南京信息工程大学

高级算法设计课程报告

题　目　基于扩散的图像生成模型与 DDPM 原理

学生姓名　周肖桐

学　　号　202312200030

学　　院　计算机学院、网络空间安全学院

专　　业　计算机技术

指导教师　余文斌

二〇二四 年 六 月 二十一 日

目录

1 前言 4
2 写在开始前 4
3 基于扩散的图像生成模型的基本流程 5
3.1 逆扩散过程 6
3.2 扩散过程 6
4 扩散过程原理与推导 7
4.1 参数设定 7
4.2 扩散过程 7
5 逆扩散过程原理与推导 11
6 噪声预测器的训练 14
6.1 极大似然函数与极大似然估计 14
6.2 利用极大似然函数进行参数迭代 15
6.2.1 复习极大似然函数的构建过程 15
6.2.2 扩散模型场景与极大似然估计 16
6.2.3 最大化似然函数的下界 17
7 总结 21
名词辨析 22
参考文献 22

基于扩散的图像生成模型与 DDPM 原理

周肖桐[1)]

南京信息工程大学计算机学院、网络空间安全学院，江苏 南京 210044

摘要：本文旨在全面概述和深入分析扩散与逆扩散过程在图像生成模型中的应用与理论基础。文章分为三个主要部分：首先总览扩散模型的图像生成过程,并作为扩散与逆扩散过程的引入。其次，详细阐述扩散过程与逆扩散过程的概念与区别，讨论这些过程在图像生成中的重要性及其工作原理。最后，通过数学公式推导，深入解析扩散与逆扩散过程的具体实现，提供系统性的理论支撑，并对其性能和有效性进行评估。本文通过对这两种过程的全面分析与推导，旨在为图像生成领域的研究者提供有价值的参考与指导。

关键词：扩散模型、原理介绍、公式推导

[1)] E-mail：xiaotongzhounuist@163.com

Diffusion Based Generation Model & Principles of DDPM

Xiaotong Zhou[2)]

School of Computer Science，NUIST，Nanjing 210044，China

Abstract： This paper aims to provide a comprehensive overview and in-depth analysis of the application and theoretical foundation of diffusion and reverse diffusion processes in image generation models. The article is divided into three main sections: First, it introduces the basic image generating process of DDPM, aiming to explore the structural similarities of these models and to serve as an introduction to diffusion and its reverse processes. Second, it elaborates on the concepts and differences between the diffusion process and the reverse diffusion process, discussing their importance and working principles in image generation. Finally, through mathematical formula derivation, it delves into the specific implementation of diffusion and reverse diffusion processes, providing systematic theoretical support and evaluating their performance and effectiveness. By providing a comprehensive analysis and derivation of these two processes, this paper aims to offer valuable reference and guidance for researchers in the field of image generation.

Key words： Diffusion models, Principles introduction, Formula derivation

[2)] E-mail: xiaotongzhounuist@163.com

1 前言

近年来，随着人工智能和深度学习技术的迅猛发展，图像生成模型在计算机视觉领域中发挥着越来越重要的作用。生成对抗网络（GAN）、变分自编码器（VAE）以及『基于扩散的图像生成模型』等，不仅在学术研究中取得了显著的进展，也在实际应用中展现出巨大的潜力。其中，『基于扩散的图像生成模型』通过模拟物理扩散过程，实现了高质量图像的生成和重建，成为当前图像生成研究的热点之一。

扩散过程和逆扩散过程是『基于扩散的图像生成模型』的核心机制。扩散过程模拟从清晰图像到噪声图像的逐步退化，而逆扩散过程则反其道而行之，通过一系列渐进的步骤，将噪声图像恢复为清晰图像。这一双向过程不仅为图像生成提供了新的思路，也在理论上为图像处理任务提供了新的视角。

本文旨在对扩散与逆扩散过程的理论基础和实际应用进行全面综述与深入解析。文章首先介绍了『基于扩散的图像生成模型』的基本工作流程，旨在为读者提供该类模型生成过程的概念框架。然后，详细阐述了『扩散过程与逆扩散过程』的原理及其在图像生成中的重要性。最后，通过数学公式推导，深入解析了这两个过程的具体实现，提供了系统性的理论支持。

以下是作者的一些碎碎念。

许多人可能认为 DDPM 是一个难以理解的过程，同时其推导也需要大量的数学知识，但其实不然。经过本人的理解，DDPM 涉及的扩散过程与逆扩散过程并非难以理解，整体过程简洁而有效。DDPM 的推导也并不需要过多的数学知识，推导过程中使用的最多的公式为贝叶斯公式，最困难的知识为 KL 散度的定义与极大似然估计的原理，所以 DDPM 并不是一个难以理解的知识。只要您仔细阅读并积极思考与理解，就能轻松获得关于 DDPM 的基础原理。为了能够获得愉快的阅读体验，本文使用口语化的表达方式；同时为了防止文字出现歧义，使用了尽可能明确的表达方式。通过以上两种方式，力求为读者提供愉悦体验的同时学到清晰明确的结论。

此外，本文以 Denoising Diffusion Porbabilistic Models（DDPM）[1] 及其原理解释论文 [2] 为基础，吸收了李宏毅老师关于 DDPM 的机器学习课程 [3] [4] [5] [6] [7] [8] 的讲解思路，参考多篇网络博客 [9]，结合自身理解与公式推导，完成了这篇关于 DDPM 的直观与深入解释文章。尽管与原始论文 [1] 在推导结果上存在的一定的差异，但是经过一定的变换后，可以得到相同的结果，且个人认为本人的推导可能在帮助理解方面能有更好的效果，所以请各位读者放心阅读。

本人花费约 2 个星期将 DDPM 的原理基本吸收，并将自己的思路与理解记录在此处，并以「力求能教会所有学过概率论的读者」的标准撰写该文档，希望本文能为图像生成领域的研究者提供有价值的参考和指导，促进该领域的进一步发展和创新。如果在阅读时遇到了名词混淆的情况，请阅读7名词辨析作为参考。

2 写在开始前

在开始前，我想先解释一下本文中两个名词『DDPM』与『基于扩散的图像生成模型』之间的差异，以解释本文内容与标题之间的联系。本文中的『DDPM』指的是文章 Denoising Diffusion Probabilistic Model [1] 中提出的模型，是最早的『基于扩散的图像生成模型』。而『基于扩散的图像生成模型』指的是以所有扩散过程与逆扩散过程作为基本原理的图像生成模型。本文的标题为『基于扩散的图像生成模型与 DDPM 原理』，实际上是在介绍两个部分，其一为『基于扩散的图像生成模型』，其二为『DDPM 原理』。第一部分将接受当前所有的『基于扩散的图像生成模型』的总体工作流程，从整体的角度为各位介绍该类模型的生成过程。第二部分以最简单的『基于扩散的图像生成模型』，也即 DDPM 作为基础，为各位推导最简单的流程。

下面开始第一部分，介绍『基于扩散的图像生成模型』的基本流程。

3 基于扩散的图像生成模型的基本流程

该部分介绍基于扩散的图像生成模型的基本工作流程，为不了解该类模型的读者简单介绍一下模型的生成过程，为后面详细介绍原理做准备。

首先需要明确的是『扩散过程与逆扩散过程』与『基于扩散的图像生成模型』这两个名词之间的差别。有些文章可能将这两个概念混淆，或直接称『基于扩散的图像生成模型』为『扩散模型』，但本文希望能将这两个概念进行辨析，以获得更明确无歧义的阅读体验。

『扩散过程与逆扩散过程』可以分成两个过程，即『扩散过程』与『逆扩散过程』。『扩散过程』其实是一个物理过程，图像生成模型中模仿这个概念，让噪声逐渐扩散到整个图像的过程被称为『扩散过程』；而『逆扩散过程』则是将扩散过程反过来，从噪声图像中恢复出一张不带有噪声的图像的过程。扩散过程可以看作『基于扩散的图像生成模型』中的一个基本原理。

而『基于扩散的图像生成模型』则是以『扩散过程与逆扩散过程』作为基本原理，进行图像生成的模型，二者是包含与被包含的关系。后者包含前者，而前者作为后者的一部分，为后者提供服务。本文中对于『扩散过程与逆扩散过程』与『基于扩散的图像生成模型』这两个概念均与上述描述保持一致，若在阅读中出现概念对这两个概念混淆的现象，可以翻阅本段文字或查阅7名词解释部分作为参考。

其次需要介绍的是『基于扩散的图像生成模型』的输入与输出。该类模型接收对于图像的文字描述作为输入，并输出一张或多张与输入文字描述相符合的图像，如下图3.1所示。

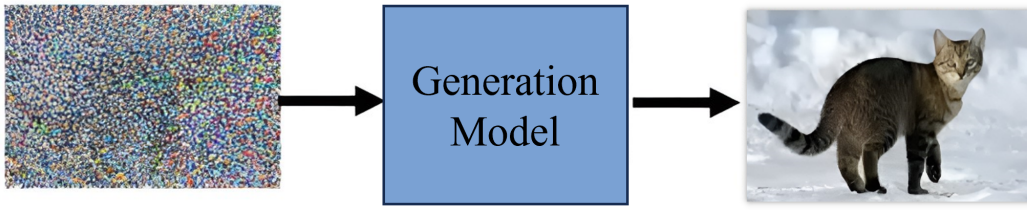
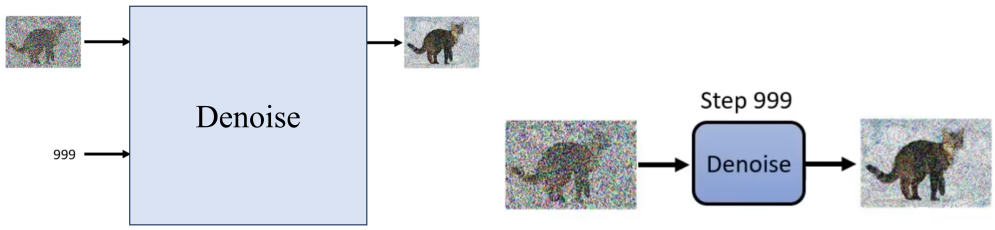


图 3.1 基于扩散的图像生成模型：输入与输出

当接收到文字输入后，『基于扩散的图像生成模型』将随机从高斯分布中采样一张噪声图像，并将该噪声图像与文字输入一同送进 Denoise 模块中。Denoise 模块将去除噪声图像中的部分噪声，从而得到一张部分降噪的图像。这个过程如下图3.2所示。



（a） Denoise 模块的输入与输出　　（b） 单个 Denoise 过程

图 3.2 基于扩散的图像生成模型：Denoise 模块

将得到的部分降噪图像重新输入到 Denoise 模块中，重复多次，便可得到一张符合输入文字描述的图像。这就是『基于扩散的图像生成模型』的基本工作流程，如图3.3所示。

总结一下，『基于扩散的图像生成模型』接收对于图像的描述作为输入，对自己生成的一张噪声图经过多次去噪操作后，得到一张与输入文字描述相同的图像。

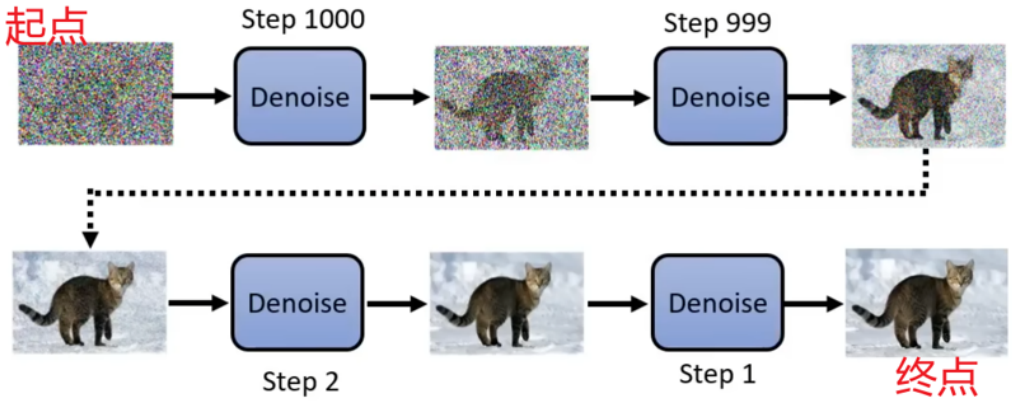


图 3.3 基于扩散的图像生成模型：整体工作流程。用一句话概括上述对『基于扩散的图像生成模型』工作流程的描述：『基于扩散的图像生成模型』接收对于图像的描述作为输入，对自己生成的一张噪声图经过多次去噪操作后，得到一张与输入文字描述相同的图像。

下面将介绍逆扩散过程与扩散过程。为了能与『基于扩散的图像生成模型』基本流程形成较好的衔接，所以先介绍逆扩散过程。

3.1 逆扩散过程

逆扩散过程可以被描述为这样一个过程：从噪声图像中恢复出不带有噪声的图像的过程。为了实现这个目标，在『基于扩散的图像生成模型』中存在多个 Denoise 过程, 该过程即图3.3表示的过程。

Denoise 过程接收当前的工作步数与上一个 Dnoise 过程的输出作为输入，输出一张在上一个 Denoise 输出基础上的部分去噪图，如下图所示。

上图中的每个 Denoise 过程都由同一个神经网络完成，即一个神经网络即可完成所有的 Denoise 过程。为了使这样的 Denoise 网络能完成这样的目标，我们需要大量成对的训练数据。这样的成对数据中应该包含两部分：当前工作步数与工作步数对应的部分去噪图。为了获取上述成对数据以供 Denoise 网络训练，我们需要扩散过程。

3.2 扩散过程

扩散过程可以被描述成这样一个过程：从清晰图像逐步添加噪声直到完全变成噪声图像的过程。为了模拟这种从无噪声到有噪声的过程，在基于扩散的图像生成模型中，通过逐步增加噪声来生成一系列带有不同程度噪声的图像。具体来说，扩散过程从一张清晰图像开始，逐步在多个步骤中加入噪声，最终得到一个完全噪声化的图像，如下图所示3.4。

图 3.4 扩散过程

在扩散过程的每一步中，图像中的噪声增加的程度是通过预定义的噪声分布实现的。每个步骤的输出即为下一个步骤的输入，逐步累积噪声直至达到预设的最大噪声水平。

上图中的每个扩散步骤都通过添加高斯噪声来实现，这些高斯噪声是通过一个简单的随机数生成器产生的。为了确保扩散过程能够准确地模拟自然图像中噪声的增加，噪声的分布和强度必须经过精确的设计和调整。

通过这种方式，扩散过程生成了大量的带噪声图像，这些图像用于训练逆扩散过程中的 Denoise 网络。每对训练数据中的一个部分是一个特定工作步数下的带噪声图像，另一个部分是该步数下去噪后的部分清晰图像。

在介绍了『基于扩散的图像生成模型』的基本工作流程后，我们将深入探讨该领域中最早提出的经典模型之一：DDPM。该模型由 Ho 等人在论文《Denoising Diffusion Probabilistic Models》[1] 中提出，是一种具有开创性意义的图像生成模型。DDPM[1] 第一个在图像生成模型中引入了『扩散过程与逆扩散过程』，奠定了基于扩散方法的图像生成技术的理论基础。通过详细推导 DDPM 的原理，我们可以更好地理解『基于扩散的图像生成模型』的原理。

接下来，让我们一起探讨 DDPM 的具体原理及其推导过程。该部分将包括三个主要内容：其一为扩散过程的原理与推导，其二为逆扩散过程的原理与推导，其三为 Dnoise 网络的训练。

4 扩散过程原理与推导

扩散过程如图3.4所示。扩散过程作为为 Denoise 网络提供训练数据的过程，在 DDPM[1] 存在重要的作用。该部分主要将扩散过程以参数化的形式表示，通过写出该过程公式并化简的方式，从公式的角度介绍扩散过程。

4.1 参数设定

- 原始图像：x_0
- 第 t 次的从标准高斯分布中采样噪声图像：z_t
- 第 t 次将噪声 z_t 加入 x_0 后的图像：x_t
- 第 t 次加噪声时, 噪声图像 z_t 与图像 x_{t-1} 的比例 $1-\alpha_t$ 与 α_t。$\alpha_1, \alpha_2, \cdots, \alpha_t$ 是一组常数，在扩散过程开始前人为设定。

4.2 扩散过程

扩散过程公式

我们可以用符号语言描述扩散过程，如下：

1. 从数据集中获取一张原始的真实图像 x_0

2. 从标准高斯分布 $\mathcal{N}(0,1)$ 中采样一张噪声图 z_1

3. 将噪声图 z_1 与原始图像 x_0 按 $\sqrt{1-\alpha_1}$ 与 $\sqrt{\alpha_1}$ 的比例混合, 可以得到第一步的加噪声结果, 如下所示

$$x_1 = \sqrt{1-\alpha_1} z_0 + \sqrt{\alpha_1} x_0 \tag{4.1}$$

4. 将噪声图 z_2 与上一步得到的结果 z_1 按 $\sqrt{1-\alpha_2}$ 与 $\sqrt{\alpha_2}$ 的比例混合, 如下所示

$$x_2 = \sqrt{1-\alpha_2} z_1 + \sqrt{\alpha_2} x_1 \tag{4.2}$$

5. 则第 t 张加噪图像 x_t 满足以下公式：

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_t \tag{4.3}$$

扩散过程公式简化

整体简化 我们考察一般情况，对第 t 张加噪图像 x_t 满足的公式 $x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_t$ 进行如下变换：

$$\begin{aligned}
x_t &= \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_t \\
&= \sqrt{\alpha_t} \left(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}} z_{t-1} \right) + \sqrt{1-\alpha_t} z_t \\
&= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{\alpha_t (1-\alpha_{t-1})} z_{t-1} + \sqrt{1-\alpha_t} z_t \\
&= \sqrt{\alpha_t \alpha_{t-1}} \left(\sqrt{\alpha_{t-2}} x_{t-3} + \sqrt{1-\alpha_{t-2}} z_{t-2} \right) + \sqrt{\alpha_t (1-\alpha_{t-1})} z_{t-1} + \sqrt{1-\alpha_t} z_t \\
&= \sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2}} x_{t-3} + \sqrt{\alpha_t \alpha_{t-1} (1-\alpha_{t-2})} z_{t-2} + \sqrt{\alpha_t (1-\alpha_{t-1})} z_{t-1} + \sqrt{1-\alpha_t} z_t \\
&= \cdots \\
&= \sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} x_0 + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_2 (1-\alpha_1)} z_1 + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_3 (1-\alpha_2)} z_2 \\
&\qquad + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_4 (1-\alpha_3)} z_3 + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_5 (1-\alpha_4)} z_4 \\
&\qquad + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_6 (1-\alpha_5)} z_5 + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_7 (1-\alpha_6)} z_6 \\
&\qquad \vdots \\
&\qquad + \sqrt{\alpha_t (1-\alpha_{t-1})} z_{t-1} + \sqrt{1-\alpha_t} z_t \\
&= \underbrace{\sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} x_0}_{\text{原图项}} + \underbrace{\sum_{i=1}^{t} \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_{i+1} (1-\alpha_i)} z_i}_{\text{叠加噪声项}}
\end{aligned} \tag{4.4}$$

通过上述公式4.4, 我们便可以计算经过 t 次加噪后获得的图像 x_t. 整个公式可以看作是两个部分,

1. 其一为包含原图 x_0 的「原图项」$\sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} x_0$
2. 其二为包含 t 个噪声 z_i 的「叠加噪声项」$\sum_{i=1}^{t} \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_{i+1} (1-\alpha_i)} z_i$

下面我们分别对「原图项」与「叠加噪声项」进行化简。

原图项的简化 实际上,「原图项」是一个化简过程较为简单的项, 而后面的「叠加噪声项」是一个化简过程较为复杂的项. 现在我们探索如何简化「原图项」.

我们定义「原图项」$\sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} x_0$ 的系数 $\sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1}$ 为 $\overline{\alpha_t}$, 有

$$\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i = \sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} \tag{5}$$

如 $\overline{\alpha_2}=\alpha_2\cdot\alpha_1$,$\overline{\alpha_4}=\alpha_4\cdot\alpha_3\cdot\alpha_2\cdot\alpha_1$

在此定义下, 将 $\overline{\alpha_t}=\prod_{i=1}^{t}\alpha_i$ 代入 x_t 中的「原图项」, 有

$$\begin{aligned}x_t&=\sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}\cdots\alpha_1}x_0+\sum_{i=1}^{t}\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}z_i\\&=\sqrt{\overline{\alpha_t}}x_0+\sum_{i=1}^{t}\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}z_i\end{aligned}\tag{4.5}$$

叠加噪声项的简化 因为我们希望整个「叠加噪声项」是一个简单的噪声, 所以我们希望最终 x_t 满足的表达式能与其一般式 $x_t=\sqrt{\alpha_t}x_{t-1}+\sqrt{1-\alpha_t}z_t$ 具有相同的形式。即，我们希望 x_t 的最终化简结果为 $x_t=\sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{\alpha_t}}\widetilde{z}$ 的形式。在这个形式中，叠加噪声项系数与原图项系数的平方和为 1，与一般式保持一致；同时，我们也希望叠加噪声项的主体部分 $\widetilde{z}$ 是一个简单的噪声，而非公式4.4中那样一堆噪声的叠加。下面正式开始对叠加噪声项的简化。

注意到「原图项」的系数现在为 $\sqrt{\overline{\alpha_t}}$, 我们期望后面的「叠加噪声项」的系数为 $\sqrt{1-\overline{\alpha_i}}$, 所以对「叠加噪声项」提取系数 $\sqrt{1-\overline{\alpha_t}}$, 从而有以下变化

$$\begin{aligned}x_t&=\sqrt{\overline{\alpha_t}}x_0+\sum_{i=1}^{t}\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}z_i\\&=\sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{\alpha_t}}\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha_t}}}z_i\end{aligned}\tag{4.6}$$

现在考察上述公式4.6中的 $\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha_t}}}z_i$ 部分 (即上面公式中最后一个分式) 以期望他是一个简单的分布。

为了方便描述，我们记这个分布为 $\widetilde{z}$，即有 $\widetilde{z}=\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha_t}}}z_i$.

这个分式部分 $\widetilde{z}$ 中的单个噪声项 z_i 均是从标准高斯分布中采样得到的, 即 $z_i\sim\mathcal{N}(0,1)$(均值为 0，方差为 1), 且由于上一次采样的噪声不会影响下一次噪声的采样, 所以 z_i 的获取是相互独立的.

同时, 我们知道, 如果 $X\sim\mathcal{N}(\mu_x,\sigma_x^2)$,$Y\sim\mathcal{N}(\mu_y,\sigma_y^2)$, 且 X 与 Y 相互独立, 则有 $aX+bY\sim\mathcal{N}(a\mu_x+b\mu_y,a^2\sigma_x^2+b^2\sigma_y^2)$。

在这样的情况下, 有 $\widetilde{z}$ 服从以下高斯分布：

$$\widetilde{z}\sim\mathcal{N}\left[\sum_{i=1}^{t}\left(\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha_t}}}\cdot 0\right),\sum_{i=1}^{t}\left(\frac{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}{1-\overline{\alpha_t}}\cdot 1\right)\right]\tag{4.7}$$

可以发现, 这个高斯分布的均值部分为 0, 因为均值部分中累加的每一项都与 0 相乘了. 从而上述高斯分布可以化简为：

$$\widetilde{z}\sim\mathcal{N}\left(0,\sum_{i=1}^{t}\frac{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}{1-\overline{\alpha_t}}\right)\tag{4.8}$$

为了搞清楚这个「叠加噪声项」到底满足什么样的高斯分布, 我们继续考察这个高斯分布的方差部分, 记为 $\sigma_s^2=\sum_{i=1}^{t}\frac{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}{1-\overline{\alpha_t}}$, 下标 s 为 sum 的缩写, 表示累加。

直接将 σ_S^2 拆开, 有

$$
\begin{aligned}
\sigma_s^2 &= \sum_{i=1}^{t} \frac{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}{1-\overline{\alpha_t}} \\
&= \frac{1}{1-\overline{\alpha_t}}\left[\sum_{i=1}^{t}\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)\right] \\
&\xlongequal{\text{去除累加符号}} \frac{1}{1-\overline{\alpha_t}}\left[(1-\alpha_t)+\alpha_t(1-\alpha_{t-1})+\alpha_t\alpha_{t-1}(1-\alpha_{t-2})+\cdots+\alpha_t\alpha_{t-1}\cdots\alpha_2(1-\alpha_1)\right] \\
&\xlongequal{\text{拆除小括号}} \frac{1}{1-\overline{\alpha_t}}\left[1-\alpha_t+\alpha_t-\alpha_t\alpha_{t-1}+\alpha_t\alpha_{t-1}-\alpha_t\alpha_{t-1}\alpha_{t-2}+\cdots+\alpha_t\alpha_{t-1}\cdots\alpha_2-\alpha_t\alpha_{t-1}\cdots\alpha_2\alpha_1\right] \\
&= \frac{1}{1-\overline{\alpha_t}}\left[1+\cancel{-\alpha_t}+\cancel{\alpha_t}\cancel{-\alpha_t\alpha_{t-1}}+\cancel{\alpha_t\alpha_{t-1}}\cancel{-\alpha_t\alpha_{t-1}\alpha_{t-2}}+\cdots+\cancel{\alpha_t\alpha_{t-1}\cdots\alpha_2}-\alpha_t\alpha_{t-1}\cdots\alpha_2\alpha_1\right] \\
&\xlongequal{\text{发现中括号中除了第一项和最后一项都可以消去}} \frac{1}{1-\overline{\alpha_t}}\left[1-\alpha_t\alpha_{t-1}\cdots\alpha_2\alpha_1\right] \\
&= \frac{1}{1-\overline{\alpha_t}}(1-\overline{a_t}) \\
&\xlongequal{\text{分子分母相同}} 1
\end{aligned}
\tag{4.9}
$$

也即 $\sigma_s^2=1$, 从而有 $\widetilde{z}=\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha_t}}}z_i\sim\mathcal{N}(0,1)$,

至此, 我们可以发现,「叠加噪声项」也是一个服从标准高斯分布的噪声, 从而我们可以得到第 t 步图像 x_t 与原图 x_0 之间的关系：

$$
\begin{aligned}
x_t &= \sqrt{\overline{\alpha_t}}x_0+\sum_{i=1}^{t}\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}z_i \\
&= \sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{a_t}}\widetilde{z}, \quad \text{其中}\widetilde{z}\sim\mathcal{N}(0,1)
\end{aligned}
\tag{4.10}
$$

进行对「叠加噪声项」的化简后，我们发现实际上 t 次添加噪声的操作与直接进行一次噪声添加的效果相同。

扩散过程的总结

在公式4.10的指导下, 可以立刻得到某个特定步骤 t 的加噪图像 x_t, 且加噪图像仅与原始图像 x_0 与当前步骤数 t 有关.

经过这样的简化, 就可以简单的获取加噪后的图像, 也即训练数据了。

概率采样的角度看扩散过程

在上面的推导中, 我们得出了直接从原图 x_0 获取第 t 次后的加噪图像 x_t 的公式如下：

$$
x_t = \sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{a_t}}\widetilde{z}, \quad \text{其中}\widetilde{z}\sim\mathcal{N}(0,1) \tag{4.11}
$$

实际上, 我们也可以将 x_t 看作是某种概率的采样结果, 推导如下：

已知 $\widetilde{z}\sim\mathcal{N}(0,1)$, 则有 $\sqrt{1-\overline{a_t}}\widetilde{z}\sim\mathcal{N}(0,1-\overline{\alpha_t})$

从而有 $\sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{a_t}}\widetilde{z}\sim\mathcal{N}(\sqrt{\overline{\alpha_t}}x_0,1-\overline{\alpha_t})$,

从而有 $q(x_t|x_0)\sim\mathcal{N}(\sqrt{\overline{\alpha_t}}x_0,1-\overline{\alpha_t})$

这个结果表明，第 t 步的加噪图像 x_t 可以看作是从一个正态分布中采样的结果，其均值和方差分别由初始图像 x_0 和累积噪声参数 $\overline{\alpha_t}$ 决定

同理, 由于有 $x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}z_t$, 所以有

$$q(x_t|x_{t-1}, x_0) \sim \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, 1-\alpha_t) \tag{4.12}$$

从概率的角度理解扩散过程有助于下面对于逆扩散过程的推导。

5 逆扩散过程原理与推导

很显然, 我们想要通过 x_t 来预测 x_{t-1}.

如果我们能够逆转上述扩散过程, 并从 $p(x_{t-1}|x_t)$ 采样，就可以从高斯噪声 $x_t \sim N(0,1)$ 还原出原图服从的分布 $x_0 \sim p(x)$。

如何获得 $p_\theta(x_{t-1}|x_t)$ 这个概率密度 (式子中的 θ 表示神经网络的参数) 就是一个需要探讨的问题. 直接计算比较困难, 所以我们可以考虑对公式进行变形. 对公式使用贝叶斯公式, 有如下结果

$$\begin{aligned} p_\theta(x_{t-1}|x_t) &\xlongequal{\text{贝叶斯公式}} \frac{p_\theta(x_{t-1}, x_t)}{p(x_t)} \\ &\xlongequal{\text{分母用全概率公式展开}} \frac{p(x_t|x_{t-1}) \cdot p_\theta(x_{t-1})}{p(x_t)} \\ &= p(x_t|x_{t-1}) \cdot \frac{p_\theta(x_{t-1})}{p(x_t)} \end{aligned} \tag{5.1}$$

通过这样的变换, 我们将一个无法计算的式子 $p_\theta(x_{t-1}|x_t)$, 改写成了一个可计算的部分 $p(x_t|x_{t-1})$ 和一个不可计算的分式 $\frac{p_\theta(x_{t-1})}{p(x_t)}$ 的乘积

自己观察这个不可计算的分式, 可以发现, 这个分式的分子与分母都是不可计算的. 因为如果我们能直接得到 $p(x_t)$ 或 $p(x_{x-1})$, 那我们就能直接得出 $p(x_0)$. 但是我们计算 $p(x_{t-1}|x_t)$ 的目的就是为了计算 $p(x_0)$, 如果可以直接得出 $p(x_0)$, 那么我们就没有计算 $p(x_{t-1}|x_t)$, 所以我们不可能直接得到 $p(x_t)$.

可以想到, 虽然直接计算 $p(x_t)$ 是不可行的, 但是计算 $p(x_t|x_0)$ 是十分简单的, 在4.2 中我们介绍过这个计算：

$$p(x_t|x_0) \sim \mathcal{N}(\sqrt{\overline{\alpha_t}}x_0, 1-\overline{\alpha_t})$$

所以我们可以考虑将 $p(x_t)$ 的计算转换成 $p(x_t|x_0)$ 用于计算, 由此可以计算 $p_\theta(x_{t-1}|x_t, x_0)$, 有

$$\begin{aligned} p_\theta(x_{t-1}|x_t, x_0) &\xlongequal{\text{贝叶斯公式}} \frac{p_\theta(x_{t-1}, x_t, x_0)}{p(x_t, x_0)} \\ &\xlongequal{\text{分子用全概率公式展开}} \frac{p(x_t|x_{t-1}, x_0) \cdot p_\theta(x_{t-1}, x_0)}{p(x_t, x_0)} \\ &= p(x_t|x_{t-1}) \cdot \frac{p(x_{t-1}, x_0)}{p(x_t, x_0)} \\ &= p(x_t|x_{t-1}) \cdot \frac{p(x_{t-1}, x_0)}{p(x_t, x_0)} \\ &= p(x_t|x_{t-1}) \cdot \frac{p(x_{t-1}|x_0) \cdot p(x_0)}{p(x_t|x_0) \cdot p(x_0)} \\ &= p(x_t|x_{t-1}) \cdot \frac{p(x_{t-1}|x_0)}{p(x_t|x_0)} \end{aligned} \tag{5.2}$$

对上述推导取等式左侧与右侧第一项, 有 $p_\theta(x_{t-1}|x_t, x_0) = p(x_t|x_{t-1}) \cdot \frac{p_\theta(x_{t-1}|x_0)}{p(x_t|x_0)}$

可能有人发现, 在第三个等式中, 我们直接将 $p(x_t|x_{t-1}, x_0)$ 替换为了 $p(x_t|x_{x-1})$, 这可以用马尔可夫性（Markov property）来解释：马尔可夫性假设指出，一个状态只依赖于前一个状态，而与更早的状态条件独立。应用于 DDPM 模型，这意味着：

$p(x_t|x_{t-1}, x_0) = p(x_t|x_{t-1})p(x_t|x_{t-1}, x_0) = p(x_t|x_{t-1})p(x_t|x_{t-1}, x_0) = p(x_t|x_{t-1})$

这三项就都是很好计算的项了, 虽然此处我们使用的是 $p(x_t|x_{t-1})$ 而非 $q(x_t|x_{t-1})$, 但他们之间表示的内容是一致的, 均表示前向过程中的加噪.

从4.2中, 我们有推导结果

$$
\begin{aligned}
p(x_t|x_{t-1}, x_0) = q(x_t|x_{t-1}, x_0) &\sim \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, 1-\alpha_t) \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\cdot(\frac{x_t-\mu}{\sigma})^2} \\
&= \frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\alpha_t}} \exp\left[-\frac{1}{2}\cdot\left(\frac{x_t-\sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}}\right)^2\right]
\end{aligned}
\tag{5.3}
$$

x_t 为随机变量

$$
\begin{aligned}
p_\theta(x_{t-1}|x_0) = q(x_{t-1}|x_0) &\sim \mathcal{N}(\sqrt{\overline{\alpha_t}}x_0, 1-\overline{\alpha_{t-1}}) \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\cdot(\frac{x_{t-1}-\mu}{\sigma})^2} \\
&= \frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\overline{\alpha_{t-1}}}} \exp\left[-\frac{1}{2}\cdot\left(\frac{x_{t-1}-\sqrt{\overline{\alpha_{t-1}}}x_0}{\sqrt{1-\overline{\alpha_{t-1}}}}\right)^2\right]
\end{aligned}
\tag{5.4}
$$

x_{t-1} 为随机变量

$$
\begin{aligned}
p(x_t|x_0) = q(x_t|x_0) &\sim \mathcal{N}(\sqrt{\overline{\alpha_t}}x_0, 1-\overline{\alpha_t}) \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\cdot(\frac{x_t-\mu}{\sigma})^2} \\
&= \frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\overline{\alpha_t}}} \exp\left[-\frac{1}{2}\cdot\left(\frac{x_t-\sqrt{\overline{\alpha_t}}x_0}{\sqrt{1-\overline{\alpha_t}}}\right)^2\right]
\end{aligned}
\tag{5.5}
$$

x_t 为随机变量

将这三个式子带入上述公式 $p_\theta(x_{t-1}|x_t, x_0) = p(x_t|x_{t-1}) \cdot \frac{p_\theta(x_{t-1}|x_0)}{p(x_t|x_0)}$, 有如下化简：

$$
\begin{aligned}
p_\theta(x_{t-1}|x_t, x_0) &= p(x_t|x_{t-1}) \cdot \frac{p_\theta(x_{t-1}|x_0)}{p(x_t|x_0)} \\
&= \frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\alpha_t}} \exp\left[-\frac{1}{2}\cdot\left(\frac{x_t-\sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}}\right)^2\right] \cdot \frac{\frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\overline{\alpha_{t-1}}}} \exp\left[-\frac{1}{2}\cdot\left(\frac{x_{t-1}-\sqrt{\overline{\alpha_{t-1}}}x_0}{\sqrt{1-\overline{\alpha_{t-1}}}}\right)^2\right]}{\frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\overline{\alpha_t}}} \exp\left[-\frac{1}{2}\cdot\left(\frac{x_t-\sqrt{\overline{\alpha_t}}x_0}{\sqrt{1-\overline{\alpha_t}}}\right)^2\right]} \\
&\xrightarrow[\text{忽略前面的系数}]{\text{指数相乘 (除) 等于幂相加 (减)}} k \cdot \exp\left\{-\frac{1}{2}\cdot\left[\left(\frac{x_t-\sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}}\right)^2 + \left(\frac{x_{t-1}-\sqrt{\overline{\alpha_{t-1}}}x_0}{\sqrt{1-\overline{\alpha_{t-1}}}}\right)^2 - \left(\frac{x_t-\sqrt{\overline{\alpha_t}}x_0}{\sqrt{1-\overline{\alpha_t}}}\right)^2\right]\right\}
\end{aligned}
\tag{5.6}
$$

这个公式看起来很复杂, 但是我们可以这样考虑：整个公式为一个常数与一个「以 e 为底的指数」的乘积, 与高斯分布的形式很像, 而我们知道, 高斯分布中的指数部分是一个完全平方：$\exp\left(-\frac{1}{2}\cdot\left[\frac{(x-\mu)}{\sigma}\right]^2\right) = \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}x^2 - \frac{2\mu}{\sigma^2}x + \frac{\mu^2}{\sigma^2}\right)\right\}$, 所以我们可以也将上面公式中的指数部分变成一个完全平方.

根据上面的思想，我们可以有如下推导：

$$p_\theta(x_{t-1}|x_t,x_0)=k\cdot\exp\left\{-\frac{1}{2}\cdot\left[\left(\frac{x-\sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}}\right)^2+\left(\frac{x-\sqrt{\overline{\alpha_{t-1}}}x_0}{\sqrt{1-\overline{\alpha_{t-1}}}}\right)^2-\left(\frac{x-\sqrt{\overline{\alpha_t}}x_0}{\sqrt{1-\overline{\alpha_t}}}\right)^2\right]\right\}$$

$\Downarrow$ 将等式右侧平方拆开

$$\text{原式}=k\cdot\exp\left\{-\frac{1}{2}\left(\frac{x_t^2-2\sqrt{\alpha_t}x_tx_{t-1}+\alpha_tx_{t-1}^2}{1-\alpha_t}+\frac{x_{t-1}^2-2\sqrt{\overline{\alpha_{t-1}}}x_0x_{t-1}+\overline{\alpha_{t-1}}x_0^2}{1-\overline{\alpha_{t-1}}}-\frac{\left(x_t-\sqrt{\overline{\alpha_t}}x_0\right)^2}{1-\overline{\alpha_t}}\right)\right\}$$

$\Downarrow$ 合并等式右侧 x_{t-1} 的同类项

$$\text{原式}=k\cdot\exp\left\{\underbrace{-\frac{1}{2}\left[\left(\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha_{t-1}}}\right)x_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}}{1-\alpha_t}x_t+\frac{\sqrt{\overline{\alpha_{t-1}}}}{1-\overline{\alpha_{t-1}}}x_0\right)x_{t-1}+C\left(x_t,x_0\right)\right]}_{\text{这部分就是高斯分布中的}-\frac{1}{2}\left(\frac{1}{\sigma^2}x^2-\frac{2\mu}{\sigma^2}x+\frac{\mu^2}{\sigma^2}\right)}\right\}$$

$$\xlongequal{\text{完全平方公式}}k\cdot\exp\left\{-\frac{1}{2}\cdot\left[\frac{x_{t-1}-\left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}x_t+\frac{\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)}{1-\overline{\alpha_t}}x_0\right)}{\sqrt{\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha_{t-1}}}}}\right]^2\right\}$$

$$=\exp\left(-\frac{1}{2}\cdot\left[\frac{(x-\mu)}{\sigma}\right]^2\right)\tag{5.7}$$

其中,$\mu=\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}x_t+\frac{\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)}{1-\overline{\alpha_t}}x_0$,$\sigma=\sqrt{\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha_{t-1}}}}$

通过以上化简, 我们轻易的得到的 $p_\theta(x_{t-1}|x_t,x_0)$ 所服从的分布实际上也是一个高斯分布, 即

$$p_\theta(x_{t-1}|x_t,x_0)\sim\mathcal{N}(\mu,\sigma^2)\sim\mathcal{N}\left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}x_t+\frac{\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)}{1-\overline{\alpha_t}}x_0,\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha_{t-1}}}\right)$$

观察方差 $\sigma^2=\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha_{t-1}}}$ 可以发现, 其中所有的值均为常数, 是一个可以直接计算的值

观察均值 $\mu=\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}x_t+\frac{\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)}{1-\overline{\alpha_t}}x_0$, 可以发现所有的 α 均是已知值,x_t 也是已知值, 而整个均值的式子中, 唯一一个不知道的值为 x_0. 如果能得知 x_0 就能很快的进行计算了.

实际上, 我们在4.2中有扩散过程的公式4.11如下

$$x_t=\sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{a_t}}\widetilde{z},\quad\text{其中}\widetilde{z}\sim\mathcal{N}(0,1)$$

从这个式子中, 我们可以通过移项的方式，反推出 x_0，即 $x_0=\frac{1}{\sqrt{\overline{\alpha_t}}}\left(x_{t-1}-\sqrt{1-\overline{\alpha_t}}\widetilde{z}\right)$，用这个值替换 μ 中的 x_0, 有如下结果：

$$\begin{aligned}\mu&=\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}x_t+\frac{\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)}{1-\overline{\alpha_t}}x_0\\&=\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}x_t+\frac{\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)}{1-\overline{\alpha_t}}\cdot\frac{1}{\sqrt{\overline{\alpha_t}}}\left(x_{t-1}-\sqrt{1-\overline{\alpha_t}}\widetilde{z}\right)\\&=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\widetilde{z}\right)\end{aligned}\tag{5.8}$$

即

$$\mu=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\widetilde{z}\right)$$

整个式子中仅有 $\tilde{z}$ 是一个不可直接获得的值了. 为何会出现一个不能直接获得的 $\tilde{z}$ 呢？回顾上述过程发现, 这个 $\tilde{z}$ 是在为了消去 x_0 时引入的一个值. 这个 $\tilde{z}$ 表示的是在获得 x_t 的过程中, 向 x_0 中加入的噪声.

为了能够获得这个噪声, 原文使用了一个神经网络进行预测.

在使用神经网络获得 $\tilde{z}$ 后,μ 与 σ^2 都变的可以计算, 从而 $p_\theta(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\mu, \sigma^2)$ 就变成了可以直接获得的高斯分布. 从而 x_{t-1} 就可以轻松的从这个分布中采样获取了. 这个神经网络就是3.1逆扩散过程中所说的 Denoise 网络，也即我们下面将要介绍的噪声预测器。

6 噪声预测器的训练

在 DDPM 中, 使用 UNet 网络进行噪声图像的预测, 并采样极大似然函数作为损失函数进行预测。

首先我们介绍一下极大似然函数与极大似然估计。

6.1 极大似然函数与极大似然估计

实际上，这一部分对于没有思考过极大似然估计背后原理的读者可能有些难懂，所以本人推荐观看 bilibili 网站中视频博主“小崔说数”关于极大似然估计的讲解[10]，此处简单记录关于视频内容的理解。

极大似然估计（Maximum Likelihood Estimation，MLE）是统计学中用于估计模型参数的一种方法，这个方法适用于已经得知样本所服从的分布时，估计这个分布中参数的情况。它基于这样一个简单而强有力的想法：小概率事件在现实中几乎不发生。因此，当我们观测到某种事件发生时，这个事件在其所服从的分布中应该是一个大概率事件。

具体来说，假设我们有一个概率分布模型，它依赖于一些未知参数，我们的任务是根据观测数据来估计这些参数。假设我们观测到了 n 个独立同分布的数据点 $x_1, x_2, \ldots, x_n$，并且我们知道数据服从一个参数为 θ 的概率分布 $P(x|\theta)$。

极大似然估计的核心思想是选择使得观测数据出现的概率最大的参数 θ。这意味着我们需要找到一个参数 θ 使得所有观测到的事件的联合概率最大化。

1. 似然函数：首先，我们定义似然函数（Likelihood Function），它表示在参数 θ 下，观测数据 $x_1, x_2, \ldots, x_n$ 出现的联合概率。对于独立同分布的数据，似然函数可以表示为：

$$L(\theta) = P(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$$

2. 最大化似然函数：我们的目标是找到参数 θ 使得似然函数 $L(\theta)$ 最大。由于乘积运算复杂，我们通常对似然函数取对数，将其转化为对数似然函数（Log-Likelihood Function）：

$$\ell(\theta) = \log L(\theta) = \log \left(\prod_{i=1}^{n} P(x_i|\theta) \right) = \sum_{i=1}^{n} \log P(x_i|\theta)$$

取对数不会改变最大值的位置，但使得计算更为简便。

3. 求解最优参数：我们通过求解对数似然函数 $\ell(\theta)$ 的最大值来找到最优参数 θ。这通常涉及对 θ 求导，并找到导数为零的点，即：

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

导数为 0 的点为函数的极值点，当 θ 为极值点时，对应的 $\ell(\theta)$ 便可能达到最大值，及联合概率密度达到最大值。这说明，当 θ 为极值点，从这个概率分布中取得样本为我们所观测数据的概率达到最大值

极大似然估计方法可以解释为：在所有可能的参数值中，我们选择这样一个参数，使得实际观测到的数据在该参数下的出现概率最大。换句话说，这个参数使得观测到的数据最符合模型的假设。因此，极大似然估计本质上是在寻找最能解释观测数据的参数。

总之，极大似然估计是一种通过最大化观测数据在模型下出现的概率来估计参数的方法。它基于一个直观的想法，即观测到的事件应该是高概率事件，因此选择使观测数据出现概率最大的参数，就是最符合数据的参数。

6.2 利用极大似然函数进行参数迭代

我们知道, 极大似然估计在参数估计时, 使用的是这样的想法：小概率事件在现实中几乎不发生. 所以如果我们观测到某种事件发生了, 则说明这个事件所服从的分布中, 已经发生的这个事件一定是一个大概率事件. 能将所有已发生的事件的概率都最大化的参数, 也就是使似然函数最大的参数, 则是我们需要的参数.

从另一个角度说, 如果似然函数 L 越大, 则说明当前取得的参数越合理.

所以我们使用似然函数 L 作为似然函数, 将最大化似然函数 $\max L$ 作为训练的目标.

6.2.1 复习极大似然函数的构建过程

我们先用一个简单例子复习一下极大似然函数的构建过程. 如果你对极大似然估计很熟悉的话, 可以跳过这一部分.

如果下一部分中构建似然函数的过程难以理解, 可以与这一个简单的例子进行类比.

1. 现在有这样一个场景
   (a) 我们有一个装了不知道多少个黑白小球的袋子 (是一个满足二项分布的模型)
      i. 取得黑球的概率为 θ, 取得白球的概率为 $1-\theta$
   (b) 现在进行 5 次采样, 有结果如下：
      i. 采样 1：黑
      ii. 采样 2：黑
      iii. 采样 3：黑
      iv. 采样 4：白
      v. 采样 5：白
   (c) 问这个二项分布的参数 θ 是多少
2. 采样与二项分布
   (a) 对于采样 1, 摸到了黑球, 在二项分布中有概率为 θ
   (b) 对于采样 2, 摸到了黑球, 在二项分布中有概率为 θ
   (c) 对于采样 3, 摸到了黑球, 在二项分布中有概率为 θ
   (d) 对于采样 4, 摸到了白球, 在二项分布中有概率为 $1-\theta$
   (e) 对于采样 5, 摸到了白球, 在二项分布中有概率为 $1-\theta$
   (f) 对于整个五次采样结果为 (黑、黑、黑、白、白) 的概率为：$\theta^3 \cdot (1-\theta)^2$

3. 极大似然估计的思想

   (a) 既然五次采样出现了这样的结果, 所以我们认为发生这种 (黑、黑、黑、白、白) 情况的概率应该是最大的 (因为小概率事件不可能发生)

4. 极大似然估计

   (a) 所以我们认为, θ 一定能使采样概率 $\theta^3 \cdot (1-\theta)^2$ 取到最大值.

   (b) 也即, 一个能使概率 $\theta^3 \cdot (1-\theta)^2$ 达到最大值的 θ 是一个合理的 θ

   (c) 所以应该求 $\theta^3 \cdot (1-\theta)^2$ 取最大值时, θ 的取值.

   (d) 因为 $\theta^3 \cdot (1-\theta)^2$ 是一个高次幂的式子, 难以计算, 我们我们转而计算 $L(\theta) = \log\left[\theta^3 \cdot (1-\theta)^2\right]$

   (e) $L(\theta) = \log\left[\theta^3 \cdot (1-\theta)^2\right]$ 即为似然函数

6.2.2 扩散模型场景与极大似然估计

当我们复习了一个简单情形的似然估计的用法后, 可以快速的类比到当前的扩散模型任务中.

1. 现在我们有扩散模型的场景

   (a) 我们有一个装了不知道多少个各异图像的图像集 (是一个满足 $p_\theta(x_{t-1}|x_t)$ 的模型)

      i. 取得图像 x_t 的概率为 $p_\theta(x_{t-1}|x_t)$, $t = 1, 2, 3, \cdots, T-1$

      ii. 取得图形 x_T 的概率为 $p(x_T) \sim \mathcal{N}(0, 1)$

   (b) 现在进行 T 次采样, 有结果如下

      i. 采样 0：x_0

      ii. 采样 1：x_1

      iii. 采样 2：x_3

      iv. $\cdots$

      v. 采样 t：x_t

      vi. $\cdots$

      vii. 采样 T：x_T

   (c) 问这个概率模型 $p_\theta(x_t|x_{t+1})$ 中的 θ 是多少

2. 采样与二项分布

   (a) 对于采样 0, 获得了图像 x_0, 在分布中有概率 $p_\theta(x_0|x_1)$

   (b) 对于采样 1, 获得了图像 x_1, 在分布中有概率 $p_\theta(x_1|x_2)$.

   (c) $\cdots$

   (d) 对于采样 t, 获得了图像 x_t, 在分布中有概率 $p_\theta(x_t|x_{t+1})$.

   (e) $\cdots$

   (f) 对于采样 $T-1$, 获得了图像 x_{T-1}, 在分布中有概率 $p_\theta(x_{T-1}|x_T)$

   (g) 对于采样 T, 获得了图像 x_T, 在分布中有概率 $p(x_T) \sim \mathcal{N}(0, 1)$.

(h) 因为 x_T 是直接从标准高斯分布中获取的, 所以下标中没有 θ

(i) 对于整个 T 次采样结果为 $(x_0, x_1, x_2, \cdots, x_T)$ 的概率为

$$p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p(x_0|x_1)$$

3. 极大似然估计的思想

(a) 既然 T 次采样出现了这样的结果, 所以我们认为发生这种 $(x_0, x_1, x_2, \cdots, x_T)$ 情况的概率应该是最大的 (因为小概率事件不可能发生)

4. 极大似然估计

(a) 所以我们认为, θ 一定能使采样概率 $p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1)$ 取到最大值.

(b) 也即, 一个能使上述概率达到最大值的 θ 是一个合理的 θ

(c) 所以应该求 $p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1)$ 取最大值时, θ 的取值.

(d) 因为 $p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1)$ 是一个高次幂的式子, 难以计算, 我们我们转而计算 $L(\theta) = \log\left[p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1)\right]$

(e) $L(\theta) = \log\left[p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1)\right]$ 即为似然函数

实际上, 我们不可能简单的仅仅采样一个 x_0, 数据集中的所有图像都应该是一个可能的 x_0
所以我们要对式子 $p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p(x_0|x_1)$ 进行积分, 有如下结果

$$P_\theta(x_0) = \int_{x_1:x_T} p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1) dx_1 : x_T \tag{6.1}$$

所以有对数似然函数

$$L(\theta) = \log P_\theta(x_0) = \log\left(\int_{x_1:x_T} p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1) dx_1 : x_T\right) \tag{6.2}$$

但实际上, 这个似然函数的计算比较困难, 所以在训练时不会直接进行计算.
那么该如何计算这个似然函数呢？请看下面的部分.

6.2.3 最大化似然函数的下界

由于 DDPM 的过程有些复杂, 一共经过了 t 次的去噪, 比较复杂. 所以我们构建一个只有一个加噪过程的 DDPM 模型用来推导 (实际上这个「只有一个加噪过程的 DDPM」就是变分自编码器 VAE)

单个 Denoise 过程的 DDPM

由于只有一个加噪过程, 所以有似然函数如下：

$$\log P_\theta(x) = \log \int_{x_1} p(x_1) p_\theta(x_0|x_1) dx_1 \tag{6.3}$$

从而可以有以下推导：

$$
\begin{aligned}
\log P_\theta(x) &= \log \int_{x_1} p(x_1) p_\theta(x_0|x_1) dx_1 \\
&= \log \int_{x_1} p_\theta(x_0, x_1) dx_1 \\
&= \log \int_{x_1} \frac{q(x_1|x_0)}{q(x_1|x_0)} p_\theta(x_0, x_1) dx_1 \\
&= \log \int_{x_1} q(x_1|x_0) \frac{p_\theta(x_0, x_1)}{q(x_1|x_0)} dx_1 \\
&= \log \mathbb{E}_{q(x_1|x_0)} \left[\frac{p_\theta(x_0, x_1)}{q(x_1|x_0)} \right] \\
&\geq \mathbb{E}_{q(x_1|x_0)} \left[\log \frac{p_\theta(x_0, x_1)}{q(x_1|x_0)} \right] \qquad \text{[琴生不等式, 如果函数}\varphi\text{为凹函数, 则有}\varphi\left(\mathbb{E}(x)\right) \geq \mathbb{E}\left(\varphi(x)\right)\text{]}
\end{aligned}
\tag{6.4}
$$

我们将 $p_\theta(x_0, x_1)$ 替换为 $p_\theta(x_0|x_1)p(x_1)$, 则有

$$
\begin{aligned}
\mathbb{E}_{q(x_1|x_0)} \left[\log \frac{p_\theta(x_0, x_1)}{q(x_1|x_0)} \right] &= \mathbb{E}_{q(x_1|x_0)} \left[\log \frac{p_\theta(x_0|x_1)p(x_1)}{q(x_1|x_0)} \right] \\
&= \mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1) + \log p(x_1) - \log q(x_1|x_0) \right] \\
&= \mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1) \right] - \mathrm{KL}\left[q(x_1|x_0) \| p(x_1) \right]
\end{aligned}
\tag{6.5}
$$

这个式子分为了两项, 其中

1. 第一项 $\mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1) \right]$ 是一个期望值，它表示的是给定潜在变量 x_1 后生成原始图像 x_0 的对数似然。这个期望, 也即其下标 $q(x_1|x_0)$ 表示前向过程, 是一个已知的过程, 所以我们可以通过采样 x_1 来近似计算这个期望值

2. 第二项 $\mathrm{KL}\left[q(x_1|x_0) \| p(x_1) \right]$ 是是两个已知分布之间的 KL 散度。而经过上面的计算, 我们知道 $q(x_1|x_0)$ 与 $p(x_1)$ 均服从正态分布, 这使得 KL 散度可以解析计算 (两个正态分布之间的 KL 散度可以用公式直接计算).

从而整个变分下界可以计算. 通过最大化这个变分下界的形式, 我们可以最大化似然函数, 找到最适合的 θ 值.

多个 Denoise 过程的 DDPM

我们只需要将「单个 Denoise 过程的 DDPM」中的 x_1 变成 $x_0 : x_T$ 即可得到 DDPM 的推导过程.
根据上面的推导, 我们知道 DDPM 有似然函数如下：

$$
\log P_\theta(x_0) = \log \left[\int_{x_1:x_T} p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1) dx_1 : x_T \right] \tag{6.6}
$$

从而有以下推导：

$$
\begin{aligned}
\log P_\theta(x_0) &= \log \left[\int_{x_1:x_T} p(x_T) \cdot p_\theta(x_{T-1}|x_T) \cdot p_\theta(x_{T-2}|x_{T-1}) \cdots p_\theta(x_1|x_2) \cdot p_\theta(x_0|x_1) dx_1 : x_T \right] \\
&= \log \left[\int_{x_1:x_T} p_\theta(x_0, x_1, \cdots, x_T) dx_1 : x_T \right] \\
&= \log \left[\int_{x_1:x_T} \frac{\prod_{t=1}^T q(x_t|x_{t-1})}{\prod_{t=1}^T q(x_t|x_{t-1})} p_\theta(x_0, x_1, \cdots, x_T) dx_1 : x_T \right] \\
&= \log \left[\int_{x_1:x_T} \prod_{t=1}^T q(x_t|x_{t-1}) \cdot \frac{p_\theta(x_0, x_1, \cdots, x_T)}{\prod_{t=1}^T q(x_t|x_{t-1})} dx_1 : x_T \right]
\end{aligned}
\tag{6.7}
$$

在上面这个推导中的第 3 个等式, 我们仿照 VAE 中的推导, 将被积函数变形为了 1· 被积函数的形式.

- 在「单个 Denoise 过程的 DDPM」中，这个 1 是分子分母均为 $q(x_1|x_0)$ 的分式.
  - q 表示的是前项过程, 也就是从 x_0 获得噪声图像 x_1 的过程
- 在「多个 Denoise 过程的 DDPM」中，这个 1 是分子分母均为 $\prod_{t=1}^T q(x_t|x_{t-1})$ 的分式
  - 同样的, q 表示的是前项过程, 也就是从 x_0 获得一系列噪声图的过程.
  - 与 VAE 中仅有一个加早过程不同, DDPM 中具有 T 个加噪过程, 即 $q(x_t|x_{t-1}), t = 1, 2, 3, \cdots, T$ 这 T 个加噪过程.
  - 所以需要将这 T 的加噪过程相乘, 也即 $\prod_{t=1}^T q(x_t|x_{t-1})$

对于上面都推导中的累乘 $\prod_{t=1}^T q(x_{t-1}|x_t)$, 我们可以有如下的化简:

$$
\prod_{t=1}^T q(x_{t-1}|x_t) = q(x_0|x_1) \cdot q(x_1|x_2) \cdots q(x_{T-1}|x_T) = q(x_1, \cdots, x_T|x_0) \tag{6.8}
$$

- 为了简便表示, 我们将化简后的结果 $q(x_1, \cdots, x_T|x_0)$ 中的 $x_1, x_2, \cdots, x_T$ 记做 $x_{1:T}$
  - 因此有 $q(x_1, \cdots, x_T|x_0)$ 可以记作 $q(x_{1:T}|x_0)$
- 类似的, 我们将分式中的分母 $p_\theta(x_0, x_1, \cdots, x_T)$ 记做 $p_\theta(x_{0:T})$

将这个累乘的化简继续代入公式进行推导, 有

$$
\begin{aligned}
\log P_\theta(x_0) &= \log \left[\int_{x_1:x_T} \prod_{t=1}^T q(x_{t-1}|x_t) \cdot \frac{p_\theta(x_{0:T})}{\prod_{t=1}^T q(x_{t-1}|x_t)} dx_1 : x_T \right] \\
&= \log \left[\int_{x_1:x_T} q(x_{1:T}|x_0) \cdot \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} dx_1 : x_T \right] \\
&= \log \mathbb{E}_{q(x_{1:T}|x_0)} \left[\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right] \\
&\geq \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right] \qquad \text{[琴生不等式, 如果函数}\varphi\text{为凹函数, 则有}\varphi\left(\mathbb{E}(x)\right) \geq \mathbb{E}\left(\varphi(x)\right)\text{]}
\end{aligned}
\tag{6.9}
$$

实际上, 我们真正想要计算的是包含参数 θ 的部分, 所以我们需要将不含参数 θ 的部分与含有 θ 的部分分开.

对于 $p_\theta(x_{0:T}) = p_\theta(x_0, x_1, \cdots, x_T)$ 而言, $p(x_T)$ 是一个与 θ 无关的量, 因为 x_T 是直接从标准正态分布中获取的. 所以我们将 $p_\theta(x_{0:T})$ 替换为 $p(x_T) \cdot \prod_{t=1}^{T-1} p_\theta(x_{t-1}|x_t)$, 有以下推导

$$\begin{aligned}
\log P_\theta(x_0) &\geq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_T)\cdot\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T) + \sum_{t=1}^{T}\log\left[p_\theta(x_{t-1}|x_t)\right] - \log q(x_{1:T}|x_0)\right]
\end{aligned} \tag{6.10}$$

此处分解联合分布 $p(x_T)\cdot\prod_{t=1}^{T-1} p_\theta(x_{t-1}|x_t)$ 使我们能够逐步处理每一个时间步的生成过程，有助于将整体问题分解为多个子问题

而 $q(x_{1:T}|x_0)$ 则是表示前向过程, 是一个与 θ 无关的值. 同样的, 我们希望分解这个联合分布以逐步处理每一个时间步的生成过程, 将问题分成多个子问题, 所以可以进行如下替换：

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}|x_0) \tag{6.11}$$

因此有：

$$\begin{aligned}
\log P_\theta(x_0) &\geq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T) + \sum_{t=1}^{T}\log\left[p_\theta(x_{t-1}|x_t)\right] - \log q(x_{1:T}|x_0)\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T) + \sum_{t=1}^{T}\log\left[p_\theta(x_{t-1}|x_t)\right] - \sum_{t=1}^{T}\log\left[q(x_t|x_{t-1}|x_0)\right]\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T) + \sum_{t=1}^{T}\log\left[\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}|x_0)}\right]\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T)\right] + \mathbb{E}_{q(x_{1:T}|x_0)}\left[\sum_{t=1}^{T}\log\left[\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}|x_0)}\right]\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T)\right] + \sum_{t=1}^{T}\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left[\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}|x_0)}\right]\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T)\right] + \sum_{t=1}^{T}\mathbb{E}_{q(x_{1:T}|x_0)}\left[D_{KL}\left(q(x_t|x_{t-1})\|p_\theta(x_{t-1}|x_t)\right)\right]
\end{aligned} \tag{6.12}$$

这个式子分为了两项, 其中

1. 第一项 $\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T)\right]$ 是一个与 θ 无关的值, 在优化时可以忽略。
2. 第二项 $\sum_{t=1}^{T}\mathbb{E}_{q(x_{1:T}|x_0)}\left[D_{KL}\left(q(x_t|x_{t-1})\|p_\theta(x_{t-1}|x_t)\right)\right]$ 是已知分布之间的 KL 散度的和。而经过上面的计算, 我们知道 $q(x_t|x_{t-1})$ 与 $p_\theta(x_{t-1}|x_t)$ 均服从正态分布, 这使得 KL 散度可以解析计算 (两个正态分布之间的 KL 散度可以用公式直接计算).

从而整个公式变得容易计算。

实际上, 本人推导的结果与原始论文中推导的结果不太相同. 原始论文中的结果中包含了 3 项, 而本人的推导仅有 2 项. 导致这种区别的原因与说明如下：

1. 原因：原始论文中, 为了能够让其更好运算, 并获得更好的效果, 将 KL 散度进一步细分；而本人并未做这样的工作, 主要是为了降低理解门槛, 并与 VAE 的推导同步.

2. 说明：虽然结果不同, 但理解起来并无区别. 也即, 本人的推导更偏向于理解, 而原始论文的推导更偏向于实践.

至此，DDPM[1]中的两个重要过程:『扩散过程』、『逆扩散过程』的推导与原理讲解完毕，DDPM[1]成为一个可以理解且实际应用的模型。

在上述两个过程的指导下，结合『噪声预测器』，使得 DDPM[1]为基础与核心的『基于扩散的图像生成模型』成为当前最为流行的图像生成模型。

7 总结

这篇文章主要讨论了『基于扩散的图像生成模型』和『DDPM 原理』，从头到尾详尽地介绍了其中涉及到的过程与各个方面。文章开篇先提了一下，近年来人工智能和深度学习技术发展的多么迅猛，然后指出扩散模型在图像生成领域是多么重要。接着就开始详细描述了这个模型的工作流程。

在解释基于扩散的图像生成模型时，文章用了很多篇幅来描述所谓的『扩散过程』和『逆扩散过程』。『扩散过程』就是把清晰的图像一步步加噪声变模糊的过程，而『逆扩散过程』则是反过来，从噪声图像一步步去噪还原清晰图像的过程。为了实现逆扩散，需要大量的训练数据，于是扩散过程就负责生成这些训练数据。

然后文章深入到 DDPM[1]的原理。通过各种数学公式推导，详细解释了扩散和逆扩散的具体实现步骤。特别是扩散过程，文章讲了很多如何一步步加噪，从公式 4.1 到 4.10，不断简化和变换，最后得出结论，原来多次加噪的效果可以简化为一次噪声添加。

逆扩散过程的部分同样详细，主要是通过贝叶斯公式和高斯分布的推导，解释了如何从带噪声的图像一步步恢复出清晰图像。为了实现这个过程，文章提出了噪声预测器的概念，并解释了如何通过极大似然估计来训练这个预测器。同时为了能够使读者无压力的理解极大似然估计，本文还从其本质原理与简单例子的角度入手，带领大家复习了极大似然估计的过程。

文章在最后总结道，扩散模型和 DDPM 为图像生成领域提供了新的思路和理论支持，既有实用价值又有学术意义。本人希望读者能认真阅读这篇文章，尽管公式和推导看起来复杂，但实际上并不涉及非常高深的知识。DDPM 的推导过程中使用最多的公式是贝叶斯公式和 KL 散度的定义与极大似然估计的原理，所以并不是难以理解的知识。

在这篇文章的写作过程中，本人花费了大约两个星期的时间，认真地理解 DDPM 的原理，并将自己的思路和理解记录下来。我希望通过这种详尽的解释和推导，能够帮助更多的研究人员理解并应用扩散模型和 DDPM，希望本文能为图像生成领域的研究者提供有价值的参考和指导，促进该领域的进一步发展和创新。

如果在阅读时遇到名词混淆的情况，请阅读7名词辨析作为参考。

名词辨析

1. 扩散过程：指噪声逐渐扩散到整张图像的过程。
2. 扩散：即扩散过程。
3. 基于扩散的图像生成模型：以扩散过程与逆扩散过程作为基本原理的图像生成模型。
4. DDPM：指文章 Denoising Diffusion Probabilistic Model[1] 中提出的模型，是最早的『基于扩散的图像生成模型』。

参考文献：

[1] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models: arXiv:2006.11239[M]. arXiv, 2020.

[2] Luo C. Understanding Diffusion Models: A Unified Perspective: arXiv:2208.11970[M]. arXiv, 2022.

[3] 李宏毅. 【生成式 AI】Diffusion Model 概念讲解 (1/2): 第 1 卷[M]. BiliBilli:BV14c411J7f2,p1, 2023.

[4] 李宏毅. 【生成式 AI】Diffusion Model 概念讲解 (2/2): 第 2 卷[M]. BiliBilli:BV14c411J7f2,p2, 2023.

[5] 李宏毅. 【生成式 AI】Diffusion Model 原理剖析 (1/4): 第 3 卷[M]. BiliBilli:BV14c411J7f2,p3, 2023.

[6] 李宏毅. 【生成式 AI】Diffusion Model 原理剖析 (2/4): 第 4 卷[M]. BiliBilli:BV14c411J7f2,p4, 2023.

[7] 李宏毅. 【生成式 AI】Diffusion Model 原理剖析 (3/4): 第 5 卷[M]. BiliBilli:BV14c411J7f2,p5, 2023.

[8] 李宏毅. 【生成式 AI】Diffusion Model 原理剖析 (4/4): 第 6 卷[M]. BiliBilli:BV14c411J7f2,p6, 2023.

[9] 郑英林. Diffusion Models：生成扩散模型[M]. https://yinglinzheng.netlify.app/diffusion-model-tutorial/, 2023.

[10] 小崔说数. 十分钟搞定最大似然估计[M]. BiliBilli:BV1Hb4y1m7rE, 2021.