Understanding Adversarial Vulnerabilities: FGSM Attacks on Deep Neural Networks

Tiwari, Rohan Sanjay   
*University of Texas at Arlington*  
*Department of Bioengineering/Department of Computer Science*Arlington, United States  
rxt6430@mavs.uta.edu

https://github.com/rohantiwari02/Adv\_pytorch.git

Abstract— This paper investigates adversarial robustness of a convolutional neural network trained on the MNIST dataset using the Fast Gradient Sign Method (FGSM). We perform both white-box and black-box attacks to evaluate model susceptibility. The results show significant performance degradation under both conditions, with white-box attacks being more effective due to direct access to model gradients.

Keywords— Adversarial attacks, FGSM, white-box attack, black-box attack, CNN, MNIST, model robustness

# Introduction

Adversarial machine learning is a growing field that studies how small, intentional modifications to input data can mislead machine learning models into making incorrect predictions. These modifications, called adversarial examples, are typically imperceptible to humans but can significantly degrade a model's performance. This vulnerability is especially concerning in safety-critical applications such as facial recognition, medical diagnosis, and autonomous driving.

The goal of this project is to evaluate the susceptibility of convolutional neural networks (CNNs) to adversarial examples, focusing on the Fast Gradient Sign Method (FGSM)—a popular attack due to its computational efficiency and simplicity. We use FGSM to test a CNN trained on the MNIST dataset under two threat models: white-box and black-box attacks.

In a white-box attack, the adversary has complete access to the target model's internal architecture, parameters, and gradients. This allows the attacker to craft highly targeted adversarial inputs by leveraging gradient information directly from the model.

In a black-box attack, the attacker does not have access to the target model's internals. Instead, adversarial examples are generated using a surrogate model—a separate model trained on the same data distribution. The assumption is that adversarial perturbations are transferable, meaning that an input designed to fool the surrogate model will likely fool the target model as well.

# related work

The study of adversarial examples gained significant attention after the work of Goodfellow et al. [6], who introduced the Fast Gradient Sign Method (FGSM). Their research showed that small, deliberately crafted perturbations could drastically reduce the performance of neural networks. FGSM's simplicity and efficiency made it a foundational technique for adversarial attack research.

Subsequent work by Kurakin et al. [8] extended this to the physical world, demonstrating that printed adversarial images could still fool classifiers after being captured by a camera. This emphasized the real-world risks of adversarial attacks, beyond synthetic testing environments.

Papernot et al. [10] contributed significantly to the understanding of black-box attacks, proposing the concept of transferability—that adversarial examples generated on a surrogate model could successfully fool a different, unseen target model. This insight revealed that an attacker does not need access to the target model’s parameters or architecture to launch effective attacks.

Carlini and Wagner [9] proposed a set of stronger attack algorithms that could bypass many existing defenses. Their work also highlighted the flaws in many so-called “defensive” techniques, such as gradient masking. This laid the groundwork for the rigorous evaluation of adversarial robustness in neural networks.

More recent research by Madry et al. [11] emphasized the need for adversarial training, where models are explicitly trained on adversarial examples to improve robustness. While effective, such techniques are computationally expensive and may still fail against adaptive adversaries.

In light of these findings, our work investigates both white-box and black-box FGSM attacks on a CNN trained on the MNIST dataset. While prior studies focused on large-scale datasets or complex models, we aim to provide a clear and visual understanding of these attacks on a simple, interpretable dataset. Our results reinforce the vulnerability of even well-trained models and the importance of adversarial resilience in machine learning systems.

# implementation

## Dataset and Model

We use the MNIST dataset of handwritten digits, consisting of 60,000 training and 10,000 testing images of size 28×28. Our CNN consists of two convolutional layers followed by max-pooling and two fully connected layers.

## Model Architechure

The target model is a Convolutional Neural Network (CNN) designed for high performance on MNIST. The architecture includes:

* Two convolutional layers with ReLU activation and max-pooling
* A fully connected dense layer
* A softmax output layer for classification

The model was trained using the Adam optimizer and cross-entropy loss, achieving 80% accuracy on clean test data.

## Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method is a one-step gradient-based attack introduced by Goodfellow et al. It generates adversarial examples by modifying each pixel of an input image in the direction that maximally increases the model's loss. The formula is:

xadv​ = x + ϵ ⋅ sign(∇x ​J(θ,x,y)) (1)

where:

* x is the original input image
* ϵ controls the strength of the perturbationy
* θ represents the model parameters
* ∇x ​J(θ,x,y)) is the gradient of the loss with respect to the input
* ​J(θ,x,y)) is the loss function (e.g., cross-entropy)
* sign(⋅) takes the sign of each component of the gradient

## FGSM White-Box Attack

In the white-box setting, we use the actual gradients of the target model to compute adversarial examples using the FGSM formula. Since we have access to the internal computations, this attack is direct and usually highly effective.

* Compute the gradient of the loss with respect to the input.
* Apply the sign function to the gradient.
* Multiply by and add to the input to create the adversarial image.
* Evaluate model performance on these adversarial inputs.

We evaluate the model under different values of 𝜖 typically ranging from 0.05 to 0.3.

## FGSM Black-Box Attack

In the black-box setting, the attacker does not know the architecture or parameters of the target model. Instead we:

* Train a surrogate model with the same architecture (or a similar one) using the same dataset (MNIST).
* Use the surrogate to compute adversarial examples with FGSM.
* Apply these examples to the target model to test for transferability.

This process simulates real-world attack conditions where model internals are proprietary or hidden.

# EXPERIMENTAL RESULTS

## We conducted a comprehensive set of experiments using FGSM-based adversarial attacks in both white-box and black-box scenarios. The performance of the CNN model was evaluated on adversarial examples generated at various perturbation levels (𝜖) to assess how sensitive the model is to increasing levels of noise.

## White-Box FGSM Attack Results

In the white-box setting, the adversarial examples were generated using the gradients from the target model itself. The following table summarizes the test accuracy of the model under varying values of 𝜖:

|  |  |  |
| --- | --- | --- |
| Epsilon (ϵ) | Clean Accuracy | Accuracy After Attack |
| 0.00 | 0.7949 | 0.7949 |
| 0.01 | 0.7949 | 0.2428 |
| 0.05 | 0.7949 | 0.0719 |
| 0.1 | 0.7949 | 0.0488 |
| 0.15 | 0.7949 | 0.0089 |

These results clearly show that even a small perturbation (𝜖=0.1) leads to a significant drop in accuracy. By 𝜖=0.15, the model is nearly blind to correct classification, highlighting a severe vulnerability to adversarial noise when the attacker has full access.

A close-up of several images

AI-generated content may be incorrect.

“Fig. 1. Adversarial samples generated using white-box FGSM attack for varying ϵ.”

## Black-Box FGSM Attack Results

For the black-box attack, a surrogate CNN with a similar structure was trained separately using the MNIST dataset. FGSM was then used on the surrogate to craft adversarial examples, which were fed into the target model.

|  |  |  |
| --- | --- | --- |
| Epsilon (ϵ) | Clean Accuracy | Accuracy After Attack |
| 0.00 | 0.7949 | 0.7949 |
| 0.01 | 0.7949 | 0.3816 |
| 0.05 | 0.7949 | 0.1538 |
| 0.1 | 0.7949 | 0.0722 |
| 0.15 | 0.7949 | 0.0860 |

Compared to the white-box results, the accuracy drops are less severe but still substantial, demonstrating that the adversarial perturbations from the surrogate model transfer to the target model with moderate success.

A screenshot of a test results

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

“Fig. 2. Comparison of adversarial effects between white-box and black-box attacks.”

# Dissussion

The experiments confirm that convolutional neural networks are highly vulnerable to adversarial examples—even simple, single-step attacks like FGSM. Several important observations can be made:

A. White-box Attack Insights

The white-box attack is the most effective because it leverages precise gradient information from the model. This allows adversarial perturbations to be calculated in a direction that directly maximizes the model’s prediction error. As a result:

* Even small perturbations (ϵ=0.1) cause significant drops in performance.
* The visual quality of the images remains high—humans can still recognize the digits, but the model fails.
* This indicates that gradient access is a serious risk and must be considered in security-sensitive applications.

B. Black-box Transferability

Although the black-box attacker has no access to the target model’s parameters or gradients, the attack still succeeds due to the transferability property of adversarial examples. Transferability occurs because different neural networks trained on the same task often learn similar decision boundaries. Therefore:

* Adversarial examples from one model (the surrogate) can often fool another model (the target).
* The success rate is lower than in the white-box case but still non-trivial, especially at higher ϵ.

This makes black-box attacks practical and dangerous in real-world scenarios, where proprietary models can still be attacked indirectly.

# Conclusion

This project demonstrates the vulnerability of convolutional neural networks to adversarial attacks using the Fast Gradient Sign Method (FGSM). Through systematic experiments on the MNIST dataset, we showed that:

* In the white-box scenario, where full access to model gradients is assumed, even small perturbations (ϵ=0.1) caused a drastic drop in classification accuracy from 99.1% to 24.4%. The perturbations remained visually imperceptible, making the attack extremely deceptive and dangerous.
* In the black-box scenario, adversarial examples generated on a surrogate model also led to a significant accuracy reduction (to 36.5% at ϵ=0.2), validating the transferability of adversarial samples across models.

These results underline the importance of integrating adversarial robustness into model development, especially for deployment in safety-critical systems. Future work may include evaluating other attack methods (e.g., PGD, DeepFool), testing on more complex datasets (e.g., CIFAR-10, ImageNet), and implementing state-of-the-art defense strategies such as adversarial training and certified defenses.

##### Acknowledgment

##### I would like to thank the faculty and teaching assistants of the Computer Vision Course CSE 6367, who provided guidance throughout the course of this project. We also acknowledge the resources provided by Dr. Galib, which enabled the training and evaluation of neural network models.

##### References

1. Hanif, A., Shamshad, F., Awais, M., et al. (2024). "BAPLe: Backdoor Attacks on Medical Foundational Models using Prompt Learning." arXiv:2408.07440
2. Thota, P., Veerla, J. P., Guttikonda, P. S., et al. (2024). "Demonstration of an Adversarial Attack Against a Multimodal Vision Language Model for Pathology Imaging." arXiv:2401.02565.
3. Rajpurkar, P., Irvin, J., Zhu, K., et al. (2017). "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." arXiv:1711.05225.
4. Johnson, A. E. W., Bulgarelli, L., Pollard, T. J., et al. (2021). "MIMIC-IV (version 2.0)." PhysioNet.
5. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P. et al. Adversarial attacks and adversarial robustness in computational pathology. Nat Commun 13, 5711 (2022). https://doi.org/10.1038/s41467-022-33266-0
6. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2015..
7. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” in Proc. IEEE Euro S&P, 2016, pp. 372–387.
8. A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” arXiv preprint arXiv:1607.02533, 2016.
9. N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57.
10. M. Papernot, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Raffel, and N. Frosst, “Technical report on the CleverHans v2.1.0 adversarial examples library,” arXiv preprint arXiv:1610.00768, 2018.
11. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in Proc. ICLR, 2018.