Fair Decision Making

David Yang†, Selena She†, Amy Feng†, Xander Goslin†

Summer 2023

Note: † denotes equal contribution.

1 Background Reading

1.1 Generative Adversarial Networks

Definition 1 (Generator)**.** *A **generator** is a neural network that learns to generate realistic samples by transforming random noise into data samples that resemble the training data.*

Definition 2 (Discriminator)**.** *A **discriminator** is a neural network that aims to distinguish between real and fake examples.*

More formally, a generative model captures the data distribution and a discriminative model estimates the probability that a sample came from the training data rather than G. "The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency."

Definition 3 (Adversarial Nets)**.** ***Adversarial nets** refer to the specific case where the generative model generates samples by passing random noise through a multilayer perceptron, and the discriminative model is also a multilayer perceptron.*

1.2 Denoising Diffusion Probabilistic Model / Denoising Diffusion Implicit Model

Other Relevant Resource(s): DDIM vs DDPM Article

Definition 4 (Diffusion Model)**.** *A **diffusion (probabilistic) model** is a parameterized Markov chain trained using variational inference to produce samples matching the data after finite time.*

DDPMs are types of generative models that use probabilistic processes to transform the data from a simple distribution to a more complex target distribution. It iteratively refines the initial random distribution, where in each step it removes the noise from the data and eventually ends up creating a realistic sample of data. In comparison to Generative Adversarial Nets (GANs), GANs may suffer from **real mode collapse** in which the generator produces just a small variety of data that is not as diverse as real-world data. On the other hand, diffusion models will not have this problem but may take longer/tend to be more computationally expensive.

Definition 5 (DDIM)**.** *A Denoising diffusion implicit model (DDIM) is a more efficient class of iterative implicit probabilistic models with the same training procedure as DDPMs.*

The generative process of a DDPM is defined as the "reverse of a Markovian diffusion process," which approximates the reverse of the forward diffusion process (from data to noise) and is much slower than a GAN, which typically requires only one pass through a network. On the other hand, a DDIM allows for much faster sampling while keeping an equivalent training objective, so that generative models using this architecture are competitive to GANs at the same model size/sample quality." This is done by "estimating the addition of multiple Markov chain jumps by estimating the sum of Gaussian Markov jumps as Gaussian."

This generalizes the Markovian forward diffusion process of DDPMs to a non-Markovian one.

One of the key differences between the DDIM and DDPMs is the "**consistency**" property (present in DDIMs but not DDPMs): if we start with the same initial latent variable and generate several samples with Markov chains of various lengths, these samples would have similar high-level features. The consistency of DDIMs allow for meaningful image interpolation.

1.3 Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness

Fair Diffusion is a strategy that attenuates biases after the deployment of generative text-to-image models. It shifts a bias, based on human instructions, in any direction yielding arbitrary proportions (e.g. identity groups).