扩散模型原理详解

日常半躺

2022 年 10 月 11 日

本文对扩散模型（Diffusion Model）做了详细的原理讲解。

1 ELBO 复习

如果你对 ELBO 还不了解，可以参考我之前写的 VAE 教程：

https://github.com/ml-researcher/VAE

回顾一下 VAE 中的 ELBO 的推导:

$$
\begin{aligned}
\log p(x) &= \log \int_{\mathcal{Z}} p(x|z)p(z)dz && (1)\\
&= \log \int_{\mathcal{Z}} q(z)\frac{p(x|z)p(z)}{q(z)}dz && (2)\\
&= \log E_{z\sim q(z)}[\frac{p(x|z)p(z)}{q(z)}] && (3)\\
&\geq E_{z\sim q(z)}[\log \frac{p(x|z)p(z)}{q(z)}] && \text{(Jensen inequity)}\ (4)\\
&= E_{z\sim q(z)}[\log p(x|z)] - KL(q(z)\|p(z)) && (5)
\end{aligned}
$$

2 扩散模型公式推导

扩散模型其实也利用了像 ELBO 一样的推导过程，只不过把 z 换成了 $x_{1:T}$，也就是说隐变量不再是单一的 z，而是 T 步的随机变量 $x_1, x_2, \ldots, x_T$。

$$\log p(x) = \log \int_{\mathcal{X}_{1:T}} p(x|x_{1:T})p(x_{1:T})dx_{1:T} \tag{6}$$

$$= \log \int_{\mathcal{X}_{1:T}} q(x_{1:T})\frac{p(x|x_{1:T})p(x_{1:T})}{q(x_{1:T})}dx_{1:T} \tag{7}$$

$$= \log E_{x_{1:T}\sim q(x_{1:T})}[\frac{p(x|x_{1:T})p(x_{1:T})}{q(x_{1:T})}] \tag{8}$$

$$\geq E_{x_{1:T}\sim q(x_{1:T})}[\log \frac{p(x|x_{1:T})p(x_{1:T})}{q(x_{1:T})}] \qquad \text{(Jensen inequity)} \tag{9}$$

上面的推导和 VAE 的推导完全一致，只是换了一下隐变量。为了和论文[1]里保持一致，我们把 x 替换为 x_0，把 $q(x_{1:T})$ 替换成 $q(x_{1:T}|x_0)$。公式如下：

$$\log p(x_0) \geq E_{x_{1:T}\sim q(x_{1:T})}[\log \frac{p(x_0|x_{1:T})p(x_{1:T})}{q(x_{1:T}|x_0)}] \tag{10}$$

为了继续推导下去，我们需要做出一些强假设——马尔可夫假设和正态分布假设：

- 假设 1：$p(\cdot)$ 和 $q(\cdot)$ 都具有马尔可夫性，也就是说对于 $p(\cdot)$ 来说，x_t 只与 x_{t+1} 有关，对于 $q(\cdot)$ 来说，x_t 只与 x_{t-1} 有关。写成表达式是这样：

$$p(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p(x_{t-1}|x_t) \tag{11}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{12}$$

- 假设 2：$p(x_{t-1}|x_t)$ 和 $q(x_t|x_{t-1})$ 都是正态分布。此外，$q(x_t|x_{t-1})$ 还满足：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{13}$$

其中，β_t 是超参数。这也是扩散模型和 VAE 的不同之处，VAE 里面 $q(\cdot)$ 是用神经网络学习的，而扩散模型里 $q(\cdot)$ 是人工指定的确定的形

[1] 《Denoising Diffusion Probabilistic Models》

式。值得注意的是，假设 2 满足的前提是 β_t 足够小，因此在选择超参的时候注意把 β_t 调小一点。

基于上面两个假设，我们继续推导下去：

$$
\begin{aligned}
\log p(x_0) &\geq E_{x_{1:T}\sim q(x_{1:T})}[\log \frac{p(x_0|x_{1:T})p(x_{1:T})}{q(x_{1:T}|x_0)}] && (14)\\
&= E_{x_{1:T}\sim q(x_{1:T})}[\log \frac{p(x_{0:T})}{q(x_{1:T}|x_0)}] && (15)\\
&= E_{x_{1:T}\sim q(x_{1:T})}[\log \frac{p(x_T)\prod_{t=1}^{T} p(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}] \qquad \text{（代入假设 1 的公式）} && (16)\\
&= E_{x_{1:T}\sim q(x_{1:T})}[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})}] && (17)
\end{aligned}
$$

基于假设 1 也就只能推到这里了，我们再看一下怎么利用假设 2。
根据假设 2，令 $\alpha_t = 1 - \beta_t$，我们可以得到：

$$
\begin{aligned}
x_t &= \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t && (18)\\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-1}) + \sqrt{1-\alpha_t}\epsilon_t && (19)\\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-1} + \sqrt{1-\alpha_t}\epsilon_t && (20)\\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1}) + (1-\alpha_t)}\epsilon'_{t-1} && (21)\\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon'_{t-1} && (22)\\
&= \cdots && (23)\\
&= \sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_1}x_0 + \sqrt{1-\alpha_t\alpha_{t-1}\cdots\alpha_1}\epsilon && (24)
\end{aligned}
$$

其中 $\epsilon \sim \mathcal{N}(0, I)$。令 $\prod_{i=1}^{t}\alpha_i = \tilde{\alpha}_t$，则有：

$$
q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\tilde{\alpha}_t}x_0, (1-\tilde{\alpha}_t)I) \qquad (25)
$$

可以看到，用 x_0 直接采样 x_t 的形式也非常简洁。

根据 $q(x_t|x_{t-1})q(x_{t-1}|x_0) = q(x_{t-1}, x_t|x_0) = q(x_{t-1}|x_t, x_0)q(x_t|x_0)$，我们可以得到：

$$q(x_t|x_{t-1}) = q(x_{t-1}|x_t, x_0) \cdot \frac{q(x_t|x_0)}{q(x_{t-1}|x_0)} \tag{26}$$

把这个式子代入公式 17，可以得到：

$$\log p(x_0) \geq E_{x_{1:T}\sim q(x_{1:T})}[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})}] \tag{27}$$

$$= E_{x_{1:T}\sim q(x_{1:T})}[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p(x_0|x_1)}{q(x_1|x_0)}] \tag{28}$$

$$= E_{x_{1:T}\sim q(x_{1:T})}[\log p(x_T) + \sum_{t=2}^{T} \log \left(\frac{p(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \cdot \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}\right) + \log \frac{p(x_0|x_1)}{q(x_1|x_0)}] \tag{29}$$

$$= E_{x_{1:T}\sim q(x_{1:T})}[\log \frac{p(x_T)}{q(x_T|x_0)} + \sum_{t=2}^{T} \log \frac{p(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \log p(x_0|x_1)] \tag{30}$$

$$= -KL(q(x_T|x_0)\|p(x_T)) - \sum_{t=2}^{T} KL(q(x_{t-1}|x_t, x_0)\|p(x_{t-1}|x_t)) + E_{x_{1:T}\sim q(x_{1:T})}[\log p(x_0|x_1)] \tag{31}$$

根据公式 26，可以得到 $q(x_{t-1}|x_t, x_0)$ 的表达式（推导过程太复杂了，先记住结论好了）：

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \frac{\sqrt{\tilde{\alpha}_{t-1}}\beta_t}{1-\tilde{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}x_t, \frac{1-\tilde{\alpha}_{t-1}}{1-\tilde{\alpha}_t}\beta_t I) \tag{32}$$

为了继续下面的推导，我们先来回顾一下两个正态分布的 KL 散度怎么计算：

$$KL(p\|q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{33}$$

到此为止，我们已经拥有了所有的条件和背景知识，已经可以推导最终的表达式了。现在我们优化的目标是公式 31，我们来分别看一下公式 31 里面每一项分别是什么。

第一项， $KL(q(x_T|x_0)\|p(x_T))$。

由公式 25 可知:

$$q(x_T|x_0) = \mathcal{N}(x_T; \sqrt{\tilde{\alpha}_T}x_0, (1-\tilde{\alpha}_T)I) \tag{34}$$

我们假设 $p(x_T) = \mathcal{N}(x_T; 0, I)$，这个条件相当于 VAE 里 z 服从高斯分布的先验。于是，第一项是一个常数（不需要优化）:

$$KL(q(x_T|x_0)\|p(x_T)) = \log\frac{1}{\sqrt{1-\tilde{\alpha}_T}} + \frac{(1-\tilde{\alpha}_T) + (\sqrt{\tilde{\alpha}_T}x_0 - 0)^2}{2} - \frac{1}{2} \tag{35}$$

$$= \frac{-\log(1-\tilde{\alpha}_T) - \tilde{\alpha}_T + \tilde{\alpha}_T x_0^2}{2} \tag{36}$$

第二项， $KL(q(x_{t-1}|x_t, x_0)\|p(x_{t-1}|x_t))$。

公式 32 可以得到 $q(x_{t-1}|x_t, x_0)$ 的均值和标准差,我们假设 $p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_t, \sigma_t^2 I)$，可得两者的 KL 散度为:

$$KL(q(x_{t-1}|x_t, x_0)\|p(x_{t-1}|x_t)) = \log\frac{\sigma_t}{\sqrt{\frac{1-\tilde{\alpha}_{t-1}}{1-\tilde{\alpha}_t}\beta_t}} + \frac{\frac{1-\tilde{\alpha}_{t-1}}{1-\tilde{\alpha}_t}\beta_t + (\frac{\sqrt{\tilde{\alpha}_{t-1}}\beta_t}{1-\tilde{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}x_t - \mu_t)^2}{2\sigma_t^2} - \frac{1}{2} \tag{37}$$

我们假设 σ_t 是确定的常数（类似于 β_t），那么上面的式子需要优化的部分只有

$$\frac{(\frac{\sqrt{\tilde{\alpha}_{t-1}}\beta_t}{1-\tilde{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}x_t - \mu_t)^2}{2\sigma_t^2} \tag{38}$$

也就是说让 μ_t 越接近 $\frac{\sqrt{\tilde{\alpha}_{t-1}}\beta_t}{1-\tilde{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}x_t$ 越好。

这里我们再梳理一下逻辑：我们训练一个神经网络来预测 μ_t，我们只需要采样 t，让神经网络的输出尽量接近 $\frac{\sqrt{\tilde{\alpha}_{t-1}}\beta_t}{1-\tilde{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}x_t$，训练出来的神经网络就具有去噪效果。(神奇的逻辑！数学的魅力!)

实际上，我们还可以进一步展开。由公式 24 可得:

$$x_0 = \frac{1}{\sqrt{\tilde{\alpha}_t}}(x_t - \sqrt{1-\tilde{\alpha}_t}\epsilon) \tag{39}$$

代入公式 38，可得优化目标变成了：

$$
\begin{aligned}
&\frac{(\frac{\sqrt{\tilde{\alpha}_{t-1}}\beta_t}{1-\tilde{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}x_t - \mu_t)^2}{2\sigma_t^2} && (40)\\
&= \frac{(\frac{\sqrt{\tilde{\alpha}_{t-1}}\beta_t}{1-\tilde{\alpha}_t}\frac{1}{\sqrt{\tilde{\alpha}_t}}(x_t - \sqrt{1-\tilde{\alpha}_t}\epsilon) + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}x_t - \mu_t)^2}{2\sigma_t^2} && (41)\\
&= \frac{(\frac{\beta_t}{1-\tilde{\alpha}_t}\frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\tilde{\alpha}_t}\epsilon) + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}x_t - \mu_t)^2}{2\sigma_t^2} && (42)\\
&= \frac{(\frac{1}{(1-\tilde{\alpha}_t)\sqrt{\alpha_t}}(\beta_t x_t + \alpha_t(1-\tilde{\alpha}_{t-1})x_t - \beta_t\sqrt{1-\tilde{\alpha}_t}\epsilon) - \mu_t)^2}{2\sigma_t^2} && (43)\\
&= \frac{(\frac{1}{(1-\tilde{\alpha}_t)\sqrt{\alpha_t}}((1-\tilde{\alpha}_t)x_t - \beta_t\sqrt{1-\tilde{\alpha}_t}\epsilon) - \mu_t)^2}{2\sigma_t^2} && (44)\\
&= \frac{(\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon) - \mu_t)^2}{2\sigma_t^2} && (45)
\end{aligned}
$$

现在我们的目标变成了让 μ_t 尽量接近 $\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon)$。我们发现，对于 μ_t 来说，x_t 是给定的（因为 μ_t 的任务就是给定 x_t 预测 x_{t-1}），所以并不用让神经网络预测 x_t 的部分，只需要让神经网络预测 ϵ 就可以了！

具体来说，令：

$$\mu_t = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_t) \tag{46}$$

代入公式 45，就得到了优化目标变成了：

$$\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\tilde{\alpha}_t)}(\epsilon - \epsilon_t)^2 \tag{47}$$

现在我们的逻辑变成了：给定 x_t，神经网络会预测 ϵ_t，代入公式 46，就可以得到 μ_t，进而可以可以根据 $p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_t, \sigma_t^2 I)$ 采样 x_{t-1}。

第三项， $E_{x_{1:T}\sim q(x_{1:T})}[\log p(x_0|x_1)]$。

前两项 KL 散度前面有负号，所以都是希望 KL 越小越好。第三项没有负号，我们期望越大越好。

第三项我们实际上可以用一个单独的 decoder 来做，这样灵活性更强。但是 DDPM 论文里的做法是让这个 decoder 和第二项里拟合 ϵ_t 的网络共享了，也就是说：

$$p(x_0|x_1) = \mathcal{N}(x_0; \mu_1, \sigma_1^2 I) \tag{48}$$

其中 $\mu_1 = \frac{1}{\sqrt{\alpha_1}}(x_1 - \frac{\beta_1}{\sqrt{1-\alpha_1}}\epsilon_1)$。因此：

$$\log p(x_0|x_1) = \log \frac{1}{\sqrt{2\pi}\sigma_1} \exp(-\frac{(x_0-\mu_1)^2}{2\sigma_1^2}) \tag{49}$$

$$= -\frac{(x_0-\mu_1)^2}{2\sigma_1^2} + C \tag{50}$$

其中 $x_0 = \frac{1}{\sqrt{\alpha_1}}(x_1 - \sqrt{1-\alpha_1}\epsilon) = \frac{1}{\sqrt{\alpha_1}}(x_1 - \frac{\beta_1}{\sqrt{1-\alpha_1}}\epsilon)$，代入 μ_1 和 x_0 可得，我们的优化目标是：

$$-\frac{(x_0-\mu_1)^2}{2\sigma_1^2} \tag{51}$$

$$= -\frac{\left(\frac{1}{\sqrt{\alpha_1}}(x_1 - \frac{\beta_1}{\sqrt{1-\alpha_1}}\epsilon) - \frac{1}{\sqrt{\alpha_1}}(x_1 - \frac{\beta_1}{\sqrt{1-\alpha_1}}\epsilon_1)\right)^2}{2\sigma_1^2} \tag{52}$$

$$= -\frac{\beta_1^2}{2\sigma_1^2\alpha_1(1-\alpha_1)}(\epsilon - \epsilon_1)^2 \tag{53}$$

形式上与第二项公式 47 一模一样!

到此，综合上面三项，我们得到了最终的优化目标，就是最小化如下表达式：

$$\sum_{t=1}^{T} \frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\tilde{\alpha}_t)}(\epsilon - \epsilon_t)^2 \tag{54}$$

3 训练过程

采样一个图片 x_0，采样一个时间 t，从高斯分布采样 ϵ，利用公式 $x_t = \sqrt{\tilde{\alpha}_t}x_0 + \sqrt{1-\tilde{\alpha}_t}\epsilon$ 得到 x_t。将 x_t 输入神经网络得到预测输出 ϵ_t，计算 $\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\tilde{\alpha}_t)}(\epsilon - \epsilon_t)^2$，反向传播。

4 预测过程

从高斯分布采样 x_T,传入神经网络得到预测 ϵ_T,通过公式 $\mu_t = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_t)$ 得到 x_{T-1} 的分布 $\mathcal{N}(\mu_T, \sigma_T^2 I)$，从这个分布里采样 x_{T-1}，以此类

推，直到 x_0。要注意的是，最后采样 x_0 的时候直接输出均值，不需要再加高斯噪声。