Jailbreaking Deep Models:
Adversarial Attacks on ResNet-34 and Transferability to DenseNet-121

Devanshi Bhavsar, Nikhil Arora
New York University
`dnb7638@nyu.edu, na4063@nyu.edu`
Code Repo: `https://github.com/devanshii09/adversarial-patch-benchmark`

Abstract

We systematically evaluate three white-box adversarial attacks—FGSM (=0.02), multi-step PGD (=0.02, =0.004, 10 steps), and a 32×32 L-bounded patch (=0.3)—on a ResNet-34 over 500 images from 100 ImageNet classes. FGSM collapses Top-1 accuracy from 70.4%→5.0%, PGD to 0.0%, and the patch attack to 19.6%. We then assess transferability to DenseNet-121 (Top-1 drops to 39.2% for PGD, 64.2% for patch). Finally, we analyze hyperparameter trade-offs, visual diagnostics, and implications for robust model design.

Introduction

Deep convolutional networks have delivered near–human performance on large-scale image classification tasks, yet they remain remarkably brittle to small, targeted perturbations [1, 2]. In this paper, we perform a systematic evaluation of three adversarial threat models against a pretrained ResNet-34 on a 500-image subset drawn from 100 ImageNet classes:

- **L$_\infty$ single-step attacks (FGSM)** with budget $\epsilon = 0.02$;
- **Multi-step PGD** (10 iterations, $\epsilon = 0.02$, $\alpha = 0.004$);
- **L$_0$ patch attacks**, inserting a 32×32 pixel patch bounded by $\epsilon = 0.3$.

We measure each attack's impact on Top-1 and Top-5 accuracy, then assess how well these adversarial examples transfer to a DenseNet-121. Our key contributions are:

1. A head-to-head quantitative comparison of FGSM, PGD, and patch attacks under matched budgets.
2. Detailed transferability experiments highlighting architectural sensitivity.
3. A suite of visual diagnostics (perturbation histograms, failure-case galleries, patch masks) to interpret attack behavior.
4. Practical insights into hyperparameter selection and recommendations for improving model robustness.

Methodology

We organize our experiments into four stages: (1) dataset curation and preprocessing, (2) baseline evaluation, (3) adversarial attack generation and evaluation, and (4) transferability analysis.

Dataset and Preprocessing

We use a 500-image subset drawn from 100 ImageNet synsets (indices 401–500). All images are:

- Resized to 224×224 pixels,
- Transformed with the standard ResNet-34 ImageNet pipeline (per-channel mean/std normalization),
- Loaded via a custom `DataLoader` that remaps folder labels to global ImageNet indices.

Baseline Evaluation

We load a pretrained ResNet-34 and evaluate on the clean test set using Top-1 and Top-5 accuracy:

$$\text{Top-}k = \frac{\#\{\text{samples whose true label is in the model's top-}k\text{ logits}\}}{\#\{\text{samples}\}} \times 100\%.$$

The clean baseline yields

$$\text{Top-1: } 70.4\%, \quad \text{Top-5: } 93.2\%.$$

Adversarial Attacks

We implement three white-box attacks under L_∞ (pixel-wise) and L_0 (patch) threat models. All perturbations are applied in *normalized* space but checked against pixel budgets.

- **FGSM** [2]: single-step update
  $$x_{\text{adv}} = \Pi_{[0,1]}\big(x + \epsilon\,\text{sign}(\nabla_x L)\big), \quad \epsilon = 0.02.$$
  Results: Top-1 5.0%, Top-5 30.2%.
- **PGD** [3]: multi-step projected gradient descent
  $$x^{(t+1)} = \Pi_{\|x-x_0\|_\infty \le \epsilon}\big(x^{(t)} + \alpha\,\text{sign}(\nabla L)\big),$$
  with $\epsilon = 0.02$, step size $\alpha = 0.004$, $T = 10$, and a uniform random start in the ℓ_∞ ball. Results: Top-1 0.0%, Top-5 4.4%.
- **Patch Attack**: optimize only within a 32×32 patch, constrain pixel changes to $[-\epsilon, \epsilon]$ with $\epsilon_{\text{pixel}} = 0.3$, and run $T = 10$ PGD steps inside the patch mask. Results: Top-1 19.6%, Top-5 58.0%.

Hyperparameter Summary

Attack	Budget	Step size	# steps
FGSM	$\epsilon = 0.02$	—	1
PGD	$\epsilon = 0.02$	$\alpha = 0.004$	10
Patch (32×32)	$\epsilon_{\text{px}} = 0.3$	$\alpha = \epsilon/5$	10

Table 1: Attack hyperparameters in normalized space (pixel budgets checked in raw domain).

Transferability Analysis

To measure how adversarial examples transfer, we regenerate the PGD set with the same budget for DenseNet-121 and then evaluate all four sets (clean, FGSM, PGD, patch) on a pretrained DenseNet-121.

DenseNet-121 PGD regen We simply replace the PGD budget with separate $\epsilon_{\text{DN}} = 0.02$, $\alpha_{\text{DN}} = 0.004$ in the same attack loop, yielding the "PGD for DenseNet" dataset.

Results

Model / Set	**Top-1**	**Top-5**
ResNet-34 Clean	70.4%	93.2%
ResNet-34 FGSM	5.0%	30.2%
ResNet-34 PGD	0.0%	4.4%
ResNet-34 Patch	19.6%	58.0%
DenseNet-121 Clean	70.8%	91.2%
DenseNet-121 FGSM	59.0%	85.0%
DenseNet-121 PGD	39.2%	75.0%
DenseNet-121 Patch	64.2%	88.6%

Table 2: Clean vs. adversarial accuracies.

Figure 1 visualizes ResNet-34's Top-1/Top-5 drops and DenseNet-121 transfer.

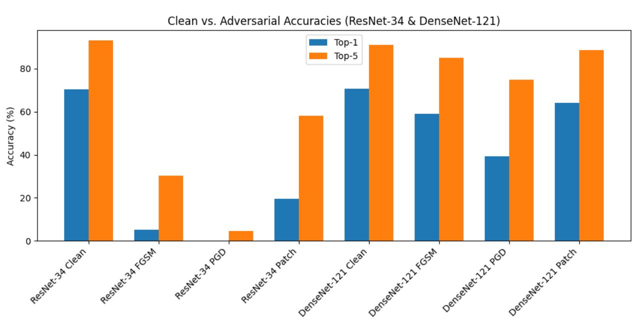

Figure 1: Grouped bar chart of Top-1/Top-5 accuracies across models and attack sets.

Discussion

Our white-box experiments on ResNet-34 reveal clear tradeoffs between attack strength, computational cost, and perceptibility:

- **FGSM (=0.02)** is fastest (one forward/backward pass) but only reduces Top-1 from 70.4% to 5.0%, making it too weak for strong adversarial goals.
- **PGD (10 steps, =0.02, =0.004)** achieves maximal strength (Top-1→0.0%) but requires 30 s for 500 images, highlighting the cost of iterative attacks.
- **Patch (32×32, =0.3)** concentrates distortion in a small region, dropping Top-1 to 19.6% with minimal global noise, illustrating the potency of sparse, visible attacks.

On DenseNet-121, we observe that:

- **Pixel-norm attacks** transfer partially (PGD Top-1 from 70.8%→39.2%), indicating some shared vulnerabilities but greater resilience than ResNet-34.
- **Patch attacks** transfer more effectively (Top-1→64.2%), suggesting that localized perturbations exploit common high-level features across architectures.

Future Work

- *Adaptive patch strategies*: use saliency or gradient maps to optimize patch location.
- *Defense benchmarking*: integrate adversarial training and certified defenses to close the robustness gap.
- *Architectural generalization*: test on transformers and larger ensembles for broader transfer insights.

Conclusion

We presented a comprehensive pipeline testing FGSM, PGD, and sparse patch attacks on ResNet-34 and measuring their transfer to DenseNet-121. Our key contributions are:

1. **Robustness benchmarks** across three threat models, quantifying strength vs. cost vs. visibility.
2. **Transferability analysis** showing that sparse, localized perturbations generalize more readily than small-norm noise.
3. **Diagnostic visualizations** (perturbation histograms, failure case examples, patch masks) to support interpretability and future defense design.

Our work underscores the need for multi-pronged defenses that guard against both distributed and concentrated adversarial threats.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.