北京航空航天大學
BEIHANG UNIVERSITY

第三十四届“冯如杯”竞赛主赛道项目论文

基于扩散模型注意力图的个性化图像编辑

2024 年 4 月

摘要

本文所述的是一种保持扩散模型整体训练数据基本不变的前提下，利用扩散模型注意力图[1]和文本反转[2]技术满足用户个性化需求的图像编辑方法，主要服务于社交方面。编辑方法的主要内容是帮助用户将某一特定人、物或者艺术风格加入到目标图片中或者替换目标图片的某一事物。创新点主要体现在支持图片输入，用户只需要输入期望插入的人物的 3-5 张图片和操作要求即可，模型可以自动提取出图片内的主要信息，并将其加入目标图片或者替换目标图片的某一事物，甚至是将目标图片的整体风格替换为插入图片类似的风格。应用场景有很多，比如往合照中加入缺席者，修改图片风格，制作常用表情包，移除风景照中多余的行人等。未来有希望嵌入到修图软件中，实现一键提取原图信息并修改目标图片，甚至根据提供图片修改目标图片画风等操作。

关键词：扩散模型，文本反转，个性化，图片输入

Abstract

The article describes an image editing method that utilizes diffusion model attention maps [1] and text reversal [2] techniques to meet users' personalized needs while keeping the overall training data of the diffusion model largely unchanged. This method primarily serves social purposes. The main content of the editing method is to help users incorporate specific individuals, objects, or artistic styles into target images or replace certain elements in target images. The innovation lies in supporting image input, where users only need to input 3-5 images of the desired character to be inserted and their operational requirements. The model can automatically extract the main information from the images and incorporate them into the target image or replace certain elements in the target image, or even replace the overall style of the target image with a style similar to the inserted image. There are many application scenarios, such as adding absentees to group photos, modifying image styles, creating commonly used emoji packs, removing unnecessary pedestrians from landscape photos, etc. In the future, there is hope to embed this method into image editing software to achieve one-click extraction of original image information and modification of target images, and even modify the artistic style of target images based on provided images.

Keywords: diffusion model, text reversal, personalized, image input

目录

绪论

近些年来，随着人工智能的发展，AI 在图像识别、生成图片等图形处理领域的技术取得了显著进步，对人们的日常生活产生了深远的影响。大到医疗诊断、自动驾驶，小到娱乐社交、零售机器，无处不在他们的身影。AI 识图可以分析医学影像来辅助诊断疾病，提高诊断的速度和准确性，也可以用于监控摄像头，识别可疑行为或监控重要区域。不仅如此，AI 图像识别技术还是自动驾驶汽车的关键组成部分。

而最近随着 chatGPT 等语言大模型的兴起，AI 生成图片也渐渐地进入了公众的视野，一些公司在需要图片的时候也使用了 AI 生成图片，甚至央视新闻联播都尝试应用了 AI。但是现在的 AI 还只是经过普遍数据训练的标准模型，会存在生成图片风格单一，图不达意的现象。而随着人们对于 AI 生成图片的需求越来越高，越来越期望利用 AI 生成符合自己特定要求的图片，如何训练出符合用户期望的个性化的模型就变的尤为重要。

本文在前人根据提示词训练扩散模型生成图片[1]和利用文本反转提取图片关键信息[2]的基础上，将二者结合在一起，完成了根据提供图片的关键特征（如人、物、图片风格等）和用户期望的操作，修改目标图片的操作。

一、相关工作

近年来，生成对抗网络（GANs）在图像生成方面取得了显著进展。GANs 通过对抗训练生成真实样本，在合成高保真度的图像方面取得了重大突破。而最近，大规模自回归或扩散模型取得了令人印象深刻的视觉效果。一些方法不是训练条件模型，而是利用测试时间优化来探索预训练生成器的潜在空间，这些模型通常优化从辅助模型派生的文本到图像相似度分数。

生成对抗网络（GANs）的发展经历了几个阶段，研究人员开始探索如何从 GANs 生成的图像中还原潜在的输入向量。最初的尝试包括使用梯度下降等优化技术，尝试最小化生成图像与原始输入之间的差异。之后研究人员开始探索 GANs 潜在空间的特性，尝试通过不同的方法和技术来实现更有效和稳定的 GANs 逆向，包括改进优化算法、探索潜在空间的结构、利用 GANs 的特定属性等。这使得 GANs 逆向得到了不断改进和扩展。研究人员开始将 GANs 逆向应用于更广泛的领域，如图像编辑、生成模型解释等。

对于基于扩散的逆向方法的灵感来自于数学和物理学中的扩散过程。这种方法最初作为一种新颖的图像逆向方法被引入，其核心思想是模拟像素级的随机扩散过程。随着研究的推进，基于扩散的逆向方法得到了算法层面的改进和优化。研究人员探索如何更有效地模拟和应用扩散过程，以实现更准确和稳定的图像逆向。近年来，基于扩散的逆向方法逐渐在图像生成、模型解释和图像处理等领域得到应用和发展。这种方法的稳定性和有效性使其在实际应用中具有广泛的潜力。

这些基于扩散的模型的应用方向超越了纯图像生成，大量研究探索了基于文本界面进行图像编辑、生成器领域适应、视频操作、运动合成、风格迁移甚至是用于 3D 对象的纹理合成等等。

但是目前的工作普遍集中于文本引导的图像生成，对于输入图像的应用方面研究较少，本文期望探索同时输入图片和文本指令的编辑图像方法，以达到满足用户生成或修改个性化图片的目的。

二、项目核心分析

1.1 项目方法

我们的目标是多图像编辑，令 $I_1I_2I_3$ 为三张包含同一主体的图片（也可以是同一风格），I_4 为一提供背景的图片，我们希望将该主体表现在 I_4 的背景上。

我们的方法主要基于如下观察：textual inversion 中不仅学习到了 psedu-word S*以及对应的 text-embedding v*，还包括了该主体对应的 cross-attention map，这允许我们可以复现原本论文中 LDM 的生成过程，以及将其应用于图片编辑模型里。

为实现这个目标，我们首先使用 textual inversion 提取多张图片共同的主体，并将其命名成一个新的名词 p*。

对于 I_4，我们使用 null-text inversion，使模型可以恢复 I_4 的信息，以此来保证编辑后的图像能够维持原图像大部分信息。

在编辑过程中，我们将原本描述 I_4 的 prompt(P)更改为包含 p*的 prompt(P*)，使用 prompt to prompt 可同时生成高保真还原的原图片 I 和带有更改主体的新图片 I*。

1.1.1 Personalizing Text-to-Image Generation using Textual Inversion (基于文本反转的个性化图像生成)

简单而言，是对 Text-embeddings 进行修改，将概念“注入”到词汇表中，再使用 Latent Diffusion 进行图像生成。

具体来说，在训练阶段，将我们定义的物品的名字输入到 text embedding 中，生成一个可以学习的新的嵌入向量 v*，之后通过 Textual-inversion，最小化 LDM 损失，直接优化计算出这个 v*，再重用 LDM 的训练方案，训练出对应的 diffusion 模型。在生成的阶段，我们需要将包含定义的物品的名字的字符串输入 text embedding 中，生成嵌入向量的索引，然后在通过 transformer 生成 LDM 的生成条件$C_\theta(y)$，指导已训练好的 diffusion 模型的去噪过程。LDM 的训练过程中，需要在 UNet 主干网络上增加 cross-attention 机制来实现条件机制。

1.1.2 Null-text Inversion for Editing Real Images using Guided Diffusion Models (基于扩散模型的“空文本”反转图像编辑引导)

对于无分类器指导扩散模型，当引导系数$\omega >= 1$时，DDIM Inversion 采样过程中会积累误差，导致结果偏离高斯分布，再经过 DDIM 采样生成的图像会严重偏离原图像，

不能准确的还原重建原有的图像，但是对原图像的重建有粗略的近似。对于扩散过程 DDIM 进行每个时间步的优化就能使重建图像尽可能的接近原图。

优化 Null-text Embedding[3]能准确还原加噪路径，同时避免了对原 DDIM 训练模型的微调和文本嵌入，从而能够保留 DDIM 模型中保存的先验信息和语义信息，进而对真实图像进行高保真的编辑。

同时在已有的实验中可以发现所有时间步共用同一个 Null-text Embedding 效果比每个时间步优化不同的 Null-text Embedding 要差。所以采用$\{\emptyset_t\}^T$替代单一的$\emptyset_t$。

1.1.3 prompt to prompt image editing with cross attention control (基于交叉注意力的文本到文本的图像编辑)

给定 prompt(P)和编辑图片的 prompt(P*)，通过文生图模型，分别获得原始图片 I 和编辑后的图片 I*。I 与 I*除了编辑区域外尽可能的近。

主要过程为通过在扩散过程中注入交叉注意映射来编辑图片，从而控制哪些像素在哪些扩散步骤中关注文本 prompt 的哪些 token。

1.2 项目创新点

本文的模型可以做到利用给定图片作为概念，将其“注入”到词汇表中，并使用 Latent Diffusion 进行图像生成。用户可以输入图片让机器识别新概念，而不仅限于文本的输入，并利用这些新概念修改目标图片，实现个性化修图。

1.3 项目难点与解决方案

整个项目的难点主要在于如何将 textual inversion 训练出来的模型应用到 prompt-to-prompt 中，解决方案是将 textual inversion 训练出来的模型作为 prompt-to-prompt 训练的基础，在此之上

textual inversion 中的主要难点在于如何优化新的嵌入向量 v*，解决方案是最小化 LDM 损失，直接优化计算出这个 v*。

null-text 中的主要难点在于一系列空文本如何产生，解决方案是利用$\emptyset_t$和z_t，产生$\emptyset_{t-1}$。

三、结果展示

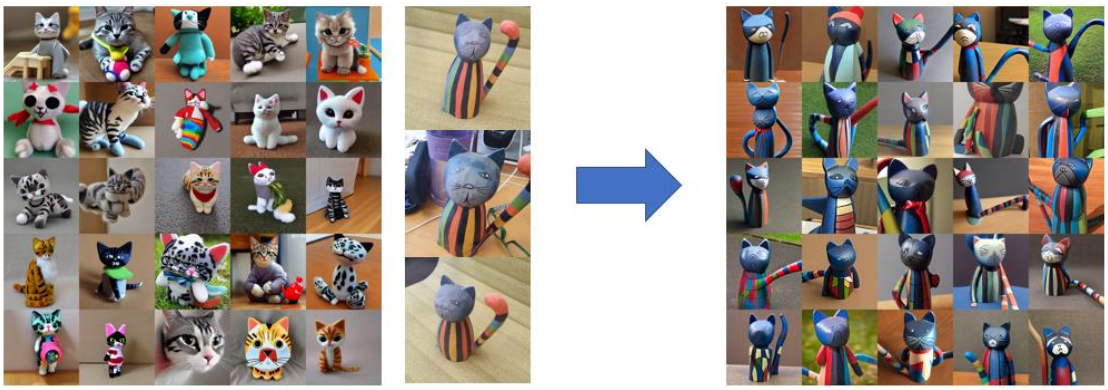

图 2　toy 例图

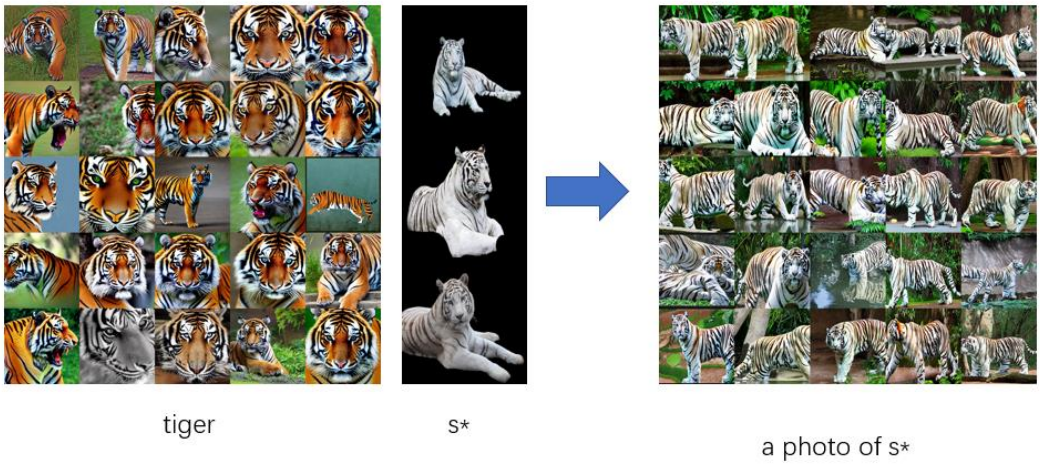

图 1　tiger 例图

图 3　替换展示

四、不足

虽然本文所训练的模型已经可以应对用户的大部分情况，但是在识别并生成精细的物体时还是会有些缺陷，比如人脸、精致的花瓶等。对于这种人或物，模型生成的图片可能会比较模糊，质量不能达到要求。除此之外，如果当提供的图片不够明确或者提供图片没有表达同一个主要信息时，模型识别生成的物体将不可控。

由于本文编辑图像前机器需要学习提供的图片，所以需要消耗一定时间，大概十分钟左右，对于商用是明显不足的，如果想要满足用户的快速的响应需求仍需要改进。

模型对于一些物理知识的判断还不能做到完全精准，比如镜子的反射，现在虽然可以做到镜子中有和镜子外相同的物体，但不能保证角度、外观等完全精确，特别是对于不对称的物体。

五、应用场景

本文制作的编辑图像的模型是聚焦于社交方向的，主要应用环境即用户的手机、电脑等，可以在未来尝试嵌入到 PS、美图秀秀、醒图等软件中，用模型帮助用户实现一键修改图片等操作，这对于频繁使用修图软件的人或者是不习惯使用手机的人来说是非常便捷的方式。这既可以帮助人们减少这种琐碎的工作，专注投入到“修改图片的什么”这种方向上，又可以让机器根据用户提供的图片明白用户的个性化需求到底是什么，从而更有针对性的实现用户的个性化需求。

具体的应用实例有很多，比如往合照中加入缺席者，只需要缺席者的几张照片和一句命令，就可以向合照中几乎完美的插入这个人，弥补缺席的遗憾。还有修改图片风格，将蒙娜丽莎改为印象派、水墨风甚至是抽象画，以及修改一本漫画的风格等等。以及制作常用表情包，因为人们（尤其是年轻人）常用的表情包大多是某段时间内的一些梗图，这些热点图被加工到各种原有的图片中，所以本文的模型可以完美契合用户需求。除此之外还有移除风景照中多余的行人、更换照片中人们穿的衣服等等。绘画的艺术就来源于想象，而本文的编辑图像方法则能帮助人们更好的想象自己喜欢的图片，并将它实现出来。

结论

本文在前人工作的基础上，提出了一种基于扩散模型注意力图的图像编辑方法，基于 latent diffusion model，使用 textual-inversion 和 prompt-to-prompt image edit 生成图片并利用 Null-text inversion 优化加噪还原过程。用户只需要提供 3-5 张自己希望修改的材料图，集中表现某一人、物或者绘画风格，并用文本的形式向机器输入指令，命令其向目标图片插入该物或者修改某物等等，机器即可自动将目标图片中的某一事物修改为提供图的事物，甚至是将目标图像的画风修改为提供图的画风。该方法主要的特点是接收用户提供的材料图并理解用户要求的究竟是什么，从而个性化地修改目标图片直至用户所期望的图片，满足了仅靠单纯的文本生成图片所不能达到的精准度。未来可以应用于各种修图软件，为人们的操作带来极大的便捷。

参考文献

[1] Hertz, A., Mokady, R., Tenenbaum, J.M., Aberman, K., Pritch, Y., & Cohen-Or, D. *Prompt-to-Prompt Image Editing with Cross Attention Control* , ArXiv, abs/2208.01626, 2022

[2] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., & Cohen-Or, D., *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion*, ArXiv, abs/2208.01618, 2022

[3] Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D., *Null-text Inversion for Editing Real Images using Guided Diffusion Models*, arXiv e-prints, doi:10.48550/arXiv.2211.09794., 2022