Note on Diffusion Model

Tianbing Xu

April 26, 2024

1 Idea

The Diffusion Probabilistic Model is a parameterized Markov Chain designed for generating high-quality data samples from noise through a sequence of invertible operations. In the forward process, random noise is incrementally added to the data, moving it closer to a standard Gaussian distribution. Conversely, in the reverse denoising process, data is generated from noise by acquiring knowledge of the reverse transition probability, aligning with the variational posterior distribution induced by the forward transition distribution.

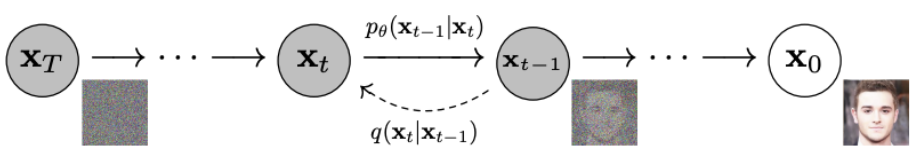
Learning

$$L_t(\theta) = E_{x_0 \sim D, \epsilon \sim N(0,I)} \left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2 \right] \tag{1}$$

Sampling

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z \tag{2}$$

where $z \sim N(0, I)$



2 Introduction

2.1 Forward Diffusion Process

Given a Markov Chain : $x_0 \to x_1 ... \to x_{t-1} \to x_t ... \to x_T$

$$\begin{aligned} x_0 &\sim q(x) \\ q(x_t|x_{t-1}) &= \mathcal{N}(\sqrt{1-\beta_t}x_t, \beta_t I) \\ x_T &\sim \mathcal{N}(0, I), T \to \infty \end{aligned}$$

with fixed schedule variance $\beta_t \in (0,1)$ and $\beta_{t-1} < \beta_t$. We add noise gradually into the process and eventually $x_T \to N(0, I)$.

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{3}$$

Let $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, we have

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{4}$$

where $\epsilon \sim N(0, I)$ and same for ϵ_{t-1}.

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)I)$$

2.2 Reverse Diffusion Process

Given the reverse sequence: $x_T ... \to x_t \to x_{t-1} ... \to x_0$

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{5}$$

with reverse transition distribution parameterized by a **learnable Neureal Network**

$$p_\theta(x_{t-1}|x_t) = N(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))) \tag{6}$$

Note that the posterior reverse transition distribution conditioned on x_0 is tractable from Bayesian rule

$$q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0) q(x_{t-1}|x_0)/q(x_t|x_0) \tag{7}$$

Then the posterior Gaussian mean $\hat{\mu}_t(x_t, t) = E[q(x_{t-1}|x_t, x_0)]$ is

$$\hat{\mu}_t(x_t, t) = \frac{x_t - \beta_t \epsilon_t / \sqrt{1-\bar{\alpha}_t}}{\sqrt{\alpha_t}} \tag{8}$$

2.3 Learning Diffusion Process with Variational Distribution Match

Variational Lower Bound

Diffusion models are latent variable models of the form

$$p_\theta(x_0) = \int p_\theta(x_{0:T})\, dx_{1:T} \tag{9}$$

The cross entropy loss (negative log-likelihood) is upper bounded by

$$L(\theta) = E_{q(x_{0:T})}\left[\frac{\log q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right] \geq -E_{q(x_0)}\left[\log p_\theta(x_0)\right] \tag{10}$$

This bound is able to be decomposed as follow

$$L(\theta) = E_q\left[D_{KL}(q(x_T|x_0)||p_\theta(x_T) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1)\right] \tag{11}$$

To minimize the variational bound, we can match $p_\theta(x_{t-1}|x_t)$ to the variatonal posterior distribution $q(x_{t-1}|x_t, x_0)$. For mean network with simplicity settings, the loss is

$$L_t(\theta) = E_{x_0 \sim D, \epsilon \sim N(0,I)}\left[||\hat{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)||^2\right] \tag{12}$$

. We may learn the noise network as well based on Eq.(8)

$$L_t(\theta) = E_{x_0 \sim D, \epsilon \sim N(0,I)}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right]$$

References

[1] Jonathan Ho, Ajay Jain, Pieter Abbeel *Denoising Diffusion Probabilistic Models, NIPS 2020*