Exploring Privacy Attacks on AI Models and Defense Strategies

Proposal

Xiaohai Wang

Master of Cybersecurity and Threat Intelligence

Professor: *Rozita Dara*

University of Guelph

May 30, 2025

1 Introduction

Artificial intelligence (AI) is increasingly integrated into systems that handle sensitive personal data, including healthcare, finance, and social platforms. However, AI models themselves are vulnerable to attacks that exploit or compromise privacy, posing a significant threat to both users and organizations. Common attacks include data poisoning, adversarial manipulation, and model inversion, which can undermine data confidentiality and model integrity. This project aims to explore the landscape of privacy attacks against AI models and develop robust defense strategies.

2 Problem Statement / Motivation

Modern AI systems are typically trained on massive datasets containing sensitive information. While these models can offer superior predictive performance, they are also vulnerable to attacks that can compromise data privacy and the model's security. For instance:

- **Data Poisoning** can bias or degrade AI performance, leading to privacy leaks or unfair outcomes.
- **Model Inversion** attacks can reveal sensitive training data.
- **Membership Inference** can disclose whether a user's data was part of the training dataset.

These threats are especially concerning as AI becomes more ubiquitous. Thus, there is an urgent need to:

1. Understand how these privacy attacks exploit AI vulnerabilities, and
2. Develop effective countermeasures to protect sensitive data and maintain the trustworthiness of AI systems.

3 Objectives

1. **Analyze Privacy Attack Techniques**: Study and simulate different privacy attacks—such as data poisoning, adversarial evasion, model inversion, and membership inference—to evaluate their impact on AI models trained with sensitive personal data.
2. **Develop Detection and Mitigation Strategies**: Design robust defense mechanisms leveraging techniques such as adversarial training and differential privacy to protect AI models against these attacks.

3. **Establish an Evaluation Framework**: Implement experimental protocols to assess the effectiveness of attack and defense strategies using quantitative and qualitative metrics (e.g., accuracy degradation, data exposure risk).

4 Rough Methodology

- **Literature Review**: Conduct a comprehensive review of current privacy attack methods in AI systems and their impact on data confidentiality.
- **Attack Simulation**: Develop Python-based implementations of selected attack types (data poisoning, adversarial attacks, model inversion, membership inference, etc.) to test vulnerabilities in AI models.
- **Defense Implementation**: Implement defense techniques—such as differential privacy mechanisms, adversarial training, and data sanitization—to mitigate these threats.
- **Evaluation**: Perform systematic testing to measure:
  - Model performance before and after attacks
  - Data exposure risk under attacks
  - Effectiveness of defense mechanisms

5 Data

For this project, we will use the **CDC Diabetes Health Indicators Dataset** from the Behavioral Risk Factor Surveillance System (BRFSS). This dataset includes healthcare statistics and lifestyle survey information about people, focusing on diabetes diagnosis, pre-diabetic states, and healthy individuals.

Key characteristics:

- **Dataset size:** 253,680 instances
- **Features:** 21 categorical and integer variables, covering demographics, lab test results, and survey answers
- **Target variable:** Classification label (`0` for no diabetes or during pregnancy, `1` for prediabetic, `2` for diabetic)

This dataset will be used to simulate and analyze privacy attacks—including membership inference, model inversion, and data poisoning—and to evaluate the effectiveness of defense strategies. Due to the dataset size and class imbalance, we will consider sampling techniques for efficient experimentation.

6 Tools

- **Python** (primary language)
- **PyTorch / TensorFlow** (for model development)
- **Scikit-learn** (for preprocessing and classical ML models)
- **Adversarial Robustness Toolbox (ART)** for attack and defense simulation
- **IDA Pro / Ghidra** (if integrating binary-level attack scenarios)

7 Experimental Settings

- Use classification tasks with 2-3 ML models (e.g., CNNs, SVMs) as baselines.
- Apply different attack methods sequentially to test model vulnerabilities.
- Evaluate model performance degradation and quantify privacy risk.
- Apply defense techniques and measure improvements in robustness and privacy protection.
- Run experiments on a local GPU or available computing resources (university cluster).

8 Expected Outcomes

- A comprehensive understanding of privacy attacks (data poisoning, adversarial evasion, model inversion, membership inference, etc.) and their mechanisms in AI systems.
- A set of Python-based simulation tools to recreate and test these attacks.
- A report evaluating the effectiveness of various defense strategies and their trade-offs in model performance and privacy protection.
- Recommendations for best practices in safeguarding AI models handling sensitive personal information.