

大语言模型低比特量化方法综述

郝志伟¹⁾ 徐冠宇¹⁾ 沈 力²⁾ 罗 勇³⁾ 曾 丹^{4),5)} 胡 晗¹⁾

¹⁾(北京理工大学信息与电子学院 北京 100081)

²⁾(中山大学深圳校区网络空间安全学院 广东 深圳 518107)

³⁾(武汉大学计算机学院 武汉 430072)

⁴⁾(杭州电子科技大学丽水研究院 浙江 丽水 323000)

⁵⁾(上海大学通信与信息工程学院 上海 200444)

摘 要 近年来,基于Transformer架构的大语言模型(Large Language Models, LLMs)在自然语言处理多个领域取得了显著进展,展现出卓越的性能。然而,随着模型参数规模的急剧扩张与计算需求的持续攀升,其推理过程对计算与存储资源的高消耗问题日益凸显,严重制约了该类模型在资源受限环境下的部署与应用。为应对上述挑战,模型量化技术应运而生。该技术通过将模型中的高精度浮点数转换为低精度表示,能够有效压缩模型存储规模、降低内存带宽需求并减少计算功耗,从而提升推理效率,为LLM在消费级硬件上的高效部署创造了条件。然而,尽管相关技术迭代迅速,目前仍缺乏对低比特量化方法的系统性梳理与总结,这在一定程度上阻碍了该领域的深入发展。为此,本文旨在对LLM的低比特量化方法展开全面而系统的综述。具体而言,本文首先阐述量化的基本概念,涵盖低精度数值表示方式、量化粒度及其对模型性能的影响;随后重点分析两大主流量化范式:训练后量化与量化感知训练。在训练后量化部分,将分别探讨多比特与单比特权重量化方法,以及权重-激活联合量化与键值缓存量化等关键技术;继而介绍量化感知训练的基本原理及其优势。最后,本文还将梳理当前支持LLM量化的主流开发框架与硬件平台,并对未来研究方向进行总结与展望,以期为该领域学术研究与实践应用的进一步发展提供参考依据。

关键词 深度学习;大语言模型;模型压缩;量化;模型推理

中图法分类号 TP183 **DOI号** 10.11897/SP.J.1016.2026.01936

A Survey of Low-Bit Quantization Methods for Large Language Models

HAO Zhi-Wei¹⁾ XU Guan-Yu¹⁾ SHEN Li²⁾ LUO Yong³⁾ ZENG Dan^{4),5)} HU Han¹⁾

¹⁾(School of Information and Electronics, Beijing Institute of Technology, Beijing 100081)

²⁾(School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong 518107)

³⁾(School of Computer Science, Wuhan University, Wuhan 430072)

⁴⁾(Lishui Institute of Hangzhou Dianzi University, Lishui, Zhejiang 323000)

⁵⁾(School of Communication and Information Engineering, Shanghai University, Shanghai 200444)

Abstract In recent years, Large Language Models (LLMs), most of which are built upon the Transformer architecture, have achieved remarkable success in a wide range of natural language processing tasks and have become a fundamental paradigm for intelligent systems. Their strong

收稿日期:2025-08-05;在线发布日期:2026-04-08。本课题得到国家自然科学基金-联合基金重点项目(U2336211)、国家自然科学基金-重大研究计划重点支持项目(92467206)、国家自然科学基金-联合基金重点项目(U23A20318)、国家自然科学基金-面上项目(62576364)、深圳市基础研究专项自然科学基金计划重点项目(JCYJ20241202124430041)和浙江省“尖兵领雁+X”研发攻关计划(2024C01024)资助。郝志伟,博士,主要研究领域为神经网络模型压缩。E-mail: haozhw@outlook.com。徐冠宇,博士研究生,主要研究领域为模型压缩与边缘智能。沈 力,博士,副教授,主要研究领域为可信人工智能、强化学习和协作学习。罗 勇,博士,教授,主要研究领域为机器学习、模式识别及其在多媒体、医学影像分析和理解等方面的应用。曾 丹(通信作者),博士,教授,主要研究领域为计算机视觉、多媒体内容分析与人工智能。E-mail: dzeng@shu.edu.cn。胡 晗,博士,教授,主要研究领域为多媒体网络、空天网络边缘智能。

capabilities in language understanding, text generation, reasoning, instruction following, and interactive applications have greatly expanded the practical scope of artificial intelligence. However, these advantages are accompanied by rapidly growing model sizes and inference costs. As the number of parameters increases from billions to tens or even hundreds of billions, LLM deployment is increasingly constrained by memory capacity, memory bandwidth, computational overhead, and energy consumption. These issues pose a major obstacle to the efficient deployment of LLMs on consumer-grade devices, edge platforms, and other resource-constrained environments. To address this challenge, low-bit quantization has emerged as one of the most effective model compression techniques. By converting high-precision floating-point representations into lower-precision numerical formats, quantization can significantly reduce storage requirements, alleviate bandwidth pressure, lower power consumption, and improve inference efficiency, thereby enabling more practical deployment of LLMs. Although recent years have witnessed rapid progress in this area, existing studies are often scattered across different algorithmic settings, evaluation protocols, and deployment targets, and a systematic overview that connects fundamental concepts, methodological development, and engineering support is still relatively lacking. Therefore, this paper presents a comprehensive survey of low-bit quantization methods for LLMs. We first introduce the basic concepts of quantization, including low-precision numerical representations, quantization granularity, and the main factors that influence quantization error and model performance. On this basis, we organize existing methods into two major paradigms: Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT). PTQ is emphasized as the dominant practical solution because it compresses pretrained models without expensive retraining, while QAT provides an important alternative by incorporating quantization simulation during training so that the model can better adapt to low-bit constraints. Within PTQ, this survey further reviews three major technical branches according to quantization targets and deployment goals. The first branch is weight-only quantization, which mainly focuses on compressing model parameters to reduce storage cost and memory traffic; representative directions include multi-bit and single-bit weight quantization, outlier handling, mixed-precision strategies, structured preprocessing, optimization-based quantization, and emerging non-uniform or ultra-low-bit methods. The second branch is joint weight-activation quantization, whose key objective is not only memory reduction but also actual low-precision computation acceleration during inference; related techniques include range migration, geometric transformation, activation-aware design, and low-bit computational optimization. The third branch is KV cache quantization, which is particularly important in autoregressive generation because the memory consumption of cached keys and values grows with sequence length and can become a major runtime bottleneck. In addition, this paper briefly reviews the software and hardware ecosystem supporting LLM quantization, including mainstream development frameworks, specialized optimization libraries, hardware accelerators, and deployment toolchains. Finally, we summarize current challenges and discuss future research directions, such as more robust ultra-low-bit quantization, better coordination between accuracy and efficiency, hardware – algorithm co-design, and more unified evaluation standards. By integrating methodological principles, representative techniques, and deployment considerations into a unified taxonomy, this survey aims to provide a clear reference for both academic researchers and industrial practitioners working on efficient LLM inference.

Keywords deep learning; large language model; model compression; quantization; model inference

1 引 言

近年来,以 GPT^[1-2]、LLaMA^[3-4]、Qwen^[5] 和 DeepSeek^[6] 等为代表的大语言模型(Large Language Model, LLM)在自然语言处理领域取得了突破性进展,重塑了自然语言任务的处理范式。通过在海量语料上的训练,这类模型能够自动学习并理解自然语言中的复杂模式。在通用语言理解与生成任务中表现卓越的同时,LLM 还展现出中小规模模型所不具备的“涌现能力”(emergent abilities),如少样本的上下文学习、自然语言指令执行,以及结合外部知识和工具实现功能增强等。这些能力使 LLMs 逐步成为各类复杂智能系统的核心组件,广泛应用于网络安全、代码生成、人机交互等多个领域。

LLM 之所以具备如此强大的能力,根本原因在于其庞大的参数规模和基于大规模数据的训练。然而,随着模型参数数量不断攀升,随之而来的计算与存储开销亦大幅增加,成为大规模部署的主要障碍。尤其是在资源受限的实际环境中,模型运行效率已成为影响其应用落地的关键因素。在实际应用中,由于参数规模庞大,LLM 推理通常面临“内存受限”而非“计算受限”的问题,即性能瓶颈更多体现在数据传输速率而非算力本身。这一现象突显了内存瓶颈的严重性,特别是在处理长序列任务时,内存利用效率成为影响模型部署效果的核心因素。LLM 推理过程中主要涉及以下几类存储开销:

(1) 模型权重(Weights)

LLM 通常包含数十亿甚至数百亿的参数。在推理阶段,这些参数需加载至显存中,占用大量内存资源。以拥有 700 亿参数的 LLaMA2-70B 模型^[4] 为例,若使用标准的半精度浮点数(FP16)表示,每个参数占用 2 字节,仅模型权重就需要 140 GB 的显存空间,如此巨大的存储需求远超大多数消费级乃至专业级 GPU 的容量,直接限制了模型在单张或少数几张 GPU 设备上的部署灵活性。

(2) 激活值(Activations)

在前向传播过程中,模型会生成大量中间激活张量,用于支持后续计算。这些激活值在推理时需暂存于内存中,构成显著的内存负担。尽管可以通过激活缓存减少 KV 缓存的重复计算,但缓存本身仍需占用显存;而若采用激活重计算策略来降低内存消耗,又会显著增加计算负担。随着批处理大小和输入序列长度的增加,这种内存-计算之间的权衡愈发突出。

(3) 键值缓存(Key-Value Cache)

KV 缓存是 LLM 推理阶段内存开销的主要来源。它用于存储历史 Token 对应的 Key 和 Value 向量,以避免在自回归生成过程中重复计算。尤其在长序列推理任务中,KV 缓存的规模随序列长度呈线性增长。同样以 LLaMA2-70B 模型为例,在批处理大小为 1、序列长度为 2048 时,其 KV 缓存的占用空间可达 5.4 GB。在更长序列的场景下,其总体内存占用可能达到数百 GB 甚至 TB 级别,成为长序列处理场景下优化难度最大的部分。

为应对上述挑战,模型压缩技术应运而生,旨在降低 LLM 的存储与计算成本。其中,量化(Quantization)作为最关键且高效的技术路径之一,其核心思想是将模型参数和中间激活值的数值精度从常见的 32 位浮点数降低为 8 位整数乃至更低精度(如 4 位整数),以实现资源效率的提升。量化技术的主要目标是在尽可能保持模型性能的前提下,提升推理速度、降低内存使用,从而实现更高效的模型部署。量化技术通过降低数值精度,带来了多方面的优势:

(1) 减少内存占用

通过采用更低位宽的数值表示,量化能够显著压缩模型的内存需求。例如,将模型权重从 32 位浮点数转换为 8 位整数,理论上可将存储开销减少约 75%。这一压缩效应对于部署在内存资源有限的设备尤为重要,使得 LLM 在资源受限环境中得以实际应用。此外,内存占用的减少也意味着可以在相同硬件资源下并行部署更多模型实例,从而提高部署效率并降低整体系统成本。

(2) 加速计算过程

低精度数值运算相较于高精度浮点运算,在现代处理器上通常具有更高的运算效率与更低的延迟。量化能够减少推理过程中所需的乘加操作复杂度,降低计算资源消耗,从而显著提升推理速度。这对于对响应时间要求较高的实时推理应用具有重要意义,如对话系统、语音助手及其他人机交互场景等。

(3) 降低访存能耗

LLM 推理阶段的能耗大部分来源于数据在存储层级及处理器之间的频繁传输。由于低比特表示可有效减少单次数据传输的位数,量化能够在保持模型功能的同时,显著降低访存带宽压力与系统整体功耗,有助于实现更绿色、节能的计算系统。

综上所述,低比特量化技术在提升大型语言模

型运行效率、推动其向边缘侧部署方面具有关键作用。其在理论研究与实际应用中均展现出广泛前景,是当前模型压缩领域的重要研究方向之一。

现有LLM低比特量化技术主要包括两种典型方法:训练后量化(Post-Training Quantization, PTQ)与量化感知训练(Quantization-Aware Training, QAT)。PTQ是在模型训练完成后对权重和激活进行直接量化的方法,近年来已涌现出许多适用于LLM的先进技术,如GPTQ^[7]。PTQ的主要优点在于无需重新训练,部署成本较低。然而,在使用极低比特(如4位以下)表示时,PTQ可能导致模型精度显著下降。相比之下,QAT通过在训练阶段引入量化噪声,模拟推理过程中的量化行为^[8],使模型在训练过程中学习对量化误差的鲁棒性,从而在保持低比特表示的同时尽可能减少精度损失。尽管QAT通常对计算资源和训练数据的要求更高,但其在对精度要求较高的实际部署场景中表现出更强的适应性和稳定性。

近年来,随着LLM在学术界和工业界的广泛应用,其量化技术逐渐成为研究热点。已有多篇高质量的综述性文献对该领域进行了系统梳理。表1展示了本文相关的代表性综述。综述[9]系统回顾了大语言模型的整体研究进展与核心技术,内容涵盖背景知识、预训练、微调、使用方式及评估方法等多个方面,具有较强的全面性与系统性。作为通用型LLM综

述,这些文章构建了完整的技术框架与发展路线图。然而,由于综述范围较为广泛,尽管提及了模型压缩技术,但对量化这一特定方向未作深入探讨。综述[10]针对LLM的模型压缩技术进行了较为全面的综述,涵盖了量化、剪枝、知识蒸馏与低秩分解等关键方法,旨在优化模型体积、提升推理效率,并支持在资源受限环境下的部署。虽然该综述涉及了一些具体量化方法,但作为压缩领域的通用综述,对于LLM量化的技术细节与最新进展仍缺乏深入剖析。综述[11]聚焦于低比特表示技术在LLM训练与推理阶段的应用,讨论了QAT与PTQ方法。然而,由于同时介绍了训练阶段的低比特技术,对推理量化技术的探讨深度仍显不足。此外,由于该综述发表时间较早,未能涵盖最新的技术进展。综述[12]专注于LLM的PTQ方法,基于实证研究对主流量化策略进行了系统评估,提供了详细的量化效果对比及部署建议。评估涵盖多种模型结构、任务类型和量化目标,具有较强的实践参考价值。但其核心贡献在于评估分析而非理论综述,因此对量化方法的算法原理、演进脉络与技术细节的讨论相对有限。此外,综述[13]聚焦于在资源受限的端侧设备上部署与应用大语言模型,系统性地梳理了模型压缩、运行时优化及端边云协同等关键技术。综述[14]从边缘智能的视角出发,全面调研了面向边缘的大模型技术,并将其划分为边缘训练与边缘推理两大板块进行深入分析与讨

表1 大语言模型低比特量化方法相关综述

文献	主要内容	侧重点			
		通用压缩技术	PTQ	QAT	性能评估 生态
A survey of large language models ^[9]	• LLM基础知识 • 预训练与微调技术 • 模型使用方式与评估方法 • 量化	✓			
A survey on model compression for large language models ^[10]	• 剪枝 • 知识蒸馏 • 低秩分解	✓			
A survey of low-bit large language models: Basics, systems, and algorithms ^[11]	• 量化基础概念 • 训练与推理中的量化技术	✓			✓
A comprehensive evaluation of quantization strategies for large language models ^[12]	• PTQ方法性能评估 • 部署建议		✓		✓
迈向泛在智能: 端侧大语言模型现状与展望 ^[13]	• 模型压缩技术 • 端边云协同技术	✓			
面向边缘智能的大模型研究进展 ^[14]	• 参数高效微调技术 • 模型压缩技术 • 量化基础概念	✓			✓
本文	• PTQ与QAT方法 • 部署生态		✓	✓	✓

论。这两篇文章也同样存在研究范围过广,未能深入探讨量化技术细节的问题。

综上所述,尽管现有综述文献在广度和应用层面具备一定参考价值,但在对LLM低比特量化技术进行系统性、层次化的梳理上仍有不足。为填补这一空白,本文旨在提供一个全面而深入的综述,其贡献主要体现在以下三个方面:

(1)构建了更系统的分类体系

本文将LLM量化技术划分为训练后量化(PTQ)和量化感知训练(QAT)两大范式。在此基础上,针对当前研究的重点PTQ,我们进一步将其解构为仅权重量化、权重-激活联合量化以及键值缓存量化三个技术分支。这种层次化的分类体系更能揭示不同技术路径的内在逻辑与演进关系。

(2)实现了对前沿技术更全面的覆盖

本文不仅系统梳理了主流的4比特/8比特量化方法,还详细介绍了在2比特乃至1比特等极低比特场景下的前沿工作,涵盖了向量量化、旋转变换、均衡化预处理等新兴技术,确保了综述内容的时效

性与前瞻性。

(3)强化了算法与硬件生态的结合

与侧重纯算法的综述不同,本文还系统性地梳理了支持LLM量化部署的主流开发框架、专用优化库与硬件平台,旨在为该领域提供从理论算法到实践部署的全方位介绍,为研究者和工程师提供更具实践指导意义的参考。

本文的整体结构如图1所示。第2节将介绍量化的基本概念,包括数值的低精度表示、量化粒度的选择,以及量化对模型性能的影响,为后续方法论探讨奠定理论基础;第3节讨论主流的训练后量化方法,重点分析其不同技术路径,包括仅对权重的量化、权重与激活联合量化,以及针对KV缓存的量化方案;第4节介绍量化感知训练的原理与实践,阐释其如何通过模拟量化过程提升模型鲁棒性与性能;第5节梳理当前支持LLM量化部署的主要软件框架与硬件生态系统,涵盖编译器优化、运行时支持与加速硬件;第6节就未来研究的潜在方向与技术发展趋势进行展望;第7节总结全文。

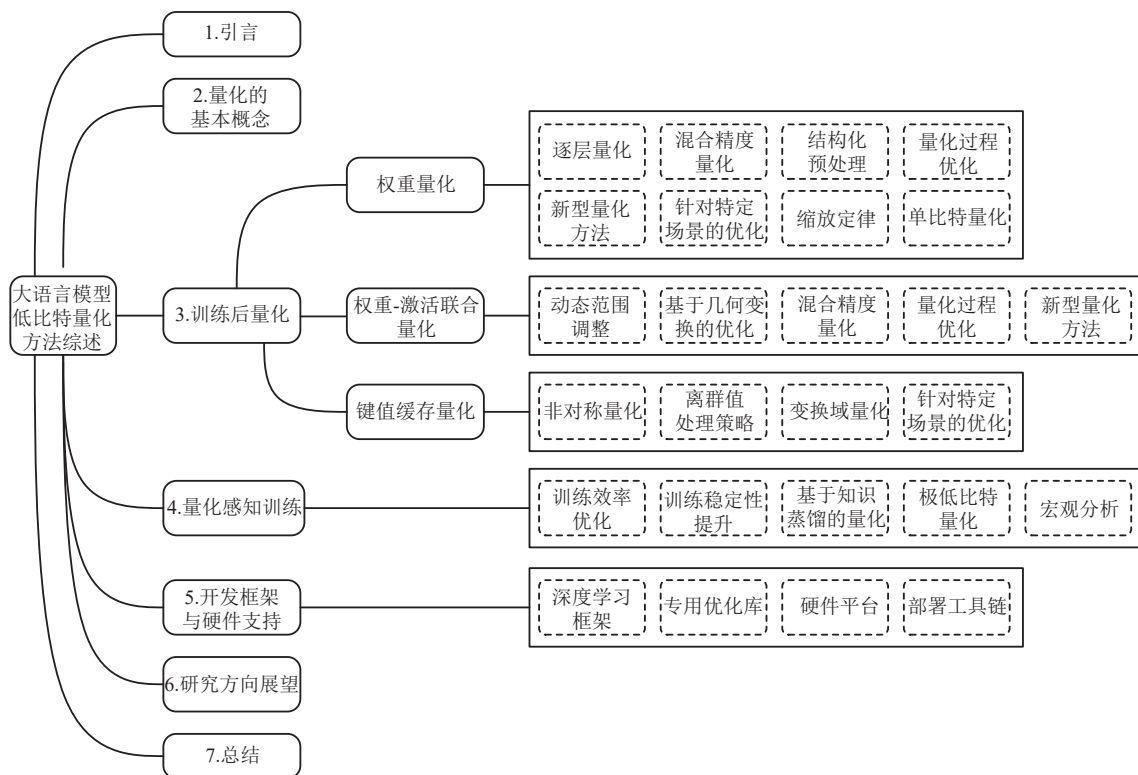


图1 本文内容组织

2 量化的基本概念

在大语言模型的实践中,低比特量化已成为

一项关键技术,它特指将模型权重与激活值从标准浮点格式转换为8位或更低位宽的数值表示的过程,对于推动参数规模庞大的LLM在资源受限环境下的部署与应用,扮演着至关重要的角色。其基本

思路在于,使用较低的数值精度来表示模型中的权重和激活值,从而实现减小模型存储体积、降低访存开销、提升计算速度并降低功耗的多重优势。相较于16位浮点数推理基准,8位整数实现了存储和内存带宽的减半,并且是当前主流AI加速硬件原生支持并能带来显著推理加速的最低整数精度之一。因此,本综述将8位量化视为低比特量化范畴的重要组成部分加以介绍,并以此为基础剖析当前更具挑战性与前沿性的4位及以下量化方法。本章将系统阐述量化的基本原理、关键技术要素及其对LLM性能的潜在影响。

2.1 浮点数与整数

在典型的LLM训练过程中,网络权重与激活值通常采用32位单精度浮点数(FP32)格式进行表示与存储。根据IEEE 754标准,一个FP32数值由三部分组成:1个符号位(Sign, S)、8个指数位(Exponent, E)和23个尾数位(Mantissa, M)。其表示的浮点数值 V 可以通过以下公式计算:

$$V = (-1)^S \times (1.M) \times 2^{E-\text{bias}} \quad (1)$$

其中, bias 是一个固定的偏移值,对于FP32而言通常是127。从该结构可以看出,指数项 $2^{E-\text{bias}}$ 使得FP32能够表示从极大到极小的数值,这对于捕捉模型训练过程中梯度和参数的剧烈变化至关重要,能保证训练的稳定收敛。此外,浮点数的精度是相对的,它在接近0的区域具有高精度,而在远离0的区域精度较低。然而,在模型推理阶段,这种高精度和宽范围往往存在显著的冗余。研究表明,LLM对参数和激活值的微小扰动具有一定的鲁棒性,这意味着我们可以使用更少的比特数来近似表示原始的FP32数值,而不过多损害模型的预测性能。

与浮点数不同,整数的表示是线性和均匀的,即任意两个相邻可表示的数值之间都具有相同的间距。在模型量化中,为了能表示正负数值通常使用有符号整数。最常用的有符号整数表示法是二进制补码(Two's Complement)。对于一个 n 位的有符号整数,其最高位作为符号位,表示的数值 V 可以通过以下公式计算:

$$V = -b_{n-1}2^{n-1} + \sum_{i=0}^{n-2} b_i 2^i \quad (2)$$

其中, $b_i \in \{0, 1\}$ 是第 i 位的比特值, b_{n-1} 是符号位。这种表示法允许 n 个比特表示的范围为 $[-2^{n-1}, 2^{n-1} - 1]$ 。这种均匀分布的特性,虽然表示范围有限,但其运算在现代CPU和专用AI芯片上的计算效率远高于浮点

运算,且硬件实现更简单、功耗更低。这为量化技术提供了理论依据。图2比较了浮点数与整数的构成。

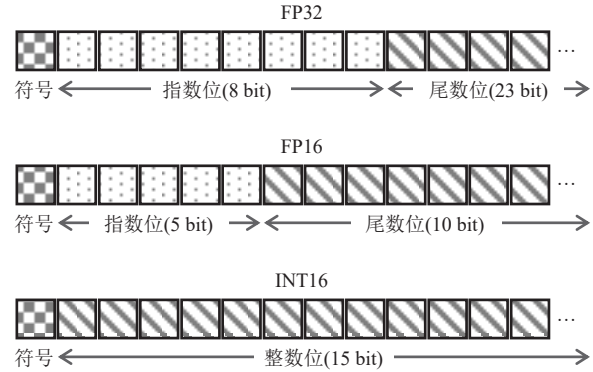


图2 浮点数与整数的构成

除标准的FP和INT格式外,研究人员还提出了诸如BF16^[15]和NF4^[16]等多种创新的数据类型。BF16是一种16位浮点格式,虽然通过缩减尾数位牺牲了部分精度,但仍保留与FP32相同的8位指数位,因此能够提供与FP32相近的动态表示范围。这使得BF16能够在保证训练稳定性的同时,显著降低内存和计算开销,因此成为模型训练的流行选择。NF4是一种更加激进的4位非均匀数据类型,主要应用于模型推理和高效微调。与传统的线性整数分布不同,NF4的量化点根据模型权重的高斯分布进行优化,在数据密集的区域分配更多的表示值。这种优化方式使得NF4能够以极低的比特数更有效地保留模型的性能,表现优于传统的INT4格式。综上所述,从FP32到BF16,再到NF4及各类整数格式的逐步演变,反映了在不同应用场景中对模型精度、内存占用和计算效率之间的灵活权衡。

2.2 从浮点到整数的映射

量化的本质,就是寻找一个最优的映射函数 f_Q ,将高精度、大范围、非均匀分布的浮点数空间,映射到一个数量有限的低精度离散值集合中。根据目标集合中离散值的分布方式,量化方法可主要分为两大类:线性量化和非线性量化。

实现这一目标的关键在于量化映射。线性量化(Linear Quantization)是目前应用最广泛的量化方法之一,其通过一个简单的线性仿射变换,将高精度浮点数 r 映射到低精度整数 q 。其数学表达式如下:

$$q = \text{round}\left(\frac{r}{s}\right) + z \quad (3)$$

其中,浮点数 s 是缩放因子(scale),定义了量化前后的映射比例。它负责将输入浮点数的动态范围压缩

到目标整数的范围内,决定了量化所能表示的数值范围的粒度。整数 z 是零点(zero-point)。它确保真实值中的0能够被精确地映射到量化值域中的某个整数,从而避免信息损失。

相应的,从量化后的整数 q 恢复到近似的浮点数被称为反量化或解量化(Dequantization),该过程表示为:

$$r' = s \times (q - z) \quad (4)$$

根据零点 z 的取值,线性量化又可分为对称量化与非对称量化。对称量化将 z 固定为0,主要用于处理以0为中心对称分布的数据;而非对称量化则允许 z 为任意整数,能够更灵活地适应非对称的数据分布,从而可能获得更小的量化误差。线性量化中量化后的整数值域中任意两个相邻数值,其对应的原始浮点数值之间的间隔是恒定的。这种均匀分布的特性使得线性量化实现简单、计算高效,且与现代计算硬件高度兼容,因此成为了当前应用最广泛的量化方法。

与线性量化相对应的是非线性量化。非线性量化不再遵循均匀的映射关系,即量化级别之间的间隔不是固定的。这种方法旨在将有限的比特数更智能地分配给数据分布中更重要的区域,从而在理论上以相同的比特数实现更低的量化误差。常见的非线性量化包括对数量化、K-均值量化以及更广义的矢量量化^[17]。与线性量化相比,非线性量化则通过更优地拟合数据分布,在极低比特场景下展现出实现更高精度的潜力,但通常伴随着更复杂的实现和潜在的硬件兼容性挑战。

2.3 量化粒度

缩放因子和零点的作用范围,即量化粒度,是决定量化精度的核心要素之一。不同的粒度代表了模型参数共享量化参数的范围,在量化误差与计算、存储开销之间形成了直接的权衡。主流的量化粒度包括张量级、通道级和分组量化。

张量级量化(Tensor-wise Quantization)是最粗糙的量化粒度。对于一个完整的权重张量(Tensor),例如某个全连接层的权重矩阵,所有的元素均共享同一对缩放因子和零点。其优点是实现简单,量化参数的存储开销极小。然而,如果张量内不同区域的数值分布差异巨大,单一的量化参数将无法有效适应这种差异,从而导致显著的量化误差,严重损害模型性能。

通道级量化(Channel-wise Quantization)为权重矩阵的每个输出通道(channel)计算独立的量化

参数。在LLM的线性层中,这通常表现为按行或按列量化。例如,对于一个形状为 $d_{out} \times d_{in}$ 的权重矩阵,按行量化会为每一行(对应一个输出神经元)计算一对独立的缩放因子和零点,共计 d_{out} 组。这种方式能更精细地捕捉不同输出特征之间的数值分布差异,是目前平衡性能与开销的主流选择,极大地提升了量化后LLM的精度。

分组量化(Group-wise Quantization)是比通道级量化更细粒度的方法,其将单个行或列的权重向量再次划分为若干个小组(group),并为每个小组计算独立的量化参数。例如,一个维度为4096的行向量可以被划分为32个组,每组包含128个权重,每组共享一对缩放因子和零点。这种细粒度策略能够以更高的保真度来表示那些包含离群点的区域,能够在极低比特量化下维持量化后LLM的性能。

总结来说,量化粒度与模型性能之间存在显著的权衡关系。随着量化粒度从张量级逐步细化至通道级乃至分组量化,模型对局部数据分布的拟合能力不断增强,从而有效降低量化误差并较好地保持原始模型精度。然而,更细的量化粒度需要存储更多的量化参数,导致存储开销增加。同时,在计算过程中,需要根据不同的粒度加载相应的量化参数进行解量化,这可能引入额外的计算延迟和控制逻辑复杂度。表2比较了不同的量化粒度。

表2 不同量化粒度的比较

量化粒度	参数共享范围	精度	存储开销	计算复杂度
张量级	整个张量	较低	极小	低
通道级	每行/每列	中等	较小	中
分组量化	行/列内的分组	较高	较大	高

需要注意的是,量化粒度这一概念描述的是量化参数(即缩放因子和零点)的共享范围,它决定了如何对一个给定的权重张量进行量化。与之正交的另一个概念是量化流程的范围,它描述了量化算法执行和优化的基本单元,即按照什么顺序来处理整个模型。量化流程的一个典型例子是逐层量化。它指代的是一种将模型逐层进行独立量化的处理策略,优化目标是最小化当前层的局部量化误差。在这个流程中,对每一层的权重进行量化时,依然需要为其选择一种具体的量化粒度,例如通道级或分组量化。

2.4 量化对模型性能的影响

量化本质上是一种有损压缩,其引入的误差

是影响模型性能的根本原因。这些误差主要来源于两个方面：

(1)截断误差(Clipping Error)

在计算量化参数之前,通常需要确定一个裁剪范围。这个范围外的数值将被强制映射到边界值。在LLM中,权重和激活值中存在的离群点对模型性能起着关键作用,若裁剪范围设置不当,将这些离群点截断会直接损害模型的重要功能。

(2)舍入误差(Rounding Error)

在将缩放后的浮点数映射到离散整数时,取整操作会引入舍入误差。误差的大小与缩放因子直接相关。粒度越粗或数据动态范围越大,缩放因子的值就越大,每个量化整数代表的浮点范围就越宽,舍入误差也就越显著。

这些误差在LLM的逐层计算中会不断累积和传播。例如,在自注意力模块中,Key和Value矩阵的微小量化误差,可能在经过Softmax和后续的矩阵乘法后被放大,最终影响到模型的输出分布,导致

生成文本的流畅度、事实准确性或指令遵循能力下降。因此,如何设计精巧的量化算法,通过改进量化参数计算、优化模型权重分布或在量化后进行微调来补偿这些误差,是LLM低比特量化领域研究的核心挑战。后续章节将围绕这一问题对各类前沿量化方法展开详细论述。

3 训练后量化

训练后量化(Post-Training Quantization, PTQ)是指在模型预训练或微调完成后,直接对模型参数进行量化的技术。它通常仅需要少量无标签的校准数据集来统计模型参数或激活值的分布信息,即可完成量化过程,有时甚至无需任何数据。本章将系统介绍主流的PTQ方法,首先从仅量化权重的技术出发,然后探讨更进一步的权重-激活联合量化,最后对支撑这些方法的理论发现进行分析。表3总结了训练后量化现有工作。

表3 训练后量化现有工作

分类	描述	现有工作
权重量化	逐层量化	逐层对权重矩阵进行量化并补偿误差 GPTQ ^[7]
	混合精度量化	针对权重中的离群点采用不同比特数分配 SpQR ^[18] , Falcon ^[19] , SqueezeLLM ^[20] , OWQ ^[21] , SEMQ ^[22]
	结构化预处理	通过剪枝、旋转等变换提升权重量化友好性 OliVe ^[23] , RSQ ^[24] , Bondarenko等 ^[25] , An等 ^[26] , AWQ ^[27]
	量化过程优化	优化舍入、参数初始化、误差补偿等环节 SignRound ^[28] , NeUQI ^[29] , QEP ^[30] , PQI ^[31]
	新型量化方法	探索均衡化预处理、向量量化等创新技术 QuIP ^[32] , QuIP ^{≠[33]} , AQLM ^[34] , QTIP ^[35]
	针对特定场景的优化	为特定模型架构和应用场景定制量化方案 DB-LLM ^[36] , MoEQuant ^[37] , Wang ^[38]
	缩放定律	通过大规模实验建立量化缩放定律 Dettmers等 ^[39]
权重-激活联合量化	单比特量化	极低比特场景下的量化 PB-LLM ^[40] , BiLLM ^[41] , PBS ² P ^[42] , BinaryMoS ^[43] , PV-Tuning ^[44]
	动态范围调整	利用等价变换将激活离群点迁移至权重 SmoothQuant ^[45] , OmniQuant ^[46] , Outlier Suppression+ ^[47] , QLLM ^[48] , MergeQuant ^[49] , per-IC ^[50]
	基于几何变换的量化	通过几何变换均匀权重与激活分布 DuQuant ^[51] , SpinQuant ^[52] , OSTQuant ^[53]
	混合精度量化	为离群激活/权重采用高精度数值计算 LLM.int8() ^[54] , QUIK ^[55] , QRazor ^[56]
	量化过程优化	引入知识蒸馏、梯度等高阶信息 ZeroQuant ^[57] , GuidedQuant ^[58] , Q-VLM ^[59]
键值缓存量化	新型量化方法	使用微调、低秩近似,采用新数值格式等 Norm Tweaking ^[60] , LLM-FP4 ^[61] , LQER ^[62] , CushionCache ^[63]
	非对称量化	针对键值缓存的分布差异采用不同量化策略 KIVI ^[64]
	离群值处理策略	动态识别并隔离缓存中的离群值 OTT ^[65] , KVQuant ^[66] , RotateKV ^[67] , IntactKV ^[68]
	变换域量化	借助SVD等技术将缓存映射到易量化空间 SVDq ^[69] , PolarQuant ^[70] , NSNQuant ^[71]
键值缓存量化	针对特定场景的优化	为特定应用设计定制缓存量化方案 AQUA-KV ^[72] , PM-KVQ ^[73] , VidKV ^[74]

3.1 权重量化

LLM的参数规模动辄达到百亿甚至千亿级别,

这使得模型的存储和加载成为巨大的开销。由于模型的权重在推理过程中是静态不变的,研究者们可

以离线地将其从高精度的浮点数转换为低精度的整数,从而在不改变模型结构和计算图的情况下,以极小的性能损失为代价,实现模型体积的数倍压缩。这种方法显著降低了对存储空间和内存带宽的需求,是当前应用最为广泛的LLM量化技术之一。

在LLM的权重量化领域,研究者们沿着多条技术路径展开了探索,旨在以更低的比特数表示模型权重,从而减小模型体积、降低内存开销并加速推理。

3.1.1 逐层量化

GPTQ^[7]是一个里程碑式的工作,它奠定了逐层量化范式的基础。其核心思想是在不改变模型输入的情况下逐层对权重矩阵进行量化。在每一层内,GPTQ采用迭代量化的方法,一次选择一部分权重进行量化并确定其取值,其目标是最小化原始层输出与量化后层输出之间的均方误差。为了做出最优的舍入决策,GPTQ利用了Hessian矩阵的近似信息来评估不同选择对该层输出重建误差的影响。具体而言,Hessian矩阵衡量了层输出对权重变化的敏感度,GPTQ利用该信息来补偿量化误差,即每量化一个权重,就相应地更新其余未量化的权重,以最小化累积误差。通过这种基于二阶信息的贪心量化策略,GPTQ首次成功地将百亿参数级别的LLM压缩至3-4比特,同时保持了极高的模型精度,证明了低比特权重量化在LLM上的可行性。

GPTQ确立了“量化-补偿”的逐层优化框架,并揭示了利用二阶信息指导量化决策的有效性。它成为了后续许多工作的基石或对比基线。未来的工作应关注如何进一步降低该框架的计算复杂度,以及如何将其与更复杂的离群点处理技术或全局优化策略相结合,以突破单层贪心优化的性能瓶颈。

3.1.2 混合精度量化

然而,实践表明,模型性能的下降往往源于权重矩阵中少数离群点的存在。这些数值的绝对值远超其他权重,对量化误差非常敏感。因此,如何妥善处理离群点成为了后续研究的焦点。一个主流且行之有效的思路是采用混合精度量化,即为离群点和普通值分配不同的比特数。在此方向上,SpQR^[18]提出了一种稀疏量化表示方法。它首先识别出那些对整体量化误差贡献最大的权重作为离群点,将这部分占比极小的权重以较高的精度独立存储,而对其余所有的非离群点权重则应用3~4比特的量化。通过这种方式,SpQR在对LLaMA^[3]和Falcon^[19]等LLM进行量化时,实现了低于1%的相对困惑度

损失。

SqueezeLLM^[20]也采用了类似的思想,其出发点是LLM推理的主要瓶颈在于内存而非计算。它将权重矩阵显式地分解为包含离群点的稀疏部分和数值范围更紧凑的密集部分。稀疏部分用高效的稀疏格式存储全精度值,而使得密集部分可以被更有效地量化。此外,SqueezeLLM还引入了一种基于二阶信息的敏感度分析方法,来为不同权重动态地、非均匀地分配最优比特数。这一系列优化使得模型即便在低至3比特的精度下也能实现近乎无损的压缩,在减小模型尺寸的同时,不牺牲模型性能并提升了推理速度。

OWQ^[21]从权重矩阵的结构敏感性入手,通过实验发现,激活值中的离群点是导致权重量化质量下降的关键因素,并导致了权重矩阵中某些特定的列对量化尤为敏感。基于此,OWQ设计了一种混合精度量化策略,它优先将这一小部分对量化敏感的权重列保持在较高的精度,同时对其余的密集权重进行高度优化的量化处理。实验证明,采用OWQ的3.1比特量化模型的表现可以实现与4比特基线模型相近的性能。

SEMQ^[22]利用权重敏感度构造误差最小化目标,并通过迭代方式优化非均匀量化点,同时结合离群值检测与隔离机制以减小极低比特量化时离群值对量化范围分配的干扰。通过对LLM权重的非均匀分布特征的适配,SEMQ在2比特和3比特量化设置下取得了具有竞争力的结果。

混合精度方法在保护模型性能和实现高压缩率之间取得了出色的平衡,已成为处理离群点问题的主流范式。未来的研究方向可能包括开发更高效、更自适应的离群点检测与比特分配算法,减少人工调参;或设计专门的硬件加速器或计算核函数,以高效处理混合精度带来的稀疏和不规则计算模式。

3.1.3 结构化预处理

除了混合精度,研究者们也提出了其他离群点处理方案,它们通过改变数据结构或分布来提升量化友好性。OliVe^[23]指出,先前依赖稀疏编码来分离离群点的量化方案需要复杂的硬件编解码逻辑和额外的计算调度控制器。为了解决这一问题,OliVe通过一种独特的嵌入策略避免了混合精度方案中复杂的稀疏计算逻辑。其核心假设在于离群点旁边的普通数值重要性不高。基于此,OliVe首先将离群点周边的普通权重剪枝为零,从而在低精度的权重矩阵中让出空间;然后,它将高精度的离群点

数据直接嵌入这些由剪枝权重让出的零值位置。这种方法在硬件层面更为友好,因为它维持了稠密矩阵的结构。

RSQ^[24]指出,传统方法在优化时平等地对待所有词元(token),忽略了它们在语义上的重要性差异。为此,RSQ提出了在量化前对权重进行预处理的“旋转-缩放-量化”方法。其首先对权重矩阵进行正交旋转变换,这种变换不改变矩阵的数学性质,但可以使权重分布更加均匀,从而有效减小离群点的峰值;接着,它根据token在注意力机制中的重要性得分对激活特征进行缩放,强化重要信号;最后,将预处理后的权重和缩放后的激活输入到GPTQ框架中进行量化。

为深入理解离群值的形成机理,近期研究对其进行了系统而细致的分析。研究发现,离群值并非随机噪声,而是在模型预训练过程中系统性地产生。通过对其在不同维度上的分布特征进行分析,研究揭示了一种与自注意力机制中的Softmax操作密切相关的离群值产生机制:当某些注意力头尝试学习一种“无操作”或仅对残差连接进行部分更新的行为时,为了在注意力矩阵中获得接近零的值,其输入会在训练过程中被推向极大,从而在网络的其他部分中引发显著的离群值现象^[25]。有趣的是,这些离群值并非完全具有负面作用。在注意力机制中,它们可能扮演着类似于上下文感知的隐式缩放因子的角色,对于维持模型性能具有重要意义^[26]。基于对离群值形成机制的深入理解,研究者提出了一些可从模型结构层面加以优化的方案。例如,通过引入裁剪Softmax或门控注意力机制,可以在预训练阶段有效抑制离群值的产生,从而得到更加量化友好的模型^[25]。

此外,AWQ^[27]从激活分布出发识别对模型更重要的权重通道,并通过等价的按通道缩放来保护这些显著权重,从而在不依赖反向传播和重构优化的情况下减小量化误差。该方法兼顾了硬件友好性与量化精度,因而已成为4比特权重量化中的代表性方法之一。

结构化预处理方法为解决离群点问题提供了新的思路,它们致力于从根本上降低量化难度,而非被动地适应离群点的存在。这些方法尤其强调硬件友好性。未来的探索可以集中于发现更多保持模型功能不变的数学变换,这些变换需要同时具备降低离群点影响和计算开销低廉的双重优点。

3.1.4 量化过程优化

除了宏观策略,还有大量研究致力于优化量化算法的各个核心环节,以提升精度和效率。例如,在最基础的舍入步骤上,传统的四舍五入策略是静态且独立的,忽略了权重之间的相互影响。为了解决这个问题,有工作提出将舍入决策本身视为一个可优化的过程,利用SignSGD等基于符号的优化器来学习每个权重的最佳舍入方向,从而在不引入任何额外推理开销的前提下,找到一个全局更优的量化权重组合^[28]。

在量化参数的初始化方面,NeUQI^[29]指出现有的均匀量化方法虽然高效和易于部署,但其参数初始化方法却未得到充分探索。具体而言,这些方法通常将量化零点约束为整数,这限制了其表示能力。针对此,NeUQI通过将零点放宽为浮点数,并将求解最优尺度和零点的问题分解为在固定尺度下确定最优零点和在给定零点下寻找最优尺度这两个高效可解的子问题。通过这种方式,该方法为均匀量化找到了更优的初始化参数,提升了其性能上限。

针对逐层量化中误差逐层累积的固有缺陷,尤其是在低比特场景下该问题会显著恶化模型性能,QEP^[30]框架提出了一种前向传播和补偿机制。它在优化当前层的量化目标时显式地引入一个由所有先前层级累积的量化误差构成的项,迫使当前层在量化时主动补偿这些误差。该框架还引入了可调节的传播强度,以防止对误差的过度拟合,并且整个过程的计算复杂度与现有的逐层量化方法相当。

PQI^[31]通过实证研究指出,先前基于梯度或Hessian的敏感度指标由于泰勒展开的局部收敛半径限制,会严重低估量化对模型最终损失的真实影响,其低估程度可达数个数量级。为此,它提出了一种更为精确的后验敏感度度量方法。该方法仅利用了模型易于满足的局部连续性假设,并且同时考虑原始权重和量化后的权重来计算每个元素的具体影响,为各种依赖于敏感度分析的量化算法提供了更精确、更可靠的基础。

对舍入、参数初始化、误差传播和敏感度度量等基础模块的精细打磨,是推动量化性能持续提升的关键。未来的研究需要继续在这些核心环节上寻求理论突破和算法创新,例如探索更全局的优化策略以替代贪心舍入,或者发展能更精确捕捉长程依赖的误差补偿机制。

3.1.5 新型量化方法

除了在现有框架内进行改进,一些研究也在探

索全新的量化范式,试图从根本上改变量化的方式。QuIP^[32]提出,当权重矩阵及其Hessian矩阵数值分布均匀时,模型对所有权重的敏感度也较为均衡,此时量化也会变得更容易。为此,QuIP首先通过一个高效的正交变换对权重和Hessian进行预处理,以增强其数值均衡性,然后再进行自适应的数值舍入。该舍入过程,旨在最小化一个关于原始权重和量化权重之间误差的二次代理目标函数。其后续改进QuIP⁺^[33]进一步将此思想完善,采用了理论性质更优且运行速度更快的随机哈达玛变换,并引入了源于数学“堆球问题”最优解的E8格(E8 lattice)来设计码本,并引入了一种层间微调算法以进一步提升量化质量,最终实现了最先进的2比特量化效果。

另一条路径是探索非线性量化方法,其典型代表是向量量化。它将多个权重打包成一个向量进行联合量化,能更高效地利用信息。AQLM^[34]将加性量化(Additive Quantization)的向量量化技术引入LLM。它将权重向量近似为从多个码本中提取的码字之和,并将其优化目标从简单地重建权重本身扩展为在联合优化跨层块码本的同时最小化层输出在真实输入分布下的误差。这种方法能够产生一种硬件友好的同构量化格式,无需对离群点进行特殊处理。

QTIP^[35]研究了传统向量量化中码本大小随维度指数爆炸的难题。它引入了通信领域的格码量化(Trellis Coded Quantization),利用状态化的解码器将码本大小与比特率和维度解耦,从而实现了超高维度的向量量化。QTIP设计的硬件友好型“位移”网络结构,使其在获得高量化质量的同时也保证了较快的推理速度。

均衡化预处理和向量量化代表了权重量化的前沿探索方向,它们为实现2比特及以下的极低比特量化带来了希望。均衡化方法的挑战在于寻找计算开销和均衡效果俱佳的变换。向量量化的主要障碍则在于如何克服码本存储和搜索的复杂性,并设计出能高效执行VQ解码的硬件架构。未来的研究需要进一步探索新型量化策略,以突破现有方法性能极限。

3.1.6 针对特定场景的优化

除了通用量化方法,还有一系列工作专注于特定模型架构、特定比特数或量化后的再处理。DB-LLM^[36]通过“对偶二值化”方法实现了高性能的2比特量化,且该过程无需任何校准数据。它使用两个独立的、并行的1比特稀疏权重矩阵进行同步矩阵乘法来共同表示2比特权重,极大地增强了

模型的表示灵活性,并结合偏差感知蒸馏技术来对量化后的模型进行微调,引导量化模型更加关注那些模糊不定的样本,从而优化量化参数。

针对混合专家模型架构,MoEQuant^[37]提出了专门的量化框架。该工作识别出MoE模型的稀疏和动态特性带来的两大挑战,即样本在不同专家间的分布不均导致的低频专家校准不足,以及不同样本与其分配到的专家之间的关联度存在差异,设计了专家均衡自采样策略,利用LLM自身的采样能力,并结合累积概率和专家均衡指标来指导采样过程,以解决不同专家被调用频率不均导致的校准数据偏差问题;并提出了亲和引导量化,通过将亲和力和整合到逐层校准中并构建加权量化误差,来处理不同样本与专家之间关联度各异的挑战。

此外,在量化后的再压缩方向,有工作发现^[38]通过简单的缩放操作可以改变已量化数据的分布,使其变得更不均匀、更具可压缩性,随后再应用无损压缩算法,可以实现对模型体积的进一步压缩。实验结果表明,该方法可以在推理时将模型内存占用进一步降低40%。

针对特定场景的优化展示了量化技术与具体应用深度结合的趋势。无论是面向2比特等极端场景,还是MoE等新兴架构,定制化的量化方案都能取得比通用方法更好的效果。量化后再压缩则开辟了一个新的优化维度。未来,随着LLM架构的持续演进,研究者们应该探索更多专用解决方案,以满足多样化的部署需求。

3.1.7 缩放定律

除了从算法层面进行分析与改进,还有研究通过大规模实证研究探索权重量化的最优策略与性能边界^[39]。该研究通过构建推理缩放定律(Scaling Laws),系统评估了不同比特精度与模型规模对零样本(Zero-shot)性能的影响,涵盖了包括OPT^[75]和BLOOM^[76]在内的多个主流模型家族,参数规模从19 M到176 B不等。这一大规模实证研究得到了具有实践指导意义的结论表明:在固定模型存储成本条件下,4比特权重量化在所有测试的模型规模与模型家族中均表现出最优性能。

大规模实证研究和缩放定律帮助实践者在面对不同模型和硬件约束时做出更优的选择。未来的工作应着眼于构建更全面的缩放定律,将更多变量(如激活量化、不同量化算法、特定硬件延迟等)纳入考量,为从业者提供一个覆盖更广、预测更准的性能评估框架。

3.1.8 单比特量化

在多比特量化方法的基础上,为追求极致的压缩率与推理效率,研究者们进一步将目光投向了单比特量化,也即模型二值化。在这种设定下,模型权重仅由+1和-1(或0和1)表示,能够将模型存储开销降低至理论极限,并可通过高效的位运算替代高成本的乘加运算。然而,如此激进的量化方式会引入巨大的信息损失,直接将预训练模型的权重进行二值化通常会导致模型性能的灾难性下降。因此,如何设计精细的量化方案,以在压缩的同时保留模型关键能力,成为关键问题。为应对上述挑战,部分研究采取了多比特量化中常用的混合精度策略进行部分二值化,该方法在量化前首先识别出小部分显著权重并为其分配较高的比特数进行存储,而将其余大部分非显著权重进行二值化。通过这种策略,模型能够在实现极高压缩率的同时有效维持其语言推理能力^[40]。

在区分权重显著性的基础上,BiLLM^[41]对不同重要性的权重设计了更为精细的差异化处理策略。该框架利用基于Hessian矩阵的度量方法来结构化地识别显著权重,并为其设计了一种残差近似方法以最大程度地恢复其动态范围;对于其余的非显著权重,BiLLM则设计了最优分裂二值化策略,通过搜索权重分布的最佳断点将权重分段后分别进行二值化,从而最小化整体量化误差。

PBS²P^[42]将剪枝与二值化优化相结合,提出一种步进式半结构化剪枝与二值化优化方法。其在每个步骤中同步地剪枝一部分元素并优化二值化参数,从而有效降低了由两种技术共同引入的误差。与BiLLM类似,该方法也利用基于Hessian矩阵的度量来指导模型识别需要剪枝的具体元素。具体而言,它设计了一种从粗到细的搜索策略:在粗粒度阶段,根据每层的冗余度确定其剪枝率;在细粒度阶段,则在该剪枝率的指导下,利用Hessian信息识别出要剪枝的具体元素,实现了剪枝和量化过程的协同优化。

与上述区分并差异化处理不同权重的方法不同,另一类工作致力于从整体上增强二值化权重本身的表示能力。BinaryMoS^[43]提出了一种尺度混合(Mixture of Scales)机制,它为模型配备了多个缩放因子。在推理过程中,模型会根据当前的输入上下文对这些缩放因子进行线性组合,生成与当前token自适应的动态缩放因子。这种设计使得固定的二值化权重矩阵能够根据输入动态地调整其所表示的数值,从而在不影响模型高压压缩率特性的前提下极大

地提升了模型的表达潜力。由于自适应过程仅涉及缩放因子而非整个权重矩阵,该方法保持了与传统静态二值化方法相近的压缩效率,其效果甚至超越了部分2比特量化方法。

此外,优化算法在极端量化中也扮演着至关重要的角色。传统的量化后微调通常依赖于直通估计器(Straight-Through Estimator, STE),但其在极低比特场景下的有效性与最优性并未得到充分验证。针对此问题,研究者提出了PV-Tuning^[44]。该框架将量化参数的优化表述为一个统一的全局优化问题,能够同时更新权重的离散整数分配和可学习的连续参数,其目标是最小化量化模型与原始模型输出之间的KL散度等全局目标函数。该方法在1至2比特的向量量化任务中展现了超越传统微调技术的性能。

现有的单比特量化方法主要围绕如何缓解二值化带来的性能退化,增强二值权重的表示能力,以及如何优化训练与量化过程以适应极端离散化带来的优化困难展开探索。该方向未来的研究方向可能包括设计专门的硬件计算单元用于高效处理由部分二值化产生的混合精度数据格式,从而将理论上的加速潜力转化为实际的推理增益。此外,还可探索其他形式的动态适应性,例如动态偏置和根据输入上下文选择不同的二值化函数,以进一步缩小与高比特模型的性能差距。

3.1.9 性能对比

为了更直观地比较多种仅权重量化方法的实际效果,表4汇总了它们在Wikitext2数据集^[77]上的困惑度(Perplexity, PPL)表现,该指标数值越低表示模型性能越优。结果显示,在4比特量化场景下,主流方法如GPTQ和AWQ均能取得与16位模型相近的性能,表明当前4比特权重量化技术已相当成熟。进一步的改进方法,如QEP和QTIP#,通过更精细的优化策略在此基础上进一步降低了性能损失。

当量化精度降低至3比特时,模型性能面临更大的挑战。GPTQ等方法出现了不同程度的性能下降,而QUIP[®]和QTIP展现出更强的鲁棒性,其性能衰减幅度相对较小。

在2比特和1比特的极端量化条件下,性能差异进一步扩大。通用方法如GPTQ在此时的表现明显劣化,而面向超低比特量化的专用算法则展现出显著优势。例如,DB-LLM在2比特量化下保持了较高的性能,PBS²P在1比特量化下同样表现突

表4 不同仅权重量化方法在 Wikitext2数据集上的困惑度 (PPL)对比

方法	位宽	困惑度(PPL)	
		LLaMa-7B	LLaMa2-7B
-	16	5.68	5.47
GPTQ ^[7]	4	5.94	6.08
SPQR ^[18]	4	5.87	-
SqueezeLLM ^[20]	4	5.79	-
OWQ ^[21]	4	5.94	-
AWQ ^[27]	4	5.97	5.83
SignRound ^[28]	4	-	7.85
NeUQI ^[29]	4	-	5.8
QEP ^[30]	4	-	5.75
QuIP ^[32]	4	-	8.43
QUIP ⁺⁺ ^[33]	4	-	5.56
QTIP ^[35]	4	-	5.17
GPTQ ^[7]	3	7.55	-
SPQR ^[18]	3	-	6.2
SqueezeLLM ^[20]	3	6.32	-
OWQ ^[21]	3	6.41	-
SEMQ ^[22]	3	6.13	-
NeUQI ^[29]	3	-	6.56
QUIP ⁺⁺ ^[33]	3	-	5.79
AQLM ^[34]	3	-	5.46
QTIP ^[35]	3	-	5.28
GPTQ ^[7]	2	22.10	20.85
SEMQ ^[22]	2	9.83	-
DB-LLM ^[36]	2	7.59	7.23
PB-LLM ^[40]	2	20.61	20.37
GPTQ ^[7]	1	1.6e5	-
BiLLM ^[41]	1	49.79	-
PBS ² P ^[42]	1	7.36	-

出,显示出在极限压缩场景中的潜在应用价值。

从效率角度看,仅权重量化主要通过压缩模型参数存储、降低权重读取开销来提升推理效率。由表5可见,主流仅权重量化方法已在不同硬件平台上表现出较稳定的加速效果,说明该类方法在实际部署中能够较好兼顾实现成本、系统兼容性与推理效率,因此仍是当前大模型高效部署中应用最广泛的量化路线之一。

总体而言,随着量化比特数的降低,不同算法之间的性能差距逐渐扩大。这一趋势表明了高级量化技术在维持模型精度方面发挥的关键作用。

3.1.10 小 结

仅对权重进行量化是当前LLM量化技术中应用最广泛、研究最成熟的分支之一。该方向的研究

表5 仅权重量化方法的收益对比

方法	位宽	环境	收益
GPTQ ^[7]	4/3	A100 / A6000	相对FP16推理,速度提升约3.25倍(A100)和4.5倍(A6000)
SqueezeLLM ^[20]	3	A6000	相对FP16推理,速度提升最高约2.3倍
AWQ ^[27]	4	RTX 4090/Jetson Orin	相对FP16推理,速度提升最高约3.9倍(RTX 4090)和3.5倍(Jetson Orin)

演进主要围绕如何有效应对权重矩阵中对模型性能具有决定性影响的离群点展开,技术发展趋势呈现出从宏观到微观的特点。

早期代表性方法GPTQ^[7]通过引入二阶信息实现逐层误差补偿,确立了“量化-补偿”的优化范式,从而验证了低比特权重量化的可行性。随后,研究者意识到离群点是量化精度下降的主要根源,提出了以SpQR^[18]、OWQ^[21]为代表的混合精度量化方法。此类方法的核心思想在于为少量关键离群权重分配更高比特精度,实现了计算资源在不同重要性权重间的差异化配置。

与被动适应离群点的策略不同,OliVe^[23]、RSQ^[24]等工作采用结构化预处理的主动方案,在量化前通过正交变换、旋转等数学操作平滑权重分布削弱离群点影响,提升数据的量化友好性。在极端压缩场景中,由于信息损失严重,研究者通常融合多种思路,例如结合混合精度机制保护关键信息,或设计专门的优化算法以增强模型的表示能力。

总体而言,权重量化技术已从早期的通用逐层贪心算法,演化为更加精细化、系统化、结构感知的算法,能够灵活利用数据结构特征、敏感度分析及数学变换来提升性能。未来的研究趋势可能集中在更高效的离群点识别与处理机制以及与硬件深度协同的预处理方案设计上。

3.2 权重-激活联合量化

仅对权重进行量化的目标在于降低模型的存储开销与内存带宽需求。这一策略对于将参数规模庞大的LLM部署在内存资源有限的环境中至关重要。然而,其局限性也同样明显。由于激活值在计算过程中仍保持高精度,在执行矩阵乘法等核心运算时,低比特的权重需要被反量化回高精度浮点数。这个额外的反量化步骤不仅会引入额外的计算开销,还

使得计算核心无法利用现代AI硬件中高度优化的低比特整数运算单元。因此,仅权重量化虽然降低了存储开销,但在计算层面的加速效果相对有限。为突破这一瓶颈,研究者们开始探索对权重和激活值进行联合量化。其核心思想是将模型中的权重和激活值同时转换为低比特整数表示,从而使整个计算流都可以在高效的整数域内完成,最大限度地发挥硬件的性能潜力。然而,与静态且分布相对稳定的权重不同,激活值的数值分布是动态的、依赖于模型输入,并且常常出现极端离群点,这给量化带来了严峻的挑战。因此,如何设计有效的方案来处理激活值的动态特性,在实现计算加速的同时避免严重的精度下降,是权重-激活联合量化研究的重点。

3.2.1 动态范围调整

为应对激活值中的离群点导致的量化性能减弱问题,一个核心思路是通过数学上的等价变换来调整数值的动态范围,从而平滑数据分布,使之更适合量化。SmoothQuant^[45]是这一方向的开创性工作,它观察到不同token通道维度上表现出相似的分模式。基于此,它提出了一种数学上等价的逐通道缩放变换,对于任意的矩阵乘法 $Y=XW$,可以引入一个对角缩放矩阵 S ,进行如下变换:

$$Y=XW=(XS^{-1})(SW)=\hat{X}\hat{W} \quad (5)$$

其中 $\hat{X}=XS^{-1}$ 是变换后的激活, $\hat{W}=SW$ 是变换后的权重。由于该变换是恒等变换,模型的数学功能在浮点精度下保持完全不变。SmoothQuant的目的是选择合适的 S ,使得 $\hat{X}$ 的动态范围变得更小、更易于量化,从而能够在保持浮点精度下模型输出不变的同时降低激活值的动态范围,将难以量化的激活值离群点迁移至更易于处理的权重张量上。

在此基础上,OmniQuant^[46]提出了一套更全面的可学习优化框架。其包含两个核心组件:可学习的权重裁剪(Learnable Weight Clipping, LWC)和可学习的等价变换(Learnable Equivalent Transformation, LET)。LWC通过优化裁剪阈值来调整权重的极端值,而非简单使用最大最小值;而LET则通过将等价变换的缩放因子变为可学习参数,以应对激活值中的离群点。整个过程在一个基于块误差最小化的可微框架下进行端到端优化,使其能够高效地为纯权重和权重激活两种量化场景找到最优量化方案。

另一项工作深入分析了离群点不仅在通道间呈现非对称性,而且还集中于特定通道的特征,并基于

此设计了更精细的迁移策略^[47]。该方法提出统一的通道级移位和缩放操作,前者用于消除激活值分布的非对称性,后者则抑制离群通道的数值范围。该工作设计了一套统一的迁移模式将移位和缩放的逆操作无缝地迁移到后续模块中。

与简单的缩放迁移不同,QLLM^[48]认为现有的迁移方法在处理低比特量化时效果有限或存在梯度不稳定的问题,因此提出了一种更为灵活的自适应通道重组技术。该技术旨在将离群值的幅度重新分配到其他通道,以平衡激活值的幅度分布。具体而言,它首先通过通道分解将包含离群值的通道拆分为多个数值范围更小的子通道;随后执行通道合并并将分布相似的通道进行聚合,以维持原有的通道数量和计算效率。此外,该方法还设计了一种自适应策略,能够自动确定通道分解时最优的子通道数量,从而更精细地处理极端离群点的影响。

为了更好地适应硬件加速,MergeQuant^[49]为解决逐通道激活量化因引入额外计算而导致的推理延迟问题,提出了一种量化步长迁移方法,它能够避免额外的量化步骤开销,使得计算过程能更好地适配高效的整数加速计算核。为了应对该迁移方法下激活值量化缩放因子的不均匀性,该工作进一步引入了维度重构和自适应裁剪。前者用于均衡不同通道间的缩放因子,后者则旨在减少每个通道内离群值的影响。

per-IC^[50]方法观察到,激活值中的离群点通常出现在特定的维度上,并沿输入通道方向对权重矩阵产生影响。因此,该方法打破了按输出通道分组成量化的常规做法,提出按输入通道对权重进行分组。这样一来,由激活离群点引发的权重离群点就能被隔离在更小的量化组内,从而获得更紧凑的量化范围,提升整体精度。

上述方法的核心思想是将激活值中难以处理的离群点问题,通过数学等价变换迁移到相对静态且更易处理的权重张量上。从最初固定的缩放因子发展到可学习的、更复杂的变换,其粒度越来越精细,优化目标也越来越明确。未来这一方向的研究应该会更加关注变换本身的复杂性与硬件实现成本之间的权衡。例如,开发更加轻量级但同样有效的自适应变换机制,或者设计能够与特定硬件指令集相匹配的变换策略,以提升该类方法实用性。

3.2.2 基于几何变换的优化

除了调整数值的动态范围,部分研究工作探索了更复杂的变换来直接优化权重和激活值的分布形

态,使其整体更加均匀,从而提升量化友好性。其中,几何变换是一个重要的研究方向。

研究发现,对激活或权重矩阵进行旋转有助于消除离群点。DuQuant^[51]针对普通离群点和极端离群点提出了双重缓解策略。它首先利用特定离群点维度的先验知识构建旋转矩阵,通过块级旋转将离群点重新分布到相邻通道;随后,它采用Z字形置换(Zigzag Permutation)来平衡不同块之间的离群点分布,以降低块间的方差,最后再进行一次旋转以进一步平滑激活值的整体分布。

SpinQuant^[52]注意到,即便是随机的旋转变换也能对量化性能产生巨大影响。因此,该工作将旋转本身变为优化目标,提出引入可学习的旋转矩阵,并将其参数化以在训练中进行优化,从而为模型寻找能够最大化量化后精度的最优旋转状态。

OSTQuant^[53]则提出了一个更为系统的框架。它引入了量化空间利用率作为度量指标,用于评估变换后的数据对量化范围的利用效率。量化空间利用率被定义为数据点占据的体积与对应量化空间体积的比率,其中量化空间被建模为一个超立方体,其边长由所有维度中的最大量化范围决定。然后,为每个全连接层分配一个由正交变换和缩放变换组成的可学习等价变换对。在优化过程中,该变换对以最大化量化空间利用率为目标,全局优化权重和激活值的分布,最后将学习到的变换及其逆变换无损地融入对应的全连接层中,从而保持原有的计算图结构。

几何变换类方法通过旋转、置换、正交变换等操作,从更高维度上重塑数据分布,旨在将集中的离群点分散到更多维度中,使整体分布更加均匀,从而降低所有通道的量化难度。这类方法提供了一个全新的视角来解决离群点问题。展望未来,该方向的挑战在于寻找最优变换的巨大计算开销。未来的研究可能会集中在如何设计低成本的、参数化的几何变换,或者利用数据本身的结构特性来指导变换矩阵的构建,从而在不进行昂贵搜索的情况下获得接近最优的变换效果。同时,如何将这些变换高效地融合到现有计算图中也是一个重要的实践问题。

3.2.3 混合精度量化

另一大类方法借鉴了权重量化中的思想,即不对所有数值进行一致的低比特量化,而是识别出对模型性能至关重要的离群点,并对它们进行隔离和特殊处理,通常保留更高精度。

LLM.int8()^[54]设计了一套两步量化流程:首

先,对于矩阵乘法中的大部分特征,采用向量级INT8量化,其核心思想是将矩阵乘法视为行向量与列向量之间一系列独立的内积运算,因此可以为每个内积操作使用独立的归一化常数以提高精度;其次,对于激活值中的离群点,该方法将这些离群特征维度分离出来,送入一个独立的FP16高精度计算路径。通过这种混合精度分解方案,绝大多数计算可以在INT8下完成,从而在几乎不损失性能的前提下实现显著加速。QUIK^[55]采用了类似的思想,通过将权重和激活矩阵都分解为可用4比特精度处理的基础分量和包含离群点、需要用INT8或FP16等更高精度处理的离群分量,将混合精度策略扩展至更低比特数。

QRazor^[56]则提出了一种先量化再压缩的两阶段方法。第一阶段中,该方法将权重、激活和键值缓存量化到一个性能几乎无损的基础精度。第二阶段中,通过位运算、截断和四舍五入等高效操作,仅保留基础精度的参数中的少数几个关键比特,从而在有效保留原始数据的关键分布特征的前提下实现高效压缩。

混合精度与离群点隔离方法核心思想是通过一个独立的、高精度的计算路径来保护离群点,从而保证了模型的整体性能。这种策略在实践中被证明非常有效。该方向未来的研究重点应着眼于如何更高效、更动态地识别和分离离群点。当前的策略大多基于固定的阈值,未来的方法可能会探索基于数据分布动态变化的自适应离群点检测机制。

3.2.4 量化过程优化

除了对数据分布本身进行处理,研究者们也从模型优化的角度出发,通过引入更高层面的信息(如知识、梯度、层间依赖)来指导和优化整个量化过程,以期达到更好的量化效果。

ZeroQuant^[57]提出了一种逐层知识蒸馏方法用于INT4和INT8混合精度量化。该方法在权重上采用分组量化、在激活值上采用逐token量化等细粒度的硬件友好量化方案,通过让每一层的低比特网络学习其对应的高精度网络层的输出来恢复量化模型的精度。这种方法的优势在于仅需极少的迭代次数,甚至在无法访问原始训练数据的情况下也能有效执行。

GuidedQuant^[58]则引入了来自任务最终损失函数的梯度信息。该方法首先利用关于层输出的最终损失梯度计算每个层输出的显著性分数,然后通过最小化由该分数加权的层输出误差来独立量化每一层。

这种方法等价于使用块对角 Hessian 矩阵进行更精确的二次误差近似,能够保留输出通道内的依赖关系,从而比传统的最小化激活误差方法更准确地估计量化对模型最终性能的影响。

考虑到简单的逐层优化忽略了层间的相互影响,有研究者提出挖掘跨层依赖性^[59]。该工作发现激活值的熵与整个模型的量化误差传播具有强相关性。因此,它利用熵作为指标对网络进行最优块划分,优化了量化策略的搜索成本和量化误差之间的权衡。

这类方法不再将量化视为一个孤立的数据重编码过程,而是将其置于整个模型优化框架内。通过利用知识蒸馏、任务梯度、层间关系等信息进行量化,从而在宏观层面最小化精度损失。在未来,构建一个能够端到端、全局性地联合优化所有层量化参数的框架将是最终目标。这需要解决巨大的优化空间和计算复杂度问题。探索更高效的二阶优化算法或者利用元学习来自动发现最优的量化策略,可能是重要的研究方向。

3.2.5 新型量化方法

除了上述主流方向,研究者们还从一些独特角度切入,提出了多种新型权重-激活联合量化方案。

Norm Tweaking^[60]聚焦于 LLM 中的层归一化 (LayerNorm) 层,通过仅微调其参数并冻结其他所有权重,同时结合受约束的数据生成方法来减少对特定校准集的依赖。通过最小化量化模型与浮点模型在激活分布上的通道级差异,在权重和权重激活联合量化上均取得了显著提升。

LLM-FP4^[61]认为相较于整数,浮点数量化格式因其更灵活的表示范围,能更好地处理激活值的长尾或钟形分布。该工作观察到激活值分布中存在“通道间方差大,通道内方差小”的模式,提出激活值的逐通道缩放因子可以通过重参数化技巧,在推理时被无缝地吸收到权重的指数偏置项中,从而在不增加额外计算开销的情况下大幅提升了 4 比特浮点量化的效果。

LQER^[62]则将量化与低秩近似技术相结合。它将一个训练好的权重矩阵近似为三个矩阵之和:一个低精度但高秩的量化矩阵,以及两个用于修正量化误差的高精度但低秩的矩阵。这两个分量都具有低计算成本的特点,其组合能够有效弥补量化带来的性能损失。

CushionCache^[63]则通过改变模型的输入来减少离群激活值。该方法首先通过贪心搜索离线地寻找

一个合适的提示词序列,该序列的目标是最小化后续 token 中的最大激活值;然后在线推理时进一步微调由该序列产生的键值缓存。其核心思想是利用该键值缓存来使得后续真正任务相关的 token 所产生的激活值分布更加平滑且易于量化,从而有效缓解离群点问题。

上述研究工作展现了量化研究的广度,其方法涵盖了模型微调、数值表示格式、与其他压缩技术的结合等多个层面。展望未来,这些新颖思路的融合可能会带来更大的突破。例如,将低秩近似与混合精度方法结合,用低秩矩阵专门处理离群点;或者将浮点数量化格式与等价变换技术结合,进一步提升对动态范围的适应性。

3.2.6 性能对比

表 6 汇总了多种主流权重-激活联合量化方法在采用 4 比特量化策略 (W4A4) 时 Wikitext2^[77]上的困惑度性能。SmoothQuant 方法验证了迁移激活离群点思路的可行性,但性能损失较大。后续工作通过引入更精细的优化策略显著提升了性能。例如,OmniQuant 通过可学习的权重裁剪和等价变换将困惑度大幅降低。基于几何变换的优化是另一个高效的思路,如 SpinQuant 利用旋转变换平滑数据分布,展现了强大的性能。最新的方法如 OSTQuant 通过正交变换和更系统的优化框架进一步将 4 比特联合量化模型的性能推向了新的高度,其困惑度已非常接近甚至优于部分仅权重量化方法。

在收益方面,与仅权重量化相比,权重-激活联合量化不仅能够减少模型存储开销,还能更充分地

表 6 主流权重-激活联合量化方法在 Wikitext2 数据集上的困惑度 (PPL) 对比

方法	困惑度 (PPL)	
	LLaMa-7B	LLaMa2-7B
SmoothQuant ^[45]	-	254.50
OmniQuant ^[46]	-	14.61
Outlier Suppression+ ^[47]	14.17	-
QLLM ^[48]	-	11.15
MergeQuant ^[49]	-	6.09
DuQuant ^[51]	6.40	6.28
SpinQuant ^[52]	6.08	8.28
OSTQuant ^[53]	6.04	5.60
QUIK ^[55]	-	5.84
QRazor ^[56]	-	6.45

发挥低比特计算的硬件优势。由表 7 可见,主流方法在推理速度和显存占用方面均已取得明显收益,表明该类方法更有利于将量化压缩进一步转化为实际的系统性能提升,因而具有更强的部署价值。

表 7 权重-激活联合量化方法的收益对比			
方法	位宽	环境	收益
SmoothQuant ^[45]	W8A8	A100	相对 FP16 推理,速度提升最高约 1.56 倍;显存占用降低约 2 倍
QUIK ^[55]	W4A4	RTX 3090	相对 FP16 推理,端到端吞吐提升最高约 3.4 倍
ZeroQuant ^[57]	W8A8	A100	相对 FP16 推理,速度提升约 3.67 倍

3.2.7 小 结

权重-激活联合量化旨在同时对权重与激活值进行低比特整数化,从而实现真正的计算加速,而不仅仅节省存储空间。其核心挑战在于激活分布会随输入内容剧烈变化,因此量化难度远高于静态权重。针对这一问题,研究者提出了多条技术路径。

其中,SmoothQuant^[45]作为开创性工作,提出了通过数学等价变换将量化难度从动态的激活值迁移至静态的权重部分,从而有效简化了量化过程。另一类方法则从更高维度重新审视量化问题,如 DuQuant^[51]与 SpinQuant^[52],它们通过可学习的旋转、置换等几何变换,直接对权重与激活的联合分布进行重塑,使整体分布更均匀、易于量化。

此外,借鉴权重量化的经验,LLM.int8()^[54]等方法引入混合精度思路,将激活中离群维度独立识别并以高精度路径计算,以确保关键语义信息不受损失。与此同时,ZeroQuant^[57]等方法引入知识蒸馏与梯度指导机制,从模型输出或任务损失角度对量化进行约束与优化,从而系统性地弥补精度损失。

总体来看,权重-激活联合量化的技术演进已从独立处理单个张量转向关注权重与激活之间的交互关系。未来的研究方向将聚焦于如何在不显著增加计算开销的前提下,实现更自适应的动态范围校准,以进一步提升联合量化的通用性与稳定性。

3.3 键值缓存量化

通过对模型权重和激活进行联合量化,研究者们成功地将计算密集型的矩阵乘法从高精度浮点运算转换为高效的低比特整数运算,显著提升了模型的计算速度。然而,LLM 的推理性能不仅受限于计

算,在推理过程中,为了避免对历史上下文的重复计算,模型会缓存每一个已生成 token 的 Key 和 Value 向量,形成所谓的 KV 缓存。这个缓存的大小与输入序列的长度成正比。随着序列长度的增长,其规模甚至会远超模型权重本身,成为限制 LLM 处理长序列能力的核心障碍。因此,对 KV 缓存进行量化是提升 LLM 长序列推理效率的关键,但这面临着与权重量化截然不同的挑战。首先,KV 缓存的数据分布具有强动态性和输入依赖性。其数值不仅随输入序列的不同而变化,更在自回归生成的每一个时间步动态更新,导致其分布难以预测,尤其容易出现与特定 token 相关的离群值。其次,量化方案的设计必须与硬件执行模型紧密适配。过于复杂的量化策略,如极细粒度的参数计算或不规则的数据访问模式,虽能提升理论压缩率,却可能引入额外的计算和控制流开销,反而抵消了因减少内存带宽而带来的性能增益。因此,对 KV 缓存进行高效量化,是提升 LLM 推理效率、拓展其应用边界的关键技术。

3.3.1 非对称量化

对 KV 缓存分布特征的分析是实现有效量化的基础。研究表明,键缓存与值缓存的数值分布呈现出显著的异质性:键缓存存在通道维度上具有相似的分

布模式,而值缓存则在 token 维度上表现出更高的一致性。基于这一关键洞察,KIVI^[64]提出了一种非对称量化方案,即对键缓存采用通道级量化,而对值缓存采用 token 级量化。然而,通道级的键缓存量化需要跨越多个 token 进行操作,与 LLM 的自回归推理不适配。为此,KIVI 将缓存动态划分为已构成完整分组的缓存和尚不满足分组条件的剩余缓存,并仅对前者进行分组量化,而后者则保留全精度,在实现量化的同时高效支持了自回归推理。尽管 KIVI 通过非对称策略有效应对了键和值缓存的分布异质性,但其设计仍存在局限性。首先,为了兼容自回归生成,KIVI 保留了末端剩余缓存的全精度,这部分未量化的数据会持续占用额外的内存和带宽。其次,KIVI 基于经验观察为键和值缓存预设了不同的量化粒度。然而不同模型架构或特定任务下的数据分布可能不完全符合此假设。当键缓存在 Token 维度或值缓存在通道维度出现显著离群点时,这种固定的策略可能无法达到最优量化效果。

此类方法的核心在于精确捕捉并利用键和值缓存内在的、非对称的数据分布特性,通过施加不同粒度的量化策略来实现更高的压缩效率。未来的研究可以进一步探索更深层次的分布规律,例如特定层、

特定注意力头乃至特定 token 位置的分布变化,从而设计出更加动态和自适应的量化策略。

3.3.2 离群值处理策略

与权重或激活量化过程类似,离群值的存在同样是限制 KV 缓存量化性能的主要障碍。针对此问题,一种直接的策略^[65]是在解码时动态地识别这些异常 token,并将其作为离群点排除在量化过程之外,以全精度形式保留,从而有效提升了模型的整体性能。然而,硬件实现成本和效率限制了该方法的实际应用。一方面,该方法在每个生成步骤中都需要运行离群点检测逻辑,带来了额外的计算负担。另一方面,离群点的隔离策略导致了 KV 缓存的存储格式高度不规则。这种非结构化的数据布局会严重降低内存吞吐量和计算单元的利用率,最终导致访存模式退化并带来性能损失。

KVQuant^[66]则首先提出应沿着通道维度对键激活值进行量化,以更好地匹配其数据分布;其次,在应用旋转位置编码(Rotary Positional Embedding, RoPE)前对键缓存进行量化,以规避 RoPE 对数值分布的复杂影响;随后,它基于对角 Fisher 信息矩阵和激活量化误差,离线计算各层的敏感度权重,并据此设计出更能有效表征数据分布的非均匀量化数据类型;最后,它为每个向量独立地执行密集-稀疏量化,将离群值从数据主体中分离,最大限度地减少了其对量化过程的干扰。

另一项工作^[67]则分析了现有旋转变换方法将相同旋转应用于所有注意力头的局限性,并提出了离群值感知旋转算法,该算法在保持快速 Walsh-Hadamard 变换计算效率的同时,增强了对不同离群值分布的适应性。

IntactKV^[68]发现了一种先前被忽视的新型离群点,它并非在所有输入中都出现,而是仅在处理某些关键 token 时,在键和值的投影权重中产生,使自注意力机制过度关注这些关键 token。这些关键 token 包括序列开始符[BOS]、逗号、句号等。为此,IntactKV 提出在推理时,使用全精度模型为这些关键 token 生成无损的键值缓存,并将其注入到量化模型的解码过程中。通过保持这部分关键信息的完整性,IntactKV 有效阻止了量化误差在后续生成步骤中的累积和放大,其有效性也得到了数学分析的支持,证明该方法能有效降低量化误差的上限。

应对离群值的策略通过动态识别并隔离,预先进行数据变换以规避离群值的影响,以及设计能够分离主体数据与离群值的量化算法来处理少数关键

数值,显著降低了量化误差,保护了模型性能。离群值的动态检测和处理会引入额外的计算开销,如何在显著影响推理速度的前提下高效完成这一过程是未来的研究重点。

3.3.3 变换域量化

部分研究通过可逆变换将 KV 缓存映射至一个更适于量化的新空间,以此来降低量化误差。SVDq^[69]利用奇异值分解将键缓存变换到由基向量构成的空间。由于数据在不同基向量上的分量迅速衰减,该方法为少数能量集中的重要通道分配更高的比特率,实现重要性感知的混合精度量化。在理论上,该工作证明了 SVDq 导致的量化误差远低于在原始空间中进行逐通道量化的方法。

PolarQuant^[70]通过递归算法将 KV 嵌入变换到极坐标系,并对角度进行量化。经过随机预处理后,变换后的角度呈现出一个具有解析形式的、紧凑且高度集中的分布,使得量化过程不再需要存储额外的缩放因子和零点参数,从而显著节省了传统量化方法中因这些参数带来的内存开销。

NSNQuant^[71]针对现有矢量量化方法对校准数据集的依赖性及其引发的分布漂移问题,设计了一个包含“归一化-平移-再归一化”的三步变换流程,并结合 Hadamard 变换,能有效地将任意输入 token 的分布对齐至标准正态分布。这使得仅需一个可重用的、无需校准数据生成的码本即可实现稳健的低比特矢量量化。

此类方法的核心思想是通过 SVD、极坐标变换或正态分布对齐等数学工具,将原始 KV 缓存数据投影到一个分布更集中、冗余度更低或更符合量化器假设的新空间,从而在根本上降低量化难度和误差。变换过程本身的计算复杂度是该技术路线需要权衡的关键点。未来的研究方向包括探索更高效的线性或非线性变换方法,以及如何将变换域量化与离群值处理、非对称量化等技术有机结合,实现性能的进一步突破。

3.3.4 针对特定场景的优化

一些研究为特定应用场景提出了专门的优化方案。有工作^[72]发现 KV 缓存在相邻层之间以及同一注意力层内部的键和值之间存在相互依赖关系。基于此,该工作通过训练线性预测器来利用一部分缓存数据来预测另一部分。通过结合了免数据矢量量化技术补偿预测过程中的误差,该方法在相同位宽下实现了更优的压缩与精度权衡。

针对长上下文推理场景,PM-KVQ^[73]指出,现

有方法的量化误差在解码过程中逐步累积,导致性能严重下降;并且基于短文本的校准无法捕捉到长序列中的低频通道分布。为此,PM-KVQ设计了渐进式混合精度量化策略,为模型中敏感度更高的层保留更多精度。此外,为解决校准问题,该方法还提出了一种利用位置插值技术,以短文本数据来近似长文本的数据分布的策略。

在多模态领域,VidKV^[74]探索了视频语言模型的KV缓存量化。具体而言,它对键缓存的异常通道和正常通道分别执行2比特量化和结合FFT的1比特量化;对值缓存则在执行1.58比特量化的同时,有选择地保留对语义理解更重要的视觉token。该研究发现,对于视频模型的值缓存采用通道级量化的效果优于LLM中通用的token级量化。

上述研究表明,通用的KV缓存量化方案无法在所有场景下达到最优。随着LLM在更多垂直领域的深度应用,为特定任务和数据模态设计专门的KV缓存压缩方案将成为一个重要趋势。未来的工作需要更深入地分析不同场景下KV缓存的行为模式,并开发出能够捕捉这些特性的高效量化算法,实现算法与应用的协同设计。

3.3.5 性能对比

为对比不同KV缓存量化方法的性能,表8比较了多种2比特KV缓存量化方法在Wikitext2^[77]困惑度和GSM8K数据集^[78]上的零样本准确率。在Wikitext2数据集上,所有先进的2比特量化方法都取得了较低的困惑度,表明其能够有效压缩显存占用而不严重影响模型的文本生成能力。而在更能体现模型复杂推理能力的GSM8K数学任务上,不同量化方法之间的性能差异更为明显。例如,KIVI采用的非对称量化策略在该任务上表现优异,超越了在Wikitext2数据集取得更好结果的RotateKV。这种多维度的评估结果表明,KV缓存量化方法的选择应充分考虑具体的应用场景,以在压缩率和任务性能之间取得最佳平衡。

表8 主流键值缓存量化方法在 Wikitext2 和 GSM8K 数据集上的性能对比

方法	位宽	Wikitext2	GSM8K
-	16	5.12	-
KIVI ^[64]	2	6.14	16.30
OTT ^[65]	2	-	21.38
KVQuant ^[66]	2	5.52	-
RotateKV ^[67]	2	5.50	13.95
NSNQuant ^[71]	2	5.29	-

在收益方面,由表9可见,KV缓存量化不仅能够有效降低显存占用,还能够在一定程度上提升吞吐量和解码速度,说明该类方法的作用不仅局限于存储压缩,而且能够进一步改善长上下文场景下的整体推理效率。

表9 KV 缓存量化方法的收益对比

方法	位宽	环境	收益
KIVI ^[64]	2	A100	峰值显存占用降低约2.6倍; 吞吐提升约2.35倍—3.47倍 相对FP16推理,KV相关矩阵-向量计算速度提升最高约1.7倍
KVQuant ^[66]	2	A100	峰值显存占用降低约3.97倍; 解码速度提升约2.32倍
RotateKV ^[67]	2	RTX 4090	

3.3.6 小 结

键值缓存量化是针对LLM在自回归推理过程中内存瓶颈问题的关键优化手段。其特殊挑战源于KV缓存的动态增长特性及其结构化分布特征,包括Key与Value分布的异质性以及不同Token在时序上的重要性差异。

针对这些问题,如KIVI^[64]的主流方法利用数据分布的非对称性,通过对Key与Value采用不同的量化粒度以适配各自分布特征,实现高效压缩。另一类方法^[65]则基于动态离群值检测与隔离机制,在推理阶段实时识别关键Token对应的离群缓存,并以全精度形式保留,以防止信息丢失对后续生成造成影响。

此外,SVDq^[69]、PolarQuant^[70]等工作提出了变换域量化的思路。它们认为原始表示空间并非最优量化域,因此通过SVD分解、极坐标变换等可逆数学映射,将KV数据映射至能量更集中、分布更紧凑的空间再进行量化,进一步降低了信息损失与量化误差。

总体而言,KV缓存量化技术正从早期的通用量化策略向深度挖掘其结构特征与动态特性的专用解决方案演进。未来研究方向可能集中在低成本的变换域量化方法以及能够感知长序列中语义重要性变化的自适应量化策略,以支持更长上下文的高效推理。

4 量化感知训练

尽管PTQ因其部署高效且无需额外训练成本而受到广泛关注,但在极低比特场景下往往面临性能急剧下降的难题。其根本原因在于,量化作为

一种有损压缩过程,随着比特数的降低,量化误差显著增大,而PTQ方法无法在量化过程中调整模型权重以补偿由此带来的误差。为克服这一挑战,量化感知训练(Quantization-Aware Training, QAT)应运而生。QAT在训练或微调过程中引入模拟量化操作,将量化误差作为训练目标的一部分进行优化,从而使得模型参数能够主动适应量化带来的影响。为了解决量化操作本身不可导、无法直接反向传播梯度的问题,QAT普遍采用直通估计器(Straight-

Through Estimator, STE)技术。该方法在反向传播时直接绕过取整函数,将其梯度近似为1,从而使误差信号能够回传并指导模型参数进行适应性更新。通过这种方式,QAT能够在低比特设置下获得显著优于PTQ方法的模型性能,但这一优势以训练开销和计算资源的大幅增加为代价。因此,如何提升QAT的效率、稳定性及其在极低比特下的性能,成为当前研究的核心议题。表10总结了量化感知训练的现有工作。

表 10 量化感知训练的现有工作

分类	描述	现有工作
训练效率优化	通过分阶段训练、参数分组和优化器改进等手段提升训练效率	EfficientQAT ^[8] , DL-QAT ^[79]
训练稳定性提升	使用正则化损失、改进梯度估计等技术提高训练稳定性	FPQ ^[80] , QuEST ^[81]
基于知识蒸馏的量化	结合知识蒸馏的方法提升低比特量化模型的性能	BitDistiller ^[82] , LLM-QAT ^[83]
极低比特量化	1比特及1.58比特等极低比特下的QAT方法	BitNet ^[84] , BitNet b1.58 ^[85] , Nielsen等 ^[86] , BitNet v2 ^[87]
宏观分析	量化性能测评框架和缩放定律	ParetoQ ^[88] , Chen等 ^[89]

4.1 训练效率优化

针对QAT高昂的训练成本,研究者提出了多种优化策略以提升其训练效率。例如,传统QAT方法需要以端到端的方式联合训练全精度权重和量化参数,导致内存开销巨大。为提升训练效率,EfficientQAT^[8]提出了一种两阶段策略:首先进行块级别的所有参数训练,即逐块地使用重构损失来训练模型参数和量化参数;随后进入端到端的量化参数训练阶段,该阶段固定已量化权重,仅在目标数据集上对量化步长进行微调,从而大幅降低了训练过程中的资源消耗。

另一项工作DL-QAT^[79]将量化权重的优化过程分解为分组幅值训练和在预定义量化空间内的权重微调两个部分。通过引入一个幅值项来校准每个量化组的整体尺度以确保模型参数更精确的表示,并利用低秩矩阵来微调量化权重,DL-QAT能够更高效地提升模型在下游任务中的适应性和性能。

此类工作的核心技术点在于分解QAT的联合优化问题。通过将训练过程划分为独立的阶段或将参数更新分解为不同部分,显著降低了单次迭代的内存占用和计算复杂度。未来的研究方向包括探索更自适应的训练调度策略,动态分配计算资源给模型的不同部分;开发专门用于QAT的内存高效优化器,减少存储中间状态的需求;以及研究混合精度训练策略在QAT中的应用,即对不同层或模块采用不同的训练效率优化技术。

4.2 训练稳定性提升

除了提高训练效率,研究者们也致力于提升

QAT过程本身的稳定性和准确性。部分研究发现,由量化函数引入的梯度不连续性和损失函数的不平滑性是导致QAT性能不稳定的关键因素。尽管理论分析表明对特征施加扰动会带来更平坦的局部最小值,但直接扰动权重或特征的简单做法反而会降低模型性能。为克服这一挑战,FPQ^[80]提出通过对Hessian矩阵施加隐式正则化以平滑损失,有效缓解了训练不稳定问题。

QuEST^[81]则从两个方面改进了QAT:首先,它通过Hadamard变换进行归一化并结合均方误差最优拟合,实现了对权重和激活值分布的精确、快速的量化;其次,它提出了一种名为“信任度估计”的新型梯度估计器,该估计器旨在最小化高精度权重真实梯度与低精度权重估计梯度之间的差异,从而实现了在1比特权重和激活下的稳定训练。

通过改善优化目标,即通过正则化等手段平滑损失曲面,降低其曲率,使优化过程对参数扰动不敏感;以及改进优化过程,即设计更精确的梯度估计器,以更准确地将梯度信息反向传播通过非可微的量化节点,能够提升QAT的稳定性。此方向未来的研究重点可能在于开发适用于LLM的二阶优化方法,以更全面地利用损失函数的信息;为梯度估计器的设计提供更强的理论保证,分析其收敛性与无偏性;以及研究量化噪声与优化器动态之间的相互作用机制。

4.3 基于知识蒸馏的量化

为进一步提升低比特模型性能,知识蒸馏成为一种与QAT结合的有效途径。BitDistiller^[82]框架即为典型代表,它将QAT与知识蒸馏相结合用于

提升低于4比特量化模型的性能。该框架不仅采用定制化的非对称量化与裁剪策略以最小化量化误差,还设计了一种简洁高效的自蒸馏方法,以全精度模型为教师指导低比特学生模型学习,显著提升了模型在语言理解及复杂推理任务中的表现。

此外,针对原始训练数据不可得的场景,有研究者提出了一种数据无关的蒸馏方法^[83]。该方法首先利用预训练的全精度模型生成数据,然后将全精度模型的预测作为标签来指导量化模型的微调过程。通过这种方式,研究者成功地将LLaMA系列模型的权重和KV缓存量化至4比特,且较好地保持了原模型的输出分布。

知识蒸馏在QAT中扮演了正则化和提供监督信息的角色。通过引入一个高精度的教师模型,低比特的学生模型可以学习到超越训练数据标签本身的暗知识(dark knowledge),从而弥补因参数量化而损失的模型容量和表达能力。数据无关蒸馏技术则解决了在数据隐私或数据受限场景下应用QAT的难题。此方向未来的研究可以探索更精细化的蒸馏目标,如匹配中间层特征图的分布或注意力矩阵;设计对量化更友好的教师模型,使其知识更易于被低比特学生模型拟合;探索多教师蒸馏或在线蒸馏框架在QAT中的应用,以进一步提升性能。

4.4 极低比特量化

在极低比特量化领域,特别是1比特模型方向,已涌现出了一系列开创性工作。BitNet^[84]首次将QAT应用于1比特LLM,其核心思想是在训练期间保持优化器状态和梯度的高精度,但将模型权重进行二值化,并将激活值量化为8比特。其实现方式相对简单,仅需替换Transformer中的线性投射层,并采用STE来处理反向传播中的梯度近似问题。在此基础上,BitNet b1.58^[85]提出了一种1.58比特的变体,将每个参数量化为-1、0或+1的三元值。它通过一种均值绝对值量化函数,先将权重矩阵按其平均绝对值进行缩放,再将每个值四舍五入到最近的三元值,从而在几乎不增加模型大小的情况下提升了模型的表达能力。针对此类模型的训练,一项研究^[86]发现,采用从16位精度预训练再过渡到1.58位QAT的混合训练策略,相比于完全从头开始进行1.58位训练,能使模型性能更接近于原始的16位模型。最新的BitNet v2^[87]进一步解决了1比特LLM中激活值量化的难题。其提出H-BitLinear模块,在激活量化前应用在线Hadamard变换,将尖锐的激活值分布平滑成更接近高斯分布的形态,使其

更适合低比特表示,从而成功实现了原生的4比特激活值量化,提升了模型的整体性能。

上述方法展示了极低比特量化场景的核心技术点,包括参数的三元化扩展、渐进式量化训练策略以及利用变换来重塑激活分布以降低量化难度。该领域的关键挑战和未来方向包括实现权重和激活均为极低比特模型的稳定训练和高性能;以及从理论上分析单比特网络的性能边界,并探索新的网络结构以弥补其与全精度模型的差距。

4.5 宏观分析

部分研究从宏观角度分析并指导QAT设计。ParetoQ^[88]提出了一个跨不同比特宽度直接且一致地比较不同方法性能的量化框架。其分析表明,与常见的4比特量化相比,1.58比特、2比特和3比特量化在模型精度和有效模型大小之间实现了更优的权衡,展示了它们在优化模型部署方面的巨大潜力。

另一项研究则提出了一个统一的QAT缩放定律^[89],将量化误差建模为模型大小、训练数据量及量化组大小的函数。该研究发现,量化误差随模型规模增大而减小,随训练数据增多与量化粒度变粗而增大。通过误差源分解,该研究指出,特别是第二层全连接层中的激活误差,是4比特权重与4比特激活(W4A4)设置下的QAT性能主要瓶颈。该研究还进一步发现,随着训练数据量的增加,权重量化误差的增长速度会超过激活量化误差,这表明在数据充足的情况下,降低权重量化误差也变得至关重要,为未来的QAT优化指明了方向。

上述研究为QAT领域提供了系统性的分析工具和理论基础。在未来,研究者们可以进一步将缩放定律扩展到更多变量,如训练步数、学习率和模型架构参数;建立从量化误差到具体下游任务性能损失的理论联系;以及利用这些理论框架进行量化策略的自动化设计,在训练前预测并选择最优的量化配置。

4.6 性能对比

量化感知训练通过在训练过程中模拟量化误差,旨在获得比训练后量化更高的模型精度。表11展示了多种主流QAT方法在不同比特设置下,于Wikitext2数据集^[77]上的困惑度和多个常识问答任务上(WG.: WinoGrande^[90]; HS.: HellaSwag^[91]; ArcC.: ARC-challenge^[92]; PQ.: PIQA^[93])的平均准确率。在4比特和3比特场景下,EfficientQAT等方法展现出较好的性能。当比特数降至2时,ParetoQ方法实现了远超其余方法的性能。此外,如BitDistiller的

知识蒸馏技术也在低比特下取得了具有竞争力的性能,证明该技术是弥补性能损失的有效手段。

表 11 主流量化感知训练(QAT)方法在多任务上的性能对比

方法	位宽	Wikitext2	WG.	HS.	ArcC.	PQ.	Avg.
EfficientQAT ^[8]	4	5.53	69.2	57.0	42.3	64.3	58.2
DL-QAT ^[79]	4	6.30	68.8	74.6	44.6	78.2	66.6
EfficientQAT ^[8]	3	5.81	69.1	55.9	42.8	77.6	61.4
DL-QAT ^[79]	3	9.40	60.2	63.6	35.9	74.8	58.6
BitDistiller ^[82]	3	5.97	68.4	55.4	41.2	77.0	60.5
LLM-QAT ^[83]	3	6.02	68.4	54.7	40.6	77.3	60.2
EfficientQAT ^[8]	2	6.86	66.0	51.6	36.9	75.3	57.4
BitDistiller ^[82]	2	8.08	61.1	43.8	29.1	73.7	51.9
LLM-QAT ^[83]	2	9.30	61.6	48.7	33.3	70.1	53.4
ParetoQ ^[88]	2	8.00	70.0	73.8	54.5	79.2	69.4

4.7 小 结

QAT作为一种在训练或微调阶段显式引入模拟量化操作的技术范式,其目标在于使模型在训练过程中主动感知并适应量化误差,从而提升对低精度表示的鲁棒性。得益于此机制,QAT在极低比特量化场景中通常能显著优于PTQ。然而,这一性能优势往往以较高的训练开销与复杂的优化难度为代价。因此,QAT研究的核心在于如何实现训练效率、优化稳定性与最终模型性能之间的平衡。

当前QAT的相关研究在训练效率方面通过引入参数分解的分阶段训练策略^{[8][79]},将复杂的联合量化优化问题拆解为可控的子任务,从而有效降低单次迭代的计算与内存成本。在训练稳定性方面,通过改进梯度估计器^[81],或引入正则化与约束项^[80]以缓解训练振荡。此外,知识蒸馏已成为提升QAT模型性能的通用且高效的辅助途径^{[82][83]}。通过引导低比特学生模型学习高精度教师模型的输出行为,知识蒸馏能够在一定程度上弥补因参数离散化导致的表达能力损失。

在极低比特量化的前沿探索中,QAT几乎是唯一可行的优化路径。相关研究的创新主要集中于设计更具表达力的数值表示形式^[85]、利用数学变换以重塑激活分布^[87]以及开发分阶段训练策略^[86],以缓解量化引入的优化不稳定性与性能退化问题。这些方法共同推动了极限压缩场景下模型性能的突破。

总体而言,QAT是推动LLM量化性能逼近理论极限的关键技术路径。未来的研究将继续围绕降低QAT应用门槛与提升通用性展开,具体方向包括探索更轻量化的QAT算法框架、建立可靠的量

化性能缩放定律以指导训练配置以及推动QAT与模型架构的协同设计,从源头上构建更易于量化、训练稳定且推理高效的模型。

5 开发框架与硬件支持

LLM低比特量化技术的发展不仅依赖于量化算法本身的进步,也得益于深度学习开发框架、硬件平台及推理优化工具链的持续演进与协同发展。本节将总结现有的相关开发框架与硬件支持现状。

5.1 深度学习框架

当前主流的深度学习框架中,PyTorch^[94]为低比特量化提供了系统而灵活的支持。其量化模块提供了多种工作流程,包括动态与静态量化以及基于FX图形表示的自动化QAT与PTQ。随着PyTorch 2.0的发布,其Export Quantization功能进一步增强了模型在跨平台部署过程中的可移植性与优化能力。PyTorch支持多种数据类型,同时提供逐张量与逐通道等多粒度量化方案,为LLM的灵活压缩与优化提供了坚实基础。

5.2 专用优化库

在PyTorch等通用训练框架基础之上,针对LLM的扩展优化也不断推进。DeepSpeed^[95]是微软开源的分布式训练与推理优化库,其核心模块虽不直接提供通用量化API,但在内存管理与分布式能力方面对量化模型部署至关重要。其ZeRO优化器通过跨GPU分片显著减少内存消耗,支持进一步将模型状态卸载至CPU或NVMe,极大地降低大规模模型在量化后部署时的资源瓶颈。同时,DeepSpeed Inference针对Transformer架构的推理过程进行并行化优化,并提供KV缓存融合机制。此外,DeepSpeed Compression库中的ZeroQuant支持对权重与激活进行细粒度的INT8/INT4量化,具有良好的硬件兼容性与精度保持能力。

为简化量化实践与部署流程,一些专用的低比特量化库应运而生。其中,bitsandbytes是广泛用于PyTorch的低比特矩阵运算库,提供高效的8位优化器与8/4位线性层支持,结合LLM.int8()^[54]方法实现对推理过程中的动态量化与离群值处理。QLoRA^[16]技术也在该生态中发挥关键作用,通过将权重量化至4比特并结合低秩适配器进行参数高效微调,大幅减少计算与显存需求。在此基础上,社区进一步发展出多种以速度和精度为导向的PTQ方

案。AutoGPTQ库^①实现了经典的GPTQ算法^[7],通过逐层量化并进行校准,能够在几乎不影响模型精度的情况下快速完成4比特权重量化。AutoAWQ库^②则通过识别并保护对模型性能至关重要的显著权重,在低比特下实现了更高的精度保持。针对CPU端推理场景,GGML^③应运而生。它不仅是一种文件格式,更是一套为CPU优化的张量库,支持从2比特到8比特的多种量化策略,通过内存映射等技术实现模型的快速加载与高效推理。为了整合多样化的压缩技术,llm-compressor^④等综合性工具包也开始出现,它将剪枝、知识蒸馏与包括GPTQ、AWQ在内的多种量化算法集于一身,为研究者和开发者提供了一站式的模型压缩解决方案。这些工具通常与HuggingFace Transformers框架深度整合,使用户能够在加载预训练模型时直接启用低比特量化,并结合Optimum工具链部署至多种硬件后端。

总体而言,专用库在量化生态中扮演了关键的桥梁角色。DeepSpeed侧重于从系统层面解决大规模模型训练与推理的资源瓶颈,而bitsandbytes、AutoGPTQ等专用工具则提供了更高层次的API抽象,通过与HuggingFace等模型社区的深度集成,显著简化了在预训练模型上应用低比特量化与参数高效微调的流程。

5.3 硬件平台

硬件平台对低精度计算的支持是低比特量化能否在实际应用中落地的关键因素。主流AI芯片厂商通过专用硬件设计和优化的软件栈,为低精度计算提供了强有力的支持。

在云端场景中,GPU、TPU及AI加速器仍是推动低精度计算的主要硬件基础。NVIDIA GPU自Volta架构引入Tensor Core以来,不断扩展对低精度数据类型(如INT8、FP8)的支持,表12总结了不同型号NVIDIA GPU对低精度的支持。最新的Hopper和Blackwell架构进一步提升了Tensor Core的运算吞吐能力,结合Transformer Engine可在保持精度的同时实现动态量化与推理加速。AMD Instinct系列GPU基于CDNA架构,采用内置Matrix Core技术,支持INT8与FP8等低精度数据类型的高效矩阵计算,并通过ROCm生态提供对主流深度学习框架的优化支持。Google的TPU则采用脉动阵列架构优化矩阵乘法的数据流,在INT8与BF16计算中兼顾高吞吐与能效比。这些高性能平台在大模型训练与云推理中验证了低精度计算的有效性,但其高功耗和高成本限制了在资源受限场景中的应用。

表12 不同NVIDIA GPU对低精度的支持

GPU型号		P4	V100	A100	H100	B100
架构		Pascal	Volta	Ampere	Hopper	Blackwell
CUDA Core	FP32	✓	✓	✓	✓	✓
	FP16	×	✓	✓	✓	✓
	BF16	×	×	✓	✓	✓
	FP8	×	×	×	×	×
	FP6/FP4	×	×	×	×	×
Tensor Core	FP32	×	×	×	×	×
	FP16	×	✓	✓	✓	✓
	BF16	×	×	✓	✓	✓
	FP8	×	×	×	✓	✓

在嵌入式与端侧设备中,计算与能耗受限更加突出,低比特量化在此类平台上尤为关键。近年,多家厂商面向边缘AI和终端设备推出了对INT8甚至INT4的原生支持硬件,显著提升了模型推理效率。NVIDIA Jetson系列在嵌入式GPU中集成了Tensor Core,可在TensorRT框架下实现INT8量化推理。结合深度学习加速器,在边缘端可实现高能效比的视觉与语音模型部署。Google Edge TPU采用专为INT8推理优化的矩阵乘法单元,实现了每瓦性能远高于通用CPU/GPU。其与TensorFlow Lite深度集成,适用于IoT、移动终端等轻量化AI场景。ARM Ethos-N系列NPU通过对INT8运算的支持实现每秒数万亿次的低比特矩阵运算,广泛应用于移动端SoC。Apple Neural Engine(ANE)同样针对低精度推理优化,在iPhone与Mac中实现端侧模型的高效执行。此外,寒武纪、地平线、华为昇腾CANN等国产AI加速器也逐步形成了完善的低精度支持体系。这些端侧平台普遍采用定点化计算单元、低带宽访存架构和片上缓存优化等技术,使低比特量化模型能够在数瓦甚至毫瓦级功耗下实现实时推理。相比依赖高算力GPU的云端推理方案,端侧硬件的量化支持有效降低了能耗与成本,使AI模型真正具备落地能力。

5.4 部署工具链

在部署工具链方面,ONNX Runtime(ORT)^⑤是一款跨平台、高兼容性的机器学习推理与训练加速器,具有良好的ONNX标准支持,能够兼容多种主流深度学习框架。ORT的通用性使其成为如LLaMA^[3]、Mistral^[96]、Phi^[97]等实现生成式AI应用广

① <https://github.com/AutoGPTQ/AutoGPTQ>
② <https://github.com/casper-hansen/AutoAWQ>
③ <https://github.com/ggml-org/ggml>
④ <https://github.com/vllm-project/llm-compressor>
⑤ <https://github.com/microsoft/onnxruntime>

泛部署的理想选择。ORT 不仅支持静态和动态量化流程,还提供了从 FP32 模型转换为 INT8 量化模型的 Python API,使得量化过程更加灵活高效。在 Transformer 模型优化方面,ORT 通过符号形状推断技术优化图结构,为高效量化推理奠定基础。此外,ORT 还与 Apache TVM 结合使用,支持 LLM 在 CPU 和 GPU 设备上的动态量化部署,展示了其在工业应用中的可扩展性与互操作性。

NVIDIA 的 TensorRT-LLM (TRT-LLM)^①则针对 NVIDIA GPU 架构进行深度优化,专注于提高 LLM 推理的性能。它在低精度部署方面领先,特别是在支持 INT8 和 FP8 精度的推理上展现出卓越的性能。TRT-LLM 深度集成了 SmoothQuant 技术,通过权重预处理将模型权重和激活同时量化为 INT8,从而保持精度的同时提高推理速度。此外,TRT-LLM 还支持权重只量化和多种训练后量化方法如 GPTQ^[7]和 AWQ^[27]。在系统优化方面,TRT-LLM 通过内核融合技术显著提升了推理效率。这些优化措施有效应对了 LLM 推理中的内存带宽和容量瓶颈,实现了量化模型在专有硬件上的性能加速。

Intel 的 OpenVINO Toolkit^②则专注于在 Intel 平台上进行 LLM 部署,提供了丰富的图优化和压缩能力,尤其在 INT4 精度的权重压缩方面具有优势。通过将权重量化到 4 位整数,OpenVINO 显著减少了 LLM 的内存占用,提高了推理速度。为了进一步优化 LLM 在工业场景中的应用,OpenVINO 通过支持状态化模型的实现,使得模型的内部状态可以在多次推理迭代中共享,从而减少内存占用并提升性能。

此外,Apache TVM^③通过引入统一的中间表示和自动调优机制,推动了深度学习模型在各种异构硬件上的最优量化部署。MLC LLM 项目则专门为 LLM 设计,利用 TVM 的内核融合与量化感知编译技术,实现了在广泛的异构边缘平台上的高效部署。通过 TensorIR 进行自动张量化程序优化,TVM 能够支持 QAT 和多种 PTQ 方法,致力于在跨设备部署中实现最优的量化效果。

总体而言,部署工具链构成了将量化模型转化为高效可执行代码的关键步骤,不同工具各有侧重。ONNX Runtime 强调跨平台可移植性,TensorRT 专注于 NVIDIA 平台上的极致性能优化,OpenVINO 面向 Intel 硬件生态,而 TVM 则提供了一种通用的、面向异构硬件的自动编译优化框架。这些工具通过图优化、算子融合及硬件感知的

代码生成,将量化带来的理论优势转化为实际的推理性能提升。

5.5 小结

当前主流的软件框架、加速库、硬件平台与部署工具链已初步构建起完整的 LLM 低比特量化支持生态,相关内容总结于表 13。从模型训练与微调,到部署与推理,各环节的协同优化显著提升了 LLM 在实际场景中的可用性与性价比,为低成本、高效能的 LLM 推理部署提供了切实可行的工程化解决方案。

6 研究方向展望

LLM 的低比特量化技术在近年取得了显著进展,从早期的高比特训练后权重量化,发展至当前在极低比特(如 2 比特甚至 1 比特)下实现权重与激活的联合量化,并成功应用于模型的各个关键组件。这些技术极大地降低了 LLM 的部署门槛,推动了其在更广泛场景中的应用。然而,实现高效、无损且具备普适性的量化方案仍面临诸多挑战。本节将首先从宏观层面分析当前量化研究所面临的核心挑战与相应机遇,然后对未来值得深入探索的研究方向进行展望。

6.1 挑战与机遇

LLM 的量化研究正处在机遇与挑战并存的发展时期。一方面,对模型性能、效率和通用性的极致追求不断催生新的技术难题;另一方面,这些难题也为研究者们开辟了新的研究空间。表 14 总结了其中的主要挑战与机遇。

6.2 未来研究方向展望

基于上述分析,未来 LLM 的低比特量化研究可在以下几个方向上进一步发展:

(1) 极低比特下的新型量化方法

随着主流模型已普遍实现高质量的 4 比特和 3 比特量化,未来的研究前沿将集中在 2 比特及以下的极低比特场景。首先,可以探索新型的数值表示与预处理技术。通过精巧的数学变换设计,探索计算开销更低且能有效均衡数据分布的变换方法;其次,继续探索向量量化技术的应用。未来的挑战在于设计存储和计算开销更低的码本及相应的高效硬

① <https://github.com/NVIDIA/TensorRT-LLM>

② <https://github.com/openvinotoolkit/openvino>

③ <https://github.com/apache/tvm>

表 13 主流开发框架与硬件平台对低比特量化的支持

类别	工具/平台	主要功能与支持特性	适用阶段	支持精度
深度学习框架	PyTorch	Eager Mode、FX Graph、QAT、PTQ、Export Quantization；支持逐张量/通道量化，qint8，float16等	训练/推理	INT8 / FP16 / BF16
训练优化框架	DeepSpeed	ZeRO 优化器、Offload、Inference、Compression；KV Cache 融合，张量/流水线并行	训练/推理	INT8 / INT4 / FP16
量化工具库	bitsandbytes	高效 8/4 比特矩阵运算；LLM.int8()；与 Hugging Face 集成，QLoRA 微调	推理/微调	INT8 / INT4
推理优化工具	TensorRT / TensorRT-LLM	图优化、核调度、KV Cache 支持；原生支持 LLM 推理；与 Triton 集成	推理	INT8 / FP8 / FP16
	ONNX Runtime	支持量化模型的跨平台高效推理；支持动态/静态/PTQ；兼容多硬件后端	推理	INT8 / FP16
	OpenVINO	INT4 支持、KV Cache 优化、图优化与内存压缩，适合 Intel 平台部署	推理	INT8 / INT4 / BF16
	Apache TVM	支持多平台量化部署，自动调优，统一中间表示；适用于嵌入式/定制硬件	推理/部署	INT8 / FP16 / BF16
硬件平台	NVIDIA GPU	Tensor Cores，Transformer Engine，原生支持 FP8 / INT8；配套 TensorRT 系统支持 LLM 加速	全流程	FP8 / INT8 / FP16
	AMD Instinct	支持 FP8、INT8；GEMM 优化与 KV Cache 融合	训练/推理	FP8 / INT8 / BF16
	Google TPU v4/v5e	脉动阵列结构，XLA 编译器优化低精度大矩阵操作；适合大规模 LLM 部署	训练/推理	INT8 / BF16 / FP32
	Intel Gaudi	支持量化、模型压缩、KV Cache；适用于本地化与边缘部署	推理	INT8 / BF16 / INT4

表 14 大语言模型低比特量化面临的挑战与机遇	
挑战	机遇
性能保持与模型压缩的矛盾	探索极限压缩的新型量化方法
离群点的普遍性存在	构建量化友好的模型训练方法
量化算法与硬件实现的脱节	推动算法与硬件的协同设计
现有量化策略的通用性不足	发展自适应的量化优化框架
评估体系与理论指导的缺失	建立多维度的系统评估体系与理论框架

件解码方案；最后，需要发展 1 比特与 1.58 比特模型。后续研究应着力于实现权重和激活值的联合极低比特量化，并探索比直通估计器更稳定、精确的梯度优化方法。

(2)量化友好的模型架构

当前量化研究的核心是应对离群点带来的量化误差，未来的研究可以从被动处理转向主动预防。一方面，需要深入探究离群点成因与模型架构的关联。如现有研究^{[25][26]}揭示的，离群点与注意力机制紧密相关。未来的工作应更系统地分析不同模型组件对数值稳定性的影响；另一方面，应进一步探索量化感知的训练与微调方法。通过在预训练或微调阶段引入对数值动态范围的正则化约束，可以引导模型收敛至一个参数分布更平滑的空间，使其具备内生的量化友好性。

(3)算法与硬件的协同设计

量化算法的最终价值体现在其在真实硬件上的加速效果。未来的研究需要更加紧密地结合硬

件特性。一方面，需要面向现有硬件进行算法优化。例如，研究如何将混合精度计算中的稀疏高精度部分高效地映射到硬件，或通过算法层面的设计来规避硬件不友好的计算模式；另一方面，需要为先进的量化算法设计新型硬件。针对向量量化、非均匀量化等硬件不友好的算法，研究并提出专用的硬件加速架构，将其理论优势转化为实际性能提升。

(4)自适应的量化优化框架

当前主流的量化流程多基于 GPTQ^[7]所奠定的逐层贪心量化范式，这在一定程度上限制了性能的上限。未来的趋势是采用更加全局化和智能化的优化策略。首先，可以发展全局优化算法，尝试将整个网络的量化建模为一个大规模组合优化问题，并利用先进的优化技术进行求解；其次，应探索输入自适应的动态量化，设计更通用的动态量化机制，例如引入一个轻量级的策略网络，根据输入特征实时决定最优的量化参数。

(5)量化理论与系统性评估体系

量化研究的发展需要更坚实的理论基础和更全面的评估标准作为指引。首先，需要完善并扩展量化缩放定律，将更多变量（如不同量化算法、训练数据规模、硬件特性等）纳入模型，并建立量化误差与具体下游任务性能损失之间更直接的理论联系；其次，需要建立多目标的评估基准，建立标准化的、涵盖精度、延迟、能耗和内存等多维度的评估基准。该

基准应覆盖多样的模型、任务和硬件平台,从而能够更公平、全面地衡量不同量化方法的综合表现,引导研究向更具实用价值的方向发展。

7 总 结

随着大语言模型在自然语言处理领域的广泛应用,其庞大的参数规模和高昂的计算资源消耗逐渐成为部署与应用的瓶颈。低比特量化作为一种有效的模型压缩技术,能够显著降低模型的存储与计算开销,同时尽可能保持其性能表现,已成为当前研究的热点之一。本文围绕大语言模型的低比特量化方法展开综述,系统介绍了量化的基本概念,包括数值的低精度表示方式、量化粒度的选择,以及量化对模型性能的潜在影响。在此基础上,重点回顾了训练后量化和量化感知训练两大主流技术路径。在训练后量化方面,本文分别从多比特和单比特权重量化方法进行梳理,并扩展至权重与激活联合量化以及针对推理效率的键值缓存量化等策略。在量化感知训练方面,讨论了在训练过程中引入量化模拟机制以提升模型精度的相关工作。此外,本文还简要介绍了当前主流的量化开发框架与硬件支持,为模型部署提供了实践指导。最后,本文展望了大语言模型的低比特量化方法的未来研究方向。

参 考 文 献

- [1] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners//Advances in Neural Information Processing Systems. Online. 2020: 1877-1901
- [2] ACHIAM, J., ADLER, S., AGARWAL, S., et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023
- [3] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023
- [4] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023
- [5] YANG An, LI An-Feng, YANG Bao-Song, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025
- [6] GUO Da-Ya, YANG De-Jian, ZHANG Hao-Wei, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025
- [7] FRANTAR E, ASHKBOOS S, HOEFLER T, et al. GPTQ: accurate post-training quantization for generative pre-trained transformers//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda. 2023: 1-16
- [8] CHEN Meng-Zhao, SHAO Wen-Qi, XU Peng, et al. EfficientQAT: efficient quantization-aware training for large language models. arXiv preprint arXiv:2407.11062, 2024
- [9] ZHAO Xin, ZHOU Kun, LI Jun-Yi, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023
- [10] ZHU Xu-Yu, LI Jian, LIU Yong, et al. A survey on model compression for large language models. Transactions of the Association for Computational Linguistics, 2024, 12: 1556-1577
- [11] GONG Rui-Hao, DING Yi-Fu, WANG Zi-Ning, et al. A survey of low-bit large language models: Basics, systems, and algorithms. arXiv preprint arXiv:2409.16694, 2024
- [12] JIN Ren-Ren, DU Jiang-Cun, HUANG Wu-Wei, et al. A comprehensive evaluation of quantization strategies for large language models//Findings of the Association for Computational Linguistics. Bangkok, Thailand. 2024: 12186-12215
- [13] DENG Yong-Heng, ZHANG Jin-Rui, YUE Sheng, et al. Towards ubiquitous intelligence: current status and prospects of on-device large language models. Bulletin of National Nature Science Foundation of China, 2025, 39(02): 263-273 (in Chinese)
(邓永恒, 章晋睿, 岳晟, 等. 迈向泛在智能: 端侧大语言模型现状与展望. 中国科学基金, 2025, 39(02): 263-273)
- [14] WANG Rui, ZHANG Liu-Yang, GAO Zhi-Yong, et al. Research progress on large models for edge intelligence. Journal of Computer Research and Development, 2025, 62(10): 2565-2582 (in Chinese)
(王睿, 张留洋, 高志涌, 等. 面向边缘智能的大模型研究进展. 计算机研究与发展, 2025, 62(10): 2565-2582)
- [15] KALAMKAR, D., MUDIGERE, D., MELLEMPUDI, N., et al. A study of BFLOAT16 for deep learning training. arXiv preprint arXiv:1905.12322, 2019
- [16] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. QLoRA: Efficient finetuning of quantized LLMs//Advances in Neural Information Processing Systems. New Orleans, Louisiana. 2023: 10088-10115
- [17] GHOLAMI, A., KIM, S., DONG, Z., et al. A survey of quantization methods for efficient neural network inference. arXiv preprint arXiv:2103.13630, 2021
- [18] DETTMERS T, SVIRSCHEVSKI R, EGIAZARIAN V, et al. SpQR: a sparse-quantized representation for near-lossless LLM weight compression//Proceedings of the International Conference on Learning Representations. Vienna, Austria. 2024: 1-29
- [19] ALMAZROUEI E, ALOBEIDLI H, ALSHAMSI A, et al. The falcon series of open language models. arXiv preprint arXiv: 2311.16867, 2023
- [20] KIM S, HOOPER C, GHOLAMI A, et al. SqueezeLLM: dense-and-sparse quantization//Proceedings of the International Conference on Machine Learning. Vienna, Austria. 2024: 23901-23923
- [21] LEE C, JIN J, KIM T, et al. OWQ: outlier-aware weight quantization for efficient fine-tuning and inference of large language models//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada. 2024: 13355-13364
- [22] LI Dong-Min, XIE Xiu-Rui, ZHANG Dong-Yang, et al.

- SEMQ: efficient non-uniform quantization with sensitivity-based error minimization for large language models. *Future Generation Computer Systems*, 2026: 108120
- [23] GUO Cong, TANG Jia-Ming, HU Wei-Ming, et al. OliVe: accelerating large language models via hardware-friendly outlier-victim pair quantization//*Proceedings of the 50th Annual International Symposium on Computer Architecture*. Orlando, USA. 2023: 1-15
- [24] SUNG Yi-Lin, YADAV P, LI Jia-Lu, et al. RSQ: learning from important tokens leads to better quantized LLMs. *arXiv preprint arXiv:2503.01820*, 2025
- [25] BONDARENKO Y, NAGEL M, BLANKEVOORT T. Quantizable transformers: removing outliers by helping attention heads do nothing//*Advances in Neural Information Processing Systems*. Vancouver, Canada. 2024: 75067-75096
- [26] AN Yong-Qi, ZHAO Xu, YU Tao, et al. Systematic outliers in large language models. *arXiv preprint arXiv: 2502.06415*, 2025
- [27] Lin, J., Tang, J., Tang, H., et al. Awq: Activation-aware weight quantization for on-device LLM compression and acceleration//*Proceedings of Machine Learning and Systems*. California, USA. 2024: 87-100
- [28] CHENG Wen-Hua, ZHANG Wei-Wei, SHEN Hai-Hao, et al. Optimize weight rounding via signed gradient descent for the quantization of LLMs//*Findings of the Association for Computational Linguistics: EMNLP*. Miami, USA. 2024: 11332-11350
- [29] LIN Li, HU Xin-Yu, WAN Xiao-Jun. NeUQI: near-optimal uniform quantization parameter initialization. *arXiv preprint arXiv:2505.17595*, 2025
- [30] ARAI Y, ICHIKAWA Y. Quantization error propagation: revisiting layer-wise post-training quantization. *arXiv preprint arXiv:2504.09629*, 2025
- [31] HU Yue-Zhou, HUANG Wei-Yu, LIANG Zi-Chen, et al. Identifying sensitive weights via post-quantization integral. *arXiv preprint arXiv:2503.01901*, 2025
- [32] CHEE J, CAI Yao-Hui, KULESHOV V, et al. Quip: 2-bit quantization of large language models with guarantees//*Advances in Neural Information Processing Systems*. New Orleans, USA. 2023: 4396-4429
- [33] TSENG A, CHEE J, SUN Qing-Yao, et al. QuIP#: even better LLM quantization with Hadamard incoherence and lattice codebooks//*Proceedings of the International Conference on Machine Learning*. Vienna, Austria. 2024: 48630-48656
- [34] EGIAZARIAN V, PANFEROV A, KUZNEDELEV D, et al. Extreme compression of large language models via additive quantization//*Proceedings of the International Conference on Machine Learning*. Vienna, Austria. 2024: 12284-12303
- [35] TSENG A, SUN Qing-Yao, HOU D, et al. Qtip: quantization with trellises and incoherence processing//*Advances in Neural Information Processing Systems*. Vancouver, Canada. 2024: 59597-59620
- [36] CHEN Hong, LV Cheng-Tao, DING Liang, et al. DB-LLM: accurate dual-binarization for efficient LLMs//*Findings of the Association for Computational Linguistics*. Bangkok, Thailand. 2024: 8719-8730
- [37] HU Xing, CHEN Zhi-Xuan, YANG Da-Wei, et al. MoEQuant: enhancing quantization for mixture-of-experts large language models via expert-balanced sampling and affinity guidance. *arXiv preprint arXiv:2505.03804*, 2025
- [38] WANG Wei-Lian, MAO Yu, TANG Dong-Dong, et al. When compression meets model compression: memory-efficient double compression for large language models//*Findings of the Association for Computational Linguistics: EMNLP*. Miami, USA. 2024: 16973-16983
- [39] DETTMERS T, ZETTLEMOYER L. The case for 4-bit precision: k-bit inference scaling laws//*Proceedings of the International Conference on Machine Learning*. Honolulu, USA. 2023: 7750-7774
- [40] YUAN Zhi-Hang, SHANG Yu-Zhang, DONG Zhen. PB-LLM: partially binarized large language models//*Proceedings of the Twelfth International Conference on Learning Representations*. Vienna, Austria. 2024: 1-14
- [41] HUANG Wei, LIU Yang-Dong, QIN Hao-Tong, et al. BiLLM: pushing the limit of post-training quantization for LLMs//*Proceedings of the International Conference on Machine Learning*. Vienna, Austria. 2024: 20023-20042
- [42] YAN Xiang-Long, ZHANG Tian-Ao, LI Zhi-Teng, et al. Progressive binarization with semi-structured pruning for LLMs. *arXiv preprint arXiv:2502.01705*, 2025
- [43] JO D, KIM T, KIM Y. Mixture of scales: memory-efficient token-adaptive binarization for large language models//*Advances in Neural Information Processing Systems*. Vancouver, Canada. 2024: 137474-137494
- [44] MALINOVSKII V, MAZUR D, ILIN I, et al. Pv-tuning: beyond straight-through estimation for extreme llm compression//*Advances in Neural Information Processing Systems*. Vancouver, Canada. 2024: 5074-5121
- [45] XIAO Guang-Xuan, LIN Ji, SEZNEC M, et al. Smoothquant: accurate and efficient post-training quantization for large language models//*Proceedings of the International Conference on Machine Learning*. Honolulu, USA. 2023: 38087-38099
- [46] SHAO Wen-Qi, CHEN Meng-Zhao, ZHANG Zhao-Yang, et al. OmniQuant: omnidirectionally calibrated quantization for large language models//*Proceedings of the International Conference on Learning Representations*. Vienna, Austria. 2024: 1-25
- [47] WEI Xiu-Ying, ZHANG Yun-Chen, LI Yu-Hang, et al. Outlier suppression+: accurate quantization of large language models by equivalent and effective shifting and scaling//*Findings of the Association for Computational Linguistics: EMNLP*. Singapore. 2023: 1648-1665
- [48] LIU Jing, GONG Rui-Hao, WEI Xiu-Ying, et al. QLLM: accurate and efficient low-bitwidth quantization for large language models//*Proceedings of the International Conference on Learning Representations*. Vienna, Austria. 2024: 1-23
- [49] WANG Jing-Guang, WANG Jing-Yu, SUN Hai-Feng, et al. MergeQuant: accurate 4-bit static quantization of large language models by channel-wise calibration. *arXiv preprint arXiv:*

- 2503.07654, 2025
- [50] HEO J H, KIM J, KWON B, et al. Rethinking channel dimensions to isolate outliers for low-bit weight quantization of large language models//Proceedings of the Twelfth International Conference on Learning Representations. Vienna, Austria. 2024: 1-19
- [51] LIN Hao-Kun, XU Hao-Chen, WU Yi-Chen, et al. DuQuant: distributing outliers via dual transformation makes stronger quantized LLMs//Advances in Neural Information Processing Systems. Vancouver, Canada. 2024: 87766-87800
- [52] LIU Ze-Chun, ZHAO Chang-Sheng, FEDOROV I, et al. SpinQuant: LLM quantization with learned rotations. arXiv preprint arXiv:2405.16406, 2024
- [53] HU Xing, CHENG Yuan, YANG Da-Wei, et al. OstQuant: refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. arXiv preprint arXiv:2501.13987, 2025
- [54] DETTMERS T, LEWIS M, BELKADA Y, et al. LLM.int8(): 8-bit matrix multiplication for transformers at scale//Advances in Neural Information Processing Systems. New Orleans, USA. 2022: 30318-30332
- [55] ASHKBOS S, MARKOV I, FRANTAR E, et al. QUIK: towards end-to-end 4-bit inference on generative large language models//Findings of the Association for Computational Linguistics: EMNLP. Miami, USA. 2024: 3355-3371
- [56] LEE D, CHOI S, CHANG I J. Qrazor: reliable and effortless 4-bit llm quantization by significant data razoring. arXiv preprint arXiv:2501.13331, 2025
- [57] YAO Zhe-Wei, YAZDANI Aminabadi R, ZHANG Min-Jia, et al. ZeroQuant: efficient and affordable post-training quantization for large-scale transformers//Advances in Neural Information Processing Systems. New Orleans, USA. 2022: 27168-27183
- [58] KIM J, HALABI M E, PARK W, et al. GuidedQuant: Large Language Model Quantization via Exploiting End Loss Guidance. arXiv preprint arXiv:2505.07004, 2025
- [59] WANG Chang-Yuan, WANG Zi-Wei, XU Xiu-Wei, et al. Q-VLM: post-training quantization for large vision-language models//Advances in Neural Information Processing Systems. Vancouver, Canada. 2024: 114553-114573
- [60] LI Liang, LI Qing-Yuan, ZHANG Bo, et al. Norm tweaking: high-performance low-bit quantization of large language models//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada. 2024: 18536-18544
- [61] LIU Shih-Yang, LIU Ze-Chun, HUANG Xi-Jie, et al. LLM-FP4: 4-bit floating-point quantized transformers//Findings of the Association for Computational Linguistics: EMNLP. Singapore. 2023: 592-605
- [62] ZHANG Cheng, CHENG Jian-Yi, CONSTANTINIDES G A, et al. LQER: low-rank quantization error reconstruction for LLMs//Proceedings of the International Conference on Machine Learning. Vienna, Austria. 2024: 58763-58779
- [63] SON S, PARK W, HAN W, et al. Prefixing attention sinks can mitigate activation outliers for large language model quantization//Findings of the Association for Computational Linguistics: EMNLP. Miami, USA. 2024: 2242-2252
- [64] LIU Zi-Rui, YUAN Jia-Yi, JIN Hong-Ye, et al. KIVI: a tuning-free asymmetric 2bit quantization for kv cache//Proceedings of the International Conference on Machine Learning. Vienna, Austria. 2024: 32332-32344
- [65] SU Yi, ZHOU Yue-Chi, QIU Quan-Tong, et al. Accurate KV cache quantization with outlier tokens tracing. arXiv preprint arXiv:2505.10938, 2025
- [66] HOOPER C, KIM S, MOHAMMADZADEH H, et al. KVQuant: towards 10 million context length LLM inference with KV cache quantization//Advances in Neural Information Processing Systems. Vancouver, Canada. 2024: 1270-1303
- [67] SU Zun-Hai, CHEN Zhe, SHEN Wang, et al. RotateKV: accurate and robust 2-bit KV cache quantization for LLMs via outlier-aware adaptive rotations. arXiv preprint arXiv:2501.16383, 2025
- [68] LIU Rui-Kang, BAI Hao-Li, LIN Hao-Kun, et al. IntactKV: improving large language model quantization by keeping pivot tokens intact//Findings of the Association for Computational Linguistics. Bangkok, Thailand. 2024: 7716-7741
- [69] HONG Yan-Kun, LI Xing, ZHEN Hui-Ling, et al. SVDq: 1.25-bit and 410x key cache compression for LLM attention. arXiv preprint arXiv:2502.15304, 2025
- [70] HAN I, KACHAM P, KARBASI A, et al. PolarQuant: quantizing KV caches with polar transformation. arXiv preprint arXiv:2502.02617, 2025
- [71] SON D, CHOI E, YOO S. NSNQuant: a double normalization approach for calibration-free low-bit vector quantization of KV cache. arXiv preprint arXiv:2505.18231, 2025
- [72] SHUTOVA A, MALINOVSKI V, EGIJAZARIAN V, et al. Cache me if you must: adaptive key-value quantization for large language models. 2025: 2501.19392, 2025
- [73] LIU Teng-Xuan, LI Shi-Yao, YANG Jia-Yi, et al. PM-KVQ: progressive mixed-precision KV cache quantization for long-cot LLMs. arXiv preprint arXiv:2505.18610, 2025
- [74] TAO Ke-Da, YOU Hao-Xuan, SUI Yang, et al. Plug-and-Play 1. x-Bit KV cache quantization for video large language models. arXiv preprint arXiv:2503.16257, 2025
- [75] ZHANG Su-San, ROLLER S, GOYAL N, et al. OPT: open pre-trained transformer language models. arXiv preprint arXiv:2205.01068, 2022
- [76] WORKSHOP B S, SCAO T L, FAN A, et al. Bloom: a 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100, 2022
- [77] Merity, S, Xiong, C, Bradbury, J, et al. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016
- [78] Cobbe, K, Kosaraju, V, Bavarian, M, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021
- [79] KE Wen-Jin, LI Zhe, LI Dong, et al. DL-QAT: weight-decomposed low-rank quantization-aware training for large language models. arXiv preprint arXiv:2504.09223, 2025
- [80] PANG Jun-Biao, CAI Tian-Yang. Stabilizing quantization-

- aware training by implicit-regularization on hessian matrix. arXiv preprint arXiv:2503.11159, 2025
- [81] PANFEROV A, CHEN Jia-Le, TABESH S, et al. QuEST: stable training of LLMs with 1-bit weights and activations. arXiv preprint arXiv:2502.05003, 2025
- [82] DU Da-You, ZHANG Yi-Jia, CAO Shi-Jie, et al. BitDistiller: unleashing the potential of sub-4-bit LLMs via self-distillation// Findings of the Association for Computational Linguistics. Bangkok, Thailand. 2024: 102-116
- [83] LIU Ze-Chun, OGUZ B, ZHAO Chang-Sheng, et al. LLM-QAT: data-free quantization aware training for large language models//Findings of the Association for Computational Linguistics. Bangkok, Thailand. 2024: 467-484
- [84] WANG Hong-Yu, MA Shu-Ming, DONG Li, et al. BitNet: scaling 1-bit transformers for large language models. arXiv preprint arXiv:2310.11453, 2023
- [85] MA Shu-Ming, WANG Hong-Yu, MA Ling-Xiao, et al. The era of 1-bit LLMs: all large language models are in 1.58 bits. arXiv preprint arXiv:2402.17764, 2024
- [86] NIELSEN J, SCHNEIDER-KAMP P, GALKE L. Continual quantization-aware pre-training: when to transition from 16-bit to 1.58-bit pre-training for BitNet language models? arXiv preprint arXiv:2502.11895, 2025
- [87] WANG Hong-Yu, MA Shu-Ming, WEI Fu-Ru. BitNet v2: native 4-bit activations with hadamard transformation for 1-bit LLMs. arXiv preprint arXiv:2504.18415, 2025
- [88] LIU Ze-Chun, ZHAO Chang-Sheng, HUANG Han-Xian, et al. ParetoQ: scaling laws in extremely low-bit LLM quantization. arXiv preprint arXiv:2502.02631, 2025
- [89] CHEN Meng-Zhao, ZHANG Chao-Yi, LIU Jing, et al. Scaling law for quantization-aware training. arXiv preprint arXiv:2505.14302, 2025
- [90] Sakaguchi, K, Bras, R, Bhagavatula, C, et al. Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM, 2021, 64(9): 99-106
- [91] Zellers, R, Holtzman, A, Bisk, Y, et al. Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830, 2019
- [92] Clark, P, Cowhey, I, Etzioni, O, et al. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018
- [93] Bisk, Y, Zellers, R, Gao, J, et al. Piqa: Reasoning about physical commonsense in natural language//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA. 2020: 7432-7439
- [94] PASZKE A. PyTorch: an imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703, 2019
- [95] RASLEY J, RAJBHANDARI S, RUWASE O, et al. DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters//ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Online. 2020: 3505-3506
- [96] Albert Q, Alexandre S, Arthur M, et al. Mistral 7B. arXiv preprint arXiv: 2310.06825, 2023
- [97] Marah A, Jyoti A, Harkirat B, et al. Phi-4 Technical Report. arXiv preprint arXiv: 2412.08905, 2024



HAO Zhi-Wei, Ph. D. His main research interests include neural network model compression.

XU Guan-Yu, Ph. D. candidate. His main research interests include model compression and edge intelligence.

SHEN Li, Ph. D., associate professor. His main research interests include model compression, edge intelligence, trustworthy AI, reinforcement learning, and collaborative

learning.

LUO Yong, Ph. D., professor. His main research interests include model compression, edge intelligence, trustworthy AI, machine learning, pattern recognition, and their applications in multimedia, medical image analysis, and understanding.

ZENG Dan, Ph. D., professor. Her main research interests include computer vision, multimedia content analysis, and artificial intelligence.

HU Han, Ph. D., professor. His main research interests include multimedia networks, space-air-ground networks, and edge intelligence.

Background

As LLMs, predominantly based on the Transformer architecture, have achieved state-of-the-art performance, their computational and memory demands have grown exponentially. This immense resource consumption has become a primary

bottleneck, severely restricting their deployment in practical, resource-constrained settings such as on-device applications and consumer-grade hardware. To overcome this barrier, model quantization has emerged as a key enabling technology.

Currently, the field of LLM quantization is highly active and rapidly evolving. Researchers have developed sophisticated techniques to reduce the precision of model weights and activations from high-precision formats like 32-bit floating-point (FP32) to low-bit integers (e. g. , INT8, INT4) or even binary representations. The current landscape is largely defined by two dominant paradigms: Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT). PTQ methods are popular for their simplicity, as they quantize a pre-trained model without requiring retraining. Advanced PTQ research has branched into specialized areas, including weight-only quantization to reduce the memory footprint, weight-and-activation quantization for computational acceleration, and KV cache quantization to optimize memory usage during long-sequence generation. QAT, while more computationally expensive, often yields superior accuracy by simulating quantization effects during the training or fine-tuning process, allowing the model to adapt to the precision loss. Despite this flurry of innovation, the field lacks a consolidated, systematic overview, making it difficult for researchers and practitioners to navigate the complex landscape of available techniques.

This paper addresses this gap by providing a comprehensive

and systematic survey of low-bit quantization methods for LLMs. Rather than introducing a novel algorithm, its contribution lies in synthesizing and structuring the existing body of knowledge. The paper systematically reviews the fundamental concepts, details the methodologies within both PTQ and QAT, and discusses supporting frameworks and hardware. By consolidating this fragmented information, the study aims to establish a clear baseline and roadmap, thereby facilitating further academic research and practical implementation in the field.

This work is supported by the Key Program of the Joint Funds of the National Natural Science Foundation of China (Grant No. U2336211), the Key Program of the Major Research Plan of the National Natural Science Foundation of China (Grant No. 92467206), the Key Program of the Joint Funds of the National Natural Science Foundation of China (Grant No. U23A20318), the General Program of the National Natural Science Foundation of China (Grant No. 62576364), the Shenzhen Basic Research Project (Natural Science Foundation) Basic Research Key Project (Grant No. JCYJ20241202124430041), and the Zhejiang “Pioneer and Leading Goose + X” R&D Program (Grant No. 2024C01024).