

Improved Denoising Diffusion Probabilistic Models 阅读笔记

作者/单位: Alexander Quinn Nichol, Prafulla Dhariwal (OpenAI)

发表: NeurIPS 2021 (改进去噪扩散概率模型)

关键词: DDPM、似然、噪声计划、采样加速、FID、模式覆盖

一、论文要解决什么问题

论文针对 DDPM (Denoising Diffusion Probabilistic Models, 去噪扩散概率模型) 在 2020 年原始工作 (Ho et al.) 中暴露的短板:

1. 对数似然不具竞争力: DDPM 尚未被证明能获得与其他基于似然的模型 (自回归模型、VAE 等) 相竞争的对数似然分数;
2. 模式覆盖能力存疑: 不清楚 DDPM 能否捕获分布的所有模式 (mode coverage);
3. 可扩展性未知: 不清楚 DDPM 如何扩展到更高多样性的数据集, 如 ImageNet;
4. 生成样本需数千步扩散过程 (文中模型用 4000 步训练), 单张采样在 GPU 上要数分钟, 采样速度极慢, 实用性差。

核心目标: 用少量修改, 让 DDPM 既能更快采样, 又能取得更好的对数似然, 同时在样本质量上逼近 GAN。

二、研究过程与主要方法

论文的核心方法是四类关键修改:

1. 学习反向过程的方差 Σ_θ (核心贡献): 原始 DDPM 把反向过程方差固定为常数, 论文改为学习方差 Σ_θ 。用混合参数化: 将 Σ_θ 参数化为上下界之间的插值 (在 β_t 与 $\tilde{\beta}_t$ 之间取 log 域插值), 并用 L_{hybrid} 目标训练。 $L_{\text{hybrid}} = L_{\text{simple}} + \lambda \cdot L_{\text{vlb}}$: 在原本的简单目标 (L_{simple}) 上加上带权重的变分下界 (L_{vlb}) 项, 让模型同时优化似然与样本质量。
2. 改进噪声计划 (cosine schedule): 原始线性噪声计划在扩散后期 α_t 下降过快, 导致训练早期噪声过大、浪费容量。论文改用 cosine 噪声计划 (余弦型 schedule), 使噪声在扩散过程中更平滑地衰减, 并专门调整了 dropout (cosine 计划下 dropout=0.3 最佳, 线性计划下 0.1 最佳)。

3. 加速采样（striding，跨步采样）：模型用 4000 步训练，但采样时可只用 25/50/100/200... 步：从训练序列 (1,2,...,T) 中取 K 个均匀间隔的子序列，重算对应采样方差。关键发现：用 L_{hybrid} 训练、学习过 $\Sigma\theta$ 的模型，即使把采样步数从 4000 降到 100 步，仍能保持接近最优的 FID——而用 L_{simple+} 固定方差的模型，降步数后样本质量大幅下降。
4. 消融与验证实验：在 CIFAR-10 / ImageNet 64×64 上评估对数似然（bits/dim）与 FID。与 GAN 对比（用精确率 precision 与召回率 recall 两个指标衡量模式覆盖）：训练了类别条件的扩散模型（small 100M 参数 / large 270M 参数）与 BigGAN-deep。

三、主要实验结果与结论

① 对数似然接近其他似然模型（表 3，bits/dim）

数据集	Glow	Sparse Transformer	NVAE	原始 DDPM	Improved DDPM
ImageNet 64	3.81	3.44	—	3.77	3.53
CIFAR-10	3.35	2.80	2.91	3.70	2.94

→ 在 CIFAR-10 上从 3.70 提升到 2.94，与最好的卷积模型竞争力相当；在 ImageNet 64 上 3.53，接近 transformer 类模型。

② 模式覆盖优于 GAN（表 4，ImageNet 64 类别条件）

模型	FID	Precision	Recall
BigGAN-deep	4.06	0.86	0.59
Improved Diffusion (small)	6.92	0.77	0.72
Improved Diffusion (large)	2.92	0.82	0.71

→ 大型扩散模型 FID 2.92 追平并超过 BigGAN (4.06)，recall (0.71 vs 0.59) 明显更高——说明扩散模型在覆盖分布全部模式上显著优于 GAN。

③ 采样速度大幅提升：学习方差的 L_{hybrid} 模型用 100 个采样步即可达到接近最优 FID（从数分钟降到数秒）。与 DDIM 对比：步数 <50 时 DDIM 更好，≥50 步时本方法更好；且本方法直接从祖先过程（ancestral process）采样，无需额外超参数。

④ 模型规模扩展（第 6 节，初步结论）：在 ImageNet 64 上训练不同规模（首层通道 64/96/128/192）的模型，发现 FID 随训练算力按幂律（power law）近似线性下降（log-log 图上接近直线）；NLL 的幂律拟合不那么干净。

最终结论：经过上述少量修改，DDPM 能更快采样、取得更好的对数似然（更接近其他似然

模型)，并且采样质量可匹配 GAN、模式覆盖远超 GAN，同时其训练目标稳定、可随算力平滑扩展——使 DDPM 成为既有好似然、又高质量、又能较快采样的有吸引力的生成模型方向。

四、相关工作与并行工作

- 音频生成：WaveGrad (Chen et al.)、DiffWave (Kong et al.) 用 DDPM 生成 mel 谱图条件的高保真音频。
- 加速采样并行工作：Song et al. (2020a) 的 DDIM (确定性隐式模型)；Song et al. (2020b) 将扩散建模为连续 SDE、用 ODE 求解器加速 (但需 Langevin 校正步并手调信噪比)。
- 对抗式得分匹配：Jolicœur-Martineau et al. 发现更好网络结构可移除对抗目标的必要性。

一句话总结：这篇论文用「学习反向过程方差 + cosine 噪声计划 + Lhybrid 混合目标」三处修改，让 DDPM 的似然逼近自回归/VAE、采样从数千步降到百步内、样本质量追平 GAN 且模式覆盖更全面，一举把扩散模型从“能生成但慢而弱”推到可与 GAN 正面竞争的位置。