基于 DreamBooth-LoRA 的文生图风格迁移模型的研究

郑思哲, 周健文, 周扬超, 赵猛, 纪冠州

{162100101, 162100103, 162100104, 162100106, 162150107}@nuaa.edu.cn

Abstract

近年来随着扩散模型的发展，涌现了一大批图片生成或者编辑的算法，其中一个重要的课题就是可控生成。在可控图片生成中一个相当有艺术性的研究课题是风格迁移问题: 旨在将一种图像的风格应用到另一张图像上, 同时保留后者的内容结构。该项目会给定一系列同风格的参考图片，给出指定的提示词指示生成相同风格的图片，主要借助 DreamBooth 或者 LoRA 对文生图扩散模型进行微调, 以实现高质量的风格迁移图片生成。实验代码见 https://github.com/ZszYmy9/NUAA-AI-Project。

项目介绍

风格迁移（Style Transfer）是计算机视觉和图像生成领域中的一个重要研究方向，旨在将一种图像的风格（如色彩、纹理、笔触等）应用到另一张图像上，同时保留目标图像的内容结构。这一技术在艺术创作、图像编辑、影视特效等领域具有广泛的应用前景。

在可控图片生成中，风格迁移不仅是一个技术挑战，更是一个极具艺术性的研究课题。通过风格迁移，我们可以将经典艺术作品的风格应用到现代照片上，或者将一种独特的视觉风格应用到一系列图像中，从而创造出具有一致风格的艺术作品。

本项目的主要目标是开发一种基于深度学习的风格迁移算法，能够在给定一系列同风格的参考图片和特定的提示词的情况下，生成具有相同风格的新图片。具体目标包括：

- 风格一致性：生成的图片在风格上应与参考图片高度一致，包括色彩、纹理、笔触等视觉元素。
- 内容保留：生成的图片应保留目标图像的内容结构，如物体的形状、位置、比例等。
- 可控性：用户可以通过提示词或其他控制参数灵活调整生成图片的风格和内容。
- 高效性：算法应能够在合理的时间内生成高质量的图片，适用于实时或近实时的应用场景。

我们先进行数据的搜集和预处理，对参考图片和目标图片进行预处理，包括尺寸调整、归一化等。设计基于深度学习的风格迁移模型，如卷积神经网络（CNN）、生成对抗网络（GAN）等，这里我们使用微调文生图扩散模型来实现我们的目标，在训练好的模型上，输入目标图片和指定的提示词，生成具有相同风格的新图片。使用定量和定性的方法评估生成图片的质量，如风格一致性、内容保留度等。根据评估结果对模型进行优化，提高生成图片的质量和可控性。

我们预期的成果有高质量的风格迁移图片：生成一系列具有高度一致风格和内容保留的图片，展示风格迁移的效果。

相关工作

基于优化的方法

Gatys 等人提出的方法：这是最早的风格迁移方法之一，通过优化一个损失函数来生成图像。损失函数包括内容损失和风格损失，分别衡量生成图像与

内容图像和风格图像的相似度。通过最小化这两个损失函数，生成图像可以同时保留内容图像的内容和风格图像的风格。

基于前馈网络的方法

1.Johnson 等人的方法：网络结构：使用一个预训练的 VGG 网络作为特征提取器，生成网络则是一个编码器-解码器结构。损失函数：内容损失使用 VGG 网络的高层特征，风格损失使用 VGG 网络的多层特征。通过最小化这些损失函数，生成网络可以学习如何生成风格迁移图像。2. 快速风格迁移（Fast Style Transfer）：网络结构：使用一个轻量级的卷积神经网络作为生成网络，通常包括几个卷积层和反卷积层。损失函数：内容损失和风格损失的定义与 Johnson 等人的方法类似，但通过使用更简单的网络结构，大大提高了生成速度。3. 自适应实例归一化（AdaIN）：网络结构：使用一个编码器-解码器结构，其中编码器提取内容和风格图像的特征，解码器生成风格迁移图像。损失函数：通过自适应实例归一化（AdaIN）将风格图像的特征转移到内容图像上，定义内容损失和风格损失。

基于生成对抗网络的方法

CycleGAN：使用生成对抗网络进行无监督的图像到图像转换，可以实现风格迁移。CycleGAN 的核心思想是通过两个生成器和两个判别器来实现两个域（Domain）之间的双向映射。例如，将照片转换为油画风格，同时也可以将油画转换回照片风格。通过循环一致性损失，CycleGAN 确保生成的图像在转换回原始域时能够恢复出原始图像的内容。

Stable Diffusion 在图像生成领域的应用

1. 文本编码：使用 CLIP 等文本编码器将用户输入的文本描述转换为向量表示。2. 扩散模型：基于扩散模型生成图像：前向过程：逐步向图像添加噪声，直到图像完全变为随机噪声。反向过程：从噪声中逐步去噪，生成符合文本描述的图像。3. 条件生成：在反向过程中，文本编码向量作为条件信息，引导模型生成与文本描述一致的图像。

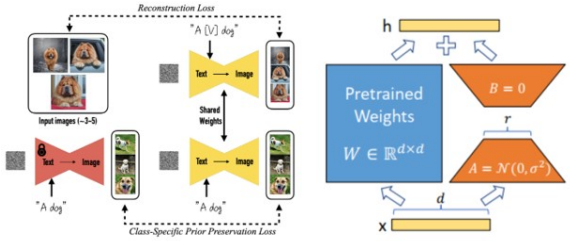

图 1: 左边为 DreamBooth 的训练过程，右边是 LoRA 微调方法的主要思想。

本项目与风格迁移的不同之处

本项目聚焦于利用 dreambooth 的相似思路，dreambooth 是利用一个主题的 3-5 张相同物体的图像来使物体与文字捆绑起来，以保证后续扩散模型中的文字生成时主体不变。本项目巧用该思路反将图像风格与文字捆绑起来，以期望得到后续扩散模型中文字生成的图片图像风格不变。不同于先前的主流方法中大量提取特征的传统思路，本项目巧妙的对文本到图像扩散模型进行微调，同时，应用了一个特定于类别的先验保存损失，该损失利用了模型在类别上的语义先验，并鼓励它使用文本提示中的类名（例如“a 纸艺”）生成属于不同主题类别的实例。

方法

为了能够顺利实现基于指定提示词生成高质量相同风格的图片，我们主要采用 DreamBooth 和 LoRA 方法去实现文本到图像的生成，从根本上讲，该方法是一种个性化的文本到图像扩散模型。虽然目前最新的文生图模型在图像生成方面大放异彩，能够生成许多高质量的图片，但我们需要生成特定对象的一些图片，也就是所谓的特定主题的图像生成，这是其他模型所不能够达到的，而 DreamBooth 方法是能够实现从几个捕获的图像中合成不同主题的图片，并在不同的背景下保持其独特特征效果的。为了减少算力损耗、提高速度，我们基于改进的 LDMs，使用了一个自编码模型，学习一个在感知上与图像空间等效的空间，引入压缩学习阶段与生成学习阶段的显式分离来解决计算复杂性高的问题。

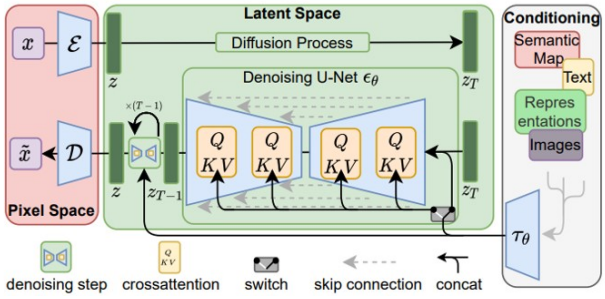

图 2: 隐变量扩散模型大致流程。

DreamBooth

为实现个性化的文本生成图像模型，首先我们的第一个任务就是要将特定主题的实例植入到模型的输出域中，这样我们就可以在模型中查询主题的各种新颖的图像。我们希望能够通过较少的图片来对模型进行微调，但是在微调的过程中应该避免模型产生过拟合和模式崩溃以及无法充分捕获目标分布等问题。为了将一个新的标识符对植入到扩散模型的字典中，我们将主题的所有输入图像标记为 identifier 和 class noun，其中 identifier 作为标识符，class noun 作为类别标签。我们可以在句子中使用类别描述符，以便将类别的先验知识与我们唯一的主题相关联，但是使用错误的类别描述符或没有类别描述符会增加训练时间和语言漂移并降低性能，所以我们利用模型对特定类别的先验知识，并将其与主题的唯一标识符嵌入在一起，以便在不同的上下文中生成主题的新姿势和表达。这样一来，我们可以借助视觉先验知识生成具有不同姿势和表达的主题图像。在选择 identifier 方面，为了避免标识符实际上是某种特定含义的单词导致性能下降的情况，我们通过在词汇表中寻找罕见的标记，然后将这些标记转换为文字。这样，我们可以最小化标识符与强先验知识的关联。

LDMs

为了减少算力损耗、提高速度，我们基于 LDMs 使用一个 Autoencoder 去学习能尽量表达原始图像空间的低维空间表达。LDMs 对速度的改进主要体现在加入 Autoencoder 使得扩散过程在 latent space 下和加入条件机制，通过 Attention 实现条件生成控制，能够使用其他模态的数据控制图像的生成两个方面。具体实施细节则分为感知图像压缩、潜在扩散模型、条件生成模型三个方面。

在感知图像压缩方面，通过感知损失和基于补丁的对抗目标的组合训练的自编码器组成感知压缩模型，编码器将给定 RGB 空间中的图像编码为潜像表示，解码器基于潜像表示重构图像，然后通过因子对图像进行下采样。在潜在扩散模型方面，设计用于通过对正态分布变量逐步去噪来学习数据分布 p(x) 的概率模型，扩散模型可以根据信噪比来规定，然后去噪扩散模型是生成模型，可用时间上向后运行的类似马尔可夫结构来恢复这个过程。然后，对与该模型相关的证据下限在离散时间步长上进行分解。最后，进行重数化。通过潜在表征的生成建模，我们可以访问高效、低维的潜在空间，在该空间中，高频、不可察觉的细节被抽象掉。此外，还能够关注数据的重要语义位，在低维、计算效率更高的空间中训练。

在条件生成模型方面，与 DDPM 不同，LDMs 在预测 noise 的过程中加入了条件机制，即通过一个编码器将条件和 Unet 连接起来。与其他类型的生成模型类似，扩散模型原则上能够对形式为条件分布进行建模，通过条件去噪自动编码器。此外，我们在 UNet 模型中使用交叉注意机制，将 Diffusion Models 转变为更灵活的条件图像生成器，以此来调节 LDM。

通过以上三个部分，我们就能够在原有基础上、不降低效果的情况下，减少算力损耗和训练时长，提高速度。

实验结果

模型设置

项目对于 Unet，使用 DreamBooth 和 Lora 组合方法进行训练，考虑到需要模型需要学习陌生的风格提示词，对 Text Encoder 使用 Lora 方法进行训练，对每种风格都要单独进行模型权重的微调，使得最终生成的图片满足结果。

可视化效果

作为训练的风格数据共有十五种，每种风格给定 10 张图片，每种风格都需要进行单独的权重微调

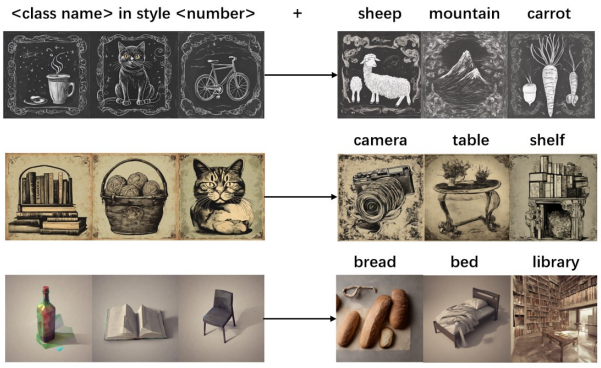

图 3: 左侧三张图片为给定的风格图片，右边是推理图片。

训练。在文中我们展示了我们训练的部分成果，完整的训练结果我们放在了 GitHub 上面，包括训练的权重。由于相关比赛已经结束，我们无法得到对结果的定量评价。

方法缺陷及改进措施

产生的部分图片与给定的文本有较大的差距，我们认为这与用 LoRA 微调 Text Encoder 进行训练有关，由于与传统意义相冲突的词出现（如 00，01），可能会使模型忘记一些先验信息，因此，可以对 Text Encoder 进行更弱强度的微调，或加强其正则化的力度，使模型的先验信息保存得更加完善。后续可以加入消融实验来验证这一部分的想法。

小组分工

郑思哲负责实验代码的实现，得出实验结果，论文的整合；周健文负责总体部分的介绍；周扬超负责相关工作的调查，阐明我们项目的不同之处；赵猛负责方法部分的阐述，创新点的分析；纪冠州负责论文的配图与编辑。