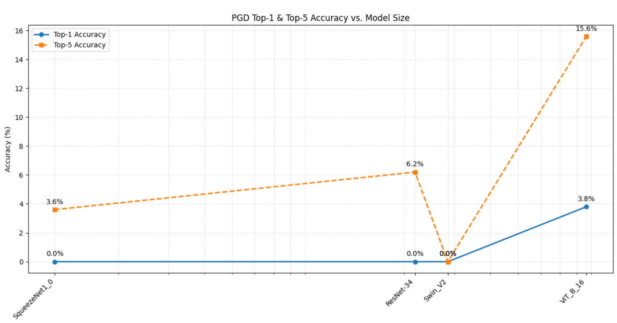

Figure 11: Parameter Vs Accuracy - PGD

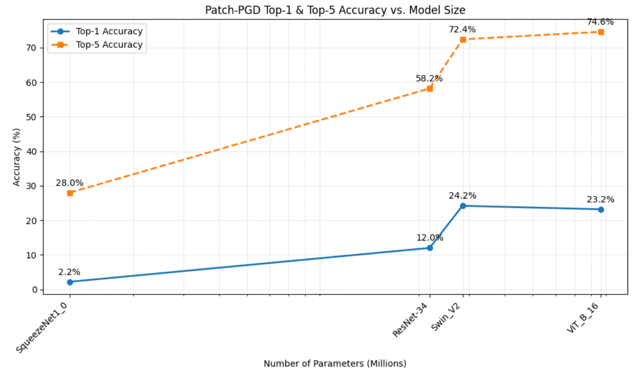

Figure 12: Parameter Vs Accuracy - Patch PGD with Momentum

Parameter-Robustness Trend: Across all attack methods, models with more parameters consistently suffer smaller accuracy drops [4]. Compact networks like SqueezeNet1_0 (1.2 M) collapse under adversarial perturbations, whereas larger architectures such as Swin-V2 (28.4 M) and ViT-B_16 (86.6 M) retain substantially higher Top-1 and Top-5 accuracies, demonstrating that increased model capacity correlates with greater adversarial resilience.

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015.

[2] Phillip Lippe. Uva deep learning tutorials - adversarial attacks. https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/tutorial10/Adversarial_Attacks.html, 2022. Accessed May 2025.

[3] PyTorch Contributors. Torchvision models documentation. https://docs.pytorch.org/vision/main/models.html, 2025. Accessed May 2025.

[4] OpenAI. Chatgpt (mar 14 version). https://chat.openai.com, 2024.