计 算 机 研 究 与 发 展
Journal of Computer Research and Development

https://crad.ict.ac.cn
63(2): 276−293, 2026

亮点论文

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

丛培壮[1] 王飞宇[1] 王国安[1] 王砚舒[1] 郑 策[2] 杨 仝[1]

[1](多媒体信息处理全国重点实验室(北京大学) 北京 100871)

[2](北京航空航天大学网络空间安全学院 北京 100191)

(congpeizhuang@pku.edu.cn)

History, Present, and Future of Low-Bit Quantization for Large Language Model: A Case Study on Complex-Domain Quantization

Cong Peizhuang[1], Wang Feiyu[1], Wang Guo'an[1], Wang Yanshu[1], Zheng Ce[2], and Yang Tong[1]

[1] (*State Key Laboratory of Multimedia Information Processing* (*Peking University*), *Beijing* 100871)

[2] (*School of Cyber Science and Technology*, *Beihang University*, *Beijing* 100191)

Abstract With the exponential growth in the parameter scale of large language models (LLMs), model deployment and inference face severe challenges in terms of memory and computational resources. Quantization, as a core compression technique, significantly reduces storage requirements and computational overhead by lowering the numerical precision of weights and activations. We first review the development of quantization techniques, from classic Int8/4 methods to cutting-edge extreme-low bit algorithms. Then we summarize the technical characteristics and performance evolution of typical methods, identifying a key challenge: traditional real-domain quantization is limited by discretization errors at extreme-low bitrates, making it difficult to break its performance ceiling. To address this limitation, we systematically review the complex-domain quantization series of work. This series introduces a quantization paradigm based on the complex domain, which significantly expands the model's expressive space by utilizing amplitude and phase as two degrees of freedom in the parameter representation. Moreover, drawing on the classic paradigm in signal processing where stable representations are achieved by applying the Fourier transform and low-pass filtering to time-domain signals, we further propose a technical path that proceeds from real-valued models through complex-domain transformation and complex-domain quantization, ultimately achieving multiplier-free stable inference. Experimental results demonstrate that iFairy outperforms existing extreme-low-bit methods on multiple benchmark datasets and effectively breaks the real-domain performance ceiling. This showcases the potential value of complex-domain quantization in achieving both efficient modeling and performance preservation. Through a systematic analysis of quantization's evolution and the case study on complex-domain quantization, we reveal development patterns and future trends, providing a reference for the theoretical research and engineering implementation of efficient LLMs.

Key words large language model; model quantization; low-bit quantization; model compression; complex valued model

摘 要 随着大语言模型(large language model, LLM)参数规模的指数级增长,模型部署和推理面临着严

收稿日期: 2025-10-31; **修回日期**: 2025-12-18

基金项目: 国家重点研发计划项目(2024YFB2906601); 国家自然科学基金项目(62372009, 62502014)

This work was supported by the National Key Research and Development Program of China (2024YFB2906601) and the National Natural Science Foundation of China (62372009, 62502014).

通信作者: 杨仝(yangtong@pku.edu.cn)

峻的内存和计算资源挑战。量化技术作为模型压缩的核心方法，通过降低权重和激活值的数值精度，显著减少了模型的存储需求和计算开销。首先回顾了量化技术的发展历程，从经典的 Int8/4 量化方法到前沿的超低比特量化算法，总结了典型方法的技术特征与性能演进规律，指出传统实数域量化在极低比特条件下存在受限于离散化误差的挑战，难以突破性能上限。为此，进而系统性地梳理了复域量化系列工作。该系列工作提出了基于复数域的量化范式，通过在参数表示中引入幅度与相位 2 个自由度，显著扩展了模型的表达空间；此外，类比信号处理中通过将时域信号进行傅里叶变换与低通滤波实现稳定表示的经典范式，进一步提出了由实数模型经复域变换与复域量化，达成了无乘法稳定推理的技术路线。实验结果表明，该方案在多个基准数据集上优于现有超低比特量化方法，有效突破了实数域模型的性能天花板，展现出复域量化在高效建模与性能保持方面的潜在价值。总体而言，通过对量化技术演进及复域量化系列研究的系统分析，旨在揭示超低比特量化的发展规律与未来趋势，为高效大模型的理论研究与工程实现提供参考。

关键词 大语言模型；模型量化；低比特量化；模型压缩；复数模型

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550788 **CSTR:** 32373.14.issn1000-1239.202550788

大语言模型（large language model, LLM）（以下简称“大模型”）的参数规模正以惊人的速度增长，从早期的千亿级别迅速发展至当前的数千亿乃至万亿级别。模型能力的提升伴随着对存储和计算资源需求的爆炸式增长[1-2]。一个万亿级参数的旗舰模型，即便以 FP16 精度存储，其权重本身也需要数 TB 的内存空间。这种资源瓶颈不仅限制了模型的应用范围，也导致了高昂的运营成本。量化技术作为模型压缩的核心方法，通过降低权重和激活值的数值精度，能够显著减少存储需求和计算开销。从信息论角度来看，神经网络参数存在大量冗余：权重分布通常呈现长尾特征，超过 95% 的权重值落在（均值±3）倍标准差的范围内；激活值在 ReLU 等函数作用后稀疏度可达 50%~70%。这些特性为量化提供了理论基础。

量化技术的发展大致经历了 3 个阶段，如图 1 所示。Int8（采用 8 位整数）量化的工程化阶段（2015—2018 年），Jacob 等人[3]提出的整数量化方法系统化实现了训练与推理的全整数算术，TensorFlow Lite[4]、TensorRT[5]等框架推动其在工业界广泛应用。Int8 量化可带来约 4 倍压缩比，在多数任务上性能损失微小。Int4 量化的突破阶段（2018—2022 年），Transformer 架构和大规模预训练模型的兴起促使 Int4 量化成为研究热点，研究工作 GPTQ[6]、AWQ[7]、SmoothQuant[8]等通过后训练量化（post-training quantization, PTQ）和量化感知训练（quantization aware training, QAT）的结合，使百亿参数模型能够在单机 GPU 上推理，实现约 8 倍压缩比。超低比特量化的探索阶段（2022—至今），BitNet[9]、OneBit[10]等 1~2 比特方案展现出惊人潜力，将模型压缩至原来的 1/16~1/32。更重要的是，这一阶段不再局限于工程优化，而是开始探索极限量化下的新架构范式和理论边界。

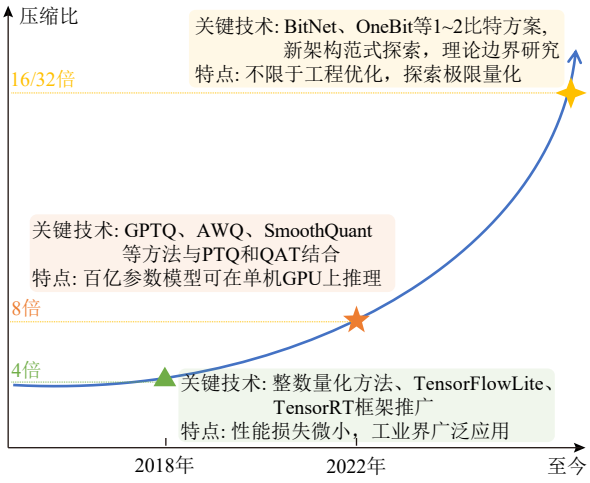


Fig. 1 Developmental trends of LLMs quantization technology

图 1 LLM 量化技术发展趋势

当前超低比特量化面临的核心难题在于如何在极低比特条件下同时保持模型精度、训练稳定性与系统加速效率。当量化精度降至 2 个比特及以下时，传统的离散映射方式会导致严重的信息损失，模型训练容易出现收敛困难与性能崩塌。尽管 PTQ 通过显著权重分治、向量量化及递推码本构建等方式提升了量化鲁棒性[11-12]，QAT 借助 BitLinear 层、符号分解与反死区机制增强了可优化性[13-14]，但这些方法仍受限于实值参数空间的离散化误差，难以在小于且等于 2 比特下维持充分的表达能力。混合方案如“N：M”稀疏二值化与自回归蒸馏虽实现亚 1 比特

压缩,却未能突破性能边界。同时,系统层面的软硬件协同问题进一步限制了推理加速。通用 CPU/GPU 针对浮点运算优化,导致低比特整数运算存在固有效率损失,即便引入改进的矩阵原语、通用类型系统或定制加速器,仍难实现性能与能效的同步提升。更深层的理论分析表明,量化误差与模型参数量、训练数据量及位宽之间呈幂律关系[15],且训练越充分的模型对量化噪声越敏感。这一现象揭示了当前量化技术的根本瓶颈:当模型受限于实数域参数表示时,其性能上限由全精度模型决定,即量化优化只能在既定天花板下权衡精度与压缩比[16]。

为突破这一结构性限制,复域量化提出了从实数域扩展至复数域的全新范式。不同于传统方法将量化视为对全精度模型的近似压缩,复域量化通过为每个参数引入幅度与相位 2 个自由度,在不增加参数数量的前提下显著提升了模型的表达空间。基于复域量化的核心思想,复数比特模型 iFairy[17] 提出了复数域 Transformer 架构,包括复数线性层(ComplexLinear)、埃尔米特注意力(Hermitian attention)机制、复数旋转位置编码 Complex RoPE(complex rotary position embedding)及 ReLU^2 激活函数,以保持幅度与相位的可分离和可训练性。在量化策略上,基于相位的四元量化{±1,±i}实现了真正的 2 比特满载编码,推理仅需加法与实虚部交换操作,从而在能效上远优于传统乘法计算。实验结果表明,iFairy 在 WikiText2 和 C4 数据集上的困惑度表现优于 BitNet b1.58,部分指标甚至超过全精度 LLaMA 模型,展示了复域量化在精度、效率与理论完备性上的潜在突破。在此基础上,该研究进一步提出了复数转换模型 Fairy2i[18],实现了将实数模型转换为 2 比特复数模型,进而使得模型的开发和部署兼顾极低成本与复数域带来的高精度优势。

为了更清晰地定位以 iFairy 为代表的复域量化范式,本文旨在提供一个系统性的分析框架。首先,我们将在第 1 节中完整梳理从 Int8/4 到超低比特量化的技术谱系与理论基础。在此之上,第 2 节将深入探讨超低比特量化面临的进展与挑战,并从扩展规律、算法改进及推理加速等维度进行多维度分析。基于这一全面的技术图景,第 3 节将重点介绍复数模型新范式,论证其作为突破实数域瓶颈的潜力,并以 iFairy 与 Fairy2i 为例,深入剖析其在架构设计、量化策略与工程实现上的关键决策与可行路径。最后,在第 4 节展望未来的研究方向,并在第 5 节进行全文总结。

1 历史与沿革:极限量化的基石

LLM 近年来在自然语言处理以及多模态理解与生成等任务中取得了显著进展。然而,其庞大的参数规模与巨大的算力消耗成为模型部署与应用中的主要瓶颈。低比特量化作为提升推理效率与降低存储开销的关键技术路径,其研究可追溯至神经网络早期的量化实践。早期研究虽主要聚焦于卷积神经网络与移动端模型压缩,但相关理论与实现经验为后续大模型的量化研究提供了重要支撑。随着 LLM 的快速发展,量化方法逐步从传统网络迁移并演化为适配大规模 Transformer 架构的新一代算法。近年来,低比特量化技术形成了 PTQ 和 QAT 两条主要的发展路线。其中,PTQ 以易用性和工程部署效率为特征,在早期模型压缩与推理加速中发挥了重要作用;而 QAT 则通过在训练过程中显式引入量化机制,使模型能够在参数更新中逐步适应低比特表示下的数值扰动,从而在极低比特条件下保持较高的精度与稳定性。随着模型规模的进一步扩大与性能需求的提升,QAT 已逐渐成为低比特量化与超低比特量化的主要研究方向,也是当前 LLM 量化领域的核心热点。

下面将系统梳理这 2 类方法的发展脉络与核心思想,重点讨论 PTQ 的发展与局限,以及 QAT 的提出、技术突破与最新研究趋势。

1.1 PTQ 的发展与局限

1.1.1 早期的 Int8 量化与移动端部署

在神经网络早期的应用中,全精度浮点(FP32)运算因存储与算力开销过大,难以在移动端与嵌入式设备上高效部署。研究逐渐发现,神经网络对数值精度的依赖相对有限,采用 Int8 即可在多数任务中维持较高性能。Jacob 等人[3] 提出的整数量化方法系统地建立了基于缩放因子与零点映射的整数算术框架,为后续 PTQ 的发展奠定了理论与工程基础,也为 QAT 等基于训练阶段量化的研究提供了启发。随后,Google 公司的 TensorFlow Lite 与 NVIDIA 公司的 TensorRT 等框架将 Int8 量化工程化,使其在移动端与 GPU 平台上得到广泛应用,成为早期深度学习模型部署的标准方案。

随着 BERT、GPT 等大规模预训练模型的兴起,PTQ 在大模型推理中的价值进一步凸显。Dettmers 等人[19] 提出的 GPT3.int8() 方法在百亿参数级 Transformer

中系统验证了 Int8 矩阵乘法的可行性，证明在保持接近 FP16 精度的同时，能够显著降低显存占用与推理延迟。这一成果被普遍认为是 Int8 在 LLM 部署中的开创性工作。目前 TensorRT-LLM、DeepSpeed-Inference、ONNX Runtime 等主流推理框架均已将 Int8 PTQ 作为核心优化选项，使其成为大规模模型落地中最成熟且稳健的压缩方案之一。

1.1.2 从 Int8 到 Int2 的拓展与应用

随着模型规模的持续扩展，Int8 压缩已无法充分满足算力与存储方面的需求，研究者开始探索更低比特的 PTQ。Banner 等人[20] 提出的 PTQ Int4 方法证明无需重新训练即可完成模型压缩，大幅简化了部署流程。进入 LLM 时代后，直接套用低比特 PTQ 往往导致性能急剧下降，因此出现了针对 LLM 特性的工程化改进。SmoothQuant[8] 通过缩放权重与激活来平衡分布，缓解了异常值激活带来的精度下降；AWQ[7] 在量化过程中引入激活分布感知机制，在对话生成和问答等任务中表现出更强的鲁棒性；ZeroQuant[21] 则提出逐层量化与在线校准，实现了无需训练数据的高效 PTQ。进一步地，QuIP[22] 与 QuIP#[23] 利用非相干性处理与哈达玛变换优化了向量量化，QTIP[24] 通过网格编码解决了高维码本瓶颈，YAQA[25] 则基于全局海森矩阵实现了端到端误差最小化。这些方法推动了低比特 PTQ 在 LLM 中的可用性，使百亿参数模型能够在单机 GPU 上完成推理，显著提升了 LLM 的部署效率与应用可行性。

1.1.3 局限性与研究转向

尽管 PTQ 具有无需额外训练、部署便捷和硬件支持完善等优势，是早期大模型落地的主要方案，但其在低于 8 比特的条件下精度保持能力明显不足，或仅适用于部分简单任务。特别是在 Transformer 的注意力模块和层归一化等对数值分布高度敏感的结构中，模型困惑度容易上升，任务性能亦显著下降。这一问题的根源在于大模型参数与激活分布的长尾特性，单纯的整数量化难以有效覆盖异常值范围。随着研究的深入，学界开始探索在训练阶段引入量化建模的新思路，使模型在训练过程中逐步适应低比特约束，以从根本上缓解精度退化问题。这一趋势表明，低比特量化的研究正从“推理后校准”逐步过渡到“训练阶段优化”，成为 QAT 兴起的重要契机。

1.2 QAT 的提出与突破

1.2.1 从 PTQ 到 QAT：思想的转变

随着低比特量化研究的不断深入，研究者逐渐认识到，仅依赖 PTQ 难以在复杂任务和更低比特条件下保持稳定性能。为此，QAT 应运而生，其核心思想是在训练阶段引入量化近似，使模型在参数更新过程中逐步适应低比特表示带来的数值扰动，从而有效提升量化稳定性和精度保持能力。Jacob 等人[3] 的研究利用缩放因子与直通估计器（straight-through estimator, STE）近似不可导的量化函数，建立了系统的 QAT 训练框架。这一思想随后被 TensorFlow Lite、TensorRT 和 ONNX Runtime 等主流系统采纳，使 QAT 从研究探索阶段逐步迈向工程实践，推动量化方法从推理端后处理转向训练端建模的范式转变。

1.2.2 向低比特量化演进：QAT 的方法突破

在 Int8 量化条件下，QAT 与 PTQ 的性能差距相对有限，但当比特宽度降低至 4 个比特及以下时，QAT 的优势愈加突出。此时，量化误差、梯度传播不稳定及激活分布失衡等问题显著加剧，传统的 PTQ 难以维持性能稳定。QAT 通过在训练阶段显式建模量化过程，使模型在参数更新中逐步适应低比特约束，从而有效缓解精度退化与收敛不稳定问题。在方法演进方面，QAT 的发展经历了从低比特可训练性验证到结构感知与稳定性优化的持续突破。

1）低比特可训练性的确立

DoReFa-Net[26] 与三值权重网络（ternary weight network, TWN）[13] 系统验证了低比特网络的可训练性，提出了同时量化权重、激活与梯度的训练框架，证明网络在低于 8 比特精度下仍能稳定收敛，为后续研究提供了方法学基础。

2）结构敏感与自适应优化的发展

早期的 HAWQ[27] 在 PTQ 框架下利用 Hessian 光谱信息实现层级精度分配，为后续 QAT 的结构敏感研究提供了启示。LSQ [28] 及其改进版本 LSQ+[29] 在训练中引入可学习的量化步长与偏移参数，通过动态调整量化区间与梯度传播，有效提升了 4 比特及以下 QAT 的稳定性与精度保持。与此同时，Q-BERT[30] 将 Hessian 信息引入 Transformer 架构的量化感知训练过程，通过层敏感度分析与精度分配策略，实现了大规模预训练模型在 4 比特条件下的高精度量化。这些研究共同推动了 QAT 从固定量化策略向自适应优化演进，成为低比特量化的重要发展方向。

综上，QAT 在低比特量化阶段已从基于全局误差补偿的粗粒度方案演进为结构感知与自适应优化的精细化训练框架。相较于依赖推理端校准的 PTQ，QAT 更注重训练动态与网络结构特性的协同建模，在 Int4 及以下量化下显著提升了模型精度与稳定性，成为低比特可训练化的核心技术路径。

1.2.3 走向超低比特：QAT 的新探索

在 LLM 场景中，研究者进一步将 QAT 推向更为极端的比特宽度。以 BitNet[9,12] 为例，该研究在 Transformer 架构中引入 BitLinear 模块，通过 QAT 实现了 1.58 比特模型的训练可行性，其在部分语言建模任务上的表现已接近半精度(FP16)基线。随后，OneBit[10] 构建了首个端到端 1 比特 QAT 框架，提出了新的 1 比特参数表示与基于矩阵分解的初始化方法，有效改善了训练不稳定的问题，并在通过蒸馏操作来弥补量化带来的显著性能损失。这一系列研究表明，QAT 已由传统的部署优化工具转向研究大模型数值极限的核心技术路径。

1.3 小 节

本节梳理了低比特量化技术的发展脉络，并从 PTQ 与 QAT 两条技术路线进行了系统分析。为便于比较不同方法的特征与性能，表 1 汇总了近年来具有代表性的研究成果，从提出时间、压缩率、推理速度与任务性能等维度展示了量化技术从 Int8 工程化部署逐步演进至 1~2 比特极限探索的发展趋势。

由此可知，PTQ 以部署简便、计算友好为优势，在大模型的实际推理中仍具有广泛应用价值；但在更低比特条件下，其精度保持能力受限。相比之下，QAT 通过在训练过程中引入量化建模与误差补偿机制，能够显著提升低比特模型的稳定性与可训练性，已成为当前低比特与超低比特量化研究的核心方向。QAT 的持续发展不仅推动了 LLM 压缩技术从“工程可用”走向“科学可控”，也成为当前 LLM 量化研究的主要热点。基于此，将在第 2 节进一步探讨超低比特量化的最新进展与主要挑战。

2 超低比特量化研究现状

2.1 超低比特量化的缩放定律

超低比特(extremely low bit, ELB)量化是提升 LLM 推理效率的重要研究方向[36]，而对量化性能损失 QiD(quantization-induced degradation)内在规律的解析已成为该领域的一个重要环节。缩放定律(scaling laws)提供了一个强大的理论框架，旨在量化模型固有属性(如规模、训练数据量)与量化效果之间的系统性交互作用。本节将系统地综述和分析该领域的关键发现，特别是如何构建 QiD 的量化模型，以及这些理论如何指导 ELB 量化算法和工程策略的突破，以推动 ELB 量化的实用化进程。

2.1.1 训练收敛与量化鲁棒性的关联机制

早期探索的量化缩放定律的研究工作证实了 QiD($\Delta_q Loss$)与模型核心属性之间的幂律关系[37]，其中模型参数量N和量化位宽P对 QiD 具有抑制作用，即$\Delta_q Loss \propto 1/N^{\alpha}$和$\Delta_q Loss \propto 1/P^{\gamma}$。然而，近期研究工作[38]揭示了一个颇具反直觉的规律：QiD 与训练数据量D之间呈现显著的正相关性，即$\Delta_q Loss \propto D^{\beta}$，如

Table 1 Comparison of Representative PTQ and QAT Low-Bit Quantization Methods in LLM

表 1 LLM 中典型 PTQ 与 QAT 低比特量化方法对比

方法类别	提出时间/应用场景	存储压缩率	推理速度	任务性能	代表性研究成果
PTQ	Int8（2018—2022 年，CNN→LLM）	≈4 倍	≈1.5~2 倍实际加速（理论 4 倍）	推理端几乎无损	Integer-only Quantization[1]；GPT3.int8()[19]
	Int4（2019 年提出，2022—2023 年，LLM 常用低比特部署）	≈8 倍	≈3~4 倍实际加速（理论 8 倍）	轻微下降，可接受	PTQ-4bit[20]；SmoothQuant[8]
	Int2（2023—2025 年，LLM 进一步压缩）	≈16 倍	≈6~7 倍实际加速（理论 16 倍）	2 比特精度明显提升	QuIP[22]; QuIP#[23]; QTIP[24]
	PTQ 工程优化（2022—2024 年，大模型常用）	4~8 倍	2~6 倍	精度略低于 QAT，但部署便捷	OBC[31]；GPTQ[6]；AWQ[7]；QuaRot[32]
	稀疏+PTQ（2022—2023 年，工业落地）	10~20 倍	10~20 倍	兼顾压缩与加速，逐渐工业化	ZeroQuant[21]；SparseGPT[33]
	混合精度 PTQ（2019—2020 年，BERT/NLP）	4~8 倍	2~6 倍	精度接近 FP32，灵活性强	HAWQ[27]；
QAT	Int8（2018—2022 年，CNN→LLM）	≈4 倍	≈1.5~2 倍	精度保持优于 PTQ	Integer-only Quantization[1]；TensorRT QAT[5]
	Int4 / 三值 QAT（2016—2023 年）	≈8~20 倍	≈3~10 倍	精度下降较小，稳定性好	DoReFa-Net[26]；TWN[13]；LSQ[28]；LSQ+[29]；Q-BERT[30]
	二值 QAT（2015—2016 年，早期探索）	≈32 倍	数十倍理论加速，依赖硬件	精度显著下降，复杂任务不足	BinaryConnect[34]；XNOR-Net[35]；
	BitNet（1.58 比特，2023 年）	≈20 倍（理论值）	≈10~20 倍	性能出现下降，无法突破原模型精度	BitNet[9]；BitNet b1.58[12]
	OneBit（1 比特，2024 年）	≈32 倍	≈20 倍	性能下降明显，高度依赖蒸馏	OneBit[10]
	复域量化（2 比特，2025）	≈16 倍	2 比特位宽与无乘法运算的双重加速	性能超越 FP16，突破 2 比特性能天花板	iFairy[17]；Fairy2i[18]

注：存储压缩率和推理速度都是相对于 FP32 计算的。

图 2 所示。也就是说，训练得越充分的模型，其对量化噪声的敏感度越高。这一现象表明，当模型在大规模数据上充分收敛时，其权重空间可能形成高度结构化且对数值扰动敏感的表示机制。关键的语义与功能信息可能被编码在权重值的微小差异中，从而使低位精度截断更容易引发高维特征空间的结构性破坏。相对地，训练不足的模型仍处于特征学习的早期阶段，其权重分布更为平滑，尚未形成精细化的表征结构，因此对量化误差具有天然的“欠训练鲁棒性”。量化噪声在此阶段往往被欠拟合特性所掩盖，从而对整体性能的影响较小。由此可见，量化损失的预测与控制不仅依赖于模型规模与精度，还需将训练收敛状态视为同等重要的因素。这一视角为超大规模大模型的训练量化部署提供了理论基准，并为工程实践中利用“欠训练鲁棒性”选择合适的量化时机提供了指导。

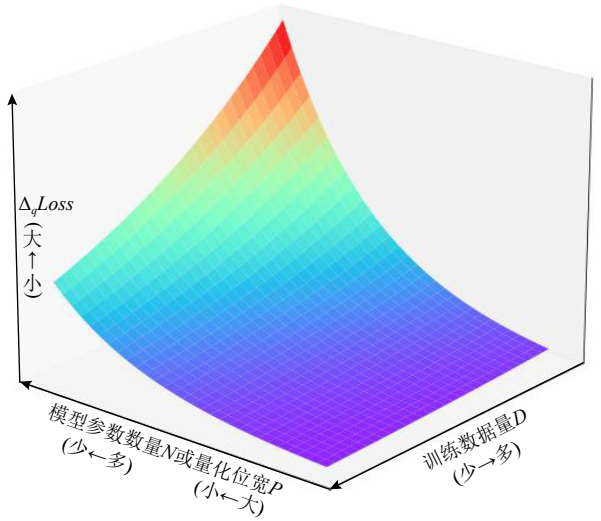


Fig. 2 Example of QiD

图 2 QiD 示例

2.1.2 极低位宽量化的算法突破

尽管基础缩放定律能够刻画$\Delta_q Loss$的总体变化趋势，但在追求极致推理效率的极低比特场景下，如$P \leqslant 2$，其预测的性能损失往往过于严重，难以在实际应用中接受。经验研究表明，Int4 精度是量化性能的关键转折点：当位宽低于 4 比特时，模型精度通常会出现陡降[39]。要突破这一指数级性能约束，关键在于算法层面的主动干预，即通过优化量化感知训练流程增强模型对量化噪声的适应能力。

针对 ELB 场景的研究普遍聚焦于在极端压缩下最大限度地保留信息，通过精细化的 QAT 调度策略与优化的量化函数设计来突破传统量化的性能极限。例如，采用非对称量化或混合位宽策略等方法可有效改善模型在低位精度下的稳定性与表示能力[40]。从缩放定律的视角来看，这类算法优化本质上是修正$\Delta_q Loss$公式中的比例常数k，从而在极小位宽条件下有效缓解指数项γ导致的损失急剧上升。通过 QAT 的迭代优化，模型权重分布被调整为更加“量化友好”的形态，使截断误差对前向传播的影响最小化。由此实现的性能改进不仅体现了量化效率与模型性能之间的帕累托最优提升，也验证了算法创新在重塑模型量化鲁棒性方面的关键作用。这一研究显著拓展了量化理论的适用边界，使得 2 比特级别的量化方案首次具备了实际可用性。

2.1.3 计算最优 QAT 的决策模型

量化缩放定律的核心应用价值在于为受限计算资源条件下的最优工程决策提供理论指导。大模型训练流程中，当计算预算（FLOPs）固定时，如何在全精度预训练阶段（影响数据规模D）与量化感知训练阶段（影响D_{QAT}）之间合理分配计算资源，成为决定模型最终性能与效率平衡的关键问题。在此背景下，Dremov 等人[41]提出了计算最优 QAT 的范式，即将N、D和P之间的相互作用纳入一个联合优化框架，实现了从描述性缩放定律向处方性决策模型的转变。并且，设定的目标函数综合考虑了D与D_{QAT}对训练 FLOPs 的消耗，以及量化位宽P对推理吞吐量的影响，从而能够在给定计算约束下全局优化模型训练路径。基于缩放定律对$\Delta_q Loss$的可预测性，该框架能够精确确定 QAT 的最优切换时机与持续时长，从而在有限计算预算内实现性能与推理效率的最大化。此方法论的提出标志着大模型基础设施的设计正从经验驱动转向量化决策驱动，使得计算资源的配置和训练策略的制定具备了可计算、可优化的理论依据。

2.2 超低比特量化算法研究进展

2.2.1 PTQ

后训练超低比特量化算法主要存在的问题是在极限压缩下避免模型性能的断崖式下跌。早期的后训练超低比特量化算法聚焦于如何最小化量化前后单个权重值的误差，但实践证明后训练超低比特量化的瓶颈并不在于单个权重的准确性，而是能否在压缩后保留原本由一群权重共同表达的协方差结构，因此研究方向逐渐从标量量化转向向量量化。

BiLLM[11]基于海森矩阵（Hessian matrix, HM）识别出数量较少的“显著权重”和其余大量的、呈钟形分布的“非显著权重”，对不同的权重进行分治处理，以更精确地减小量化误差。TesseraQ[42]提出了渐进

式自适应舍入方法,逐步将软舍入参数硬化为离散值,同时结合反量化缩放因子调优,从而显著减少块粒度的重构误差。PTQ 1.61[43]通过引入一种基于输入激活值的 1 维结构化掩码,解决了低于 2 比特的 PTQ 中因存储非结构化掩码而产生的巨大元数据开销问题,并进一步引入低秩自适应微调,将权重分布调整为更易于量化的模式,从根本上降低了量化难度。

VPTQ[44]算法标志着后训练向量量化方法的兴起。该工作基于海森矩阵将权重分组为向量,利用权重间的相关性实现比标量量化更高的压缩效率,并引入对残差和离群点的多阶段量化策略进一步提升精度。CCQ[45]算法借鉴了通信理论中的“卷积码”思想,通过构建具有内在递推结构的码本来解决传统向量量化因解码时需要查表而导致的推理效率低下问题,成功地结合了向量量化的高精度优势与标量量化的高效率优势。BTC-LLM[46]则提出了一个高效的后训练亚 1 比特量化框架,通过挖掘二值化权重向量中的模式冗余来实现压缩,且不依赖于特定的稀疏硬件,最终实现了远低于 1 比特的平均每权重比特率。

2.2.2 QAT

尽管 PTQ 因其便捷性备受青睐,但量化感知训练通过在训练过程中模拟量化效应,为模型提供了适应低精度表示的机会,从而在理论上证明对性能悬崖问题的解决思路更接近本质。该范式不仅催生了多种高效的微调方法,更推动了对大模型底层架构的革新。

BitNet 系列工作开启了大模型超低比特量化感知训练的时代[9,46-47]。如图 3 所示,BitNet b1.58[12]使用三值化约束的 BitLinear 层取代了传统线性层,从而将代价较大的矩阵乘法操作转变为高效的加减法。通过量化感知训练范式和直通估计器进行端到端训练,BitNet b1.58 在同等规模下实现了与 FP16 LLaMA 模型相媲美的性能,这不仅提出了一条新的高能效缩放定律,也催生了为超低比特量化模型设计专用硬件的新计算范式。

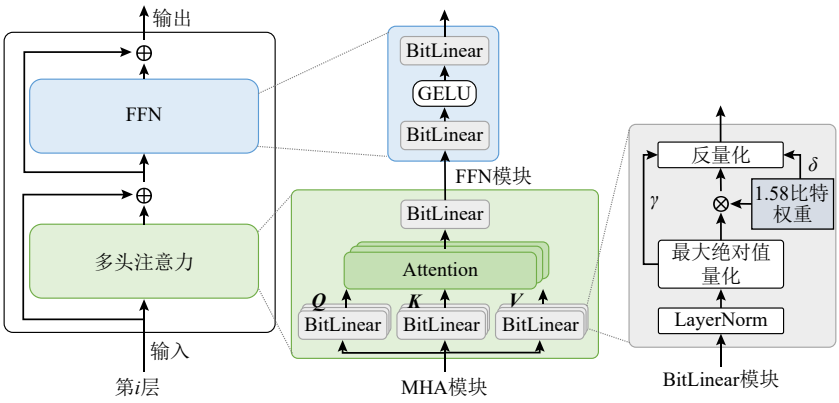
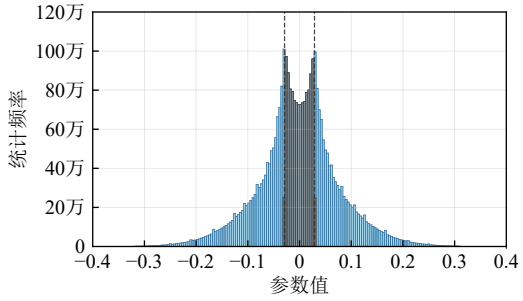


Fig. 3 Model architecture of BitNet b1.58

图 3 BitNet b1.58 模型架构

后续研究主要集中在如何进一步弥补 BitNet 训练带来的量化损失。LittleBit[48]在 OneBit[10]的基础上提出了基于双重符号-值独立分解(dual sign-value independent decomposition, Dual-SVID)的多尺度补偿机制,利用低秩分解的思想弥补超低比特量化损失,以在训练中稳定并恢复模型性能,最终取得了显著的压缩与性能平衡。Tequila[49]聚焦三值量化中的“死区陷阱”问题。如图 4 所示第 1 层全连接层的权重张量的数值分布,三值量化中被量化为 0 的大部分权重倾向于向舍入操作的边界靠近,从而形成较为显著的双峰分布。该工作通过对这些权重重新硬化为动态偏置,使得这些权重在前向传播中能提供连续信号,在几乎不增加推理开销的情况下,显著提升了模型的学习能力。Multi-Boolean-LLM[50]则从数据结构层面提出了一种更为彻底的布尔化方案,通



注:图中虚线的值分别为正负二分之一的缩放因子

Fig. 4 Full-precision weights distribution of pre-trained tenary model

图 4 预训练三元模型的全精度权重分布

过将权重表示为多个布尔核(Boolean kernels, BK)的线性组合，并使用专门的布尔优化器直接在布尔域{−1, 1}对参数进行更新，从而完全摆脱量化感知训练对全精度隐权重的依赖以及直通估计器带来的梯度近似误差，使得优化过程更加直接和稳定。

2.2.3 混合方法

为了在PTQ的低成本和量化感知训练的高性能之间取得平衡，许多工作探索了混合式的训练路径。这些工作通常以一个预训练好的全精度模型为起点，通过轻量级的量化感知微调或知识蒸馏等方法，以较低的代价恢复因量化而损失的性能。

OneBit[10]创新性地提出了SVID，将权重矩阵分解为一个包含高秩结构信息的1比特符号矩阵和2个高精度的值向量，从而为量化感知训练提供一个优秀的起点以加速模型收敛，并引入蒸馏操作以弥补量化带来的显著精度损失。BitDistiller[51]算法进一步引入超低比特模型对全精度模型的自蒸馏方法，并设计了一种正向相对熵和反向相对熵调和的蒸馏损失函数，在保障模型通用能力的同时提升模型的推理能力，最终以更少的训练资源达到更优的模型性能。BitNet Distillation[52]使用子层归一化(sub-layernorm, SubLN)结构稳定低精度模型的激活方差，并将超低比特量化引入继续预训练阶段，同时整合logits蒸馏和多头注意力蒸馏以恢复超低比特量化模型的精度，最终在文本分类和文本摘要任务中实现了接近全精度的性能。

STBLLM[53]和PBS^{2}P[54]则通过将硬件友好的“N：M”结构化稀疏与二值化相结合，成功将模型压缩至1比特以下。STBLLM提出了一种新的标准化重要性(standard importance, SI)度量，同时考虑权重大小和输入激活值范数来更准确地决定保留哪些权重，并允许不同层采用不同的稀疏率。PBS^{2}P则聚焦于联合优化剪枝和二值化过程，逐步协同地减少2种操作带来的误差。这些方法证明了即使在二值化之后，模型中依然存在可被结构化移除的冗余，为极限压缩开辟了新路径。

2.3 面向超低比特量化的推理加速

将大模型权重压缩至二元或三元等超低比特形式在理论上可显著降低内存占用与计算复杂度，从而使模型在边缘设备上部署成为可能。然而，由于通用硬件(如CPU和GPU)缺乏对低比特运算的原生支持，其在速度与能效上的潜在收益往往难以兑现，这是由于低比特权重需反量化后计算、内存对齐导致填充浪费、固定数据通路对位操作与查表类计算效率低下，这些因素共同造成了严重的架构与算法不匹配。为充分释放低精度模型在现有通用计算架构下的潜能，学术界与工业界逐渐致力于软件栈的深度创新与底层计算范式的重构，力求在不依赖专用硬件定制的前提下打破性能瓶颈。

要构建高效的硬件加速器，首先需对底层算法进行优化，使其具备不依赖特定硬件即可加速的特性。现有工作通过改进矩阵乘法原语[55]、编译器创新[56]结合查找表(lookup table, LUT)等方案[57-58]，在抽象栈的每一层挖掘性能提升空间，实现对模型架构本身的彻底重构。

神经网络推理的核心是计算密集型的矩阵乘法运算。研究工作[55]针对这一基础瓶颈，在离线阶段将每个三元矩阵预处理为2个二元矩阵并进行分区归类，在推理过程中复用同类模式的计算结果，从而显著降低冗余计算，使得向量-矩阵乘法的理论时间复杂度从标准的$O(n^2)$降至$O(n^2/\log n)$，在端到端推理中实现了高达5.24倍的实际加速。

除了直接改进计算核心，Ladder[56]通过编译器设计消除了新兴自定义数据类型与固定硬件支持之间的矛盾。该工作在未引入新计算单元的前提下构建了一个通用类型系统和一套硬件感知的张量变换原语，使得非标准格式的张量运算能被智能地映射到现有GPU的固定计算单元上，从而在不修改硬件的前提下，为多种低比特模型带来了显著的端到端加速。

在此基础上，超低比特推理的计算范式逐渐从传统的算术中心型运算转向基于查找表的内存中心型运算。如图5所示，查找表算法通过预先计算和存储运算结果来避免低比特操作在通用硬件上所需的额外反量化步骤，从而实现推理加速。LUT-DLA[57]采用向量量化技术将输入特征映射为索引，从而将神经网络转换为查找表表示的通用架构，实现了功耗效率和面积效率的提升。T-MAC[58]面向通用CPU平台进一步消除反量化瓶颈，通过位运算重构矩阵乘法，从而将整个推理过程转化为查表与加法操作，最终在低功耗的树莓派上也能实现实用的推理速度。LUT-TensorCore[59]则重点关注如何通过硬件架构充分发挥LUT范式的潜力。通过软件栈侧重查找表管理和硬件侧重查找表计算的合理分工，同时定义基于查找表的矩阵乘加指令集以融入现有AI软件生态，最终在功耗、性能与面积均实现了显著提升。

2.4 超低比特量化技术的扩展应用

超低比特量化的研究已不再局限于LLM的参数压缩与推理加速，而是逐步演化为面向多领域任务

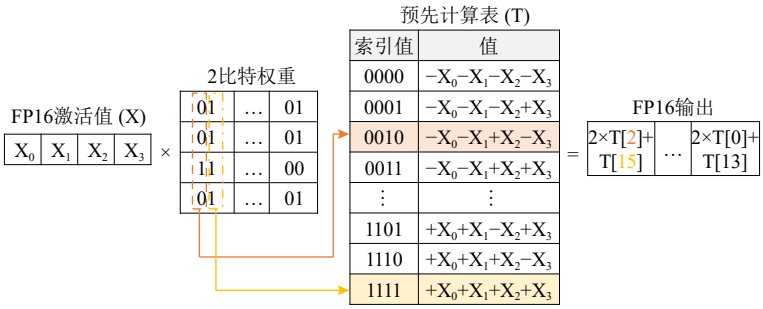


Fig. 5 Illustration of LUT algorithm

图 5 LUT 算法示意图

的定制化量化与架构协同重构。其核心目标在于，在计算与存储资源极度受限的条件下，实现模型的高效普适化部署，同时最大限度地保持甚至优化任务性能。具体而言，超低比特量化的跨领域扩展应用可以总结为如下 3 个方面。

2.4.1 应对高维模态的精度与效率的冲突

多模态和高维数据处理领域的核心冲突是复杂的特征表示需要高精度，而模型部署时对效率有极致要求。低位宽解决方案通过非均匀量化和定制层替换来应对。在视觉模型方面，DeepCompress-ViT 旨在将 Vision Transformer 引入移动设备[60]。该模型在结构上引入了全新的 BitLinear 层，以实现对线性层的彻底重构。其中，权重量化采用绝对均值量化（absmean quantization）来确定量化缩放因子 α，并将权重限制在{−1,0,1}；激活函数量化则采用最大绝对值量化（absmax quantization），将激活值缩放至固定的 8 比特内。在矩阵乘法前加入层归一化（layer normalization）进一步稳定了激活分布，使得在 1.58 比特下仍能保持图像特征提取的高效性。这一思路在多模态任务中得到进一步扩展。代表性工作如 LLaVaOLMoBitNet1B[61]，将 1 比特或三值量化架构嵌入视觉-语言模型中，通过深度压缩跨模态 Transformer 结构中的线性投影与注意力机制，大幅降低了跨模态推理的带宽与延迟需求。这些研究表明，即便面对复杂的语义关联任务，BitNet 衍生架构依然能在极低精度下保持竞争性能，为资源受限设备上的多模态智能奠定了基础。

2.4.2 缓解序列模型的量化失真

语音处理领域要求极低的延迟和极高的压缩率，但极低位宽量化极易造成序列建模关键信息的失真。因此，研究重点转向了协同训练与参数分解等增强鲁棒性的策略。针对自动语音识别（automatic speech recognition, ASR）中的 Conformer 模型，文献 [62] 提出了一种多精度模型协同训练框架，以在 1 比特权重下仍保持了接近全精度的性能。该框架利用 1 比特模型、2 比特模型与全精度教师模型的权重共享及知识蒸馏，构建出一种渐进式量化学习路径。特别地，引入的随机精度策略在训练过程中随机选择部分层进行二值化，相当于隐式模型集成，显著增强了 1 比特模型的泛化和鲁棒性。最终，该方法在维持词错误率无显著增加的前提下，实现了极高的压缩比。

在更复杂的语音翻译任务中，SVCD[63] 通过符号向量协同分解克服了单纯符号化造成的性能退化。其核心思想是将权重矩阵 $\boldsymbol{W}$ 近似分解为 $\boldsymbol{W}\approx sgn(\boldsymbol{W})\odot \boldsymbol{Vector}_1\cdot \boldsymbol{Vector}_2^{\mathrm{T}}$，其中 $sgn(\boldsymbol{W})$ 是 1 比特符号矩阵，$\boldsymbol{Vector}_1$ 和 $\boldsymbol{Vector}_2$ 为可学习的浮点向量，用作缩放和校准项。该设计以极少的浮点参数补偿了 1 比特化带来的信息损失，使模型在保持高压缩率的同时，依然能准确建模复杂的跨模态序列依赖。这一思路成为序列任务中低位宽定制化量化的典型范式。

2.4.3 驱动非传统计算架构的能效极限

低位宽量化与非传统计算架构的深度融合正在加速实现超低功耗和下一代边缘智能，重点在于全架构量化和核心机制的低精度化。对于新型状态空间模型 Mamba, Slender-Mamba[64] 实现了端到端的 1.58 比特全量化。该工作突破了 BitNet 仅量化主体层的局限，通过定制化的 QAT 将 1.58 比特扩展到了 Mamba 架构中参数占比较大的嵌入层（embedding layer）和投影层（projection layer）。这种全方位的架构量化显著提高了 Mamba 模型的实际参数压缩率和能效。在脉冲神经网络（spiking neural network, SNN）这一仿生计算范式中，SpQuant-SNN[65] 直接解决了 SNN 核心机制的能耗瓶颈，其针对 SNN 内部膜电位这一高精度且频繁读写的部分，实现了超低精度表示。同时，它引入了自适应可学习电位阈值机制，通过动态调整神经元的激发条件来补偿膜电位低精度化带来的误差，从而使 SNN 具备了在物联网设备上

实现超低功耗部署的能力。

2.5 挑　战

在超低比特量化技术中，模型表达能力的受限通常被认为是导致精度下降的主要原因之一。现有量化方法均存在不同程度的局限性。一方面，如 1 比特和 2 比特方案缺少关键状态“0”的表达或不具有对称性；另一方面，如 1.58 比特方案未能充分利用比特位进行信息编码，从而导致信息表达的不完整。然而，基于复数表示的 iFairy 模型的提出，解决了上述的不足，实现了精度与效率的双重优化。其核心原理与实现机制将在下文中详细介绍。

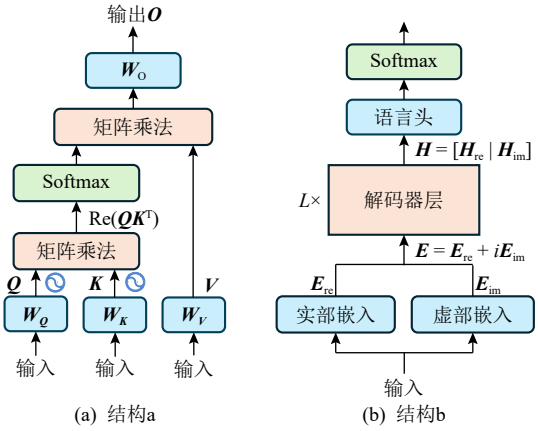


Fig. 6 Transformer model design under complex-valued domain

图 6 复数域下的 Transformer 模型设计

3 复数模型新范式

在量化研究持续向极限推进的背景下，现有实值量化模型的性能始终受制于其对应的全精度模型，即“量化模型性能=全精度模型性能−量化误差”。换言之，当前研究的主要目标集中于最小化量化误差，而默认全精度模型的性能是构成量化模型的理论上限。为突破这一固有限制，iFairy 提出了一种复数模型范式，通过系统性地将 Transformer 的关键组件扩展至复数域，利用复数表示中固有的幅度与相位双自由度，在不增加参数数量的前提下显著增强线性变换的建模能力，从而使量化后的模型具备超越传统全精度模型上限的潜力。该方法的核心思想并非仅在于抑制量化误差，而在于主动提升全精度模型的表达能力与精度表现，为极低比特量化的高效实现奠定理论与结构基础。

3.1 iFairy: 复数域下的极低比特量化模型

3.1.1 复数域下的 Transformer 模型设计

为实现复数域下的大型语言模型，iFairy 从底层架构上对传统 Transformer 进行了系统性扩展，构建了全复数化的 Transformer 神经网络架构。该架构的核心思想是：在保持原有 Transformer 整体结构不变的前提下，将所有关键线性层替换为复数线性层，使模型能够在复平面上进行幅度与相位的联合建模，同时针对其他核心组件设计进行复数化适配与改造。图 6 展示了复数域下的 Transformer 改造的核心创新点，包括复数线性层、Hermitian 注意力机制、复数域下的旋转位置编码、双路嵌入和语言建模头。

1）复数线性层。该结构是复数线性变换的基础单元。iFairy 首先将所有实值线性层替换为复数线性层（ComplexLinear），其输入、权重与输出均为复数。具体而言，一个权重矩阵 $\boldsymbol{W}\in\mathbb{C}^{m\times n}$ 及其输入 $\boldsymbol{x}\in\mathbb{C}^{m}$ 中的元素均为复数，可由其实部与虚部分别表示为输入向量 $\boldsymbol{x}=\boldsymbol{x}_{\mathrm{re}}+\mathrm{i}\boldsymbol{x}_{\mathrm{im}}$、权重矩阵 $\boldsymbol{W}=\boldsymbol{W}_{\mathrm{re}}+\mathrm{i}\boldsymbol{W}_{\mathrm{im}}$，线性映射定义为 $\boldsymbol{Y}=\bar{\boldsymbol{x}}\boldsymbol{W}$，其中乘法采用了 Hermitian 内积的语义。在复数域下，Hermitian 内积具有对称性、线性性和正定性，在复空间中是良定义的。ComplexLinear 层是复数 Transformer 的基本构件。

2）Hermitian 注意力机制。在自注意力机制中，iFairy 采用 Hermitian 内积的实部来计算查询（$\boldsymbol{Q}$）与键（$\boldsymbol{K}$）之间的相似度：

$$S=Re(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})=\boldsymbol{Q}_{\mathrm{re}}\boldsymbol{K}_{\mathrm{re}}^{\mathrm{T}}+\boldsymbol{Q}_{\mathrm{im}}\boldsymbol{K}_{\mathrm{im}}^{\mathrm{T}}\ 。$$

这一设计同时考虑了实部与虚部的交互，既保持了注意力分数的实值性，可直接用于后续的 Softmax 计算操作，又保留了复空间的结构信息。该形式具有 3 个方面的理论合理性：首先，该机制使用的核函数具有正定性，对于任意 2 个复数向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$，其 Hermitian 内积的实部等价于将复向量的实部和虚部拼接之后嵌入到实向量空间后的标准内积，保证了 Gram 矩阵的正半定性与数值稳定性；其次，该机制具有几何一致性，Hermitian 内积的实部对应于复向量的欧几里得夹角，保留了方向信息与可解释的几何结构；最后，该机制在实现效率上具有优势，通过将实部与虚部拼接为实矩阵，可直接复用现有的为实数 Transformer 设计的高效 GPU 内核，例如 FlashAttention 等，实现与实数模型同等级的计算效率。

3）基于相位旋转的位置编码（Complex RoPE）。旋转位置编码[66]通过二维旋转矩阵实现特征维度间的相对位置信息，而在复数域中，这一旋转可以被自然地表达为复数的相位旋转。对于位置 m,n 的查询与键，其编码定义为

$$q'_m = q_m e^{\theta}_{im},\ k'_n = k_n e^{\theta}_{in},$$

其中 θ 为旋转角度，从而注意力分数满足

$$score(q_m, k_n) = \mathrm{Re}(\langle q_m, k_n \rangle e^{\mathrm{i}(n-m)\theta}) 。$$

这意味着位置 m 和 n 的查询和键的注意力得分仅依赖于相对位移 $n-m$，实现了相对位置编码。与实值 RoPE 需要成对维度模拟旋转不同，Complex RoPE 在复空间中对每一维均施加统一的相位变换，数学上更为简洁，旋转语义更为自然，工程实现更为简单。

4）双路嵌入与语言建模头。为了在离散的 token 空间与连续的复数表示空间之间建立桥梁，iFairy 采用了一种双路嵌入方法，其通过 2 个并行的嵌入层实现。对于给定的输入 token，一个嵌入层生成嵌入向量的实部 $\boldsymbol{E}_{\mathrm{re}}$，另一个嵌入层生成虚部 $\boldsymbol{E}_{\mathrm{im}}$，从而构成最终的复数嵌入表示：$\boldsymbol{E} = \boldsymbol{E}_{\mathrm{re}} + \mathrm{i}\boldsymbol{E}_{\mathrm{im}}$。尽管这 2 个通道在结构上是独立的，但它们在后续的复数运算中被联合学习优化，促使模型在端到端训练过程中学习统一的复数表示。

为了将最终的复数隐藏状态投影回词汇空间以进行 token 预测，iFairy 设计了与输入嵌入层对称的语言模型输出头。最终隐藏状态 $\boldsymbol{H} = \boldsymbol{H}_{\mathrm{r}} + \mathrm{i}\boldsymbol{H}_{\mathrm{i}}$ 首先通过将其实部与虚部拼接为一个更大的实值矩阵来进行转换，即 $\tilde{\boldsymbol{H}} = [\boldsymbol{H}_{\mathrm{r}} | \boldsymbol{H}_{\mathrm{i}}]$。随后，该矩阵通过一个标准的实值线性层进行投影，从而计算出在词汇表上的最终 $logits = \tilde{\boldsymbol{H}}\boldsymbol{W}_{\mathrm{out}}$，其中 $\boldsymbol{W}_{\mathrm{out}}$ 是输出投影层的可学习权重矩阵。这种结构利用了最终隐藏状态的实部激活与虚部激活，同时确保输出概率仍可通过标准 Softmax 计算。

5）复数前馈网络与激活函数。在位置前馈网络（feed-forward network, FFN）部分，iFairy 采用 ReLU^2[67] 取代传统 ReLU。对于复输入 $\boldsymbol{Z} = \boldsymbol{Z}_{\mathrm{re}} + \mathrm{i}\boldsymbol{Z}_{\mathrm{im}}$，定义 $f(\boldsymbol{Z}) = ReLU^2(\boldsymbol{Z}_{\mathrm{re}}) + \mathrm{i}ReLU^2(\boldsymbol{Z}_{\mathrm{im}})$。ReLU^2 保留了 ReLU 的稀疏性特征，同时其导数在原点连续，可显著提升优化的稳定性；相较于依赖象限判定的复数域下的激活函数，$ReLU^2$ 不会引入相位不连续性，避免了复平面上的信息丢失，从而在大规模训练中表现出更平滑的收敛特性。

6）iFairy 总体架构。如图 7 所示，将上述复数组件系统化集成于 Transformer 主干即可完成复数化改造。图中所示的矩阵乘法均为复数线性层。整体模型在语义表达上更为丰富，同时保持 Transformer 结构上的对称性与复数语义的兼容性。

3.1.2 PhaseQuant：复数权重的 2 比特量化机制

在复数 Transformer 主干的基础上，iFairy 提出了

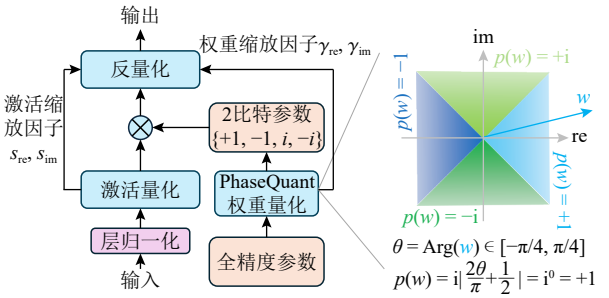


Fig. 7 Complex linear layer using PhaseQuant algorithm

图 7 应用 PhaseQuant 算法的复数线性层

首个专门针对复数权重设计的 PhaseQuant 量化框架。传统的极低比特量化方法主要针对实值模型设计，在对称性、稀疏性与表达能力之间往往存在权衡，难以充分发挥复数域的潜在优势。PhaseQuant 通过在复平面上直接进行相位投影，建立起一种对称、信息论最优的两比特量化方案，弥补了这一空白。图 7 展示了应用了 PhaseQuant 量化算法的复数线性层的推理流程以及 PhaseQuant 算法的基本原理。

1）从三值实域到四元复域的拓展。在实值量化中，BitNet 及其后续方法通常使用三元集合 $\{-1, 0, +1\}$ 来编码权重，虽然可在平均 1.58 比特的存储代价下实现近似 2 比特的压缩，但该集合在几何结构上是不对称的，且未充分利用 2 比特量化的表达能力。PhaseQuant 则在复数平面上构造了新的四元集合：$\{\pm 1, \pm \mathrm{i}\}$，即单位圆上的四次单位根。该集合在几何上等角分布且在统计上完全对称，从信息论角度看可实现 $\mathrm{lb}4 = 2$ 比特的最大熵，是真正意义上的 2 比特编码。这一设计既避免了实域量化的非对称性，又充分利用了复空间的相位信息。

2）基于相位的二比特量化算法。PhaseQuant 的量化过程以权重的相位角为依据进行确定性划分。设复数权重 $w = w_{\mathrm{re}} + \mathrm{i}w_{\mathrm{im}}$，其相位为 $\theta = \mathrm{Arg}(w) \in [-\pi, \pi]$。PhaseQuant 定义投影算子：

$$P(w) = \mathrm{i}^{\left\lfloor \frac{2\theta}{\pi} + \frac{1}{2} \right\rfloor},$$

即将复平面分为 4 个大小均等的扇区，每个扇区分别对应+1, −1, +i, −i 之一，图 7 是 PhaseQuant 算法的 1 个例子。量化后的权重可写为

$$w_{\mathrm{b}} = P(w) = w_{\mathrm{b,re}} + \mathrm{i}w_{\mathrm{b,im}} 。$$

在前向传播中，为恢复幅值信息，PhaseQuant 会分别计算实部与虚部的缩放因子：

$$\gamma_{\mathrm{re}} = \frac{1}{\mathbb{E}[|\boldsymbol{W}_{\mathrm{re}}| | \boldsymbol{P}(\boldsymbol{W}) \in \{\pm 1\}]},$$

$$\gamma_{\mathrm{im}} = \frac{1}{\mathbb{E}[|\boldsymbol{W}_{\mathrm{im}}| | \boldsymbol{P}(\boldsymbol{W}) \in \{\pm 1\}]} 。$$

从而得到反量化形式：

$$w_{\mathrm{q}}=\frac{w_{\mathrm{b,re}}}{\gamma_{\mathrm{re}}}+\mathrm{i}\frac{w_{\mathrm{b,im}}}{\gamma_{\mathrm{im}}}。$$

反向传播中采用直通估计器，使梯度能够直接传递至原始全精度权重，避免非连续投影带来的优化阻塞。值得注意的是，在 iFairy 的原始实现中，在对权重进行 PhaseQuant 量化之外，还对激活的实部和虚部分别进行了 Int8 量化。

3）无乘法推理设计。PhaseQuant 的一个独特性质是：量化后每个权重的实部或虚部均为零。因此，在复数乘法中，仅需进行符号取反或实虚部交换即可完成运算，而无需真实乘法操作。由此，PhaseQuant 保证在推理阶段可通过仅使用加法与实部虚部交换完成复数线性变换，实现完全的无乘法计算。这一特性在保持相位结构完备的同时，极大地降低了计算复杂度与能耗，使得 2 比特复数模型在硬件上具备与 1.58 比特实值模型相当甚至更高的效率。

4）PhaseQuant 的范式意义。从信息论角度来看，PhaseQuant 的四元码字在熵意义上达到 2 比特的理论上限；从几何角度看，{±1,±i}在单位圆上构成完美对称的正方形，具有最大角分离度，从而提升量化稳定性与抗噪声鲁棒性。此外，该方案与复数网络的表达机制天然契合，即复数的相位部分主要编码方向与相关性，而幅度反映信号强度。PhaseQuant 通过独立缩放实、虚 2 部分保留幅值分布，使量化聚焦于相位信息，从而在 Int2 条件下仍能精确刻画复杂的线性变换。PhaseQuant 不仅在精度上优于实域三值方案，更重要的是，它开创了复数神经网络的极低比特新范式。

3.1.3 案例研究——复域量化 iFairy

基于复数域下的 Transformer 模型设计和 PhaseQuant 2 比特复数量化算法，iFairy 实现了第一个复数域下的 Int2 极低比特量化模型 iFairy，其模型架构如图 8 所示。为验证 iFairy 框架在极低比特条件下的有效性，研究团队在预训练与零样本评测 2 个层面开展了系统实验。实验结果表明，iFairy 在语言建模能力和下游任务表现中可以实现甚至超越全精度模型的性能。PhaseQuant 量化算法可以保证均匀稳定的量化后参数分布，并且保证稳定的模型训练。

如表 2 所示，在语言建模能力方面，基于 WikiText2[68] 与 C4[69] 数据集对 iFairy 和对比模型进行了评测。结果显示，iFairy 在 7×10^8 参数规模下的平均困惑度分别为 11.13，在 1.3×10^9 参数下进一步降低至 10.05，均显著优于 BitNet b1.58，甚至在 1.3×10^9 规模上超越了全精度 LLaMA 模型。在零样本（zero-shot）推理任务中，iFairy 在多个常识推理基准（ARC-Easy[70]、ARC-Challenge[71]、HellaSwag[72]、BoolQ[73]、OpenBookQA[74]、PIQA[75]、Winogrande[76]）进行评测。表 2 展示了不同模型的平均得分。iFairy 1.3B 模型的平均准确率达到 46.52%，略高于 BF16 LLaMA（46.32%），并显著超过所有三值量化模型。这一结果首次表明：在相同计算与存储预算下，极低比特模型可以实现甚至超越全精度模型的性能。

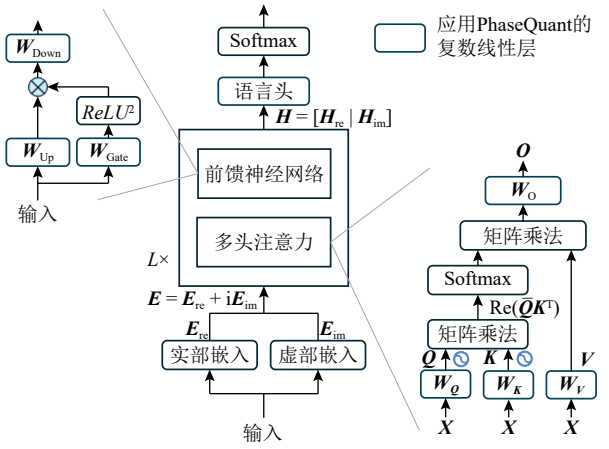


Fig. 8 Model architecture of iFairy

图 8 iFairy 模型架构

Table 2 Evaluation Results of Language Modeling Capability and Downstream Tasks

表 2 语言建模能力和下游任务评测结果 %

模型	量化	平均 PPL↓	平均准确率↑
BF16 LLaMA 700M	否	11.02	44.61
iFairy○ 700M	否	**10.08**	**46.18**
BitNet b1.58★ 700M	是	12.87	44.26
BitNet b1.58† 700M	是	11.51	44.23
iFairy 700M	是	**11.13**	**44.70**
BF16 LLaMA 1.3B	否	10.21	46.32
iFairy○ 1.3B	否	**9.34**	**48.23**
BitNet b1.58★ 1.3B	是	11.29	45.39
BitNet b1.58† 1.3B	是	10.32	46.05
iFairy 1.3B	是	**10.05**	**46.52**

注：PPL 指困惑度；“↑”表示越高越好，“↓”表示越低越好；“○”表示没有 PhaseQuant 量化的 iFairy；“★”表示 BitNet 原始论文汇报的结果；“†”表示我们复现的 BitNet 的结果。黑体数值表示在同类算法（模型参数量、是否量化）中表现最好。

3.2 Fairy2i: 实数模型复数化的极低比特量化方法

iFairy 模型验证了复数域下的 Transformer 模型在极低比特量化上的潜力，打破了量化模型性能受限于全精度模型的传统认知。但它与其他基于 QAT

的极低比特量化方法类似,都需要从零开始进行量化感知训练。对于动辄千亿甚至万亿参数的大模型而言,从头训练的计算和时间成本是天文数字,这极大地阻碍了此类新架构的实际应用。为此,进一提出了 Fairy2i 框架,其目标是利用已有海量的全精度实数预训练模型,在不从零开始训练的前提下,将其高效转换为 2 比特的复域量化模型,从而兼顾极低的开发和部署成本以及复数域带来的高精度优势。利用 Fairy2i 框架,从任意实数预训练模型出发,经复数化转化、相位量化感知训练后即可获得高效可部署的极低比特量化模型,为后续的复数神经网络低比特化提供了完整路径。

3.2.1 整体思路

Fairy2i 的核心目标是在不重新训练全模型的前提下,将实数预训练模型映射到复数域,使得其线性层可被等价替换为复数线性层,并在复数权重上实施基于相位的 PhaseQuant。类比于“时域信号通过傅立叶变换进入频域,再经由低通滤波实现稳定有效表示”的递进处理范式,Fairy2i 引导实数权重矩阵经由复域变换等价映射至复数域,随后通过复域量化滤除数值冗余并锁定核心特征,最终实现稳定无乘法运算的高效推理。这样,模型参数被约束在复平面四次单位根集合$\{\pm1, \pm\mathrm{i}\}$,从而在保持模型表达能力的同时实现 Int2 极低比特表示。整个流程包含 2 个核心步骤,分别是实数模型的复数化等价转换和基于相位的量化感知训练。

3.2.2 实数模型的复数化等价转换

如图 9 所示,在复数化步骤中,考虑任意实数线性映射的参数矩阵$\boldsymbol{A}\in\mathbb{R}^{2n\times 2m}$(假设参数矩阵$\boldsymbol{A}$的维度均为偶数,如遇奇数,可以填充为偶数),首先将其成对重排成 4 块,记

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{pmatrix},\ \boldsymbol{A}_{ij}\in\mathbb{R}^{n\times m}。$$

把实数输入$\tilde{\boldsymbol{x}}=[\boldsymbol{x}_{\mathrm{re}};\boldsymbol{x}_{\mathrm{im}}]\in \mathrm{x}^{2n}$视为复向量$\boldsymbol{x}=\boldsymbol{x}_{\mathrm{re}}+\mathrm{i}\cdot\boldsymbol{x}_{\mathrm{im}}\in\mathbb{C}^n$,Fairy2i 以 2 组复权重$\boldsymbol{W}=\boldsymbol{W}_{\mathrm{re}}+\mathrm{i}\boldsymbol{W}_{\mathrm{im}}\in\mathbb{C}^{n\times m}$,$\boldsymbol{U}=\boldsymbol{U}_{\mathrm{re}}+\mathrm{i}\boldsymbol{U}_{\mathrm{im}}\in\mathbb{C}^{n\times m}$定义复数线性层的前向为$\boldsymbol{Y}=\tilde{\boldsymbol{x}}\boldsymbol{W}+\boldsymbol{x}\boldsymbol{U}$,其中第 1 项是反线性支路(共轭支路),第 2 项是复线性支路(非共轭支路),两者构成一对并联通路。将上式在实域等价展开,可对应到 2 个 2×2 的实块矩阵

$$\boldsymbol{B}_1(\boldsymbol{W})=\begin{pmatrix}\boldsymbol{W}_{\mathrm{re}} & \boldsymbol{W}_{\mathrm{im}}\\ \boldsymbol{W}_{\mathrm{im}} & -\boldsymbol{W}_{\mathrm{re}}\end{pmatrix},\ \boldsymbol{B}_2(\boldsymbol{U})=\begin{pmatrix}\boldsymbol{U}_{\mathrm{re}} & -\boldsymbol{U}_{\mathrm{im}}\\ \boldsymbol{U}_{\mathrm{im}} & \boldsymbol{U}_{\mathrm{re}}\end{pmatrix}。$$

若满足精确恒等式$\boldsymbol{A}=\boldsymbol{B}_1(\boldsymbol{W})+\boldsymbol{B}_2(\boldsymbol{U})$,即可保证以复数线性层替换原实数线性层时功能完全等价。

Fairy2i 给出一组复权重$\boldsymbol{W}$和$\boldsymbol{U}$的初始化方式

$$\boldsymbol{W}_{\mathrm{re}}=\frac{1}{2}(\boldsymbol{A}_{11}-\boldsymbol{A}_{22}),\ \boldsymbol{W}_{\mathrm{im}}=\frac{1}{2}(\boldsymbol{A}_{12}+\boldsymbol{A}_{21});$$

$$\boldsymbol{U}_{\mathrm{re}}=\frac{1}{2}(\boldsymbol{A}_{11}+\boldsymbol{A}_{22}),\ \boldsymbol{U}_{\mathrm{im}}=\frac{1}{2}(\boldsymbol{A}_{21}-\boldsymbol{A}_{12})。$$

代入可逐项验证$\boldsymbol{A}=\boldsymbol{B}_1(\boldsymbol{W})+\boldsymbol{B}_2(\boldsymbol{U})$成立。由此,任何实数模型的线性层(含多头注意力内的$\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$, $\boldsymbol{O}$映射与前馈网络的 2 层线性变换)均可无缝复数化为$\boldsymbol{y}=\tilde{\boldsymbol{x}}\boldsymbol{W}+\boldsymbol{x}\boldsymbol{U}$的形式,且参数规模保持与原层相同。

3.2.3 基于相位的量化感知训练

在量化阶段,Fairy2i 首先对复权重$\boldsymbol{W}$和$\boldsymbol{U}$采用相位量化(PhaseQuant),将每个复参数的相位映射到四次单位根集合$\{\pm1,\pm\mathrm{i}\}$。为弥补量化带来的感知误差,Fairy2i 提出在量化之后进行量化感知训练,与 QAT 方法类似,Fairy2i 对量化后的模型在少量预训练语料上进行训练,在训练时进行量化和反量化操作,通过直通估计器方式对原始权重进行优化,通过量化

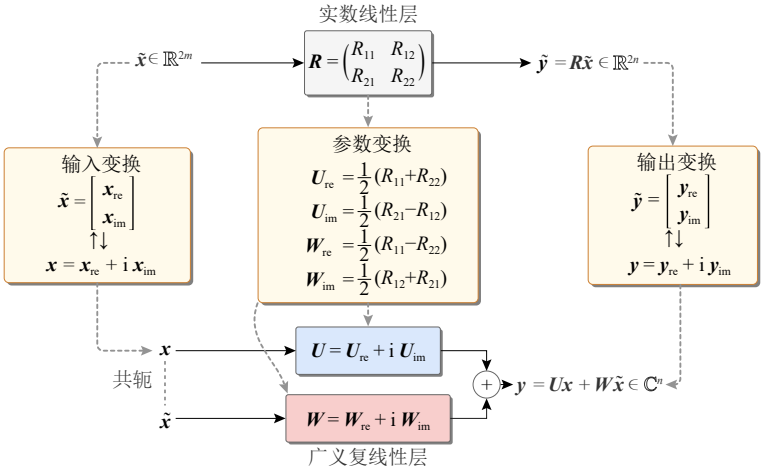


Fig. 9 Complexification equivalent transformation of Fairy2i real number model

图 9 Fairy2i 实数模型的复数化等价转换

感知的训练让模型快速学习和适应相位量化带来的的计算模式的改变。由于每一个权重占用空间位 2 比特，通过计算可知，Fairy2i 所需的最终码字存储所需的比特数量与原始矩阵$\boldsymbol{A}$的参数量相同，因此 Fairy2i 可以视为 1 比特量化算法。如表 3 所示，实验结果验证了该框架的有效性：在 LLaMA2-7B 模型上，Fairy2i 在平均困惑度和下游任务平均得分上均显著优于 BitNet 1 比特量化方案，证明了通过复数化转换与相位量化感知训练，能够使实数预训练模型在极低比特下焕发更高的性能潜力。

Table 3 Experimental Results of Quantitative Perception Training of Real Pre-Training Model

表 3 实数预训练模型的量化感知训练实验结果 %

原始模型	量化方法	平均 PPL↓	平均准确率↑
LLaMA2-7B	BitNet 1bit	11.75	46.21
	Fairy2i	**11.03**	**48.66**

注：PPL 指困惑度；“↑”表示越高越好，“↓”表示越低越好。黑体数值表示在同类算法（模型参数量、是否量化）中表现最好。

3.3 小 结

综合来看，iFairy 在 Int2 条件下兼具高精度、强收敛性与几何稳定性，打破了传统“量化精度受制于全精度模型”的长期假设。在性能层面，iFairy 在语言建模与下游任务中均超越现有最优三值方案，部分指标甚至超过全精度模型；在结构层面，复数 Transformer 保持了幅度分布的稳定与相位编码的均衡；在理论层面，PhaseQuant 实现了复数域的对称量化，使表达能力与效率同时达到信息论最优。Fairy2i 基于 iFairy 所提出的核心创新点，进一步提出了一种实数模型复数化框架，通过将实数模型等价变换到复数域并进行量化感知训练，均衡了复用已有实数模型的成本优势与复数域带来的精度优势。

这些结果共同验证了复数域量化的可行性与优越性：极低比特模型并非精度受限的折衷方案，而可成为性能与能效兼得的新范式。

4 展 望

随着人工智能模型规模的持续扩大，计算与存储成本正成为制约其广泛应用的重要瓶颈。低比特量化作为降低模型复杂度和能耗的关键手段，已取得显著进展，但主流方法仍局限于实数域。在极低比特条件下，实数域量化常面临精度急剧下降与训练不稳定等问题。复域量化作为一种超越传统实数框架的新兴范式，正展现出独特的潜力。它不仅可能推动量化精度进一步降低，还为构建高效、节能和可扩展的人工智能系统提供了新的研究方向。

复域量化的核心优势源于复数域的信息表达特性。通过将数据表示为幅度与相位 2 个正交维度，模型能够在极低比特条件下保持较高的信息密度与结构表达能力。这种双自由度表示机制使复域量化在相同比特宽度下比实数量化具备更强的表示力和鲁棒性。例如，部分研究工作表明，仅在相位空间中进行离散化即可在极低比特下维持较高的模型性能，有效缓解传统量化方法中信息坍塌的问题。此外，复数表示与 Transformer 等高度非线性模型结构具有良好的适配性，能够在低精度条件下保持语义一致性和特征稳定性，为进一步降低量化比特提供理论支撑。

要将这一理论潜力转化为实际性能，仍需在算法与系统层面进行深入的协同优化。目前主流计算硬件仍以实数运算为中心，对复数操作的支持相对有限，这在一定程度上制约了复域量化在推理效率和能耗控制方面的潜力。未来的研究可在不依赖大幅硬件改造的前提下，探索以算法结构重构、运算图优化或算子近似替代等方式提升复域量化的计算效率，同时在通用硬件架构上研究轻量化的复数算子映射策略，使复域量化能够平滑融入现有人工智能生态体系。值得注意的是，复域量化的理论框架与新兴计算范式具有天然关联。量子计算中的基本单位量子比特以复数幅度和相位共同描述其状态，其信息处理机制与复数神经网络的表示形式高度契合。这种一致性为未来量子加速器与复域量化模型的协同演化提供了新的研究空间。随着量子硬件的逐步成熟，复数形式的计算可能获得原生支持，从而进一步释放其在能效与性能方面的潜能。

总体而言，复域量化不仅是量化技术的一种扩展，更代表着从表示学习到计算体系的系统性革新。通过在复数域中重构信息编码方式，模型可在极限压缩下保持性能稳定；通过算法与系统层面的协同优化，可推动高能效人工智能的发展；同时，与量子计算等新兴方向的潜在融合，也为智能计算的长期演进提供了新的理论支点。未来，复域量化有望成为实现高效、可持续和普适人工智能的重要基础技术。

5 总 结

本文通过系统梳理低比特量化技术的演进逻辑，

指出当压缩逼近 1~2 比特极限时，传统实数域方案已触及信息表达的理论下界，单纯的误差优化已难以消除离散化带来的性能天花板。相比之下，复域量化展现了更为本质的突破路径，通过引入幅度与相位双重自由度，在不增加参数规模的基础上实现了模型表达能力的“升维”扩展。这种从实数域向复数域的视角转换，不仅在算法层面解决了极低比特下的信息匮乏难题，更在理论层面证伪了量化必然导致精度衰减的传统认知。可以预见，未来的高效智能模型构建应当超越对全精度的简单逼近，转向基于复数计算体系的架构创新，从而为突破现有计算架构限制、实现计算能效与智能水平的双重跃升提供了新的演进方向。

作者贡献声明：丛培壮提出了论文整体框架；王飞宇、王国安、王砚舒撰写了第 2~5 节内容；郑策撰写了第 1 节内容；杨仝提出了复域量化的新范式，提出了 iFairy 和 Fairy2i 的具体算法创新，完善了论文写作思路并修改论文。

参 考 文 献

[1] Chen Huimin, Liu Zhiyuan, Sun Maosong. The social opportunities and challenges in the era of large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1094−1103
(陈慧敏, 刘知远, 孙茂松. 大语言模型时代的社会机遇与挑战[J]. 计算机研究与发展, 2024, 61(5): 1094−1103)

[2] Kimi Team. Kimi k2: Qpen agentic intelligence[J]. arXiv preprint arXiv: 2507.20534, 2025

[3] Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 2704−2713

[4] TensorFlow Lite Team. Tensorflow lite: An open source deep learning framework for on-device inference[EB/OL]. Google AI Blog, 2017[2025-10-01]. https://blog.tensorflow.org/2017/11/announcing-tensorflow-lite.html

[5] NVIDIA. TensorRT Developer Guide[M/OL]. NVIDIA Corporation, 2018[2025-10-01]. https://docs.nvidia.com/deeplearning/tensorrt/

[6] Frantar E, Ashkboos S, Hoefler T, et. al. GPTQ: Accurate post-training quantization for generative pre-trained transformers[C]//Proc of the 11th Int Conf on Learning Representations (ICLR). New York: Curran Associates, 2023

[7] Lin Ji, Tang Jiaming, Tang Haotian, et al. AWQ: Activation-aware weight quantization for LLM compression and acceleration[C]//Proc of Machine Learning and Systems 6 (MLSys) New York: Curran Associates, 2024, 6: 87−100

[8] Xiao Guangxuan, Lin Ji, Seznec M, Wu H, et al. SmoothQuant: Accurate and efficient post-training quantization for large language models[C]//Proc of the 40th Int Conf on Machine Learning (ICML). New York: ACM, 2023: 38087−38099

[9] Wang Hongyu, Ma Shuming, Dong Li, et al. BitNet: Scaling 1-bit transformers for large language models[J]. arXiv preprint arXiv: 2310.11453, 2023

[10] Xu Yuzhuang, Han Xu, Yang Zonghan, et al. OneBit: Towards extremely low-bit large language models[C]//Proc of the 38th Int Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2024, 37: 66357−66382

[11] Huang Wei, Liu Yangdong, Qin Haotong, et al. BiLLM: Pushing the limit of post-training quantization for LLMs[C]//Proc of the 41st Int Conf on Machine Learning (ICML). New York: PMLR, 2024: 19889−19908

[12] Ma Shuming, Wang Hongyu, Ma Lingxiao, et al. The era of 1-bit LLMs: All large language models are in 1.58 bits[J]. arXiv preprint arXiv: 2402.17764, 2024

[13] Li Fengfu, Liu Bin, Wang Xiaoxing. Ternary weight networks[J]. arXiv Preprint, arXiv: 1605.04711, 2016

[14] Zhu Feng, Gong Ruihao, Yu Fengwei, et al. Towards unified int8 training for convolutional neural network[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 1969−1979

[15] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[C]//Proc of the 36th Int Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2022: 30016−30030

[16] Trabelsi C, Bilaniuk O, Zhang Y, et al. Deep complex networks[C]//Proc of the 6th Int Conf on Learning Representations (ICLR). New York: Curran Associates, 2018: 3570−3601

[17] Wang Feiyu, Wang Guoan, Zhang Yihao, et al. iFairy: The first 2-bit complex LLM with all parameters in {±1, ±i}[J]. arXiv preprint arXiv: 2508.05571, 2025

[18] Wang Feiyu, Tan Xinyu, Huang Bokai, et al. Fairy2i: Training complex LLMs from real LLMs with all parameters in {±1, ±i}[J]. arXiv preprint arXiv: 2512.02901, 2025

[19] Dettmers T, Lewis M, Belkada Y, et al. GPT3.int8(): 8-bit matrix multiplication for transformers at scale[C]//Proc of the 36th Int Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2022, 35: 30318–30332

[20] Banner R, Nahshan Y, Hoffer E, et. al. Post-training 4-bit quantization of convolutional networks for rapid-deployment[C]//Proc of the 33rd Int Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2019: 1149–1159

[21] Yao Zhewei, Yazdani Aminabadi R, Zhang Minjia, et al. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers[C]//Proc of the 36th Int Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2022, 35: 27168−27183

[22] Chee J, Cai Y, Kuleshov V, et al. QuIP: 2-bit quantization of large language models with guarantees[C]//Proc of the 37th Int Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2023, 36: 4396−4429

[23] Tseng A, Chee J, Sun Q, et al. QuIP# even better LLM quantization with hadamard incoherence and lattice codebooks[C]//Proc of the 41st Int Conf on Machine Learning (ICML). New York: ACM, 2024:

48630−48656

[24] Tseng A, Sun Q, Hou D, et al. QTIP: Quantization with trellises and incoherence processing[C]//Proc of the 38th Int Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2024: 59597−59620

[25] Tseng A, Sun Z, De Sa C. Model-preserving adaptive rounding[J]. arXiv preprint arXiv: 2505.22988, 2025

[26] Zhou Shuchang, Wu Yuxin, Ni Zekun, et al. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients[J]. arXiv preprint arXiv: 1606.06160, 2016

[27] Dong Zhen, Yao Zhewei, Gholami A, et al. HAWQ: Hessian aware quantization of neural networks with mixed-precision[C]//Proc of the IEEE/CVF Int Conf on Computer Vision (CVPR). Piscataway, NJ: IEEE, 2019: 293–302

[28] Esser S K, McKinstry J L, Bablani D, et al. Learned step size quantization[C]//Proc of the 8th Int Conf on Learning Representations (ICLR). New York: Curran Associates, 2020: 6910–6921

[29] Bhalgat Y, Lee J, Nagel M, et al. LSQ+: Improving low-bit quantization through learnable offsets and better initialization[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops (CVPR workshop). Piscataway, NJ: IEEE, 2020: 696−697

[30] Shen Sheng, Dong Zhen, Ye Jiayu, et al. Q-BERT: Hessian based ultra low precision quantization of BERT[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 8815−8821

[31] Frantar E, Alistarh D. Optimal brain compression: A framework for accurate post-training quantization and pruning[C]//Proc of the 36th Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2022, 35: 4475−4488

[32] Ashkboos S, Mohtashami A, Croci M L, et al. QuaRot: Outlier-free 4-bit inference in rotated LLMs[C]//Proc of the 38th Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2024

[33] Frantar E, Alistarh D. Sparsegpt: Massive language models can be accurately pruned in one-shot[C]//Proc of the 40th Int Conf on Machine Learning (ICML). New York: PMLR, 2023: 10323−10337

[34] Courbariaux M, Bengio Y, David J P. BinaryConnect: Training deep neural networks with binary weights during propagations[C]//Proc of the Advances in Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2015: 3123–3131

[35] Rastegari M, Ordonez V, Redmon J, et al. XNOR-Net: ImageNet classification using binary convolutional neural networks[C]//Proc of the of European Conf on Computer Vision (ECCV). Berlin: Springer, 2016: 525–542

[36] Yu T, Gupta G, Gopalswamy K, et al. Collage: Light-weight low-precision strategy for LLM training[C]//Proc of the 41st Int Conf on Machine Learning (ICML). New York: ACM, 2024: 57459−57479

[37] Kumar T, Ankner Z, Spector B F, et al. Scaling laws for precision[C]//Proc of the 13th Int Conf on Learning Representations (ICLR). New York: Curran Associates, 2025: 71833−71864

[38] Ji Jiaming, Wang Kaile, Qiu Tianyi, et al. Language models resist alignment: Evidence from data compression[C]//Proc of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA: ACL, 2025: 23411−23432

[39] Dettmers T, Zettlemoyer L. The case for 4-bit precision: *k*-bit inference scaling laws[C]//Proc of the 11th Int Conf on Machine Learning (ICML). New York: PMLR, 2023: 7750−7774

[40] Liu Zechun, Zhao Changsheng, Huang Hanxian, et al. ParetoQ: Scaling laws in extremely low-bit llm quantization[C]//Proc of the 39th Int Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2025

[41] Dremov A, Grangier D, Katharopoulos A, et al. Compute-optimal quantization-aware Training[J]. arXiv preprint arXiv: 2509.22935, 2025

[42] Li Yuhang, Panda P. TesseraQ: Ultra low-bit LLM post-training quantization with block reconstruction[J]. arXiv preprint arXiv: 2410.19103, 2024

[43] Zhao Jiaqi, Zhang Miao, Wang Ming, et al. PTQ 1.61: Push the real limit of extremely low-bit post-training quantization methods for large language models[J]. arXiv preprint arXiv: 2502.13179, 2025

[44] Liu Yifei, Wen Jichen, Wang Yang, et. al. VPTQ: Extreme low-bit vector post-training quantization for large language models[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2024: 8181−8196

[45] Zhou Zhaojing, Li Xunchao, Li Minghao, et al. CCQ: Convolutional code for extreme low-bit quantization in LLMs[J]. arXiv preprint arXiv: 2507.07145, 2025

[46] Gu Hao, Li Lujun, Wang Zheyu et al. BTC-LLM: Efficient sub−1-bit LLM quantization via learnable transformation and binary codebook[J]. arXiv preprint arXiv: 2506.12040, 2025

[47] Wang Hongyu, Ma Shuming, Wei Furu. Bitnet v2: Native 4-bit activations with hadamard transformation for 1-bit LLMs[J]. arXiv preprint arXiv: 2504.18415, 2025

[48] Lee B, Kim D, You Y, et al. LittleBit: Ultra low-bit quantization via latent factorization[C]//Proc of the 39th Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2025

[49] Huang Hong, Wu Decheng, Cen Rui, et al. Tequila: Trapping-free ternary quantization for large language models[J]. arXiv preprint arXiv: 2509.23809, 2025

[50] Tran B H. Ultra-efficient and effective large language models with Multi-Boolean architectures[C]//Proc of the ICML 2025 Workshop on Machine Learning for Wireless Communication and Networks (ML4Wireless). New York: PMLR, 2025: 1–31

[51] Du Dayou, Zhang Yijia, Cao Shijie, et al. BitDistiller: Unleashing the potential of sub−4-bit LLMs via self-distillation[C]//Proc of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA, ACL, 2024: 102–116

[52] Wu Xun, Huang Shaohan, Wang Wenhui, et al. BitNet Distillation[J]. arXiv preprint arXiv: 2510.13998, 2025

[53] Dong Peijie, Li Lujun, Zhong Yuedong, et al. STBLLM: Breaking the 1-bit barrier with structured binary LLMs[C]//Proc of the 13th Int Conf on Learning Representations (ICLR). New York: Curran Associates, 2025

[54] Yan Xianlong, Zhang Tianao, Li Zhiting, et al. Progressive binarization with semi-structured pruning for LLMs[J]. arXiv preprint arXiv: 2502.01705, 2025

[55] Dehghankar M, Erfanian M, Asudeh A. An efficient matrix multiplication algorithm for accelerating inference in binary and ternary neural networks[C]//Proc of the 42nd Int Conf on Machine Learning (ICML). New York: PMLR, 2025: 12969−12986

[56] Wang Lei, Ma Lingxiao, Cao Shijie, et al. Ladder: Enabling efficient low-precision deep learning computing through hardware-aware tensor transformation[C]//Proc of the 18th USENIX Symp on Operating Systems Design and Implementation (OSDI). Berkeley, CA: USENIX Association, 2024: 307−323

[57] Li Guoyu, Ye Shengyu, Chen Chunyun, et al. LUT-DLA: Lookup table as efficient extreme low-bit deep learning accelerator[C]//Proc of 2025 IEEE Int Symp on High Performance Computer Architecture (HPCA). Piscataway, NJ: IEEE, 2025: 671−684

[58] Wei Jianyu, Cao Shijie, Cao Ting, et al. T-MAC: CPU renaissance via table lookup for low-bit LLM deployment on edge[C]//Proc of the 20th European Conf on Computer Systems (Eurosys). New York: ACM, 2025: 278−292

[59] Mo Zhiwen, Wang Lei, Wei Jianyu, et al. LUT-TensorCore: A software-hardware co-design for LUT-base low-bit LLM inference[C]//Proc of the 52nd Annual Int Symp on Computer Architecture (ISCA). Piscataway, NJ: IEEE, 2025: 514−528

[60] Ahmed S, Al Arafat A, Najafi D, et al. Deepcompress-vit: Rethinking model compression to enhance efficiency of vision transformers at the edge[C]//Proc of the Computer Vision and Pattern Recognition Conf (CVPR). Piscataway, NJ: IEEE, 2025: 30147−30156

[61] Sundaram J, Iyer R. LLaVaOLMoBieNet1B: Ternary LLM goes multimodal![J]. arXiv preprint arXiv: 2408.13402, 2024

[62] Li Zhaoqing, Xu Haoning, Jin Zengrui, et al. Towards one-bit ASR: Extremely low-bit conformer quantization using co-training and stochastic precision[C]//Proc of Interspeech. Baixas, France: International Speech Communications Association, 2025: 1973−1977

[63] Morrison M, Churchwell C, Pruyne N, et al. Sign value constraint decomposition for efficient 1-bit quantization of speech translation tasks[C]//Proc of 2024 Interspeech. New York: Curran Associates, 2024: 842−846

[64] Yu Z, Kojima T, Matsuo Y, et al. Slender-Mamba: Fully quantized Mamba in 1.58 bits from head to toe[C]//Proc of the 31st Int Conf on Computational Linguistics (COLING). Stroudsburg, PA: ACL, 2025: 4715−4724

[65] Hassan A, Meng J, Anupreetham A, et al. SpQuant-SNN: Ultra-low precision membrane potential with sparse activations unlock the potential of on-device spiking neural networks applications[J]. Frontiers in Neuroscience, 2024, 18: 1440000

[66] Su Jianlin, Murtadha A, Lu Yu, et al. Roformer: Enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063

[67] Zhang Zhenyan, Song Yixin, Yu Guanghui, et al. Relu^2 wins: Discovering efficient activation functions for sparse LLMs[J]. arXiv preprint arXiv: 2402.03804, 2024

[68] Merity S, Xiong C, Bradbury J, et al. Pointer sentinel mixture models[J]. arXiv:1609.07843v1

[69] Raffel C, Shazeer N, Roberts A. et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(140): 1−67

[70] Yadav V, Bethard S, Surdeanu M. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering[C]//Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, ACL, 2019: 2578−2589

[71] Clark P, Cowhey I, Etzioni O, et al. Think you have solved question answering? try ARC, the AI2 reasoning challenge[J]. arXiv preprint arXiv: 1803.05457, 2018

[72] Zellers R, Holtzman A, Bisk Y, et al. HellaSwag: Can a machine really finish your sentence?[C]//Proc of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA: ACL, 2019: 4791−4800

[73] Clark C, Lee K, Chang M W, et al. BoolQ: Exploring the surprising difficulty of natural yes/no questions[C]//Proc of NAACL-HLT. Stroudsburg, PA: ACL, 2019: 2924−2936

[74] Mihaylov T, Clark P, Khot T, et al. Can a suit of armor conduct electricity? A new dataset for open book question answering[C]//Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 2381−2391

[75] Bisk Y, Zellers R, Bras R, et al. PIQA: Reasoning about physical commonsense in natural language[C]//Proc of the AAAI Conf on Artificial Intelligence (AAAI). Palo Alto, CA: AAAI, 2020: 7432−7439

[76] Sakaguchi K, Bras R, Bhagavatula C, et al. Winogrande: An adversarial winograd schema challenge at scale[J]. Communications of the ACM, 2021, 64(9): 99–106



Cong Peizhuang, born in 1994. PhD, postdoctoral researcher. Member of CCF. His main research interests include LLM quantization and inference optimization, and new architecture of networking.

丛培壮, 1994 年生。博士，博士后研究员。CCF 会员。主要研究方向为大语言模型量化与推理优化、新型网络架构。



Wang Feiyu, born in 1999. PhD candidate. His main research interests include LLM quantization technology and inference acceleration.

王飞宇, 1999 年生。博士研究生。主要研究方向为大语言模型量化技术、推理加速。



Wang Guo'an, born in 2004. Master candidate. His main research interests include edge intelligence and model compression algorithms.

王国安, 2004 年生。硕士研究生。主要研究方向为端侧智能、模型压缩算法。



Wang Yanshu, born in 1994. PhD, postdoctoral researcher. His main research interests include high-performance network measurement algorithms, distributed machine learning systems, and LLM quantization and inference acceleration.

王砚舒, 1994 年生。博士，博士后研究员。主要研究方向为高性能网络测量算法、分布式机器学习系统、大语言模型量化与推理加速。



Zheng Ce, born in 1994. PhD candidate. Her main research interests include network measurement algorithms and LLMs.

郑　策, 1994 年生。博士研究生。主要研究方向为网络测量算法、大语言模型。



Yang Tong, born in 1982. PhD, associate professor, PhD supervisor. His main research interests include LLM, network measurement, and new data structures.

杨　仝, 1982 年生。博士，副教授，博士生导师。主要研究方向为大语言模型、网络测量、新型数据结构。