# Stable Diffusion原理解读

zhuanlan.zhihu.com/p/583124756

SeanAI/NLP/AIGC



## 引言

最近大火的AI作画吸引了很多人的目光，AI作画近期取得如此巨大进展的原因个人认为有很大的功劳归属于Stable Diffusion的开源。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，大大降低了落地门槛，也带来了文图生成领域的大火。所以，如果你想了解Stable Diffusion的背后原理，可以跟我一起深入解读一下其背后的论文**High-Resolution Image Synthesis with Latent Diffusion Models（Latent Diffusion Models），**同时这篇文章后续也会针对ppdiffusers的相关代码进行讲解。该论文发表于CVPR2022，第一作者是Robin Rombach，来自德国慕尼黑大学机器视觉与学习研究小组。

## Stable Diffusion

再解读论文之前，首先让我深入了解一下Stable Diffusion。

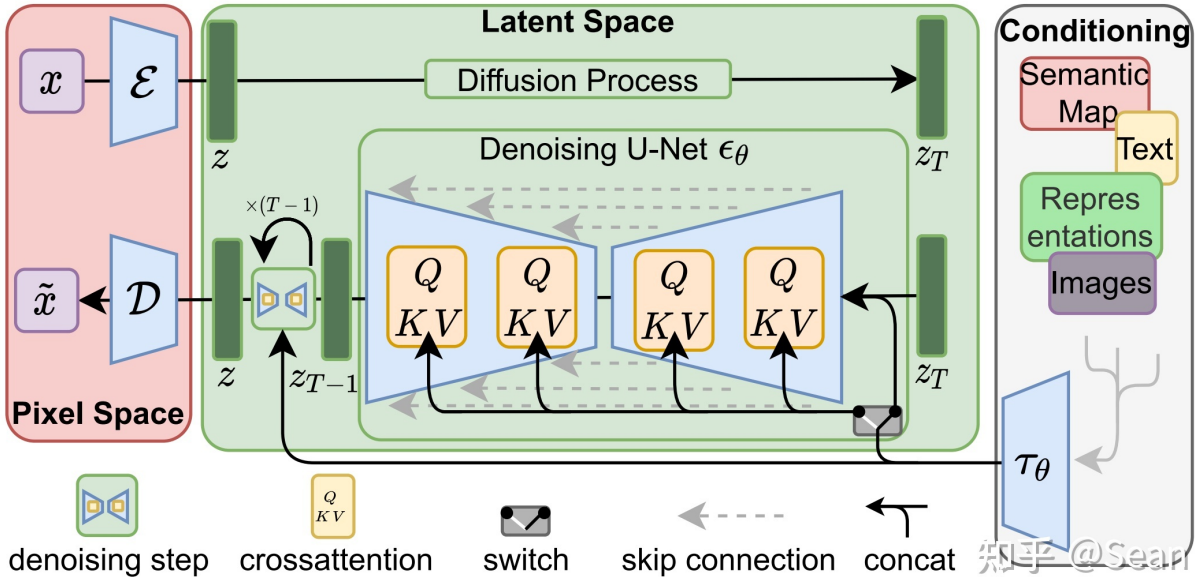
Stable Diffusion基于Latent Diffusion Models，专门用于文图生成任务。目前，Stable Diffusion发布了v1版本，即Stable Diffusion v1，它是Latent Diffusion Models的一个具体实现，具体来说，它特指这样的一个模型架构设置：自动编码器下采样因子为8，UNet大小为860M，文本编码器为CLIP ViT-L/14。官方目前提供了以下权重：

1. `sd-v1-1.ckpt` : 237k steps at resolution `256x256` on laion2B-en. 194k steps at resolution `512x512` on laion-high-resolution (170M examples from LAION-5B with resolution `>= 1024x1024` ).
2. `sd-v1-2.ckpt` : Resumed from `sd-v1-1.ckpt` . 515k steps at resolution `512x512` on laion-aesthetics v2 5+ (a subset of laion2B-en with estimated aesthetics score `> 5.0` , and additionally filtered to images with an original size `>= 512x512` , and an estimated watermark probability `< 0.5` . The watermark estimate is from the LAION-5B metadata, the aesthetics score is estimated using the LAION-Aesthetics Predictor V2).
3. `sd-v1-3.ckpt` : Resumed from `sd-v1-2.ckpt` . 195k steps at resolution `512x512` on "laion-aesthetics v2 5+" and 10% dropping of the text-conditioning to improve classifier-free guidance sampling.
4. `sd-v1-4.ckpt` : Resumed from `sd-v1-2.ckpt` . 225k steps at resolution `512x512` on "laion-aesthetics v2 5+" and 10% dropping of the text-conditioning to improve classifier-free guidance sampling.

## 论文贡献

- Diffusion model相比GAN可以取得更好的图片生成效果，然而该模型是一种自回归模型，需要反复迭代计算，因此训练和推理代价都很高。论文提出一种在潜在表示空间（latent space）上进行diffusion过程的方法，从而能够大大减少计算复杂度，同时也能达到十分不错的图片生成效果。
- 相比于其它空间压缩方法（如），论文提出的方法可以生成更细致的图像，并且在高分辨率图片生成任务（如风景图生成，百万像素图像）上表现得也很好。
- 论文将该模型在无条件图片生成（unconditional image synthesis）, 图片修复（inpainting）,图片超分（super-resolution）任务上进行了实验，都取得了不错的效果。
- 论文还提出了cross-attention的方法来实现多模态训练，使得条件图片生成任务也可以实现。论文中提到的条件图片生成任务包括类别条件图片生成（class-condition），文图生成（text-to-image）, 布局条件图片生成（layout-to-image）。这也为日后Stable Diffusion的开发奠定了基础。

## 方法




Latent Diffusion Models整体框架如图，首先需要训练好一个自编码模型（AutoEncoder，包括一个编码器 $E$ 和一个解码器 $D$）。这样一来，我们就可以利用编码器对图片进行压缩，然后在潜在表示空间上做diffusion操作，最后我们再用解码器恢复到原始像素空间即可，论文将这个方法称之为感知压缩（Perceptual Compression）。个人认为这种将高维特征压缩到低维，然后在低维空间上进行操作的方法具有普适性，可以很容易推广到文本、音频、视频等领域。

在潜在表示空间上做diffusion操作其主要过程和标准的扩散模型没有太大的区别，所用到的扩散模型的具体实现为 time-conditional UNet。但是有一个重要的地方是论文为diffusion操作引入了条件机制（Conditioning Mechanisms），通过cross-attention的方式来实现多模态训练，使得条件图片生成任务也可以实现。

下面我们针对感知压缩、扩散模型、条件机制的具体细节进行展开。

### 图片感知压缩（Perceptual Image Compression）

感知压缩本质上是一个tradeoff，之前的很多扩散模型没有使用这个技巧也可以进行，但原有的非感知压缩的扩散模型有一个很大的问题在于，由于在像素空间上训练模型，如果我们希望生成一张分辨率很高的图片，这就意味着我们训练的空间也是一个很高维的空间。引入感知压缩就是说通过VAE这类自编码模型对原图片进行处理，忽略掉图片中的高频信息，只保留重要、基础的一些特征。这种方法带来的的好处就像引文部分说的一样，能够大幅降低训练和采样阶段的计算复杂度，让文图生成等任务能够在消费级GPU上，在10秒级别时间生成图片，大大降低了落地门槛。

感知压缩主要利用一个预训练的自编码模型，该模型能够学习到一个在感知上等同于图像空间的潜在表示空间。这种方法的一个优势是只需要训练一个通用的自编码模型，就可以用于不同的扩散模型的训练，在不同的任务上使用。这样一来，感知压缩的方法除了应用

在标准的无条件图片生成外，也可以十分方便的拓展到各种图像到图像（inpainting，super-resolution）和文本到图像（text-to-image）任务上。

由此可知，基于感知压缩的扩散模型的训练本质上是一个两阶段训练的过程，第一阶段需要训练一个自编码器，第二阶段才需要训练扩散模型本身。在第一阶段训练自编码器时，为了避免潜在表示空间出现高度的异化，作者使用了两种正则化方法，一种是KL-reg，另一种是VQ-reg，因此在官方发布的一阶段预训练模型中，会看到KL和VQ两种实现。在Stable Diffusion中主要采用AutoencoderKL这种实现。

具体来说，给定图像 $x \in R^{H \times W \times 3}$，我们可以先利用一个编码器 $E$ 来将图像编码到潜在表示空间 $z = E(x)$，其中 $z \in R^{h \times w \times c}$，然后再用解码器从潜在表示空间重建图片 $\tilde{x} = D(z) = D(E(x))$。在感知压缩压缩的过程中，下采样因子的大小为 $f = H/h = W/w$，它是2的次方，即 $f = 2^m$。

## 潜在扩散模型（Latent Diffusion Models）

首先简要介绍一下普通的扩散模型（DM），扩散模型可以解释为一个时序去噪自编码器（equally weighted sequence of denoising autoencoders）$\epsilon_\theta(x_t, t); t = 1 \dots T$，其目标是根据输入 $x_t$ 去预测一个对应去噪后的变体，或者说预测噪音，其中 $x_t$ 是输入 $x$ 的噪音版本。相应的目标函数可以写成如下形式：

$$L_{DM} = E_{x, \epsilon \sim N(0,1), t}[\| \epsilon - \epsilon_\theta(x_t, t) \|_2^2]$$

其中 $t$ 从 $\{1, \dots, T\}$ 中均匀采样获得。

而在潜在扩散模型中，引入了预训练的感知压缩模型，它包括一个编码器 $E$ 和一个解码器 $D$。这样就可以利用在训练时就可以利用编码器得到 $z_t$，从而让模型在潜在表示空间中学习，相应的目标函数可以写成如下形式：

$$L_{LDM} := E_{E(x), \epsilon \sim N(0,1), t}[\| \epsilon - \epsilon_\theta(z_t, t) \|_2^2]$$

## 条件机制（Conditioning Mechanisms）

除了无条件图片生成外，我们也可以进行条件图片生成，这主要是通过拓展得到一个条件时序去噪自编码器（conditional denoising autoencoder） $\epsilon_\theta(z_t, t, y)$ 来实现的，这样一来我们就可通过 $y$ 来控制图片合成的过程。具体来说，论文通过在UNet主干网络上增加cross-attention机制来实现 $\epsilon_\theta(z_t, t, y)$。为了能够从多个不同的模态预处理 $y$，论文引入了一个领域专用编码器（domain specific encoder） $\tau_\theta$

，它用来将 $y$ 映射为一个中间表示 $\tau_\theta(y) \in R^{M \times d_\tau}$，这样我们就可以很方便的引入各种形态的条件（文本、类别、layout等等）。最终模型就可以通过一个cross-attention层映射将控制信息融入到UNet的中间层，cross-attention层的实现如下：

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, with$$

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), K = W_K^{(i)} \cdot \tau_\theta(y), V = W_V^{(i)} \cdot \tau_\theta(y)$$

其中 $\varphi_i(z_t) \in R^{N \times d_\epsilon^i}$ 是UNet的一个中间表征。相应的目标函数可以写成如下形式：

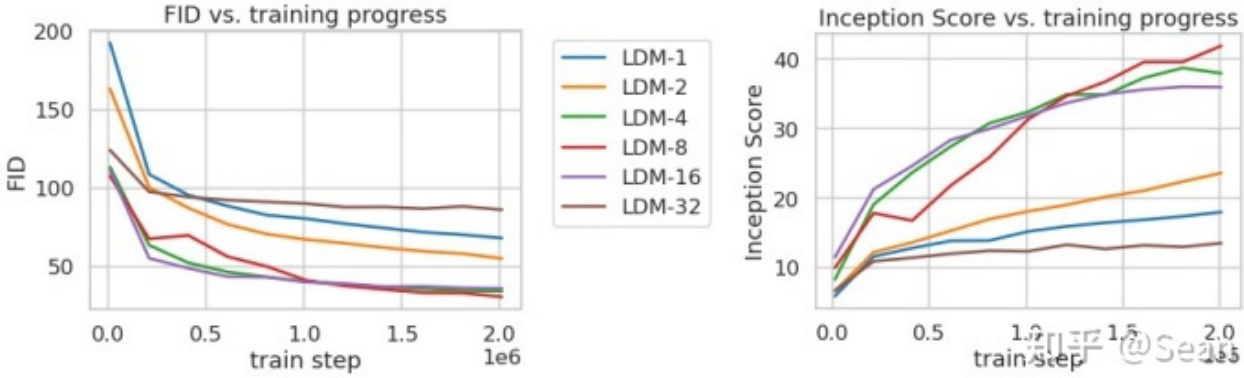
$$L_{LDM} := E_{E(x), y, \epsilon \sim N(0,1), t}\left[\| \epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y)) \|_2^2\right]$$

## 实验

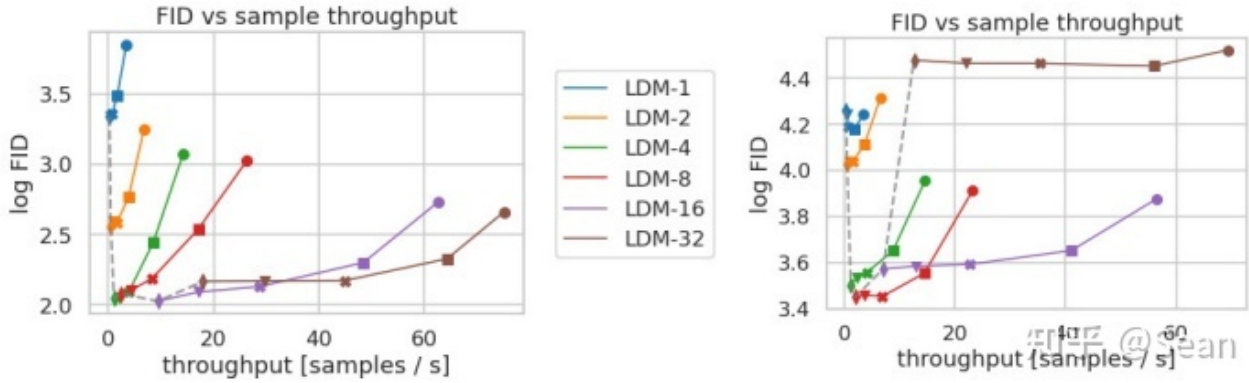
论文的所用到的模型为LDMs，在无条件图片生成任务上用到的数据集为LSUN、FFHQ以及CelebA-HQ，在类别条件图片生成用到的数据集为ImageNet，在文图生成任务上用到的数据集为Conceptual Captions、LAION。论文设计了大量的对比实验，并分别对感知压缩权衡（下采样因子）、LDM生成效果对比进行了分析验证。并且还在其他任务上进行了实验，包括Super-Resolution、Inpainting、layout-condition在内的多种图片生成等任务，这说明说明LDMs中的学习到的潜在表示空间确实具备很强的分布拟合能力，能够够适配各种下游任务。

### 感知压缩权衡（Perceptual Compression Tradeoffs）

前面提到过下采样因子 $f$ 的大小为 $f = H/h = W/w$，如果 $f = 1$ 那就等于没有对输入的像素空间进行压缩，如果 $f$ 越大，则信息压缩越严重，可能会噪声图片失真，但是训练资源占用的也越少。论文对比了 $f$ 在分别 $\{1, 2, 4, 8, 16, 32\}$ 下的效果，发现 $f$ 在 $\{4-16\}$ 之间可以比较好的平衡压缩效率与视觉感知效果。作者重点推荐了*LDM-4*和*LDM-8*。




下采样因子对比实验，横坐标train step，左纵坐标FID，右纵坐标Inception Score




下采样因子对比实验，横坐标throughtput，纵坐标log FID，左CelebA-HQ数据集，右ImageNet数据集

## LDM生成效果（Image Generation with Latent Diffusion）

论文从FID和Precision-and-Recall两方面对比LDM的样本生成能力，实验数据集为CelebA-HQ、FFHQ和LSUN-Churches/Bedrooms，实验结果如下：

| CelebA-HQ 256 × 256 | | | | FFHQ 256 × 256 | | | |
|---|---|---|---|---|---|---|---|
| **Method** | FID ↓ | Prec. ↑ | Recall ↑ | **Method** | FID ↓ | Prec. ↑ | Recall ↑ |
| DC-VAE [61] | 15.8 | - | - | ImageBART [21] | 9.57 | - | - |
| VQGAN+T. [23] (k=400) | 10.2 | - | - | U-Net GAN (+aug) [75] | 10.9 (7.6) | - | - |
| PGGAN [38] | 8.0 | - | - | UDM [42] | 5.54 | - | - |
| LSGM [90] | 7.22 | - | - | StyleGAN [40] | 4.16 | 0.71 | 0.46 |
| UDM [42] | 7.16 | - | - | ProjectedGAN [74] | **3.08** | 0.65 | 0.46 |
| *LDM-4* (ours, 500-s†) | **5.11** | 0.72 | 0.49 | *LDM-4* (ours, 200-s) | 4.98 | **0.73** | **0.50** |

| LSUN-Churches 256 × 256 | | | | LSUN-Bedrooms 256 × 256 | | | |
|---|---|---|---|---|---|---|---|
| **Method** | FID ↓ | Prec. ↑ | Recall ↑ | **Method** | FID ↓ | Prec. ↑ | Recall ↑ |
| DDPM [29] | 7.89 | - | - | ImageBART [21] | 5.51 | - | - |
| ImageBART [21] | 7.32 | - | - | DDPM [29] | 4.9 | - | - |
| PGGAN [38] | 6.42 | - | - | UDM [42] | 4.57 | - | - |
| StyleGAN [40] | 4.21 | - | - | StyleGAN [40] | 2.35 | 0.59 | 0.48 |
| StyleGAN2 [41] | 3.86 | - | - | ADM [15] | 1.90 | **0.66** | **0.51** |
| ProjectedGAN [74] | **1.59** | 0.61 | 0.44 | ProjectedGAN [74] | **1.52** | 0.61 | 0.34 |
| *LDM-8** (ours, 200-s) | 4.02 | **0.64** | **0.52** | *LDM-4* (ours, 200-s) | 2.95 | **0.66** | 0.48 |

其效果超过了GANs和LSGM，并且超过同为扩散模型的DDPM。

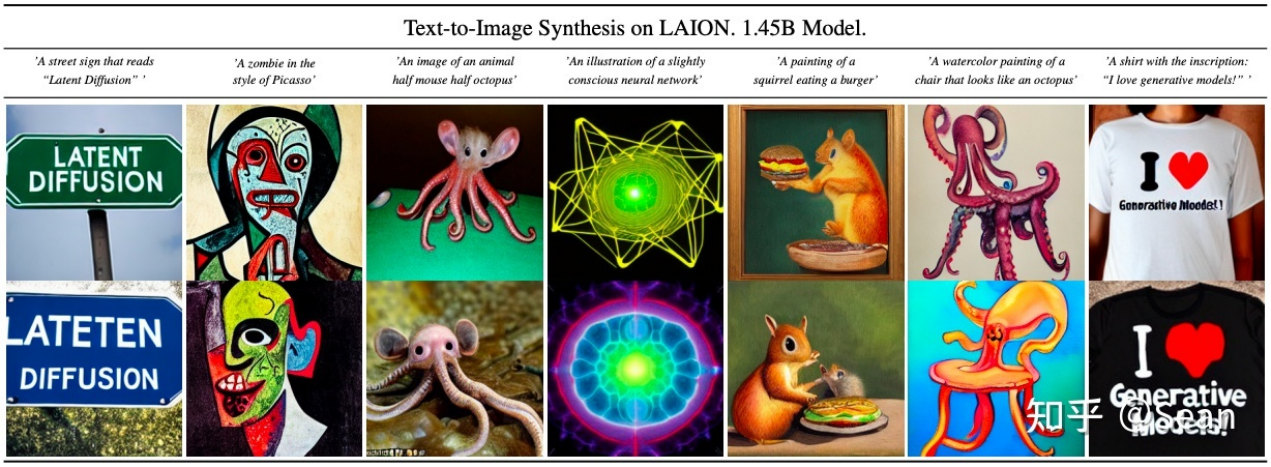
## 效果展示

看一下在各个任务上的效果。

无条件图片生成（unconditional-image ）：


类别条件图片生成（unconditional-image ）：

文图生成（text-to-image）：


布局图片生成（layout-to-image）：


超分辨率（super-resolution）：

图片修复（inpainting, object removal）：


风景图语义合成（semantic-to-image，semantic synthesis of landscape images）：


## 参考