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: Ruthvik Reddy Gaddam

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.

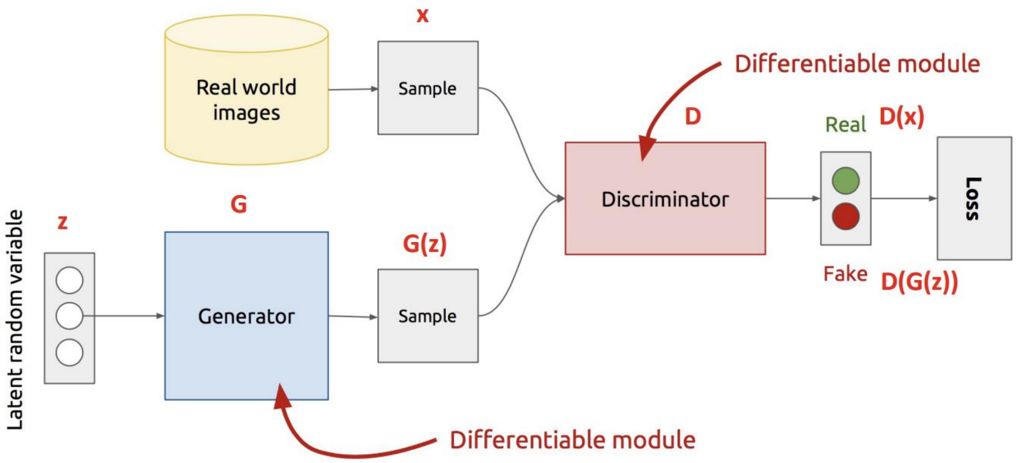
Ans. In GANs, the adversarial process is where two neural networks, the generator and discriminator, are trained simultaneously. They both have different objectives and goals from each other. The generator's goal is to generate data indistinguishable from real data. A random noise is given as input to the generator, and a fake data record, like image, text, audio, etc., is given as output. The generator's objective is to make the discriminator believe that the output is real data.

The discriminator's goal is to correctly classify inputs as real training data or generated data from the generator. The real or the generated data is given as input to the discriminator, and the data has to be classified as real or fake. The discriminator's objective is to detect the fake (generated) data.

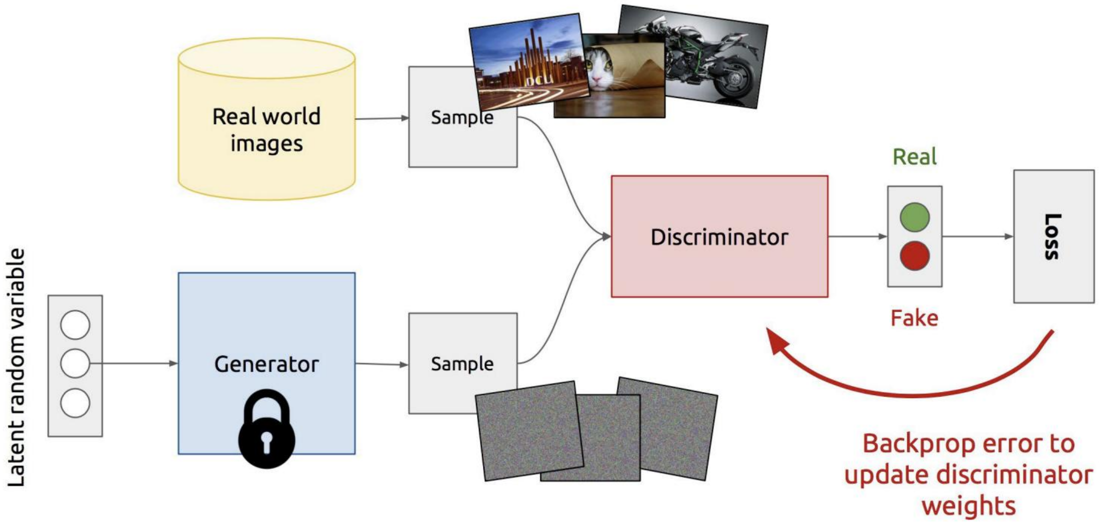
How the generator and discriminator improve though competition:

1. Initially, the discriminator easily spots the fakes.
2. As training goes on, the generator learns to make better fake data, which is more realistic.
3. The discriminator is then forced to perform more accurately, which improves its ability to detect the fakes.
4. This feedback loop creates continuous competition, which improves the generator and the discriminator.

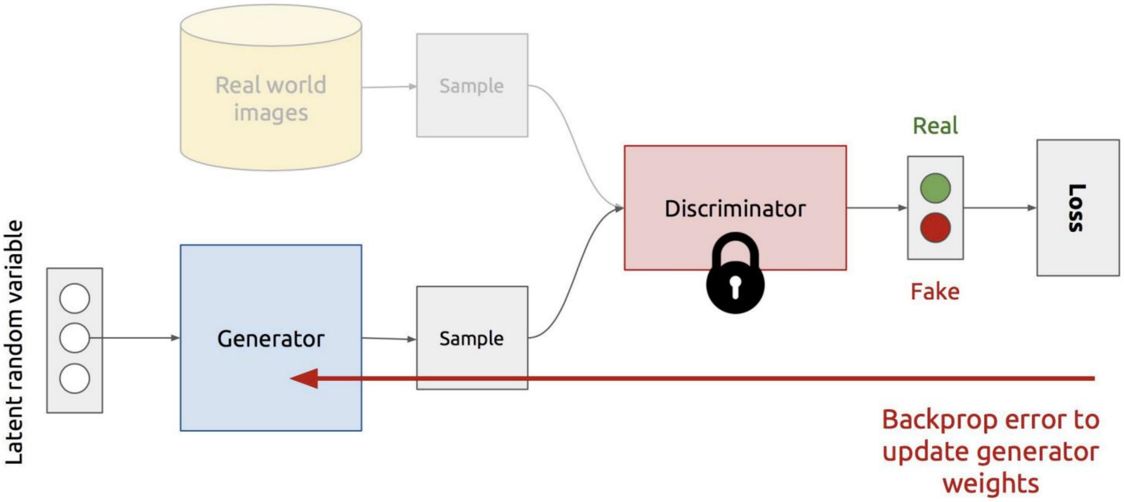
This is the Architecture of the GAN. Z is random noise(gaussian/uniform)



Discriminator training



Generator training



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. Let's take the example of an automated resume screening system used for hiring. A company uses an AI system to filter job applicants. The model is trained on historical hiring data that has past human biases like favoring candidates from certain universities or male candidates for leadership roles. Due to this, qualified candidates from less famous or prestigious schools or women with strong leadership skills may be systematically downgraded or filtered out, leading to unfair denial of job opportunities. This is allocational harm because it directly impacts access to opportunities like employment.

Mitigation strategies:

1. Check for model responses to different groups of people
2. Implement algorithms that promote fairness to every candidate
3. Train the model on various groups or demographics so it can be fair to everyone
4. Incorporate a human in the decision making loop

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)

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

Generative AI models like GPT-2 can memorize large amounts of private data from their training data. Investigations showed that GPT-2 can output strings of real names, phone numbers, and email addresses when a prompt was provided to it in specific ways. These models have also been found to generate copyrighted content like Harry Potter text, etc. Memorization raises ethical concerns and violates the privacy of people without their consent. Taking training data from blogs and websites can lead to the spread of misinformation and legal issues because the content is created by other parties.

I believe Generative AI models should have restrictions because people's privacy has to be protected, which can lead to data leaks and is unethical. Freely using the content generated by other parties is disrespectful to the creators, as it is their intellectual property. Regulating training data builds trust in the model and stops the spread of misinformation.

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.

1. False Negative Rate parity evaluates the rate at which different demographic groups receive false negatives is approximately equal.
2. It is important because it is used to see if there is fair access to opportunity. In important job sectors like healthcare, a high False Negative Rate Parity for a certain group shows that the group is unfairly being denied a job.
3. A model might fail this metric due to an imbalanced dataset, biased features, historical bias like fewer jobs done by a group in the past, etc.