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: M. Siddartha Reddy
700774070

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.

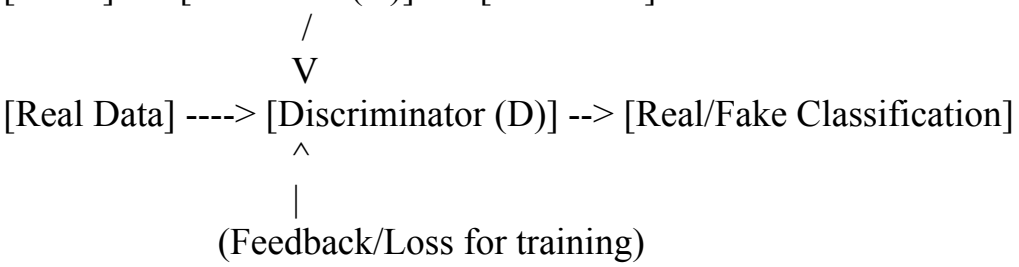
Ans:

GAN training is an **adversarial process** between two neural networks:

- **Generator (G):** Creates fake data from noise. Its goal is to **fool** the discriminator into thinking its fakes are real.
- **Discriminator (D):** Distinguishes between real data and the generator's fakes. Its goal is to **accurately classify** them.

They improve through **competition**: G gets better at generating fakes as D gets better at detecting them, and vice-versa, pushing both towards higher performance until G can create highly realistic data.

Diagram:



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.

Ans:

I will choose **Allocational harm**.

Application: A loan application AI that disproportionately denies loans to minority groups due to historical biases in its training data, even if it doesn't explicitly use race as a feature.

Harm Mitigation Strategies:

1. **Fairness-aware AI Development:** Employ algorithms and training techniques designed to reduce bias, such as adversarial debiasing or re-

weighting training data to ensure equitable representation across sensitive attributes.

2. **Regular Auditing and Impact Assessments:** Periodically and thoroughly audit the AI system's decisions for disparate impact on different demographic groups, and conduct pre-deployment impact assessments to identify and address potential biases before widespread use

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

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

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.

Ans:

AI-generated content faces legal and ethical challenges:

1. **Memorizing Private Data:** Raises privacy violations (e.g., GDPR) if models reproduce personal information from training data. Ethically, it's a breach of trust and can lead to misuse.
2. **Generating Copyrighted Material:** Poses copyright infringement risks if AI output is substantially similar to existing works. Ethically, it devalues human creativity and raises concerns about fair compensation for creators.

Yes, generative AI models should be restricted from certain data during training.

Justification:

- **Protect Privacy:** Uphold data protection laws and individual rights.
- **Uphold Copyright:** Ensure fair compensation for creators and prevent devaluing original works.
- **Promote Responsible AI:** Encourage ethical development and public trust by mitigating legal risks and potential harm.

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.

Ans:

Bias Metric: False Negative Rate Parity (FNR Parity)

- **What it measures:** FNR parity measures whether the **false negative rate** (the proportion of actual positive cases that the model incorrectly predicts as negative) is approximately the same across different demographic groups. For example, in a medical diagnosis model for a disease, a false negative means the model said a patient *doesn't* have the disease, but they *actually do*.
    - **Formula:** FNR = (False Negatives) / (False Negatives + True Positives)
    - FNR parity means FNR for Group A ≈ FNR for Group B.
- **Why it's important:** FNR parity is crucial in applications where **missing a positive case has high stakes or negative consequences** for an individual. This is common in "assistive" or "preventative" interventions.
    - **Examples:** Denying a loan to someone who would have paid it back, failing to diagnose a serious illness, or overlooking a qualified job applicant. When FNR is higher for a specific group, that group is disproportionately denied access to opportunities or critical services.
- **How a model might fail this metric:**
    - **Underrepresentation in Training Data:** If a particular group (e.g., a minority demographic) is underrepresented in the training data, the model might not learn their specific patterns as well, leading to higher FNR for that group.
    - **Feature Bias:** Features used by the model might inadvertently correlate with sensitive attributes, causing the model to make more false negatives for certain groups. For instance, if a loan application model indirectly uses zip codes correlated with race, it might disproportionately deny loans (false negatives) to certain racial groups.

- **Algorithm Design:** The optimization objective of the model might prioritize overall accuracy, or another metric, over equitable error rates across groups, leading to disparities in FNR. For example, if a model is optimized to minimize overall errors, it might sacrifice performance for smaller groups if it improves overall accuracy.