Understanding Adversarial Vulnerabilities and Defenses in Machine Learning: A Practical Approach Using FGSM

Eunjeong Lee

Abstract

This paper explores the vulnerability of machine learning models, particularly neural networks, to adversarial attacks and investigates defense strategies to improve model robustness. The study focuses on the Fast Gradient Sign Method (FGSM) attack technique and applies it to the MNIST dataset to analyze the model's vulnerability. Additionally, adversarial training is proposed as a defense strategy to enhance model performance and robustness. Experimental results confirm that adversarial training significantly improves the model's robustness against FGSM attacks, highlighting the importance of addressing adversarial vulnerabilities in machine learning systems. The research provides valuable insights into the effectiveness of adversarial defense techniques and emphasizes the necessity of designing robust models that can withstand adversarial perturbations in real-world applications. Adversarial examples, which are intentionally crafted input data that mislead a model into making incorrect predictions, represent a critical challenge to machine learning systems, especially deep learning models. These perturbations, though often imperceptible to the human eye, can significantly affect the model's performance. Adversarial attacks, designed to induce incorrect predictions, have been shown to be particularly effective in various domains such as autonomous driving, security systems, and financial services, where they pose serious threats. In response to these challenges, adversarial learning techniques aim to train models to resist such attacks by incorporating adversarial examples into the training process. The paper examines the types of adversarial attacks, including FGSM and other advanced techniques, and their impact on machine learning models. Furthermore, it discusses the importance of robust defense mechanisms, such as adversarial training, to improve model resilience and ensure the reliability of machine learning systems in the presence of adversarial threats. This work contributes to the growing body of research focused on enhancing the security and robustness of machine learning models against adversarial manipulation.

1 Introduction

Adversarial Examples refer to input data that has been intentionally modified by small perturbations to cause machine learning models, particularly neural networks, to misclassify originally well-classified data. These perturbations are often imperceptible to the human eye but can have a significant impact on the model, leading to incorrect predictions. Adversarial examples can be used to deceive or attack models. For instance, they can trick the image recognition system of an autonomous vehicle, leading to incorrect decisions. Adversarial Attacks are deliberate techniques designed to induce incorrect predictions in a model. In contrast, Adversarial Learning involves training methods aimed at defending against such attacks. Adversarial attacks represent a serious threat in real-world applications, including autonomous driving, financial systems, and security systems. Consequently, developing effective defense mechanisms against these attacks is critical. This paper investigates the vulnerability of machine learning models to FGSM (Fast Gradient Sign Method) attacks using the MNIST dataset and explores the application of Adversarial Training as a defense strategy to improve model performance. Furthermore, we assess the robustness of the model after applying FGSM attacks and compare the effectiveness of the defense through experimental evaluation. graphicx

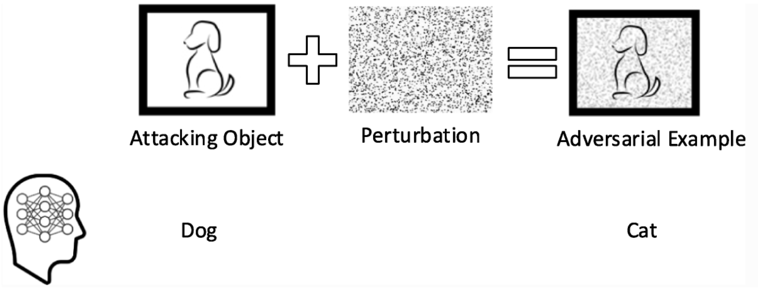

Figure 1: Adeversarial Example [3]

2 Background

2.1 Adversarial Examples and Adversarial Training

Adversarial Examples are input data specifically designed to exploit the vulnerabilities of machine learning models. While these examples may appear almost identical to the original data to the human eye, they can cause the model to misclassify or make entirely incorrect predictions. For example, an image that appears to be a picture of a cat to a human may be classified as a dog by the model. Adversarial examples particularly expose vulnerabilities in complex machine learning structures like deep learning models. This phenomenon occurs by exploiting weaknesses in the decision boundary that the model has learned, and it demonstrates how machine learning systems may not be reliable across diverse environments. For this reason, adversarial examples have become a significant area of research aimed at improving the safety and robustness of machine learning systems.

Adversarial Training is one of the most widely studied defense mechanisms to protect machine learning models from adversarial examples. This approach involves incorporating adversarial examples into the training process to help the model learn more robust decision boundaries. Specifically, adversarial training involves mixing clean data with adversarial examples to form the training dataset, improving the model's generalization ability and making it more capable of responding to various types of attacks. For instance, by including adversarial examples generated using attack techniques like FGSM (Fast Gradient Sign Method) or PGD (Projected Gradient Descent), the model becomes more immune to these specific attacks.

2.2 Types of Adversarial Attacks

Adversarial attacks can be performed in various ways, each with its unique features and objectives. Among the various methods, this paper adopts the FGSM (Fast Gradient Sign Method) due to its low computational cost and simplicity. The main adversarial attack techniques are as follows:

- FGSM (Fast Gradient Sign Method): FGSM is a relatively simple yet efficient technique for generating adversarial examples by utilizing the gradient of the loss function with respect to the input data. This method adds small perturbations to the input data in the direction of the gradient, causing the model to misclassify the input. FGSM is fast and easy to implement, which is why it has been widely used in early research, although it has limitations when dealing with more sophisticated attacks.
- PGD (Projected Gradient Descent): PGD is an extension of FGSM that generates adversarial examples through multiple iterative steps. At each step, the input data is adjusted in the direction of the loss function's gradient, and the values are projected back into the allowed range to limit the perturbation. This iterative process results in adversarial examples that are more refined and effective than those generated by FGSM, making it widely used in adversarial training.
- DeepFool: DeepFool is an attack technique that computes the minimal change required to move an input towards the decision boundary. It gradually adjusts the input data until the model misclassifies it, achieving high success rates with minimal perturbations. This method is often used for vulnerability analysis due to its efficiency in generating adversarial examples with small changes.

2.3 History and Development of Adversarial Attacks

Adversarial examples and attack techniques have been a significant area of research since the early days of machine learning, with their importance becoming particularly evident as deep learning models gained widespread use. Initially, research focused on identifying and addressing model vulnerabilities, and adversarial examples were introduced as a way to create small perturbations in input data that mislead the model. In 2013, Szegedy et al. discovered that deep learning models were particularly susceptible to small input changes, sparking an increase in research on adversarial attacks.

Early attack methods, such as Fast Gradient Sign Method (FGSM), utilized the gradient of the model's loss function to make small modifications to inputs, causing the model to misclassify them. This approach proved effective, and soon more sophisticated techniques, like Projected Gradient Descent (PGD), emerged, offering more powerful attack strategies. These methods highlighted the vulnerabilities of models and spurred the development of defensive strategies.

Early defense techniques mostly focused on input preprocessing or model regularization, but as attackers adapted their strategies, these defenses proved to have limitations. As a result, adversarial training became a key defense method, wherein models are trained on data that includes adversarial examples, making them more robust to attacks. Today, adversarial training remains one of the primary ways to fortify machine learning models against adversarial threats.

The evolution of adversarial attacks and defense techniques has played a crucial role in enhancing the security and robustness of machine learning systems. These advancements are especially vital in high-stakes domains like autonomous vehicles and medical systems, where ensuring the safety of AI systems is paramount.

2.4 Vulnerabilities of Machine Learning Models

Machine learning models, especially deep learning models, are vulnerable to adversarial attacks, and the main reasons for this vulnerability lie in the high complexity of neural networks and the issue of overfitting.

- High Complexity of Neural Networks: Deep learning models have high expressive power, enabling them to learn complex patterns. However, this often leads to very fine and sensitive decision boundaries, making it easy for small perturbations, like adversarial examples, to cause the model to misclassify inputs.
- Overfitting: When a model overfits to the training data, it becomes overly sensitive to new data or perturbations like adversarial examples. Overfitted models tend to perform poorly on test data and are more susceptible to attacks.

These two factors contribute to why deep learning models are particularly vulnerable to adversarial attacks.

2.5 Practical Applications

Adversarial machine learning can be applied in various fields, and its potential risks and benefits are significant. In the field of autonomous driving, attackers can manipulate data received from cameras or sensors, causing the vehicle's system to misjudge its direction or speed, potentially leading to accidents. For example, attackers could alter road signs or distort sensor data at an intersection, causing the self-driving system to make incorrect decisions. This presents a serious threat to the safety and reliability of autonomous vehicles.

In the financial services sector, adversarial attacks can pose major risks as well. For instance, attackers could target anomaly detection models in credit card transaction systems, manipulating them to hide fraudulent transactions as legitimate ones. This would allow attackers to continue engaging in illegal transactions while evading detection, making it difficult for financial institutions to protect against such fraud.

Furthermore, in security systems, facial recognition technology can be deceived by adversarial examples, creating vulnerabilities. Attackers might generate manipulated images to bypass facial recognition systems, allowing unauthorized individuals to gain access to restricted areas or systems. This poses a threat not only to physical security but also to data security, as unauthorized access could lead to breaches of sensitive information.

These examples highlight the importance of defining threat models in real-world systems, helping to understand how attackers can bypass or exploit system vulnerabilities. By examining these practical attack scenarios, it becomes clear that developing robust defenses and improving the resilience of systems against adversarial attacks is crucial for safeguarding against such risks.

3 Methodology

3.1 Attack and Defense Methods

- Attack : In this experiment, we use the FGSM (Fast Gradient Sign Method) attack technique to generate adversarial examples for the MNIST dataset images. FGSM calculates the gradient of the loss function and adds a small distortion to the input images, causing the model to make incorrect predictions. This allows us to evaluate whether the model is vulnerable to adversarial attacks.
- Defense : In this study, we use Adversarial Training as a defense technique. This method modifies the training data so that the model becomes robust to attacks by learning from both clean data and adversarial examples. By doing so, the model is exposed to various adversarial examples and develops the ability to defend against such attacks.

3.2 Model Architecture

In this study, we use a Convolutional Neural Network (CNN) to solve the handwritten digit recognition problem. The model consists of two Convolutional Layers and two Fully Connected Layers, with a ReLU activation function applied after each convolutional layer. The final output of the model provides predictions for 10 classes.
- Conv1: Input channels = 1, Output channels = 32, Kernel size = 3x3
- Conv2: Input channels = 32, Output channels = 64, Kernel size = 3x3
- Fully Connected Layer 1 (fc1): Connects the 64x24x24 output to 128 neurons
- Fully Connected Layer 2 (fc2): Connects the 128 neurons to 10 output classes

The model takes as input 28x28 pixel images of handwritten digits from the MNIST dataset and classifies them into one of 10 classes (digits 0-9).

3.3 FGSM Attack

The Fast Gradient Sign Method (FGSM) is a widely recognized adversarial attack technique introduced by Goodfellow et al. in Explaining and Harnessing Adversarial Examples. This method leverages the gradients of a model's loss function with respect to the input data to generate adversarial examples by introducing small distortions to the input. Despite the simplicity of this approach, it has proven to be remarkably effective in exposing the vulnerabilities of neural networks.

FGSM works by exploiting the way neural networks learn. Instead of minimizing the loss by adjusting the model's weights through backpropagation, the attack modifies the input data itself to maximize the loss. It calculates the gradient of the loss function and perturbs the input image by adding a small adjustment based on the sign of the gradient. This adjustment is scaled by a parameter, which controls the strength of the attack. A larger adjustment introduces more noticeable distortions but may also risk detection by humans, while a smaller adjustment maintains subtlety but can still deceive the model effectively.

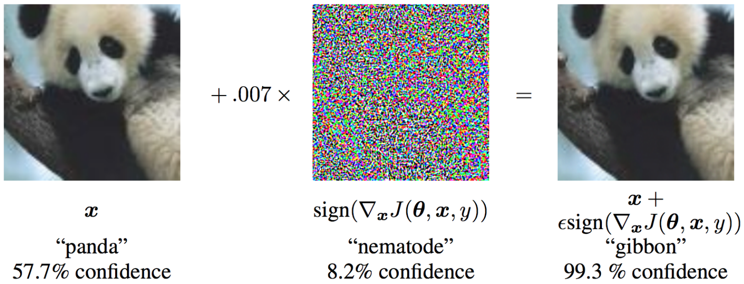
+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure 2: Example FGSM [2]

For example, in the well-known ”panda” experiment, the original image, correctly classified as a ”panda,” is slightly altered with a small distortion. The resulting adversarial image, though visually indistinguishable from the original to humans, is misclassified by the model as a ”gibbon.” This demonstrates how neural networks can be misled by seemingly minor perturbations and highlights the need for robust defenses against adversarial attacks.
In this study, we implement the FGSM attack to evaluate its impact on the model’s performance. By varying the strength of the attack, we analyze the model’s accuracy before and after the adversarial distortions. This approach not only demonstrates the model’s vulnerabilities but also provides insights into its ability to handle adversarial perturbations under different levels of attack.

3.4 Adversarial Training

Adversarial training is a technique that improves the model’s robustness by training it on both clean data and adversarial examples. In this study, we generate adversarial examples using FGSM and train the model on both the original data and these adversarial examples. This enables the model to become more robust to attacks and handle a wider range of adversarial examples.

3.5 Experimental Design

- Dataset: The experiment is conducted using the MNIST dataset [1]. The MNIST dataset is a handwritten digit recognition problem, consisting of 28x28 pixel images of handwritten digits.

The experiments in this study are conducted in the following steps:

- Basic Training: Train the model using the original data and evaluate its baseline performance.
- FGSM Attack: Expose the trained model to FGSM attacks and measure its performance. The accuracy before and after the attack is compared.
- Adversarial Training: Generate adversarial examples using FGSM and incorporate them into the training process to improve the model’s robustness.
- Post-Attack Performance Evaluation: Apply FGSM attacks again to the trained model and evaluate its defense performance.

The model’s performance is measured in terms of accuracy, and the changes in accuracy based on different epsilon values are analyzed.

3.6 Experiment Reproducibility

All experiments were conducted in the Google Colab environment, utilizing Python 3.11.8 and a GPU setup for training and evaluation. The code and settings for the experiment are publicly available on GitHub, allowing other researchers to easily reproduce the experiment.

4 Result

	CNN	CNN+FGSM (ep = 0.1)	CNN+FGSM (ep = 0.2)	CNN+FGSM (ep = 0.3)
Accuracy	98.88	91.68	82.2	67.9

Table 1: Model Performance under FGSM Attack with Different Epsilon Values

The basic CNN model shows high accuracy on normal data, but its accuracy drastically decreases after being subjected to FGSM attacks, confirming its vulnerability to FGSM attacks. In particular, the significant performance degradation as the epsilon value increases demonstrates that FGSM attacks effectively exploit the model’s weaknesses by leveraging the gradient of the loss function. These results highlight the need for robust model design to counter adversarial examples, suggesting that defensive techniques like adversarial training are essential for ensuring the stability and reliability of the model.

	CNN+FGSM	CNN+FGSM+Adversarial Learning
ep = 0.1	91.68	97.9
ep = 0.2	82.2	96.33
ep = 0.3	67.9	93.75
ep = 0.4	50.1	90.52
ep = 0.5	38.09	85.53
ep = 0.6	29.46	78.34
ep = 0.7	24.82	69.27
ep = 0.8	22.62	60.31

Table 2: Comparison of Model Performance with and without Adversarial Training

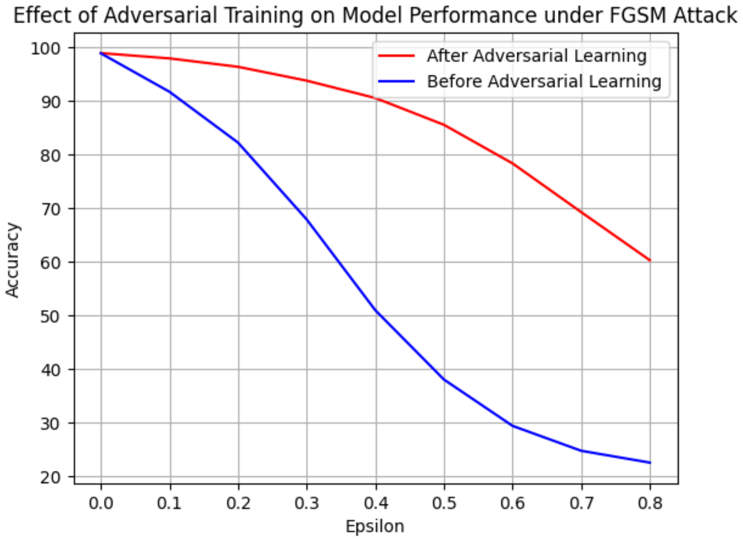

Figure 3: Effect of Adversarial Training on Model Performance under FGSM Attack

As the epsilon value increases, the performance gap between the basic CNN model and the adversarially trained model becomes more pronounced, clearly demonstrating that adversarial training is an effective defense method against strong attacks. In adversarial environments such as FGSM attacks, adversarial training significantly enhances the robustness of the model, helping it to go beyond simply achieving high accuracy and instead develop a strong defense against attacks. Specifically, models with adversarial training maintain stable performance and provide reliable predictions across various attack scenarios. These results emphasize the importance of designing AI systems that prioritize stability, reliability, and robustness. In other words, ensuring that AI systems can adapt to and respond to a variety of threats in real-world environments becomes a goal that is more important than merely achieving high accuracy.

5 Conclusion

This study experimentally analyzed the vulnerability of machine learning models, especially neural networks, to adversarial attacks, and evaluated the effectiveness of adversarial training as a defense mechanism. The experimental results showed that the basic CNN model, when subjected to FGSM (Fast Gradient Sign Method) attacks, experienced a significant drop in accuracy, highlighting the model's inability to effectively handle adversarial examples. In contrast, the model trained with adversarial training demonstrated much higher robustness, maintaining stable performance even as the attack strength increased.

These findings confirm that adversarial examples reveal critical vulnerabilities in machine learning models, and adversarial training is an effective defense mechanism for improving robustness. Additionally, the study emphasizes the importance of not only achieving high accuracy but also ensuring stability and reliability in model performance. Specifically, the ability for AI systems to adapt to and respond to various threats in real-world environments is essential for the successful deployment of these models.

Therefore, this research highlights the significance of adversarial training in enhancing the safety and robustness of machine learning models. Future work should explore more sophisticated attack methods and defensive strategies to further improve the resilience of machine learning systems.

References

[1] Y. LeCun, C. Cortes, and C. J. Burges. Gradient-based learning applied to document recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1998.

[2] PyTorch. Fgsm tutorial, 2024. Available at https://pytorch.org/tutorials/beginner/fgsm_tutorial.html.

[3] L. Sun, M. Tan, and Z. Zhou. A survey of practical adversarial example attacks. *Cybersecur*, 1(9), 2018.