Denoising Diffusion Implicit Models

xinli*

September 15, 2024

Abstract

这里记录了我在学习 DDIM 过程中的一些过程和思考。(仅供参考)

1 Preliminary

在开始学习之前，我们先看一些需要用到的知识，帮助我们后面更好的理解内容。

1.1 边缘和条件高斯分布

给定 x 的边缘高斯分布和给定 x 条件下 y 的条件高斯分布，它们的形式如下：

$p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$ 表示 x 的边缘分布为高斯分布，均值为 μ，协方差矩阵为 Λ^{-1} 的逆矩阵。$p(y \mid x) = \mathcal{N}(y \mid Ax + b, L^{-1})$ 表示在给定 x 的条件下 y 的条件分布也是高斯分布，均值为 $Ax + b$，协方差矩阵为 L^{-1} 的逆矩阵。

我们可以推导得到：y 的边缘分布 $p(y)$ 和在给定 y 条件下 x 的条件分布 $p(x \mid y)$ 分别为：$p(y) = \mathcal{N}(y \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^\top)$ 表示 y 的边缘分布依然是高斯分布，均值为 $A\mu + b$，方差为 $L^{-1} + A\Lambda^{-1}A^\top$。$p(x \mid y) = \mathcal{N}(x \mid \Sigma\{A^\top L(y - b) + \Lambda\mu\}, \Sigma)$，其中 $\Sigma = (\Lambda + A^\top LA)^{-1}$

1.2 简单的数学归纳法

题目： 证明不等式 $5^n + 3^n \geq 2^{2n+1}$ 对于所有 $n \in \mathbb{Z}^+$ 成立。

解答： 我们使用数学归纳法来证明该不等式。

步骤 1：基础步骤 (n = 1)

首先，我们检查当 $n = 1$ 时不等式是否成立：

$$5^1 + 3^1 = 5 + 3 = 8 \tag{1}$$

$$2^{2(1)+1} = 2^3 = 8 \tag{2}$$

显然，$8 \geq 8$，所以 $n = 1$ 时不等式成立。

步骤 2：归纳假设

假设当 $n = k$ 时不等式成立，即

$$5^k + 3^k \geq 2^{2k+1} \tag{3}$$

*https://github.com/xinli2008

步骤 3：归纳步骤 ($\mathbf{n = k + 1}$)

我们需要证明当 $n = k + 1$ 时不等式也成立，即证明

$$5^{k+1} + 3^{k+1} \geq 2^{2(k+1)+1} \tag{4}$$

注意到：

$$\begin{aligned} 5^{k+1} + 3^{k+1} &= 5 \cdot 5^k + 3 \cdot 3^k \\ &= (4+1) \cdot 5^k + (4-1) \cdot 3^k \\ &= 4 \cdot (5^k + 3^k) + 5^k - 3^k \\ &\geq 2^2 \cdot 2^{2k+1} + 5^k - 3^k \\ &\geq 2^{2(k+1)+1} \end{aligned}$$

因此，对于 $n = k + 1$ 时，不等式也成立。

结论：根据数学归纳法原理，$5^n + 3^n \geq 2^{2n+1}$ 对于所有正整数 n 都成立。

2 Introduction

2.1 DDIM 的发现

在 DDIM 的论文中，作者提到了一个核心发现，即**DDPM 的损失函数 *Loss* 只依赖于边缘分布 $q(x_t \mid x_0)$，而不直接依赖于联合分布 $q(x_{1:T} \mid x_0)$**。也就是说，联合分布 $q(x_{1:T} \mid x_0)$ 以什么形式出现，并不会影响我们训练 DDPM 的 loss 函数。在 DDPM 的论文中，联合分布 $q(x_{1:T} \mid x_0)$ 是可以拆散成很多条件分布相乘的，即 $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$，这是因为 DDPM 的前向和反向过程都是基于马尔可夫假设的。

于是，DDIM 的作者开始思考，是否存在着非马尔可夫的加噪过程呢？也就是说，如果**我们设计出一种非马尔可夫的扩散过程，并且我们只需要保证 $q(x_t \mid x_0)$ 和 DDPM 中的 $q(x_t \mid x_0)$ 是一样的，而 $q(x_{1:T} \mid x_0)$ 和 DDPM 的是不一样的，那么我们就可以直接复用训练好的 DDPM 模型，然后使用新的概率分布来进行逆过程的采样，来完成推理过程的加速**。

2.2 设计非马尔可夫链的扩散过程

DDIM 的团队设计了一种非马尔可夫链的分布，前向扩散过程如下：

$$q_\sigma(x_{1:T}|x_0) := q_\sigma(x_T|x_0) \prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t, x_0) \tag{5}$$

其中，σ 是超参数，我们可以自己设置。另外，$q_\sigma(x_T|x_0) = \mathcal{N}(\sqrt{\alpha_T}x_0, (1-\alpha_T)\mathbf{I})$。当 $t > 1$ 时，$q_\sigma(x_{t-1}|x_t, x_0)$ 的计算公式如下：

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\mathbf{I}\right) \tag{6}$$

按照我们之前所说的，**只有当这个分布在任意时刻的 $q(x_t \mid x_0)$ 都要和 DDPM 中的 $q(x_t \mid x_0)$ 一致，我们才可以复用 DDPM 模型**。虽然作者提到了，$q_\sigma(x_T|x_0) = \mathcal{N}(\sqrt{\alpha_T}x_0, (1-\alpha_T)\mathbf{I})$，但是我们接下来仍然要证明，这个公式对于任意的 t 都成立。

题目： 在给定公式(5)和公式(6)的前提下，证明 $q_\sigma(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I})$ 对于任意时刻的 t 成立。

解答： 我们使用数学归纳法来证明该不等式。

对于 T 时刻，$q_\sigma(x_T|x_0) = \mathcal{N}(\sqrt{\alpha_T}x_0, (1-\alpha_T)\mathbf{I})$ 已经成立。（作者给出的条件）

我们假设 t 时刻时，$q_\sigma(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I})$

那么，对于时刻 $t-1$ 来说，首先

$$q_\sigma(x_{t-1}|x_0) := \int_{x_t} q_\sigma(x_t|x_0)q_\sigma(x_{t-1}|x_t, x_0)dx_t \tag{7}$$

$$q_\sigma(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I}) \tag{8}$$

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\mathbf{I}\right) \tag{9}$$

接下来，根据(1.1)中的公式，我们已经有了 $q_\sigma(x_t|x_0)$ 和 $q_\sigma(x_{t-1}|x_t, x_0)$，那么我们可以得到 $q_\sigma(x_{t-1}|x_0) = \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$，其中：

$$\mu_{t-1} = \sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\sqrt{\alpha_t}x_0 - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}} = \sqrt{\alpha_{t-1}}x_0 \tag{10}$$

$$\Sigma_{t-1} = \sigma_t^2\mathbf{I} + \frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}(1-\alpha_t)\mathbf{I} = (1-\alpha_{t-1})\mathbf{I} \tag{11}$$

所以，我们可以得到：

$$q_\sigma(x_{t-1}|x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}x_0, (1-\alpha_{t-1})\mathbf{I}) \tag{12}$$

到此，我们已经证明对于任何时刻的 t，

$$q_\sigma(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I}) \tag{13}$$

这个分布形式是和 DDPM 中的分布形式一致的，那我们就可以继续沿用 DDPM 中的 loss 函数来训练模型。

2.3 对比非马尔可夫扩散后验分布于 DDPM 马尔可夫扩散的后验分布

在我们之前推导 DDPM 的时候，我们知道

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}) \tag{14}$$

其中，

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \tag{15}$$

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{16}$$

我们可以看到，DDPM 的后验分布 $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ 中的方差是一个常数。而在 DDIM 的后验分布中 $q_\sigma(x_{t-1}|x_t, x_0)$ 中，多了一个超参数 σ_t:

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\mathbf{I}\right) \tag{17}$$

这个超参数 σ_t 就给了 DDIM 发挥的空间，我们可以自己来设置 σ_t。σ_t 的不同，就决定了后验分布的不同；而后验分布的不同，就影响了模型采样的重参数化的计算方式。

2.4 非马尔可夫扩散逆过程的采样

接下来，我们可以定义一个可训练的生成过程 $p_\theta(X_{0:T})$，θ 代表模型的可训练的参数。其中每一步的迭代 $p_\theta^{(t)}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 都是利用了公式 $q_\sigma(x_{t-1}|x_t, x_0)$。

因为我们前面(13)已经证明了在 DDIM 中 $q(x_t \mid x_0)$ 的形式，因此只要给定了 x_0 和 ϵ_t 的话，x_t 是可以根据重参数技巧计算得到的。在 DDPM 中，模型预测的是噪声，也就是 ϵ_t，但是得到了 ϵ_t 和 x_t，我们可以根据公式(13)来反推得到 x_0。虽然我们模型预测的是 ϵ_t，但是我们可以将它和 x_t 带入到公式(13)中得到 $f_\theta^{(t)}(x_t)$，这个值是去噪观测量，我们也可以认为它是在当前时刻我们能够预测得到的 x_0:

$$f_\theta^{(t)}(x_t) := \left(x_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(x_t)\right) / \sqrt{\alpha_t} \tag{18}$$

既然 x_0 已经出来了，那我们就可以当它代入公式(6)中得到生成过程:

$$p_\theta^{(t)}(x_{t-1}|x_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(x_1), \sigma_1^2 \mathbf{I}) & \text{if } t = 1 \\ q_\sigma(x_{t-1}|x_t, f_\theta^{(t)}(x_t)) & \text{otherwise} \end{cases} \tag{19}$$

2.5 一种特殊的采样-DDIM

刚刚我们已经写出 DDIM 的采样公司，见公式(19)，但是因为采样过程中(19)(6)存在着超参数 σ。我们是可以自己设置超参数 σ

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta^{(t)}(x_t) + \sigma_t \epsilon_t \tag{20}$$

3 Questions

3.1 为什么 DDIM 可以复用 DDPM 模型？

DDIM 的作者团队发现，**DDPM 的损失函数是依赖于边缘分布 $q(x_t \mid x_0)$，而不依赖于联合分布 $q(x_{1:T} \mid x_0)$**。那也就是说，只要我们设计了一个分布，它在任何时刻 t 的边缘分布都是等于 DDPM 的边缘分布的，而联合分布不一定需要等于 DDPM 的联合分布，那么我们就可以直接复用 DDPM 的模型。

3.2 简单介绍一下 DDIM？

DDIM 是从一个更一般的，非马尔可夫的扩散过程。这个扩散过程的加噪过程的边缘分布和 DDPM 一样，但是联合分布和 DDPM 的不一样，因此，它可以复用 DDPM 的模型。另外，在采样过程中，通过将采样分布中的方差设置为 0，也就是在重参数化技巧的时候，不需要从从标准正态分布中来随机采样噪声，此时的采样过程已经变成一个确定性的了，这就是 DDIM。

3.3 DDIM 中的确定性采样指的是什么？

我们知道，DDIM 设计了一种非马尔可夫链的扩散过程，它的采样过程如(20)所示，当我们将这个分布的标准差 σ_t 设置为 0 的时候，公式中的最后一项就会变为零。因此，此时在采样的过程中就不再需要从标准正太分布中采样一个随机噪声了，即此时的采样过程已经变成确定的了。

3.4 介绍一下 DDIM 中的 respacing？

DDIM 中使用了 respacing 的技巧来加速采样的步骤，因为 DDIM 训练扩散模型的过程中，Loss 函数和扩散过程的联合分布没有什么具体的关系。这样的话，就允许 **DDIM 在推理的过程中可以跳过一些步骤，它可以以五倍少的步骤来获得跟 DDPM 一样的采样质量**。当然了，**DDPM 也可以用 respacing 策略，但是 DDPM 使用 respacing 后，图片生成的质量效果不太好**。