**University of Central Missouri**

**Department of Computer Science & Cybersecurity**

**CS5720 Neural network and Deep learning**

**Spring 2025**

**Home Assignment 5. (Cover Ch 11, 12)**

**Student name:**

**VIDHAI DIAMOND 700771948**

**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 the BB.
* 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.

In GAN training, the generator and discriminator engage in an adversarial process where the generator aims to create realistic data indistinguishable from real data, while the discriminator tries to distinguish between real and generated data. This competition forces both networks to improve; the generator learns to produce higher-quality data, and the discriminator becomes more adept at detecting fakes.

Goals

A diagram of a model

Description automatically generated

GANs improve through a competitive process where two neural networks, a generator and a discriminator, work against each other.

Goals of Generator and Discriminator:

* **Generator:**

The generator's goal is to produce synthetic data samples that are indistinguishable from real data. It does this by learning to map random noise to realistic data.

* **Discriminator:**

The discriminator's goal is to correctly classify data as either real or generated by the generator. It learns to differentiate between real data and the generator's outputs.

**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.

Chosen Harm: **Misinformation in Generative AI**

An example of misinformation harm is the use of generative AI to create fake news articles during an election. AI-generated content can look credible, making it easy to spread false information that influences public opinion and damages trust in democratic processes.

Two mitigation strategies are:

1. Content Verification Systems – using AI tools and human reviewers to detect, label, and limit the spread of misinformation.
2. Transparency and Provenance Marking – embedding metadata or visible disclaimers to show that the content was AI-generated, helping users verify authenticity.

**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

**Graphs**

* Confusion matrix **before** poisoning
* Confusion matrix **after** poisoning

**Results**

* Print accuracy before and after
* Clearly show the drop in performance or misclassifications especially related to "UC Berkeley" phrases

**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.

**Legal and Ethical Concerns of AI-Generated Content**

There are major legal and ethical concerns around how generative AI models handle sensitive and copyrighted data. One example is **memorizing private data**, such as how earlier versions of models like GPT-2 sometimes unintentionally reproduced names, phone numbers, or confidential information from their training sets. This raises serious privacy issues, since individuals' personal data could be exposed without their consent, violating privacy laws like GDPR.

Another major concern is **generating copyrighted material**. For instance, if an AI can produce text that closely mimics or even directly copies parts of *Harry Potter*, it risks infringing on the intellectual property rights of the original author. This could lead to legal action and undermines the rights of creators to control how their work is used and profited from.

**Should AI models be restricted from certain data during training?**

Yes, I believe generative AI models should be restricted from using certain kinds of data, especially private information and copyrighted materials without permission. Training on sensitive or protected content without proper safeguards not only exposes companies to legal risks but also erodes public trust in AI technology. Restricting such data ensures that models respect individuals' privacy, uphold copyright laws, and promote the responsible use of AI.

**6. Bias & Fairness Tools**

Visit [Aequitas Bias Audit Tool](http://www.datasciencepublicpolicy.org/projects/aequitas/).  
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.

**Bias Metric: False Negative Rate Parity**

**What the metric measures:**  
False Negative Rate (FNR) parity checks whether different groups (e.g., by race, gender) have similar false negative rates. A **false negative** occurs when the model incorrectly predicts a negative outcome for a case that is actually positive (for example, saying someone is "not eligible" when they actually are). FNR parity means no group should experience more false negatives than others.

**Why it's important:**  
It is important because false negatives can deny people critical opportunities — like being denied a loan, job, or medical treatment — when they deserve it. If certain groups have higher false negative rates, it can systematically disadvantage them and reinforce existing inequalities.

**How a model might fail this metric:**  
A model might fail FNR parity if, for example, it predicts "loan denial" much more often for one racial group even when they meet eligibility criteria, compared to others. This would show the model is unfairly biased, punishing one group more with wrongful negative outcomes.