A Very Short Introduction to Diffusion Model

January 5, 2023

Preface

This blog gives a very high level diffusion model which only requires basic knowledge on terms including stochastic differential equation (SDE), Brownian motion, and evidence lower bound (ELBO). The content can be beneficial to readers who still struggle to fully understand the model after reading 5 papers.

Disclaimer: the notations are not rigorous and are used for illustration purpose only.

Basically, diffusion models have a forward process that gradually adds Brownian noise to source input, x_0, until it becomes pure noise, $x_T \sim \mathcal{N}(0,1)$, and a backward (a.k.a. inference) process $p_\theta(x_{t-1}|x_t)$ to gradually generate new inputs from pure noise. This backward process is parameterized by a neural network $p_\theta(x_t, t)$, and is learned by maximizing the ELBO that approximates the true data distribution of x_0.

Feel free to skip the paragraph above since I assume you have read similar ones everywhere.

Forward Process

The forward process can be regarded as a stochastic differential equation:

$$\begin{cases} dX_t &= f(t)dtX_t + g(t)dB_t \quad \text{(continuous)} \\ X_{t+1} &= f(t)dtX_t + g(t)dB_t \quad \text{(discrete)} \end{cases} \tag{1}$$

where dB_t is the standard Brownian motion. With some derivations it can be showed that

$$\begin{cases} x_t \leftarrow q(x_t|x_0) \sim \mathcal{N}(\cdot,\cdot) \\ q(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\hat{\alpha}x_t + \hat{\beta}x_0, \hat{\sigma}) \end{cases} \tag{2}$$

Generative Model (Training)

The goal for generative models is to learn (maximize) $P(x_0)$, which is the distribution of the data. For diffusion models,

$$P(x_0) \;\sim\; \text{ELBO} \;\sim\; \text{KL}\Big(q(x_{t-1}|x_t, x_0)\,||\,p_\theta(x_{t-1}|x_t)\Big)$$

where KL represents KL divergence. $p_\theta(x_{t-1}|x_t)$ is a function of $\{x_{t-1}, t\}$ and parameterized by θ. Also, $p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$, so that the KL term is reduced to mean matching, i.e.,

$$\min_\theta \;||p_\theta(x_t, t) - x_0||_2$$

and the parameter θ is learned accordingly.

Backward Process (Inference)

Given (every) forward SDE has a corresponding backward SDE (BSDE), we can use this BSDE to generate new samples iteratively starting from pure noise $\sim \mathcal{N}(0, 1)$. This implicitly requires $T = 1$ for the forward process. Importantly, the inference can be done in stochastic as well as deterministic format.

I will add more details on this aspect.