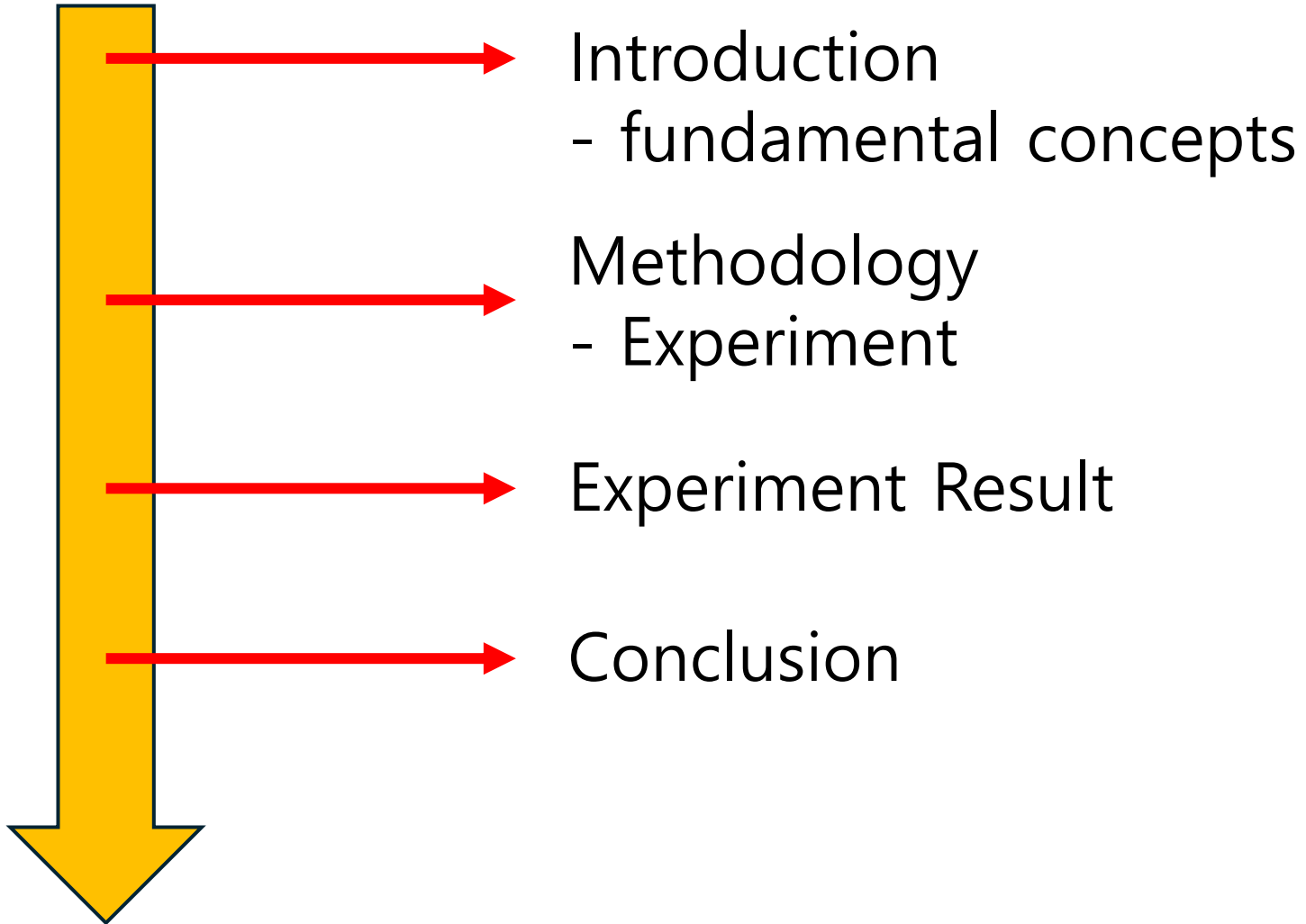# Adversarial Learning

Eunjeong Lee

Contents

Introduction : **Adversarial Examples**

- **Adversarial Examples** are inputs intentionally modified with small perturbations to mislead machine learning models, especially neural networks.

- These perturbations are **imperceptible to humans** but significantly impact model predictions.

- While the original and adversarial images appear identical to humans, the model **misclassifies** them.

- Adversarial examples **highlight vulnerabilities in deep learning models** and pose security risks.

Introduction : **Adversarial Examples**
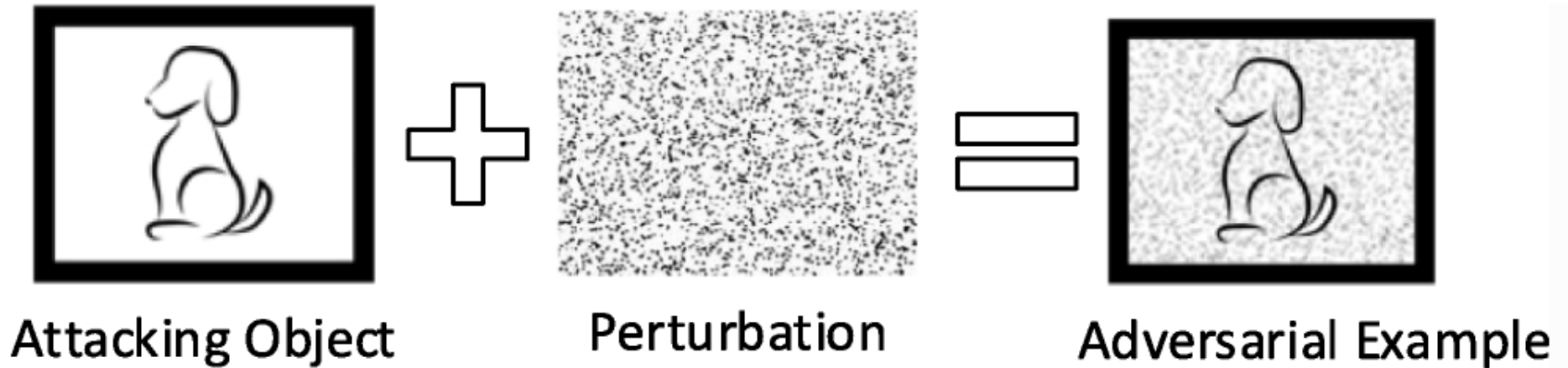
- Example: A model **misidentifying a dog as a cat** due to adversarial noise.

Introduction : **Adversarial Attacks**

- Adversarial attacks intentionally manipulate inputs to mislead machine learning models.

- These attacks can take various forms, each with unique characteristics and objectives.

- In our experiment, we chose FGSM (Fast Gradient Sign Method) due to its low computational cost and simplicity.

Introduction : **Adversarial Attacks(FGSM)**

📌 **FGSM (Fast Gradient Sign Method)**

- A simple yet effective method for generating adversarial examples.

- Uses the **gradient of the loss function** to perturb input data.

- Adds a small change **in the direction of the gradient** to mislead the model.

◆ **Strengths & Limitations**

✅ Fast and easy to implement

✅ Widely used in early research

❌ Less effective against advanced defenses

Introduction : **Adversarial Learning**

- **Adversarial Learning** is a machine learning approach that improves model robustness by **training it with adversarial examples.**

- It is used to **defend against adversarial attacks** and enhance model security.

- The main idea is to expose the model to **perturbed inputs** during training, helping it recognize and resist adversarial manipulations.

◆ **Key Aspects of Adversarial Learning**

✅ Defense against adversarial attacks

✅ Enhances model robustness & security

✅ Commonly used in deep learning applications

❌ May reduce model accuracy on clean data

Example) Autonomous Vehicles

- **Adversarial attacks** can deceive autonomous vehicles' **image recognition and sensor systems**, leading to incorrect decisions.

- **LiDAR spoofing attacks**: Can cause the vehicle to misinterpret or fail to detect obstacles.

- **Adversarial patch attacks**: Adding patterns to objects like pedestrians can make the vehicle fail to recognize them.

Solutions

- **Adversarial Training**: Training models to be more resistant to attacks.

- **Improved Feature Extraction**: Ensuring the model learns the true features, not adversarial ones.

The purpose of the paper

- **Analyzing model vulnerability** using **FGSM attack** on the **MNIST dataset**

- **Improving model performance** with **Adversarial Learning** using **FGSM**

- **Evaluating model robustness** and comparing the effectiveness of **Adversarial Training**

Methodology : **MNIST Dataset**

📝 **MNIST Dataset**

- It is a large collection of handwritten digits, commonly used for training and testing machine learning models.

- The dataset consists of **60,000 training images** and **10,000 testing images** of handwritten digits (0–9).
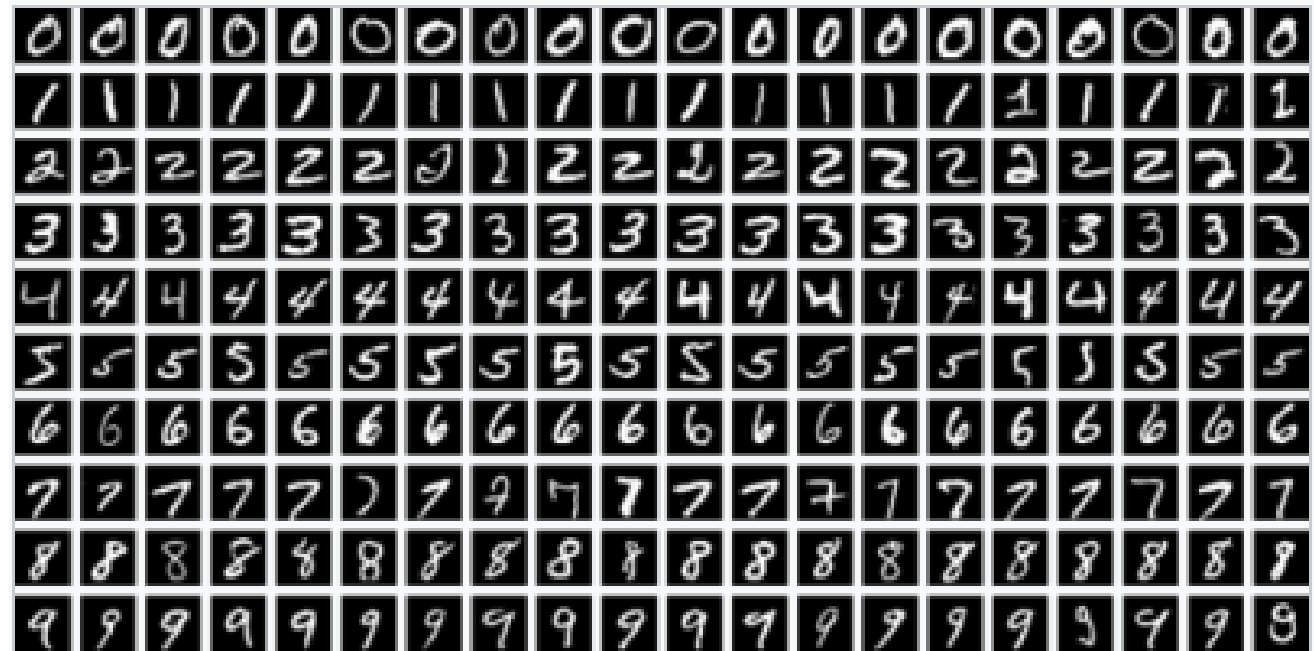
📊 **Features**

- **Image size**: 28x28 pixels, grayscale

- **Classes**: 10 classes (digits 0–9)

Experiment

How can we apply FGSM attack to the MNIST dataset?

Original Image (e.g., Number 3)

- Original data: MNIST handwritten digit "3"

- The model accurately recognizes the image

- No alterations to the image

- The model can classify it correctly

Image with FGSM Applied

- Image with added noise

- Differences are barely visible to the human eye

- The original image is almost unchanged, but small perturbations are added to mislead the model

- Tiny changes lead to a significant impact on the model's prediction

Experiment

1) **Performance Variation in CNN Model Before and After FGSM Attack**

   Objective: Measure the performance change of the CNN model before and after the

   FGSM attack is applied.

   Comparison details:

   - Base Model: CNN model trained on MNIST data

   - After FGSM Attack: Model performance after being attacked by FGSM (accuracy drop)

Experiment

2) Performance Change Before and After Adversarial Training

    Objective: Measure the performance change before and after applying Adversarial

                  Training on FGSM-attacked data

    Comparison details:

     - Before the Attack: Vulnerable model to FGSM attack

     - After the Attack: Improved robustness through adversarial training (enhanced

    defense performance against attacks)

Experiment

Performance Evaluation

Evaluation Metric: Accuracy

Variable: Analysis of model accuracy changes based on epsilon values

Experiment Result

1) Performance Variation in CNN Model Before and After FGSM Attack

	Basic 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

ep => epsilon, Noise Intensity

As the epsilon value increases, stronger noise is applied, and as the epsilon value decreases,

the noise becomes weaker.

Experiment Result

1) Performance Variation in CNN Model Before and After FGSM Attack

	Basic 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

- CNN model shows high accuracy on normal data but **significantly drops** after FGSM attack, indicating vulnerability.

- **Epsilon** increases lead to larger performance degradation, proving FGSM targets model weaknesses effectively.

- Results highlight the need for **robust models** and defense techniques like **adversarial training** for **stability** and **reliability**.

Experiment Result

2) Performance Change Before and After Adversarial Training

	CNN + FGSM	CNN + FGSM + **Adversarial Learning**
ep = 0.1	91.68%	97.90%
ep = 0.2	82.20%	96.33%
ep = 0.3	67.90%	93.75%
ep = 0.4	50.10%	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%

Experiment Result

2) Performance Change Before and After Adversarial Training

Effect of Adversarial Training on Model Performance under FGSM Attack
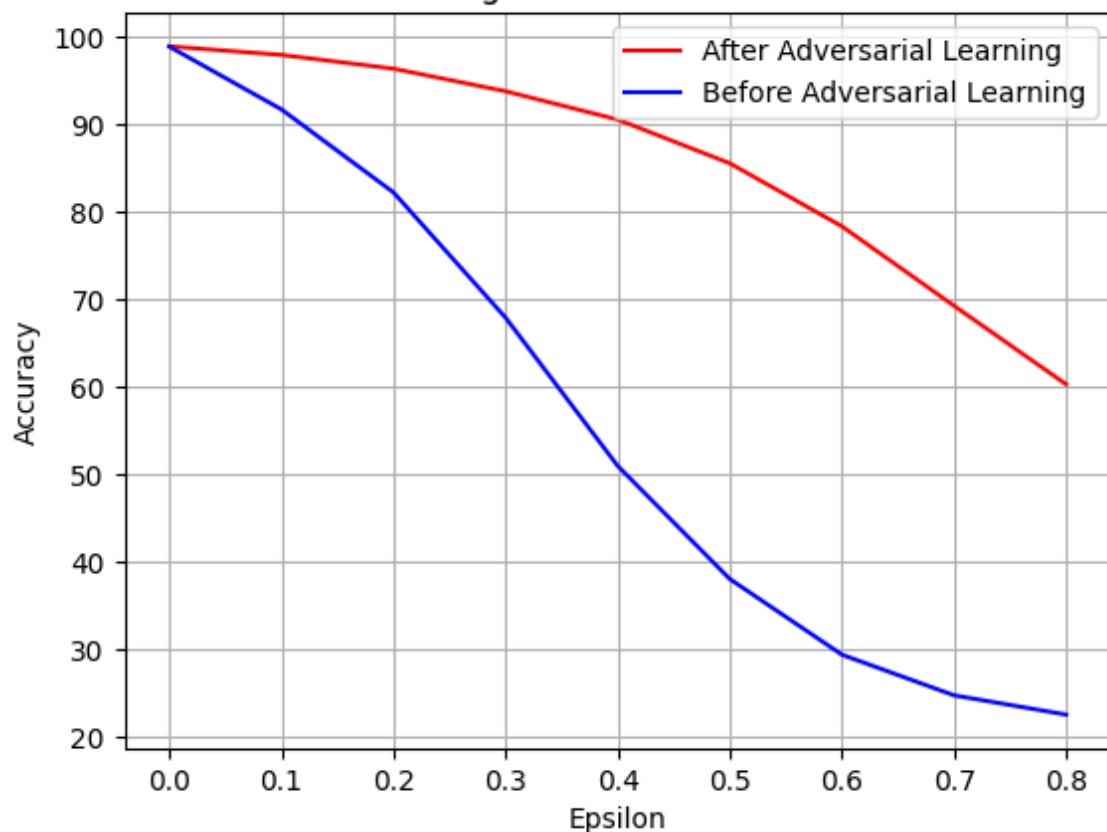
- As **epsilon** increases, the performance gap between the CNN and adversarially trained models grows, proving that **adversarial training** is an effective defense.

- Adversarial training **enhances robustness**, helping models defend against attacks while maintaining stable performance.

Conclusion

Study on Adversarial Attacks and Defense Mechanisms

- This study experimentally analyzes the vulnerability of machine learning models, especially neural networks, to adversarial attacks and evaluates the effectiveness of **adversarial training** as a defense.

- **FGSM Attack** on the basic CNN model caused a sharp drop in accuracy, showing the model's inability to handle adversarial examples effectively.

- In contrast, the model with **adversarial training** maintained high robustness, showing stable performance even as the attack intensity increased.

Conclusion

Key Findings and Implications

- Adversarial examples reveal the vulnerabilities of machine learning models, and adversarial training is an effective defense mechanism.

- Emphasizes the need for AI systems to focus on **stability, reliability,** and **robustness**, not just accuracy, to adapt and defend against real-world threats.

- Future research should explore advanced attack techniques and defense strategies to further improve model robustness.