图像生成研究综述

皮佳豪 [1)]

[1)] 南京邮电大学计算机学院南京市中国

摘　要：本文主要聚焦于图像生成领域的四大主流技术：变分自编码器（VAE）、自回归模型、生成对抗网络（GAN）和扩散模型。VAE 通过编码器和解码器框架，结合变分推断，学习数据的潜在表示，用于图像生成和重建。自回归模型基于像素级别的条件概率，通过顺序生成每个像素值来构建图像，与 VAE 结合的 VQ-VAE 模型有效利用两者优势进行图像合成。GAN 通过生成器和鉴别器的对抗训练生成高质量图像，但可能面临训练不稳定和模式崩溃的问题。扩散模型作为较新的技术，通过逐步去噪过程生成图像。此外，FID（Frechet Inception Distance）和 IS（Inception Score）作为评估指标，衡量生成图像与真实图像的相似度及多样性。

关键词：变分自编码器；自回归模型；生成对抗网络；扩散模型

A Survey of Image Generation Research

Pi JiaHao[1)]

[1)]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing

Abstract: This paper primarily focuses on four mainstream technologies in the field of image generation: Variational Autoencoders (VAE), autoregressive models, Generative Adversarial Networks (GAN), and diffusion models. VAEs learn the latent representation of data through an encoder-decoder framework combined with variational inference, which is used for image generation and reconstruction. Autoregressive models construct images by sequentially generating each pixel value based on pixel-level conditional probabilities, and the VQ-VAE model effectively leverages the advantages of both VAE and autoregressive models for image synthesis. GANs generate high-quality images through adversarial training between a generator and a discriminator, but may face issues such as training instability and mode collapse. Diffusion models, as a relatively new technology, generate images through a gradual denoising process. Additionally, Frechet Inception

Distance (FID) and Inception Score (IS) are used as evaluation metrics to measure the similarity and diversity of generated images compared to real images.

Key words: Variational Autoencoder; Autoregressive Model; Generative Adversarial Network; Diffusion Model

1 引言

基于人工智能技术的生成内容（artificial intelligence generated content, AIGC）已成为当下的热门话题。随着深度学习的提出与发展，生成模型在图像生成领域已取得显著成果。它可以帮助人类和计算机更好的交互，帮助人类完成很多应用，被广泛应用于电影、游戏、绘画和虚拟现实等领域。图像生成中最关键的问题是选取合适的生成模型，不同的生成模型往往适用于不同的生成任务。早期的图像生成基于能量优化模型，随着深度学习的兴起，诞生了一系列基于神经网络的生成模型，如变分自编码器（Variational Auto-Encoder, VAE）[1]，生成对抗网络（Generative Adversarial Network, GAN）[2]，自回归网络（AutoRegressive Model, AR Model）[3]，去噪扩散模型（Denoising Diffusion Probabilistic Model, DDPM）[4] 等。

2 基本理论和方法

早期的图像生成工作大多基于特征表达方式来生成图像，受限于特征表达能力，模型只能生成简单的图像。深度神经网络的出现极大地提高了模型对复杂特征的表达能力，使得图像生成技术得到了突破性的发展。本节将依次介绍基于深度学习的变分自编码器 (VAE) [1]，生成对抗网络 (GAN) [2]，自回归模型 (AR Model) [3]，以及去噪扩散模型 (DDPM) [4]。

2.1 变分自编码器

变分自编码器（Variational Auto-Encoder, VAE）[1] 是在自编码器（AutoEncoder, AE）架构上发展起来的深度学习模型。自编码器因其在降维和特征提取等领域的广泛应用而闻名，它由编码器和解码器两部分组成：在训练过程中，编码器将输入样本编码到一个低维的隐空间，而解码器则将这些隐变量解码回原始样本。尽管如此，传统的自编码器并不具备生成模型的特性，因为它没有对隐空间中的变量分布进行建模，因而无法通过在隐空间内采样来创造新的样本。为了克服这一限制，变分自编码器在训练过程中对隐空间的变量施加了约束，最普通的做法是利用 KL 散度（Kullback-Leibler divergence）来强制隐空间变量遵循高斯分布。这样，在生成新样本时，可以按照高斯分布从隐空间中抽取样本，并将这些样本送入解码器以产生新的图像。

其中，第一项表示编码的分布与目标高斯分布的距离，第二项表示重构图像与输入图像在欧氏空间的距离，它让重构图像向真实图像靠近。当变分自编码模型训练结束，可以直接从标准正态分布 $N(0,1)$ 中采样得到隐变量 z，再将 z 输入解码器网络即可得到生成图像。约束隐空间符合正态分布往往过于困难，这会使生成图像的质量不理想。为了解决这个问题，Oord 等人提出矢量量化变分自编码器（Vector-Quantized Variational Auto-Encoder, VQ-VAE）[5]。它约束变分自编码器的隐空间满足离散分布。如图2所示，编码器 E 将输入图像 X 编码成隐向量 Z_x，再将其通过一个可学习的码本 C 量化成离散的编码。解码器 D 将量化后的离散编码解码成重建图像 X'。由于矢量化的不可微性，

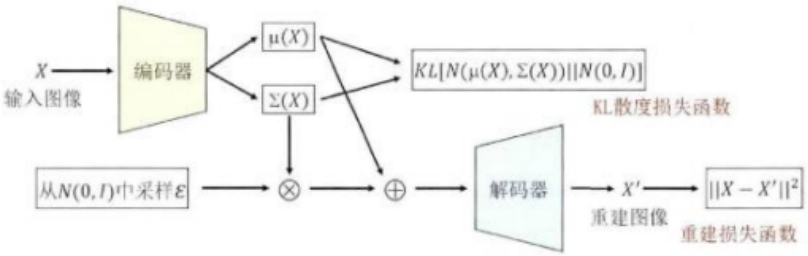

Figure 1: 变分自编码器（VAE）示意图

码本无法与编码器和解码器一起进行梯度回传。训练时的损失函数为：

$$\text{VQVAE} = \|X' - X\|^2 + \|\text{sg}[Z_x(x)] - e\|^2 + \|z_x(x) - \text{sg}[e]\|^2 \tag{2.1}$$

其中，e 表示码本中的编码，sg 是梯度截断符号，代表此变量之前的模型在梯度回传时不变。上述中的第一项是重建损失，由于矢量化不可微，梯度更新时解码器的输入端梯度将直接复制到编码器输出端；后两项约束编码器编码的向量和码本的向量尽可能相似。

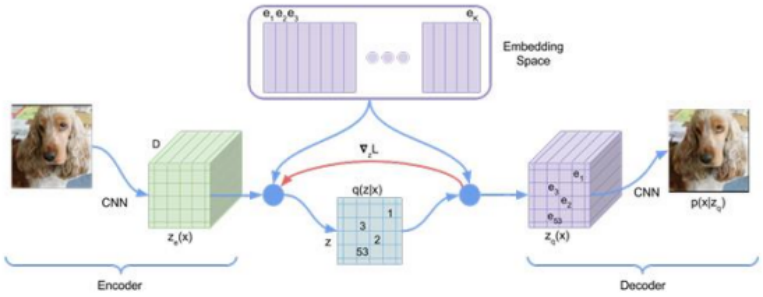

Figure 2: 量化变分自编码器（VQVAE）示意图

2.2 生成对抗网络

生成对抗网络（Generative Adversarial Network, GAN）[2] 是当今应用最为广泛的生成模型，在 2014 年由 Ian Goodfellow 等人首次提出。生成对抗网络由一个生成器和一个判别器构成。生成器用于拟合数据分布，判别器用于判别输入数据是真实数据还是生成数据。

如图3所示，假设真实数据 x 符合 P_x 分布，隐空间变量 z 符合 P_z 分布，其中 P_z 是一个已知的可采样的分布，如高斯分布或均匀分布。整个生成对抗网络的损失函数为：

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_x}[\log D(x)] + \mathbb{E}_{z \sim P_z}[\log(1 - D(G(z)))] \tag{2.2}$$

从理论上可以证明，假设生成器和判别器均有无限拟合能力，则基于该训练算法，收敛后生成网络和判别网络将达到纳什均衡状态，生成数据分布和真分布相同，即 $P_g = P_r$。

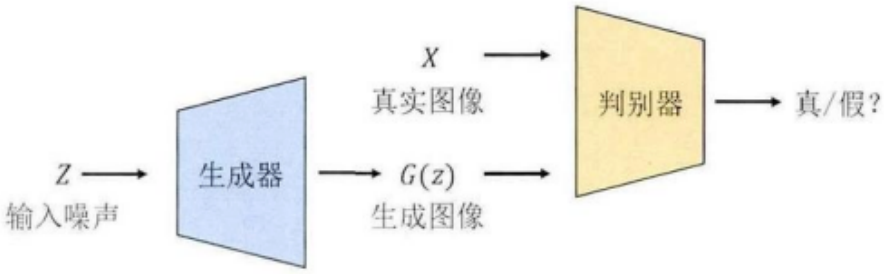

Figure 3: 生成对抗网络（GAN）示意图

然而，在实际训练过程中，由于生成器和判别器的拟合能力有限，并且在迭代训练中不能保证优化到最优，因此实际生成对抗网络并不能达到纳什均衡的状态。此外，生成对抗网络往往还伴随训练不稳定，模式崩塌，梯度难以回传等问题。

在损失函数改进的方向上，Mao 等人 [7] 提出最小平方生成对抗网络（Least Square GAN, LSGAN）来提高对抗训练的稳定性。Wasserstein GAN（WGAN）[8] 对训练不稳定提出解释并引入 Lipschitz 连续的限制。

在网络结构改进方面，SAGAN [9] 提出将自注意力机制引入生成对抗网络中。PG-GAN [10] 提出从低分辨率到高分辨率逐步优化的图像生成策略。StyleGAN [11] 提出用可调节的实例归一化注入调制信号。

2.3 自回归模型

自回归模型（AutoRegressive Model）是一种常用的生成模型，它将图像生成过程转换为逐像素点序列化生成的问题：

$$P(x) = \prod_{i=1}^{n} p(x_i|x_1, ..., x_{i-1}) \tag{2.3}$$

在训练和测试的过程中，一般在初始化置输入占位符 <S>，用于表示初始输入（start of token），然后依次预测后续的结果，在测试阶段，依次输出至终止字符 <E>（end of token）。

不同的自回归模型可能会选取不同的生成网络结构。PixelRNN [12] 采用递归神经网络处理预测过程。PixelCNN [6] 采用带掩模的卷积层来近似递归神经网络。DALL-E [13] 借助量化变分自编码器将图像编码到隐空间的离散编码上，再用基于 Transformer 结构的自回归模型拟合离散编码的分布。

然而，自回归模型用于图像生成不可避免地会具有以下几个问题：

1. 图像生成消耗大量时间
2. 位置依赖假设不合理
3. 误差累积问题

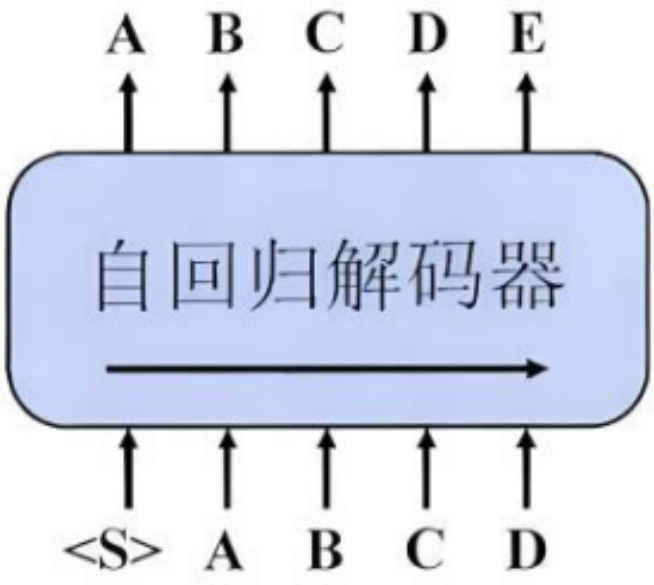

Figure 4: 自回归模型示意图

2.4 扩散模型

去噪扩散模型（Denoising Diffusion Probabilistic Model, DDPM）[4] 由 Jonathan Ho 在 2020 年提出，它包含一个正向噪声扩散马尔科夫过程和一个反向去噪马尔科夫过程。

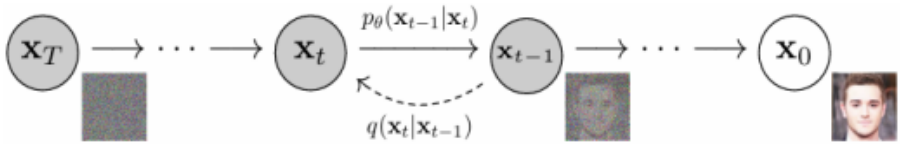

Figure 5: 去噪扩散模型（DDPM）示意图

正向过程：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I) \tag{2.4}$$

反向过程：

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t), \Sigma_\theta(x_t)) \tag{2.5}$$

扩散模型在文本到图像的生成任务中表现出了卓越的性能，如 Imagen [14]、Stable Diffusion [15] 和 DALL-E3 [16]。除了文本条件，扩散模型还支持图像条件，例如传输图像、深度图或人体骨架 [17, 18]。

3 生成模型评估

生成模型的输出需要同时满足两个关键标准：高质量和多样性。

3.1 评估指标

Inception Score (IS) [19] 定义如下：

$$IS(p_g) = \exp(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x)\|p(y))) \tag{3.1}$$

Frechet Inception Distance (FID) [20] 定义如下：

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{3.2}$$

3.2 实验结果

Table 1: ImageNet 数据集上不同生成方法的对比

方法	IS ↑	FID ↓
BigGAN	224.5	6.95
StyleGan-XL	265.1	2.30
DiT-L/2	167.2	5.02
ViTVQ	175.1	4.17
L-DiT-3B	304.4	2.10
VAR-d16	274.4	3.30

4 总结

图像生成是计算机视觉和计算机图形学的重要研究方向。本文介绍了图像生成领域的四类方法：变分自编码器、生成对抗网络、扩散模型、自回归模型。随着深度学习的发展，图像质量的多样性得到提升，特别是 GAN 和扩散模型已经在生成高分辨率和高质量的图像方面取得了显著进展。这些模型能够生成更加多样化和真实的图像。而矢量量化技术的引入减少了模式崩溃的问题。

尽管如此，图像生成模型仍面临一些挑战和未来的研究方向。尽管技术已经显著提高了生成图像的质量，但在图像的真实感方面仍然存在挑战。在图像生成内容上存在较大的不确定性，对于图像本身可控性要求较高的领域来说，生成图像是否与预期目标相符，以及对图像精度的精准控制十分关键。这些挑战和研究方向将推动图像生成模型技术的发展，并在未来的应用中发挥越来越重要的作用。

References

[1] KINGMA D P, WELLING M. Auto-encoding variational bayes[A]. 2013.

[2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014: 2672-2680.

[3] LAROCHELLE H, MURRAY I. The neural autoregressive distribution estimator[C]// Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. 2011: 29-37.

[4] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[A]. 2020.

[5] OORD A V D, VINYALS O, KAVUKCUOGLU K. Neural discrete representation learning [A]. 2017.

[6] VAN DEN OORD A, KALCHBRENNER N, ESPEHOLT L, et al. Conditional image generation with pixelcnn decoders[C]//Advances in Neural Information Processing Systems. 2016: 4790-4798.

[7] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2017: 2794-2802.

[8] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein gan[A]. 2017.

[9] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks[C]//International Conference on Machine Learning. 2019: 7354-7363.

[10] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[A]. 2017.

[11] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4401-4410.

[12] OORD A V D, KALCHBRENNER N, KAVUKCUOGLU K. Pixel recurrent neural networks [A]. 2016.

[13] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[A]. 2021.

[14] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. arXiv:2205.11487, 2022.

[15] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2022: 10684-10695.

[16] OpenAI. Dall-e 3, 2023. URL: https://openai.com/dall-e-3.

[17] BROOKS T, HOLYNSKI A, EFROS A A. Instruct-pix2pix: Learning to follow image editing instructions[C]// CVPR. 2023: 18392-18402.

[18] ZHANG L, RAO A, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]// ICCV. 2023: 3836-3847.

[19] SALIMANS T, KARPATHY A, CHEN X, et al. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications[A]. 2017.

[20] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]// Advances in Neural Information Processing Systems. 2017: 6626-6637.