

23 生成扩散模型漫谈（二十四）：少走捷径，更快到达

Apr By 苏剑林 | 2024-04-23 | 30138位读者 引用

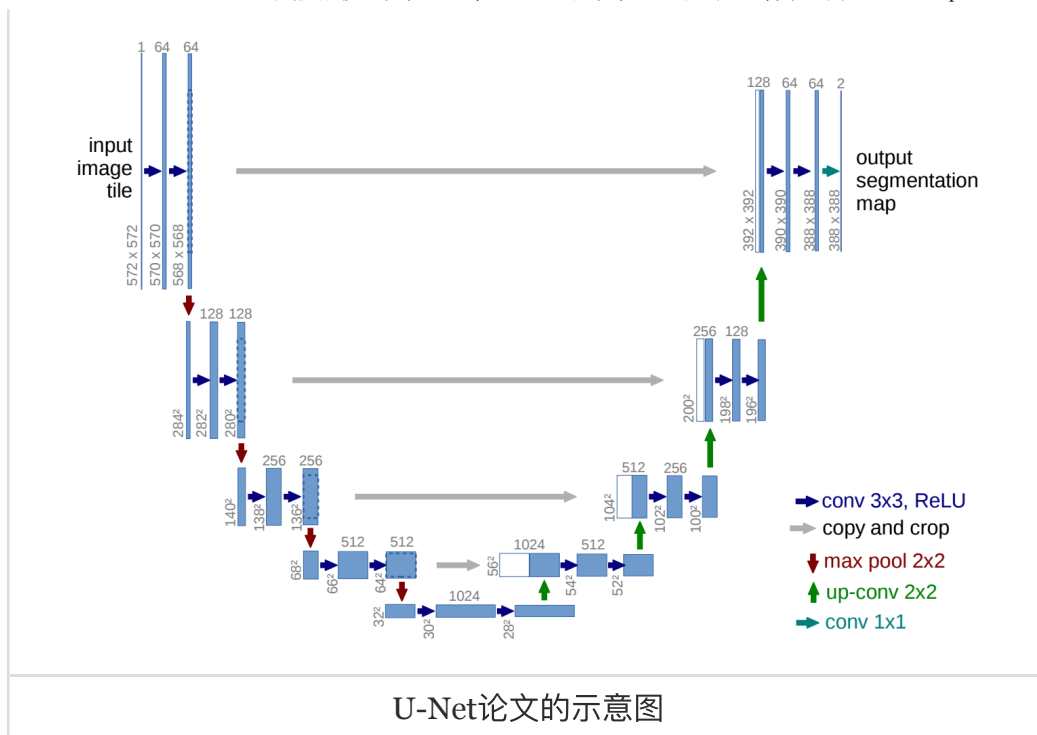
如何减少采样步数同时保证生成质量，是扩散模型应用层面的一个关键问题。其中，《生成扩散模型漫谈（四）：DDIM = 高观点DDPM》介绍的DDIM可谓是加速采样的第一次尝试。后来，《生成扩散模型漫谈（五）：一般框架之SDE篇》、《生成扩散模型漫谈（五）：一般框架之ODE篇》等所介绍的工作将扩散模型与SDE、ODE联系了起来，于是相应的数值积分技术也被直接用于扩散模型的采样加速，其中又以相对简单的ODE加速技术最为丰富，我们在《生成扩散模型漫谈（二十一）：中值定理加速ODE采样》也介绍过一例。

这篇文章我们介绍另一个特别简单有效的加速技巧——Skip Tuning，出自论文《The Surprising Effectiveness of Skip-Tuning in Diffusion Sampling》，准确来说它是配合已有的加速技巧使用，来一步提高采样质量，这就意味着在保持相同采样质量的情况下，它可以进一步压缩采样步数，从而实现加速。

模型回顾

一切都要从U-Net说起，这是当前扩散模型的主流架构，后来的U-Vit也保持了大致相同的形式，只不过将CNN-based的ResBlock换成了Attention-based。

U-Net出自论文《U-Net: Convolutional Networks for Biomedical Image Segmentation》，最早是为图像分割设计。它的特点是输入和输出的大小一致，这正好契合了扩散模型的建模需求，所以自然地被迁移到了扩散模型之中。形式上看，U-Net跟常规的AutoEncoder很相似，都是逐步下采样然后又逐步上采样，但它补充了额外的Skip Connection，来解决AutoEncoder的信息瓶颈：



不同的论文实现的U-Net在细节上可能不一样，但都有相同的Skip Connection，大致上就是第一层（block）的输出有一条“捷径”直达倒数第一层，第二层的输出有一条“捷径”直达倒数第二层，依此类推，这些“捷径”就是Skip Connection。如果没有Skip Connection，那么由于木桶效应，模型的信息流动就受限于分辨率最小的feature map，那么对于要用到完整信息的任务如重构、去噪等，就会得到模糊的结果。

除了避免信息瓶颈外，Skip Connection还起到了线性正则化的作用。很明显，如果靠近输出的层只使用Skip Connection作为输入，那么等价于后面的层都白加了，模型愈发接近一个浅层模型甚至线性模型。因此，Skip Connection的加入鼓励模型优先使用尽可能简单（即越接近线性）的预测逻辑，只有在必要情况下才使用更复杂的逻辑，这就是inductive bias之一。

寥寥几行

了解U-Net之后，Skip Tuning其实几句话就可以说完了。我们知道，扩散模型的采样是一个多步递归地从 \mathbf{x}_T 到 \mathbf{x}_0 的过程，这构成了 \mathbf{x}_T 到 \mathbf{x}_0 的一个复杂的非线性映射。出于实用的考虑，我们总希望减少使用采样步数，而不管具体用哪种加速技术，最终都在无形之中降低了整个采样映射的非线性能力。

很多算法如ReFlow的思路是通过调整noise schedule让采样过程走尽量“直”的路线，这样它采样函数本身就尽可能线性，从而减少加速技术带来的质量下降。而Skip Tuning则反过来想：**既然加速技术损失了非线性能力，我们可不可以从其他地方将它补回来？**答案就在Skip Connection上，刚才我们说了它的出现鼓励模型简化预测逻辑，如果Skip Connection越重，那么越接近一个简单的线性模型甚至恒等模型，那么反过来降低Skip Connection的权重，就可以增加模型的非线性能力。

当然，这只是增加模型非线性能力的一种方式，不能保证它增加的非线性能力正好是采样加速损失掉的非线性能力，而Skip Tuning的实验结果表明两者正好一定的等价性！所以顾名思义，对Skip Connection的权重做一定的Tuning，就可以进一步提高加速后的采样质量，或者在保持采样质量的前提下减少采样步数。Tuning的方式很简单，假设有 $k + 1$ 个Skip Connection，我们将最靠近输入层的Skip Connection乘以 ρ_{top} ，最远离输入层的Skip Connection乘以 ρ_{bottom} ，剩下的按照深度均匀变化就行，多数情况下我们都设 $\rho_{\text{top}} = 1$ ，所以基本上就只有 ρ_{bottom} 一个参数需要调。

Skip Tuning的实验效果也是相当不错的，下面摘录了两个表格，更多实验效果图可以自行阅读原论文。

Table 2. EDM Skip-Tuning with few-step sampling. ρ stands for the linear interpolation from the bottom to the top layer.

	SAMPLER	STEP	NFE	FID
EDM	HEUN	5	9	35.12
EDM (ρ :0.55 TO 1.0)	HEUN	5	9	18.71
EDM	UNI-PC	9	9	5.88
EDM (ρ :0.68 TO 1.0)	UNI-PC	9	9	2.92

Skip Tuning效果1

Table 3. Skip-Tuning in distilled EDM (CD: Consistency Distillation, DI: Diff-Instruct). *: results reported in original papers. †: In our reproduction, we replaced flash attention with standard attention for better compatibility.

	NFE	FID
CD EDM*	1	6.20
CD EDM†	1	6.85
CD EDM†(ρ : 0.91 TO 0.96)	1	5.56
DI EDM*	1	4.24
DI EDM†	1	4.16
DI EDM†(ρ_{TOP} : 0.98)	1	3.98

Skip Tuning效果2

个人思考 #

这应该是扩散系列最简单的一篇文章，没有冗长的篇幅，也没有复杂的公式，读者直接去读原论文肯定也容易搞懂，但笔者仍然愿意去向介绍一下它。跟上一篇文章《生成扩散模型漫谈（二十三）：信噪比与大图生成（下）》一样，它体现的是作者别出心裁的想象力和观察力，这是笔者自觉相当欠缺的。

跟Skip Tuning比较相关的一篇论文是《FreeU: Free Lunch in Diffusion U-Net》，它分析了U-Net的不同成分在扩散模型中的作用，发现Skip Connection主要负责添加高频细节，主干部分则主要负责去噪。这样一来我们似乎可以从另一个角度来理解Skip Tuning了：Skip Tuning主要实验的是ODE-based的扩散模型，这种扩散模型在缩减采样步数时往往噪点会增加，所以缩小Skip Connection，相对来说也就是加大了主干的权重，增强了去噪能力，属于“对症下药”。反过来，如果是SDE-based的扩散模型，可能

要减少Skip Connection的缩小比例，甚至可能要反过来增加Skip Connection的权重，因为此类扩散模型在缩减采样步数时往往会生成过度平滑的结果。

Skip Tuning调整的是Skip Connection，那么像DiT这种没有Skip Connection的是不是就没有机会应用呢？应该也不至于，DiT虽然没有Skip Connection，但还是有残差，Identical分支的设计本质上也是线性正则化的inductive bias，所以如果没有Skip Connection，调调残差可能也会有所收获。

文章总结

这篇文章介绍了一个能有效地提高扩散模型加速采样后的生成质量的技巧——降低U-Net的“捷径”（即Skip Connection）的权重。整个方法框架非常简单明快，直观易懂，值得学习一番。

转载到请包括本文地址：<https://spaces.ac.cn/archives/10077>

更详细的转载事宜请参考：《科学空间FAQ》

如果您需要引用本文，请参考：

苏剑林. (Apr. 23, 2024). 《生成扩散模型漫谈（二十四）：少走捷径，更快到达》[Blog post]. Retrieved from <https://spaces.ac.cn/archives/10077>

```
@online{kexuefm-10077,
  title={生成扩散模型漫谈（二十四）：少走捷径，更快到达},
  author={苏剑林},
  year={2024},
  month={Apr},
  url={\url{https://spaces.ac.cn/archives/10077}},
}
```