导向扩散模型

摘要

该论文的研究表明扩散模型在图像样本质量上能够超越当前最先进的生成模型。在无条件图像合成方面，他们通过一系列剖析找到了更好的架构，从而提高了性能。本文在原文所述的扩散模型上引入对步骤噪声方差的经过神经网络的弹性配置，使得扩散模型在较少步骤的场景下灵活性和性能得到提升；UNet 架构在Jolicoeur-Martineau等人的研究中被发现，相比之前用于去噪得分匹配的架构，这种架构的调整能够显著提升在更大、更多样化、更高分辨率数据集上的样本质量；研究了生成模型架构改变对结构带来的影响；讨论了对扩散模型进行条件训练的方法，通过利用类别标签来改进生成过程，结合分类引导和上采样扩散模型，FID会有更好的表现。

关键词： 扩散模型；梯度分类引导

1 引言

在过去的几年里，生成模型已经取得了生成人类自然语言、高质量合成图像以及多样化的人类语音和音乐等方面的能力。这些模型可以通过多种方式使用，例如从文本提示生成图像或学习有用的特征表示。尽管这些模型已经能够生成逼真的图像和声音，但在当前技术水平之上仍有很大的改进空间，更好的生成模型可能对图形设计、游戏、音乐制作等领域产生深远影响。

目前，生成对抗网络（GANs）在大多数图像生成任务上处于最先进水平，通过样本质量度量如FID、Inception Score和Precision等进行评估。然而，一些度量标准并不能完全捕捉到多样性，而且已经证明相对于基于似然的模型，GANs捕捉到的多样性较少。此外，GANs通常很难训练，需要精心选择的超参数和正则化项。

虽然GANs目前处于最先进水平，但它们的缺点使得它们难以扩展和应用到新的领域。因此，许多工作已经致力于使用基于似然的模型实现类似GAN样本质量的目标。虽然这些模型捕捉到更多的多样性，并且通常比GANs更容易扩展和训练，但在视觉样本质量方面仍然存在不足。此外，除了变分自编码器（VAEs）外，从这些模型中采样的速度在墙钟时间上比GANs慢。

扩散模型是一类基于似然的模型，最近已被证明能够生成高质量的图像，具有分布覆盖、固定训练目标和易于扩展等有益特性。这些模型通过逐渐从信号中去除噪声来生成样本，其训练目标可以表示为一个重新加权的变分下界。虽然这类模型在CIFAR-10等数据集上已经达到最先进水平，但在LSUN和ImageNet等难度较大的生成数据集上仍然落后于GANs。

作者假设扩散模型与GANs之间的差距至少有两个因素导致：首先，最近GAN文献中使用的模型架构已经得到深入探讨和改进；其次，GANs能够在多样性和保真度之间做出权衡，生成高质量样本但未覆盖整个分布。为了弥补这些不足，作者首先通过改进模型架构，然后通过设计一种交换多样性和保真度的方案，将这些优势引入到扩散模型中。

2 相关工作

在这一部分，我们将回顾与本文课题内容相关的先前工作，其中涵盖了得分基础生成模型、扩散模型及其改进、以及其他生成模型和技术的研究。

2.1 得分基础生成模型

得分基础生成模型最初由Song和Ermon 引入，作为使用数据分布梯度建模的一种方式，然后使用Langevin动力学进行采样。Ho等人发现这种方法与扩散模型之间存在联系，并通过利用这种联系实现了出色的样本质量。随着这一突破性工作的推出，许多后续工作取得了更加令人期待的结果，如Kong等人和Chen等人证明扩散模型在音频方面表现出色，Jolicoeur-Martineau等人发现类似GAN的设置可以改善这些模型的样本，Song等人]探索了利用随机微分方程技术改善得分基础模型获得的样本质量的方法，Nichol和Dhariwal提出了改进采样速度的方法，Nichol和Dhariwal以及Saharia等人展示了在困难的ImageNet生成任务上使用上采样扩散模型取得了令人期待的结果。

2.2 多样性与保真度权衡

在先前的扩散模型研究中，一个缺失的元素是在多样性和保真度之间进行权衡的方法。其他生成技术提供了自然的权衡杠杆。Brock等人为GAN引入了截断技巧，其中潜在向量是从截断正态分布中采样的。他们发现增加截断自然导致多样性的减少，但提高了保真度。最近，Razavi等人提出使用分类器拒绝采样，以过滤掉基于自回归似然模型的坏样本，并发现这种技术提高了FID。大多数基于似然的模型还允许低温采样，这提供了一种强调数据分布模式的自然方法。

2.3 基于分类器的生成模型控制

其他工作使用预训练分类器来控制生成模型。例如，一系列工作旨在使用预训练的CLIP模型优化GAN潜在空间以适应文本提示。与我们的工作更相似的是，Song等人使用分类器生成具有扩散模型的类条件CIFAR-10图像。有时，分类器可以作为独立的生成模型。例如，Santurkar等人证明了强大的图像分类器可以用作独立的生成模型，而Grathwohl等人训练了一个同时是分类器和能量基模型的模型。

3 本文方法

3.1 本文方法概述

在这一部分，我们将概述本文要复现的方法。本文的方法主要包括改进的扩散模型架构以及引入的分类器引导技术。首先，我们介绍了改进的扩散模型架构，基于UNet架构，其中包括残差层、下采样卷积、上采样卷积以及跳跃连接等组件。这一架构在之前的工作中已被证明在提高样本质量方面具有显著效果。

3.2 架构改进

我们通过进行多种架构削减实验，探索了几种可能的模型改进，以找到在扩散模型中提供最佳样本质量的架构。在这些实验中，我们尝试了以下架构改变：

- 增加深度与增加宽度相比，保持模型大小相对稳定。
- 增加注意头的数量。
- 在32×32、16×16和8×8的分辨率上使用注意力，而不仅仅在16×16上使用。
- 使用BigGAN的残差块进行上采样和下采样操作。
- 使用 √1/2来重新缩放残差连接，按照先前的研究。

通过比较不同架构变化的效果，我们观察到增加注意头的数量或减少每个头的通道数都可以改善FID。在选择最终的架构时，我们考虑到性能和时间的平衡，最终选择使用64个通道每个头作为默认配置。

Number of heads	Channels per head	FID
1		14.08
2		-0.50
4		-0.97
8		-1.17
	32	-1.36
	64	-1.03
	128	-1.08

Table 2: Ablation of various attention configurations. More heads or lower channels per heads both lead to improved FID.

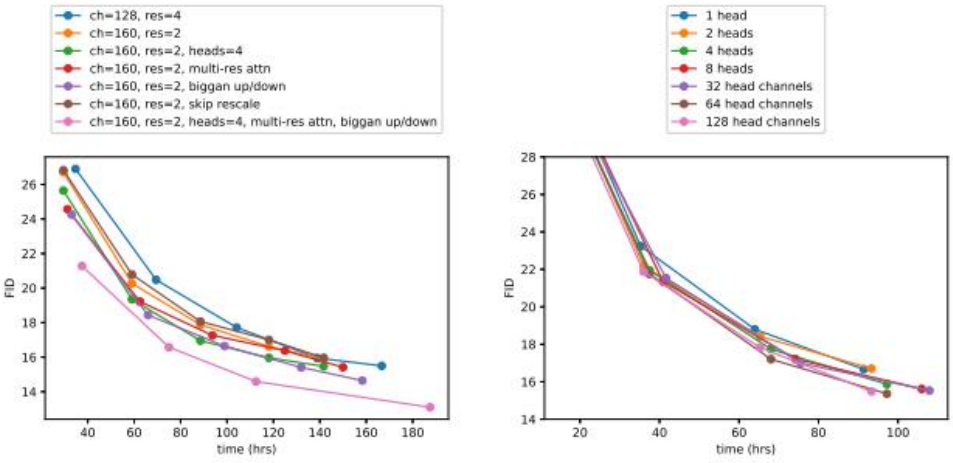

Figure 2: Ablation of various architecture changes, showing FID as a function of wall-clock time. FID evaluated over 10k samples instead of 50k for efficiency.

图1.论文原文方法效能比对图，分布改变注意力头数目，通道数后测量样本FID

3.3 自适应组规范

我们还尝试了一种自适应组规范（AdaGN）层，该层将时间步和类别嵌入引入到每个残差块中。这一层的定义为AdaGN(h, y) = ys GroupNorm(h)+yb，其中h是残差块第一卷积之后的中间激活，y = [ys, yb]是从时间步和类别嵌入的线性投影中获得的。实验证明，引入这一自适应组规范层可以提升FID。

在接下来的章节中，我们将使用这一改进的模型架构作为默认配置，其中包括可变宽度、每个分辨率2个残差块、64个通道每个头的多头注意力、32、16和8分辨率的注意力、BigGAN残差块用于上采样和下采样，以及自适应组规范用于将时间步和类别嵌入引入残差块。

3.4 分类器引导

除了使用精心设计的架构外，条件图像合成的GANs还广泛使用类别标签。为了进一步支持这一观察，本文探索了使用分类器引导扩散模型的方法。具体而言，我们使用分类器来指导扩散采样过程，从而生成特定类别的样本。在本节中，我们首先回顾了两种使用分类器导出条件采样过程的方法，然后描述了如何在实践中利用这些分类器来提高样本质量。

4 复现细节

4.1 与已有开源代码对比

在本部分中，由于个人可使用算力限制，对所需代码进行了整合调参，其中：

ADM (Ablated Diffusion Model):

对于64x64模型，--classifier_scale 1.0 表示使用完整的分类器引导，即ADM-G。

对于256x256模型，--classifier_scale 1.0 表示再次使用完整的分类器引导，即ADM-G。

对于256x256无条件模型，--classifier_scale 10.0 表示在无条件模式下使用强烈的分类器引导，相当于ADM-G。

对于512x512模型，--classifier_scale 4.0 表示使用相对较强的分类器引导，相当于ADM-G。

1. 出于算力限制，对大部分的样本生成采用64×64的分辨率。然后使用diffusion model进行unsampling，将低分辨率图像转换成较高分辨率图像。此处使用脚本，对两部操作整合。

```
import os
import subprocess

# Step 1: Generate Samples using classifier_sample.py
classifier_cmd = [
    "python",
    "classifier_sample.py",
    "--attention_resolutions", "32,16,8",
    "--class_cond", "True",
    "--diffusion_steps", "1000",
    "--dropout", "0.1",
    "--image_size", "64",
    "--learn_sigma", "True",
    "--noise_schedule", "cosine",
    "--num_channels", "192",
    "--num_head_channels", "64",
```

图2.脚本局部，统合二过程

2. 根据算力资源调整参数

根据原文理论，DDIM技术可以帮助生成确定性的图像，但是会极大的增加计算代价，由此此处不使用DDIM技术。

为了尽可能保持图像质量，对扩散模型的扩散步骤取尽量大值，并且根据原文测算结果，测算梯度引导因子的改变给样本质量带来的影响，在实验中，我们采用64×64，256×256，512×512的基于ImageNet的分类器进行实验，得到图像结果并计算FID值，把梯度引导的因子控制在1.0，5.0，10.0进行实验。同时，我们新增了一个参数‘_cond_type’来控制生成类型，这个类型对应ImageNet中的一个标签，我们建立一个映射表来对应类别。

由于Unsampling方法的开销过大，我们优先对该方法的效能参数进行尝试，在确定可用的参数后只在64×64的样本上进行测试。在给定的代码片段中，

classifier_sample.py 脚本用于从生成模型中进行采样。根据提供的标志和模型参数，该脚本可用于两种情况：基本扩散模型（ADM）和使用分类器引导的扩散模型（ADM-G）。

4.2 实验环境搭建

从git上克隆仓库并根据requirements文件安装即可。

5 实验结果分析

5.1 Unsampling函数参数调整

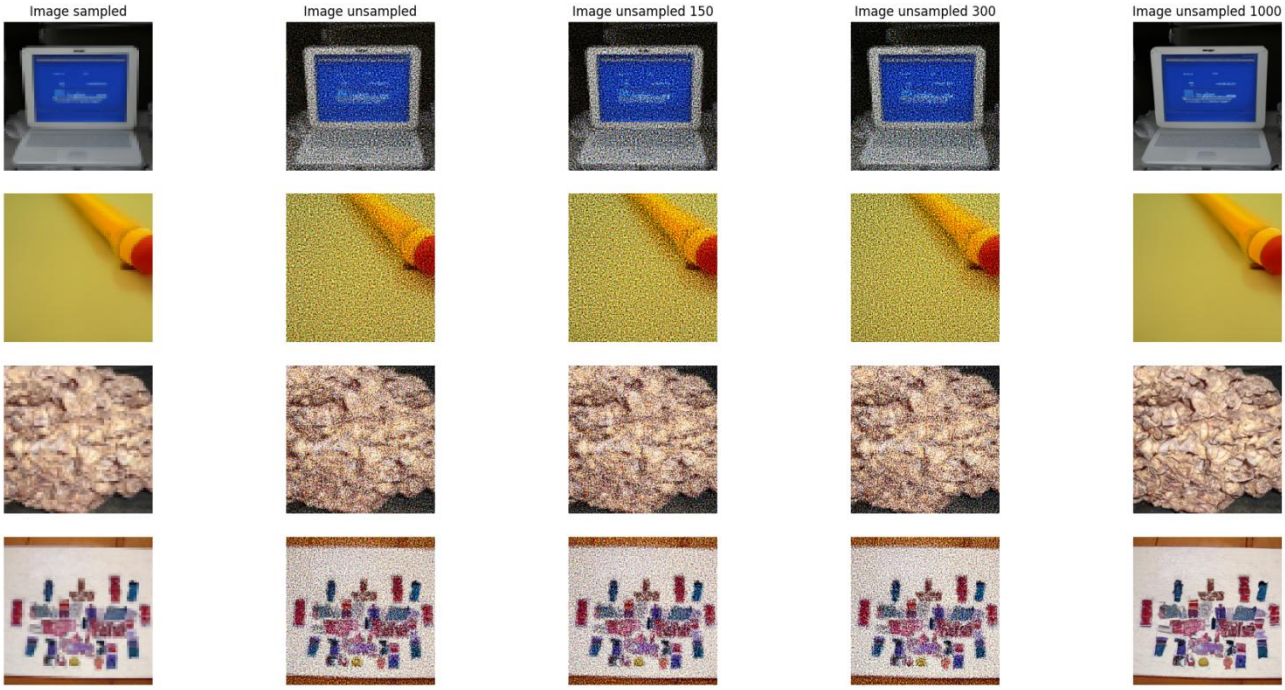

图3.上采样代码模块效果

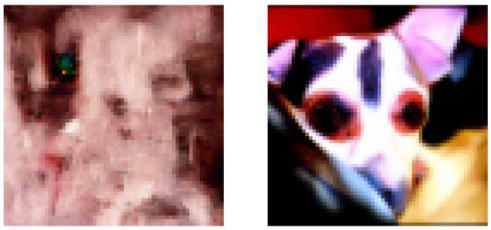

图4.吉娃娃图像生成样本64×64，其中分类梯度因子大小为30.0，由于过强的强调特征的引导，图像显然出现了过拟合。

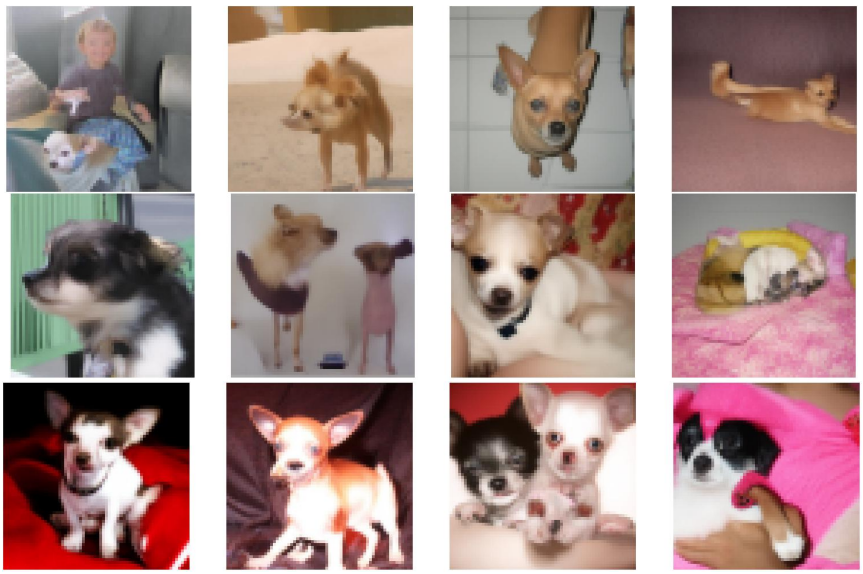

图5.吉娃娃图像生成样本64×64，其中分类梯度因子大小为1.0(上)5.0(中)10.0(下)，可以明显观察到随着分类器梯度引导强度的增加，生成的图像特征更加突出浮夸，与背景色对比更加强烈。

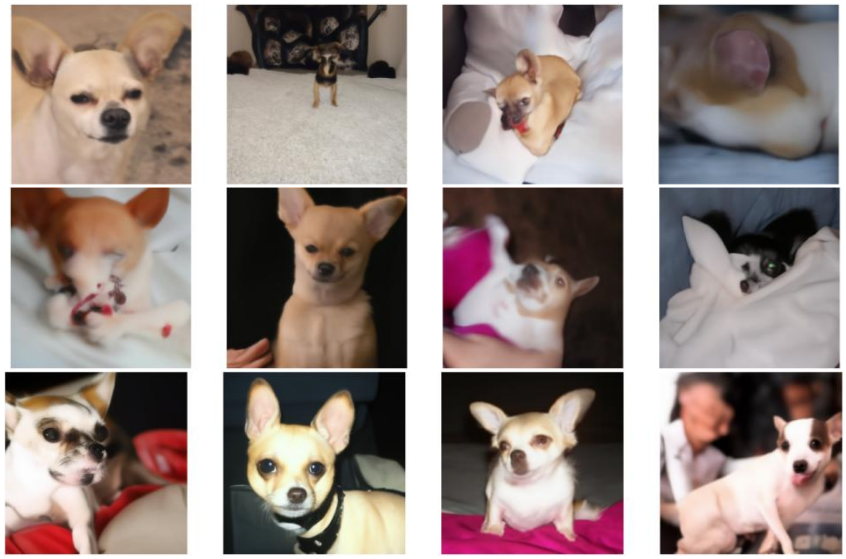

图6.吉娃娃图像生成样本256×256，其中分类梯度因子大小为1.0(上)5.0(中)10.0(下)

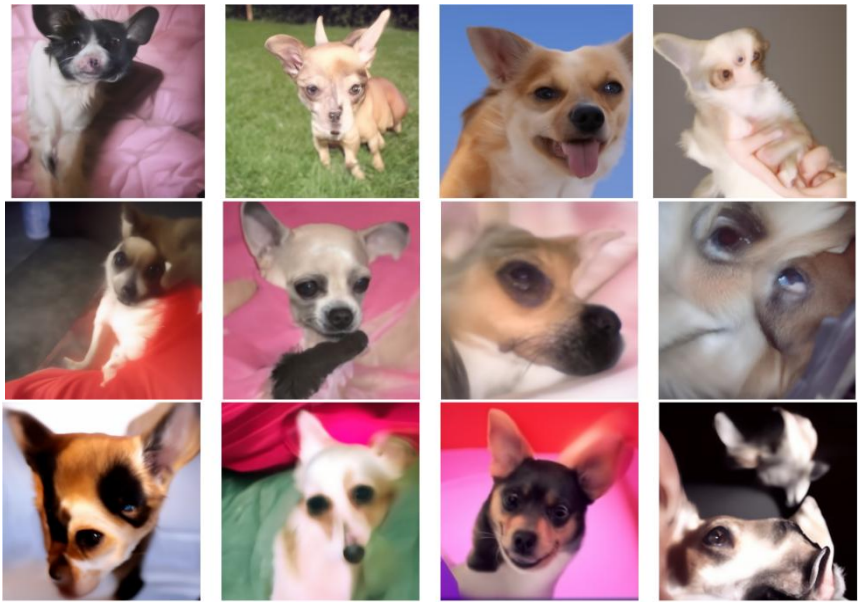

图7.吉娃娃图像生成样本512×512，其中分类梯度因子大小为1.0(上)，5.0(中)，10.0(下)

可以发现，随着生成图像的规模增加，模型对特征的控制能力在下降，在更大的图像规模里，特征分布的空间越大，所能做到的特征控制效果越差。我们绘制如下的平均FID折线图。发现得到了相同的FID趋势。

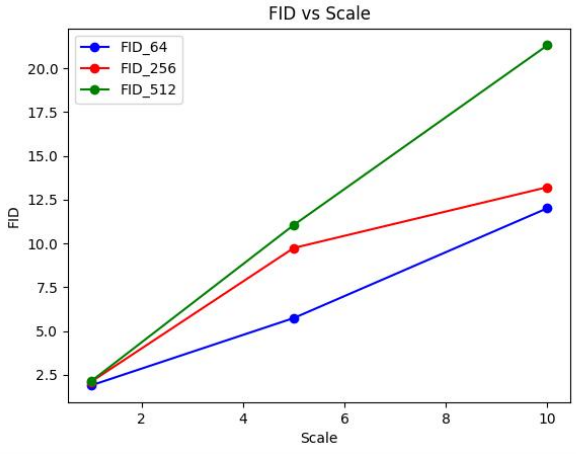

图8. 各图像规模下的平均FID值折线图统计

6 总结与展望

虽然我们认为扩散模型是生成建模的一个极具潜力的方向，但由于使用了多个去噪步骤（因此需要多个前向传递），它们在采样时仍然比GAN慢。在这个方向上的一项有希望的工作来自Luhman和Luhman，他们探索了一种将DDIM采样过程蒸馏成单步模型的方法。单步模型的样本虽然还不足以与GAN竞争，但比以前的基于单步似然的模型要好得多。未来在这个方向上的工作可能能够完全缩小扩散模型和GAN之间的采样速度差距，而不牺牲图像质量。

我们提出的分类器引导技术目前仅限于带标签的数据集，并且我们没有提供在无标签数据集上在多样性和保真度之间进行权衡的有效策略。将来，我们的方法可能通过对样本进行聚类以生成合成标签，或通过训练判别模型来预测样本是否来自真实数据分布或采样分布的方式，从而扩展到无标签数据。

分类器引导的有效性表明我们可以通过分类函数的梯度获得强大的生成模型。这可以用各种方式来条件化预训练模型，例如通过使用CLIP的嘈杂版本来将图像生成器与文本标题进行条件化，类似于最近使用文本提示来引导GAN的方法。这也暗示着将来可以利用大型无标签数据集来预训练强大的扩散模型，并通过使用具有良好特性的分类器进行改进。

参考文献

[1] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Research*, 32(11):1231– 1237, 2013.