iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.72811

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Adversarial Attacks on Neural Networks: Approach, Findings, and Analysis

Amaan Mithani[1], Amaan Ansari[2], Kunaal Vadgama[3], Gaurav Ghop[4]
Department of Electrical and Computer Engineering, New York University

Abstract: ***This report explores adversarial attacks targeting deep image classification models, specifically ResNet-34 and DenseNet-121 trained on the ImageNet-1K dataset. We implement and assess the effectiveness of Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and localized patch-based attacks, all constrained by strict and spatial limitations. Our results reveal significant declines in both top-1 and top-5 accuracy, and emphasize the cross-architecture transferability of adversarial inputs. We detail our approach, share findings and insights.***
Keywords: Adversarial Attacks, Neural Networks, Image Classification, Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD).

I. INTRODUCTION

While deep neural networks have shown impressive performance in image classification tasks, they are still susceptible to adversarial attacks; subtle, intentional modifications to input images that can lead to incorrect predictions. Here, we target a ResNet-34 model trained on ImageNet-1K by generating adversarial examples within and localized patch constraints. We also examine how well these attacks transfer to a DenseNet-121 model. The objective is to reduce model accuracy without introducing noticeable changes to the images.

II. METHODOLOGY

1) Dataset and Preprocessing: We work with a curated subset of ImageNet-1K, consisting of 500 test images spanning 100 distinct classes. The images are normalized using the standard ImageNet mean and standard deviation values and loaded via PyTorch's ImageFolder utility. Class names are assigned based on the mappings provided in the labels_list.json file.

2) Baseline Evaluation: We assess the pretrained ResNet-34 model on the test set, recording top-1 and top-5 accuracy as baseline performance metrics. To study attack transferability, we also evaluate the DenseNet-121 model.

3) FGSM (L∞ attack): We implement the Fast Gradient Sign Method, where each pixel is perturbed by a maximum of $\varepsilon = 0.02$ in normalized space. The adversarial image x' is generated as:

$$x' = x + \varepsilon \cdot sign(\nabla_x \mathscr{L})$$

where $\mathscr{L}$ is the cross-entropy loss with respect to the input x.

4) PGD (L∞ multi-step): Projected Gradient Descent extends FGSM by applying iterative updates with a step size $\alpha = 0.01$ for 20 steps. After each update, the perturbed image is projected back onto the ε-ball to ensure the perturbation stays within bounds.

5) Patch Attack: In this method, we modify only a randomly selected 32 × 32 patch within each image. A larger perturbation budget of $\varepsilon = 0.5$ is used, with noise sampled randomly within the patch region.

6) Evaluation Metrics: For every attack method, we store the generated adversarial examples and verify that the L∞ constraint is satisfied. We report both top-1 and top-5 classification accuracy for ResNet-34 and DenseNet-121. Additionally, we provide visual comparisons between original and adversarial predictions for selected representative samples to highlight the impact of the perturbations.

III. RESULTS

We begin by assessing the clean (unperturbed) performance of both models:

1) ResNet-34: Top-1 Accuracy: 76.00%, Top-5 Accuracy: 94.20%

2) DenseNet-121: Top-1 Accuracy: 74.80%, Top-5 Accuracy: 93.60%

A. *FGSM Attack ($\varepsilon = 0.02$)*

- ResNet-34: Top-1 Accuracy: 2.00%, Top-5 Accuracy: 4.80%
- DenseNet-121: Top-1 Accuracy: 3.40%, Top-5 Accuracy: 6.00%

The FGSM attack leads to a drastic drop in accuracy, reducing performance by over 70 percentage points. Despite this, the perturbed images remain visually similar to the original inputs.

B. *PGD Attack ($\varepsilon = 0.02$, 20 steps)*

- ResNet-34: Top-1 Accuracy: 1.80%, Top-5 Accuracy: 3.40%
- DenseNet-121: Top-1 Accuracy: 3.00%, Top-5 Accuracy: 4.80%

PGD performs even more aggressively than FGSM, demonstrating the heightened effectiveness of iterative adversarial attacks.

C. *Patch Attack (32×32 patch, $\varepsilon = 0.5$)*

- ResNet-34: Top-1 Accuracy: 3.80%, Top-5 Accuracy: 5.40%
- DenseNet-121: Top-1 Accuracy: 3.80%, Top-5 Accuracy: 6.40%

Even when restricted to a small, localized region, the patch-based attack results in significant accuracy degradation—especially with the larger perturbation budget.

D. *Transferability*

Adversarial examples crafted for ResNet-34 also generalize to DenseNet-121, causing substantial reductions in accuracy for both models. This cross-model vulnerability is consistent across all attack types, as illustrated in Table 1.

Attack	ResNet-34		DenseNet-121	
	Top-1	Top-5	Top-1	Top-5
Original	76.0	94.2	74.8	93.6
FGSM	2.0	4.8	3.4	6.0
PGD	1.8	3.4	3.0	4.8
Patch	3.8	5.4	3.8	6.4

Table 1: Top-1 and Top-5 accuracy (%) for each attack and model.

IV. DISCUSSION: LESSONS LEARNED AND MITIGATION

Our experiments revealed several key insights:

1) Even basic attack strategies like FGSM can lead to a severe drop in model accuracy.
2) While multi-step methods such as PGD are generally more effective, their improvement over FGSM is relatively limited for the chosen ε.
3) Patch-based attacks prove surprisingly powerful, especially when a higher perturbation budget is allowed.
4) Adversarial examples exhibit strong transferability across model architectures, highlighting a significant challenge for real-world deployment.

A. *Mitigation Strategies*

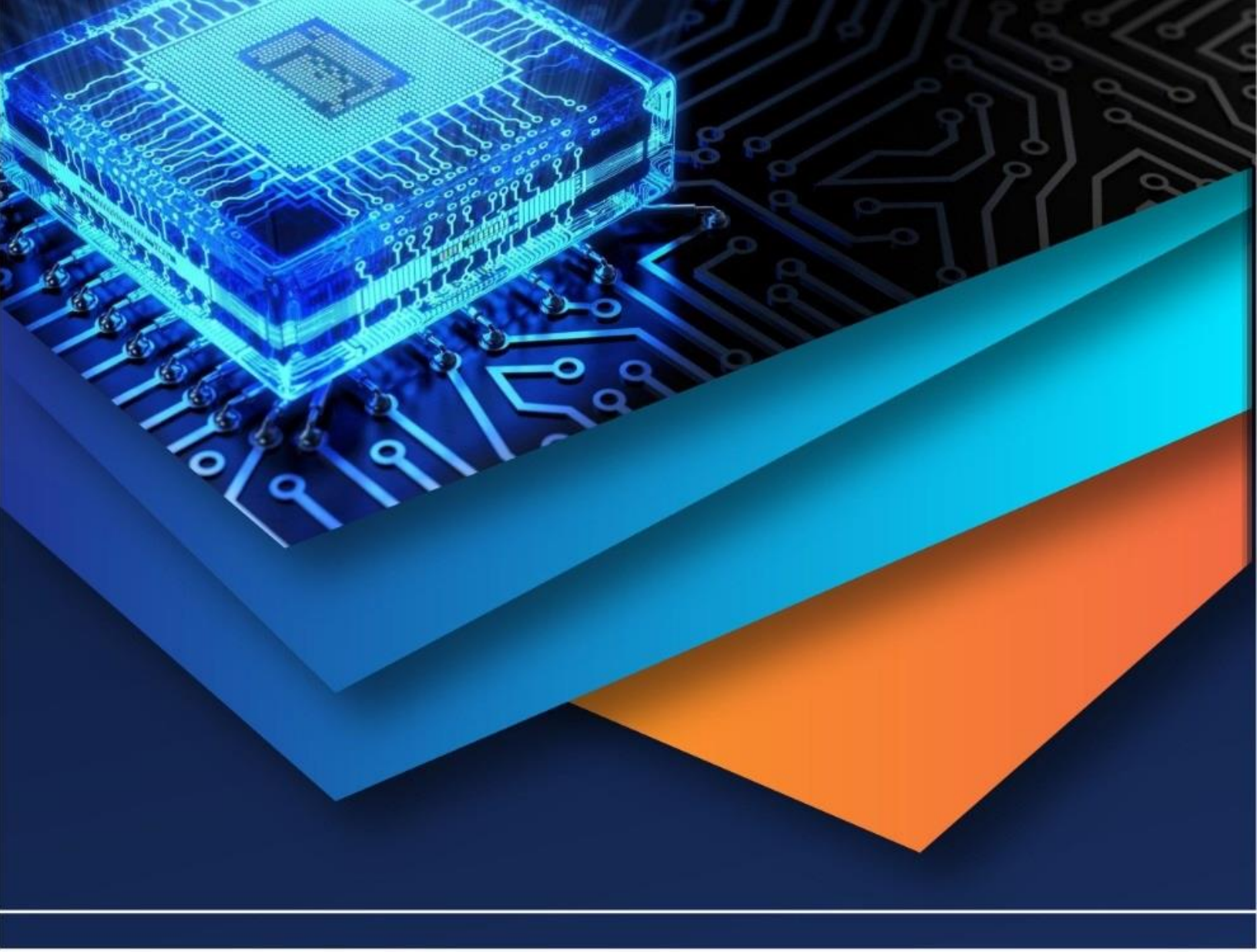
Potential defenses include adversarial training, input transformation techniques, and designing inherently robust architectures. However, these methods often fall short when facing adaptive adversaries, indicating that truly robust defenses remain an open research problem.

V. CONCLUSION

The vulnerability of deep image classification models to adversarial attacks have been systematically evaluated. Using ResNet-34 and DenseNet-121 trained on a subset of ImageNet-1K, we implemented and analyzed FGSM, PGD, and localized patch attacks under strict perturbation constraints. Our results demonstrated that even small perturbations can significantly degrade model performance, with adversarial examples transferring effectively across architectures. While methods such as adversarial training offer partial defences, the persistence of these vulnerabilities highlights the need for continued research into more robust and resilient learning systems.

REFERENCES

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. arXiv preprint arXiv:1412.6572, 2014.
[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. International Conference on Learning Representations (ICLR), 2018
[3] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. arXiv preprint arXiv:1605.07277, 2016
[4] Tom B. Brown, D. Mane, A. Roy, M. Abadi, and J. Gilmer. Adversarial Patch. arXiv preprint arXiv:1712.09665, 2017.
[5] Naveed Akhtar and Ajmal Mian. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access, 6:14410–14430, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY