8 生成扩散模型漫谈（二十二）：信噪比与大图生成（上）

Apr By 苏剑林 | 2024-04-08 | 45361位读者 引用

盘点主流的图像扩散模型作品，我们会发现一个特点：当前多数做高分辨率图像生成（下面简称“大图生成”）的工作，都是先通过Encoder变换到Latent空间进行的（即LDM，Latent Diffusion Model），直接在原始Pixel空间训练的扩散模型，大多数分辨率都不超过64*64，而恰好，LDM通过AutoEncoder变换后的Latent，大小通常也不超过64*64。这就自然引出了一系列问题：扩散模型是不是对于高分辨率生成存在固有困难？能否在Pixel空间直接生成高分辨率图像？

论文《Simple diffusion: End-to-end diffusion for high resolution images》尝试回答了这个问题，它通过“信噪比”分析了大图生成的困难，并以此来优化noise schdule，同时提出只需在最低分辨率feature上对架构进行scale up、多尺度Loss等技巧来保证训练效率和效果，这些改动使得原论文成功在Pixel空间上训练了分辨率高达1024*1024的图像扩散模型。

LDM回顾

在进入正题之前，我们不妨先反过来想一想：为什么LDM能成功成为主流的扩散模型做法？笔者认为，主要原始是两方面：

> 1、不管是应用还是学术，用LDM的主要原因想必是效率：当前主流的工作都直接重用了LDM论文所开源的训练好的AutoEncoder，它的Encoder部分会将512*512的图像变成了64*64的Latent，相当于说只用到64*64分辨率这个级别的算力和时间，就可以生成512*512的图像，这个效率显然是非常吸引人的；
>
> 2、LDM契合了FID这个指标，这让它看起来是效果无损的：FID全称是“Fréchet Inception Distance”，其中Inception是指用ImageNet预训练的InceptionV3模型作为Encoder编码图片，然后假设编码特征服从高斯分布来算$\mathcal{W}$距离，而LDM也是先Encoder编码，两个Encoder虽然不完全相同，但也有一定共性，因此在FID上表现为几乎无损。

我们还可以稍微展开一下。LDM的AutoEncoder在训练阶段组合了很多内容——它的重构Loss并不只有常规的MAE或者MSE，还包括对抗Loss和Perceptual Loss，对抗Loss用来保证重构结果的清晰度，而Perceptual Loss用来保证重构结果的语义和风格的相似性。Perceptual Loss跟FID很相似，都是用ImageNet模型的特征计算的相似性指标，只不过用的不是InceptionV3而是VGG-16，由于训练任务的相似性，可以猜测两者特征有很多共性，因此Perceptual Loss的加入变相地保证了FID的损失尽可能少。

此外，LDM的Encoder对原始图像来说是降维的，比如原始图像大小为512*512*3，直接patchify的话结果是64*64*192，但LDM的Encoder出来的特征是64*64*4，降低到了1/48，同时为了进一步降低编码特征的方差，避免模型“死记硬背”，LDM还对Encoder出来的特征加了相应的正则项，可选的有VAE的KL散度项或VQ-VAE的VQ正则化。降维和正则的设计，都会压缩特征的多样性，提高特征的泛化能力，但也会导致重构难度增加，最终导致了有损的重构结果。

到这里，LDM能成功的原因其实就“豁然开朗”了：“降维 + 正则”的组合，降低了Latent的信息量，从而降低了在Latent空间学习扩散模型的难度，同时Perceptual Loss的存在，保证了重构虽然有损但FID几乎无损（Perceptual Loss的Encoder跟FID一样都用InceptionV3理论上更好）。这样一来，对于FID这个指标来说，LDM几乎就是免费午餐了，因此不管是学术和工程都乐意沿用它。

信噪之比

尽管LDM简单高效，但它毕竟是有损的，其Latent只能保持宏观上的语义，局部细节可能会严重缺失。而在之前的文章《“闭门造车”之多模态思路浅谈（一）：无损输入》中，笔者表达过一个观点：当作为输入时，图像最好的表示方式就是原始Pixel数组。基于这个观点，笔者最近都比较关注直接在Pixel空间上训练的扩散模型。

然而，将低分辨率（比如64*64）图像的扩散模型配置直接应用于高分辨率（比如512*512）的大图生成时，会存在算力消耗过大、收敛速度太慢等问题，而且效果上也比不上LDM（至少FID指标如此），Simple diffusion逐一分析了这些问题并提出了相应的解决方案。其中，笔者认为利用“信噪比（Signal-to-Noise Ratio，SNR）”的概念来分析高分辨率扩散模型的学习效率低问题最为精彩。

具体来说，Simple diffusion观察到，如果我们给高分辨率图像加上某个方差的noise，那么相对于加上同样方差的noise的低分辨率图像来说，它的信噪比其实更高，原论文的Figure 3非常直观地演示了这一点，如下图所示。第一行图片，是由512*512的图片加了特定方差的noise后再降采样（平均Pooling）到64*64的，而第二行则是直接在64*64的图片加上同样方差的noise，很明显第一行的图片更加清晰，也就是相对信噪比更高了。

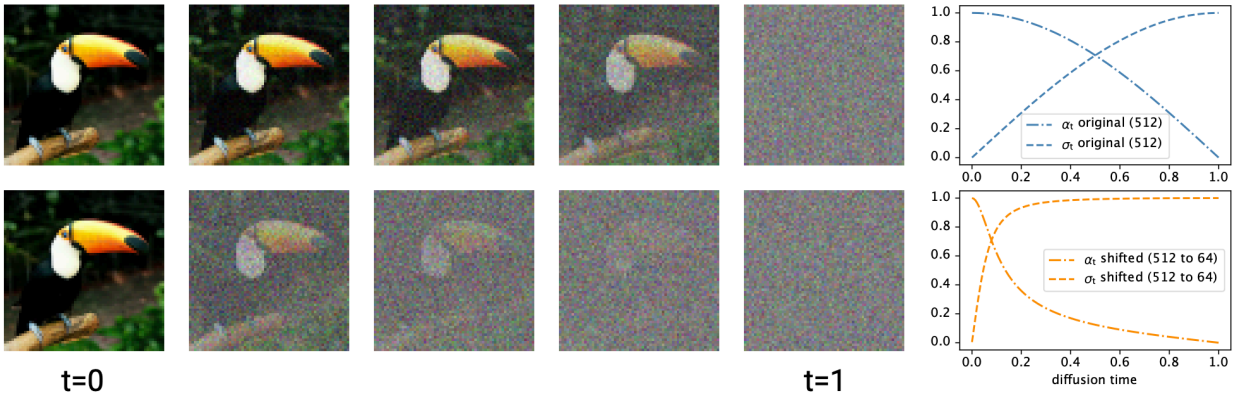

Figure 3: The standard and shifted diffusion noise on an image of 512×512, that is visualized by average pooling to a resolution of 64×64. The top row shows a conventional cosine schedule, the bottom row shows our proposed shifted schedule.

同一noise不同分辨率的信噪比

所谓“信噪比”，顾名思义即“信号与噪声的强度之比”，信噪比更高（即噪声的占比更低）意味着去噪更容易，换言之训练阶段Denoiser面对的更多是简单样本，但实际上大图生成的难度显然更高，也就是说我们的目标是一个更难的模型，但却给了更简单的样本，因此导致了学习效率的低下。

向低看齐

我们也可以从数学上描述这一点。沿用本系列的记号，通过加噪来构造$\boldsymbol{x}_t$的运算可以表示为

$$\boldsymbol{x}_t = \bar{\alpha}_t \boldsymbol{x}_0 + \bar{\beta}_t \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{1}$$

其中$\bar{\alpha}_t, \bar{\beta}_t$就称为noise schedule，满足$\bar{\alpha}_0 = \bar{\beta}_T = 1, \bar{\alpha}_T = \bar{\beta}_0 = 0$，此外一般来说它们还有额外的约束，比如在DDPM中通常是$\bar{\alpha}_t^2 + \bar{\beta}_t^2 = 1$，本文将沿用这个约束。

对于一个随机变量来说，信噪比是均值平方与方差之比。给定$\boldsymbol{x}_0$，$\boldsymbol{x}_t$的均值显然是$\bar{\alpha}_t\mathbb{E}[\boldsymbol{x}_0]$，方差则是$\bar{\beta}_t^2$，于是信噪比为$\frac{\bar{\alpha}_t^2}{\bar{\beta}_t^2}\mathbb{E}[\boldsymbol{x}_0]^2$，由于我们总是在给定$\boldsymbol{x}_0$下讨论，因此我们也可以简单地说信噪比就是$SNR(t)=\frac{\bar{\alpha}_t^2}{\bar{\beta}_t^2}$

当我们对$\boldsymbol{x}_t$使用$s\times s$大小的平均Pooling时，每个$s\times s$的patch通过取平均变成了一个标量，即

$$\frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\boldsymbol{x}_t^{(i,j)} = \bar{\alpha}_t\left(\frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\boldsymbol{x}_0^{(i,j)}\right) + \bar{\beta}_t\left(\frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\boldsymbol{\varepsilon}^{(i,j)}\right),\quad \boldsymbol{\varepsilon}^{(i,j)}\sim\mathcal{N}$$

平均Pooling不改变均值，但会降低方差，从而提高信噪比，这是因为正态分布的可加性得出

$$\frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\boldsymbol{\varepsilon}^{(i,j)}\sim\mathcal{N}(0,1/s^2) \tag{3}$$

所以在同一noise schedule下，如果我们将高分辨率图像通过平均Pooling来对齐低分辨率，那么就会发现信噪比更高，是原来的s^2倍：

$$SNR^{w\times h\to w/s\times h/s}(t) = SNR^{w/s\times h/s}(t)\times s^2 \tag{4}$$

反过来想，如果我们已经有一个在低分辨率图像上调好了的noise schedule $\bar{\alpha}_t^{w/s\times h/s},\bar{\beta}_t^{w/s\times h/s}$，那么当我们想要scale up到更高分辨率时，应该要调整noise schedule为$\bar{\alpha}_t^{w\times h},\bar{\beta}_t^{w\times h}$，使得它降采样到低分辨率后，其信噪比能够跟已经调好的低分辨率的noise schedule对齐，这样才能最大程度上“传承”已经低分辨率扩散模型的学习效率，即

$$\frac{(\bar{\alpha}_t^{w\times h})^2}{(\bar{\beta}_t^{w\times h})^2}\times s^2 = \frac{(\bar{\alpha}_t^{w/s\times h/s})^2}{(\bar{\beta}_t^{w/s\times h/s})^2} \tag{5}$$

如果加上约束$\bar{\alpha}_t^2+\bar{\beta}_t^2=1$，那么就可以从$\bar{\alpha}_t^{w/s\times h/s},\bar{\beta}_t^{w/s\times h/s}$中唯一地解出$\bar{\alpha}_t^{w\times h},\bar{\beta}_t^{w\times h}$。这就解决了高分辨率扩散的noise schedule设置问题。

架构拓展

为了做好大图的扩散生成，除了要调整noise schedule之外，我们还需要把架构也scale up上去，因为前面我们也已经说了，大图生成是一个更难的问题，因此理应需要更加重量级的架构。

扩散模型常用的就是U-Net或者U-Vit，两者都是先逐渐降采样然后逐渐上采样，比如512*512的输入，一般先进行一个block的运算，然后降采样到256*256，接着进行新一个block的运算，在降采样到128*128，依此类推，降采样到一个最低的分辨率16*16，接下来再次重复这个过程，但将降采样改为上采样，直到分辨率恢复512*512。默认设置下，我们会将参数平均分到每一个block中，但这样一来靠近输入和输出的block由于输入尺寸都很大，因此计算量会急剧增加，导致模型训练效率低下甚至不可行。

Simple diffusion提出了两个应对方案。第一，它提出可以直接在第一层（而不是第一个block，每个block有多个层）之后就降采样，并且考虑一步到位低降到128*128甚至64*64，最后输出的时候，也是在最后一层之前才从64*64或者128*128直接上采样到512*512，这样模型的大部分block所处理的分辨率都降低了，从而降低了整体计算量；第二，它提出将模型所scale up的层都放到最低分辨率（即16*16）之后，而不是平摊到每一个分辨率的block，即新增的层处理的都是16*16的输入，包括Dropout也都只加入到低分辨率的层中，这样一来分辨率增加带来的计算压力就明显减少了。

此外，为了进一步稳定训练，论文提出了“多尺度Loss”的训练目标。默认情况下，扩散模型的Loss等价于MSE损失

$$\mathcal{L} = \frac{1}{wh}\|\boldsymbol{\varepsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t \boldsymbol{x}_0 + \bar{\beta}_t \boldsymbol{\varepsilon}, t)\|^2 \tag{6}$$

Simple diffusion将它泛化为

$$\mathcal{L}_{s\times s} = \frac{1}{(w/s)(h/s)}\left\|\mathcal{D}_{w/s\times h/s}[\boldsymbol{\varepsilon}] - \mathcal{D}_{w/s\times h/s}[\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t \boldsymbol{x}_0 + \bar{\beta}_t \boldsymbol{\varepsilon}, t)]\right\|^2 \tag{7}$$

其中$\mathcal{D}_{w/s\times h/s}[\cdot]$是通过平均Pooling将输入变换到$w/s \times h/s$的下采样算子，原论文取了多个$s$对应的Loss进行平均，作为最终的训练目标。这个多尺度Loss的目标也很明

显，跟通过信噪比对齐来调整noise schedule一样，都是保证训练出来的高分辨率扩散模型至少不差于直接训练的低分辨率模型。

至于实验部分，大家自行看原论文就好。Simple diffusion实验的最大分辨率是1024*1024（在附录中提到），效果都尚可，并且对比实验表明上述提出的一些技巧都是有提升的，最终直接在Pixel空间中训练出来的扩撒模型，相比LDM也取得了有竞争力的效果。

文章小结

在这篇文章中，我们介绍了Simple diffusion，这是一篇探索如何直接在Pixel空间中端到端地训练图像扩散模型的工作，利用了信噪比的概念介绍了高分辨率扩散模型的训练效率低问题，并由此来指标调整新的noise schedule，以及探索了如何尽可能节约算力成本地scale up模型架构。

如果您需要引用本文，请参考：

苏剑林. (Apr. 08, 2024). 《生成扩散模型漫谈（二十二）：信噪比与大图生成（上）》[Blog post]. Retrieved from https://spaces.ac.cn/archives/10047

```
@online{kexuefm-10047,
    title={生成扩散模型漫谈（二十二）：信噪比与大图生成（上）},
    author={苏剑林},
    year={2024},
    month={Apr},
    url={\url{https://spaces.ac.cn/archives/10047}},
}
```