文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

基于 Diffusion 模型的图像生成技术研究

张永佳
武汉东湖学院 电子信息工程学院
DOI:10.12238/acair.v3i3.15588

[摘　要] 图像生成技术是人工智能技术的一个细分领域。传统的图像生成技术主要是使用生成对抗网络,通过大量对抗训练来生成逼近真实的图像。但随着算力的增长,扩散模型被更多的用于图像生成。扩散模型是通过正、反向扩散来生成图像,具有更丰富的细节表达能力。它也是未来图像生成技术领域主要的研究方向。
[关键词] 图像生成技术；生成对抗网络；扩散模型
中图分类号：TV149.2　文献标识码：A

Research on image generation technology based on Diffusion model

Yongjia Zhang
School of Electronic and Information Engineering, Wuhan East Lake University
[Abstract] Image generation technology is a subdivision of artificial intelligence technology. Traditional image generation techniques mainly use generative adversarial networks to generate images that approximate the real world through a large number of adversarial training. However, with the growth of computing power, diffusion models are increasingly used for image generation. The diffusion model generates images through forward and reverse diffusion, and has richer ability to express details. It is also the main research direction in the field of image generation technology in the future.
[Key words] image generation technology; generative adversarial network; diffusion model

前言

人工智能(Artificial Intelligence),英文缩写为AI,是当下最热门的技术热点之一。该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。人工智能从诞生以来,理论和技术日益成熟,应用领域也不断扩大[1]。

近年以来,以OpenAI公司为代表的研究机构,在探索通用人工智能方面取得了长足的进步。先后诞生了AlphaGo,ChatGPT等人工智能模型。特别是2022年起,以ChatGPT发布为标志,代表人工智能的研究全面进入到AI大模型时代。AI大模型是指一个庞大复杂的神经网络,需要通过存储更多的参数来增加模型的深度和宽度,从而提高模型的表现能力,参数从百亿起步,对大量数据进行训练并产生高质量的预测结果。著名的AI大模型有OpenAI的GPT-3模型,参数规模达1750亿,PaLM-E 的参数规模更是达到了5620亿。

AI图像生成技术作为人工智能技术的一个细分领域,它的出现得益于AI大模型技术的进步,近年来也取得了巨大发展,诞生了很多开源AI绘画模型,如OpenAI的DALL·E2模型,只需输入简单的文本(prompt),它就可以生成多张1024*1024的高清图像。谷歌的文本生成图像AI模型Imagen,它能够通过给定的文本描述生成该场景下逼真的图像。而Stability.Ai公开发布的文本生成图像模型Stable diffusion是其中最具有代表性的模型之一。Stable diffusion是一个基于Latent Diffusion Models(潜在扩散模型,LDMs)的文图生成(text-to-image)模型。具体来说,得益于Stability AI的计算资源支持和LAION的数据资源支持,Stable Diffusion在LAION-5B的一个子集上训练了一个Latent Diffusion Models,该模型专门用于文图生成。Latent Diffusion Models通过在一个潜在表示空间中迭代“去噪”数据来生成图像,然后将表示结果解码为完整的图像,让文图生成能够在消费级GPU上,在10秒级别时间生成图片,大大降低了应用落地的门槛。

1 GAN(生成对抗网络)

在Stable Diffusion(扩散模型)出现之前,计算机视觉与机器学习领域最重大的突破是GAN(Generative Adversarial Networks生成对抗网络)。

2014年Goodfellow等人启发自博弈论中的二人零和博弈,开创性地提出了GAN(生成对抗网络)。生成对抗网络包含一个生成模型和一个判别模型。其中,生成模型负责捕捉样本数据的分布,而判别模型一般情况下是一个二分类器,判别输入是真实数

据还是生成的样本。这个模型的优化过程是一个“二元极小极大博弈”问题,训练时固定其中一方(判别网络或生成网络),更新另一个模型的参数,交替迭代,最终,生成模型能够估测出样本数据的分布[2]。

GAN的核心思想就是通过大量的攻防训练来尽量模拟出一个逼近真实的图像(如图1所示)。其训练步骤一般如下：

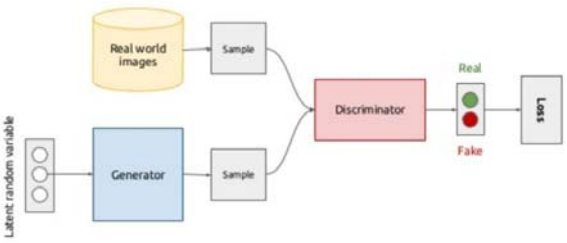

图1 生成对抗网络训练模型

(1)给定一个真实数据的数据集(如图片集),以及一个生成器(Generator,以下简称G)和一个判别器(Discriminator,以下简称D)。生成器G负责生成假的图片,判别器D负责鉴别一张图片是否为真,并输出0/1二分类结果(判断真伪)。

(2)随机初始化一个向量z,G以z作为输入生成一张新的图片X',从真实数据集中随机一张图片X,将(X,X')这两张图送进D,由它来判断哪张图是真的,哪张是假的,并把判断依据反馈给G。

(3)G的目的是不断生成更像真实数据集里的图片以试图骗过D,而D学习如何判断送过来的两张图片哪张是真的、哪张是假的。

GAN的出现对无监督学习,图片生成的研究起到极大的促进作用。GAN已经从最初的图片生成,被拓展到计算机视觉的各个领域,如图像分割、视频预测、风格迁移等。

但随着技术发展,GAN开始暴露出瓶颈和弊病,主要集中在图像生成缺乏多样性、模式崩溃、多模态分布学习困难以及因问题表述的对抗性而不易训练。近几年,随着算力增长,一些过去因算力不足无法实现的复杂算法得以实现,其中“扩散模型”从气体扩散物理过程汲取灵感,试图在多个科学领域模拟相同现象,它在图像生成领域展现出巨大潜力,成为如今Stable Diffusion的基础。

2 Diffusion models(扩散模型)

diffusion models(扩散模型)是深度生成模型中新的SOTA,在图片生成任务中超越原SOTA：GAN,且在计算机视觉、NLP等诸多应用领域表现出色。扩散模型是用于生成与训练数据相似数据的生成模型,和GAN基于对抗的思路不同,它是给真实图像不断增加高斯噪声直至其分布为高斯分布,再逆序从高斯分布重建图像,简单来说,其工作方式是迭代添加高斯噪声“破坏”训练数据,再学习消除噪声恢复数据。

一个标准扩散模型有两个主要过程：正向扩散和反向扩散。在正向扩散阶段,通过逐渐引入噪声来破坏图像,直到图像变成完全随机的噪声。在反向扩散阶段,使用一系列马尔可夫链逐步去除预测噪声,从高斯噪声中恢复数据[3]。如下图2所示。

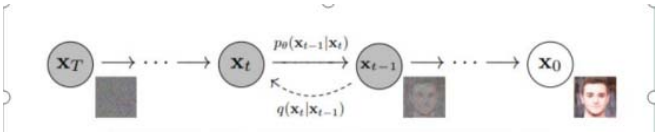

图2 标准扩散模型

2.1 Diffusion正向扩散

前向过程$x_0 \sim q(x)$,即在图片上添加噪声的过程。尽管这一步骤无法实现图片生成,但它是理解Diffusion model以及构建训练样本GT极为关键的一步。

给定真实图片,diffusion前向过程通过T次累计对其添加高斯噪声,得到 $x_1, x_2, \ldots, x_T$。这里需要给定一系列的高斯分布方差的超参数 $\{\beta_t \in (0,1)\}_{t=1}^{T}$。前向过程由于每个时刻t只与t-1时刻有关,所以也可以看作马尔科夫过程：

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right), q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \quad (1)$$

这个过程中,随着t的增大, x_t 越来越接近纯噪声。当 $t \to \infty$, x_t是完全的高斯噪声。且实际中 β_t随着t增大是递增的,即 $\beta_1 < \beta_2 < \ldots < \beta_T$。

2.2 Diffusion反向扩散

如果说正向扩散(forward)是加噪的过程,那么反向扩散(reverse)就是Diffusion的去噪推断过程。如果我们能够逐步得到逆转后的分布 $q(x_{t-1}|x_t)$,就可以从完全的标准高斯分布 $x_T \sim \mathcal{N}(0,\mathbf{I})$还原出原图分布 x_0。我们使用深度学习模型(目前主流是U-Net+attention的结构)去预测这样的一个逆向的分布 p_θ：

$$p_\theta(X_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \quad (2)$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)) \quad (3)$$

虽然我们无法得到逆转后的分布 $q(x_{t-1}|x_t)$,但是如果知道 x_0,是可以通过贝叶斯公式得到：

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I}) \quad (4)$$

2.3 Diffusion训练

对于噪声的估计和去除,最常使用的是U-Net。该神经网络的架构看起来像字母U,由此得名。U-Net是一个全连接卷积神经网络,这使得它对图像处理非常有用。U-Net的特点在于它能够将图像作为入口,并通过减少采样来找到该图像的低维表示,这使得它更适合处理和查找重要属性,然后通过增加采样将图像恢复。

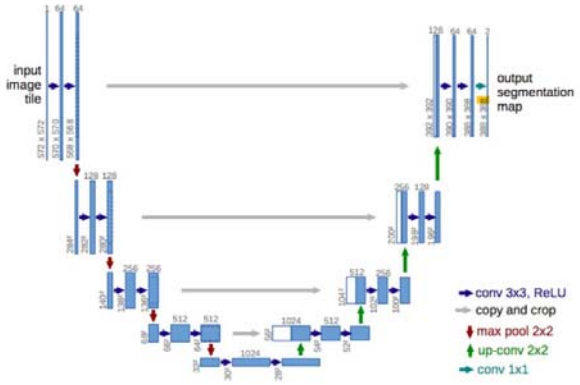

图3 一个典型的U-Net架构

训练过程可以看作：

(1)获取输入 x_0，从1.....T随机采样一个t。

(2)从标准高斯分布采样一个噪声 $z_t \sim \mathcal{N}(0,\mathbf{I})$ 。

(3)最小化 $\|z_t - z_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}z_v, t\right)\|$ 。

下图4就是训练/测试流程图。

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1,\dots,T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\|\boldsymbol{\epsilon} - \mathbf{z}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
2: **for** $t = T,\dots,1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{z}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

图4 Diffusion model训练/测试流程图

3 Stable Diffusion

扩散模型的最大问题是时间和经济成本“昂贵”，Stable Diffusion旨在解决该问题。生成1024×1024尺寸图像时，U-Net用相同尺寸噪声生成图像，一步扩散计算量大，循环迭代多次成本更高。解决办法之一是将大图片拆分为小分辨率图片训练，再用额外神经网络产生更大分辨率图像(超分辨率扩散)。

2021年发布的Latent Diffusion模型给出不同方法，它不在图像上操作，而是在潜在空间操作，将原始数据编码到更小空间，让U-Net在低维表示上添加和删除噪声。潜在空间是对压缩数据的表示，压缩是用更少数位编码信息的过程，如用单颜色通道表示RGB图片。降维会丢失部分信息，但某些情况下可过滤不重要信息，保留重要信息。

“潜在扩散模型”结合了GAN的感知能力、扩散模型的细节保存能力和Transformer的语义能力，创造出更稳健高效的生成模型。与其他方法相比，Latent Diffusion节省内存，生成的图像有多样性和高细节度，同时还保留了数据的语义结构。

4 结束语

扩散模型是深度生成模型中新的SOTA，在图片生成任务中超越原SOTA的GAN，且在计算机视觉、NLP等诸多领域表现出色，还与稳健学习等研究领域联系密切。不过，原始的扩散模型存在缺点，如采样速度慢、最大似然估计不佳、泛化能力差。

如今，很多研究已从实际应用角度解决这些限制，或从理论角度分析模型能力。Stable Diffusion模型本质上属于潜在扩散模型，潜在扩散模型在生成高分辨率图像方面稳健，保留图像语义结构，是图像生成及深度学习领域的重大进步。Stable Diffusion将潜在扩散模型用于高分辨率图像，用CLIP作文本编码器，其未来研究方向主要有以下几个方面：

(1)diffusion model已经成为一个强大的框架，可以在大多数应用中与生成对抗性网络(GAN)竞争，而无需诉诸对抗性训练。对于特定的任务，我们需要了解为什么以及何时扩散模型会比其他网络更加有效，理解扩散模型和其他生成模型的区别将有助于阐明为什么扩散模型能够产生优秀的样本同时拥有高似然值。另外，系统地确定扩散模型的各种超参数也是很重要的。

(2)diffusion model如何在隐空间中提供良好的latent representation，以及如何将其用于data manipulation的任务也是值得研究的。

(3)将diffusion model和generative foundation model结合，探索更多类似于ChatGPT，GPT-4等有趣的AIGC应用。

[参考文献]

[1]王珂，翟婷婷.人工智能及计算智能在物联网方面的应用[J].数字技术与应用，2014(8):2.

[2]张营营.生成对抗网络模型综述[J].电子设计工程，2018，26(05):34-37+43.

[3]李婧文，李雅文.深度合成技术应用与风险应对[J].网络与信息安全学报，2023，9(02):184-190.

作者简介：

张永佳(1982--)，男，汉族，湖南溆浦人，助教，硕士研究生，研究方向：深度学习、人工神经网络、大模型应用。