

扩散模型奠基与DDPM

前向过程

1. 数据分布: $q(x^{(0)})$
2. 目标分布: 一个容易解析的简单分布 $\pi(x^{(T)})$
3. 马尔可夫内核 T_π : 一个一般性的马尔可夫矩阵, 将当前状态映射到下一个状态的概率

$$q(x^{(t)}|x^{(t-1)}) = T_\pi(x^{(t)}|x^{(t-1)}; \beta_t)$$
$$q(x^{(0 \dots T)}) = q(x^{(0)}) \prod_{t=1}^T q(x^{(t)}|x^{(t-1)})$$

反向过程

1. 原始分布: 简单分布 $\pi(x^{(T)})$
2. 目标分布: 数据分布

$$p(x^{(T)}) = \pi(x^{(T)})$$
$$p(x^{(0 \dots T)}) = p(x^{(T)}) \prod_{t=1}^T p(x^{(t-1)}|x^{(t)})$$

关键假设

当每一次加噪的幅度非常小, 如果正向过程是高斯分布 (或者二项分布), 则反向过程也是高斯分布 (或者二项分布) 因此我们可以直接根据非常少的参数就能够把反向的过程确定下来。

详解反向过程

$$p(x^{(0 \dots T)}) = p(x^{(T)}) \prod_{t=1}^T p(x^{(t-1)}|x^{(t)})$$
$$p(x^{(0)}) = \int dx^{(1)} \int dx^{(2)} \dots \int dx^{(T)} p(x^{(0 \dots T)})$$
$$= \int dx^{(1 \dots T)} p(x^{(0 \dots T)}) \frac{q(x^{(1 \dots T)}|x^{(0)})}{q(x^{(1 \dots T)}|x^{(0)})}$$
$$= \int dx^{(1 \dots T)} q(x^{(1 \dots T)}|x^{(0)}) \frac{p(x^{(0 \dots T)})}{q(x^{(1 \dots T)}|x^{(0)})}$$
$$= \int dx^{(1 \dots T)} q(x^{(1 \dots T)}|x^{(0)}) \cdot p(x^{(T)}) \prod_{t=1}^T \frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t)}|x^{(t-1)})}$$

对数似然及其ELBO

现在我们已经有了 $p(x^{(0)})$ 的一个比较好的解析方法，因此可以用来比较模型的输出值 $p(x^{(0)})$ 和真实数据 $q(x^{(0)})$ 之间的匹配程度，用对数似然函数即可：

$$\begin{aligned} L &= \int dx^{(0)} q(x^{(0)}) \log p(x^{(0)}) \\ &= \int dx^{(0)} q(x^{(0)}) \log \left[\int dx^{(1 \dots T)} q(x^{(1 \dots T)} | x^{(0)}) \cdot p(x^{(T)}) \prod_{t=1}^T \frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] \end{aligned}$$

由Jenson不等式可知，对于任意的上凸函数，都有 $f(E[x]) \geq E[f(x)]$ ，而 \log 函数就是一个上凸函数，因此我们可以得到对数似然函数的下界：

$$\begin{aligned} L &= \int dx^{(0)} q(x^{(0)}) \log E_{x^{(1 \dots T)} \sim q(x^{(1 \dots T)} | x^{(0)})} \left[p(x^{(T)}) \prod_{t=1}^T \frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] \\ &\geq \int dx^{(0)} q(x^{(0)}) E_{x^{(1 \dots T)} \sim q(x^{(1 \dots T)} | x^{(0)})} \log \left[p(x^{(T)}) \prod_{t=1}^T \frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] \\ &= \int dx^{(0)} dx^{(1 \dots T)} q(x^{(0)}) q(x^{(1 \dots T)} | x^{(0)}) \log \left[p(x^{(T)}) \prod_{t=1}^T \frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] \\ &= \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[p(x^{(T)}) \prod_{t=1}^T \frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] \\ &= \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \left[\sum_{t=1}^T \log \frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} + \log p(x^{(T)}) \right] \\ &= K \end{aligned}$$

接下来我们可以对这个下界 K 继续进行处理：由(1)式可知，我们可以把 $q(x^{(0 \dots T)})$ 转换为 $q(x^{(0)}) \prod_{t=1}^T q(x^{(t)} | x^{(t-1)})$ ，后面的条件概率刚好可以和前面的积分抵消形成边缘概率分布，最终得到：

$$\begin{aligned} K &= \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \sum_{t=1}^T \log \left[\frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] + \int dx^{(T)} q(x^{(T)}) \log p(x^{(T)}) \\ &= \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \sum_{t=1}^T \log \left[\frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] + \int dx^{(T)} \pi(x^{(T)}) \log \pi(x^{(T)}) \\ &= \sum_{t=1}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] - H_p(X^{(T)}) \end{aligned}$$

接下来我们要考虑边界点，这需要进行单独考虑，就好比抓住绳子的两端之后要并在一起才能把绳子完美地对折：

$$p(x^{(0)} | x^{(1)}) = q(x^{(1)} | x^{(0)}) \frac{\pi(x^{(0)})}{\pi(x^{(1)})} = T_\pi(x^{(0)} | x^{(1)}; \beta_1)$$

$$\begin{aligned} K &= \sum_{t=2}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] + \int dx^{(0)} \int dx^{(1)} q(x^{(0)}, x^{(1)}) \log \left[\frac{p(x^{(0)} | x^{(1)})}{q(x^{(1)} | x^{(0)})} \right] - H_p(X^{(T)}) \\ &= \sum_{t=2}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] + \int dx^{(0)} \int dx^{(1)} q(x^{(0)}, x^{(1)}) \log \left[\frac{q(x^{(1)} | x^{(0)}) \pi(x^{(0)})}{q(x^{(1)} | x^{(0)}) \pi(x^{(1)})} \right] - H_p(X^{(T)}) \\ &= \sum_{t=2}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)} | x^{(t)})}{q(x^{(t)} | x^{(t-1)})} \right] + \int dx^{(0)} \int dx^{(1)} q(x^{(0)}, x^{(1)}) \log \left[\frac{\pi(x^{(0)})}{\pi(x^{(1)})} \right] - H_p(X^{(T)}) \end{aligned}$$

由于前向加噪的初始步长非常小，实际上通过积分得到了两个Entropy可以直接消去，最终得到：

$$\begin{aligned}
K &= \sum_{t=2}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t)}|x^{(t-1)})} \right] - H_p(X^{(T)}) \\
&= \sum_{t=2}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t)}|x^{(t-1)}, x^{(0)})} \right] - H_p(X^{(T)}) \\
&= \sum_{t=2}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t-1)}|x^{(t)}, x^{(0)})} \frac{q(x^{(t-1)}|x^{(0)})}{q(x^{(t)}|x^{(0)})} \right] - H_p(X^{(T)}) \\
&= \sum_{t=2}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t-1)}|x^{(t)}, x^{(0)})} \right] + \sum_{t=2}^T [H_q(X^{(t)}|X^{(0)}) - H_q(X^{(t-1)}|X^{(0)})] - H_p(X^{(T)}) \\
&= \sum_{t=2}^T \int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t-1)}|x^{(t)}, x^{(0)})} \right] + H_q(X^{(T)}|X^{(0)}) - H_q(X^{(1)}|X^{(0)}) - H_p(X^{(T)})
\end{aligned}$$

接下来处理前面的大块头：由于log函数里面只和 $x^{(t)}, x^{(t-1)}, x^{(0)}$ 有关，所以对于每一项积分，我们都可以把其他的积分全部消掉变成1:

$$\begin{aligned}
&\int dx^{(0 \dots T)} q(x^{(0 \dots T)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t-1)}|x^{(t)}, x^{(0)})} \right] \\
&= \int dx^{(0)} dx^{(t-1)} dx^{(t)} q(x^{(0)}) q(x^{(t-1)}) q(x^{(t)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t-1)}|x^{(t)}, x^{(0)})} \right] \\
&= \int dx^{(0)} dx^{(t-1)} dx^{(t)} q(x^{(0)}, x^{(t)}) q(x^{(t-1)}|x^{(t)}, x^{(0)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t-1)}|x^{(t)}, x^{(0)})} \right] \\
&= \int dx^{(0)} dx^{(t)} q(x^{(0)}, x^{(t)}) \int dx^{(t-1)} q(x^{(t-1)}|x^{(t)}, x^{(0)}) \log \left[\frac{p(x^{(t-1)}|x^{(t)})}{q(x^{(t-1)}|x^{(t)}, x^{(0)})} \right] \\
&= - \int dx^{(0)} dx^{(t)} q(x^{(0)}, x^{(t)}) D_{KL}(q(x^{(t-1)}|x^{(t)}, x^{(0)}) \| p(x^{(t-1)}|x^{(t)}))
\end{aligned}$$

因此我们可以用KL散度来表示K:

$$\begin{aligned}
K &= - \sum_{t=2}^T \int dx^{(0)} dx^{(t)} q(x^{(0)}, x^{(t)}) D_{KL}(q(x^{(t-1)}|x^{(t)}, x^{(0)}) \| p(x^{(t-1)}|x^{(t)})) \\
&\quad + H_q(X^{(T)}|X^{(0)}) - H_q(X^{(1)}|X^{(0)}) - H_p(X^{(T)})
\end{aligned}$$

由于后面几个熵的值都是常数，我们只能尝试通过最小化 $D_{KL}(q(x^{(t-1)}|x^{(t)}, x^{(0)}) \| p(x^{(t-1)}|x^{(t)}))$ 使得K尽可能达到最大，因此L的下界尽可能增大。

前向扩散核、反向扩散核

在前向扩散的时候，我们可以自定义加噪公式：

$$\begin{aligned}
q(x^{(t)}|x^{(t-1)}) &= \mathcal{N}(x^{(t)}; \sqrt{1 - \beta_t} \cdot x^{(t-1)}, \mathbf{I}\beta_t) \\
x^{(t)} &= \sqrt{1 - \beta_t} \cdot x^{(t-1)} + \beta_t \cdot \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{I})
\end{aligned}$$

这里的 β_t 是自定义的，你可以选择线性增长或者三角函数型增长。通过每一步加噪，我们将分布的均值逐渐往0逼近，并且方差越来越大，最后形成一个 $\mathcal{N}(x^{(T)}; 0, \mathbf{I})$ 的标准正态分布。

反向过程中需要学习两个参数：

$$p(x^{(t-1)}|x^{(t)}) = \mathcal{N}(x^{(t-1)}; f_\mu(x^{(t)}; t), f_\Sigma(x^{(t)}; t))$$

即每一次反向过程中的均值和方差。

简化表达 (DDPM开始发力)

1. α 的引入

很显然，如果我们真的按照每一步加噪、去噪来训练模型的话，那么模型训练的时间相当长，因为每一次我都要以上一次加噪（去噪）的结果为输入，因此在前面的时间步进行训练的时候，后面的时间步只能等待。

我们发现从 $x^{(0)}$ 到 $x^{(t)}$ 的累积过程可以递归展开：

$$\begin{aligned}x^{(1)} &= \sqrt{1 - \beta_1} \cdot x^{(0)} + \beta_1 \cdot \epsilon_1 \\x^{(2)} &= \sqrt{1 - \beta_2} \cdot x^{(1)} + \beta_2 \cdot \epsilon_2 \\&= \sqrt{1 - \beta_2} \cdot \left(\sqrt{1 - \beta_1} \cdot x^{(0)} + \beta_1 \cdot \epsilon_1 \right) + \beta_2 \cdot \epsilon_2 \\&= \sqrt{(1 - \beta_2)(1 - \beta_1)} \cdot x^{(0)} + \sqrt{(1 - \beta_2)\beta_1} \cdot \epsilon_1 + \sqrt{\beta_2} \cdot \epsilon_2\end{aligned}$$

由于 ϵ_1, ϵ_2 相互独立并且服从 $\mathcal{N}(0, \mathbf{I})$ ，因此我们可以把他们俩合并成一个高斯分布：

$$\begin{aligned}\left(\sqrt{(1 - \beta_2)\beta_1} \cdot \epsilon_1 + \sqrt{\beta_2} \cdot \epsilon_2 \right) &\sim \mathcal{N}(0, [(1 - \beta_2)\beta_1 + \beta_2]\mathbf{I}) \\&= \mathcal{N}(0, [1 - (1 - \beta_1)(1 - \beta_2)]\mathbf{I})\end{aligned}$$

接下来就可以推广到 t 步：定义 $\alpha_t = 1 - \beta_t$ ， $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ ，则 $x^{(t)}$ 可以表示为：

$$\begin{aligned}x^{(t)} &= \sqrt{\bar{\alpha}_t} \cdot x^{(0)} + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \\q(x^{(t)}|x^{(0)}) &= \mathcal{N}\left(x^{(t)}; \sqrt{\bar{\alpha}_t} \cdot x^{(0)}, (1 - \bar{\alpha}_t)\mathbf{I}\right)\end{aligned}$$

这个表达式仅依赖于 $\bar{\alpha}_t$ ，即 α_t 的累积乘积，避免了逐步计算噪声项的复杂性。

2. D_{KL} 在高斯分布条件下的简化与损失函数

上文已经给出，我们的目标是最小化 $D_{KL}(q(x^{(t-1)}|x^{(t)}, x^{(0)}) \parallel p_\theta(x^{(t-1)}|x^{(t)}))$ 使得 K 尽可能达到最大，其中 θ 表示神经网络的参数

对于两个高斯分布 $\mathcal{N}(\mu_1; \sigma^2 \mathbf{I}), \mathcal{N}(\mu_2; \sigma^2 \mathbf{I})$ ，他们之间的 D_{KL} 可以用以下方式表示：

$$D_{KL}(\mathcal{N}(\mu_1; \sigma^2 \mathbf{I}) \parallel \mathcal{N}(\mu_2; \sigma^2 \mathbf{I})) = \frac{1}{2\sigma^2} \|\mu_1 - \mu_2\|^2$$

接下来我们需要计算前向过程的真实后验分布，即 $q(x^{(t-1)}|x^{(t)}, x^{(0)})$ ，它也遵循高斯分布，不过比较复杂：

$$q(x^{(t-1)}|x^{(t)}, x^{(0)}) = \mathcal{N}\left(x^{(t-1)}; \tilde{\mu}_t(x^{(t)}, x^{(0)}), \tilde{\beta}_t \mathbf{I}\right)$$

在这里，均值和方差都有解析解，接下来推导其过程：

目标是求 $q(x^{(t-1)}|x^{(t)}, x^{(0)})$ ，即已知 $x^{(t)}$ 和 $x^{(0)}$ 时， $x^{(t-1)}$ 的条件分布。

根据贝叶斯定义：

$$q(x^{(t-1)}|x^{(t)}, x^{(0)}) = \frac{q(x^{(t)}|x^{(t-1)}, x^{(0)}) \cdot q(x^{(t-1)}|x^{(0)})}{q(x^{(t)}|x^{(0)})}$$

由于前向过程是马尔可夫链， $q(x^{(t-1)}|x^{(t)}, x^{(0)}) = q(x^{(t-1)}|x^{(t)})$ ，因此：

$$q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) \propto q\left(x^{(t)}|x^{(t-1)}\right) \cdot q\left(x^{(t-1)}|x^{(0)}\right)$$

接下来把各个高斯分布显式地写出：

1. 前向单步转移概率：

$$q\left(x^{(t)}|x^{(t-1)}\right) \propto \exp\left(-\frac{\left\|x^{(t)} - \sqrt{\alpha_t}x^{(t-1)}\right\|^2}{2\beta_t}\right)$$

2. 从 $x^{(0)}$ 到 $x^{(t-1)}$ 的累积分布：

$$q\left(x^{(t-1)}|x^{(0)}\right) \propto \exp\left(-\frac{\left\|x^{(t-1)} - \sqrt{\bar{\alpha}_{t-1}}x^{(0)}\right\|^2}{2(1 - \bar{\alpha}_{t-1})}\right)$$

3. 联合分布的指数项，即后验分布的指数项为：

$$-\frac{\left\|x^{(t)} - \sqrt{\alpha_t}x^{(t-1)}\right\|^2}{2\beta_t} - \frac{\left\|x^{(t-1)} - \sqrt{\bar{\alpha}_{t-1}}x^{(0)}\right\|^2}{2(1 - \bar{\alpha}_{t-1})}$$

将指数项展开并合并同类项之后可以得到关于 $x^{(t-1)}$ 的二次项系数和一次项系数：

1. 二次项系数：

$$-\frac{1}{2}\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}}\right)\left(x^{(t-1)}\right)^2$$

2. 一次项系数：

$$-\frac{1}{2}\left(\frac{\sqrt{\alpha_t}}{\beta_t}x^{(t)} + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}}x^{(0)}\right)x^{(t-1)}$$

那么分别对应到一个典型高斯分布的指数项形式：

$$-\frac{(x - \mu)^2}{2\sigma^2} = -\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}$$

通过对比，就可以得到后验分布的均值和方差：

1. 方差：

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

2. 均值：

$$\frac{\tilde{\mu}_t}{\tilde{\beta}_t} = \frac{\sqrt{\alpha_t}}{\beta_t}x^{(t)} + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}}x^{(0)} \Rightarrow \tilde{\mu}_t = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}x^{(t)} + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}x^{(0)}$$

再根据前向过程 $x^{(t)} = \sqrt{\bar{\alpha}_t} \cdot x^{(0)} + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon$ ，解得：

$$x^{(0)} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x^{(t)} - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon\right)$$

将 $x^{(0)}$ 带入均值表达式可得：

$$\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}\left(x^{(t)} - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon\right)$$

因此，如果神经网络预测的反向过程的均值 μ_θ 为：

$$\mu_{\theta}\left(x^{(t)}, t\right)=\frac{1}{\sqrt{\bar{\alpha}_t}}\left(x^{(t)}-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_{\theta}\left(x^{(t)}, t\right)\right)$$

则 D_{KL} 就可以简化为：

$$\begin{aligned} D_{KL}\left(q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) \| p_{\theta}\left(x^{(t-1)}|x^{(t)}\right)\right) \\ =\frac{1}{2 \sigma^2}\left\|\tilde{\mu}_t-\mu_{\theta}\left(x^{(t)}, t\right)\right\|^2 \\ =\frac{\beta_t^2}{2 \tilde{\beta}_t \alpha_t\left(1-\bar{\alpha}_t\right)}\left\|\epsilon-\epsilon_{\theta}\left(x^{(t)}, t\right)\right\|^2 \end{aligned}$$

忽略权重系数之后，最终损失函数为：

$$\mathcal{L}_t(\theta)=\mathbf{E}_{x^{(0)}, \epsilon, t}\left[\left\|\epsilon-\epsilon_{\theta}\left(x^{(t)}, t\right)\right\|^2\right]$$

- 高斯假设：**前向和反向过程均为高斯分布，且方差固定。
- 均值匹配：**KL散度简化为均值差的MSE。
- 噪声预测：**通过参数化技巧，将均值预测转换为对噪声 ϵ 的直接预测。
- 损失函数：**最终损失函数为预测噪声与真实噪声的MSE。

代码思路

编写扩散模型代码的主要步骤包括：

- 定义前向过程的噪声调度 β_t 。
- 实现前向过程的加噪步骤。
- 设计反向过程的神经网络模型(U-Net)。
- 定义损失函数（基于KL散度）。
- 实现训练循环，包括随机采样时间步、前向加噪、反向去噪、计算损失和参数更新。
- 实现采样过程，从噪声中逐步生成样本。

1. 时间步骤嵌入的作用

在扩散模型中，前向过程包含多个时间步（如 $T=1000$ ），每个时间步 t 对应不同的噪声强度（由 β_t 控制）。反向过程需要知道当前处于哪个时间步，以调整去噪的幅度。例如，在时间步 $t=500$ 和 $t=900$ ，模型需要去除不同程度的噪声。时间嵌入（Time Embedding）通过将时间步 t 编码为向量，输入到网络中，使模型能够区分不同时间步的噪声分布。

关键点：时间嵌入让模型感知当前去噪的进度，避免对所有时间步使用相同的参数。

2. U-Net的作用与噪声预测

U-Net的核心功能

在扩散模型中，U-Net的直接目标是预测前向过程中添加的噪声，而非直接生成去噪后的数据。其设计特点如下：

- 编码器-解码器结构：**
 - 编码器：**通过卷积和下采样提取多尺度特征（如边缘、纹理等全局信息）。
 - 解码器：**通过上采样和跳跃连接恢复细节（如局部像素级信息）。
 - 跳跃连接：**将编码器的高分辨率特征直接传递到解码器，避免信息丢失。

2. 时间嵌入的融合：

- 时间步 t 被编码为向量，并与特征图结合，使模型能够区分不同噪声强度下的去噪策略。

全局特征与细节恢复的含义

- 全局特征：**编码器提取的高层语义信息（如物体形状、整体布局），帮助模型理解图像的整体结构。
- 细节恢复：**解码器通过上采样逐步重建像素级细节，确保生成图像的局部一致性。

为什么预测噪声而非直接去噪？

1. 理论推导的简化：

根据ELBO优化目标，反向过程的KL散度可转换为预测噪声的均方误差（MSE）：

$$\mathcal{L} = \mathbf{E}_{t, x^{(0)}, \epsilon} \left[\left\| \epsilon - \epsilon_{\theta} \left(x^{(t)}, t \right) \right\|^2 \right]$$

直接预测噪声与理论目标一致，且计算更稳定。

2. 训练稳定性：

预测噪声的尺度与输入噪声 ϵ 一致（均值为0，方差为1），而直接预测去噪后的数据可能涉及复杂的分布变换。

3. 隐式数据生成：

通过逐步去除预测的噪声，最终生成的 $x^{(0)}$ 自然满足数据分布，无需显式建模。