中国科学 : 信息科学 2026 年 第 56 卷 第 4 期: 1027–1028
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
精选文章导读

面向存内计算架构的模型量化

孙思凡[1,2], 白金宇[1], 陈翰廷[1], 邓凯文[1], 谢智威[2], 李晶晶[2], 曹斌[3], 张和[1], 康旺[1,2*], 赵巍胜[1,2]

1. 北京航空航天大学集成电路科学与工程学院, 北京 100191
2. 北京航空航天大学国际创新研究院自旋芯片与技术全国重点实验室, 杭州 311115
3. 河北工业大学人工智能与数据科学学院, 天津 300401
* 通信作者. E-mail: wang.kang@buaa.edu.cn

随着人工智能技术的快速演进, 深度神经网络 (DNN) 在视觉、语言与多模态任务中取得了卓越表现, 但其庞大的参数规模与计算量也将传统冯 · 诺依曼架构推向瓶颈. 大量数据在处理器与存储器之间频繁往返, 使得能耗和时延难以持续压降, 尤其在移动终端与高密度边缘场景中更为突出. 存算一体 (CIM) 架构通过在存储阵列内部直接执行矩阵乘加运算, 显著减少数据搬移, 在能效上展现出数倍乃至数十倍的潜在优势, 正在成为下一代智能计算体系结构的重要方向 [1]. 然而, 受限于模拟存储器的物理特性、器件噪声以及 ADC/DAC 精度等约束, 当前 CIM 架构难以支持高精度推理. 因此, 模型量化成为在 CIM 平台上部署 DNN 的核心技术路径, 通过以低比特整数替代浮点表示 [2], 大幅降低存储需求、数据接口开销与模拟域累积误差, 是实现 CIM 高能效推理的关键.

SCIENCE CHINA Information Sciences 在 2026 年第 3 期上发表了综述文章 "Model quantization for computing-in-memory: a survey". 本文系统分析了模型量化在 CIM 加速器上的研究进展, 并围绕固定精度量化 (FPQ)、混合精度量化 (MPQ) 以及量化模型优化三条主线构建了整体技术脉络, 如图 1 所示.

在固定精度量化方面, 本文首先分析了 CIM 平台的独特映射方式, 讨论了不同映射策略对计算性能和存储效率的影响. 特别地, 本文对比了传统的权重展平映射与基于空间位置的权重映射方法, 指出映射策略不仅影响阵列的利用率, 还与模拟噪声对计算精度的放大效应密切相关. 接着, 本文详细讨论了固定精度量化中不同量化对象 (如权重、激活和部分和) 的处理方式 [3], 特别强调了部分和量化在 CIM 中的重要性. 由于部分和的计算需要经过模拟到数字转换 (ADC), 其能耗在整体功耗中占比非常高, 因此低比特部分和量化是提升 CIM 效率的关键环节. 本文进一步介绍了几种常见的量化粒度, 包括层级、组级、通道级以及更贴近硬件结构的阵列级粒度, 并对均匀量化与非均匀量化方法进行了对比, 尤其是非均匀量化在应对模拟电路非线性和噪声方面的优势. 此外, 本文还讨论了二值和三值网络在 CIM 平台上的应用, 分析了二值和三值量化通过将乘法操作替换为位移操作, 显著提升了硬件效率, 尤其是在计算需求低、能效要求高的应用场景中.

对于混合精度量化, 本文进一步探讨了如何根据 CIM 平台的硬件特点, 为不同层或通道分配不同的精度, 从而在不显著降低模型准确度的情况下, 减少计算和存储开销 [4]. 混合精度量化能够有效地平衡性能和硬件效率, 尤其适用于需要同时兼顾精度和硬件资源限制的场景. 本文介绍了 CIM 平台上如何实现混合精度映射, 通过为不同层或通道分配不同的位宽, 以优化计算和存储效率. 接着讨论了 MPQ 的搜索空间设计, 重点分析了位宽候选和量化粒度的选择, 如何根据硬件特性和任务需求进行合理配置. 随后, 本文介绍了几种常见的搜索策略, 如进化算法、强化学习和基于梯度的方法, 来有效探索广泛的搜索空间. 最后, 讨论了 MPQ 配置评估的方法, 强调了如何在硬件约束下, 通过仿真和

英文原文: Sun S F, Bai J Y, Chen H T, et al. Model quantization for computing-in-memory: a survey. Sci China Inf Sci, 2025, 68: 211401, doi: 10.1007/s11432-024-4522-8

© 2026《中国科学》杂志社

https://www.sciengine.com/doi/10.1360/SSI-2025-0503

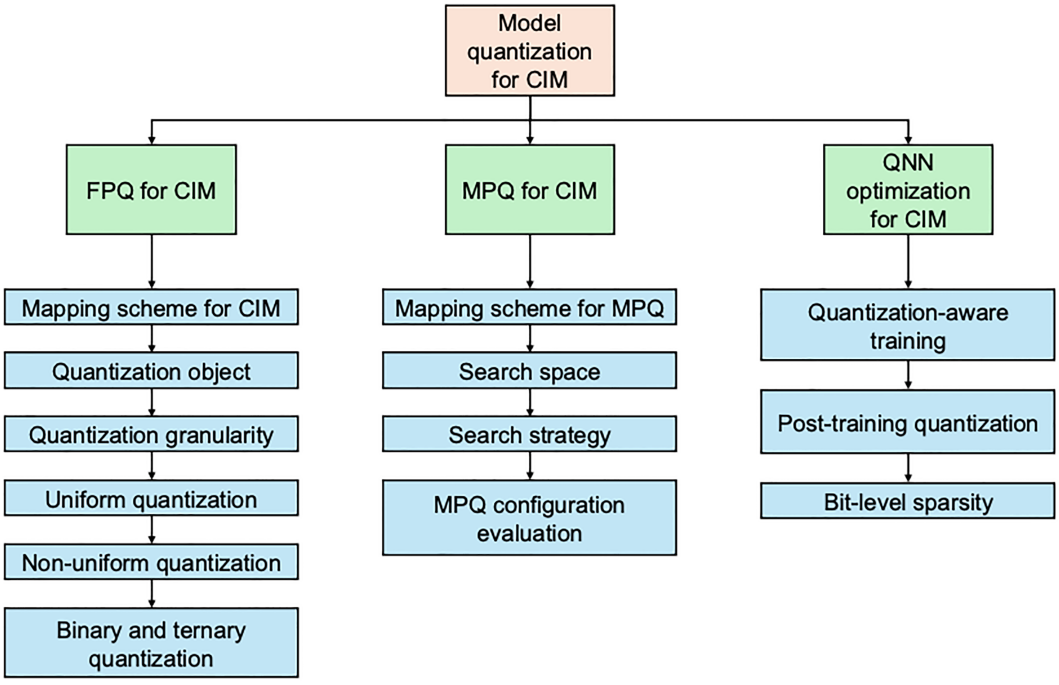


图 1　(Color online) 面向存内计算的模型量化技术分类.

Figure 1　(Color online) Classification of model quantization for CIM.

误差度量来加速配置的评估过程, 确保优化的量化配置能够平衡性能和硬件效率.

在量化模型优化方面, 本文介绍了量化感知训练 (QAT) 和后训练量化 (PTQ) 两种常见的优化策略. 量化感知训练通过在训练过程中模拟低精度计算, 使得模型能够适应量化带来的精度损失; 而后训练量化则是在训练完成后, 将训练得到的高精度模型进行量化. 两种方法在 CIM 平台上的应用效果不同, 本文详细分析了它们在不同应用场景中的优缺点, 并提出了一些新的优化思路. 特别地, 针对 CIM 的硬件特性, 本文还探讨了比特级稀疏化的优化方法, 通过动态稀疏化减少计算量和存储开销 [5], 进一步提升 CIM 加速器的效率.

最后, 本文总结了目前在 CIM 平台上进行模型量化的研究成果, 并展望了未来的发展方向. 尽管 CIM 架构为加速深度学习提供了前所未有的机会, 但由于其模拟计算特性和硬件精度限制, 如何设计出既高效又精准的量化策略仍是一个具有挑战性的问题. 未来的研究应聚焦于如何进一步优化量化方法, 以充分发挥 CIM 硬件的计算潜力, 并推动该领域在实际应用中的落地与发展.

参考文献

1　Yu S, Jiang H, Huang S, et al. Compute-in-memory chips for deep learning: recent trends and prospects. IEEE Circ Syst Mag, 2021, 21: 31–56

2　Rokh B, Azarpeyvand A, Khanteymoori A. A comprehensive survey on model quantization for deep neural networks in image classification. ACM Trans Intell Syst Technol, 2023, 14: 1–50

3　Bai J, Sun S, Zhao W, et al. CIMQ: a hardware-efficient quantization framework for computing-in-memory-based neural network accelerators. IEEE Trans Comput-Aided Design Integrated Circ Syst, 2024, 43: 189–202

4　Sun S, Bai J, Shi Z, et al. CIM^{2}PQ: an arraywise and hardware-friendly mixed precision quantization method for analog computing-in-memory. IEEE Trans Comput-Aided Design Integrated Circ Syst, 2024, 43: 2084–2097

5　Xue W, Bai J, Sun S, et al. Hierarchical non-structured pruning for computing-in-memory accelerators with reduced ADC resolution requirement. In: Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE), 2023. 1–6

https://www.sciengine.com/doi/10.1360/SSI-2025-0503