Diffusion 相关技术调研与探讨

Jianhua Guo

摘要

本文结合《Understanding Diffusion Models: A Unified Perspective》等文献，深入调研并系统探讨了扩散模型的数学基础、核心机制及其与变分自编码器和分数基生成模型的联系。文章首先阐述扩散模型的基本原理，说明变分扩散模型作为马尔可夫层次变分自编码器的特殊形式，通过逐步加噪和去噪实现数据生成。接着分析了其三种等价优化目标：预测原始图像、源噪声和分数函数，并探讨了分数基生成模型通过朗之万动态采样的机制。对于条件生成任务，文章介绍了分类器引导和无分类器引导方法，分析其在生成结果控制与样本多样性之间的平衡。最後，总结了扩散模型的优势与局限性，指出其在可解释性、潜变量维度和采样效率方面的挑战，并展望了层次变分自编码器的未来发展潜力。

关键词：扩散模型；变分自编码器；分数基生成模型；条件生成

1. Diffusion 模型技术路线

扩散模型的理论灵感源于非平衡热力学，通过模拟数据的逐步加噪和去噪过程实现生成。Jascha Sohl-Dickstein 等人于 2015 年首次提出基于非平衡热力学的深度无监督学习框架，奠定了扩散模型的理论基础。扩散模型可以被视为马尔可夫层次变分自编码器的特殊形式，通过定义固定或可学习的加噪过程，将输入数据逐步转化为标准高斯噪声，并通过逆向去噪过程生成样本。这种方法的核心在于优化证据下界，通过最大化似然估计来学习数据的潜在分布。

扩散模型的数学推导通常基于两个关键假设：一是潜变量维度与数据维度相同，确保模型能够捕捉输入的完整信息；二是编码器采用预定义的线性高斯模型，使得加噪过程可控且易于优化。此外，扩散模型与变分自编码器的联系在于它们均通过潜变量建模数据分布，但扩散模型的潜变量通常是数据的噪声版本，而非压缩的语义表示。

分数基生成模型是扩散模型的另一重要视角，通过学习数据分布的分数函数来实现生成。扩散模型与分数基生成模型的等价性指出扩散模型的优化目标可以通过预测原始图像、源噪声或分数函数来实现。Yang Song 等人提出通过朗之万动态进行采样，利用分数函数引导数据从随机噪声逐步逼近真实分布。为解决分数函数在低密度区域和高维空间中的问题，研究者引入了多尺度噪声扰动，通过逐步增加高斯噪声来增强模型的鲁棒性和采样效率。

分数基生成模型的优势在于无需显式计算归一化常数，但其挑战在于分数函数在低维流形上的定义不清以及低密度区域的训练信号不足。扩散模型通过多层次噪声扰动有效缓解了这些问题，使得模型在图像生成等复杂任务中表现更佳。

条件生成是扩散模型的重要应用方向，广泛用于文本到图像生成、图像超分辨率等任务。两种主要的条件生成引导机制分别是分类器引导和无分类器引导。分类器引导通过训练一个额外的分类器来预测条件信息，利用其梯度调整分数函数，从而控制生成结果。然而该方法需额外训练分类器，且对噪声输入的适应性较差。无分类器引导则通过在单一模型中联合学习条件和无条件分布，利用条件信息随机丢弃来模拟无条件生成，从而实现更高效的条件控制。

近期研究进一步扩展了条件扩散模型的应用。

DALL-E 2 和 Imagen 通过结合文本编码和扩散模型，显著提升了文本条件图像生成的逼真度和多样性。Cascaded Diffusion Models 则通过多阶段扩散过程实现了高分辨率图像生成，展示了扩散模型在复杂任务中的潜力。

扩散模型在多个领域展现了强大的生成能力。在图像生成方面，去噪扩散概率模型显著提升了生成图像的质量，超越了传统的生成对抗网络。在语音合成领域，DiffWave 通过扩散模型实现了高质量的语音波形生成。在跨模态任务中，GLIDE 和 Stable Diffusion 通过结合文本和图像信息，实现了高效的文本引导图像生成。

此外，扩散模型还在视频生成、分子设计和医学图像处理等领域得到应用。Video Diffusion Models 通过扩展扩散过程到时序数据，实现了高质量视频生成；DiffDock 则将扩散模型应用于分子对接任务，展示了其在科学计算中的潜力。

尽管扩散模型在生成质量上表现出色，但仍面临若干挑战。首先，采样过程通常需要多次迭代，导致生成速度较慢。研究者提出了一些加速方法，如 DDIM 通过非马尔可夫扩散过程减少采样步骤。其次，扩散模型的潜变量通常是高维噪声，缺乏语义可解释性，限制了其在需要结构化表示的任务中的应用。此外，条件生成中的引导机制需要在生成质量和样本多样性之间权衡，需进一步优化。

2. Diffusion **模型的原理和机制**

扩散模型（Diffusion Models）是一种强大的生成模型，通过前向过程逐步向数据添加高斯噪声，再通过反向过程从噪声中恢复原始数据，生成高质量样本。

2.1. 扩散模型基本概念

扩散模型是马尔可夫层次变分自编码器（MHVAE）的特例，具有以下限制：

- 潜变量维度与数据维度相同。
- 编码器为预定义的线性高斯模型。
- 高斯参数随时间变化，最终潜变量分布为标准高斯分布。

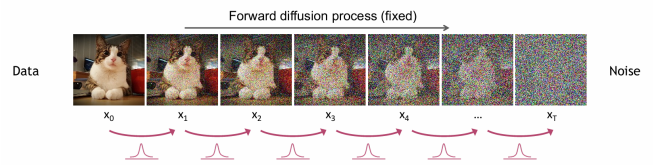

图 1. Forward Diffusion Process

2.2. 前向过程（加噪过程）

前向过程通过马尔可夫链逐步向数据添加高斯噪声。令原始数据为 $\boldsymbol{x}_0$，时间步 $t = 1, \ldots, T$，潜变量 $\boldsymbol{x}_t$ 由 $\boldsymbol{x}_{t-1}$ 生成，遵循高斯分布：

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathrm{I}) \tag{1}$$

其中 α_t 是噪声调度参数。联合分布为：

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \tag{2}$$

直接从 $\boldsymbol{x}_0$ 到 $\boldsymbol{x}_t$ 的分布为：

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathrm{I}), \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s \tag{3}$$

最终，$\boldsymbol{x}_T \approx \mathcal{N}(0, \mathrm{I})$。

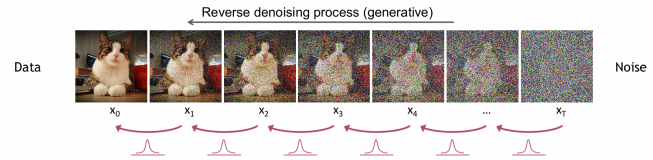

图 2. Reverse Denoising Process

2.3. 反向过程（去噪过程）

反向过程从纯噪声 $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathrm{I})$ 开始，逐步去噪生成 $\boldsymbol{x}_0$。联合分布为：

$$p(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \tag{4}$$

去噪分布建模为高斯分布：

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_q(t)) \tag{5}$$

目标是优化参数 θ，使 p_θ 逼近真实逆向分布 $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$。

2.4. 优化目标：证据下界（**ELBO**）

扩散模型通过最大化证据下界（ELBO）优化：

$$\log p(\boldsymbol{x}_0) = \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \tag{6}$$

引入变分分布 $q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)$，应用 Jensen 不等式：

$$\log p(\boldsymbol{x}_0) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \tag{7}$$

代入联合分布和后验分布，ELBO 为：

$$\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \tag{8}$$

优化目标等价于最小化 KL 散度 $D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \| p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t))$。

2.5. 三种等价优化目标

变分扩散模型可通过以下三种方式优化：

1. **预测原始图像** $\boldsymbol{x}_0$：学习神经网络 $\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t) \approx \boldsymbol{x}_0$。
2. **预测源噪声** $\boldsymbol{\epsilon}_0$：利用重参数化技巧：

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t} \boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_0 \tag{9}$$

推导出：

$$\boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}} \tag{10}$$

优化目标：

$$\arg\min_\theta \frac{1}{2\sigma_q^2(t)} \left\| \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} (\boldsymbol{\epsilon}_0 - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t)) \right\|_2^2 \tag{11}$$

3. **预测分数函数** $\nabla \log p(\boldsymbol{x}_t)$：利用 Tweedie 公式：

$$\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \frac{\boldsymbol{x}_t + (1 - \bar{\alpha}_t) \nabla \log p(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}} \tag{12}$$

优化目标：

$$\arg\min_\theta \frac{1}{2\sigma_q^2(t)} \left[\frac{1 - \alpha_t}{\bar{\alpha}_t} \left\| \boldsymbol{s}_\theta(\boldsymbol{x}_t, t) - \nabla \log p(\boldsymbol{x}_t) \right\|_2^2 \right] \tag{13}$$

分数函数与源噪声的关系：

$$\nabla \log p(\boldsymbol{x}_t) = -\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_0 \tag{14}$$

2.6. 基于分数的生成模型

扩散模型与基于分数的生成模型等价，分数函数 $\nabla \log p(\boldsymbol{x}_t)$ 表示增加对数似然的方向。使用朗之万动力学生成样本：

$$\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + c \nabla \log p(\boldsymbol{x}_i) + \sqrt{2c} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathrm{I}) \tag{15}$$

通过多尺度噪声扰动，定义：

$$p_{\sigma_t}(\boldsymbol{x}_t) = \int p(\boldsymbol{x}) \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}, \sigma_t^2 \mathrm{I}) d\boldsymbol{x} \tag{16}$$

优化目标：

$$\arg\min_\theta \sum_{t=1}^{T} \lambda(t) \mathbb{E}_{p_{\sigma_t}(\boldsymbol{x}_t)} \left[\left\| \boldsymbol{s}_\theta(\boldsymbol{x}, t) - \nabla \log p_{\sigma_t}(\boldsymbol{x}_t) \right\|_2^2 \right] \tag{17}$$

2.7. 条件扩散模型

条件扩散模型生成条件分布 $p(\boldsymbol{x}_{0:T} \mid y)$：

$$p(\boldsymbol{x}_{0:T} \mid y) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, y) \tag{18}$$

两种引导方法：

- **分类器引导**：分数函数为：

$$\nabla \log p(\boldsymbol{x}_t \mid y) = \nabla \log p(\boldsymbol{x}_t) + \gamma \nabla \log p(y \mid \boldsymbol{x}_t) \tag{19}$$

- **无分类器引导**：分数函数为：

$$\nabla \log p(\boldsymbol{x}_t \mid y) = \gamma \nabla \log p(\boldsymbol{x}_t \mid y) + (1 - \gamma) \nabla \log p(\boldsymbol{x}_t) \tag{20}$$

2.8. 优缺点

优点：

- 生成高质量样本，数学优雅。
- 多尺度噪声解决低维流形问题。

缺点：

- 采样需要多次迭代，计算成本高。
- 潜变量缺乏可解释性和压缩性。

3. Diffusion 模型的总结和探讨

扩散模型通过一个马尔可夫链结构，结合前向加噪和反向去噪过程，实现了从纯高斯噪声到高质量数据的生成。其核心思想是将数据逐步扰动为噪声，随后通过学习逆向过程恢复原始数据分布。优化过程以证据下界为基础，通过最小化变分分布与真实后验分布之间的差异，训练一个神经网络来预测去噪步骤。这种方法在数学上与基于分数的生成模型等价，允许通过朗之万动力学从噪声中逐步生成样本。条件扩散模型进一步引入控制信号，如文本或低分辨率图像，扩展了模型在生成任务中的灵活性，例如在文本引导图像生成中的出色表现。扩散模型的优雅数学框架和强大的生成能力使其成为生成式人工智能领域的核心支柱。

从理论视角看，扩散模型的成功源于其对复杂数据分布的建模能力，尤其是在处理高维数据时，通过多尺度噪声扰动克服了低维流形上的分数函数未定义问题。这种方法本质上通过构造一系列中间分布，平滑了数据空间的概率密度，使得生成过程更加稳定。然而，当前模型的潜变量仅为数据的噪声版本，缺乏语义结构，与变分自编码器相比，其表示能力受到限制。未来的理论创新可聚焦于设计非线性编码器或引入层次化潜变量，以捕捉更丰富的语义信息。例如，可以探索结合流模型的可逆变换特性，构建更灵活的潜空间表示。此外，扩散模型的连续时间形式通过随机微分方程描述，为噪声调度提供了新的自由度。如何优化 SDE 的参数以最小化训练成本，同时保持生成质量，是理论研究的开放问题。另一个潜在方向是统一扩散模型与其他生成模型的框架，通过融合各自优势，构建更鲁棒的生成模型。

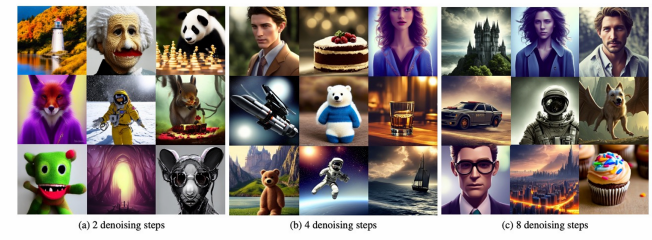

图 3. Progressive Distillation in Latent Space

在应用层面，扩散模型已在图像生成、音频合成和分子设计等领域展现了巨大潜力。例如，基于条件扩散模型的 DALL-E 2 和 Imagen 在文本引导图像生成中实现了接近真实的表现，显著超越了传统生成模型。然而，扩散模型的采样过程需要多次迭代，导致计算成本高，限制了其在实时场景（如视频生成或移动设备应用）中的部署。近期研究提出了快速采样方法，如基于确定性微分方程的去噪策略，通过减少采样步数显著提升效率，但可能以牺牲生成多样性为代价。此外，扩散模型对大规模训练数据和高性能计算资源的依赖使其在资源受限环境中的应用受到挑战。针对这一问题，知识蒸馏、模型压缩或设计轻量级架构成为研究热点。例如，可以通过将大模型的知识迁移到小型网络，保留生成质量的同时降低推理成本。在跨模态生成方面，扩散模型可进一步扩展到多模态任务，如联合图像、文本和音频生成，为虚拟现实和增强现实提供更丰富的生成能力。然而，条件生成中的多样性与控制精度之间的权衡仍需优化，尤其在需要高保真输出的场景中，如何确保生成样本既符合条件又保持多样性是一个关键问题。

展望未来，扩散模型的发展需应对多重挑战。首先，采样效率的提升仍是核心问题。尽管快速采样算法和 ODE 方法取得进展，但如何在极少步数内实现高质量生成仍需探索。一种潜在方向是自适应采样，通过动态调整去噪步数以适应不同数据复杂性。其次，潜变量的可解释性和压缩性是改进方向。当前的潜变量设计限制了模型在表示学习中的能力，未来可通过结合变分自编码器或自回归模型，构建更具语义意义的潜空间，用于生成控制或数据压缩。此外，扩散模型在跨领域应用中的泛化能力需进一步验证。例如，在医疗影像生成或金融时间序列建模中，如何适应稀疏数据或非平稳分布是一个挑战。针对资源受限场景，开发高效的轻量级扩散模型或利用迁移学习将是大势所趋。最后，扩散模型的成功启发了对生成模型统一理论的思考。通过整合扩散模型、GAN 和流模型的优势，可能催生更高效、灵活的生成框架，为人工智能在创意、科学和工业领域的应用开辟新路径。

总之，扩散模型以其数学优雅性和强大的生成

能力，重新定义了生成式人工智能的边界。尽管面临采样效率、潜变量可解释性和计算成本等挑战，其在多领域的突破性应用和理论上的扩展潜力表明，扩散模型将继续引领生成模型的研究。