# A Review of Adversarial Attacks on Image Classification Models

Noam Yakar

February 2025

## Abstract

Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated state-of-the-art performance in image classification tasks. However, these models are highly vulnerable to adversarial attacks—small, imperceptible perturbations crafted to mislead the model's predictions. This paper provides an overview of common adversarial attack techniques, their mathematical foundations, and their implications for the security of deep learning systems. A detailed examination of perturbation methods, including their theoretical underpinnings and empirical performance, is presented. Finally, we discuss defense strategies aimed at improving model robustness.

## 1 Introduction

Convolutional neural networks (CNNs) have achieved significant success in computer vision applications, including object recognition, medical imaging, and autonomous driving. Despite their effectiveness, these models remain susceptible to adversarial perturbations [**?**]. An adversarial perturbation is a carefully crafted noise vector $\delta$ added to an input image $x$, generating a perturbed image $x^* = x + \delta$, which remains indistinguishable to the human eye but results in misclassification by the model:

$$\arg\max f(x^*) \neq y, \quad \text{subject to } ||x - x^*||_p \leq \epsilon. \quad (1)$$

Here, $\epsilon$ is a constraint ensuring the perturbation remains imperceptible within a predefined norm space.

## 2 InceptionV3 Model

The InceptionV3 model was chosen for this study due to its widespread use in image classification tasks and its susceptibility to adversarial attacks. This architecture, known for its efficiency and accuracy, employs multiple convolutional layers with varying kernel sizes to capture multi-scale features. However, this complex structure introduces vulnerabilities, as small perturbations can exploit feature extraction mechanisms to induce incorrect classifications. The high non-linearity and reliance on learned features make InceptionV3 more sensitive to subtle changes in input data, increasing its susceptibility to adversarial perturbations.

## 3 Perturbation Methods

### 3.1 Fast Gradient Sign Method

The FGSM is a single-step gradient-based attack designed to maximize the model's loss by perturbing the input image in the direction of the gradient of the loss function. Given an image $x$ and its true label $y$, the adversarial example $x^*$ is computed as:

$$x^* = x + \epsilon \cdot \text{sign}(\nabla_x J(f(x), y)). \quad (2)$$

The perturbation magnitude $\epsilon$ is carefully chosen to balance imperceptibility and attack effectiveness. FGSM is computationally efficient but often produces perturbations that are easily detected by robust models.

### 3.2 Projected Gradient Descent

PGD refines FGSM by applying multiple iterative steps of small perturbations while ensuring that the resulting adversarial example remains within a bounded perturbation space. Each iteration follows:

$$x_{t+1} = \mathrm{Proj}_{x+\mathcal{S}}(x_t + \alpha \cdot \mathrm{sign}(\nabla_x J(f(x_t), y))), \quad (3)$$

where $\alpha$ is the step size, and $\mathrm{Proj}_{x+\mathcal{S}}$ projects the perturbed image back onto the $\epsilon$-ball around the original image. PGD is considered one of the strongest attacks, as it iteratively refines adversarial perturbations to mislead models effectively.

### 3.3 Momentum Iterative Method

MIM improves PGD by incorporating a momentum term, which stabilizes gradient updates and enhances attack transferability. The accumulated gradient is updated iteratively as follows:

$$g_{t+1} = \mu g_t + \frac{\nabla_x J(f(x_t), y)}{||\nabla_x J(f(x_t), y)||_1}, \quad (4)$$

where $\mu$ represents the momentum factor. The adversarial example is then refined using:

$$x_{t+1} = x_t + \alpha \cdot \mathrm{sign}(g_{t+1}). \quad (5)$$

This approach mitigates oscillatory updates and improves the effectiveness of the attack across different models.

### 3.4 Carlini & Wagner (C&W) Attack

The C&W attack formulates adversarial perturbation as an optimization problem that minimizes a combination of perturbation magnitude and classification loss:

$$\min ||\delta||_p + c \cdot J(f(x + \delta), y^*), \quad (6)$$

where $c$ is a trade-off parameter. Unlike FGSM and PGD, which rely on simple gradient sign updates, the C&W attack uses optimization techniques such as Adam or L-BFGS to iteratively refine perturbations, making it highly effective against various defenses.

## 4 Results of Perturbations

To evaluate the effectiveness of these perturbation methods, experiments were conducted on a pretrained InceptionV3 model. The classification confidence and loss function values were tracked over multiple iterations. The adverse images generated with FGSM, PGD, and MIM were visually indistinguishable from the original image, but caused significant misclassification.

Figure 1 presents the results of our experiments, displaying the perturbed images alongside their predicted labels and classification confidence scores. The original image, correctly classified as a "tabby" with 78.25% confidence, remains unchanged under FGSM, but PGD and MIM successfully misclassify it as a "toaster" with 100.00% confidence. The corresponding loss progression plot illustrates the rapid convergence of PGD and MIM, highlighting their robustness in generating adversarial examples.

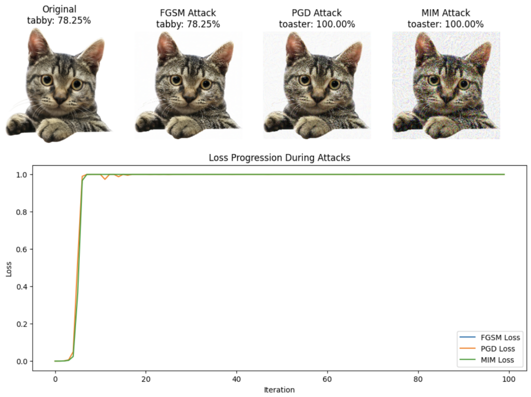

Figure 1: Comparison of adversarial attack effectiveness across FGSM, PGD, and MIM, with loss progression over iterations.

## 5 Implications of Adversarial Attacks

The vulnerabilities demonstrated by these adversarial attacks highlight significant security concerns in deep learning applications. In real-world scenarios, adversarial perturbations could be leveraged to mislead autonomous vehicles, bypass facial recognition systems, or generate misleading medical diagnoses. The ability to manipulate a model's decision-making process with minimal perturbation emphasizes the need for robust adversarial defenses.

The implications extend beyond classification errors, affecting ethical considerations and trust in AI-driven decision making. Attackers could exploit adversarial examples to deceive financial fraud detection systems or manipulate automated surveillance. Addressing these threats requires ongoing research on adversarial training, anomaly detection, and certified defenses that provide verifiable robustness guarantees.

# References

1. Goodfellow et al., "Explaining and Harnessing Adversarial Examples," 2014.

2. Madry et al., "Towards Deep Learning Models Resistant to Adversarial Attacks," 2017.

3. Kurakin et al., "Adversarial Machine Learning at Scale," 2016.