arXiv:2601.09451v1 [cs.LG] 14 Jan 2026

最新研究成果：Quamba-SE：状态空间模型中激活值的软边缘量化器

Yizhi Chen {yizhic@kth.se} 和 Ahmed Hemani {hemani@kth.se}电子与嵌入式系统系，瑞典皇家理工学院，瑞典斯德哥尔摩

摘要—我们提出 Quamba-SE，一种用于状态空间模型（SSM）激活值量化的软边缘量化器。与现有方法不同，Quamba-SE 使用标准 INT8 运算，并采用三种自适应尺度：对小值使用高精度尺度，对正常值使用标准缩放因子，对离群值使用低精度尺度。这样既能保留离群值信息，避免硬裁剪，又能维持其他值的精度。我们在 Mamba-130M 上使用 6 个零样本基准测试进行评估。结果表明，Quamba-SE 始终优于 Quamba，在单项基准测试上最高提升 +2.68%，在 6 个数据集的平均准确率上最高提升 +0.83%。

索引词—量化，状态空间模型，Quamba

I. 引言

大语言模型（LLM）正在迅速发展，其中以 GPT 等 Transformer 架构为主。近期，Mamba [1], 等状态空间模型（SSM）已成为重要的替代方案。Chiang 等 [2] 表明，Mamba 系列的表现优于同等规模的 Pythia [3], Transformer。

量化对于提升速度、减少存储需求至关重要。对于大语言模型，训练后量化（PTQ） [4]得到了广泛应用，因为量化感知训练（QAT） [5]需要大量计算训练成本。然而，相关研究 [2], [4], [6] 揭示了训练后量化面临的独特挑战：状态空间模型的激活值包含大量离群值。尽管离群值仅占一小部分，但将其纳入会损害正常值的量化精度。

现有方法包括哈达玛变换 [2], [6], [7],、百分位裁剪 [2], 和分组缩放 [6]。然而，这些方法依赖 CUDA，且缺乏硬件层面的优化。这些方法使用标准 INT8 运算，所有数据仅使用一个尺度，并会对离群值进行硬裁剪。

这也是一个非常新的问题。Quamba [2]和 Quamba2[6] 分别于 2025 年 4 月和 2025 年 7 月在会议上发表。我们提出了 Quamba-SE，这是一种在硬件层面运行的软边缘量化器，可为数值提供软边缘，而非进行硬裁剪。虽然该量化器是在 Quamba 上评估的，但其设计适用于任何状态空间模型。

II. 背景与相关工作

A. 背景

我们在**图1**中展示了 Quamba 和 Quamba-SE 的数据流。一些组件可以轻松量化，例如 INT8 线性投影。状态空间模型的输入和输出是量化面临的主要挑战。对于输出 Y，我们使用哈达玛变换来平滑离群值分布。

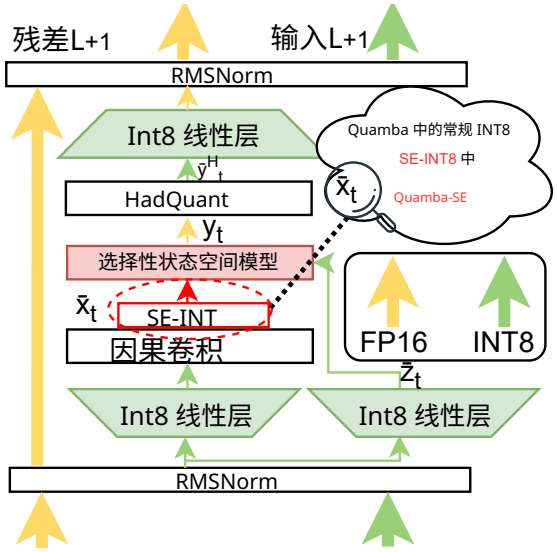

图 1. Quamba-SE：数据流与数据精度

B. 相关工作

我们重点关注状态空间模型的输入 X_t （在**图1**中突出显示）。Mamba-PTQ [4] 指出，状态空间模型量化的难点源于激活异常值。Quamba [2]通过校准（包括百分位裁剪）排除离群值。在推理过程中，使用存储的尺度执行常规 INT8 量化，其性能优于 Mamba-PTQ。Quamba2 [6]在 Quamba 的基础上增加了分组缩放因子，提高了 Mamba 2 [9]模型的准确率，但对 Mamba1 模型的提升较小。此外，受 CUDA 和 PyTorch 框架限制，Quamba 和 Quamba2 专注于为标准 INT8 运算寻找更优的尺度。

III. 软边缘量化方法

与 Quamba 和 Quamba2 相同，我们在线下校准并保存尺度。不同之处在于推理阶段，如**图 1** 中红色部分所示：Quamba 使用单一尺度并裁剪离群值，而我们在推理过程中自适应地应用三种尺度。

除对正常值使用相同的存储尺度外，我们在**图 2**中引入了另外两种尺度：用于小值的高精度缩放因子（scale/4）、用于中等值的标准缩放因子，以及用于离群值的低精度缩放因子（scale×4）。推理过程中，首先对数值进行分类：正常值使用标准 INT8，且禁用软边缘（SE）标识符。对于小于阈值 L 或大于阈值 H的值，则启用 SE 标识符。为避免额外存储，我们使用 INT8 数据的第二位来区分小值和大值，剩余

表 I MAMBA-130M 上零样本准确率（%）对比。量化模型中最佳结果以**粗体**标示，次佳结果以下划线标示。 †选定基准测试的平均值：LAMBADA、ARC-C、WinoGrande。

设置	方法	LAMBADA	HellaSwag	PIQA	ARC-E	ARC-C	WinoGrande	**Avg.**	Δ	平均值†	Δ†
已报告 [2]	FP16	44.20	35.30	64.50	48.00	24.30	51.90	44.70	–	40.13	–
	Quamba	40.60	35.00	63.00	46.50	23.00	53.10	43.50	–	38.90	–
FP16	FP16	44.24	35.22	64.57	48.05	24.35	52.00	44.74	–	40.20	–
校准（99.99%）	Quamba	42.44	35.30	63.04	45.92	23.98	52.47	43.86	–	39.63	–
	Quamba-SE	43.42	34.90	62.84	**46.80**	25.05	51.51	**44.08**	+0.22	39.99	+0.36
校准 (99.999%)	Quamba	42.23	34.69	62.16	42.23	24.18	52.45	42.99	–	39.62	–
	Quamba-SE	**44.56**	35.03	**63.06**	42.44	24.37	52.03	43.58	+0.59	**40.32**	+0.70
官方权重 [8]	Quamba	39.06	35.17	61.94	45.13	24.90	50.73	42.82	–	38.23	–
	Quamba-SE	40.40	**35.54**	62.08	45.13	**25.32**	**53.41**	43.65	+0.83	39.71	+1.48

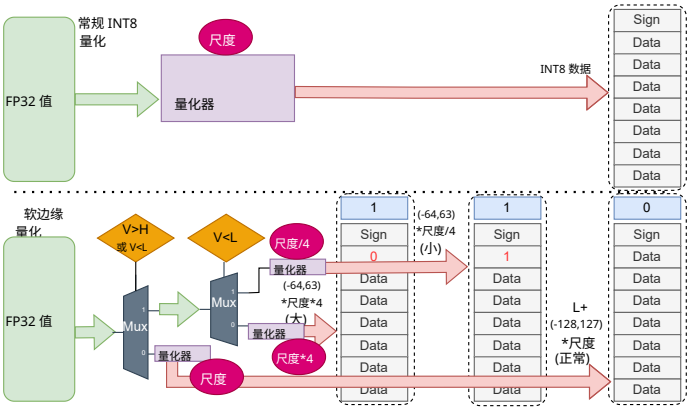

图 2. Quamba-SE 的硬件架构

6 位用于特殊范围数据。结果表明，特殊区域采用 6 位精度已足够。我们的方法提供了卓越的动态范围：小值具有高精度，中等值的精度与 Quamba 相同且毫无损失，而离群值则采用低精度处理（而非截断）。

总之，尽管 Quamba 提供了有效的尺度，但它采用的是标准 INT8 量化。我们沿用 Quamba 的有效尺度，并设计了一种专用硬件量化器，为不同数值范围提供软边缘。

IV. 实验

我们使用 RTX 5090 和 CUDA 12.8，在开源 Quamba 代码库 [10] 上开展实验。由于该框架不支持自定义 INT8，我们将 FP32 值限制为 Quamba-SE 可表示的数值，以模拟定制软边缘量化器硬件设计。我们在最小的 Mamba 模型（130M）上进行评估，因为量化广泛应用于边缘部署，而此类场景更青睐较小的模型。我们使用官方 Quamba 权重以及通过不同百分位校准生成的权重进行测试。与 Quamba 一致，我们报告每个数据集上 5 次运行的平均准确率，共评估 6 个数据集。

我们在**表 I**中展示了与 Quamba 的结果对比，而报告的结果 [2] 仅供参考。在 LAMBADA 上，Quamba-SE 的准确率最高提升 +2.33%（校准 99.999%）。单个数据集的最大提升为 WinoGrande 上的 +2.68%。

在所有校准设置下，Quamba-SE 的表现始终优于 Quamba，平均准确率分别提升 +0.22%、 +0.59% 和 +0.83%。由于边缘设备上的小型模型通常针对特定用例进行定制，我们进一步报告了模型表现尤为出色的部分基准测试，结果显示其较 Quamba 分别提升 +0.36%、 +0.70% 和 +1.48%。整体最大提升（平均值 +1.48%†）出现在使用官方预训练权重时。即使在 Quamba 中，采用 99.99% 百分位数的 Quamba 略优于报告结果 [2], , Quamba-SE 仍然优于 Quamba。

A. 讨论

离群值与精度困境：覆盖离群值会增大量化步长；舍弃离群值则会造成信息损失。我们的软边缘在保留离群值（而非进行硬裁剪）的同时，确保大多数数值的精度。

延迟： 我们的软边缘量化器会增加延迟，但与繁重的矩阵运算相比，量化只是模型中计算量较小的一部分。这点额外延迟可由其带来的准确率提升所抵消。如果应用分支预测，还能进一步降低额外的延迟开销。

Quamba2 与其意义：Quamba2 在 Mamba1 模型上的提升较小（1.4B Mamba1 上，Quamba1 为 57.9%→ ，Quamba2 为 58.1%）。我们的 Quamba-SE 在 130M Mamba1 上比 Quamba1 高出 0.83%。在大语言模型领域，提升通常较为有限；考虑到多数据集评估的复杂性， +0.83% 的提升意义重大。对 Mamba2 使用 Quamba2 进行评估将留待未来工作。

V. 结论

我们提出了 Quamba-SE，这是一种用于 Mamba 激活值量化的软边缘量化器，通过自适应尺度保留离群值，而非进行硬裁剪。我们在 130M Mamba模型上，采用不同设置对六个数据集进行了评估，并与 2025 年发表的最先进水平工作进行了比较。实验表明，相较于 Quamba，准确率提升了+0.22% 至 +0.83%。

由于状态空间模型独特的激活特征需要采用超越标准 INT8 运算的专用量化方法。尽管现有方法着重于在 CUDA 约束下进行改进，我们证明定制硬件量化器能够实现更高的准确率。硬件综合留待未来工作完成。本研究展现了一个有前景的方向——针对特定层和特定模型设计量化器。

参考文献

[1] A. Gu 和 T. Dao，“Mamba：采用选择性状态空间进行线性时间序列建模”，载于第一届语言建模会议，2024。[2] H.-Y. Chiang、C.-C. Chang、N. Frumkin、K.-C. Wu 和 D. Marculescu，“Quamba：面向选择性状态空间模型的训练后量化方案”，载于第十三届国际学习表征会议，2025。[3] S. Biderman、H. Schoelkopf、Q. G. Anthony、H. Bradley、K. O’Brien、E. Hallahan、M. A. Khan、S. Purohit、U. S. Prashanth、E. Raff等，“Pythia：用于分析大语言模型训练与扩展的套件”，载于国际机器学习会议。PMLR，2023，第2397–2430页。[4] A. Pierro 和 S. Abreu，“Mamba-ptq：循环大语言模型中的离群通道”，*arXiv* 预印本 *arXiv:2407.12397*，2024。[5] Z. Liu、B. Oguz、C. Zhao、E. Chang、P. Stock、Y. Mehdad、Y. Shi、R. Krishnamoorthi 和 V. Chandra，“Llm-qat：面向大语言模型的无数据量化感知训练”，载于计算语言学协会 *ACL 2024* 年会论文集，2024，第467–484页。[6] H.-Y. Chiang、C.-C. Chang、N. Frumkin、K.-C. Wu、M. S. Abdelfattah 和 D. Marculescu，“Quamba2：面向选择性状态空间模型的鲁棒、可扩展训练后量化框架”，载于第四十二届国际机器学习会议，2025。[7] Z. Xu、Y. Yue、X. Hu、Z. Yuan、Z. Jiang、Z. Chen、J. Yu、C. Xu、S. Zhou 和 D. Yang，“Mambaquant：采用方差对齐旋转方法量化 Mamba系列”。[8] UT-ENYAC，“Quamba-130m-w8a8”，2025，Hugging Face模型中心：https://huggingface.co/ut-enyac/quamba-130m-w8a8。[9] T. Dao 和 A. Gu，“Transformer 即状态空间模型：通过结构化状态空间对偶实现广义模型与高效算法”，*arXiv* 预印本 *arXiv:2405.21060*，2024。[10] H.-Y. Chiang、C.-C. Chang、N. Frumkin、K.-C. Wu 和 D. Marculescu，“Quamba官方GitHub代码仓库”，2025，https://github.com/enyac-group/Quamba。