**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: ALCHURI NIKHITHA**

**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:

### **1.Adversarial Process in GAN Training (with Diagram):**

Generative Adversarial Networks (GANs) consist of two competing neural networks:

1. **Generator (G)**: Tries to produce realistic data (e.g., images).
2. **Discriminator (D)**: Tries to distinguish between real data (from the dataset) and fake data (produced by the generator).

### **2. Goals of Each Component:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Component** | **Input** | **Goal** |  | | --- | --- | --- | --- | | **Generator** | Random noise vector (z) | Fool the discriminator by generating realistic samples |  | | **Discriminator** | Real data or generated data | Correctly classify real vs. fake samples |  | |

### **Adversarial Training Loop:**

1. **Discriminator Training:**
   * Show real samples → output should be **1** (real).
   * Show fake samples (from G) → output should be **0** (fake).
   * Update D to minimize classification error.
2. **Generator Training:**
   * Generate fake samples.
   * Pass them through D.
   * Update G to maximize D’s mistake (i.e., D classifies fake as real).

The two networks **compete**:

* G improves at fooling D.
* D improves at spotting fakes.

Eventually, the generator produces samples indistinguishable from real data.

### **4.GAN Objective Functions:**

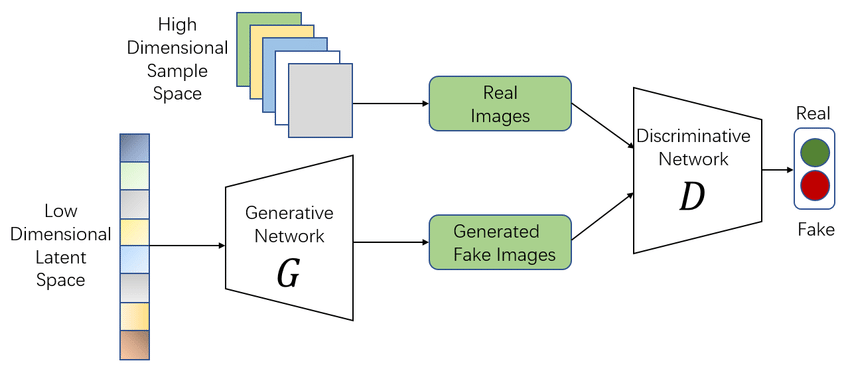
* **Discriminator Loss:**

LD=−Ex∼pdata[log⁡D(x)]−Ez∼pz[log⁡(1−D(G(z)))]\mathcal{L}\_D = -\mathbb{E}\_{x \sim p\_{\text{data}}}[\log D(x)] - \mathbb{E}\_{z \sim p\_z}[\log(1 - D(G(z)))]LD​=−Ex∼pdata​​[logD(x)]−Ez∼pz​​[log(1−D(G(z)))]

* **Generator Loss:**

LG=−Ez∼pz[log⁡D(G(z))]\mathcal{L}\_G = -\mathbb{E}\_{z \sim p\_z}[\log D(G(z))]LG​=−Ez∼pz​​[logD(G(z))]

### **5.Diagram: GAN Architecture:**



### **Data Flow & Objectives:**

* **Generator (G):**
  + **Input:** Random noise vector zzz sampled from a simple distribution (e.g., Gaussian).
  + **Output:** Synthetic (fake) data sample G(z)G(z)G(z) meant to resemble real data.
  + **Objective:** Learn to generate realistic data that **fools the discriminator** into classifying generated data as real.
* **Discriminator (D):**
  + **Input:** Either a real data sample xxx or a fake sample G(z)G(z)G(z).
  + **Output:** A probability score indicating whether the input is real (close to 1) or fake (close to 0).
  + **Objective:** Accurately **classify real vs fake data**.
* **Training process:**
  + **Discriminator:** Maximizes the probability of correctly classifying real and fake samples.
  + **Generator:** Tries to minimize the discriminator’s ability to detect fake samples, i.e., maximize the discriminator’s error on generated samples.

**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:**

### **Definition: Representational Harm:**

Representational harm occurs when AI systems reinforce or amplify stereotypes, biases, or misrepresent certain groups, often leading to social and cultural marginalization.

### **Real-World Example: Image Generation Tool Bias:**

**Application:**  
An AI-powered image generator (like those used to create avatars, illustrations, or stock photos) trained primarily on Western-centric and male-dominated datasets.

**Harm:**  
When prompted with professions like "doctor," "engineer," or "scientist," the generator produces mostly images of white men, while prompts like "nurse" or "teacher" produce mostly women or people of color.  
This **reinforces gender and racial stereotypes,** which can distort public perception and limit inclusivity.

### **Harm Mitigation Strategies:**

1. **Diverse and Inclusive Training Data**
   * Curate training datasets that **intentionally balance gender, race, culture, and appearance** across various professions and contexts.
   * Audit datasets to identify underrepresented groups and correct imbalances.
2. **Bias Audits and Model Evaluation**
   * Regularly evaluate model outputs using **fairness and representational metrics**.
   * Involve multidisciplinary teams, including ethicists and representatives of affected groups, to assess whether generated content aligns with social equity goals.

**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:

### **Legal and Ethical Implications of Generative AI:**

Generative AI models like GPT and diffusion-based tools (e.g., image generators) raise significant **legal** and **ethical** concerns. Let's examine two key examples:

### 1. **Memorizing Private Data (e.g., names, phone numbers in GPT-2):**

**Concern:**

* Large language models (LLMs) trained on public internet data sometimes memorize **private or sensitive information**.
* For example, GPT-2 was shown to output real names, phone numbers, or emails present in its training data.

**Legal Implications:**

* May violate **data protection laws** like **GDPR** or **CCPA**, especially if personally identifiable information (PII) is exposed.
* Data subjects may not have **consented** to their information being used in training.

**Ethical Implications:**

* Undermines **user trust**.
* Breaches the principle of **data minimization and purpose limitation**.

### 2. **Generating Copyrighted Material (e.g., Harry Potter passages):**

**Concern:**

* LLMs can reproduce long sequences of copyrighted works seen in training, such as passages from Harry Potter*.*
* Raises issues of **unauthorized reproduction** and **derivative works.**

**Legal Implications:**

* Potential **copyright infringement** if the generated output is substantially similar to protected content.
* U.S. and international laws grant exclusive reproduction rights to copyright holders.

**Ethical Implications:**

* Exploits authors’ creative labor without compensation or credit.
* Blurs lines of **originality** and **authorship** in creative work.

### **Should GenAI Models Be Restricted from Certain Data During Training?**

**Yes, with clear boundaries.**

**Justification:**

1. **Privacy Protection:**  
   Restricting training on datasets containing **PII** helps ensure compliance with data laws and maintains user trust.
2. **Respect for Intellectual Property:**  
   Copyrighted content should be used under **explicit licenses** or not at all unless transformed sufficiently under fair use, which is often ambiguous with AI.
3. **Transparency & Accountability:**  
   Knowing what data goes into training enables better audits, legal clarity, and the ability to **challenge harmful or biased outputs**.
4. **Mitigating Harm:**  
   Ethical AI should avoid training on content that reinforces stereotypes, misinformation, or violates privacy.

### **Nuanced Consideration:**

However, **total restriction** may **hinder progress:**

* **Publicly available** copyrighted content (e.g., news, books) may be useful for generalization and fluency.
* A middle-ground solution is using **data filtering, redaction, and differential privacy** during training.

### **Conclusion:**

Generative AI should **not be trained indiscriminately** on all available data. Legal compliance, ethical responsibility, and respect for creators and individuals must guide dataset curation. Implementing **data use policies, licensing frameworks**, and **technical safeguards** is crucial to prevent AI harms while maintaining innovation.

**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

ANS:

**False Negative Rate (FNR)**:

This measures the proportion of actual positive cases (e.g., qualified loan applicants, people who will not reoffend) that a model **incorrectly classifies as negative** (e.g., loan denied, predicted to reoffend).

FNR=False Negatives False Negatives + True Positives\text {FNR} = \frac{\text{False Negatives}}{\text{False Negatives + True Positives}}FNR=False Negatives + True Positives False Negatives​

**False Negative Rate Parity**:

These checks whether the **FNR is roughly equal across different demographic groups** (e.g., race, gender, age group). If it’s not, one group may be unfairly denied opportunities or services.

* Why it's important

ANS:

Failing this metric means that the model is disproportionately denying opportunities to one group more often than another when they actually deserve a positive outcome.

* In criminal justice: wrongly labeling someone as high-risk for reoffending.
* In lending: denying loans to qualified individuals from a minority group.
* In healthcare: not identifying patients who need urgent treatment.

Unequal FNR can worsen systemic discrimination and erode public trust in automated systems.

* How a model might fail this metric

ANS:

Let’s say you're using a model to screen job applicants. If:

* Group A (e.g., men) has an FNR of **10%**
* Group B (e.g., women) has an FNR of **40%**

It means that qualified women are being rejected 4× more often than qualified men, even though they are equally fit for the job. This failure could stem from:

* Imbalanced or biased training data
* Feature leakage related to group membership
* Historical patterns of discrimination being learned and repeated

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