Deep Learning Project 3: Jailbreaking-Deep-Models

Team SPAR: Aryan Mamidwar, Spandan Rout
Github Codebase

arm9337@nyu.edu, sr7729@nyu.edu

Abstract

This report summarizes an evaluation of adversarial robustness across four state-of-the-art image classifiers: DenseNet-121, EfficientNet-B0, MobileNet-V2, and ViT-L-32—on a 100-class subset of ImageNet. We crafted three adversarial attack methods (FGSM, PGD, and PatchPGD) by specifically building and fine-tuning them on a ResNet-34 backbone. After identifying optimal parameters through a targeted hyperparameter sweep for patch-based attacks, we tested transferability to the other models. The experiments reveal severe performance degradation under even single-step attacks and demonstrate that iterative and patch-based perturbations.

Introduction

Deep neural networks achieve high accuracy on image classification benchmarks but remain vulnerable to small, human-imperceptible perturbations. In this project, we investigate:

- One-step attack (FGSM) to get baseline vulnerability.
- Iterative PGD attack for strong adversarial examples.
- PatchPGD attack limiting perturbations to a small region (32 x 32).
- Transferring the attacks on different models.

Methodology

Data Preperation

A subset of 500 images from the 100 ImageNet classes was loaded with normalization (mean=[0.485,0.456,0.406], std=[0.229,0.224,0.225]).
The Adversarial datasets were then pre-saved as tensors: FGSM, PGD and PatchPGD variants.

Model Architecture

We used 5 pretrained model architecture: ResNet-34, DenseNet-121, EfficientNet-B0, MobileNet-V2, and Vision Transformer ViT-L-32, each set to evaluation mode on GPU.

Attack Algorithms

- FGSM: A single-step $L\infty$ perturbation with normalized epsilon per channel.
- PGD: Multi-step gradient descent with $L\infty$ constraint, ensuring normalized bounds.
- PatchPGD: Iterative PGD constrained to a 32×32 patch; initialization via adversarial patterns and per-channel normalization .

Hyperparameter (PGD)

- Defined range: iterations: [1, 5, 10, 20, 50].
- Tracked top-1 accuracy, logging the best under lowest top-1 as seen in the Fig1

Hyperparameter (PatchPGD)

- Defined ranges: ϵ: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0], α: [4, 8, 16, 20, 25], iterations: [40, 50, 80, 100, 120, 150, 170, 200]
- Sampled strategic combinations (63 total) on a 64-image subset to reduce compute time.
- Tracked top-1 and top-5 accuracy, logging the best under lowest top-1.

Optimal PatchPGD Parameters

The hyperparameter grid search identified $\epsilon = 0.08$, $\alpha = \epsilon/16 = 0.05$, and 200 iterations the most destructive patch attack. Applying these hyperparameter settings on the full 500-image yielded:

- Optimal PatchPGD Top-1: 6.4%
- Optimal PatchPGD Top-5: 29.6%

Lessons Learned

Hyperparameter tuning is essential: A thorough sweep over ϵ, α and iterations can uncover the most damaging attack settings, what's "best" for one model or datasets may not transfer to another as seen in the last part.
Patches don't save you: The need of taking into account both dense and localized perturbations is highlighted by the fact that even patch attacks that are limited to a small area can significantly reduce accuracy.
Visualization helps: Inspecting original vs. adversarial

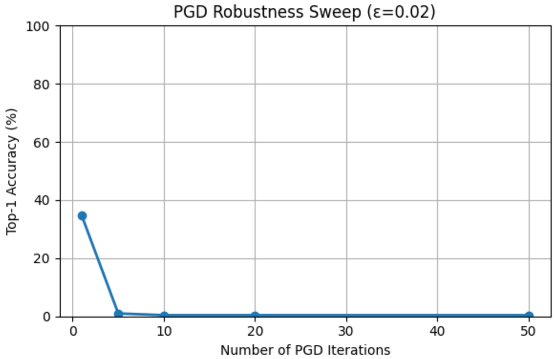

Figure 1: PGD robustness sweep (ϵ=0.02) for ResNet-34

examples reveals how subtle changes in pixel space translate to misclassifications, guiding more intuitive defense strategies.

Visualization

- FGSM with $\epsilon = 0.02$ introduces imperceptible noise that not noticeable to the human eye but fools ResNet-34 across a variety of images. All five perturbed samples are misclassified, underscoring FGSM's ability to induce errors with minimal distortion as seen in Fig2
- A 10-step PGD attack ($\epsilon = 0.02, \alpha = 0.005$) drives the Top-1 accuracy of ResNet-34 to zero. On the displayed samples (In Fig3), even though each perturbations remains nearly invisible. Accuracy collapses only after a handful iterations.
- A small 32 x 32 patch, optimized via PGD, is added onto each input and remains visually inconspicuous while causing ResNet-34 to misclassify all five examples (In Fig4).

Results & Discussion

On the clean test set, the ResNet-34 model achieves a Top-1 accuracy of 76.20% and Top-5 accuracy of 94.20%.
After FGSM perturbation (ϵ=0.02), the model accuracy reduces the Top-1 down to about 3.4% and Top-5 to 21.2%.
After 10 iterations of PGD attack the model accuracy reduces the Top-1 down to nearly 0.0% and Top-5 to 1.8%.
For patch PGD, even after hyperparameter tuning, with a best Top-1 of 6.4% and Top-5 of 29.6%.
Under the attacks FGSM, PGD and Patch based attack, all 4 architectures i.e. DenseNet-121, EfficientNet-B0, MobileNet-V2, and Vision Transformer ViT-L-32 suffered steep Top-1 and Top-5 accuracy drops. With MobileNet-V2 being most vulnerable in top-1 accuracies as seen in Fig5 and Fig6.
ViT_L_32 maintains the highest robustness across perturbation types, losing far fewer correct predictions than the other CNN counterparts.

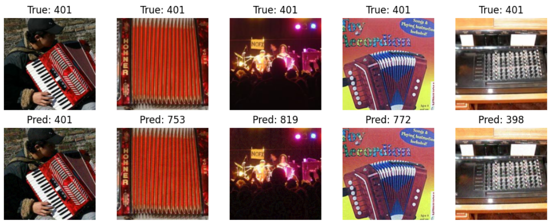

Figure 2: FGSM adversarial examples on ResNet-34

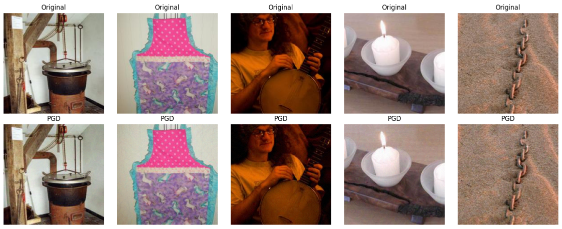

Figure 3: PGD adversarial examples on ResNet-34

References

[1] Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2018. Adversarial Patch. arXiv:1712.09665.

[2] Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. arXiv:1608.04644.

[3] Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. arXiv:2003.01690.

[4] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

[5] Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

[6] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

[7] Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083.

[8] Papernot, N.; McDaniel, P.; and Goodfellow, I. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. arXiv:1605.07277.

[9] Tan, M.; and Le, Q. V. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946.

[10] Yuan, X.; He, P.; Zhu, Q.; and Li, X. 2019. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9): 2805–2824.

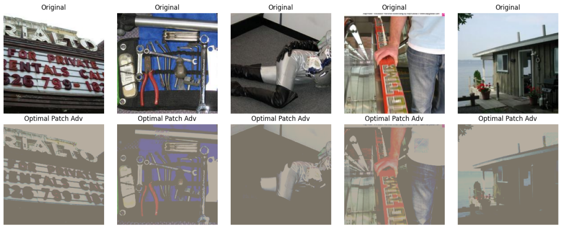

Figure 4: Patch-PGD adversarial attacks on ResNet-34

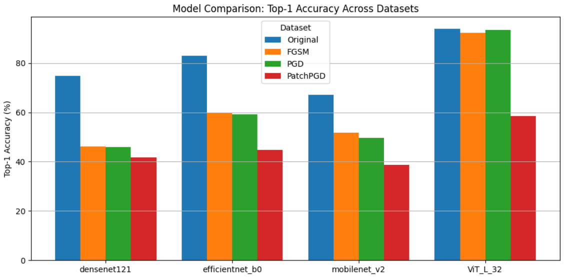

Figure 5: Top-1 accuracy comparison across models and datasets

Disclousre

We have taken inspiration from the assignments/labs of another course (Intro to ML @ NYU). We have also taken inspiration of the above cited papers, articles, code bases. We have also used some LLM models to understand more about the project (ChatGPT, Grok, Perplexity). We have also used some more online resources like StackOverflow, official documentation of the packages used.

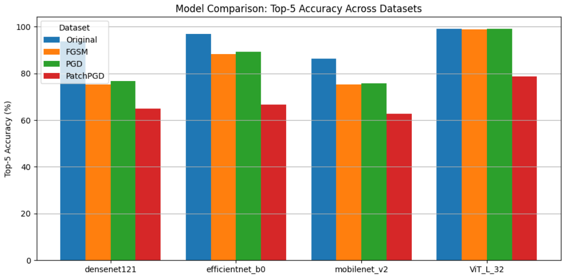

Figure 6: Top-5 accuracy comparison across models and datasets