图像生成研究报告

张铭徐

报告日期：2023.6.9

生成模型为什么这么火？

- 生成模型是目前爆火的一个研究方向。

生成模型为什么这么火？

- 生成模型是目前爆火的一个研究方向。
- 同时也是人工智能通往 AGI 的必由之路。

生成模型为什么这么火？

- 生成模型是目前爆火的一个研究方向。
- 同时也是人工智能通往 AGI 的必由之路。
- 《Segment Anything》这篇论文，也是图像分割领域的里程碑。

生成模型为什么这么火？

- 生成模型是目前爆火的一个研究方向。
- 同时也是人工智能通往 AGI 的必由之路。
- 《Segment Anything》这篇论文，也是图像分割领域的里程碑。
- 本文从四篇经典的图像生成相关的文献入手，对论文的动机，研究方法，实验结果进行相应的总结分析，并试图总结出图像生成领域目前的局限性以及未来的研究方向。

《Generative Adversarial Nets》[GPAM+14] 是由 Ian Goodfellow 等人于 2014 年提出的重要论文，提出了一种新颖的生成模型，称为生成对抗网络（GAN）。

- 提出了一种通过对抗过程估计生成模型的框架。
- 其中包含两个模型：一个生成模型 G 和一个判别模型 D。
- 生成模型 G 用来生成伪造数据，而判别模型 D 用来评估一个数据样本是真实数据还是伪造数据。
- 这两个模型通过对抗过程相互训练，最终得到一个能够生成类似于真实数据的生成模型。

网络介绍

- 我们以一个通俗易懂的例子来考虑生成模型和判别模型。
- 生成模型可以看成是印假钞的罪犯。
- 判别模型是抓罪犯的验钞机。
- 我们最终的效果是想要制造出与真钞无异的假钞。
- 这意味着，在外人眼里与真钞毫无区别；验钞机区分不出来是真是假。

网络介绍

我们从更加专业的角度考虑：

- GAN 的诞生开创了生成模型的一个全新的世界。
- 其主要利用了博弈论的原理：训练两个神经网络分别是 D 和 G，分别是判别网络 D 和生成模型 G。
- 生成模型是学习给定样本的数据分布，并尽可能的生成出符合给定样本数据分布的全新数据；
- 判别器的作用是判断给定的样本是生成器生成出来的还是原始数据。
- 我们期望于判别模型尽可能的能够分清给定数据到底是生成器生成的，还是原始的数据分布；并且期望于生成器能够尽可能的逼近原始数据，做到以假乱真的效果。

研究方法

- GAN 由两个主要组件组成：生成器（Generator）和判别器（Discriminator），通过对抗训练的方式相互竞争，从而达到生成逼真样本的目的。
- 生成器的目标是学习生成与真实样本相似的数据，而判别器的目标是区分生成器生成的样本和真实样本。两者通过博弈过程相互学习，直到达到一个平衡点，生成器能够生成逼真的样本，判别器无法区分真实样本和生成样本。
- 可以证明，博弈过程的纳什平衡点是 50%，即判别器对于给定的数据有 50% 的概率认为是原始数据，有 50% 的概率认为是生成器生成的数据。

研究方法

- 生成器（Generator）：生成器将一个随机噪声向量 z 作为输入，通过一系列神经网络层生成一个与真实样本相似的数据样本 x。生成器可以表示为函数 $G(z;\theta_g)$，其中 θ_g 是生成器的参数。
- 判别器（Discriminator）：判别器接收一个数据样本 x，并输出一个介于 0 和 1 之间的概率，表示 x 是真实样本的概率。判别器可以表示为函数 $D(x;\theta_d)$，其中 θ_d 是判别器的参数。

研究方法

对抗过程？

$$\min_{\theta_g} \max_{\theta_d} V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

其中，第一项表示判别器将真实样本判别为真的期望，第二项表示判别器将生成样本判别为假的期望。

研究方法

- 对于我们生成样本的训练目标：生成器的目标是最大化生成样本被判别器判别为真实样本的概率，可以表示为：

$$\min_{\theta_g} V(G,D) = \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

- 对于判别器的训练目标：判别器的目标是最小化生成样本被判别器判别为真的概率，同时最大化真实数据被判别为真的概率，则对应了：

$$\max_{\theta_d} V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)$$

研究方法

- 生成网络和判别网络共享一个损失函数。
- 我们在训练时，首先更新判别网络 D 的参数：
- 首先需要最大化 $D(x)$，同时最最大化 $1 - D(G(z))$，值得注意的是，对于后面的 $1 - D(G(z))$，代表的是判别器成功将生成数据判定为生成数据的概率。
- 同理，在更新生成器时，需要最小化 $1 - D(G(z))$，也就是说最大化 $D(G((z))$，这与我们上面所讨论的原理是相一致的。

训练过程

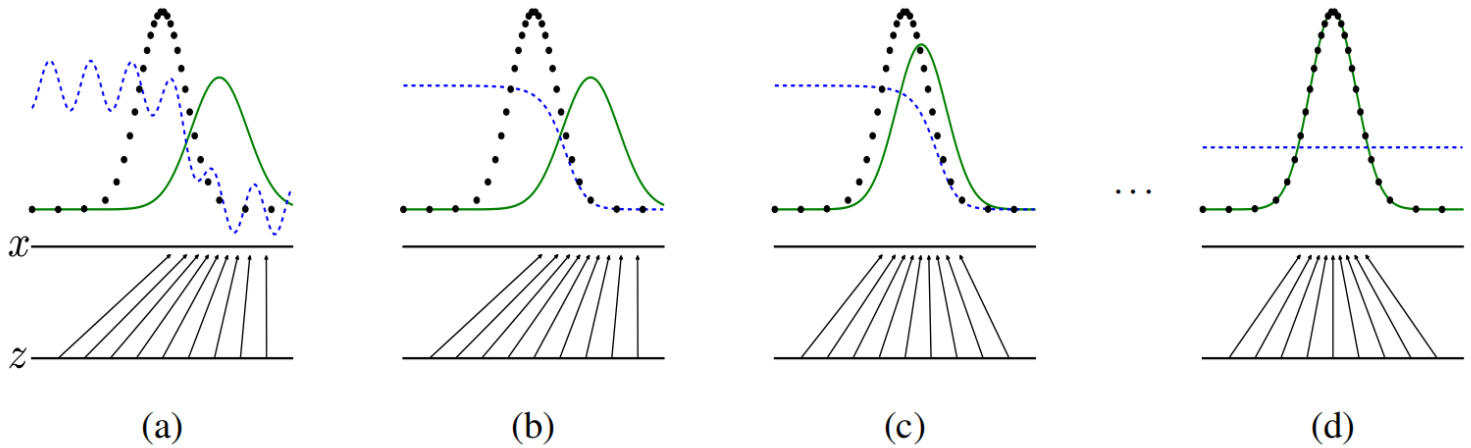

图 1: GAN 训练图

训练过程

- (a),(b),(c),(d) 代表训练的不同阶段。
- 蓝色的虚线代表判别网络 D，黑色的点线代表真实数据的分布 p_x，绿色的实线代表生成的数据分布 p_g。
- 先更新判别网络，使其判别能力尽可能的强，然后通过判别网路来更新生成网络，使生成网络生成的数据分布尽可能的拟合原始数据分布；最后经过若干轮的迭代，达到收敛点即纳什平衡点。

背景知识

自编码器是一种无监督学习的神经网络模型，用于学习输入数据的低维表示。它由两部分组成：编码器（Encoder）和解码器（Decoder）。编码器将输入数据映射到潜在空间中的低维表示，而解码器则将该低维表示映射回原始数据的重构。

- 编码器（Encoder）：接收输入数据并将其映射到潜在空间中的低维表示。编码器的目标是将输入数据压缩成更紧凑的表示，捕捉输入数据的重要特征。
- 解码器（Decoder）：解码器接收编码器的输出，也就是低维表示，并将其映射回原始数据的重构。解码器通常与编码器对称。解码器的目标是尽可能准确地还原原始数据。

背景知识

原理图可以参照下图：

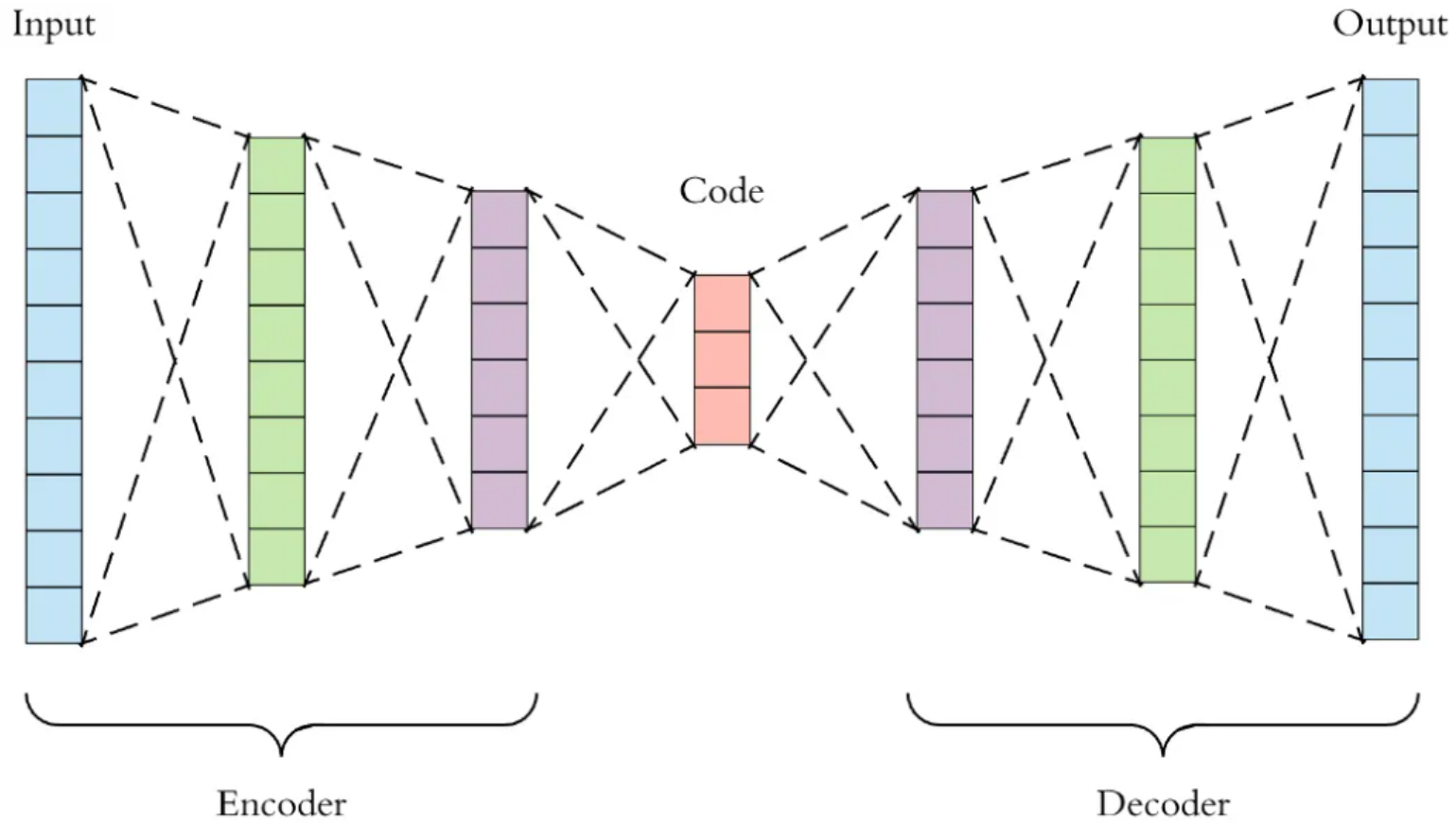

图 2: Auto-encoder 原理图

VAE 研究动机

- 自编码器是将原输入数据通过一系列非线性变换映射到潜在空间中，每一个输入的向量对应潜在空间的一个向量点，换言之，潜在空间中的点是通过编码器对每个输入数据进行独立映射得到的。
- 通过 decoder，可以将任意一个潜在空间上的点重新升维成原数据规模的向量，也就是说，我们可以通过对样本空间进行采样，将采样后的点通过 decoder 生成出与原本数据分布类似的样本。但是这样存在一些问题：auto-encoder 并不是学习整个潜在空间的样本分布，而是训练出了一个映射模型，所以它无法通过直接从空间中随机采样来生成新的样本。换言之，无法随机采样。

VAE 研究动机

- 变分自编码器 (VAE)[KW22] 的动机便是通过引入概率推断的思想，允许对潜在空间中的点进行随机采样，并通过解码器生成与这些采样点对应的新样本。这种方法使得生成的样本具有更大的多样性，并能够在更广泛的数据分布中进行采样，这使得 VAE 能够更好地描述数据的分布，并在生成任务中展现出更强的表现。

Solution

我们考虑学习潜在空间的分布即 $p(z)$ 的分布，我们考虑生成数据的过程，可以认为是：

- 对先验分布 $p(z)$ 进行采样得到一个 z_i。
- 根据上面得到的 z_i，从条件分布 $P(X|z_i)$ 中采样得到一个数据点 x_i。值得注意的是，在这里，我们可以认为该条件分布是一个 decoder 的过程。

如果能够对这个过程加以建模，学习到这个过程，那么就能够解决上面自编码器存在的问题即只能对部分数据点进行采样映射，生成有用的数据。

How to do?

VAE 的目标是最大化对数似然函数的下界，称为变分下界，即：

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x)||p(z))$$

- $\log p(x)$ 表示数据点 x 的对数似然函数，用于衡量 VAE 模型生成原始数据的能力。
- $\mathbb{E}q\phi(z|x)[\log p_\theta(x|z)]$ 是重构项，它计算在给定潜变量 z 的情况下，重构数据点 x 的期望对数似然函数。该项鼓励 VAE 生成与原始数据相似的样本。
- $\mathrm{KL}(q_\phi(z|x)||p(z))$ 是近似后验分布 $q_\phi(z|x)$ 与先验分布 $p(z)$ 之间的 KL 散度。它度量这两个分布之间的差异。这一项作为正则化项，鼓励近似后验分布与先验分布接近。

损失函数

VAE 的损失函数包括两个部分：重构损失和正则化损失。重构损失：用于衡量解码器生成的数据样本与原始数据之间的差异。通常使用平均平方误差作为重构损失函数。对于给定的数据样本 x 和解码器生成的数据样本 $\hat{x}$，重构损失可以表示为：

$$\mathcal{L}_{\text{rec}} = ||x - \hat{x}||^2$$

正则化损失：用于衡量编码器输出的潜在变量分布与先验分布之间的差异。通常使用 KL 散度作为正则化损失函数。对于给定的编码器输出的潜在变量分布 $q_\phi(z|x)$ 和先验分布 $p(z)$，正则化损失可以表示为：

$$\mathcal{L}_{\text{KL}} = \text{KL}(q_\phi(z|x)||p(z))$$

损失函数

最终，VAE 的损失函数是重构损失和正则化损失的组合，通常是将它们加权求和：

$$\mathcal{L} = \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}}$$

其中，λ_{rec} 和 λ_{KL} 是控制两个损失项权重的超参数。

研究动机

前面模型的特点：

- 通过一个隐变量 z 来生成样本数据分布的 x，其主要思想是通过 decoder 或者是对抗生成网络等模型来学习样本的分布。
- 通过不断最小化生成样本与输入样本的误差来训练出一个映射，我们期望于生成样本与输入样本的概率分布尽可能的相似，将给定的数据映射到原来的数据分布中。

有什么问题吗？

研究动机

我们该如何度量这个相似？

- KL 散度？

研究动机

我们该如何度量这个相似？

- KL 散度？
- KL 散度是度量两个概率分布已知的分布之间的差异，而我们的输入数据的分布和生成数据的分布都是未知的，这就导致了我们的可解释性较差。

研究动机

我们该如何度量这个相似？

- KL 散度？
- KL 散度是度量两个概率分布已知的分布之间的差异，而我们的输入数据的分布和生成数据的分布都是未知的，这就导致了我们的可解释性较差。
- 且 VAE 中对于变分后验的计算量难以承受。

研究动机

我们该如何度量这个相似？

- KL 散度？
- KL 散度是度量两个概率分布已知的分布之间的差异，而我们的输入数据的分布和生成数据的分布都是未知的，这就导致了我们的可解释性较差。
- 且 VAE 中对于变分后验的计算量难以承受。
- 扩散模型 [HJA20] 应运而生。

基本性质

扩散过程是一个马尔科夫链，具有平稳性。

- 一个概率分布如果随时间变化，那么在马尔可夫链的作用下，它一定会趋于某种平稳分布。

基本性质

扩散过程是一个马尔科夫链，具有平稳性。

- 一个概率分布如果随时间变化，那么在马尔可夫链的作用下，它一定会趋于某种平稳分布。
- 只要终止时间足够长，概率分布就会趋近于这个平稳分布。

基本性质

扩散过程是一个马尔科夫链，具有平稳性。

- 一个概率分布如果随时间变化，那么在马尔可夫链的作用下，它一定会趋于某种平稳分布。
- 只要终止时间足够长，概率分布就会趋近于这个平稳分布。
- 扩散模型的本质是不断在当前的状态下对图像加噪声，也就是说，在这里的马尔科夫的转移过程，每一次转移都是在对图像加噪。

基本性质

扩散过程是一个马尔科夫链，具有平稳性。

- 一个概率分布如果随时间变化，那么在马尔可夫链的作用下，它一定会趋于某种平稳分布。
- 只要终止时间足够长，概率分布就会趋近于这个平稳分布。
- 扩散模型的本质是不断在当前的状态下对图像加噪声，也就是说，在这里的马尔科夫的转移过程，每一次转移都是在对图像加噪。
- 扩散模型最后的稳定状态是一个各向同性的噪声图片。

基本性质

扩散过程是一个马尔科夫链，具有平稳性。

- 一个概率分布如果随时间变化，那么在马尔可夫链的作用下，它一定会趋于某种平稳分布。
- 只要终止时间足够长，概率分布就会趋近于这个平稳分布。
- 扩散模型的本质是不断在当前的状态下对图像加噪声，也就是说，在这里的马尔科夫的转移过程，每一次转移都是在对图像加噪。
- 扩散模型最后的稳定状态是一个各向同性的噪声图片。
- 类似于向清水中滴入墨水，墨水在水中随时间晕染的过程。

扩散模型逆向过程

事实上，如果我们可以知道如何从当前状态 x_t 转移到前一个状态 x_{t-1}，那么根据马尔科夫链的传递性，我们就可以从各向同性的高斯分布的噪声中逐渐去噪，得到服从原本概率分布的生成数据。这样做有很大的裨益：

- 逐步演化的过程可以提高生成样本的质量，使其更接近目标分布，从而避免了传统生成模型中可能出现的模式崩溃或低质量样本的问题。
- 去噪效果：扩散模型的马尔科夫链通过逆向扩散过程，可以实现对噪声样本的去噪。
- 训练效率：扩散模型的马尔科夫链在训练过程中可以使用有效的数值方法进行迭代更新。相比于传统生成模型中的优化算法，如梯度下降，马尔科夫链的更新方法可以更高效地训练模型，并且具有更好的收敛性。

优化目标

在扩散模型中，训练的目标是通过最小化生成样本和真实样本之间的差异来优化模型参数。为此，我们可以引入一个损失函数，被称为 Denoising Score Matching (DSM) 损失函数。
DSM 损失函数可以通过比较生成样本 x_t 和真实样本 x 的局部对数密度来定义，如下所示：

$$\mathcal{L}(x_t, x) = \frac{1}{2}\|\nabla_x \log p(x_t) - \nabla_x \log p(x)\|^2$$

其中，∇_x 表示对 x 求梯度的操作。

简介

在去年“AI 作画”火出了圈，用户惊叹于人工智能的创造力，只需要输入自然语言的 prompt，人工智能就可以输出一幅满足自然语言意境指导的画作。AI 作画近期取得如此巨大进展的原因有很大的功劳归属于 Stable Diffusion 的开源。Stable diffusion 是一个基于 Latent Diffusion Models[RBL+21] 的文图生成模型。

Stable 在哪里?

- stable diffusion 旨在通过不停去除噪音来获得期望结果的一个生成式模型。在 AI 绘画早期，扩散是发生在像素空间 pixel space 的，不仅效果不好而且单张图大约需要 10-15 分钟，后来英国初创公司 Stability AI 对模型进行了改进，把核心计算从像素空间改到了潜空间 (latent space) 中，使得稳定性与像素质量都得到了极大提升，并且速度提高了近 100 倍，故名 stable diffusion。
- 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，大大降低了落地门槛，也带来了文图生成领域的大火。

研究动机

原来的 Diffusion Model 有什么问题？

- 直接在像素空间中对图像进行操作，性能较强的扩散模型通常需要消耗极大的计算资源，推断过程极为昂贵。

研究动机

原来的 Diffusion Model 有什么问题？

- 直接在像素空间中对图像进行操作，性能较强的扩散模型通常需要消耗极大的计算资源，推断过程极为昂贵。
- 同时，合成高分辨率的图像也面临一定的挑战。

研究动机

原来的 Diffusion Model 有什么问题？

- 直接在像素空间中对图像进行操作，性能较强的扩散模型通常需要消耗极大的计算资源，推断过程极为昂贵。
- 同时，合成高分辨率的图像也面临一定的挑战。
- 训练过程可能产生模式坍缩的问题，导致生成的样本缺乏多样性。

研究动机

原来的 Diffusion Model 有什么问题？

- 直接在像素空间中对图像进行操作，性能较强的扩散模型通常需要消耗极大的计算资源，推断过程极为昂贵。
- 同时，合成高分辨率的图像也面临一定的挑战。
- 训练过程可能产生模式坍缩的问题，导致生成的样本缺乏多样性。
- 难以精确控制生成样本的某些特征、样式或属性。

研究动机

Stable Diffusion 背后的模型 Latent Diffusion Models 解决了上述的问题。

- 解决传统扩散模型的资源消耗问题。
- 解决传统模型无法合成高分辨率的问题。
- 考虑通过 VAE 和 DM 的结合，将训练复杂度和模型生成质量达到一个最优的平衡。

研究方法

在原始的 Diffusion Model 中，通过多个时间步骤逐渐" 扩散" 噪声，将噪声逐渐转化为生成样本。这个过程中的每个时间步骤既包含了数据的压缩（通过学习潜在表示），也包含了数据的生成（通过逆向操作从潜在表示重建样本）。在这个过程中，压缩和生成是同时进行的，没有明确的分离阶段。

研究方法

在 LDMs 中，作者明确的将数据压缩和数据生成阶段分离，通过该方法，可以解决传统模型的问题。

- 压缩学习阶段：在这个阶段，LDMs 通过对数据进行编码，学习到一个低维的潜在空间表示。这个阶段类似于一个编码器（Encoder）的训练过程。压缩学习阶段的目标是将输入数据映射到一个潜在空间中的低维表示，以便有效地捕捉数据的重要特征。
- 生成学习阶段：在这个阶段，LDMs 通过对潜在空间进行解码，学习到如何从潜在空间生成高质量的样本。这个阶段类似于一个解码器（Decoder）的训练过程。生成学习阶段的目标是学习如何从潜在空间中采样，并通过解码器生成与原始数据相似的样本。
- 我们考虑，对于数字图像大部分的像素值实际上都对应着我们无法感知的细节。本文的 LDMs 可以有效的删除无法感知的细节以节约算力。

阶段学习分离方法

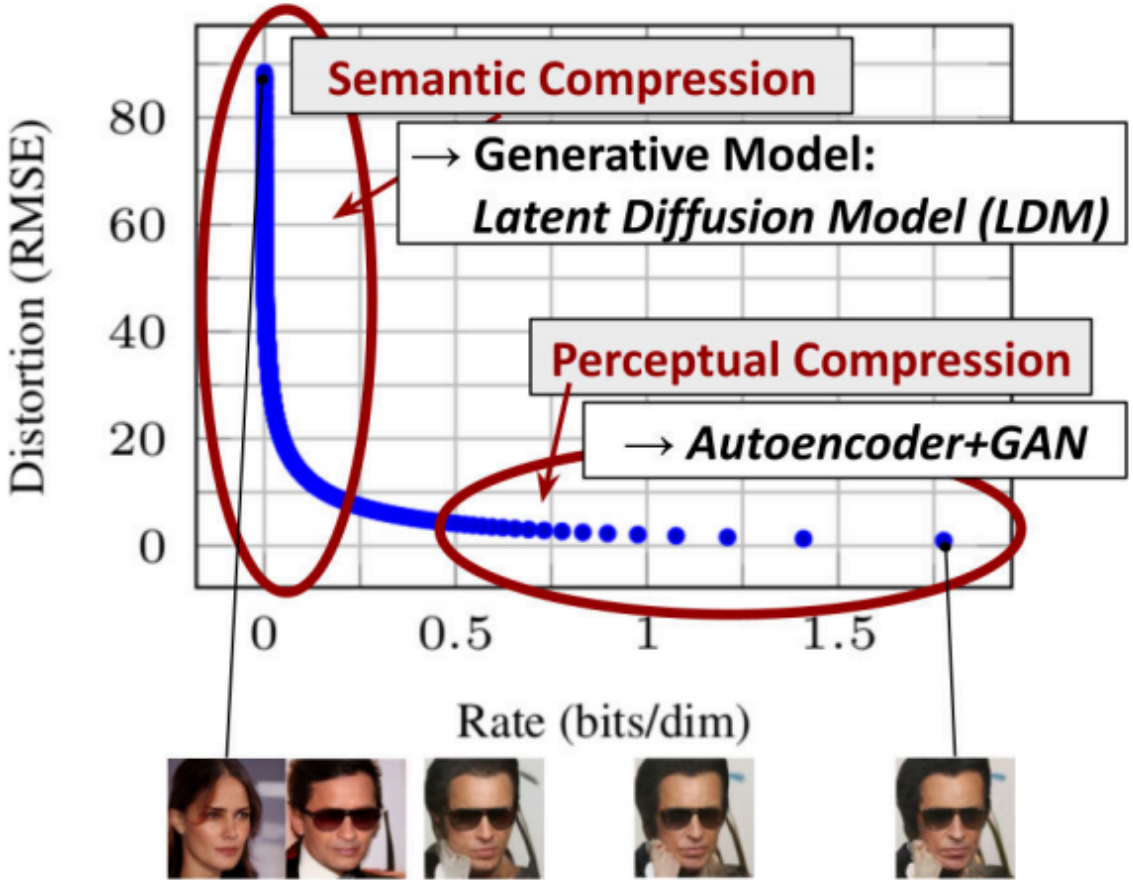

图 3: 阶段学习分离方法

实验结果

在图像修复，图像超分，图像生成等实验上取得了极好的效果：

实验结果

在图像修复，图像超分，图像生成等实验上取得了极好的效果：

四种模型各自特点

GAN：GAN 是一种基于对抗训练的生成模型，通过训练一个生成器和一个判别器来实现图像生成。GAN 具有以下特点：

- 通过生成器和判别器之间的对抗过程，可以逐渐提升生成器的生成能力。
- 通过最小化生成器和判别器之间的损失函数，可以实现生成图像的多样性和质量。
- 可以通过引入条件信息或其他扩展技术，实现有条件的图像生成。

四种模型各自特点

VAE：VAE 是一种基于变分推断的生成模型，通过学习潜在空间中的编码和解码过程来生成图像。VAE 具有以下特点：

- 通过学习潜在空间的分布，可以实现对图像生成过程的建模和控制。
- 通过引入 KL 散度项或其他正则化方法，可以约束潜在空间的结构和平滑性。
- 可以使用重参数化技巧有效地进行训练和推断。

四种模型各自特点

DMs：DMs 也是一种基于扩散过程的生成模型，通过在像素空间中进行渐进的反向扩散来生成图像。DMs 与 LDMs 的区别在于采样空间的不同，DMs 采样像素空间，而 LDMs 采样潜在空间。DMs 的特点包括：

- 通过最小化与感知无关的细节的损失项，可以忽略感知上不相关的细节。
- 在像素空间中需要进行昂贵的函数评估，导致计算时间和能源资源的巨大需求。
- 具有 UNet 架构的归纳偏差，对具有空间结构的数据特别有效。

四种模型各自特点

LDMs：LDMs 是一种基于扩散过程的生成模型，通过在潜在空间中进行渐进的扩散来生成图像。LDMs 具有以下特点：

- 通过在低维空间上进行采样，提供了更高效的图像生成过程。
- 提供通用的压缩模型，其潜在空间可用于训练多个生成模型，也可用于其他下游应用。
- 对于多种应用如修复，超分等方向具有极为先进的效果。

存在的问题

- 图像质量与多样性平衡。
- 模式崩溃和模式坍缩。
- 长期依赖和一致性。
- 解释性和可控性。
- 训练和推理的效率。

未来的方向

- 提高生成模型的质量和多样性：研究人员可以探索新的网络架构、损失函数和训练方法，以改善生成模型生成图像、文本、音频等领域的质量和多样性。
- 可控性和解释性的生成：开发可控的生成模型，使用户能够在生成过程中精确控制生成结果的某些属性。同时，提高生成模型的解释性，使生成结果的生成过程更加可解释和可理解。
- 长期依赖和一致性建模：可以探索新的架构和算法来捕捉长期依赖，以生成更连贯和一致的序列结果。
- 训练和推理的效率：可以通过剪枝、量化、并行计算等方法来减少生成模型的计算资源需求，以实现更快速、更高效的训练和推理过程。

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial nets.
pages 2672–2680, 2014.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel.
Denoising diffusion probabilistic models, 2020.

[KW22] Diederik P Kingma and Max Welling.
Auto-encoding variational bayes, 2022.

[RBL+21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer.
High-resolution image synthesis with latent diffusion models, 2021.

Thanks!