Denoising Diffusion Probabilistic Models

这是一份来自UC Berkeley（加州大学伯克利分校）的工作，发布时间为2020年。

本文地址为：https://arxiv.org/pdf/2006.11239.pdf

本文的代码仓库位于：https://github.com/hojonathanho/diffusion

零：概述

DDPM可以用来生成高质量的图片，这是一种被非平衡热力学启发的隐变量模型。基于扩散随机模型和Langvin动力学去噪分数的奇妙联系，我们得以训练出最好的结果。而且我们的模型天然接受渐进式有损压缩的方案，这可以被理解为是为一种自回归编码器。在无条件的CIFAR10数据集中，我们使用DDPM取得了Inception 9.64分的成绩，而目前最为先进的FID模型只有3.17分。在256 $\times$ 256的LSUN数据集中，我们得到的质量与ProgressiveGAN相近。

一：介绍

各种的深度生成模型最近都在不同的数据形式上展现了高质量的结果。GAN，自回归模型，flows，以及VAE都得以生成令人惊艳的图片和音频样例。而且energy-based模型和score matching生成了较之GAN模型更好的图片。

本文展现了扩散概率模型的过程。一个扩散概率模型（之后简称为DM）是一种参数化的马尔可夫链，被训练为使用变量推测来在有限时间内生成符合数据的样品。这一链条学习如何去逆转一个扩散过程（一个逐步向原图中添加噪声直到所有的信息都被摧毁）。因为扩散是由一系列的高斯噪声所组成的，所以使用简单的神经网络就可以进行这一过程。

扩散模型很容易去定义和训练，但据我们所知，并没有他们可以生成高质量样例的证明。本文证明了其生成高质量图像的能力甚至超过现有的许多模型。此外，我们还展示了，特定DM的参数化能揭示一种有关去噪分数和Langevin动力学的一致性。而且这种参数可以得到最好的结果，因此这种一致性也将成为本文的一个贡献。

尽管可以生成高质量的图片，我们的模型难以得到与其他模型竞争的对数似然性。此外我们还发现大部分的无损的代码长度用于描述图片中难以察觉到的细节。

二：背景

具体的公式推导问题可以见：
此处为语雀内容卡片，点击链接查看：https://www.yuque.com/chunfen-njv0x/ncn3vq/ehud1yd4iptzalde?view=doc_embed

最终得出的结论如下图所示

$$P(x_{t-1}|x_t) = N(\frac{\sqrt{\alpha_t}(1-\hat{\alpha}_{t-1})}{1-\hat{\alpha}_t}x_t + \frac{\sqrt{\hat{\alpha}_{t-1}}(1-\alpha_t)}{1-\hat{\alpha}_t} \times \frac{x_t - \sqrt{1-\hat{\alpha}_t} \times \epsilon}{\sqrt{\hat{\alpha}_t}}, (\frac{(1-\alpha_t)(1-\hat{\alpha}_{t-1})}{1-\hat{\alpha}_t}))$$

此时，只需要一个 ϵ 作为输入，就可以将 x_t 进行一次降噪变为 x_{t-1} ，而预测这个噪声，就是DDPM中的Unet所主要承担的工作。

三：扩散模型和去噪自编码器

DM类似一种受限制的隐变量模型，但其在实现的时候会有一系列可以自由调整的结构。比如需要确定一组 β_t 来实现前向过程（同样会用于反向过程）。为了指导这一选择，我们提出了一种在扩散模型和去噪分数检验之间的明确联系，这使得我们得到了简单有效的变量。最后，我们的模型通过一些简单的经验方式证明进行检验。

隐变量模型：将原有信息压缩为较低纬度的隐变量再进行处理，如VAE

3.1 前向过程和 L_t

我们放弃了将 β_t 作为可学习的参数，而是将其固定，这使得前向过程也是一个确定的过程，在训练中可以忽略。

3.2 反向过程和 $L_{1:T-1}$

首先我们讨论最后得到的公式部分：

$$P(x_{t-1}|x_t) = N(\frac{\sqrt{\alpha_t}(1-\hat{\alpha}_{t-1})}{1-\hat{\alpha}_t}x_t + \frac{\sqrt{\hat{\alpha}_{t-1}}(1-\alpha_t)}{1-\hat{\alpha}_t} \times \frac{x_t - \sqrt{1-\hat{\alpha}_t} \times \epsilon}{\sqrt{\hat{\alpha}_t}}, (\frac{(1-\alpha_t)(1-\hat{\alpha}_{t-1})}{1-\hat{\alpha}_t}))$$

可以看到最终的 $\sigma^2 = \frac{1-\hat{\alpha}_{t-1}}{1-\hat{\alpha}_t}\beta_t$,追加定义为 $\hat{\beta}_t$,但如果直接使用更加简化的 $\sigma^2 = \beta_t$,实际上会得到接近的结果，前者更适合 x_0 趋近集中于一点，而后者更适合 x_0 趋近于正态分布。

然后对于均值 μ 部分，我们提出了一种参数化的表达方式，最终使用的计算方式为：

$$\mu_t = \frac{1}{\sqrt{\alpha}_t}\left(x_t - \frac{\beta_t}{\sqrt{1-\hat{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$$

3.3 数据范围，反向过程的decoder，以及 L_0

我们假定图像的数据由0~255范围的整数组成，并将其线性压缩到[−1,1]范围。这可以保证神经网络在一个稳定的范围内进行运算（输入为高斯分布的 x_0 ）。为了处理个别的近似问题，在最后一步 $x_1 \to x_0$ 中，进行了一个边界情况的处理

decoder使用类似VAE中的decoder

3.4 简单的训练

为了进一步简化训练，所使用的损失函数为：

$$L_{simple} = E_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right]$$

其中， x_t 由 x_0 和固定的参数 β 求得

四：训练

设置的步数T=1000，参数 β 由 $\beta_1 = 10^{-4}$ 到 $\beta_T = 0.02$ 线性递增，这是因为数据范围限定在[−1,1]之间，所以需要取较小的值。并且保证 x_t 的信噪比足够低（在实验中与标准正态分布的loss约为 10^{-5} ）。

为了进行反向过程，使用Unet作为基础网络。参数在时间维度上共享，且使用16*16的自注意力机制。

4.1 生成质量

IS分数：用于衡量生成图片的多样性和质量。FID分数：用于评价生成图像与真实图像分布的相似程度

这两个评价分数都是早期用于评价GAN生成效果的检测方式

NNL则是负对数似然

Table 1: CIFAR10 results. NLL measured in bits/dim.

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	**10.06**	**2.67**	
Unconditional			
Diffusion (original) [53]			≤ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			**2.80**
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
NCSNv2 [56]		31.75	
NCSN [55]	8.87 ± 0.12	25.32	
SNGAN [39]	8.22 ± 0.05	21.7	
SNGAN-DDLS [4]	9.09 ± 0.10	15.42	
StyleGAN2 + ADA (v1) [29]	$\mathbf{9.74} \pm 0.05$	3.26	
Ours (L, fixed isotropic Σ)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	**3.17**	≤ 3.75 (3.72)

可以看到表中，DDPM的分数相对较高（尤其是使用简化过后的Loss函数）。

我们起初认为真实的变分界限上进行训练会较之简化物体上训练更好，但事实正好相反。

4.2 消融实验

Table 2: Unconditional CIFAR10 reverse process parameterization and training objective ablation. Blank entries were unstable to train and generated poor samples with out-of-range scores.

Objective	IS	FID
$\tilde{\mu}$ prediction (baseline)		
L, learned diagonal Σ	7.28 ± 0.10	23.69
L, fixed isotropic Σ	8.06 ± 0.09	13.22
$\|\tilde{\mu} - \tilde{\mu}_\theta\|^2$	–	–
ϵ prediction (ours)		
L, learned diagonal Σ	–	–
L, fixed isotropic Σ	7.67 ± 0.13	13.51
$\|\tilde{\epsilon} - \epsilon_\theta\|^2$ (L_{simple})	$\mathbf{9.46 \pm 0.11}$	**3.17**

进行了一些消融实验，总之还是简化的loss有着更好的效果。

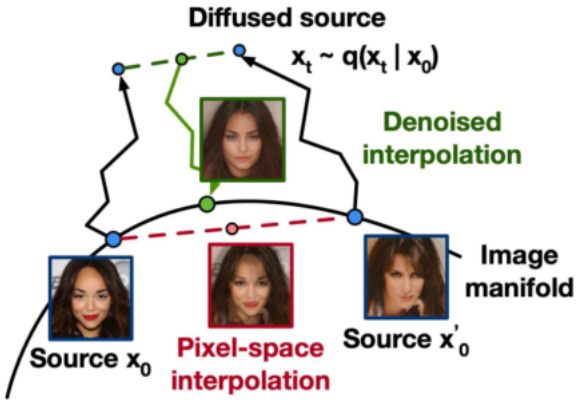
4.3 渐进编码

表一还展示了我们模型的码长（？），训练集和测试集的结果差距十分微小，可以认为没有发生过拟合现象。

渐进损失压缩 似乎都是信号学的内容，暂且搁置。

总之大概的意思是，在T较大时，主要改变整体轮廓，而在T较小时，主要修改细节。

4.4 插入图片



两张图片线性混合，出来的图片效果不佳，但如果放到隐空间上，再求取中心点，再映射回图像空间，就可以得到效果较好的中间图片。

五：相关工作

尽管DM和flow以及VAE模型类似，但其设计使得q（应该指的是前向过程）不需要任何参数，而且最初的输入值 x_t 几乎不含有任何的信息（VAE的生成也是类似思路）。反向过程的参数选择和物理中的扩散过程有所联系。

六：总结

本文使用扩散模型，实现了高质量的图像生成，为图像生成这一方向奠定了新的基础。