University of Central Missouri

Department of Computer Science & Cybersecurity

CS5720 Neural Networks and Deep Learning

Summer 2025

Home Assignment 4. (Cover Ch 11, 12)

Student name: Ajay Muppa

ID: 700769264

Submission Requirements:

- Total Points: 100
- Once finished your assignment push your source code to your repo (GitHub) and explain the work through the ReadMe file properly. Make sure you add your student info in the ReadMe file.
- Submit your GitHub link and video on BrightSpace.
- Comment your code appropriately ***IMPORTANT.***
- Make a simple video about 2 to 3 minutes which includes demonstration of your home assignment and explanation of code snippets.
- Any submission after provided deadline is considered as a late submission.

1. GAN Architecture

Explain the adversarial process in GAN training. What are the goals of the generator and discriminator, and how do they improve through competition? Diagram of the GAN architecture showing the data flow and objectives of each component.

Answer: Adversarial Process in GANs

GANs (Generative Adversarial Networks) involve two neural networks:

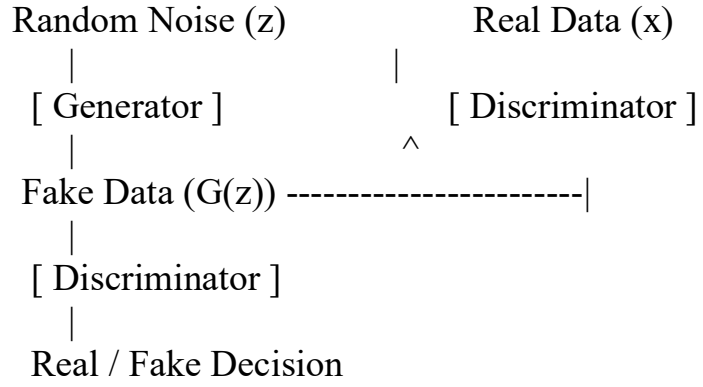
- **Generator (G)**: Learns to generate realistic fake data from random noise.
- **Discriminator (D)**: Learns to distinguish between real data and fake data.

Training Process (Adversarial Loop):

1. **G tries to fool D** by generating data that mimics real data.
2. **D tries to detect** whether a sample is real or generated.
3. **Loss Feedback**:
    - G improves by minimizing the probability that D correctly classifies its outputs as fake.
    - D improves by maximizing classification accuracy.

The training is a two-player minimax game:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$



Generator Goal: Fool Discriminator → Generate Realistic Data
Discriminator Goal: Distinguish Real from Fake

2. Ethics and AI Harm

Choose one of the following real-world AI harms discussed in Chapter 12:

- Representational harm
- Allocational harm
- Misinformation in generative AI

Describe a real or hypothetical application where this harm may occur. Then, suggest **two harm mitigation strategies** that could reduce its impact based on the lecture.

Answer: Scenario: Loan Approval System
A financial AI model allocates loans based on historical credit data.
Harm: If trained on biased data (e.g., denying loans to certain zip codes), the system allocates resources unfairly, perpetuating systemic discrimination.
Mitigation Strategies:

1. **Bias-Aware Training: Use fairness-aware algorithms that regularize against group-level disparities.**
2. **Auditing & Transparency: Regular audits for disparate impact metrics, explainability tools to review decisions.**

3. Programming Task (Basic GAN Implementation)

Implement a simple GAN using PyTorch or TensorFlow to generate handwritten digits from the MNIST dataset.

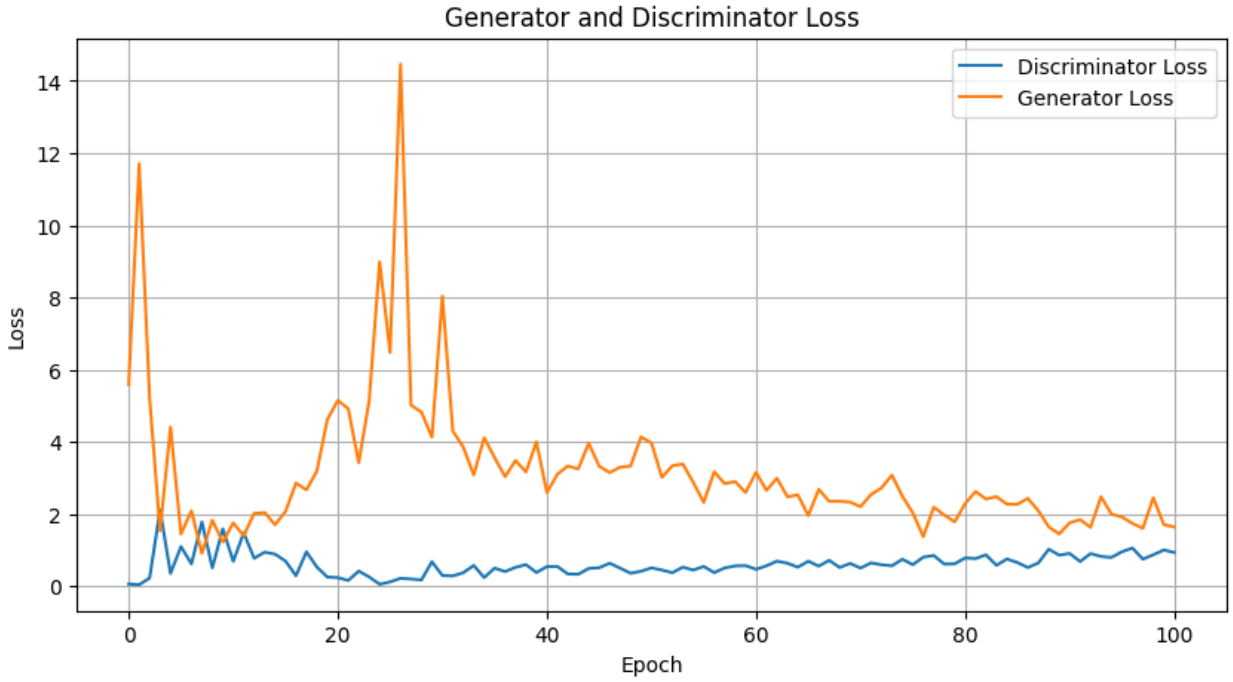
Requirements:

- Generator and Discriminator architecture
- Training loop with alternating updates
- Show sample images at Epoch 0, 50, and 100

Deliverables:

- Generated image samples
- Screenshot or plots comparing losses of generator and discriminator over time

Output:



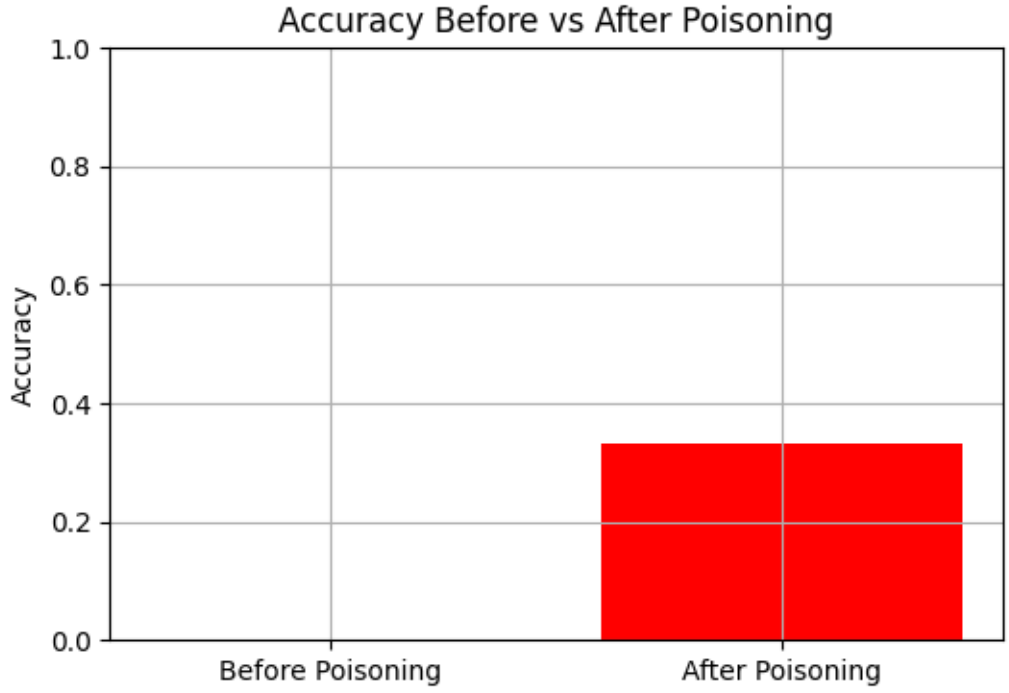
4. Programming Task (Data Poisoning Simulation)

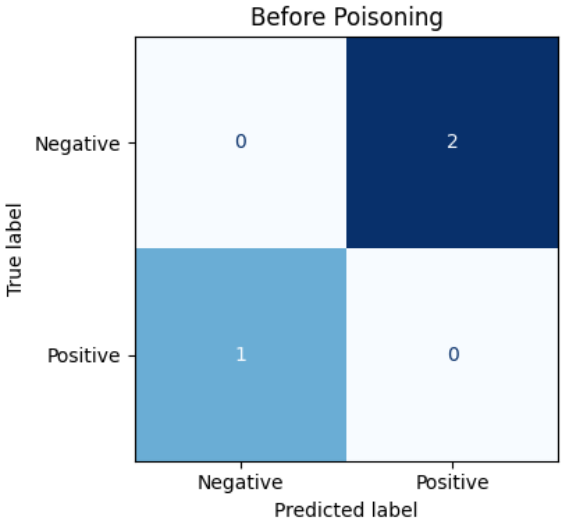
Simulate a data poisoning attack on a sentiment classifier.
Start with a basic classifier trained on a small dataset (e.g., movie reviews). Then, poison some training data by flipping labels for phrases about a specific entity (e.g., "UC Berkeley").

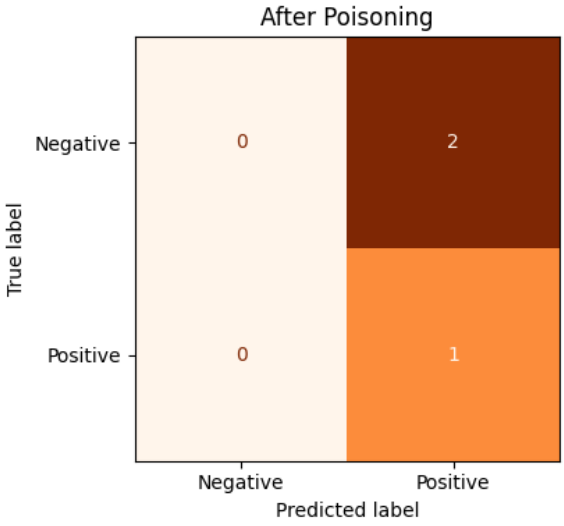
Deliverables:

- Graphs showing accuracy and confusion matrix before and after poisoning
- How the poisoning affected results

Output:







5. Legal and Ethical Implications of GenAI

Discuss the legal and ethical concerns of AI-generated content based on the examples of:

- Memorizing private data (e.g., names in GPT-2)
- Generating copyrighted material (e.g., Harry Potter text)

Do you believe generative AI models should be restricted from certain data during training? Justify your answer.

Answer: **Concerns:**

1. **Memorizing Private Data**:
    - Early models (like GPT-2) memorized names/emails from training corpora.
    - Raises privacy and GDPR compliance issues.
2. **Generating Copyrighted Material**:
    - GPT-like models may regenerate copyrighted text (e.g., paragraphs from *Harry Potter*).
    - This violates IP laws, especially if used for profit.

Opinion:

Yes, **restrictions are necessary**:

- Excluding personal/private and copyrighted content helps maintain ethical and legal standards.
- Alternatives: Use licensed, synthetic, or anonymized data.

6. Bias & Fairness Tools

Visit Aequitas Bias Audit Tool.
Choose a bias metric (e.g., false negative rate parity) and describe:

- What the metric measures
- Why it's important
- How a model might fail this metric

Optional: Try applying the tool to any small dataset or use demo data.

Answer: **Metric: False Negative Rate Parity**

- What it Measures:
  Whether the rate of false negatives (failing to identify a positive case) is consistent across demographic groups.
- Why it Matters:
  In criminal justice or healthcare, high FNR in one group can deny critical services or misjudge risk.
- Failure Example:
  A model underpredicts high-risk cases in one group (e.g., certain ethnicities), leading to under-monitoring.