生成扩散模型学习（一）：DDPM = 拆楼 + 建楼

朱梦

初稿于 2025-08-25，修改于 2025-08-28

1. 生成模型类比拆楼建楼

生成模型可以定义为将一个随机噪声 $\boldsymbol{z}$ 变换成一个样本数据 $\boldsymbol{x}$ 的过程。可以将此过程类比为“建设”，其中随机噪声 $\boldsymbol{z}$ 类比为砖瓦泥等建筑原材料，样本数据 $\boldsymbol{x}$ 类比为建筑好的高楼大厦，所以生成模型就是一支用建筑原材料建设为高楼大厦的施工队。

$$
\begin{tikzcd}
\text{随机噪声}\boldsymbol{z} \arrow[r, "\text{变换}"] \arrow[d, "\text{类比}"'] & \text{样本数据}\boldsymbol{x} \arrow[d, "\text{类比}"'] \\
\text{砖瓦水泥} \arrow[r, "\text{建设}"] & \text{高楼大厦}
\end{tikzcd}
$$

图 1 生成模型类比建楼过程

建楼过程肯定很难的，所以才用了那么多关于生成模型的研究。但是，俗话说“破环容易建设难”，可以考虑将高楼大厦一步步地拆为砖瓦水泥的过程。设 $\boldsymbol{x}_0$ 为建设好的高楼大厦（即样本数据），$\boldsymbol{x}_T$ 为拆好的砖瓦水泥（即随机噪声），每一步拆楼的过程定义为：

$$f_{\text{拆楼}}: \boldsymbol{x}_{t-1} \to \boldsymbol{x}_t \tag{1}$$

假定“拆楼”至砖瓦水泥需要 T 步，那么循环执行式(1)共 T 步，就可以得到完整的拆楼过程：

$$\text{for } i = 1 \to T: f_{\text{拆楼}}: \boldsymbol{x}_{t-1} \to \boldsymbol{x}_t, \text{即} \boldsymbol{x}_0 \to \boldsymbol{x}_1 \to ... \to \boldsymbol{x}_{T-1} \to \boldsymbol{x}_T \tag{2}$$

建造高楼大厦的难度在于从砖瓦水泥 $\boldsymbol{x}_T$ 到最终高楼大厦 $\boldsymbol{x}_0$ 的跨度过大，大多数人很难理解 $\boldsymbol{x}_T$ 是如何一步变成 $\boldsymbol{x}_0$ 的。但是，当有了拆楼的中间过程 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$，...，$\boldsymbol{x}_{T-1}$ 后，那么反过来 $\boldsymbol{x}_t \to \boldsymbol{x}_{t-1}$ 不就表示建楼的一步？如果能建模每一步建楼的过程 $f^{-1}_{\text{建楼}}: \boldsymbol{x}_t \to \boldsymbol{x}_{t-1}$，那么循环执行 T 步，最终不就造出高楼大厦了？

2. DDPM 如何拆

具体来说，去噪扩散概率模型（Denoising Diffusion Probabilistic Models，DDPM）将每一步“拆楼”的过程定义为：

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_{t-1} + \beta_t \boldsymbol{\epsilon}_t,\ \text{s.t. } \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{3}$$

其中，$\alpha_t, \beta_t > 0$ 且满足 $\alpha_t^2 + \beta_t^2 = 1$，$\beta_t$ 通常接近于 0。$\beta_t \boldsymbol{\epsilon}_t$ 表示每一步“拆楼”过程中对“原楼体” $\boldsymbol{x}_{t-1}$ 的破坏程度。式(3)也可以理解为每一步“拆楼”过程中将“原楼体” $\boldsymbol{x}_{t-1}$ 分解为 $\alpha_t \boldsymbol{x}_{t-1}$ 的“楼体”加上 $\beta_t \boldsymbol{\epsilon}_t$ 的“砖瓦水泥”

反复迭代式(3)有：

$$\begin{aligned}
\boldsymbol{x}_t &= \alpha_t \boldsymbol{x}_{t-1} + \beta_t \boldsymbol{\epsilon}_t \\
&= \alpha_t(\alpha_{t-1}\boldsymbol{x}_{t-2} + \beta_{t-1}\boldsymbol{\epsilon}_{t-1}) + \beta_t \boldsymbol{\epsilon}_t \\
&= \dots \\
&= (\alpha_t \dots \alpha_1)\boldsymbol{x}_0 + (\alpha_t \dots \alpha_2)\beta_1 \boldsymbol{\epsilon}_1 + (\alpha_t \dots \alpha_3)\beta_2 \boldsymbol{\epsilon}_2 + \cdots + \alpha_t \beta_{t-1}\boldsymbol{\epsilon}_{t-1} + \beta_t \boldsymbol{\epsilon}_t \\
&= \prod_{i=1}^{t} \alpha_i \boldsymbol{x}_0 + \prod_{i=2}^{t} \alpha_i \beta_1 \boldsymbol{\epsilon}_1 + \prod_{i=3}^{t} \alpha_i \beta_2 \boldsymbol{\epsilon}_2 + \cdots + \alpha_t \beta_{t-1}\boldsymbol{\epsilon}_{t-1} + \beta_t \boldsymbol{\epsilon}_t \\
&= \prod_{i=1}^{t} \alpha_i \boldsymbol{x}_0 + \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} \alpha_i \beta_j \boldsymbol{\epsilon}_j + \beta_t \boldsymbol{\epsilon}_t
\end{aligned} \tag{4}$$

首先，上式中 $\sum_{j=1}^{t-1} \prod_{i=j+1}^{t} \alpha_i \beta_j \boldsymbol{\epsilon}_j + \beta_t \boldsymbol{\epsilon}_t$ 正好为多个独立的正太噪声之和，其均值为 0，方差分别为 $(\alpha_t \dots \alpha_2)^2 \beta_1^2$、$(\alpha_t \dots \alpha_3)^2 \beta_2^2$、…、$\beta_t^2$。然后，根据正太分布的可叠加性，即 $\sum_{j=1}^{t-1} \prod_{i=j+1}^{t} \alpha_i \beta_j \boldsymbol{\epsilon}_j + \beta_t \boldsymbol{\epsilon}_t$ 服从均值为 0、方差为 $(\alpha_t \dots \alpha_2)^2 \beta_1^2 + (\alpha_t \dots \alpha_3)^2 \beta_2^2 + \cdots + \beta_t^2$ 的正太分布。最后，在 $\alpha_t^2 + \beta_t^2 = 1$ 恒成立之下，可以得到：

$$(\alpha_t \dots \alpha_1)^2 (\alpha_t \dots \alpha_2)^2 \beta_1^2 + (\alpha_t \dots \alpha_3)^2 \beta_2^2 + \cdots + \beta_t^2 = 1 \tag{5}$$

因此，式(4)可以等价于：

$$\begin{aligned}
\boldsymbol{x}_t &= (\alpha_t \dots \alpha_1)\boldsymbol{x}_0 + \sqrt{1 - (\alpha_t \dots \alpha_1)^2}\,\bar{\boldsymbol{\epsilon}}_t \\
&\triangleq \bar{\alpha}_t \boldsymbol{x}_0 + \bar{\beta}_t \bar{\boldsymbol{\epsilon}}_t,\ \text{s.t. } \bar{\boldsymbol{\epsilon}}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})
\end{aligned} \tag{6}$$

这为计算 $\boldsymbol{x}_t$ 提供了极大的便利。且 DDPM 会选择适当的 α_t 形式，使得 $\bar{\alpha}_T \approx 0$，这意味着经过 T 步的拆楼后，所剩的“楼体”几乎所剩无几，已经全部转换为“砖瓦水泥”。

3. DDPM 又如何建与损失函数

“拆楼”是 $\boldsymbol{x}_{t-1} \to \boldsymbol{x}_t$ 的过程，此过程可以得到很多的数据对 $(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)$，那么“建楼”自然就是从这些数据对中学习一个 $\boldsymbol{x}_t \to \boldsymbol{x}_{t-1}$ 的模型。首先，“拆楼”的式(3)可以改写为 $\boldsymbol{x}_{t-1} = (\boldsymbol{x}_t - \beta_t \boldsymbol{\epsilon}_t)/\alpha_t$，这启发可以将建楼模型 $g(\boldsymbol{x}_t)$ 设计为：

$$g(\boldsymbol{x}_t) = \frac{1}{\alpha_t}(\boldsymbol{x}_t - \beta_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)) \tag{7}$$

其中，$\boldsymbol{\theta}$ 表示待优化参数。容易想到的损失函数就是最小化 $\boldsymbol{x}_{t-1}$ 和 $g(\boldsymbol{x}_t)$ 之间的欧式距离：

$$\begin{aligned} \|\boldsymbol{x}_{t-1} - g(\boldsymbol{x}_t)\|^2 &= \|\frac{1}{\alpha_t}(\boldsymbol{x}_t - \beta_t \boldsymbol{\epsilon}_t) - \frac{1}{\alpha_t}(\boldsymbol{x}_t - \beta_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))\|^2 \\ &= \frac{\beta_t^2}{\alpha_t^2}\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}_t\|^2 \end{aligned} \tag{8}$$

结合式(3)、(6)有：

$$\begin{aligned} \boldsymbol{x}_t &= \alpha_t \boldsymbol{x}_{t-1} + \beta_t \boldsymbol{\epsilon}_t = \alpha_t(\bar{\alpha}_{t-1}\boldsymbol{x}_0 + \bar{\beta}_{t-1}\bar{\boldsymbol{\epsilon}}_{t-1}) + \beta_t \boldsymbol{\epsilon}_t \\ &= \bar{\alpha}_t \boldsymbol{x}_0 + \alpha_t \bar{\beta}_{t-1}\bar{\boldsymbol{\epsilon}}_{t-1} + \beta_t \boldsymbol{\epsilon}_t \end{aligned} \tag{9}$$

将式(9)代入式(8)，得到的损失函数形式为：

$$\frac{\beta_t^2}{\alpha_t^2}\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t \boldsymbol{x}_0 + \alpha_t \bar{\beta}_{t-1}\bar{\boldsymbol{\epsilon}}_{t-1} + \beta_t \boldsymbol{\epsilon}_t, t) - \boldsymbol{\epsilon}_t\|^2 \tag{10}$$

为什么式(9)要回退一步给出 $\boldsymbol{x}_t$？因为已经事先采样了 $\boldsymbol{\epsilon}_t$，而 $\boldsymbol{\epsilon}_t$ 和 $\bar{\boldsymbol{\epsilon}}_t$ 并不是相互独立的，因此给定 $\boldsymbol{\epsilon}_t$ 的情况下，不能完全独立采样 $\bar{\boldsymbol{\epsilon}}_t$。

4. 降低方差与超参设置

原则上，损失函数(10)已经可以完成 DDPM 的训练了，但它在实践中可能有方差过大的风险，从而导致收敛过慢的问题。要理解这一点并不困难，只需观察到式(10)包含了四个需要采样的随机变量：

（1） 从所有训练样本采样一个数据样本 $\boldsymbol{x}_0$。

（2） 从正太分布 $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 中采样 $\bar{\boldsymbol{\epsilon}}_{t-1}$ 和 $\boldsymbol{\epsilon}_t$。

（3） 从 $1 \sim T$ 中采样一个 t。

要采样的随机变量越多，就越难对损失函数做准确估计，即每次对损失函数进行估计的波动（方差）过大。幸运的是，可以通过积分技巧来将 $\bar{\boldsymbol{\epsilon}}_{t-1}$、$\boldsymbol{\epsilon}_t$ 合并在单个正太随机变量中，从而缓解方差过大的问题。这个积分确实有点技巧性，但也不算复杂。由于正太分布的叠加性，可得 $\alpha_t\bar{\beta}_{t-1}\bar{\boldsymbol{\epsilon}}_{t-1} + \beta_t\boldsymbol{\epsilon}_t$ 实际相当于单个随机变量 $\bar{\beta}_t\boldsymbol{\epsilon}|\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$。同理，$\beta_t\bar{\boldsymbol{\epsilon}}_{t-1} - \alpha_t\bar{\beta}_{t-1}\boldsymbol{\epsilon}_t$ 实际相当于单个随机变量 $\bar{\beta}_t\boldsymbol{\omega}|\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$。并且可以证明 $\mathbb{E}[<\boldsymbol{\epsilon}, \boldsymbol{\omega}>] = 0$。接下来，用 $\boldsymbol{\epsilon}$ 和 $\boldsymbol{\omega}$ 重新表示 $\boldsymbol{\epsilon}_t$

$$\boldsymbol{\epsilon}_t = \frac{(\beta_t\boldsymbol{\epsilon} - \alpha_t\bar{\beta}_{t-1}\boldsymbol{\omega})\bar{\beta}_t}{\beta_t^2 + \alpha_t^2\bar{\beta}_{t-1}^2} = \frac{\beta_t\boldsymbol{\epsilon} - \alpha_t\bar{\beta}_{t-1}\boldsymbol{\omega}}{\bar{\beta}_t} \tag{11}$$

代入式(10)得：

$$\begin{aligned} &\mathbb{E}_{\bar{\boldsymbol{\epsilon}}_{t-1}, \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}\left[\frac{\beta_t^2}{\alpha_t^2}\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t\boldsymbol{x}_0 + \alpha_t\bar{\beta}_{t-1}\bar{\boldsymbol{\epsilon}}_{t-1} + \beta_t\boldsymbol{\epsilon}_t, t) - \boldsymbol{\epsilon}_t\|^2\right] \\ &= \frac{\beta_t^2}{\alpha_t^2}\mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t\boldsymbol{x}_0 + \bar{\beta}_t\boldsymbol{\epsilon}, t) - \frac{\beta_t\boldsymbol{\epsilon} - \alpha_t\bar{\beta}_{t-1}\boldsymbol{\omega}}{\bar{\beta}_t}\|^2\right] \end{aligned} \tag{12}$$

由于式(12)关于 $\boldsymbol{\omega}$ 只是二次的，所以可以展开将它的期望直接算出来，结果为：

$$\frac{\beta_t^2}{\alpha_t^2}\left(\frac{\beta_t^2}{\bar{\beta}_t^2}\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}\left[\|\frac{\bar{\beta}_t}{\beta_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\bar{\alpha}_t\boldsymbol{x}_0 + \bar{\beta}_t\boldsymbol{\epsilon}, t\right) - \boldsymbol{\epsilon}\|^2\right] + \text{常数}\right) \tag{13}$$

省略常数和损失函数的权重，便得到 DDPM 最终所用的损失函数：

$$\|\frac{\bar{\beta}_t}{\beta_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\bar{\alpha}_t\boldsymbol{x}_0 + \bar{\beta}_t\boldsymbol{\epsilon}, t\right) - \boldsymbol{\epsilon}\|^2 \tag{14}$$

DDPM 定义 α_t 为：

$$\alpha_t = \sqrt{1 - \frac{0.02t}{T}} \tag{15}$$

在推导式(6)时有要求 $\alpha_T \approx 0$，大致估算：

$$\begin{aligned} \log\bar{\alpha}_T &= \sum_{t=1}^{T}\log\alpha_t = \frac{1}{2}\sum_{t=1}^{T}\log\left(1 - \frac{0.02t}{T}\right) \\ &< \frac{1}{2}\sum_{t=1}^{T}\left(-\frac{0.02t}{T}\right) = -0.005(T+1) \end{aligned} \tag{16}$$

代入 $T = 1000$ 大致为 $\bar{\alpha}_T = 1 \times 10^{-5}$，这就刚好达到约为 0 的标准。

在式(7)，显示地写出来 t，这是因为原则上不同的 t 处理的是不同层次的对象，所以应该使用不同的重构模型，即应该有 T 个不同的重构模型，于是共享所有重构模型的参数，将 t 作为条件传入。按照附录的说法，t 是转换成位置编码后，直接加到残差模块上的。

5. 递归生成

训练完后，可以从一个随机噪声 $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 出发执行 T 步式(7)来进行生成。这对应于自回归解码中的 Greedy Search。如果需要进行 Random Sample，那么需要补上噪声项：

$$\boldsymbol{x}_{t-1} = \frac{1}{\alpha_t}(\boldsymbol{x}_t - \beta_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)) + \sigma_t \boldsymbol{z},\ \text{s.t. } \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{17}$$

其中，$\sigma_t = \beta_t$，即正向和反向的方差保持同步。从生成过程来看，生成速度为一个瓶颈，即每生成一张图像，需要反复将 $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))$ 执行 T 步，因此 DDPM 的一大缺陷就是采样速度慢。

引用源

转载于：

```
@online{kexuefm-9119,
    title={生成扩散模型漫谈（一）：DDPM = 拆楼 + 建楼},
    author={苏剑林},
    year={2022},
    month={Jun},
    url={https://spaces.ac.cn/archives/9119}
}
```