生成式AI与随机微分方程 (Lec 4)

Conditional Image Generation

笔记整理：Gemini

MIT IAP 2025 | Jan 27, 2025

2025 年 7 月 30 日

目录

1 条件生成与引导 (Conditional Generation and Guidance)

到目前为止，我们讨论的都是**无条件生成**，即模型从数据分布 p_{data} 中随机采样，我们无法控制生成的内容。本节课的目标是实现**条件生成**，即根据给定的条件 y（例如一段文本描述），从条件数据分布 $p_{\text{data}}(z|y)$ 中采样。

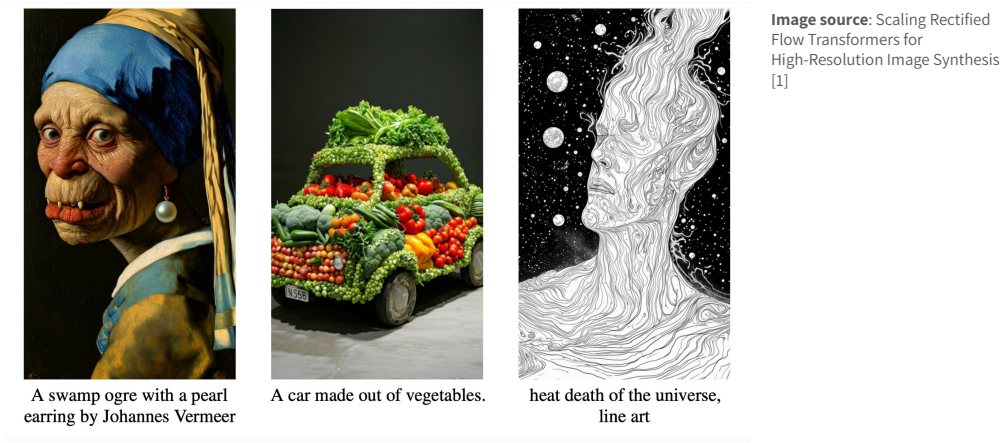

图 1: 条件生成的例子：根据文本提示生成高度具体的图像 (源自幻灯片 Lec4, Page 5)。

1.1 如何修改我们的框架？

为了实现条件生成，我们需要将之前的整个理论框架“条件化”。这非常直观，只需在所有相关的概率分布和向量场上加上条件 y 即可。

- **目标分布**: $p_{\text{data}}(z|y)$
- **边际概率路径**: $p_t(x|y)$
- **边际向量场 (训练目标)**: $u_t^{\text{target}}(x|y)$
- **神经网络模型**: $u_t^\theta(x, y)$ (注意，现在模型需要接收条件 y 作为额外输入)

相应的，我们的流匹配损失函数也需要更新，以包含对条件 y 的期望：

$$\mathcal{L}_{\text{CFM}}^{\text{guided}}(\theta) = \mathbb{E}_{(z,y)\sim p_{\text{data}},t\sim\text{Unif},x\sim p_t(\cdot|z)}\left[\|u_t^\theta(x,y) - u_t^{\text{target}}(x|z)\|^2\right]$$

1.2 分类器无关引导 (Classifier-Free Guidance, CFG)

直接使用上述方法训练出的模型虽然可以生成符合条件的图像，但有时效果并不惊艳。实践发现，一种名为**分类器无关引导 (Classifier-Free Guidance, CFG)** 的技术可以显著提升生成质量和与提示的符合度。

1.2.1 CFG的核心思想

CFG 的思想非常巧妙：它认为，一个好的“引导向量场” $u_t(x|y)$，应该是在“无引导向量场” $u_t(x)$ 的基础上，朝着更能体现条件 y 特征的方向进行“夸大”或“加强”。

修改后的引导向量场 $\tilde{u}_t(x|y)$ 由以下公式定义：

$$\tilde{u}_t(x|y) = (1-w)u_t^{\text{target}}(x) + w \cdot u_t^{\text{target}}(x|y)$$

或者等价地写成：

$$\tilde{u}_t(x|y) = u_t^{\text{target}}(x) + w \cdot \left(u_t^{\text{target}}(x|y) - u_t^{\text{target}}(x)\right)$$

- w 是**引导强度 (guidance scale)**。
- 当 $w = 1$ 时，退化为普通的条件生成。
- 当 $w > 1$ 时，模型会沿着“从无条件到条件”的方向 $(u_t(x|y) - u_t(x))$ 进行外插，从而生成更符合提示、特征更鲜明的图像。

1.2.2 CFG的训练与采样

为了使用CFG，我们需要一个模型能同时预测**条件向量场** $u_t^\theta(x, y)$ 和**无条件向量场** $u_t^\theta(x, \emptyset)$。这里的 $\emptyset$ 是一个特殊的“空”或“无条件”令牌。

- **训练**: 我们训练一个通用的条件模型 $u_t^\theta(x, y)$。在每个训练步骤中，我们以一定的概率（例如10%）将真实的条件 y 替换为无条件令牌 $\emptyset$。这样，同一个模型就同时学会了条件和无条件两种情况。
- **采样**: 在采样时，我们使用一个修改后的ODE，将模型对条件和无条件的预测结合起来。

Algorithm 1 分类器无关引导 (CFG) 采样算法

1: **需要:** 训练好的模型 $u_t^\theta(x, y)$, 引导强度 $w > 1$
2: 选择一个提示 y (例如“一只猫在烤蛋糕”)。
3: 初始化 $X_0 \sim p_{\text{init}}$。
4: 通过求解以下ODE来进行采样：

$$dX_t = \left[(1-w)u_t^\theta(X_t, \emptyset) + w \cdot u_t^\theta(X_t, y)\right] dt$$

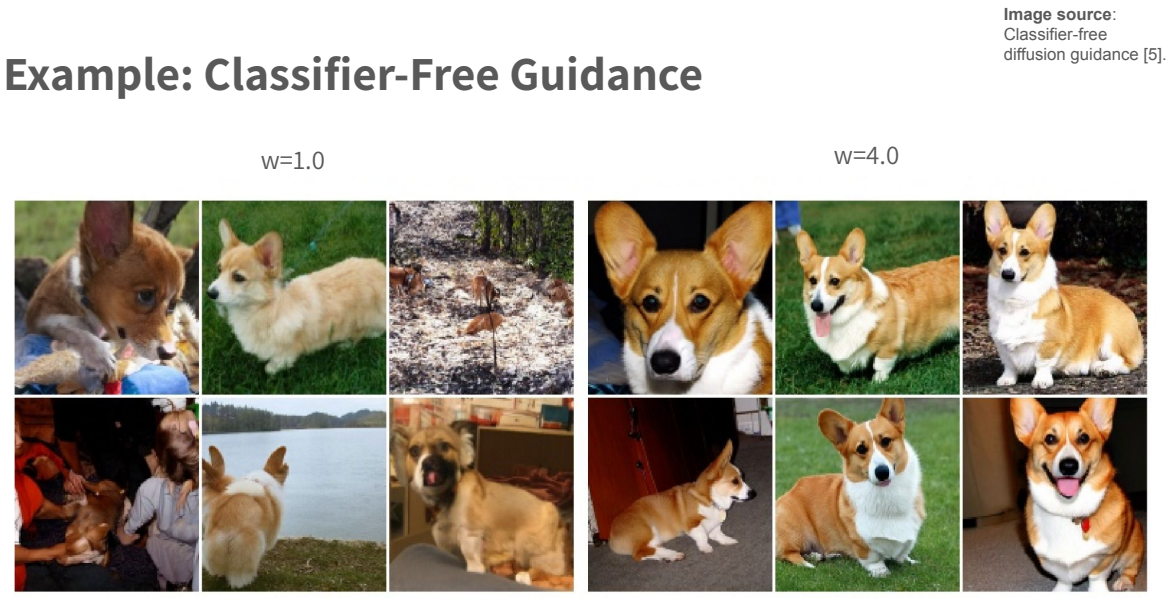

图 2: CFG效果对比：随着引导强度 w 的增加，生成的柯基犬图像质量更高、特征更典型 (源自幻灯片 Lec4, Page 13)。

2 图像生成模型的网络架构

理论和算法都有了，但我们的神经网络 $u_t^\theta(x, y)$ 具体应该长什么样呢？对于高维的图像数据，简单的多层感知机 (MLP) 显然是不够的。本节介绍了两种主流的架构。

2.1 U-Net 架构

U-Net 是在图像生成领域（尤其是扩散模型早期）占主导地位的卷积神经网络架构。

- **编码器-解码器结构**: 左侧的编码器通路通过一系列卷积和下采样操作，逐步提取图像的深层语义特征。右侧的解码器通路则通过上采样和卷积，将这些特征逐步恢复为与输入同尺寸的输出。
- **U形结构与跳跃连接 (Skip Connections)**: U-Net的精髓在于从编码器到解码器对应层级的“跳跃连接”（图中的虚线）。这些连接允许解码器直接利用编码器浅层的、高分辨率的细节特征，这对于恢复图像的精细纹理至关重要，避免了信息瓶颈。
- **条件注入**: 时间步 t 和条件 y 通常被编码成向量，然后通过仿射变换（或交叉注意力机制）注入到网络的各个残差块 (Residual Layer) 中，从而在不同尺度上指导生成过程。

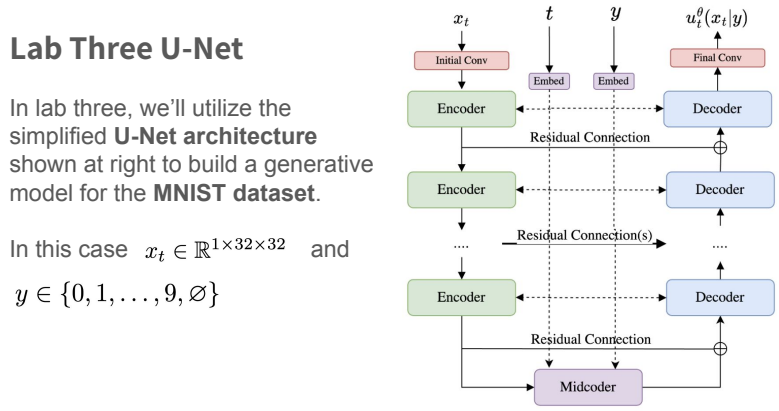

图 3: U-Net 架构示意图。它接收带噪声的图像 x_t、时间步 t 和条件 y 作为输入 (源自幻灯片 Lec4, Page 17)。

2.2 Diffusion Transformer (DiT) 架构

随着Transformer在NLP和CV领域的巨大成功，研究者们也将其引入了扩散模型，提出了Diffusion Transformer (DiT)。

- **抛弃卷积，拥抱注意力**: DiT完全摒弃了U-Net中的卷积操作，其核心是Transformer中的自注意力机制 (Self-Attention)。
- **图像分块 (Patchify)**: 与Vision Transformer (ViT) 类似，DiT首先将输入的图像（或特征图）分割成一系列不重叠的小块 (patches)。每个小块被线性投影成一个token。
- **全局信息交互**: 这些tokens（连同位置编码、时间编码和条件编码）被送入一系列Transformer模块中。自注意力机制使得模型能够捕捉图像中任意两个小块之间的长距离依赖关系，具有比卷积更好的扩展性 (scalability)。

2.3 潜在空间扩散 (Latent Diffusion)

直接在像素空间（例如 512x512 图像）上运行扩散模型计算成本极高。像Stable Diffusion这样的模型采用了一种更高效的策略：在压缩的**潜在空间 (Latent Space)** 中进行扩散。

工作流程:

1. 使用一个预训练好的**变分自编码器 (VAE)**。
2. **编码**: VAE的编码器 $\mathcal{E}$ 将高分辨率的真实图像 x 压缩成一个维度小得多的潜在表示 z。
3. **潜在空间扩散**: 所有的扩散和去噪过程（由U-Net或DiT完成）都在这个低维、信息密集的潜在空间中进行，极大地降低了计算复杂度。

Image sources: Vision transformer paper [2] (left), diffusion transformer paper [3] (right).

Diffusion Transformer (DiT)

Idea: Divide an image into **patches** and **attend** between the patches. Based on the **vision transformer** (ViT).

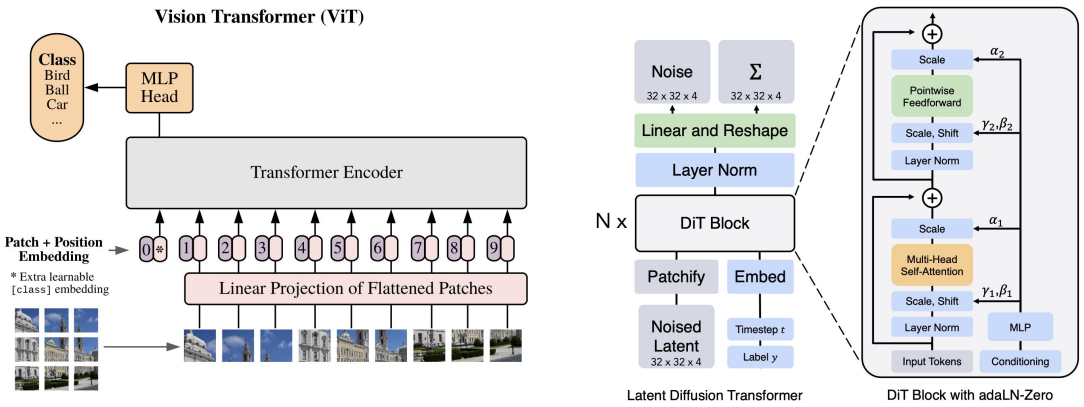

图 4: 左侧为Vision Transformer (ViT) 原理，右侧为Diffusion Transformer (DiT) 模块设计 (源自幻灯片 Lec4, Page 22)。

Image source: .High Resolution Image Synthesis with Latent Diffusion Models [4]

Generative Modeling in Latent Space

Idea: Train the generative model in the **latent space** of a pre-trained (variational) autoencoder.

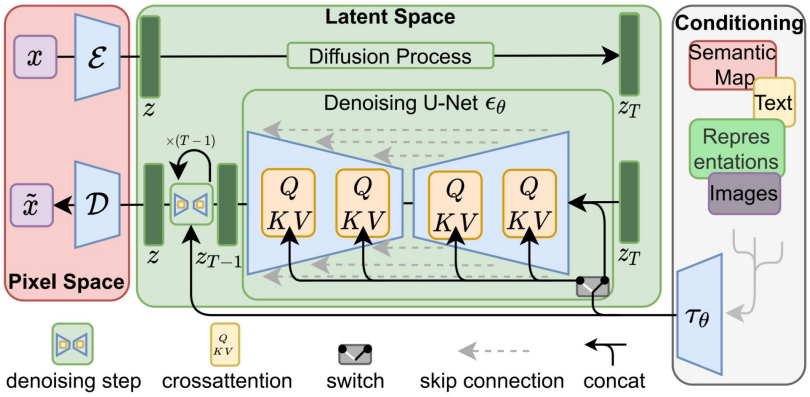

图 5: Latent Diffusion Model (LDM) 的工作流程 (源自幻灯片 Lec4, Page 23)。

4. **解码**: 生成过程结束后，使用VAE的解码器 $\mathcal{D}$ 将最终生成的潜在表示 $\tilde{z}$ 恢复成高分辨率的像素图像 $\tilde{x}$。

3 总结

本节课我们完成了从无条件生成到可控的条件生成的跨越。核心技术**分类器无关引导 (CFG)** 极大地提升了生成质量。同时，我们探讨了实现这些模型的两大主流架构——**U-Net** 和 **DiT**，以及通过**潜在空间扩散**来提升效率的关键技巧。至此，我们已经掌握了构建一个现代高性能图像生成模型所需的全套核心理论和架构知识。