Jailbreaking Deep Models: Adversarial Attack Implementation and Analysis

Yifei Xu*

Tandon School of Engineering, New York University
5 Metrotech Center,
Brooklyn, New York 11201 USA
yx3882@nyu.edu

Abstract

Deep neural networks have achieved remarkable success in image classification tasks but remain surprisingly susceptible to adversarial examples. In this report, we explore adversarial vulnerabilities of a pre-trained ResNet-34 model on a subset of the ImageNet-1K dataset, focusing on four main attack scenarios: single-step pixel-wise (FGSM), multi-step iterative (PGD), patch-based attacks, and transferability to other architectures. Our FGSM experiments highlight that even a small ℓ_∞ budget can reduce Top-1 accuracy significantly. PGD attacks exacerbate this vulnerability, driving accuracy to near 0% with multiple gradient steps. Although patch attacks constrain perturbations to a small image region, they can still severely compromise classification accuracy when given higher pixel budgets. Finally, we evaluate transferability on DenseNet-121, finding partial cross-model effectiveness: adversarial inputs crafted against ResNet-34 maintain noticeable performance drops on DenseNet-121, albeit less dramatic. Our results highlight the persistent brittleness of production-grade networks and underscore the importance of robust defenses, especially as deep learning models are increasingly deployed in safety-critical applications.

Introduction

Deep neural networks have achieved unprecedented success in image classification. However, even as deep networks excel in accuracy, researchers discovered a glaring vulnerability: *adversarial examples*. Szegedy et al. first revealed that state-of-the-art classifiers can be "fooled" by inputs that differ only slightly from correctly classified examples (Szegedy et al. 2013). In particular, by adding carefully crafted small perturbations to an image – often so subtle that the changes are imperceptible to human observers – one can cause the model's predictions to change dramatically, leading to severe misclassifications. Goodfellow et al. attributed this brittleness in part to the high-dimensional linear behavior of these models and introduced the Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2015), a one-step gradient-based attack that efficiently generates such adversarial perturbations. Since then, numerous attack strategies (e.g., iterative multi-step methods) have been proposed, and studies have found that the adversarial perturbation often transfers across different network architectures, highlighting a fundamental blind spot in current deep learning models.

These adversarial vulnerabilities pose significant concerns for real-world deployment of deep models. In safety-critical applications, a maliciously crafted input could lead to catastrophic outcomes. For example, a perturbed traffic sign image (with only subtle sticker modifications) might be misidentified as a different sign, potentially causing an autonomous vehicle to misbehave (Eykholt et al. 2018). The existence of such easily "jailbroken" models underscores the security and reliability risks of deep neural networks. Recent research has therefore focused on *adversarial robustness*: developing defense techniques and robust training frameworks to harden models against attacks. While approaches like adversarial training (e.g., using multi-step projected gradient attacks during training) can significantly improve a model's resilience (Madry et al. 2018), achieving reliable robustness without sacrificing accuracy remains an open challenge.

In this context, our report examines the brittleness of a production-grade image classifier through the implementation and analysis of multiple adversarial attacks. We target a pre-trained ResNet-34 model (ImageNet-1K) and evaluate how various attack methods can "jailbreak" its performance. By analyzing these attack scenarios on an industry-grade model, this report provides a comprehensive study of deep network vulnerability.

Methodology

In this section, we detail our methodology for evaluating adversarial attacks against deep image classification models. We will provide an overview of the dataset, the models we considered, and the different attack strategies implemented.

Dataset

We use a subset of the ImageNet-1K dataset consisting of 500 images selected from 100 classes. This subset allows a more focused examination of adversarial vulnerability while maintaining enough class diversity to mimic the complexity of ImageNet (Deng et al. 2009).

*The code and notebook are available at Github repository: https://github.com/yifeihsu/DL_Project_3.

Models

We evaluate attacks on the ResNet-34 architecture (He et al. 2016), which is pre-trained on the ImageNet-1K dataset. This model represents a standard production-grade network with robust performance on ImageNet and related vision tasks. To study transferability of adversarial examples, we also deploy a DenseNet-121 model (Huang et al. 2017) on the same dataset without any retraining.

Attacks

Our goal is to generate adversarial examples for the image classifier such that the model's predictions are incorrect while the perturbation remains visually subtle. Below, we outline each attack variant.

Fast Gradient Sign Method (FGSM) We first implement a single-step ℓ_∞ attack known as FGSM (Goodfellow, Shlens, and Szegedy 2015). Given an original image $\mathbf{x}$ and the model's loss $\mathcal{L}$, the adversarial example is produced via

$$\mathbf{x}_{\text{adv}} \leftarrow \mathbf{x} + \epsilon \operatorname{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}\right),$$

where ϵ is the maximum perturbation allowed in each pixel. We set $\epsilon = 0.02$ (in normalized pixel space), corresponding to ± 1 on a 0–255 scale. Smaller ϵ values preserve image quality but may yield weaker attacks, while larger values ensure bigger accuracy drops but risk perceptible artifacts.

Projected Gradient Descent (PGD) We then extend the single-step FGSM to a multi-step approach using Projected Gradient Descent (Madry et al. 2018), iterating the gradient update multiple times:

$$\mathbf{x}_{\text{adv}}^{(t+1)} \leftarrow \Pi_{\|\mathbf{x}-\mathbf{x}_{\text{adv}}^{(t)}\| \le \epsilon}\big(\mathbf{x}_{\text{adv}}^{(t)} + \alpha \operatorname{sign}(\nabla_{\mathbf{x}_{\text{adv}}} \mathcal{L})\big),$$

where $\Pi(\cdot)$ projects the perturbation onto the ℓ_∞-ball of radius ϵ around the original image, and α is the step size.

We explored the number of iterations (*e.g.*, 20–70) and step sizes ($\alpha = 0.005$–0.02). Fewer iterations reduce computation but sometimes yield less potent attacks, while more iterations may overshoot or cause redundancy if α is not tuned properly. Introducing momentum $\beta \approx 0.9$ typically improves convergence, but we noticed that excessive momentum can lead to over-concentrated perturbations in certain image regions.

From our trials, PGD consistently outperformed FGSM in terms of reducing model accuracy, confirming prior findings that iterative attacks better exploit the local loss surface. Nonetheless, PGD is computationally heavier. In scenarios with limited GPU time, we found a trade-off between the iteration count and final attack strength: about 40–70 steps gave us a good accuracy drop without excessively long training loops.

Patch Attacks To restrict perturbations to a small image region, we implement patch attacks (Brown et al. 2017). Instead of perturbing all pixels, we randomly select a 32×32 patch within the image and allow gradient-based perturbations only in that region. Since fewer pixels can be modified, we increase ϵ (e.g., up to 0.3–1.5 in normalized space) for a more visible effect. We also experiment with targeted objectives by choosing a class label that is different from the ground truth, thereby guiding the classifier to misclassify the image into a specific target class.

The key parameters for the patch attacks are the patch size, ℓ_∞ budget, and iteration count. A larger patch can more easily disrupt predictions, but also becomes more visually obvious. We also experimented with targeted variations, selecting a target class least likely given the original image. Though targeted approaches sometimes required more steps (e.g., 200–500) to converge, they could force extremely counterintuitive misclassifications.

Results and Discussion

In this section, we report quantitative findings for each attack and provide qualitative insights on model robustness and adversarial transferability. We create three adversarial test sets by applying each attack to all 500 images. For each set, we compute:

- Top-1 Accuracy (%): the fraction of images for which the model's top predicted class matches the true label.
- Top-5 Accuracy (%): the fraction of images for which the model's top-5 predictions include the true label.

We also verify that the ℓ_∞ distance between the perturbed and original images does not exceed the specified ϵ. Finally, to assess transferability, we evaluate the same adversarial images on a DenseNet-121 model and measure how much performance degrades compared to a clean baseline.

Baseline Performance

We first evaluated the unperturbed test set on both ResNet-34 and DenseNet-121.

- **ResNet-34**: Achieved a Top-1 accuracy of **76.00%** and a Top-5 accuracy of **94.20%**.
- **DenseNet-121**: Reached a Top-1 accuracy of **74.60%** and a Top-5 accuracy of **93.60%**.

These values set the baseline for subsequent adversarial comparisons.

FGSM Attack

FGSM applies a one-step gradient update with $\epsilon = 0.02$ in the normalized domain. Key observations:

- **ResNet-34 Accuracy:** Dropped to **6.00%** (Top-1) and **35.60%** (Top-5), indicating an over 50% reduction in Top-1 accuracy from the baseline.
- **Perturbation Size:** The maximum ℓ_∞ distortion in normalized space was 0.02, with an approximate raw pixel change of 1.17 (on a 0–255 scale). Despite the limited budget, the attack was highly effective.

Overall, FGSM proved surprisingly destructive for a single-step method, confirming that even small, carefully directed gradients can significantly alter classifier predictions.

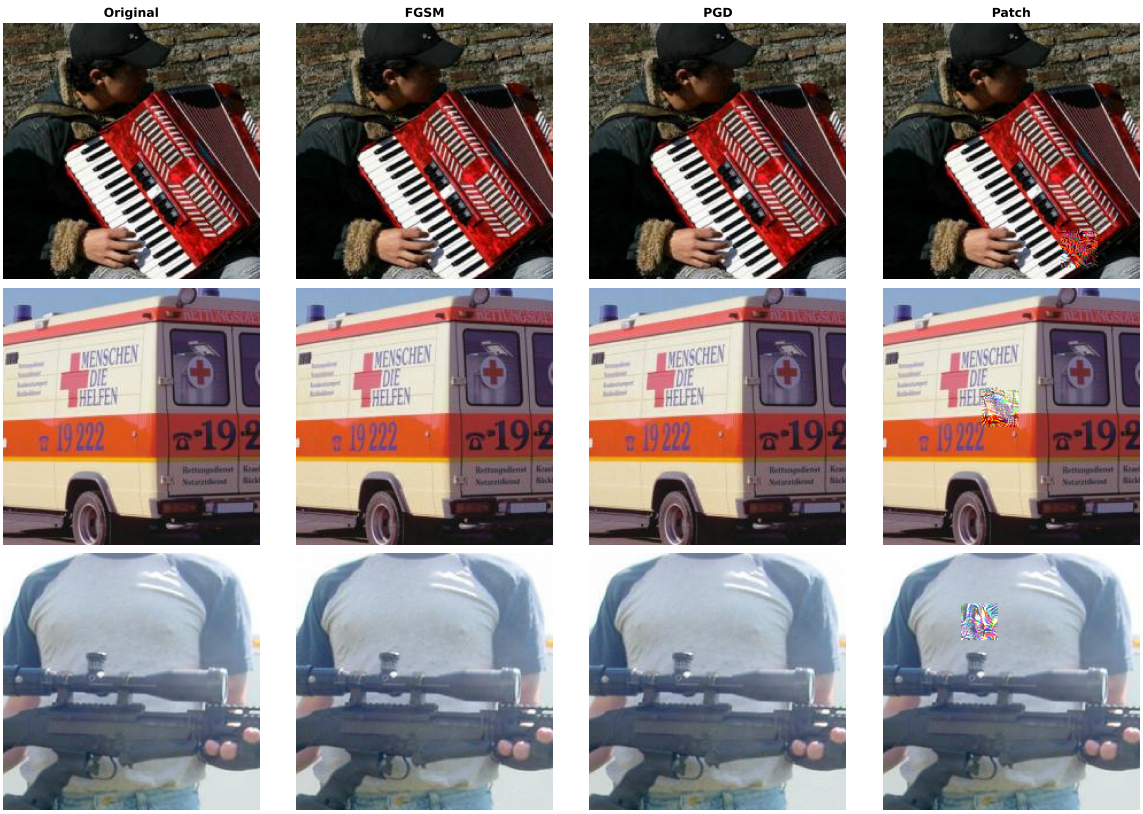

Figure 1: Adversarial examples generated by FGSM, PGD, and Patch attacks. (a) Original image, (b) FGSM perturbation, (c) PGD perturbation, (d) Patch perturbation. The original image is misclassified as a different class in all cases.

PGD Attack

We next evaluated a PGD attack with momentum, using $\epsilon = 0.02$, $\alpha = 0.008$, and $\text{iterations} = 70$. Because of the iterative nature, generating 500 adversarial images took longer time on our setup. The final adversarial set delivered:

- **ResNet-34 Accuracy: 0.00%** (Top-1) and **4.20%** (Top-5). This is a complete collapse of Top-1 performance and underscores PGD's effectiveness as a "strong" first-order adversary (Madry et al. 2018).
- **Perturbation Size:** Confirmed to stay within $\ell_\infty = 0.02$ normalized (1.17 raw pixel change).

Relative Top-1 accuracy dropped by 76% from the baseline. Given that FGSM already achieved a large accuracy drop, PGD's multi-step refinement was able to push it to near-complete misclassification.

Patch Attack

For patch-based attacks, we constrained the perturbation to a 32×32 subregion. Due to the fewer available pixels, we increased the normalized ϵ to 1.5, ran 500 iterations with $\alpha = 0.2$, and allowed extended GPU time to process 500 samples. The final outcomes:

- **ResNet-34 Accuracy: 23.80%** (Top-1) and **56.60%** (Top-5), yielding a 52% absolute Top-1 drop compared to the baseline.
- **Perturbation Size:** The maximum raw pixel change observed for the patch region was $\mathbf{87.59}$. Since we allowed a large ϵ, the patch is often visually more noticeable yet remains confined to a relatively small region.

Despite the smaller spatial extent, a well-placed patch substantially disrupted predictions, illustrating that highly localized modifications can still break a classifier if the perturbation magnitude is large.

Transferability

We evaluated each adversarial set on an unmodified DenseNet-121 model to assess transferability. Table 2 summarize these cross-model tests:

- **DenseNet-121 Baseline:** 74.60% (Top-1) and 93.60% (Top-5) on the clean set.
- **FGSM Set:** Accuracy dropped to 63.60% (Top-1), suggesting the FGSM examples, while tuned for ResNet-34, retain moderate efficacy on DenseNet-121.
- **PGD Set:** Accuracy dropped to 64.80% (Top-1), similar to FGSM in cross-model potency.
- **Patch Set:** Accuracy dropped to 69.80% (Top-1). This smaller drop suggests that localized, high-ϵ patches exploit more model-specific receptive field patterns in ResNet-34 and thus do not transfer as strongly.

Table 1: Summary of Key Results for ResNet-34.

Configuration	Top-1 (%)	Top-5 (%)
Baseline (Clean)	76.0	94.2
FGSM ($\epsilon = 0.02$)	6.0	35.6
PGD ($\epsilon = 0.02$)	0.0	4.2
Patch ($\epsilon = 1.5$)	23.8	56.6

Table 2: Transferability of ResNet-34 Attacks to DenseNet-121 (Top-1).

Attack (on ResNet-34)	DenseNet-121 Acc. (%)
Baseline (Clean)	74.6
FGSM	63.6
PGD	64.8
Patch	69.8

Interpretation. The relatively modest drops in DenseNet-121's accuracy (compared to ResNet-34's near-complete failures) highlight that transferability depends on architectural similarities. Strong multi-step attacks (like PGD) do not always guarantee superior transfer performance, suggesting that certain perturbations overfit to the victim model's decision boundary.

From a defender's perspective, these findings emphasize the need for adversarial robustness strategies (e.g., adversarial training, input preprocessing) and highlight that attacks can exploit even small perturbation budgets or confined image regions.

Conclusion

In this report, we investigated the adversarial vulnerabilities of deep image classifiers by launching multiple attacks against a pre-trained ResNet-34 model. Our experiments demonstrated that even modest ℓ_∞ perturbations can drastically reduce model accuracy (often to near zero), underscoring how brittle these networks can be. Iterative attacks like PGD proved more potent than single-step FGSM. Patch attacks, while spatially localized, still induced significant misclassifications at higher per-pixel budgets. Transferability tests on DenseNet-121 revealed partial cross-model effectiveness, highlighting the need for defenses that generalize beyond single architectures. Overall, our findings reinforce the importance of designing more robust deep networks, as real-world applications demand resilience against even subtle adversarial threats.

Acknowledgments

We would like to acknowledge the valuable help of Large Language Models (LLMs), such as OpenAI ChatGPT and Google Gemini, for enhancing the code readability, suggesting improvements for our reports, and assisting with experiment design.

This work would not have been possible without the collaborative effort and support from both computational and AI-based resources, to whom we extend our deepest gratitude.

References

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Eykholt, K.; Evtimov, I.; Fernandess, E.; Li, B.; Rahmati, A. R.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1625–1634.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.

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*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.