6.7900 Fall 2024: Lecture Notes 22

We are going to talk about one of the most important generative models today — diffusion models. We will put more emphasis on intuition and insights.

1 Idea and Intuition

The idea of diffusion model is to build complex objects ($Q_D(x)$) through simple samples ($P(z)$) via hierarchical stacks. Note that

$$\int Q_D(x)Q(z|x,\phi)dx \approx P(z)$$
$$\int P(z)P(x|z,\theta)dz \approx Q_D(x)$$

In VAEs, we hope to jointly learn the encoder $Q(z|x,\phi)$ and the decoder $P(x|z,\theta)$ with $P(z)$ and $Q_D(x)$ fixed. In diffusion models, We further fix the encoder $Q(z|x)$ and only learn the decoder $P(x|z,\theta)$. Note that the fixed encoder still need to be carefully chosen (which is the art of diffusion models) to satisfy the above first approximated equality to ensure consistency.

To be more concrete, the encoder follows a simple fixed structure where noise is added in each step as a forward process, while the decoder de-noises gradually as a reverse process. When adding noise, intuitively we are washing out signals first from high frequency then to low frequency. The de-noising steps uncover those lost signals. Once we have a de-noising model, we can draw a noisy sample and iterate reversely to generate a new sample. The question essentially lies in how to de-noise — there can be multiple choices and directions to reverse, and multiple inputs could have similar noisy forward outputs.

2 Forward Process

Let $q(x_t|x_{t-1})$ follow a Gaussian distribution such that

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

Through the property of Gaussian, we know that

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \text{ for some } \epsilon \sim \mathcal{N}(0, I)$$

where $\bar{\alpha}_t = \Pi_{s=1}^{t}(1-\beta_s)$. As $T \to +\infty$, x_T becomes Gaussian. The latent class z of x_0 becomes a trajectory $x_{1:T}$. Thus, diffusion models are generally over-parameterized models that can overfit on small datasets but can perform quite well on large (structured) datasets such as images.

3 Reverse Process

We want to further *approximate* $p(x_{0:T})$ as

$$\begin{aligned} p(x_{0:T}) &= p(x_T)\Pi_{t=1}^{T} p(x_{t-1}|x_t) \\ &\approx \mathcal{N}(0,I)\Pi_{t=1}^{T}\mathcal{N}(x_{t-1}|\mu_\theta(x_t,t),\Sigma_\theta^2(x_t,t)) \end{aligned}$$

under the assumption that (i) step-size β_t is sufficiently small, and (ii) T is large enough such that $p(x_T) \sim \mathcal{N}(0,I)$. In fact, with the two assumptions, $\Sigma_\theta^2(x_t,t)$ can be approximated as $\beta_t I$. It suffices to learn the de-noising "vector field" $\mu_\theta(x_t,t)$. In principle, this can be done by maximizing ELBO:

$$\log p_\theta(x_0) \geq \mathbb{E}_q\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right].$$

Keep in mind that the latent structure is characterized via $x_{1:T}$

In the following we provide a different perspective of the argument above that may bring improved insights. Note that

$$q(x_{1:T}|x_0) = q(x_T|x_0)\Pi_{t=2}^{T} q(x_{t-1}|x_t,x_0).$$

Since

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon)$$

and

$$q(x_{t-1}|x_t,x_0) = \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)} \qquad \text{which is Gaussian,}$$

we can get

$$q(x_{t-1}|x_t,x_0) \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t I \approx \beta_t I\right).$$

Essentially we should not use x_0, but the derivation shows that we can in fact approximate $\mu_\theta(x_t,t)$ by *predicting* the (marginal) noise ϵ as $\epsilon_\theta(x_t,t)$. Predicting the noise may be better since the noise is always in the same scale and thus may make the training easier and more stable.

See Slides pp.34 for detailed algorithms for training and sampling.

Continuous diffusion models very shortly covered in class. See Slides pp.35-46 for details.