扩散模型

主讲人： 苏勤亮

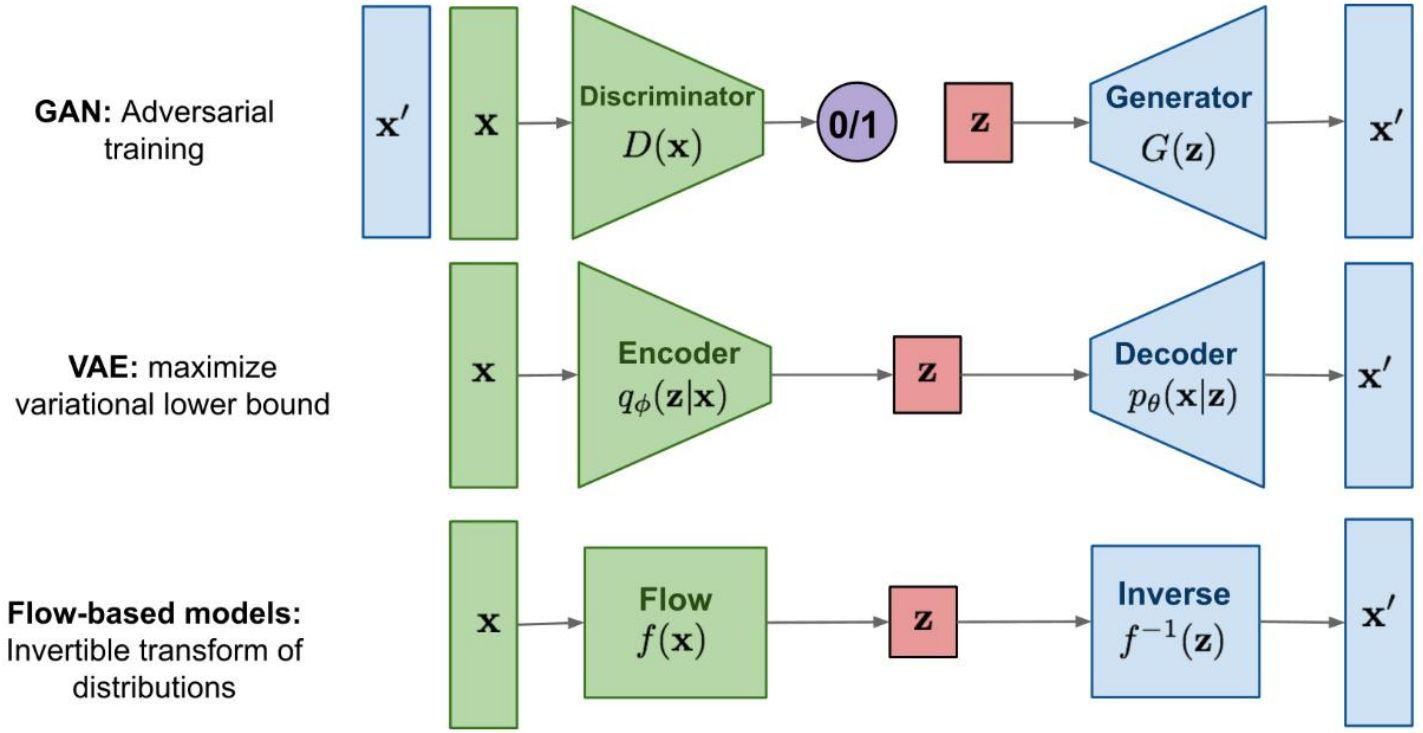
Outline

- 背景
- 扩散模型原理总览
- DDPM原理剖析

背景

生成模型

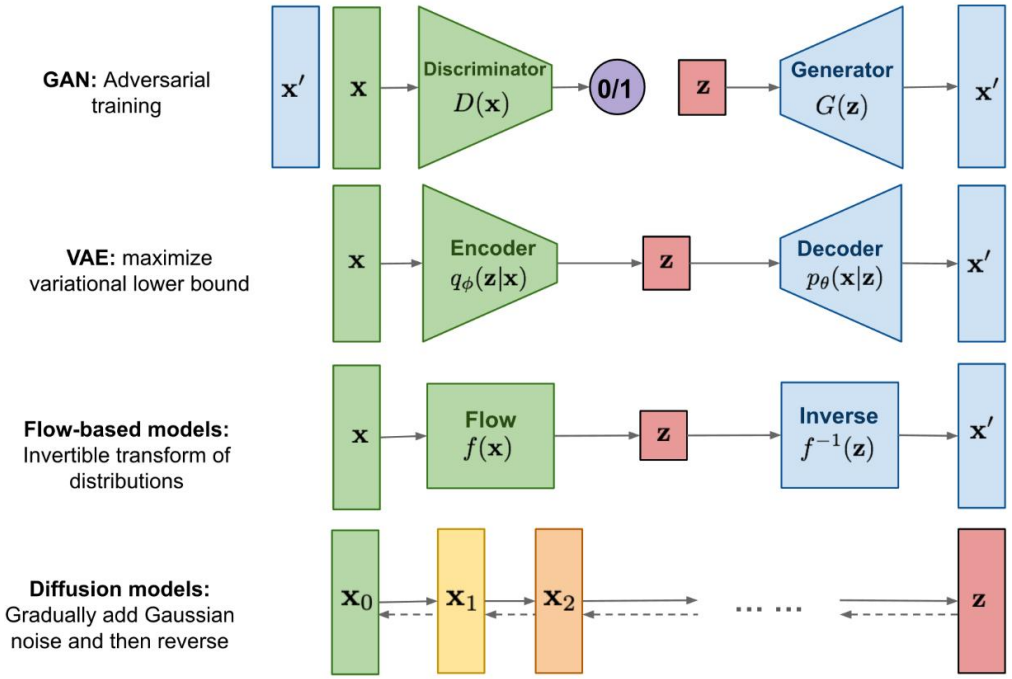
- GAN：训练不稳定，极难达到纳什均衡，对超参数敏感，训练容易崩塌
- VAE：生成图像模糊，均方误差和变分近似导致生成的细节丢失，清晰度不如GAN
- Flow-based Models：需要专门设计的可逆网络结构，参数量巨大，生成高维数据困难

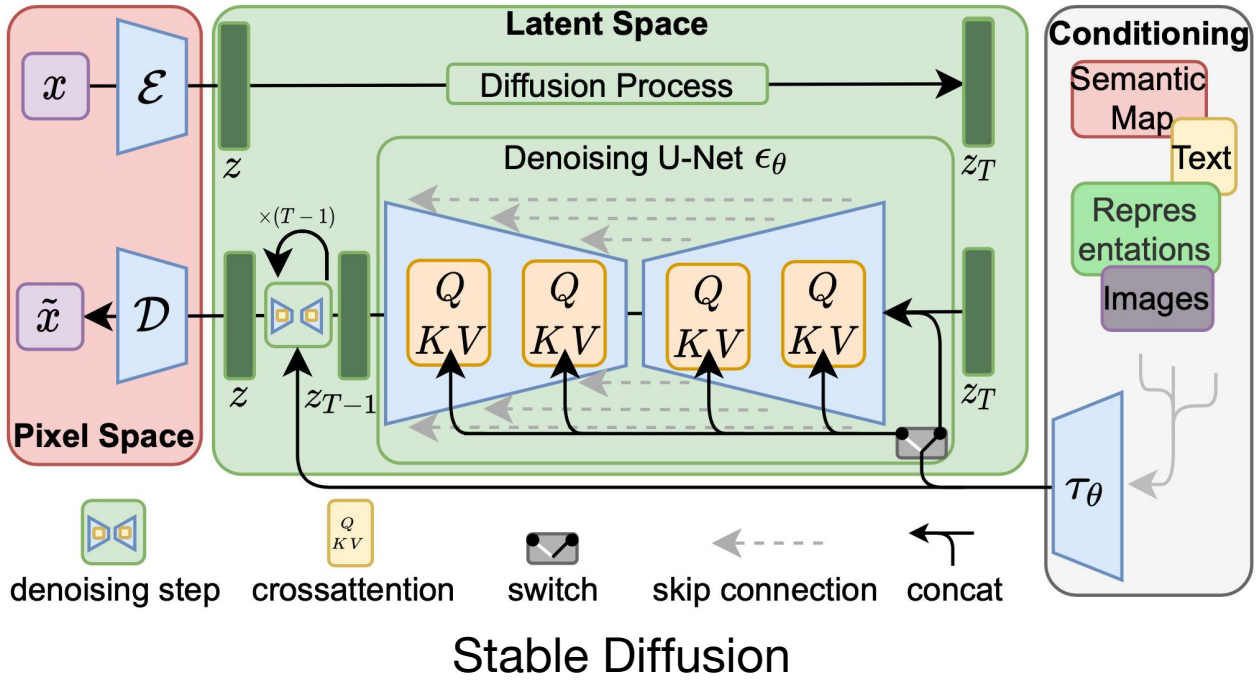


背景

❑ 扩散模型的优势

- **训练稳定性**：不同于GAN的对抗训练，扩散模型的训练本质上是回归问题
- **生成质量**：扩散模型的生成质量已超越GAN，能够生成极其精细的纹理和复杂的场景
- **可扩展性**：DALL-E 3, Stable Diffusion, Sora 等现代AIGC应用的基石





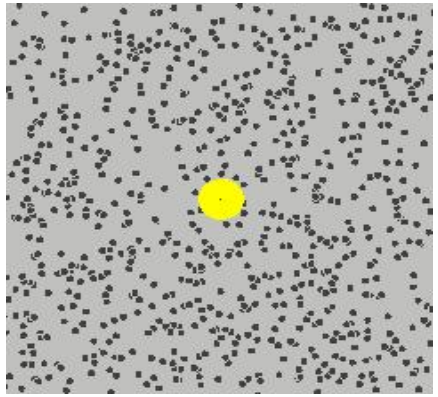
Stable Diffusion

Outline

- 背景
- 扩散模型原理总览
- DDPM原理剖析

扩散过程

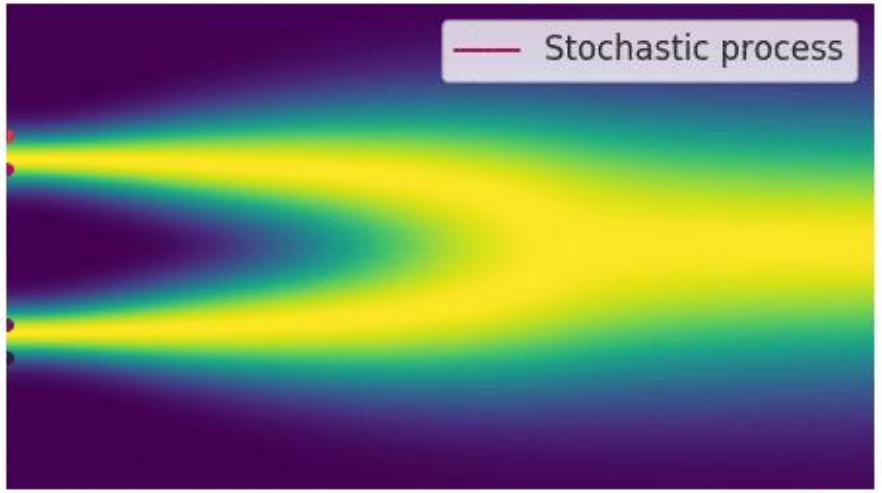
- 物理学中，扩散过程用于描述分子从高浓度区域向低浓度区域自发迁移、最终达到动态平衡的过程



- 统计学中，扩散过程指描述随机变量随时间演化的过程

$$X_{t+\Delta t} - X_t = \mu(X_t,\ t)\Delta t + \sigma(X_t,\ t)\Delta t \cdot \epsilon_t$$

  - ϵ_t: 高斯随机噪声 $\epsilon_t \sim \mathcal{N}(0,\ I)$
  - $\mu(X_t,\ t)$: 漂移项，表示系统的平均趋势



图像扩散过程

1000步加噪过程

给定图像$\boldsymbol{x}$，对其逐渐添加高斯噪声，同时减弱其强度

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t \qquad \text{其中 } \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{I})$$

其中，β_t表示每次加入噪声的强度，值非常小，如：0.01；$t = 1, 2, \cdots, 1000$

想一想，随着时间的增加，$\boldsymbol{x}_t$会逐渐变成什么？

高斯噪声，即：$\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$

扩散模型的基本原理

加噪过程

前向扩散过程：$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t$

训练一个模型，用于近似扩散过程的逆过程

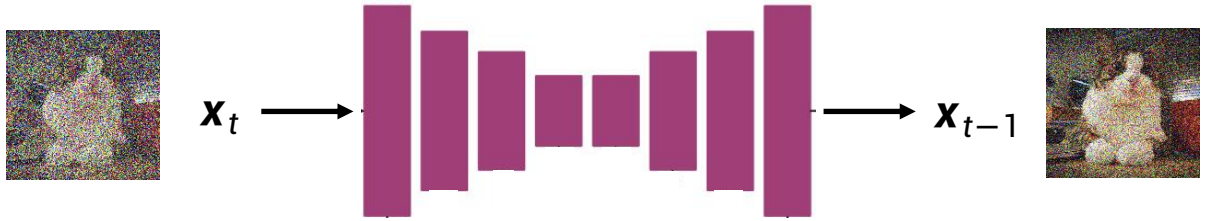
去噪过程

近似逆扩散过程：$\hat{\boldsymbol{x}}_{t-1} = G_\theta(\boldsymbol{x}_t)$

扩散模型的启发式训练方法（I）

近似逆扩散过程：$\hat{\boldsymbol{x}}_{t-1} = G_\theta(\boldsymbol{x}_t)$

- 思路一：训练一个神经网络，用$\boldsymbol{x}_t$直接预测$\boldsymbol{x}_{t-1}$



$$L = \|G_\theta(\boldsymbol{x}_t) - \boldsymbol{x}_{t-1}\|^2$$

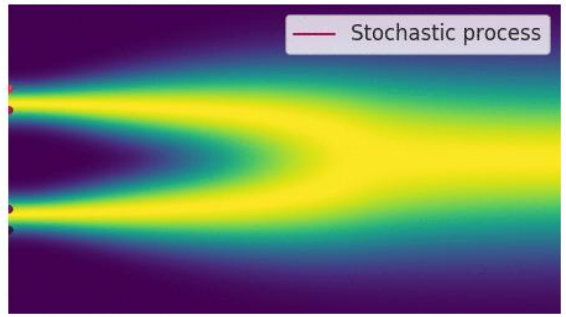
存在的问题：没有利用扩散过程蕴含的结构信息，模型被主要用于预测扩散过程中不可预测的随机性，导致最后效果不好

扩散模型的启发式训练方法（II）

- 思路二：充分利用扩散过程所蕴含的结构信息

  - 扩散过程 $\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t$实质上描述的是一个马尔科夫随机过程

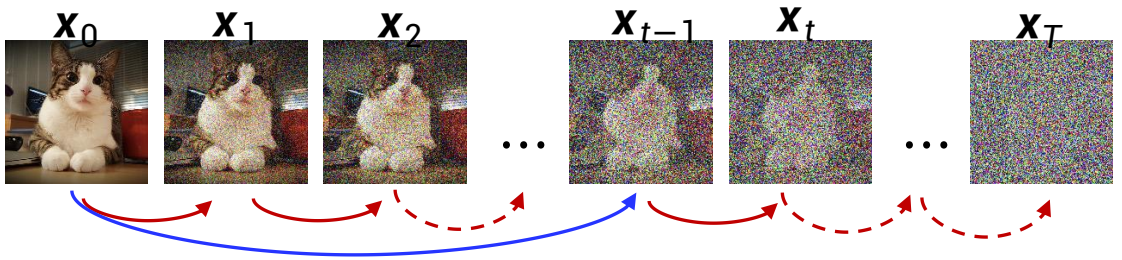
$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right)$$



  - 0到$t-1$时刻的条件转移概率

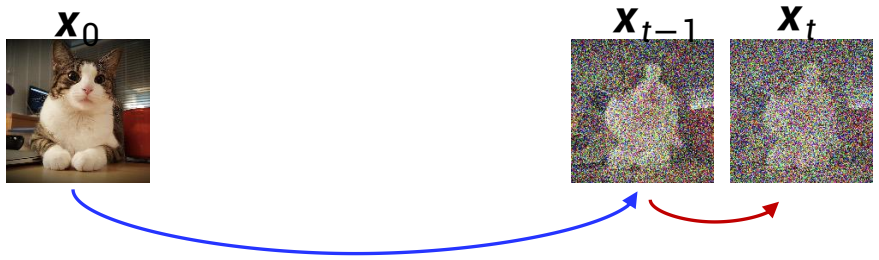
$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\overline{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\overline{\alpha}_{t-1})\boldsymbol{I}\right)$$

其中 $\overline{\alpha}_{t-1} \triangleq \prod_{s=1}^{t-1}(1-\beta_s)$



扩散模型的启发式训练方法（III）

扩散过程只指定了如下前向加噪过程



$$q(x_{t-1}|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_{t-1}}x_0, (1-\bar{\alpha}_{t-1})I\right)$$

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

逆扩散过程： 根据贝叶斯定理，在t时刻的状态 x_t已知时，$t-1$时刻状态x_{t-1}分布为

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_{t-1}|x_0) \times q(x_t|x_{t-1})}{q(x_t|x_0)} = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I\right)$$

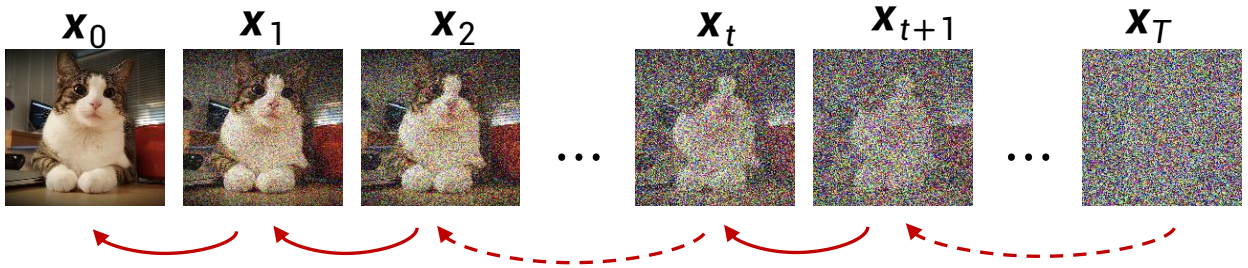
其中 $\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$ $\quad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$

扩散模型的启发式训练方法（IV）

后验分布 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \boldsymbol{I}\right)$ 揭示了扩散过程逆过程的采用过程

$$\boldsymbol{x}_{t-1} = \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) + \tilde{\beta}_t \cdot \boldsymbol{\epsilon}_t$$

- $\boldsymbol{\epsilon}_t$: 高斯随机噪声 $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{I})$



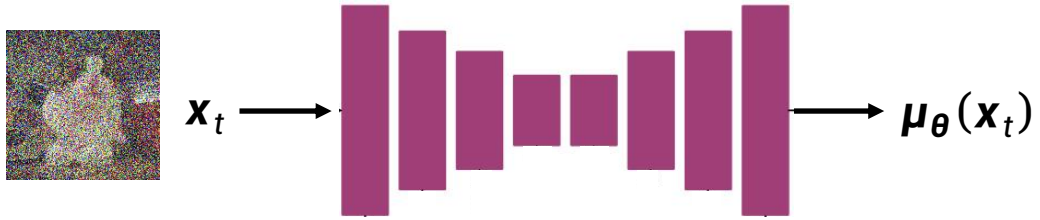
思考：是否可以使用上述采用过程来生成数据呢？

不可以，$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$含数据 $\boldsymbol{x}_0$，而在数据生成阶段，只有$\boldsymbol{x}_t$可使用

如何解决？

扩散模型的启发式训练方法（V)

解决思路：训练一个输入为x_t的神经网络，要求其输出值$\mu_\theta(x_t)$尽可能逼近 $\tilde{\mu}_t(x_t, x_0)$

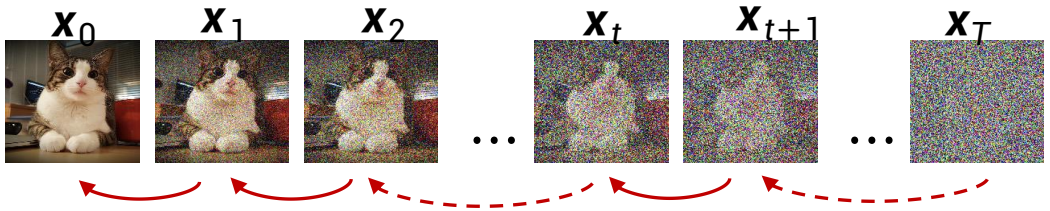


$$Loss = \sum_{x_0 \in \mathcal{D}} \|\mu_\theta(x_t) - \tilde{\mu}_t(x_t, x_0)\|^2$$

用$\mu_\theta(x_t)$代替 $\tilde{\mu}_t(x_t, x_0)$实现数据的采用

$$x_{t-1} = \tilde{\mu}_t(x_t, x_0) + \tilde{\beta}_t \cdot \epsilon_t$$

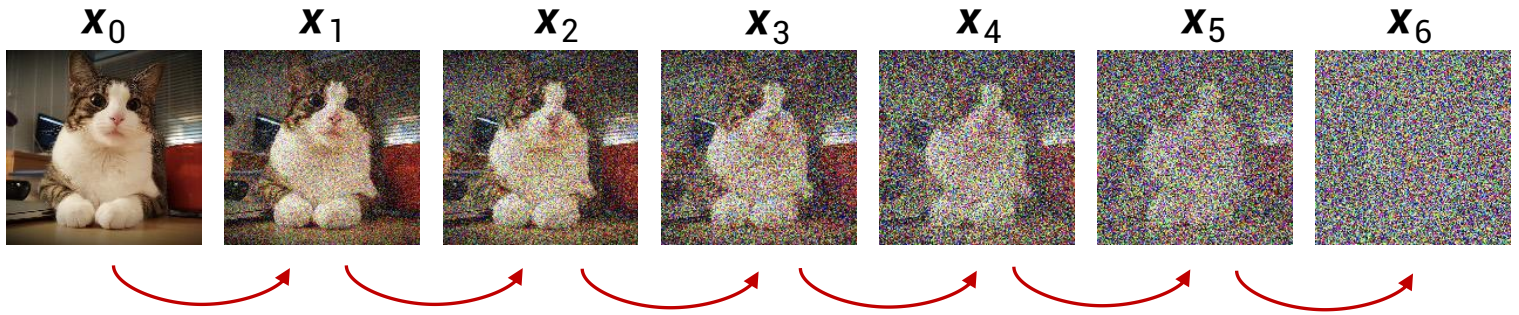
$$x_{t-1} = \mu_\theta(x_t) + \tilde{\beta}_t \cdot \epsilon_t$$



Outline

- 背景
- 扩散模型原理总览
- DDPM原理剖析

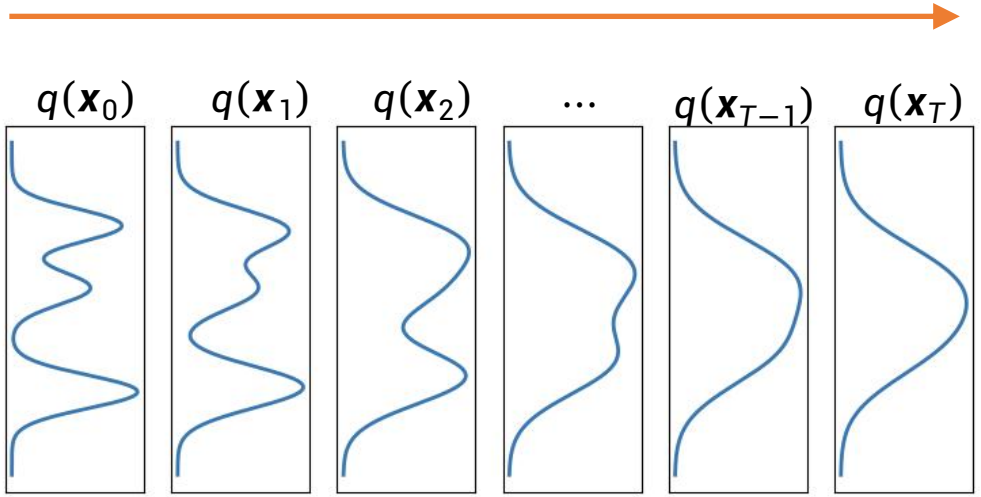
前向扩散过程



- 假设初始的数据分布为 $q(\boldsymbol{x}_0) = p_{data}(\boldsymbol{x})$
- 扩散过程可以由转移概率分布来描述

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right)$$

其中 β_t 通常是一个非常小的值，例如 0.01。



- 若已知 $\boldsymbol{x}_{t-1}$，则样本 $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ 可通过以下采样过程来获得：

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t \qquad \text{其中 } \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{I})$$

问题 1: 为什么我们要在 $\boldsymbol{x}_{t-1}$ 的前面乘上系数 $\sqrt{1-\beta_t}$？（防止方差爆炸）

问题 2: 当 T趋于无穷大时， $q(\boldsymbol{x}_T)$ 是什么分布？（正态分布）

- 若已知 $\boldsymbol{x}_{t-1}$，则样本 $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ 可通过以下采样过程来获得：

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t \qquad \text{其中 } \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{I})$$

问题 1: 为什么我们要在 $\boldsymbol{x}_{t-1}$ 的前面乘上系数 $\sqrt{1-\beta_t}$?

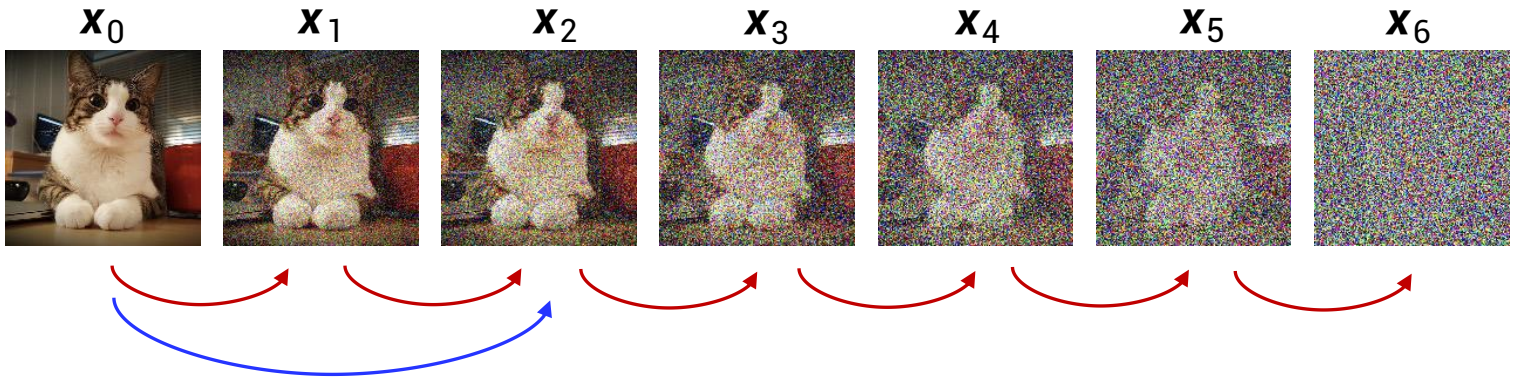
在扩散模型的前向过程中，我们在每一步都向图像中添加高斯噪声。为了确保数据在经过多次加噪后数值范围不会无限扩大（即防止方差爆炸），我们需要对原始信号$\boldsymbol{x}_{t-1}$进行缩放

假设$\boldsymbol{x}_{t-1}$的的方差已经归一化为 1，如果我们希望$\boldsymbol{x}_t$的方差保持为 1，计算如下

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

系数$\sqrt{1-\beta_t}$的作用是抵消加入噪声$\sqrt{\beta_t}\boldsymbol{\epsilon}_t$后带来的方差增加。如果不乘这个系数，随着$t$的增加，图像的像素值方差会越来越大（方差爆炸）

已知 $q(\boldsymbol{x}_1|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_1; \sqrt{1-\beta_1}\boldsymbol{x}_0, \beta_1 \boldsymbol{I})$



分布$q(\boldsymbol{x}_2|\boldsymbol{x}_0)$的形式是什么?

$$\boldsymbol{x}_2 = \sqrt{1-\beta_2}\boldsymbol{x}_1 + \sqrt{\beta_2}\boldsymbol{\epsilon}_2$$

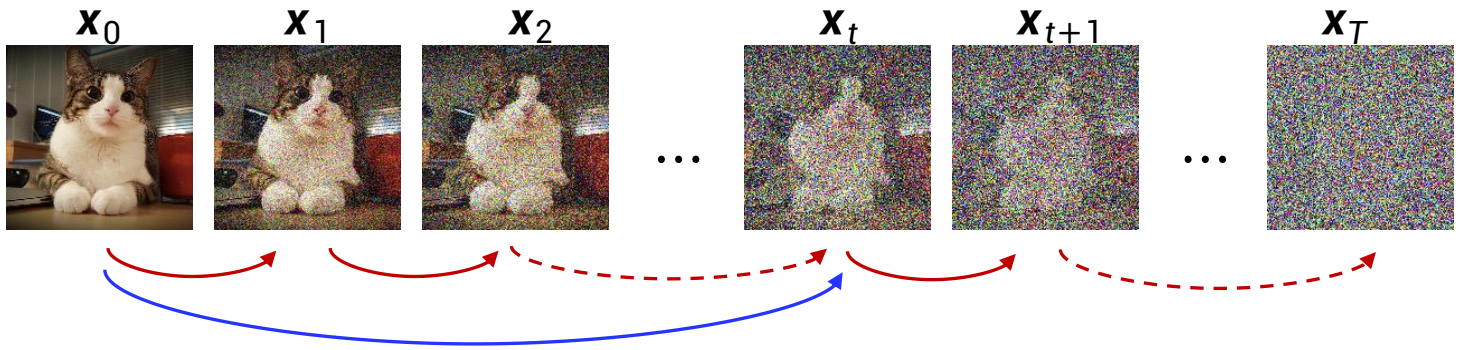
$$= \sqrt{1-\beta_2}\left(\sqrt{1-\beta_1}\boldsymbol{x}_0 + \sqrt{\beta_1}\boldsymbol{\epsilon}_1\right) + \sqrt{\beta_2}\boldsymbol{\epsilon}_2$$

$$= \sqrt{(1-\beta_2)(1-\beta_1)}\boldsymbol{x}_0 + \sqrt{(1-\beta_2)\beta_1}\boldsymbol{\epsilon}_1 + \sqrt{\beta_2}\boldsymbol{\epsilon}_2$$

$$\rightarrow \quad \mathbb{E}[\boldsymbol{x}_2] = \sqrt{(1-\beta_2)(1-\beta_1)}\boldsymbol{x}_0 \qquad Var[\boldsymbol{x}_2] = 1-(1-\beta_2)(1-\beta_1)$$

$$\Rightarrow \quad q(\boldsymbol{x}_2|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_2; \sqrt{(1-\beta_1)(1-\beta_2)}\boldsymbol{x}_0,\ \left(1-(1-\beta_2)(1-\beta_1)\right)\boldsymbol{I}\right)$$

那么分布 $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ 是什么形式?



$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\prod_{s=1}^{t}(1-\beta_s)}\boldsymbol{x}_0,\ \left(1-\prod_{s=1}^{t}(1-\beta_s)\right)\boldsymbol{I}\right)$$

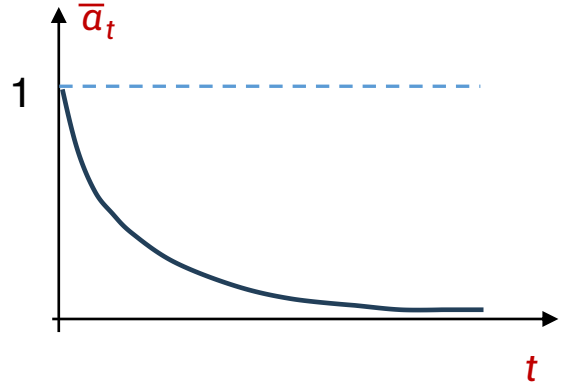
$$= \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0,\ (1-\overline{\alpha}_t)\boldsymbol{I}\right) \qquad \text{其中 } \overline{\alpha}_t \triangleq \prod_{s=1}^{t}(1-\beta_s)$$

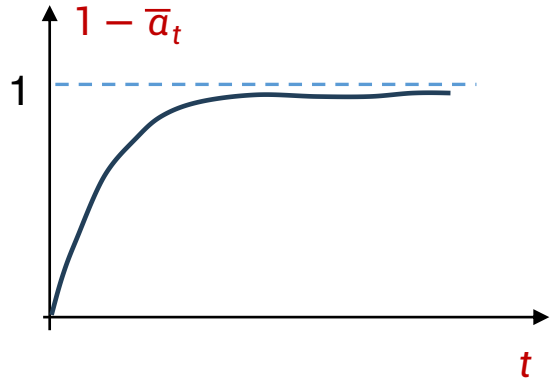
$$\boldsymbol{x}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon} \qquad \text{其中 } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

$\overline{\alpha}_t$ 的含义是什么?

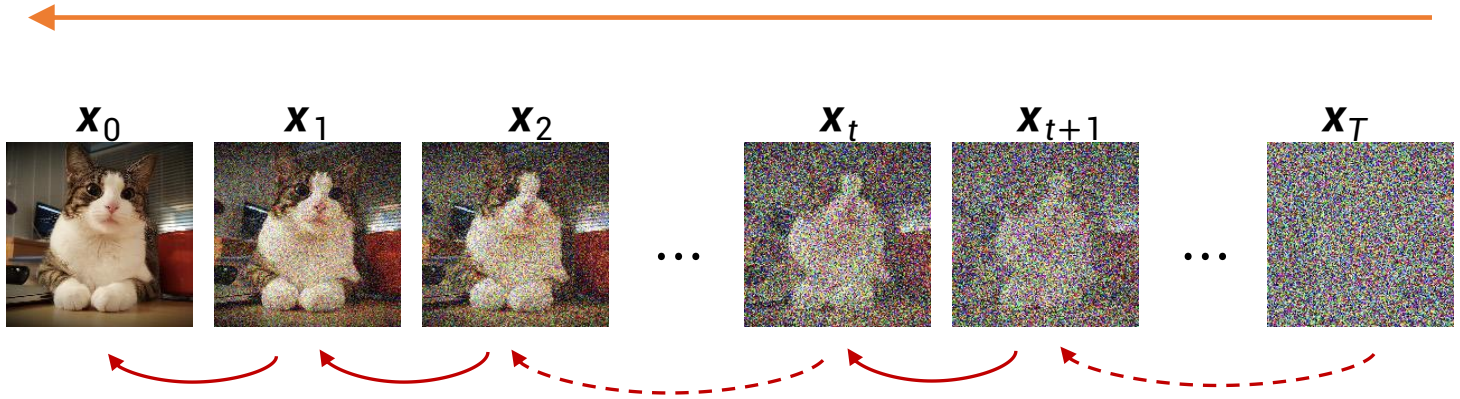
$1-\overline{\alpha}_t$ 代表图像 $\boldsymbol{x}_t$ 中噪声的强度（方差）

$\overline{\alpha}_t$ 代表图像 $\boldsymbol{x}_t$ 中原图像 $\boldsymbol{x}_0$ 的强度





通过逆转扩散过程进行生成



扩散过程规定了一个联合分布：

$$q(\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) = q(\boldsymbol{x}_0)q(\boldsymbol{x}_1|\boldsymbol{x}_0)\cdots q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\cdots q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})$$

由于其马尔可夫结构，该分布可以等价地以逆形式表示

$$q(\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) = q(\boldsymbol{x}_T)q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)\cdots q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\cdots q(\boldsymbol{x}_0|\boldsymbol{x}_1)$$

$$q(\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) = q(\boldsymbol{x}_T)q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)\cdots q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\cdots q(\boldsymbol{x}_0|\boldsymbol{x}_1)$$

已知当 $T \to +\infty$ 时，$q(\boldsymbol{x}_T) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 。因此，为了从分布 $q(\boldsymbol{x}_0)$ 中采集样本，我们可以通过以下过程进行采样：

1) 初始化 $\boldsymbol{x}_T \sim q(\boldsymbol{x}_T)$
2) 采样 $\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 对于 $t = T, T-1, \cdots, 1$
3) 保留 $\boldsymbol{x}_0$ ，同时舍弃样本 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T$

问题是如何得到转移概率分布 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

- 但是 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 是未知的，因为分布 $q(\boldsymbol{x}_0)$ 是未知的

$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$分布

- 尽管 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 未知，但在进一步给定 $\boldsymbol{x}_0$ 条件下，我们可以得到分布 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$
- 根据$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\overline{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\overline{\alpha}_{t-1})\boldsymbol{I})$及 $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\boldsymbol{I})$，我们可以得到联合分布：

$$q(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\boldsymbol{I}\right) \cdot \mathcal{N}\left(\boldsymbol{x}_{t-1}; \sqrt{\overline{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\overline{\alpha}_{t-1})\boldsymbol{I}\right)$$

- 从该联合概率密度函数中，我们可以得到后验分布（推导过程见后页补充材料）：

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t\boldsymbol{I}\right)$$

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

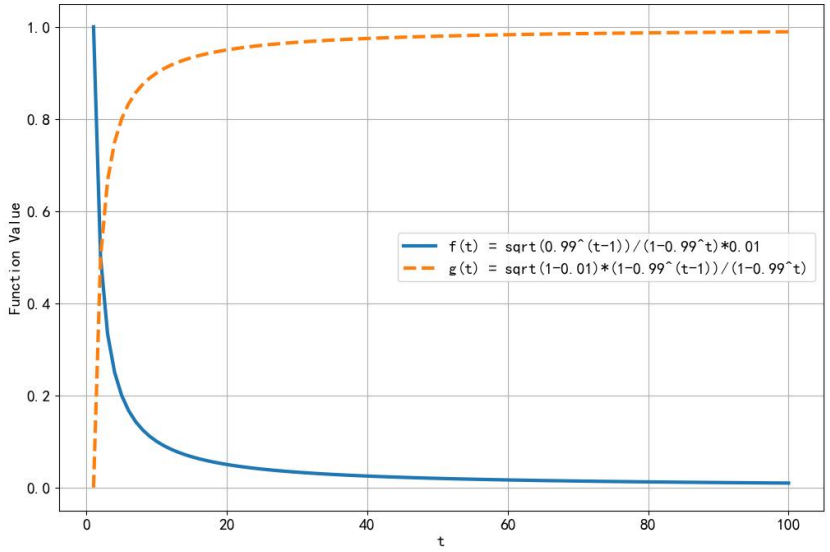
因此，后验分布的形式为

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \boldsymbol{I}\right)$$

其中

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

- 下图显示了当设置 $\beta_t = 0.01$ 时，系数 $\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}$ 和 $\frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$ 随 t 变化的函数关系



- 当 $t \geq 20$ 时，后验均值 $\tilde{\boldsymbol{\mu}}_t$ 主要由 $\boldsymbol{x}_t$ 决定
- 当 $t < 20$ 时，后验均值 $\tilde{\boldsymbol{\mu}}_t$ 由 $\boldsymbol{x}_t$ 和 $\boldsymbol{x}_0$ 共同决定

补充材料：使用“完全平方”技巧推导后验概率

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\propto\exp\left\{-\frac{1}{2\beta_t}\left(\boldsymbol{x}_t-\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}\right)^2\right\}\cdot\exp\left\{-\frac{1}{2(1-\bar{\alpha}_{t-1})}\left(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0\right)^2\right\}$$

$$\propto\exp\left\{-\frac{1}{2}\left[\left(\frac{1-\beta_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{1-\beta_t}\boldsymbol{x}_t}{\beta_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\}$$

➡ 后验分布 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ 的方差和均值为：

$$Var(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)=\frac{1}{\frac{1-\beta_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}}=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$$\mathbb{E}[\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0]=\frac{\frac{\sqrt{1-\beta_t}\boldsymbol{x}_t}{\beta_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}}{\frac{1-\beta_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}}=\frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t}{1-\bar{\alpha}_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\boldsymbol{x}_0}{1-\bar{\alpha}_t}$$

一种启发式生成方法

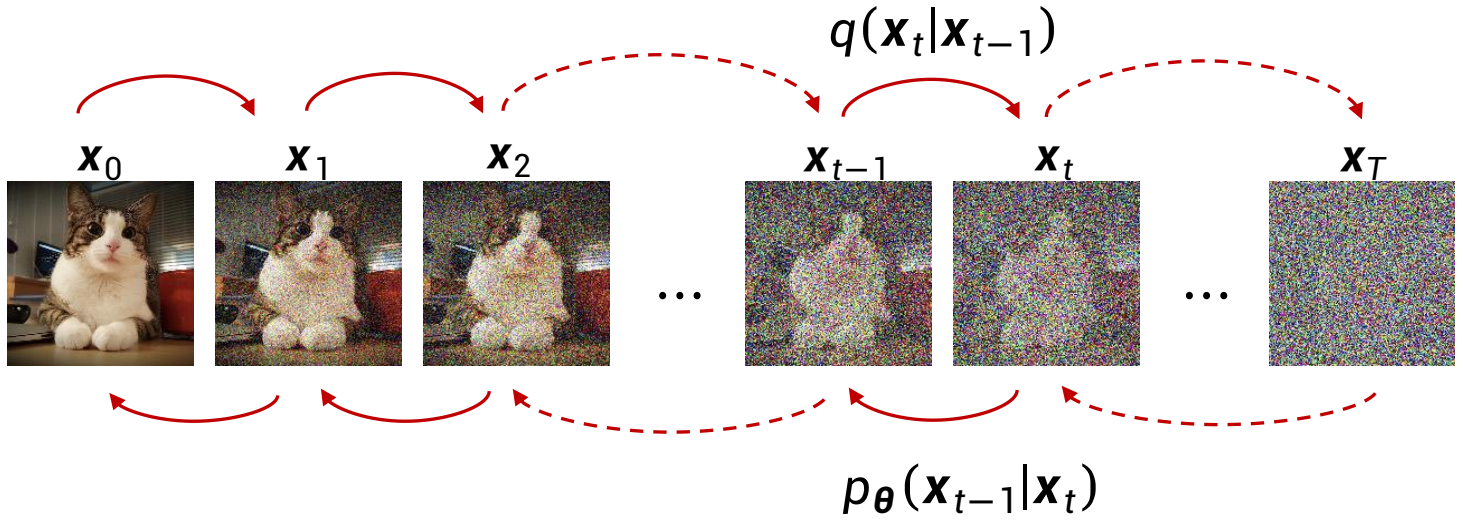
- 由于无法获得真实图像 $\boldsymbol{x}_0$ ，我们不能使用真实的后验分布 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ 来生成图像
- 一个直观的想法
  - 对于所有 $\boldsymbol{x}_0 \sim q(\boldsymbol{x})$，学习一个分布 $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 来近似真实的后验分布 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ ，其中$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$设定为如下形式

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \tilde{\beta}_t \boldsymbol{I}\right)$$

  - 使用 $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 来生成图像
- 为最小化$KL(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))$，可以验证等价于$\min \|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t) - \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)\|^2$

能否更严谨一点？

一个更严谨的视角



将 $t \geq 1$ 时的 $\boldsymbol{x}_t$ 视为隐变量，扩散生成模型就是一个多层隐变量模型，其分布为

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)p(\boldsymbol{x}_T)$$

其中 $p(\boldsymbol{x}_0|\boldsymbol{z}) \triangleq p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)$ 并且 $p(\boldsymbol{z}) \triangleq p_{\boldsymbol{\theta}}(\boldsymbol{x}_1|\boldsymbol{x}_2)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)p(\boldsymbol{x}_T)$

- 对于一个生成模型，一个广泛使用的训练目标是最大化对数似然

$$\frac{1}{|\mathcal{D}|}\sum_{\boldsymbol{x}_0\in\mathcal{D}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0)$$

或其变分下界（仅考虑单个样本 $\boldsymbol{x}_0$ ）

$$\mathbb{L}=\int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_0)\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_0,\,\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_0)}d\boldsymbol{z}$$

$$=\int q_{\boldsymbol{\phi}}(\boldsymbol{x}_1,\,\cdots,\,\boldsymbol{x}_T|\boldsymbol{x}_0)\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)p(\boldsymbol{x}_T)}{q_{\boldsymbol{\phi}}(\boldsymbol{x}_1,\,\cdots,\,\boldsymbol{x}_T|\boldsymbol{x}_0)}d\boldsymbol{z}$$

其中隐变量$\boldsymbol{z}\triangleq\{\boldsymbol{x}_1,\,\cdots,\,\boldsymbol{x}_T\}$

- 直接将近似后验分布$q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\boldsymbol{x}_0)$设定为如下扩散过程的分布

$$q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\boldsymbol{x}_0) = q(\boldsymbol{x}_1|\boldsymbol{x}_0)\cdots q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\cdots q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}),$$

- 变分下界可以写成如下形式

$$\mathfrak{l} = \int q(\boldsymbol{x}_1|\boldsymbol{x}_0)\cdots q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\cdots q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} d\boldsymbol{X}$$

- 这里，后验分布是固定且不可学习的，我们只能学习生成模型中的参数 $\boldsymbol{\theta}$

跟之前学习的VAE有什么区别？

变分下界$\mathfrak{l}$的表达式

- 通过一定概率变换操作，可推导变分下界的表达式为（具体见下页补充材料）：

$$\mathfrak{l} = \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)] - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[KL(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]}_{\mathfrak{l}_{t-1}} - KL(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$$

- 将 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \boldsymbol{I}\right)$ 和 $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \tilde{\beta}_t \boldsymbol{I}\right)$ 代入 KL 散度，可以得到

$$\mathfrak{l}_{t-1} = \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{1}{2\tilde{\beta}_t}\|\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|^2\right] + C$$

- 因此，最大化变分下界1等价于最大化

$$\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)] - \sum_{t=2}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[\frac{1}{2\tilde{\beta}_t} \|\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|^2 \right]$$

- 该目标与前述启发式训练目标（最小化 $\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t) - \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)\|^2$ ）非常类似，除了两点区别
  - 直观方法不包含数据项 $\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)]$
  - 直观方法没有指定如何选择 $\boldsymbol{x}_t$ ，而变分下界目标明确要求 $\boldsymbol{x}_t$ 来自于 $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ ，即 $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)$

补充材料：变分下界的推导过程

- 重写后验分布

$$q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\boldsymbol{x}_0) = q(\boldsymbol{x}_1|\boldsymbol{x}_0)\cdots q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\cdots q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})$$

$$= q(\boldsymbol{x}_T|\boldsymbol{x}_0)q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T, \boldsymbol{x}_0)\cdots q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\cdots q(\boldsymbol{x}_1|\boldsymbol{x}_2, \boldsymbol{x}_0)$$

$q(\boldsymbol{x}_T|\boldsymbol{x}_0)$ 和 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ 的分布是什么?

$$q(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_T; \sqrt{\overline{\alpha}_T}\boldsymbol{x}_0, (1-\overline{\alpha}_T)\boldsymbol{I}) \text{ 其中 } \overline{\alpha}_T \triangleq \prod_{s=1}^{T}(1-\beta_s)$$

$$\approx \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$$

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t\boldsymbol{I}\right)$$

补充材料：变分下界的推导过程 Cont'

将后验分布代入变分下界，我们得到

$$\mathcal{L} = \int q(\boldsymbol{x}_1|\boldsymbol{x}_2, \boldsymbol{x}_0)\cdots q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T, \boldsymbol{x}_0)q(\boldsymbol{x}_T|\boldsymbol{x}_0) \times \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)p_{\boldsymbol{\theta}}(\boldsymbol{x}_1|\boldsymbol{x}_2)\cdots p_{\boldsymbol{\theta}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_1|\boldsymbol{x}_2, \boldsymbol{x}_0)\cdots q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T, \boldsymbol{x}_0)q(\boldsymbol{x}_T|\boldsymbol{x}_0)} d\boldsymbol{x}_1 \cdots d\boldsymbol{x}_T$$

$$= \int q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\boldsymbol{x}_0) \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\, d\boldsymbol{X} \quad \Leftrightarrow \quad \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)]$$

$$-\sum_{t=2}^{T} \int q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\boldsymbol{x}_0) \log \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)} d\boldsymbol{X} \quad \Leftrightarrow \quad \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[KL(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

$$-\int q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\boldsymbol{x}_0) \log \frac{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}{p(\boldsymbol{x}_T)} d\boldsymbol{X} \quad \Leftrightarrow \quad KL(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$$

- 因此，变分下界可以写为

$$\mathcal{L} = \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)] - \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[KL(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))] - KL(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$$

重写 $\tilde{\mu}_t(x_t, x_0)$ 的表达式

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t$$

- 回顾 $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ 其中 $\epsilon \sim \mathcal{N}(0, I)$。将 $x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}}$ 代入 $\tilde{\mu}_t(x_t, x_0)$ 得到

$$\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

- 因此，我们可以将估计的均值 $\mu_\theta(x_t, t)$ 显式地约束为如下形式

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

- 也就是说，为估计$\tilde{\mu}_t$，我们只需训练一个神经网络 $\epsilon_\theta(x_t, t)$来估计所添加的噪声 ϵ

重新表示 $\mathfrak{l}_{t-1}$

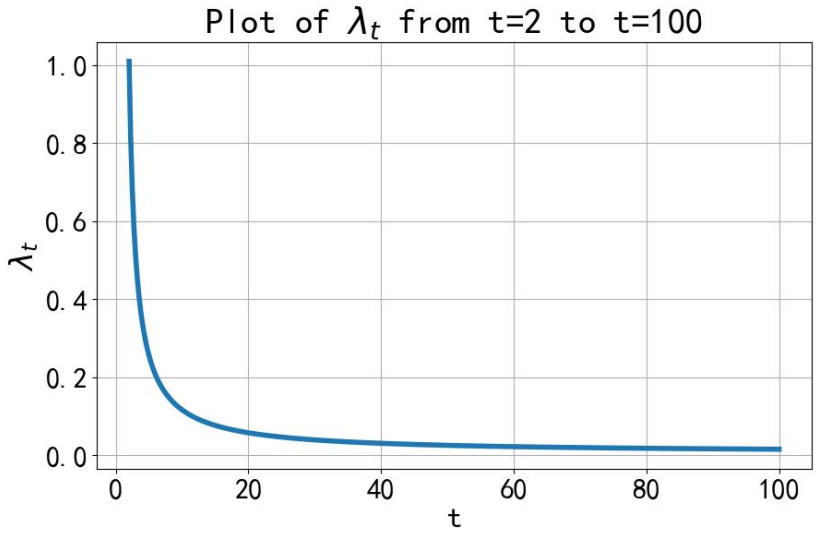
- 将 $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ 和 $\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$ 的表达式代入 $\mathfrak{l}_{t-1}$ 得到

$$\mathfrak{l}_{t-1} = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0}, \boldsymbol{I})}\left[\lambda_t \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\underbrace{\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}_{\boldsymbol{x}_t}, t\right)\right\|^2\right] + C$$

其中 $\lambda_t \triangleq \dfrac{\beta_t^2}{2\tilde{\beta}_t\,(1-\beta_t)(1-\bar{\alpha}_t)} = \dfrac{\beta_t}{(1-\beta_t)(1-\bar{\alpha}_{t-1})}$

当最小化损失函数 $\tilde{\mathfrak{l}} = \sum_{t=1}^{T} \mathfrak{l}_{t-1}$ 时，我们实际上是在尝试最大化对数似然 $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0)$

系数 λ_t 过分重视了较小的 t 值



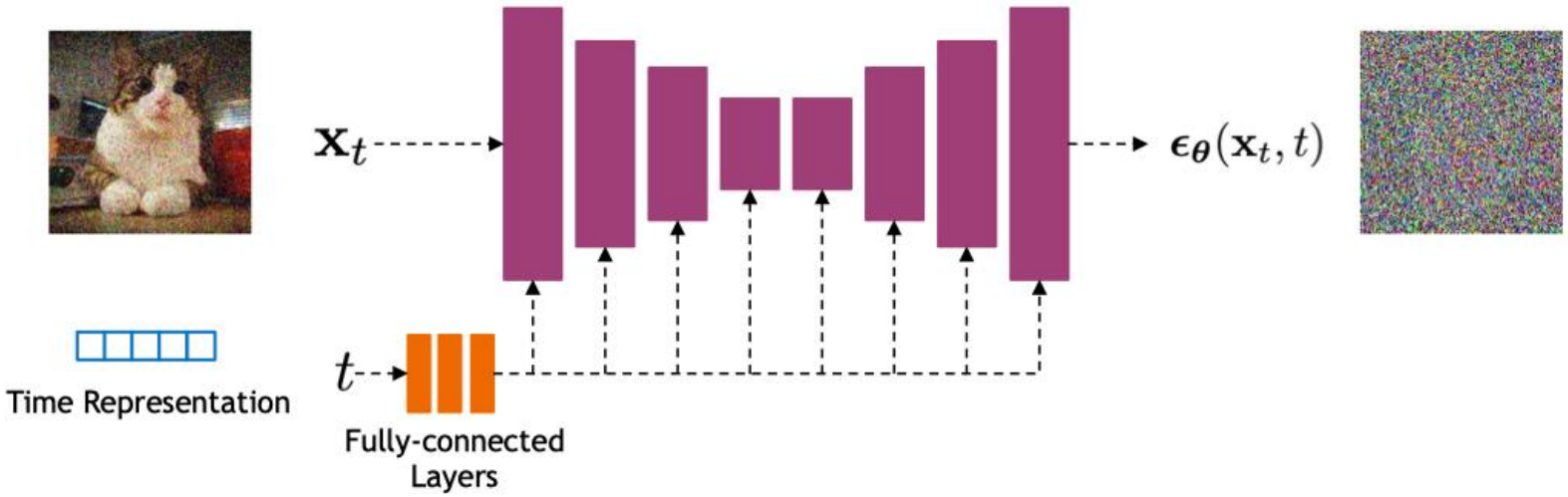
研究发现，通过简单地将所有 t 的 λ_t 设置为 1，我们可以生成更高质量的样本。

也就是说，最大化以下损失函数

$$\mathcal{L}_{simple} = \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0),\, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}),\, t \sim U(1,T)} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}} \left(\sqrt{\overline{\alpha}_t} \boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t} \boldsymbol{\epsilon},\ t \right) \right\|^2 \right]$$

$\epsilon_\theta(x_t,\ t)$ 的神经网络架构

使用包含 ResNet 模块的 U-Net 架构来表示 $\epsilon_\theta(x_t,\ t)$



- 时间步编码：正弦函数位置编码或随机傅里叶特征
- 时间特征通过空间加法或自适应组归一化（AGN）输入到残差块中

训练过程

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

采样过程

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right) + \sqrt{\tilde{\beta}_t}\boldsymbol{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

课堂小结

- 扩散模型原理总览
- DDPM原理剖析