# Diffusion Models

## A Brief Introduction

Shounak Shirodkar

MINRO, IIIT-Bangalore

May 2022

# Table of Contents


1. Introduction
   - Background
   - GAN
   - VAE
   - Flow-based Models
2. Diffusion Models
   - Concept
   - Forward Diffusion Process
   - Reverse Diffusion Process
   - Enhancements

# Background

1. Among the variety of statistical methods, generative modeling is used to learn data distribution in an unsupervised manner.
2. Capturing the distribution behind raw data allows us to generate new data points.
3. With the success of neural nets across several domains, a new family of models called Deep Generative Models (DGM) was proposed.
4. Large-scale DGMs have been deployed in computer vision and natural language processing.

# Generative Adversarial Network

A Generative Adversarial Network (GAN) has two parts:

- The **generator** learns to generate plausible data. The generated instances become negative training examples for the discriminator.
- The **discriminator** learns to distinguish the generator's fake data from real data. The discriminator penalizes the generator for producing implausible results.

The GAN is based upon the game theoretic concept of actor-critic models. However, the GAN has limitations: difficulty in achieving Nash equilibrium, instability due to potentially low-dimensional supports, and the vanishing gradient problem.

# Variational Autoencoder

The Variational Autoencoder (VAE) is an ANN architecture belonging to the family of probabilistic graphical models and variational Bayes methods.

The basic scheme is such that the model is fed samples from a particular distribution, which is compressed into the latent space. The decoder receives samples from this latent space and attempts to reconstruct the original input by minimizing the objective, typically the KL divergence.

Like most other competitive deep learning models, VAE follows the encoder-decoder architecture.

# Flow-based Models

Flow-based models explicitly model the respective input distribution through the use of *normalizing flow*: a method for constructing complex distributions by transforming a probability density through a series of invertible mappings. By repeatedly applying the rule for change of variables, the initial density 'flows' through the sequence of invertible mappings.

# What are Diffusion Models?

- Diffusion Models (DM) progressively corrupt training data, and then learn to denoise the provided data.

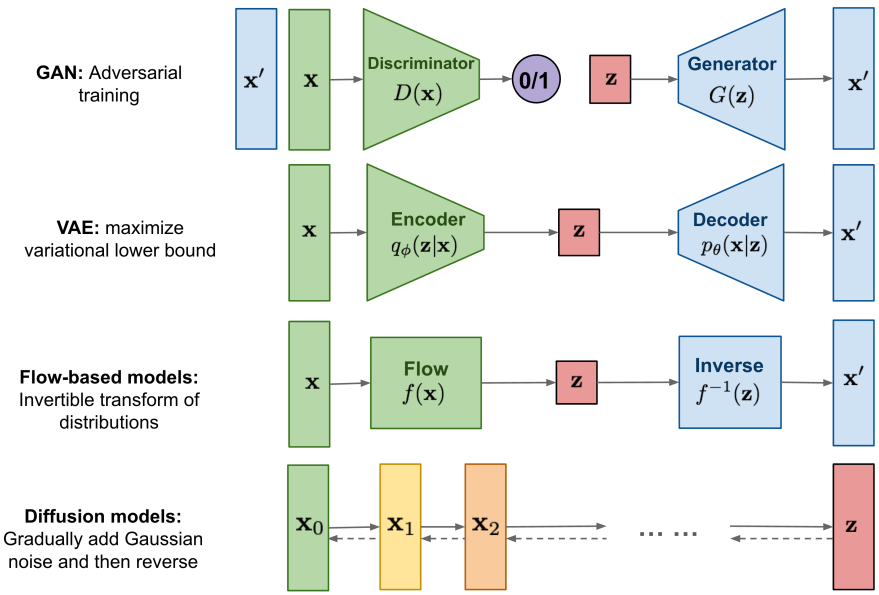
# What are Diffusion Models?

- Diffusion Models (DM) progressively corrupt training data, and then learn to denoise the provided data.
- The data corruption is achieved through progressive addition of Gaussian noise to the training data through a series of Markovian diffusion steps. The learning process attempts to reverse this forward diffusion by reverse Markovian diffusion.

# What are Diffusion Models?

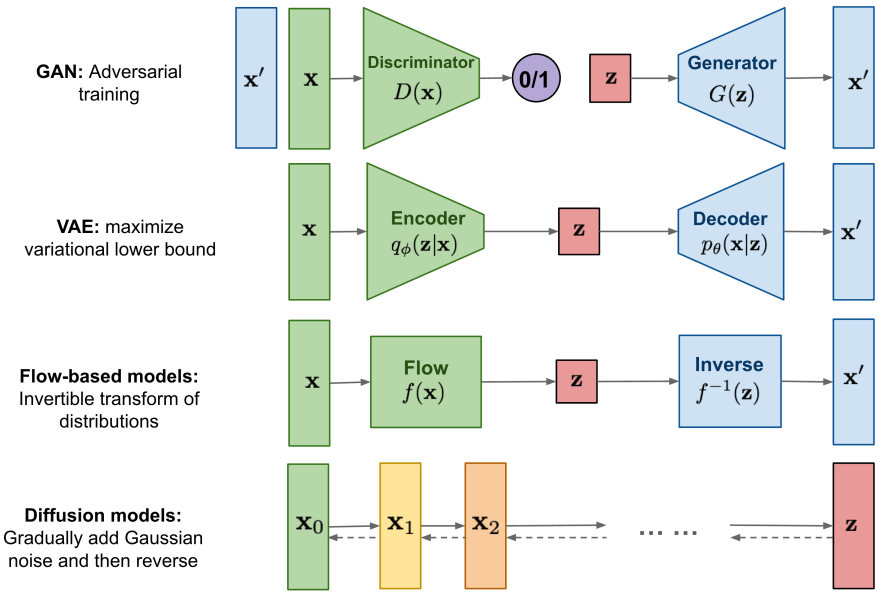
- Diffusion Models (DM) progressively corrupt training data, and then learn to denoise the provided data.
- The data corruption is achieved through progressive addition of Gaussian noise to the training data through a series of Markovian diffusion steps. The learning process attempts to reverse this forward diffusion by reverse Markovian diffusion.
- Unlike VAE or flow models, DMs are learned with a fixed procedure and the latent variable has high dimensionality (same as the original data).

# What are Diffusion Models?




1. GAN has a discriminator and a generator. The former learns to distinguish the real data from the fake samples produced by the latter. Based on zero-sum minimax.
2. VAE inexplicitly optimizes the log-likelihood of the data by maximizing the evidence lower bound (ELBO).
3. A flow-based model is a sequence of invertible transformations. The model explicitly learns the distribution; the loss function is the negative log-likelihood.

# What are Diffusion Models?




The intuition for DM follows from the idea that by reversing the forward diffusion process of injecting Gaussian disturbance into the input, one might be able to sample from the true sample before the addition of Gaussian noise.
Note the contrast of the mechanism behind DM with GAN, VAE and Flow-based models in the neighboring diagram.

# Diffusion Process

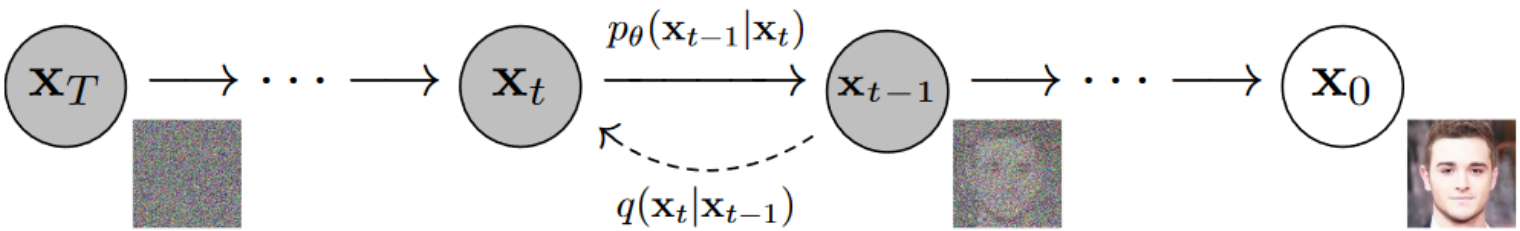



Figure: The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. Sourced from Ho et al. as is.

# Forward Diffusion Process

Let a data point be sampled from a distribution as $x_0 \sim q(x)$. Let there be a Markovian forward diffusion process which adds noise to this sample in $T$ steps, producing a sequence of noisy samples $x_1, \ldots, x_T$ such that

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}) \quad q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

controlled by a variance schedule $\beta_1, \ldots, \beta_T$.
The data sample gradually loses its distinguishable features and tends to an isotropic Gaussian distribution as the timestep $T$ becomes bigger.

# Closed-form Sampling

Forward diffusion can be reduced to a closed-form equivalent, such that a sample $x_t$ at timestep $t$ can directly be sampled without the extensive diffusion process. This is possible through a **reparameterization trick**. Let $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{i=1}^{T} \alpha_i$, then

$$
\begin{aligned}
x_t &= \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_{t-1} \quad ; z_i \sim \mathcal{N}(0, \mathbf{I}) \\
&= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}} \overline{z}_{t-2} \quad ; \overline{z}_{t-2} = z_{t-1} z_{t-2} \\
&= \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t} z \\
q(x_t|x_0) &= \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t} x_0, (1-\overline{\alpha}_t)\mathbf{I})
\end{aligned}
$$

# Reverse Diffusion Process

If forward diffusion is *reversed*, and samples are drawn from $q(x_{t-1}|x_t)$, the original sample before the addition of Gaussian noise might be obtained. It is difficult to estimate $q(x_{t-1}|x_t)$, since the entire dataset is required; instead, a model $p_\theta$ is learnt to approximate the conditional probabilities to drive the reverse process.

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

given $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$.

# Tractability

When conditioned on $x_0$, the prior ie. $q(x_{t-1}|x_t, x_0)$ is tractable, so that

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

Expanding using the Bayes rule, the mean and variance may be parameterized as

$$\tilde{\beta}_t = \frac{1 - \overline{\alpha}_{t-1}}{1 - \overline{\alpha}_t} \cdot \beta_t$$

$$\tilde{\mu}(x_t, x_0) = \frac{\sqrt{\alpha_t}(1 - \overline{\alpha}_{t-1})}{1 - \overline{\alpha}_t} x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1 - \overline{\alpha}_t} x_0$$

# Tractability

By using the earlier *reparameterization trick* and substituting $x_0$,

$$\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}} z_t \right)$$

which is similar to parameters in VAE. Thus, a variational lower bound might be used to optimize the negative log likelihood.

$$\begin{aligned} -\log p_\theta(x_0) &\leq -\log p_\theta(x_0) + D_{KL}(q(x_{1:T}|x_0)) \\ &= -\log p_\theta(x_0) + \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)} \left[ \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})/p_\theta(x_0)} \right] \\ &= \mathbb{E}_q \left[ \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})} \right] = L_{VLB} \end{aligned}$$

# Tractability

We have that the cross-entropy loss $L_{CE} \leq L_{VLB}$. Sohl-Dickstein et al. detail a way to rewrite all terms in the equation to be analytically computable, so that the objective is a sum of KL-divergences and entropy terms.
In brief,

$$L_{VLB} = L_T + L_{T-1} + \cdots + L_0$$

where $L_T = D_{KL}(q(x_T|x_0)||p_\theta(x_T))$,
$L_t = D_{KL}(q(x_t|x_{t+1}, x_0)||p_\theta(x_t|x_{t+1}))$ for $t \leq T-1$ and
$L_0 = -\log p_\theta(x_0|x_1)$.
All KL divergences in $L_{VLB}$ except initial term $L_0$ compare two Gaussians and can thus be computed in closed-form. $L_T$ is constant and can be ignored during training, as $q$ has no learnable parameters and $x_T$ is simply Gaussian noise.

# Enhancements

Nichol & Dhariwal suggest several enhancements to the vanilla DM for better performance on real world data.

- Improvements to the parameterization system for $\beta_t$ have been proposed.
- Reverse process variance was optimized with better parameterization.
- Markovian sampling being slow, especially given the larger datasets of today, speed-up diffusion sampling has been suggested. Higher quality samples are thus made possible with much fewer steps.
- Since DM primarily operates in the unsupervised realm, with some modifications, it was shown that DM proved better performing than GAN on class-conditioned data.