

扩散模型的量化

李无邪

南开大学统计与数据科学学院

2025 年 4 月

目录

- ① 扩散模型
- ② PTQ4DM
- ③ Q-Diffusion

本文档是自己学习过程中的笔记，仅用于记录知识、学习理解论文

本周研读论文

- ① 2020-NeurIPS Denoising Diffusion Probabilistic Models (去噪概率扩散模型)
- ② 2023-CVPR Post-training Quantization on Diffusion Models (扩散模型的训练后量化)
- ③ 2023-ICCV Q-Diffusion: Quantized Diffusion Models (量化扩散模型)

方便起见，以上三篇论文后续分别简称为 DDPM、PTQ4DM 和 QDM

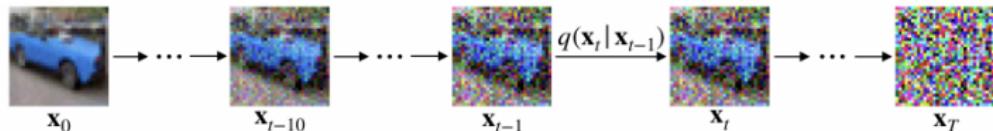
扩散模型

扩散模型（Diffusion Models）是一种基于概率的图像生成模型，此领域的爆火源于Jonathan Ho在 2020年提出的“去噪扩散概率模型”（DDPM），当今的AI生图算法普遍基于扩散模型。

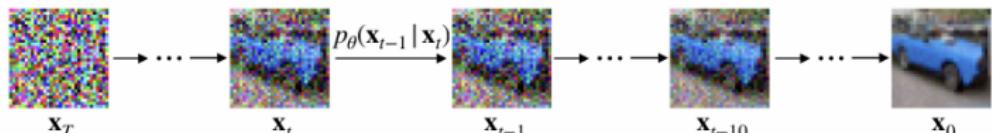
该领域的一些著名的工作包括：DDPM(2020), DDIM(2021), Score-Based Models(2021)

扩散模型的训练阶段分为前向和反向两过程，其中前向过程向图像中注入高斯噪声，使图像最终收敛于 $\mathcal{N}(0, I)$ ，反向过程则根据带噪声的图像和时间步信息进行去噪。接下来我们以 DDPM 模型为例进行介绍。

去噪概率扩散模型：前向



(a) Forward diffusion process



(b) Reverse diffusion process

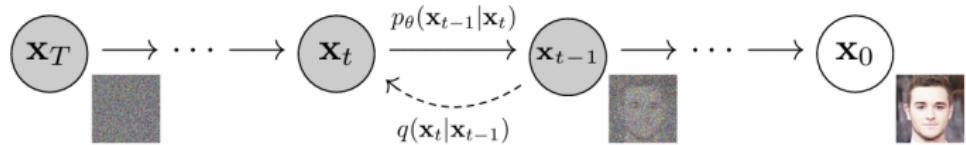
我们假设原图像服从某一复杂分布 $x_0 \sim p(x_0)$, 并且图像的加噪过程是服从马尔可夫性质的。则 x_t 的分布仅依赖于 x_{t-1} , 那么加入高斯噪声可以表示为如下的转移概率

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (1)$$

利用重参数化技巧, 得

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1 - \alpha_t}\epsilon_t \text{ where } \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2)$$

去噪概率扩散模型：前向



递推(2)式，可得

前向扩散公式

$$\mathbf{x}_T = \sqrt{\prod_t^T \alpha_t} \mathbf{x}_0 + \sqrt{1 - \prod_t^T \alpha_t} \epsilon_0 = \sqrt{\bar{\alpha}} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_0 \quad (3)$$

扩散过程的收敛性

当 T 充分大时， $\prod_t^T \alpha_t \rightarrow 0$

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (4)$$

去噪概率扩散模型：反向



反向过程从纯噪声图出发，进行逐步的去噪。可以理解为，反向过程在模拟前向过程的“倒放”，这个倒放的转移分布也是正态的，并且具有马尔可夫性。反向过程需要构建神经网络并进行训练，逐步地恢复原始的高清图像。

为什么要先加噪再去噪呢？

- ① 原始图像的分布是十分复杂的，而DDPM采用的概率建模将整个过程视为数步正态分布的转移，正态分布是易于研究的。
- ② 在宋飏的论文中，DDPM被解释为一个朗之万动力学视角下的随机微分方程，它相当于一个势能场，将正态分布中的噪声点引向“高概率密度区域”。

去噪概率扩散模型：反向

反向过程的转移概率如下

反向过程的转移分布

$$q_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (5)$$

根据贝叶斯统计，我们的模型实质上在构建一个未知真实图片的后验分布，而我们的目的是让其逼近真实情况下的后验分布 $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$ 。可以证明 q_θ 和 q 均为正态分布，故我们只需要固定方差，比较两者均值的差异即可。

采取合理的参数化方法，可以证明，均值的差异等价于预测噪声与真实噪声的差异。

去噪概率扩散模型：训练目标

可以导出下面的训练目标

训练目标

可以证明前向和反向转移分布的KL散度 $KL(p, q_\theta)$ 等价于

$$\mathbb{E}_{\mathbf{x}_0, \epsilon, t \sim \mathcal{U}[0, T]} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_0, t)\|^2 \right] \quad (6)$$

其中， $\epsilon_\theta(\sqrt{\bar{\alpha}}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_0, t)$ 表示噪声预测器对正向第 t 步的噪声预测值。

去噪概率扩散模型：采样

在实际实现中，我们采用如下的方程进行去噪

反向采样算法

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z} \quad (7)$$

其中， $\mathbf{z} \sim \mathcal{N}(0, I)$ ，上式为一个随机差分方程。

以上就是扩散模型的理论基础，接下来我们将对DDPM的神经网络架构进行介绍，引出影响其运算效率的问题所在。

U-Net in DDPM

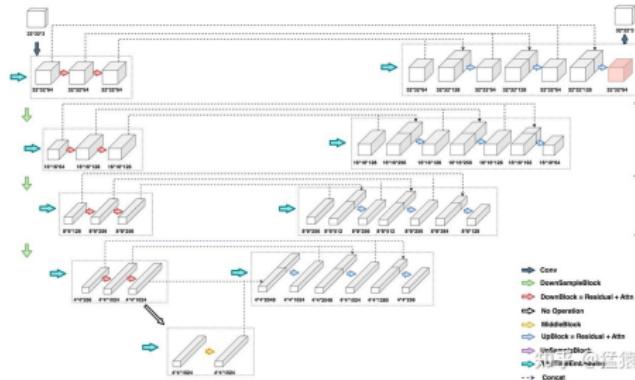


图: DDPM的U-Net结构

- ① 输入噪声图像 x_t 并嵌入时间步 t
- ② 下采样: 卷积和池化, 压缩图像特征
- ③ 上采样: 反卷积和上采样, 恢复图像特征
- ④ 跳跃连接: 将高分辨率图像的信息保留并传给对应的解码器
- ⑤ 输出: 预测噪声 $\epsilon_\theta(x_t, t)$

DDPM的效率问题

- ① 严格依赖于马尔可夫性，生成过程无法跳步处理
- ② 需要大量的时间步才能还原图像
- ③ U-Net 的计算量很大，每一步都需要进行大量的计算
- ④ 难以快速生成高分辨率图像

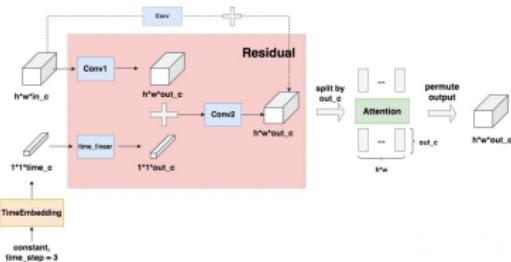
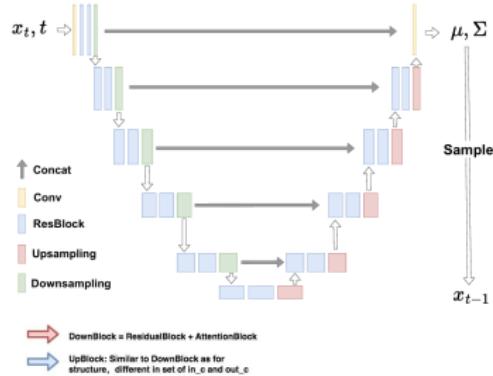
如何对扩散模型进行 PTQ 量化

对于一个普通的神经网络，我们可以研究其激活分布来决定量化的参数，并直接从其训练集中收集校准数据集。而扩散模型不同，整个模型虽只有一个参数化的神经网络——即 U-Net，但是由于时间步的嵌入，每一步输入图像的噪声含量差距，以及 U-Net 内部的分块结构，导致该神经网络在不同时间步和不同层的激活分布差距很大。

我们将以上问题归结为如下三个方面的问题：

- ① 选择合适的量化算子
- ② 分析激活分布的差异，采取恰当的校准数据集收集方法
- ③ 采取合适的校准方法

量化算子的选择



模型的架构中大量使用了残差块，而残差块中又进行了多次卷积操作以及注意力机制（非必需），显然，这些计算密集的层应当被量化。

图: ResBlock(残差块)的内部结构

量化算子的选择

能否对模型的输出采用量化呢？

Table 1. Exploration on operation selection for 8-bit quantization. The diffusion model is for unconditional ImageNet 64x64 image generation with a cosine noise schedule. DDIM (250 timesteps) is used to generate 10K images. IS is the inception score.

	IS	FID	sFID
FP	14.88	21.63	17.66
quantize μ	15.51	21.38	17.41
quantize Σ	15.47	21.96	17.62
quantize x_{t-1}	15.26	21.94	17.67
quantize $\mu+\Sigma+x_{t-1}$	14.94	21.99	17.84

图：不同量化操作的效果评估，其中 IS 为正向指标，其余为负向

根据 PTQ4DM 的实验结果，对 μ, Σ, x_{t-1} 三者之一采取量化的效果较好，但同时量化的效果较差。因此 PTQ4DM 选择量化三者之一

量化误差分析

QDM 中提到，由于扩散模型的多时间步去噪架构，量化误差会随着时间步累积，因此在量化时需要确保每层的量化误差都尽可能小。

校准数据集的搜集：激活分布的分析

激活分布

神经网络的隐藏层中常常设有激活函数，张量经过隐藏层的计算后，输出的数据分布被称为激活分布。

在量化多层神经网络时，为了找到合适的量化区间和缩放因子，需要对其激活分布进行研究。

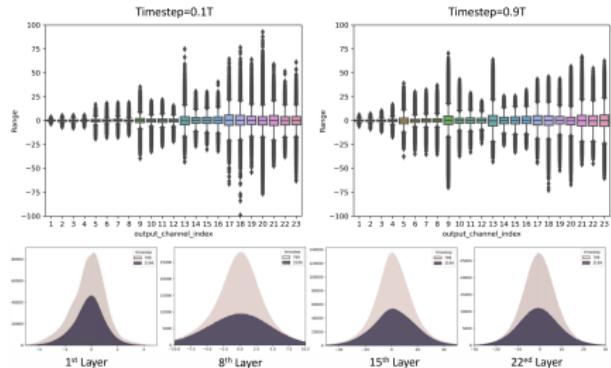
对于一般的神经网络，其通常只接收类似的输入，产生类似的激活分布。而在扩散模型中，模型接收不同信噪比的输入和时间步，每一步的激活分布都是截然不同的。传统 PTQ 的方法不再适用。

问题提出

- ① 选择什么样本作为校准数据集？
- ② 选择哪一时间步的样本更能使得全局量化损失最小？

为解决该问题，PTQ4DM 的作者进行了三个实验。

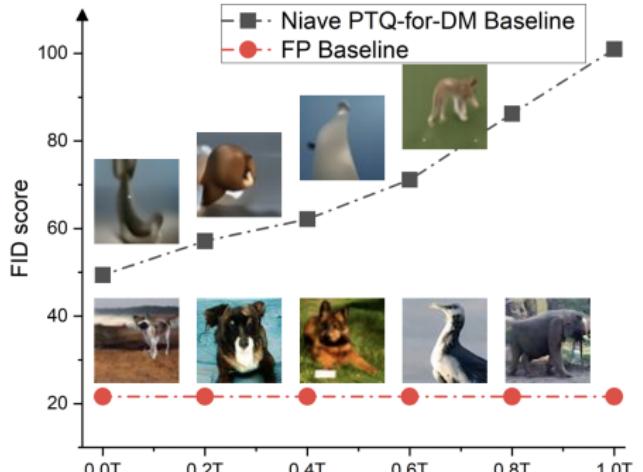
PTQ4DM 的三个重要实验



图：U-Net 不同时间步下各通道的分布箱线图&不同时间步下各层的分布差异

	IS ↑	FID ↓	sFID ↓
Noise Samples	13.92	33.15	20.38
Image Samples	6.90	128.63	90.04
Training-mimic	12.91	34.55	25.18

图：去噪生成样本、原图像、加噪生成样本三者分别作为校准集的效果评估



图：不同时间步的样本作为校准集的效果评估：信噪比越高效果越好

结论

- ① 时间步、U-Net 层间的分布差异巨大，不能简单地采取某一固定时间步生成的样本来校准
- ② 应当采用反向去噪生成的样本作为量化校准集
- ③ 应当使采用的样本时间步更小，更接近真实分布

根据以上结论，作者提出了“正态分布时间步长校准算法”(Normally Distributed Time-step Calibration)

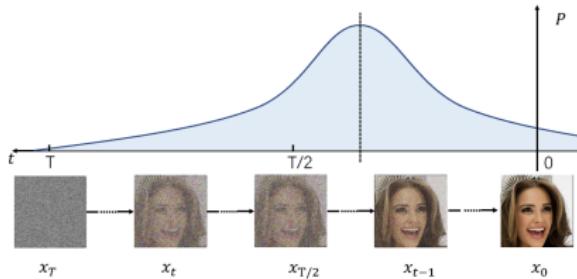


Figure 5. A general illustration of sampling time-steps following a distribution over the range of the denoising time-step.

Algorithm

- ① 从正态分布 $\mathcal{N}(\mu < \frac{T}{2}, \frac{T}{2})$ 中采样 $t_{1:N}$
- ② 进行取整、截断操作，使得 $t_{1:N}$ 是时间步
- ③ 对 $\forall t_i \in \{t_{1:N}\}$ 使用全精度网络求得模型在该步的输出 x_{t_i}
- ④ 得到校准集 $\mathcal{C} = \{x_{t_i}\}_{1:N}$

该算法是本篇论文的核心内容，它以一种十分简洁的方式解决了量化数据集的选取问题，同时也是扩散模型量化理论的开山之作：首次将 PTQ 应用于扩散模型，压缩至 8 位。
接下来的 QDM 理论，将从更深入的结构入手研究模型量化的挑战。

Q-Diffusion: Layer角度的思考

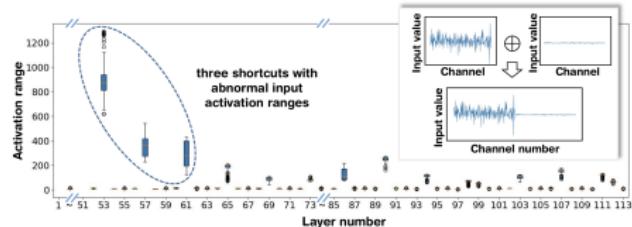


图: 全时间步下各层的激活分布, 可以观察到三个异常值

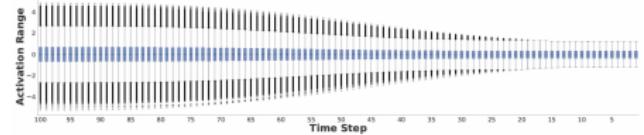


图: 各时间步的激活分布, 可见相邻步的差异并不大

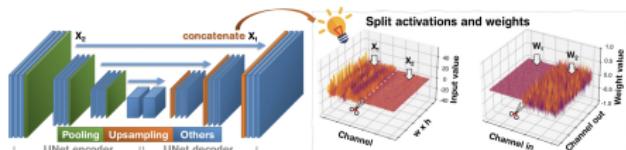


图: 跳跃连接造成某些层拥有特殊的激活分布, 若统一量化, 会造成严重的误差

观察 U-Net 架构

U-Net的每一层并非独立存在的, 而是按特定功能被分为了多个块(重建块), 因此在校准时, 传统 PTQ 的逐层校准往往会忽略块结构, 无法处理层间依赖性和泛化问题。

QDM：问题提出

针对以上的研究，我们需要解决如下的问题：

- ① 开发一种合理的采样算法，获得校准数据集
- ② 采用更合理的校准算法，解决逐层校准带来的误差
- ③ 解决跳跃连接带来的问题

对于问题2，我们可以将 U-Net 中的某些重要结构作为一个整体校准，其他则使用逐层校准

按块校准

设重建误差函数为 $L(X_{cons}, X_0)$ ，按块校准的意思是，不再关注各独立层的重建误差 $L(X_{i_{cons}}, X_{i_0})$ ，而是直接计算块的重建误差。

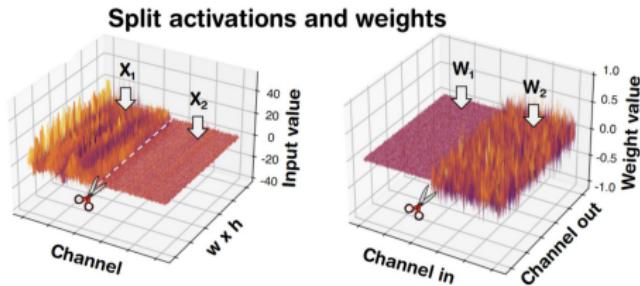
Step-Aware 采样算法

由于相近时间步的激活分布差异不大，作者引入了一种等间距采样的算法，既能减少算法的复杂度，又能对不同时间步的分布进行完整的覆盖。

Algorithm

- ① 设 c 为采样间隔，对时间步 $t \% c = 0$ ，从全精度去噪网络中采样 n 次，得到 $\{\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)} \dots \mathbf{x}_t^{(n)}\}$ ，添加到校准集 \mathcal{D} 中
- ② 对于分块 $i \in \{1, \dots, N\}$ ：使用 \mathcal{D} 和全精度网络更新量化网络中第 i 块的权重量化器
- ③ (若要进行激活函数的量化)对于分块 $i \in \{1, \dots, N\}$ ：使用 \mathcal{D} 和全精度网络更新量化网络中第 i 块的激活量化器步长

连接层拆分量化



深入分析问题3：

U-Net 的跳跃连接将深层特征通道 (X_1) 和浅层特征通道 (X_2) 相连接。由于深层和浅层特征通道的激活值范围差异较大，这会导致相应通道中的权重分布也表现出双峰特性。而对于分布极度不均匀的权重进行量化，极有可能导致较大误差。因此，作者认为此处不再适合 PTQ 量化

Shortcut-splitting Method

为解决连接导致的量化问题，对该特殊层采用先量化再连接的方法：

$$\mathcal{Q}_X(X) = \mathcal{Q}_{X_1}(X_1) \oplus \mathcal{Q}_{X_2}(X_2)$$

$$\mathcal{Q}_W(W) = \mathcal{Q}_{W_1}(W_1) \oplus \mathcal{Q}_{W_2}(W_2)$$

分别量化两者的激活分布和权重，再搭建连接，最终分别计算全连接层的量化结果。

总结

- ① PTQ4M 的工作首次将训练后量化技术用于扩散模型，探索出了量化扩散模型的着手方向，并提出了简便的 NDTC 算法。
- ② QDM 的工作深入到了 U-Net 的架构上，提出了分块校准、先量化后连接、等间距步长采样等方法，解决了量化扩散模型中的许多细节问题。