淡蓝小点直播系列：DDPM原理推导及代码实现

淡蓝小点Bluedotdot

微信：**bluedotdot_cn**

2024 年 4 月 3 日

目录

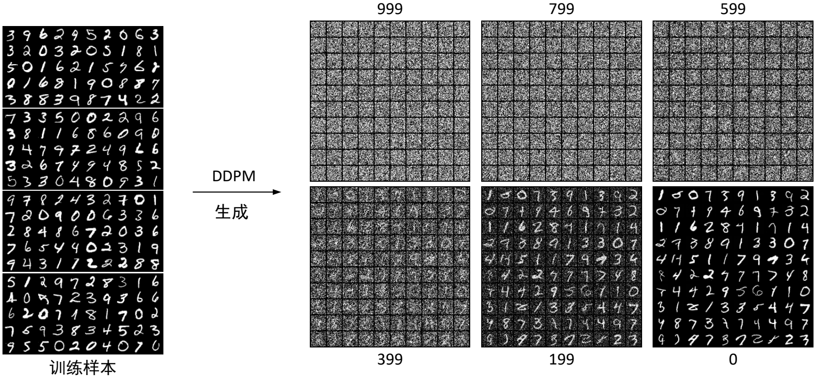
DDPM背景介绍

- DDPM全称为去噪声扩散概率模型（Denoising Diffusion Probabilistic Models），它是GAN之后最优秀的生成模型之一。扩散模型的灵感起源于非平衡态热力学（arXiv:1503.03585），它的第一篇有重要影响力的论文是2020年发表的（arXiv:2006.11239）。
- DDPM最典型的应用是图像生成，如著名的Midjourney、DALL-E等都是基于DDPM的。
- 扩散模型是一个庞大、复杂的算法家族，它被广泛的应用于各个领域如计算机视觉、自然语言处理、时序数据建模等。从技术方向上看，扩散模型主要包括三类：DDPM、SGM（Score-Based Generative Model）、SDE（Stochastic Differential Equations）。因此，DDPM只是扩散模型家族中的重要一类。
- DDPM包括三个主要步骤：前向过程（forward process，或者也称为扩散过程diffusion process）、反向过程（reverse process，或者也称为denoising process）以及采样过程（sampling procedure）。其中前向过程可设置为确定的，一般仅在训练阶段需要；反向过程是图像生成过程；采样过程是参数训练的重点。

DDPM背景介绍

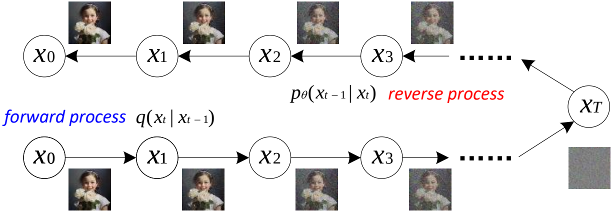
当前几乎所有著名的图像/视频生成式AI应用都是基于Diffusion模型或其变种的，如Midjourney、DALL-E、Sora等。

本次介绍的最后会基于pytorch实现一个最基本的DDPM程序，主干网络为4个注意力头的Unet，总参数量约1000万即0.01B。以六万张MNIST图像为训练数据，训练64个epoch，在RTX3050上约需训练3小时左右，在RTX4090上约需训练30分钟。



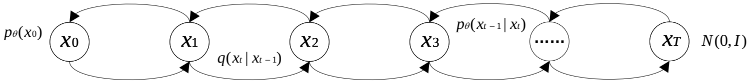
DDPM的**前向过程**是将一幅正常的图像逐渐加噪声使其成为纯噪声，**反向过程**则是从纯噪声开始逐渐去噪声使其成为一幅有意义的图像。

- 前向过程中每次基于$\mathbf{x}_{t-1}$得到$\mathbf{x}_t$的分布$q(\mathbf{x}_t|\mathbf{x}_{t-1})$，再从分布中采样得到$t$时刻数据$\mathbf{x}_t$（反向过程亦然）
- 前向加噪声过程是确定的，后向过程的噪声由神经网络计算输出（该网络的参数就是学习的重点）
- 前后向扩散过程都遵从一阶马尔科夫链，扩散可持续几百步上千步



模型推导

$\mathbf{x}_0$是训练数据，我们期望经反向扩散后恢复得到$\mathbf{x}_0$的可能性是最大的即最大化$\ln p_\theta(\mathbf{x}_0)$。通过对$\ln p_\theta(\mathbf{x}_0)$的变形，得到优化的目标函数。



$$
\begin{aligned}
\ln p_\theta(\mathbf{x}_0) &= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\ln p_\theta(\mathbf{x}_0)\mathrm{d}\mathbf{x}_{1:T} \\
&= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\ln \frac{p_\theta(\mathbf{x}_0)p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}\mathrm{d}\mathbf{x}_{1:T} \\
&= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\ln \frac{p_\theta(\mathbf{x}_{0:T})}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}\mathrm{d}\mathbf{x}_{1:T} \\
&= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\ln \frac{p_\theta(\mathbf{x}_{0:T})}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\mathrm{d}\mathbf{x}_{1:T} \\
&= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}\mathrm{d}\mathbf{x}_{1:T} \\
&= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\mathrm{d}\mathbf{x}_{1:T} + \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\ln \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}\mathrm{d}\mathbf{x}_{1:T} \\
&= \mathrm{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}] + \mathrm{KL}(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\|p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0))
\end{aligned}
$$

因为KL[1]散度是非负的，所以有

$$\ln p_\theta(\mathbf{x}_0) \geq \mathrm{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}]$$

又因为有

$$\ln p_\theta(\mathbf{x}_0) = \int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \ln p_\theta(\mathbf{x}_0) \mathrm{d}\mathbf{x}_{1:T} = \mathrm{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[\ln p_\theta(\mathbf{x}_0)]$$

所以有

$$\mathrm{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[\ln p_\theta(\mathbf{x}_0)] \geq \mathrm{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}]$$

右边相当于左边的变分下界[2]

[1] 有关KL散度的内容可参考PRML 1.6.1或PRML Page-by-page项目的1-066

[2] 关于变分下界可参考PRML 9.4或PRML Page-by-page项目的9-051

继续对变分下界做推导变形

$$
\begin{aligned}
\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} &= \ln \frac{p_\theta(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})} \\
&= \ln p_\theta(\mathbf{x}_T) + \ln \prod_{t=1}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \\
&= \ln p_\theta(\mathbf{x}_T) + \sum_{t\geq 1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \\
&= \ln p_\theta(\mathbf{x}_T) + \sum_{t>1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} + \ln \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}
\end{aligned}
$$

对第二项继续做变形，首先注意到分母上有$q(\mathbf{x}_t|\mathbf{x}_{t-1})$，它是前向过程的第$t$步。前向过程服从一阶马尔科夫链，所以理论上$\mathbf{x}_t$只跟$\mathbf{x}_{t-1}$有关，跟$\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{t-2}$都无关。但是因为$\mathbf{x}_1$由$\mathbf{x}_0$得到，$\mathbf{x}_2$由$\mathbf{x}_1$得到，依此类推$\mathbf{x}_t$最终是跟$\mathbf{x}_0$有关的，所以$q(\mathbf{x}_t|\mathbf{x}_{t-1})$也可以显示的写成$q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)$[3] 所以有

$$\begin{aligned}
q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) \\
&= \frac{q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0)}{q(\mathbf{x}_{t-1}, \mathbf{x}_0)} \\
&= \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q(\mathbf{x}_t, \mathbf{x}_0)}{q(\mathbf{x}_{t-1}, \mathbf{x}_0)} \\
&= \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)q(\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)q(\mathbf{x}_0)} \\
&= \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}
\end{aligned}$$

[3] 这二者本质上是一样的，只不过前者表达式只跟$\mathbf{x}_{t-1}$有关，后者经过变形后同时跟$\mathbf{x}_{t-1}, \mathbf{x}_0$有关

将变形结果代入第二项后有

$$\ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} = \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)}$$

所以变分下界有

$$\begin{aligned}
\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} &= \ln p_\theta(\mathbf{x}_T) + \sum_{t>1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} + \ln \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)} \\
&= \ln p_\theta(\mathbf{x}_T) + \sum_{t>1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)} + \ln \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)} \\
&= \ln p_\theta(\mathbf{x}_T) + \sum_{t>1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} + \sum_{t>1} \ln \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} + \ln \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}
\end{aligned}$$

对于上式第三项有

$$\sum_{t>1} \ln \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} = \{\ln q(\mathbf{x}_1|\mathbf{x}_0) + \ln q(\mathbf{x}_2|\mathbf{x}_0) + \cdots + \ln q(\mathbf{x}_{T-1}|\mathbf{x}_0)\}$$
$$- \{\ln q(\mathbf{x}_2|\mathbf{x}_0) + \ln q(\mathbf{x}_3|\mathbf{x}_0) + \cdots + \ln q(\mathbf{x}_T|\mathbf{x}_0)\}$$
$$= \ln q(\mathbf{x}_1|\mathbf{x}_0) - \ln q(\mathbf{x}_T|\mathbf{x}_0)$$

代回去之后有

$$\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} = \ln p_\theta(\mathbf{x}_T) + \sum_{t>1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} + \sum_{t>1} \ln \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} + \ln \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}$$
$$= \ln p_\theta(\mathbf{x}_T) + \sum_{t>1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} + \ln q(\mathbf{x}_1|\mathbf{x}_0) - \ln q(\mathbf{x}_T|\mathbf{x}_0) + \ln \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}$$
$$= \ln \frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} + \sum_{t>1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} + \ln p_\theta(\mathbf{x}_0|\mathbf{x}_1)$$

$$
\begin{aligned}
&\mathrm{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}] \\
&= \int \{q(\mathbf{x}_{1:T}|\mathbf{x}_0) \ln \frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} + q(\mathbf{x}_{1:T}|\mathbf{x}_0) \sum_{t>1} \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} + q(\mathbf{x}_{1:T}|\mathbf{x}_0) \ln p_\theta(\mathbf{x}_0|\mathbf{x}_1)\} \mathrm{d}\mathbf{x}_{1:T} \\
&= \int q(\mathbf{x}_T|\mathbf{x}_0) \ln \frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} \mathrm{d}\mathbf{x}_T + \sum_{t>1} \int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \mathrm{d}\mathbf{x}_{1:T} + \int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \ln p_\theta(\mathbf{x}_0|\mathbf{x}_1) \mathrm{d}\mathbf{x}_{1:T} \\
&= -\mathrm{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \| p_\theta(\mathbf{x}_T)) + \sum_{t>1} \int q(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{x}_0) \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \mathrm{d}\mathbf{x}_{t-1} \mathrm{d}\mathbf{x}_t + \int q(\mathbf{x}_1|\mathbf{x}_0) \ln p_\theta(\mathbf{x}_0|\mathbf{x}_1) \mathrm{d}\mathbf{x}_1 \\
&= -\mathrm{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \| p_\theta(\mathbf{x}_T)) + \sum_{t>1} \{\int q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \ln \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \mathrm{d}\mathbf{x}_{t-1}\} q(\mathbf{x}_t|\mathbf{x}_0) \mathrm{d}\mathbf{x}_t + \mathrm{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}[\ln p_\theta(\mathbf{x}_0|\mathbf{x}_1)] \\
&= -\mathrm{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \| p_\theta(\mathbf{x}_T)) - \sum_{t>1} \int q(\mathbf{x}_t|\mathbf{x}_0) \mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \mathrm{d}\mathbf{x}_t + \mathrm{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}[\ln p_\theta(\mathbf{x}_0|\mathbf{x}_1)] \\
&= -\underbrace{\mathrm{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \| p_\theta(\mathbf{x}_T))}_{L_T} - \sum_{t>1} \mathrm{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \underbrace{[\mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))]}_{L_{t-1}} + \mathrm{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \underbrace{[\ln p_\theta(\mathbf{x}_0|\mathbf{x}_1)]}_{L_0}
\end{aligned}
$$

- 因为$p_\theta(\mathbf{x}_T)$是纯噪声服从标准正态分布，所以可认为它是确定量，跟待估参数无关，所以L_T在优化过程中可看作常量。可以为L_0选择特殊的形式让它跟待估参数相独立，这样L_0也可以作看跟优化过程无关。所以要极大化变分下界，最重要的是处理好L_{t-1}项。
- 很明显，要极大化$\mathrm{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[\ln \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}]$就要极小化$\mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$。这相当于要求前向后向两个过程中$\mathbf{x}_t \longmapsto \mathbf{x}_{t-1}$尽量相似。
- 前面的推导过程中没有对q和p_θ所服从的分布做任何假设。

接下来为q和p_θ指定特定的分布（高斯分布）并做进一步推导。先看q，令

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathrm{N}(\mathbf{x}_t|\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

很明显，可以将$\mathbf{x}_t$写作

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_t, \quad \epsilon_t \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$$

根据递推关系式有

$$\mathbf{x}_{t-1} = \sqrt{1-\beta_{t-1}}\mathbf{x}_{t-2} + \sqrt{\beta_{t-1}}\epsilon_{t-1}, \quad \epsilon_{t-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$$

将$\mathbf{x}_{t-1}$的表达式代入到$\mathbf{x}_t$的表达式中则有

$$\begin{aligned}\mathbf{x}_t &= \sqrt{1-\beta_t}\left(\sqrt{1-\beta_{t-1}}\mathbf{x}_{t-2}+\sqrt{\beta_{t-1}}\boldsymbol{\epsilon}_{t-1}\right)+\sqrt{\beta}_t\boldsymbol{\epsilon}_t \\ &= \sqrt{1-\beta_t}\sqrt{1-\beta_{t-1}}\mathbf{x}_{t-2}+\sqrt{1-\beta_t}\sqrt{\beta_{t-1}}\boldsymbol{\epsilon}_{t-1}+\sqrt{\beta}_t\boldsymbol{\epsilon}_t\end{aligned}$$

令

$$\alpha_t = 1-\beta_t$$

代入之后有

$$\mathbf{x}_t = \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2}+\sqrt{\alpha_t(1-\alpha_{t-1})}\boldsymbol{\epsilon}_{t-1}+\sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t$$

注意ϵ_t和ϵ_{t-1}是两个独立的高斯变量，所以$\sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-1}+\sqrt{1-\alpha_t}\epsilon_t$ 就是两个高斯的卷积，可套用卷积公式求结果[4]。这个公式是：设$x\sim \mathrm{N}(x|\mu_x,\sigma_x^2)$，$y\sim \mathrm{N}(y|\mu_y,\sigma_y^2)$相互独立，令$z=x+y$则有

$$z\sim \mathrm{N}(z|\mu_x+\mu_y,\sigma_x^2+\sigma_y^2)$$

因为

$$\sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-1}\sim \mathrm{N}(\mathbf{0},\alpha_t(1-\alpha_{t-1})\mathbf{I}),\quad \sqrt{1-\alpha_t}\epsilon_t\sim \mathrm{N}(\mathbf{0},(1-\alpha_t)\mathbf{I})$$

所以

$$\sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-1}+\sqrt{1-\alpha_t}\epsilon_t\sim \mathrm{N}(\mathbf{0},[\alpha_t(1-\alpha_{t-1})+1-\alpha_t]\mathbf{I})=\mathrm{N}(\mathbf{0},(1-\alpha_t\alpha_{t-1})\mathbf{I})$$

[4]关于高斯卷积的内容可参考PRML Page-by-page项目的2-030

所以

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\boldsymbol{\epsilon}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \\ &= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})\end{aligned}$$

将递推式继续递推下去则有

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_1}\mathbf{x}_0 + \sqrt{1-\alpha_t\alpha_{t-1}\cdots\alpha_1}\boldsymbol{\epsilon} \\ &= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad (\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}), \bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i)\end{aligned}$$

这是一个重要公式即我们可以基于$\mathbf{x}_0$得到第t次加噪声后的结果$\mathbf{x}_t$ （为什么$\mathbf{x}_t$最终能演变成纯噪声？）

$$\boxed{\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}$$

很明显$\mathbf{x}_t$可以写成只跟$\mathbf{x}_0$有关的概率分布

$$q(\mathbf{x}_t|\mathbf{x}_0) \sim \mathrm{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

注意，在上面的推导中可能有人会犯这样的错误：对于$\sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-1}+\sqrt{1-\alpha_t}\epsilon_t$，因为$\epsilon_{t-1}$和$\epsilon_t$都服从标准正态分布，所以它们都可以用$\epsilon$来表示，所以

$$
\begin{aligned}
\sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-1}+\sqrt{1-\alpha_t}\epsilon_t &= \sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon+\sqrt{1-\alpha_t}\epsilon \\
&= [\sqrt{\alpha_t(1-\alpha_{t-1})}+\sqrt{1-\alpha_t}]\epsilon \\
&\sim \mathrm{N}(\mathbf{0},[\sqrt{\alpha_t(1-\alpha_{t-1})}+\sqrt{1-\alpha_t}]^2\mathbf{I})
\end{aligned}
$$

这种做法的错误之处在于：原本式子中的ϵ_t和ϵ_{t-1}是两个独立的高斯变量，统一用ϵ表示之后它们就不独立了，所以计算结果是错的。

前面已经指定了$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$服从高斯分布，并且推导了$q(\mathbf{x}_t|\mathbf{x}_0)$服从的分布（相当于也知道了$q(\mathbf{x}_{t-1}|\mathbf{x}_0)$服从的分布），所以根据条件概率公式

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}$$

只需要将右边三项的概率密度公式代进去就能得到$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$，但我们也可以用更简单的方法来求其表达式。设$p(\mathbf{x}) \sim \mathrm{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$，则它的概率密度函数指数项部分可以写成关于$\mathbf{x}$的二次项、一次项、常数项三部分：

$$p(\mathbf{x}) \sim -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + C$$

所以，只需要写出有关$\mathbf{x}_{t-1}$的二次项、一次项，就能写出$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$的均值和协方差，就相当于写出了它服从的分布。所以我们只需要关注$q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)$即可（分母项与$\mathbf{x}_{t-1}$无关）。

模型推导

$$
\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) &\propto q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0) \\
&\propto \exp\{-\frac{1}{2\beta_t}(\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_{t-1})^2\}\exp\{-\frac{1}{2(1-\bar{\alpha}_{t-1})}(\mathbf{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2\}
\end{aligned}
$$

用$Q_2(\mathbf{x}_{t-1})$表示$\mathbf{x}_{t-1}$的二次项，则有

$$
\begin{aligned}
Q_2(\mathbf{x}_{t-1}) &= -\frac{\alpha_t}{2\beta_t}\mathbf{x}_{t-1}^2-\frac{1}{2(1-\bar{\alpha}_{t-1})}\mathbf{x}_{t-1}^2 \\
&= -\frac{1}{2}(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}})\mathbf{x}_{t-1}^2
\end{aligned}
$$

用$Q_1(\mathbf{x}_{t-1})$表示$\mathbf{x}_{t-1}$的一次项，则有

$$
\begin{aligned}
Q_1(\mathbf{x}_{t-1}) &= \frac{\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1}}{\beta_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0\mathbf{x}_{t-1}}{1-\bar{\alpha}_{t-1}} \\
&= (\frac{\sqrt{\alpha_t}\mathbf{x}_t}{\beta_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0}{1-\bar{\alpha}_{t-1}})\mathbf{x}_{t-1}
\end{aligned}
$$

若令

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \sim \mathrm{N}(\mathbf{x}_{t-1}|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0),\tilde{\beta}_t\mathbf{I})$$

根据

$$p(\mathbf{x}) \sim -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + C$$

对应则有

$$\begin{aligned}
\frac{1}{\tilde{\beta}_t} &= \frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}} \\
&= \frac{\alpha_t - \alpha_t\bar{\alpha}_{t-1} + \beta_t}{\beta_t(1-\bar{\alpha}_{t-1})} \\
&= \frac{1-\bar{\alpha}_t}{\beta_t(1-\bar{\alpha}_{t-1})} \quad (\alpha_t = 1-\beta_t, \alpha_t\bar{\alpha}_{t-1} = \bar{\alpha}_t)
\end{aligned}$$

所以

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

模型推导

对于一次项有

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) &= \tilde{\beta}_t Q_1(\mathbf{x}_{t-1}) \\
&= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\left(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{\beta_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0}{1-\bar{\alpha}_{t-1}}\right) \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0
\end{aligned}$$

所以最终有

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \sim \mathrm{N}(\mathbf{x}_{t-1}|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

$$\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0, \quad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

回忆一下前面对优化目标的推导，我们最终优化的目标是极小化$\mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$，并且我们已经写出了$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$所服从的分布。一种最简单的思路就是令$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$也服从高斯分布，并且它的均值、协方差尽量和$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$相近。已知

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\sim \mathrm{N}(\mathbf{x}_{t-1}|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0),\tilde{\beta}_t\mathbf{I})$$

令

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\sim \mathrm{N}(\mathbf{x}_{t-1}|\boldsymbol{\mu}_\theta(\mathbf{x}_t,\mathbf{t}),\boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t))$$

注意这里对$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$的建模是跟时间$t$有关的。先看协方差，最简单的办法是直接令

$$\boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t)=\sigma_t^2\mathbf{I}=\tilde{\beta}_t\mathbf{I},\quad \sigma_t^2=\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

或者更简单一点

$$\boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t)=\sigma_t^2\mathbf{I}=\beta_t\mathbf{I},\quad \sigma_t^2=\beta_t$$

这种两设定最终的实际效果可能差不太多，后面代码实现时我们按第一种方式实现。

因为最终要求的是$\mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$，而这里的$q$和$p_\theta$都服从高斯分布，所以我们先求两个高斯分布的KL散度。设$p(x)=\mathrm{N}(\mathrm{x}|\mu_1,\sigma_1^2), q(x)=\mathrm{N}(x|\mu_2,\sigma_2^2)$，那么有

$$
\begin{aligned}
\mathrm{KL}(p||q) &= \int p(x)\ln\frac{p(x)}{q(x)}\mathrm{d}x \\
&= \int p(x)\ln\frac{\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_1}\exp(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2)}{\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_2}\exp(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2)}\mathrm{d}x \\
&= \int p(x)\ln\frac{\sigma_2}{\sigma_1}\mathrm{d}x + \int p(x)\ln\exp\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}+\frac{(x-\mu_2)^2}{2\sigma_2^2}\}\mathrm{d}x \\
&= \ln\frac{\sigma_2}{\sigma_1} + \int p(x)\{-\frac{1}{2\sigma_1^2}(x^2-2x\mu_1+\mu_1^2)+\frac{1}{2\sigma_2^2}(x^2-2x\mu_2+\mu_2^2)\}\mathrm{d}x \\
&= \ln\frac{\sigma_2}{\sigma_1} - \frac{1}{2\sigma_1^2}\{\int p(x)x^2\mathrm{d}x - 2\mu_1\int p(x)x\mathrm{d}x + \mu_1^2\int p(x)\mathrm{d}x\} + \frac{1}{2\sigma_2^2}\{\int p(x)x^2\mathrm{d}x - 2\mu_2\int p(x)x\mathrm{d}x + \mu_2^2\int p(x)\mathrm{d}x\} \\
&= \ln\frac{\sigma_2}{\sigma_1} - \frac{1}{2\sigma_1^2}\{\sigma_1^2+\mu_1^2-2\mu_1^2+\mu_1^2\} + \frac{1}{2\sigma_2^2}\{\sigma_1^2+\mu_1^2-2\mu_2\mu_1+\mu_2^2\} \\
&= \ln\frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}
\end{aligned}
$$

而对于$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$和$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$，前面已经假定了$\boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t)=\sigma_t^2\mathbf{I}=\tilde{\beta}_t\mathbf{I}$，相当于它们的方差相等，所以套用前面的公式则有

$$\begin{aligned}\mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) &= \ln\frac{\sigma_2}{\sigma_1}-\frac{1}{2}+\frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} \\ &= \ln 1-\frac{1}{2}+\frac{\sigma_t^2+(\tilde{\mu}_t-\mu_\theta)^2}{2\sigma_t^2} \\ &= \frac{1}{2\sigma_t^2}\|\tilde{\mu}_t-\mu_\theta\|^2\end{aligned}$$

至此，我们优化的目标函数最终变成了

$$\mathrm{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}[\mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))]=\mathrm{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0)-\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\|^2]$$

模型推导

要想极小化优化目标函数，最简单的办法就是令

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0)$$

其中$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0)$前面已经推导过了

$$\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0$$

前面我们已经做过递推推导得到了$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$，稍作变形则有

$$\mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t}}$$

将这一结果代入$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0)$则有

模型推导

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t}} \\
&= \{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\sqrt{\bar{\alpha}_t} + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}\}\mathbf{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}\boldsymbol{\epsilon} \\
&= \frac{\alpha_t\sqrt{\bar{\alpha}_{t-1}} - \bar{\alpha}_t\sqrt{\bar{\alpha}_{t-1}} + \sqrt{\bar{\alpha}_{t-1}} - \alpha_t\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}\mathbf{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{\sqrt{\bar{\alpha}_t}\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon} \\
&= \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\bar{\alpha}_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}\mathbf{x}_t - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon} \\
&= \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon} \\
&= \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t\boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}})
\end{aligned}
$$

根据$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0)$的思路可以对$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$建模为

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}})$$

这里的$\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$就成了整个问题的关键，我们正是为它建立了神经网络模型（后面代码实现时我们构建带注意力头的U-Net）。这个网络的输入是$\mathbf{x}_t$和t，输出是一个噪声值。此时损失函数为：

$$\begin{aligned}
\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2 &= \frac{1}{2\sigma_t^2}\|\frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t \boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}}) - \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}})\|^2 \\
&= \frac{1}{2\sigma_t^2}\|\frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta)\|^2 \\
&= \frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2
\end{aligned}$$

利用这一损失函数就能求θ的梯度并进而更新θ的值[5]。

[5]这里主要通过BP算法进行更新，相关内容可参考PRML 5.3或PRML Page-by-page项目的5-027

总结一下DDPM的训练过程：

1. 从训练集中选取一个样本$\mathbf{x}_0$：$\mathbf{x}_0 \sim q(\mathbf{x}_0)$
2. 确定扩散步数t：$t \sim \mathrm{Uniform}(\{1, \cdots, T\})$
3. 获取前向采样的噪声ϵ：$\epsilon \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$
4. 利用梯度对神经网络中的参数进行更新，梯度为损失函数关于参数的导数：$\nabla_\theta \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2$ [6]
5. 如果达到收敛条件就停止计算否则返回第1步

问题：第2步为什么要从$t \sim \mathrm{Uniform}(\{1, \cdots, T\})$中随机采样一个$t$了？

[6] 这里主要通过BP算法进行更新，相关内容可参考PRML 5.3或PRML Page-by-page项目的5-027

总结一下DDPM的生成过程：

1. 从纯噪声中采样获得$\mathbf{x}_T$：$\mathbf{x}_T \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$，并令$t = T$
2. 若$t \neq 0$：将$\mathbf{x}_t$和t送入训练得到的神经网络，计算输出ϵ_θ；从$\mathrm{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$中采样得到$\mathbf{z}$；再计算得到$\mathbf{x}_{t-1}$。计算方法为：根据$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$可将$\mathbf{x}_{t-1}$再参数化为

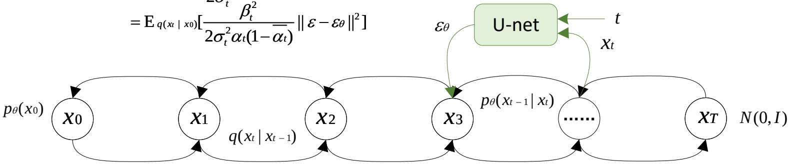
$$
\begin{aligned}
\mathbf{x}_{t-1} &= \boldsymbol{\mu}_\theta(\mathbf{x}_t, \epsilon_\theta(\mathbf{x}_t, t)) + \sigma_t \mathbf{z}, \quad \mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \\
&= \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta) + \sigma_t \mathbf{z}
\end{aligned}
$$

3. 若$t = 0$则将$\mathbf{x}_t$返回给用户

Python代码实现

假设我们构建一个带注意力机制的U-net，根据前面的推导这个网络接受$\mathbf{x}_t$和t两个输入，生成噪声$\epsilon_\theta(\mathbf{x}_t, t)$作为输出。所以我们的训练工作是根据前面推导的Loss函数，利用BP算法更新网络中的参数$\boldsymbol{\theta}$。

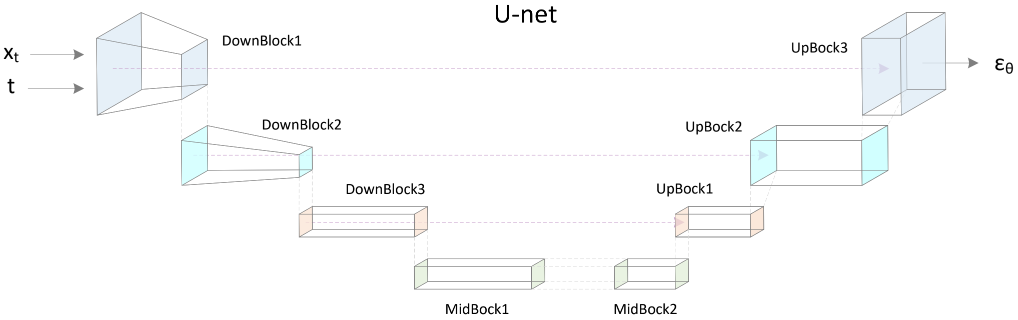
$$
\begin{aligned}
Loss &= \mathrm{E}_{q(x_t \mid x_0)}\left[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2\right] \\
&= \mathrm{E}_{q(x_t \mid x_0)}\left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t(1-\bar{\alpha}_t)}\|\varepsilon - \varepsilon_\theta\|^2\right]
\end{aligned}
$$



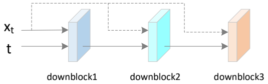
代码实现主要参考：https://github.com/explainingai-code/DDPM-Pytorch

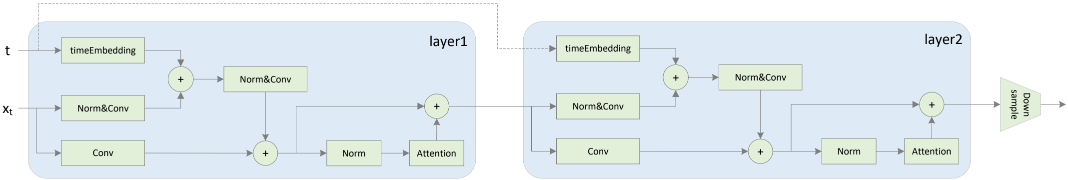
具体到U-net内部，我们设计一个较为简单的网络来实现，这里面比较重要的点包括

- 下采样阶段channel不断增加但图像的size不断减小
- 上采样channel不断减少同时利用反向卷积恢复图像的size
- 下采样的输入会直连到上采样作为输入



U-net的下采样阶段，我们设计三个Block，每个Block的结构几乎一样，也主要包括规范化、卷积、time embedding、残差、自注意力等计算步骤。它的下采样是通过卷积实现的。





Python代码实现

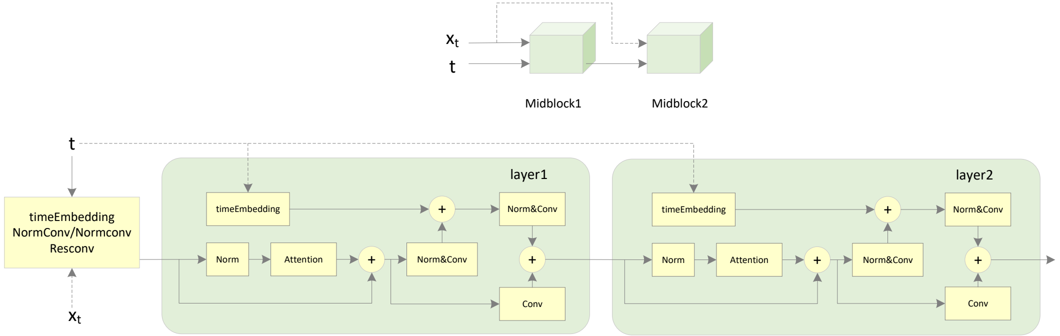
在卷积计算时要注意数据维度，假设数据输入为$(B, C_{in}, H_{in}, W_{in})$（分别代表Batch_size，Channels，Height，Width），当前指定的卷积核大小为$kernel_size = [k1, k2]$，衬垫填充大小为$padding = [p1, p2]$，前进步伐为$stride = [s1, s2]$，那么卷积后数据维度为$(N, C_{out}, H_{out}, W_{out})$

- C_{out}由用户在定义卷积层时指定
- $H_{out} = \frac{H_{in} - kernel\_size[0] + 2 * padding[0]}{stride[0]} + 1$
- $W_{out} = \frac{W_{in} - kernel\_size[1] + 2 * padding[1]}{stride[1]} + 1$

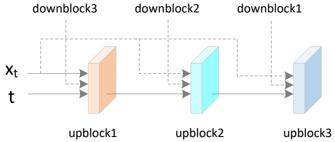
例如以下情况：

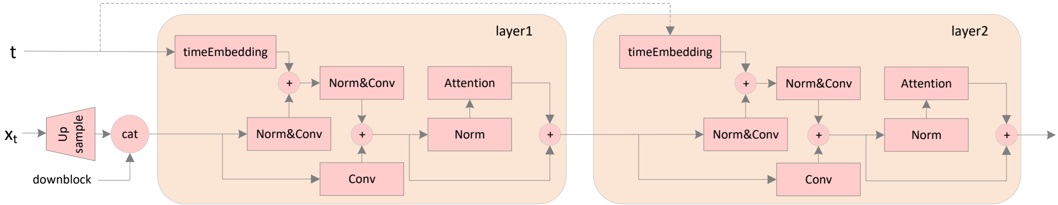
- 若$kernel = 3, stride = 1, padding = 1$，那么经卷积后有$H_{out} = \frac{H_{in} - 3 + 2}{1} + 1 = H_{in} - 1 + 1 = H_{in}$
- 若$kernel = 4, stride = 2, padding = 1$，那么经卷积后有$H_{out} = \frac{H_{in} - 4 + 2}{2} + 1 = \frac{1}{2}H_{in} - 1 + 1 = \frac{1}{2}H_{in}$
- 若$kernel = 1$（采用默认值$stride = 1, padding = 0$），经卷积后有$H_{out} = \frac{H_{in} - 1}{1} + 1 = H_{in}$

U-net的中间层我们设计两个Block，每个Block主要包括规范化、卷积、time embedding、残差、自注意力等计算步骤。



U-net的上采样阶段设计三个Block，每个Block也包括规范化、卷积、time embedding、残差、自注意力等，但它有两个特殊之处：和下采样对应block的拼接以及反向卷积恢复数据维度。





利用反向卷积扩增数据维度时，假设数据输入为$(B, C_{in}, H_{in}, W_{in})$（分别代表Batch_size，Channels，Height，Width），指定的反向卷积参数为$kernel_size = k$，$stride = s$，$padding = p$，其它参数取默认值如$output_padding = 0$，$dilation = 1$，那么反向卷积后数据维度为$(N, C_{out}, H_{out}, W_{out})$：

$$H_{out} = (H_{in} - 1) * s - 2 * p + k = s * H_{in} - 2 * p - s + k$$

假设$kernel_size = 4, stride = 2, padding = 1$，那么反向卷积后数据维度为：

$$H_{out} = 2 * H_{in} - 2 * 1 - 2 + 4 = 2 * H_{in}$$

可见此时数据规模将被扩大至原来的两倍

时间参数t的embedding算法：时间t本身是标量，假设要将t嵌入到100维的空间中（即将标量t转换为100维的向量）：

- 将t扩增为50维的向量
- 定义50维的因子数
- 用t的各个维度与因子数的各个维度相除（将t的数值映射到较小的范围内）
- 对t的各个元素做sin和cos计算并将结果拼接起来，得到100维的向量

```
t_emb = time_steps[:,None].repeat(1, t_emb_dim // 2)
factor = 10000 ** ( torch.arange(start=0, end=t_emb_dim // 2,
    dtype=torch.float32) / ( t_emb_dim // 2))
t_emb = t_emb / factor
t_emb = torch.cat([torch.sin(t_emb), torch.cos(t_emb)], dim=-1)
```

在处理噪声时，首先应确定α_t、β_t的值。可设定$\beta_0 = 0.0001, \beta_T = 0.02$，其中$T$为最大扩散步数。再根据$\alpha_t = 1 - \beta_t$和$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ 求出α_t、$\sqrt{\bar{\alpha}_t}$、$\sqrt{1-\bar{\alpha}_t}$

```
betas = torch.linspace(beta_0, beta_T, T)
alphas = 1. - betas
alpha_cum_prod = torch.cumprod(alphas, dim=0)
sqrt_alpha_cum_prod = torch.sqrt(alpha_cum_prod)
sqrt_one_minus_alpha_cum_prod = torch.sqrt(1 - alpha_cum_prod)
```

在训练时首先利用训练数据$\mathbf{x}_0$和$\mathbf{x}_t$的关系式得到$\mathbf{x}_t$（作为反向去噪声时第t时刻的数据）

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

然后利用待训练的网络基于$\mathbf{x}_t$和t求出反向噪声ϵ_θ

$$\epsilon_\theta = \mathrm{model}(\mathbf{x}_t, t)$$

再利用均方误差函数$\nabla_\theta \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2$求梯度更新参数

在预测阶段，首先采样一个纯噪声作为$\mathbf{x}_T$，然后从$\mathbf{x}_T$开始做T次去噪声。假设当前在第t步，根据前面推导的公式

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta) + \sigma_t \mathbf{z}$$

将$\mathbf{x}_t$和t送入训练后的U-net中得到这里的ϵ_θ，从$N(\mathbf{0}, \mathbf{I})$中采样得到$\mathbf{z}$。这里的$\sigma_t$采用前面介绍的第一种方式

$$\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

就这样不断迭代，直到$t=0$时即得到生成的图像$\mathbf{x}_0$

小结

对于原理部分，有以下几点需要注意：

- DDPM的正反向扩散过程本身是确定的，不确定的是待减去的噪声
- 较长的扩散步数会使得模型训练时间很长

对于实现部分，有以下几点需注意：

- 注意正确设定各layer输入输出的channel数
- 所有参与计算的变量要位于同一设备上（CPU或GPU）

网友问题：论文《Understanding Diffusion Models: A Unified Perspective》（arXiv:2208.11970）推导过程如何从等式（44）等到等式（45）?

这里最关键的一点在于要把握一阶马尔科夫链的特性：前向过程中$\mathbf{x}_t$的取值只跟$\mathbf{x}_{t-1}$有关，即$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)$。注意前向过程是从$\mathbf{x}_0$得到$\mathbf{x}_1$，再从$\mathbf{x}_1$得到$\mathbf{x}_2$并这样迭代下去的，所以虽然理论上$\mathbf{x}_t$只跟$\mathbf{x}_{t-1}$有关，但是因为$\mathbf{x}_{t-1}$是由$\mathbf{x}_0$ 逐步演化得到的，所以经过变形也可以把$\mathbf{x}_t$写成跟$\mathbf{x}_0$有关的形式，或者写成同时跟$\mathbf{x}_{t-1}$和$\mathbf{x}_0$有关的形式（把一部分$\mathbf{x}_{t-1}$保留，另一部分$\mathbf{x}_{t-1}$ 替换成$\mathbf{x}_0$的形式）[7]

$$\mathrm{E}_{q(\mathbf{x}_{T-1}, \mathbf{x}_T|\mathbf{x}_0)}[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})}] = \int q(\mathbf{x}_{T-1}, \mathbf{x}_T|\mathbf{x}_0) \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \mathrm{d}\mathbf{x}_{T-1}\mathrm{d}\mathbf{x}_T$$

根据条件概率公式及前面的分析有

$$\begin{aligned} q(\mathbf{x}_{T-1}, \mathbf{x}_T|\mathbf{x}_0) &= q(\mathbf{x}_T|\mathbf{x}_{T-1}, \mathbf{x}_0) q(\mathbf{x}_{T-1}|\mathbf{x}_0) \\ &= q(\mathbf{x}_T|\mathbf{x}_{T-1}) q(\mathbf{x}_{T-1}|\mathbf{x}_0) \end{aligned}$$

[7]这一点前面的Slide11也解释过

所以代入之后有

$$
\begin{aligned}
\mathrm{E}_{q(\mathbf{x}_{T-1},\mathbf{x}_T|\mathbf{x}_0)}[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})}] &= \int q(\mathbf{x}_{T-1},\mathbf{x}_T|\mathbf{x}_0) \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \mathrm{d}\mathbf{x}_{T-1}\mathrm{d}\mathbf{x}_T \\
&= \int q(\mathbf{x}_T|\mathbf{x}_{T-1}) q(\mathbf{x}_{T-1}|\mathbf{x}_0) \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \mathrm{d}\mathbf{x}_{T-1}\mathrm{d}\mathbf{x}_T \\
&= \int q(\mathbf{x}_{T-1}|\mathbf{x}_0) \int q(\mathbf{x}_T|\mathbf{x}_{T-1}) \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \mathrm{d}\mathbf{x}_T\mathrm{d}\mathbf{x}_{T-1} \\
&= -\int q(\mathbf{x}_{T-1}|\mathbf{x}_0) \mathrm{KL}(q(\mathbf{x}_T|\mathbf{x}_{T-1}) \| p(\mathbf{x}_T)) \mathrm{d}\mathbf{x}_{T-1} \\
&= -\mathrm{E}_{q(\mathbf{x}_{T-1}|\mathbf{x}_0)}[\mathrm{KL}(q(\mathbf{x}_T|\mathbf{x}_{T-1}) \| p(\mathbf{x}_T))]
\end{aligned}
$$

后面一项的推导是类似的，只需注意到

$$\begin{aligned} q(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}|\mathbf{x}_0) &= q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \mathbf{x}_0) q(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0) \\ &= q(\mathbf{x}_t|\mathbf{x}_{t-1}) q(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0) \end{aligned}$$

所以

$$\begin{aligned} \mathrm{E}_{q(\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{x}_{t+1}|\mathbf{x}_0)}[\log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}] &= \int\int\int q(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \mathrm{d}\mathbf{x}_{t-1}\mathrm{d}\mathbf{x}_{t+1}\mathrm{d}\mathbf{x}_t \\ &= \int\int\int q(\mathbf{x}_t|\mathbf{x}_{t-1}) q(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \mathrm{d}\mathbf{x}_{t-1}\mathrm{d}\mathbf{x}_{t+1}\mathrm{d}\mathbf{x}_t \\ &= \int\int q(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0) \int q(\mathbf{x}_t|\mathbf{x}_{t-1}) \log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{x}_{t-1}\mathrm{d}\mathbf{x}_{t+1} \\ &= -\int\int q(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0) \mathrm{KL}(q(\mathbf{x}_t|\mathbf{x}_{t-1}) \| p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \mathrm{d}\mathbf{x}_{t-1}\mathrm{d}\mathbf{x}_{t+1} \\ &= -\mathrm{E}_{q(\mathbf{x}_{t-1},\mathbf{x}_{t+1}|\mathbf{x}_0)}[\mathrm{KL}(q(\mathbf{x}_t|\mathbf{x}_{t-1}) \| p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}))] \end{aligned}$$

The end