https://crad.ict.ac.cn
63(2)：275−275, 2026

DOI: 10.7544/issn1000-1239.ps20260201 **CSTR:** 32373.14.issn1000-1239.ps20260201

专家评述

从大模型量化问题揭示复域数制之妙

徐志伟

（大湾区大学　广东东莞　523000）

数制（numerical system）是计算机科学技术的基础之一，经历了4个发展阶段。第1阶段的里程碑是莱布尼茨在1703年发表的论文《仅使用数字0和1的二进制算术阐述，兼论其效用，及其对古代中国伏羲符号的揭示》，它展示了可用二进制符号实现数的加减乘除运算。第2阶段的里程碑是冯·诺依曼架构，可用二进制实现算术、逻辑运算、字符甚至指令。第3阶段的里程碑是IEEE 754浮点数标准，它实现了浮点数运算的代数完备性，显著降低了科学计算的错误率，卡汉教授因其贡献获得图灵奖。当前我们正处于第4阶段，其中一个重要研究领域是量化（quantization）：使用更少比特的数制来改善计算速度和能效，同时保证结果性能。量化技术20多年前在超级计算领域率先发展（参见Jack Dongarra在ISPA 2006的报告），但直到最近十余年才在人工智能领域逐步得到广泛应用。

北京大学杨仝团队的“大语言模型低比特量化的历史、现状和未来：以复域量化为例”前沿研究论文，为读者带来了3个惊喜：提出了令人惊奇的研究问题与挑战，介绍了令人惊喜的洞察、理论与技术，开启了“复域量化”的研究方向。

量化压缩后的模型能否取得比全精度模型更好的结果性能？杨仝团队提出了这个貌似不可能完成的挑战。业界已有很多量化工作，大都采用了“量化压缩模型的性能=全精度模型性能−量化误差”的思维模式。例如，有工作只用1比特，似乎将量化做到了极致。但这种思维模式自然就将“全精度模型性能”设成了上限。杨仝团队突破了这个研究问题的天花板，并在理论和技术方面做到了“量化模型性能>全精度模型性能”。

杨仝团队如何攻克这个貌似不可能完成的挑战呢？有什么奇妙的学术点？该论文妙在展示了团队的数学洞察、理论创新、技术突破。他们继承了从黎曼到彭罗斯的智慧：复数有其独特的奇妙。他们洞察到，复数不只是一对实数，而是一个整体，进而用其“幅度”与“相位”的双自由度特性来扩展信息表达空间。他们基于数学原理提出了“复域量化”理论，其相位量化算法优美地利用了复数平面的几何对称性，实现了无乘法的高效推理。在技术方面，团队构建了全复数化的Transformer架构，并打通了实数大模型向复域量化大模型迁移的路径。

该论文开辟了“复域量化”的研究新方向，尚有很多问题值得研究：1)“复域量化”能否用于大模型之外的其它应用领域？2）能否集成到处理器芯片、软件库、操作系统、应用框架等各类产业生态中，得到实践验证，甚至启发新型计算架构？ 3）能否拓展到更多比特，如16比特，进而提出一种媲美IEEE 754浮点数标准的复数数制？

随着中国研究者日益活跃在全球学术界，《计算机研究与发展》特别欢迎学者在本刊率先发表杨仝团队这样的原创性研究成果，进而将论文写在中国和世界的大地上。

评述专家



徐志伟，中国科学院计算技术研究所研究员，大湾区大学讲席教授，中国计算机学会会士，《计算机研究与发展》主编。研究方向为高性能计算机体系结构、分布式系统、算力网前沿技术。

亮点论文

丛培壮, 王飞宇, 王国安, 王砚舒, 郑策, 杨仝. 大语言模型低比特量化的历史、现状和未来：以复域量化为例 [J]. 计算机研究与发展, 2026, 63(2)：276−293. DOI: 10.7544/issn1000-1239.202550788