计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(8): 1950−1987, 2026

# 学习驱动的模型量化:方法、挑战与展望

黄馨怡[1] 屈志昊[1] 叶保留[2] 胡世红[1] 贾柠晖[1] 钱 琳[3]
[1](河海大学水利部水利大数据重点实验室 南京 211100)
[2](南京大学计算机软件新技术国家重点实验室 南京 210046)
[3](中国南方电网电算科技数字工程(广东)有限公司 广州 510000)
(241307010018@hhu.edu.cn)

# Learning-Driven Model Quantization: Methods, Challenges and Prospects

Huang Xinyi[1], Qu Zhihao[1], Ye Baoliu[2], Hu Shihong[1], Jia Ninghui[1], and Qian Lin[3]
[1](*Key Laboratory of Water Big Data Technology of Ministry of Water Resources*, *Hohai University*, *Nanjing* 211100)
[2](*State Key Laboratory for Novel Software Technology*, *Nanjing University*, *Nanjing* 210046)
[3](*China Southern Power Grid Electro-Computational Technology Digital Engineering* (*Guangdong*) *Co.*, *Ltd.*, *Guangzhou* 510000)

**Abstract** The increasing parameter scale and structural complexity of deep neural networks (DNNs) pose significant challenges for their efficient deployment in cloud-edge-end collaborative computing architectures, where edge and terminal nodes must support real-time inference under stringent constraints on computation, storage, inference latency, and communication overhead. Although model quantization effectively reduces resource consumption, traditional approaches based on fixed rules or empirical heuristics exhibit notable limitations in accuracy preservation, training stability, and adaptability across diverse architectures and application scenarios, particularly under very low bit-width conditions (≤4 bit). To mitigate these issues, learning-based quantization incorporates learnable components and diverse supervisory signals, substantially improving accuracy preservation, optimization robustness, and hardware adaptability under low-bit conditions. In this paper, we systematically review representative studies on learning-based quantization, centered on two primary paradigms: post-training quantization (PTQ) and quantization-aware training (QAT). We summarize performance enhancement strategies including learnable parameter modeling, reconstruction and approximation optimization, auxiliary supervision, and learning-driven methods under data and hardware constraints. We further organize existing technical approaches and their interrelationships from the perspectives of quantization process stages and learning signal mechanisms, and analyze the applicability of different strategies under varying data conditions, model scales, and application constraints. Finally, we discuss future research directions in learning-based quantization from three aspects: error-analysis-based optimization of quantization parameters and learning signals, the trade-off between accuracy and resources in learning-based quantization, and a unified quantization framework across model architectures.

**Key words** model quantization; post-training quantization (PTQ); quantization-aware training (QAT); learning-driven; large language models (LLMs)

**收稿日期**: 2026-03-12; **修回日期**: 2026-04-29

**基金项目**: 江苏省基础研究计划重点项目(BK20253011); 国家自然科学基金项目(62441225, 62572171); 教育部基础学科和交叉学科突破计划(JYB2025XDXM118); 高等学校学科创新引智基地资助(B26023)

This work was supported by the Key Project of Basic Research Program of Jiangsu Province (BK20253011), the National Natural Science Foundation of China (62441225, 62572171), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM118), and the "111 Center" (B26023).

**通信作者**: 屈志昊(quzhihao@hhu.edu.cn)

**摘　要**　随着深度神经网络参数规模与结构复杂度持续提升，以及云边端协同计算架构的广泛部署，模型在边端侧高效推理面临计算存储受限、推理延迟高及通信开销大等关键技术挑战。模型量化能够有效降低资源开销，但在极低比特（≤4 bit）条件下，基于固定规则或经验启发的传统量化方法在精度保持、训练稳定性以及跨架构与应用场景适配性方面仍存在明显局限。为缓解上述问题，学习驱动量化通过引入可学习机制与多样化监督信号，显著提升了极低比特条件下模型的精度保持能力、优化鲁棒性及异构硬件适配效率。围绕学习驱动量化的核心方法与关键机制，系统梳理并分析了相关代表性研究工作，以后训练量化（post-training quantization, PTQ）与量化感知训练（quantization-aware training, QAT）两类典型范式为主线，总结了可学习参数建模、重构与近似优化、辅助信号监督以及数据和硬件约束下的学习驱动量化的性能增强方法。此外，从量化流程阶段与学习信号作用机制2个角度，对现有技术路径及其内在联系进行了系统整理，分析不同策略在数据条件、模型规模与应用约束下的适用特征。在此基础上，从基于误差解析的量化参数与学习信号优化、学习驱动量化中精度与资源的权衡，以及跨模型架构的学习驱动量化统一框架等方面，对学习驱动量化的未来发展趋势进行了展望。

**关键词**　模型量化；后训练量化；量化感知训练；学习驱动；大语言模型

**中图法分类号**　TP183

**DOI:** 10.7544/issn1000-1239.202660128　　**CSTR:** 32373.14.issn1000-1239.202660128

深度学习已在计算机视觉、自然语言处理、语音识别等领域取得显著突破，推动了人工智能应用的快速发展。然而，随着深度神经网络参数规模持续增长，模型在存储占用、计算开销与能耗需求等方面的资源消耗急剧上升，使其在边缘设备、移动终端乃至数据中心的高效部署面临严峻挑战。以LLaMA-3-70B为例，该模型参数规模约为700亿，在16位浮点数（FP16）精度下的显存占用高达140 GB，边缘硬件难以支撑模型的存储开销与计算负载，难以满足实时推理的时延要求。在云边端协同计算场景中，边端节点须在有限算力与内存约束下承担实时推理任务，同时与云端协同完成模型更新与知识同步，对模型轻量化提出了更为严格的要求。

模型压缩能够有效降低深度神经网络的存储空间、计算开销及传输带宽，是突破算力网边端侧节点资源瓶颈的关键使能技术。主流的方法包括网络剪枝（network pruning）[1]、知识蒸馏（knowledge distillation）[2]、低秩分解（low-rank decomposition）[3]以及模型量化（model quantization）[4-5]等。相较于其他方法，模型量化通过将连续的高精度浮点参数映射为离散的低位宽整数（如INT8、INT4），利用通用硬件的整数指令集大幅提升推理吞吐率[5]，显著降低分布式环境下的数据通信带宽占用，从而在压缩效率与异构算力适配方面展现出巨大优势。该机制有效突破了边缘侧的存储与传输瓶颈，支持大规模神经网络在资源受限终端上的高效部署、协同推理与跨节点协同计算，极大地拓展了智能计算在泛在场景下的应用边界。

现有量化方法可分为后训练量化（post-training quantization, PTQ）与量化感知训练（quantization-aware training, QAT）2类。PTQ在模型训练完成后对已训练好的模型进行量化，仅需少量校准数据即可完成参数映射，适用于快速部署或数据隐私受限场景；QAT则在训练过程中显式引入量化操作，通过反向传播微调参数以适应低比特表示，该方法通常能够获得更高的量化精度，但需要额外的训练开销[6]。

模型量化技术的发展历程经历了由规则驱动向学习驱动逐步演进的过程，如图1所示。早期量化方法主要聚焦于统一量化规则、整数推理实现和基础校准策略[5-6]，通常采用固定的量化映射形式与预设位宽，在INT8场景下已能取得较好的部署效果，但对更低比特条件下复杂分布偏移和跨层误差累积的建模仍较为有限。随着对精度要求的提升，研究进入基于统计特性的校准量化阶段[7-8]，通过引入少量校准数据来统计激活值的分布特征，如利用KL散度或分布假设搜索合适的截断阈值与量化范围，一定程度上缓解了量化带来的精度退化。为突破极低比特压缩中精度损失不可控的难题，演进出基于重构目标的优化量化[9-10]，该阶段侧重于建立层级特征对齐目标，利用迭代优化算法寻找最优离散化参数，以最小化量化前后特征图的重构误差，为学习驱动方法奠定了算法基础。

尽管传统模型压缩方法能够基本满足中高位宽的部署需求，但在极低比特（≤4 bit）及异构计算场景

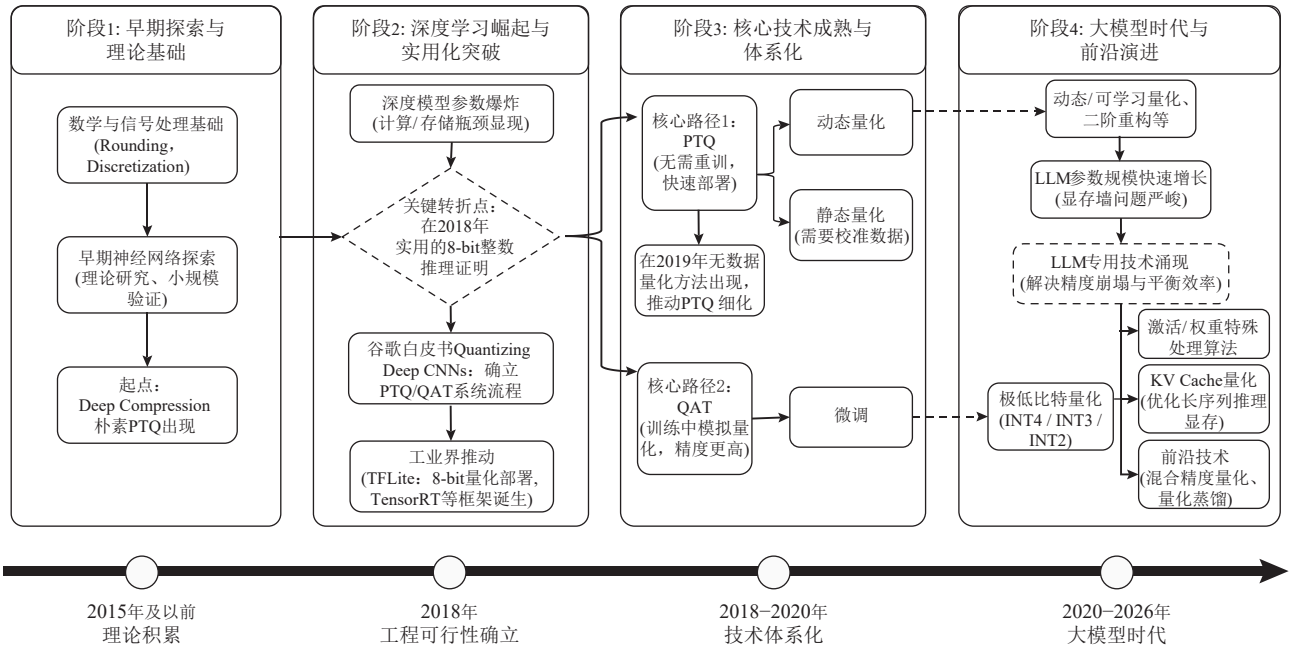



Fig. 1 Timeline of model quantization development

图 1 模型量化发展时间线

中仍面临严重的精度下降与适配性瓶颈。针对此挑战，学习驱动量化近年来得到国内外学术界和工业界的广泛关注[9,11-12]。如图 2 所示，其核心思想是将步长、偏移等关键量化参数转化为连续的可学习参数并纳入可微分计算图，结合全局任务损失进行联合训练与梯度回传。在 PTQ 与 QAT 两大框架下，该范式均能通过端到端的梯度优化或自适应校准，有效融合重构误差、蒸馏知识及硬件反馈等多元约束信息，从而显著提升 2~4 bit 条件下模型对复杂任务与异构架构的适配能力，并增强其在资源受限算力网节点上的部署效能与鲁棒性。

具体而言，相较于传统量化方法，学习驱动量化在以下 3 个方面展现出显著优势：

1）精度保持能力增强。针对固定规则难以适配低比特分布的局限，学习驱动量化将步长、偏移、量化区间或舍入决策等关键量化参数纳入可优化框架。在 PTQ 中，通过引入自适应软舍入与逐层重构等机制[9,13-14]可实现离散权重的精细校准与误差最小

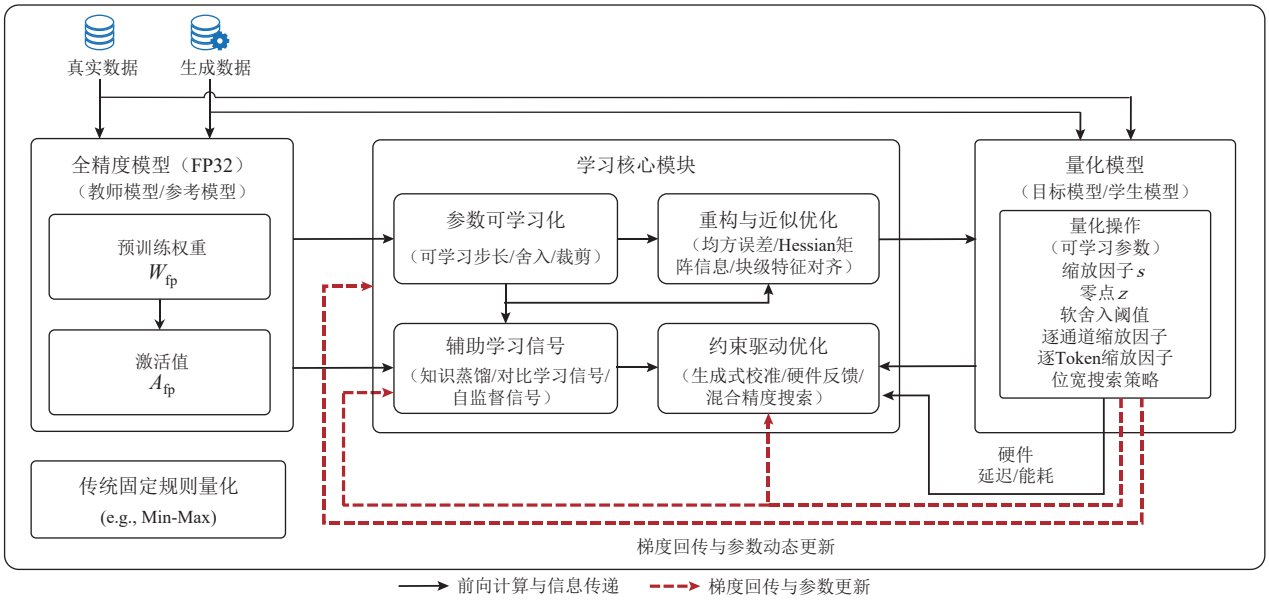



Fig. 2 General scenarios and mechanisms of learning-driven quantization

图 2 学习驱动量化通用场景与机制

化；在 QAT 中，通过可微量化算子与端到端联合训练[11-12,15-16]可实现 2~4 bit 极低位宽模型性能接近全精度性能。

2）训练与校准稳定性改善。为克服小样本校准偏差及梯度失配难题，学习驱动量化构建了鲁棒的优化路径。在 PTQ 中，基于二阶 Hessian 信息的块级误差对齐[9,17]能够有效抑制层间误差传播并降低对校准数据的依赖；在 QAT 中，通过可学习量化器、平滑可微近似及渐进式训练策略[18-19]可缓解低比特量化导致的梯度估计偏差和优化不稳定性，实现训练过程的快速收敛与稳定。

3）硬件适配性提升。面向云边端算力网异构节点的资源差异，学习驱动量化将延迟、能耗等硬件约束显式融入优化目标。通过基于敏感度分析的混合精度配置[15,20]可实现计算资源的最优分配；通过结合强化学习的自动位宽搜索[21-22]可在硬件反馈闭环中自动寻找满足特定资源约束的帕累托最优量化方案，大幅提升跨平台协同部署效率。

学习驱动量化研究方向主要包括 PTQ、QAT 及面向大模型的架构感知量化。针对 PTQ，研究重心从早期的统计校准转向基于重构目标的优化范式，利用块级特征对齐、自适应舍入及二阶 Hessian 信息引导等机制，实现了在无需全量重训练前提下的参数精准估计与误差最小化[9-10,17,23]。在 QAT 中，通过将步长、截断阈值等纳入端到端微分计算图并与网络权重协同优化，显著提升了极低比特条件下的训练稳定性与性能上限[11-12,16,18]。针对视觉 Transformer（vision Transformer, ViT）与大语言模型（large language model, LLM）的长尾激活分布与注意力机制敏感性等特异性挑战，近期研究重点转向架构感知的专用策略，通过激活平滑、权重等效变换及分组量化等技术，有效解决了大规模参数模型在算力网边缘节点的部署难题[20,24-27]。

传统模型量化领域已存在多篇综述性工作，但现有文献多侧重于通用量化流程、特定模型架构或硬件加速器的宽泛式梳理。然而，针对学习驱动量化这一前沿范式，目前尚缺乏系统性的归纳与总结。为直观呈现现有综述在研究侧重与方法覆盖方面的差异，表 1 对近年来具有代表性的模型量化综述文献进行了对比分析。

表 1 系统对比了近年来代表性综述在量化范式、模型架构、硬件适配及理论深度等维度的研究侧重。现有综述大多围绕量化技术的分类体系、特定架构的适配方法或硬件部署优化展开。在通用神经网络量化方面，文献 [28−29] 重点梳理了量化技术的分类体系，以及其在模型压缩整体框架中的定位与精度转换流程；在硬件协同优化方面，文献 [28, 30−32] 聚焦于量化模型在实际计算平台上的运行效率，分析了不同策略下的能耗与推理速度表现；在特定架构方面，面向 Transformer 与大模型，文献 [33−34] 着重讨论了针对大规模参数的 PTQ 策略及其误差控制技术。此外，文献 [32, 35−39] 进一步补充了国产硬件适配与自动化设计等视角。然而，现有综述缺乏对学习驱动量化范式的系统性梳理。具体而言，现有工作尚未深入揭示由规则向数据驱动的演进逻辑，对量化过程中学习机制的引入方式、监督信号的来源与作用以及优化目标的设计原则等核心问题缺乏系统性讨论，且缺少对 PTQ 与 QAT 共性策略的对比及跨模型架构的适用性分析。

基于此，本文围绕学习驱动的量化方法，系统梳理该领域的代表性工作与关键机制，分析监督信号、优化目标与硬件约束的协同权衡，主要挑战如下：

1）在 PTQ 范式下，传统方法高度依赖校准数据，在数据隐私受限或无标签场景下难以通过简单的统计校准维持高精度；

2）在 QAT 范式下，极低比特（≤4 bit）引入的高噪声导致梯度估计偏差严重，传统固定步长策略难以保证训练的收敛性与稳定性；

3）在大模型适配场景下，传统方法难以应对大语言模型巨大的参数规模及特殊的网络结构，导致通用量化策略直接迁移时性能严重退化。

针对上述挑战，本文总结了学习驱动量化领域的突破性进展：1）针对 PTQ 数据依赖难题，归纳了基于重构目标与生成式数据合成的学习化校准方法，实现了无数据条件下精度保持的模型压缩；2）针对 QAT 训练失稳难题，梳理了参数化梯度近似与可微位宽学习策略，通过端到端联合优化显著提升了极低比特模型的训练鲁棒性；3）针对大模型适配难题，总结了架构感知的专用量化策略，通过平滑异常激活值与分组量化等手段，实现了大模型在异构算力网节点上的高效协同部署与推理加速。

本文进一步展望了学习驱动量化的未来研究方向：1）深化可学习参数与学习信号的优化研究，刻画量化误差对模型精度影响的逐层传播机理，并在此基础上建立量化参数之间的显式耦合关系，协调多辅助信号训练中的梯度更新；2）改善学习驱动量化中的精度与资源权衡，降低量化校准对外部数据的依赖，提升资源分配效率，并增强混合精度量化在实

**Table 1 Comparison of Surveys on Existing Model Quantization**

**表 1 现有模型量化综述对比**

| 来源 | 领域 | 量化类型 | 对象覆盖 | 精度范围 | 硬件协同 | 理论分析 | 关注点 |
|---|---|---|---|---|---|---|---|
| 文献 [28] | 图像分类 CNN 量化 | ✔ | ✔ | ✔ | ✔ | ○ | ① 按网络组件、QAT/PTQ、映射方式、分布特性等维度建立系统分类体系<br>② 深入分析 STE 直通估计器及首/末层量化敏感性问题<br>③ 详细推导 MAC 优化、对数量化 Bit-shift、XNOR+bitcount 等硬件实现 |
| 文献 [29] | 通用神经网络量化 | ✔ | ✔ | ✔ | ✔ | ✔ | ① 建立统一量化公式，明确涵盖 CNN、Transformer、BERT 等 CV/NLP 模型<br>② 系统讨论零样本/数据无关量化、混合精度搜索及硬件-算法协同设计 |
| 文献 [33] | LLM 压缩与推理 | ✔ | ○ | ✔ | ✔ | ○ | ① 系统总结 GPTQ/AWQ/SmoothQuant 等 LLM 量化方法及剪枝策略<br>② 覆盖 DeepSpeed/FlexGen/PowerInfer 等加速框架实践 |
| 文献 [34] | 分布式深度学习 | ✔ | ○ | ✔ | ✔ | ✔ | ① 建立梯度量化三级体系，包括确定性+误差反馈、随机量化、矩阵分解<br>② 系统讨论梯度稀疏化（top-$k$/DGC）及联邦学习专用量化压缩<br>③ 提供 QSGD/top-$k$ 等方法的理论收敛保证 |
| 文献 [40] | 通信压缩技术 | ○ | ✗ | ○ | ✗ | ○ | ① 算法视角，量化分类仅分截断表达式和码本映射 2 类<br>② 建立稀疏化三级分类，包括基于相对值、绝对值和矩阵分解 |
| 文献 [41] | 梯度压缩量化评估 | ○ | ○ | ✔ | ✔ | ○ | ① 大规模量化评估，讨论真实系统环境下压缩方法实际表现<br>② 提出 GRACE 统一框架，强调系统级开销可能抵消通信收益<br>③ 发现误差反馈对 SignSGD 可能有害，架构依赖性显著 |
| 文献 [30] | 二值神经网络 | ✔ | ○ | ○ | ✔ | ✔ | ① 从信息论、优化理论、硬件协同三维度剖析 1bit 极端量化<br>② 分类，最小化量化误差、改进损失函数、减少梯度误差三大优化策略<br>③ 提供 FPGA/ASIC 部署实践及任务泛化性分析 |
| 文献 [31] | 低比特量化 | ✔ | ✔ | ✔ | ✔ | ○ | ① 提出八大类 24 小类分类体系，覆盖 2020—2025 年最新进展<br>② 系统化梳理离群值处理方法（SmoothQuant/QuIP#）及混合精度搜索方法<br>③ 开创性总结扩散模型量化，探索 FP8/NormalFloat 等高级格式 |
| 文献 [42] | 量化方法与理论 | ○ | ○ | ✔ | ○ | ✔ | ① 建立确定性与随机量化基础分类框架<br>② 讨论 STE 直通估计器数学原理及 BinaryConnect 二值化方法收敛性理论证明<br>③ 分析二值化向量几何性质（角度保持 37°）及码本设计权衡 |
| 文献 [35] | DNN 压缩与加速 | ✔ | ✔ | ✔ | ○ | ✔ | ① 系统性理论分析，深度剖析各压缩方法内在机理、优缺点及组合可能性<br>② 重点讨论极低精度量化，包括二值化（XNOR-Net）、三值化（TWN/TTQ）<br>③ 强调软硬件协同及二值网络收敛性理论边界 |
| 文献 [36] | 模型轻量化 | ✔ | ✔ | ✔ | ✔ | ○ | ① 应用导向综述，强调边缘设备部署的数据安全、硬件适配和自动化设计<br>② 系统梳理非均匀量化方法与无数据压缩方法 |
| 文献 [37] | 深度网络模型压缩 | ○ | ✔ | ○ | ○ | ○ | ① 联合讨论网络剪枝、知识蒸馏、网络分解<br>② 分层压缩效用评价：量化分析卷积层 vs 全连接层差异 |
| 文献 [38] | 模型压缩与加速 | ✔ | ✔ | ✔ | ✔ | ○ | ① 建立量化分类体系，涵盖二值化、混合位宽、自适应位宽等<br>② 讨论模型压缩混合策略，包括剪枝+量化、量化+蒸馏等<br>③ 分析量化比特宽度对模型性能的影响规律 |
| 文献 [32] | 边缘智能轻量 DNN | ✔ | ✔ | ✔ | ✔ | ○ | ① 详细分类校准方法，包括均匀/非均匀量化、模拟/纯整型量化<br>② 讨论混合精度量化，为不同层自适应分配位宽，建立精度-存储帕累托前沿<br>③ 实测数据详实，涵盖训练/验证时间、Top-1/5 精度、不同 CPU/GPU 配置 |
| 文献 [39] | 轻量化 CNN 设计方法 | ○ | ✔ | ○ | ✔ | ○ | ① 设计方法论视角，系统对比人工设计与 AutoML/NAS 两大范式<br>② 编译框架及硬件适配视角，包括 PyTorch/TensorFlow/FPGA/TPU<br>③ 分析权值量化算法原理，涵盖 Hashed Nets 哈希共享、K-means 聚类 |

注：✔完全讨论，○部分讨论，✗该工作中基本没有讨论。

际硬件上的加速收益；3)探索跨模型架构的学习驱动量化统一框架，挖掘不同模型间的共性规律与结构差异，为不同模型架构下的可学习量化方法设计提供参考。

本文第 1 节阐述模型量化的基础理论、主流方法及典型流程；第 2 节构建学习驱动量化的总体分类框架，从参数可学习化、重构与近似优化、辅助学习信号及约束驱动优化等维度概述量化方法的分类与总览；第 3 节与第 4 节分别总结并分析学习驱动的 PTQ 与 QAT 关键技术，重点对比各类方法在精度保持、训练稳定性及跨场景适配等方面的性能表现与适用条件；第 5 节聚焦跨架构泛化，横向对比 CNN、ViT 与 LLM 的量化敏感性特征与适配方案；第 6 节总结学习驱动量化面临的核心技术挑战，并从基于误差解析的量化参数与学习信号优化、学习驱动量化中精度与资源的权衡，以及跨模型架构的学习驱动量化统一框架 3 个方面展望未来研究方向；第 7 节总结全文贡献并探讨应用前景。

## 1 背景知识

### 1.1 深度神经网络模型量化的基本概念

深度神经网络（deep neural network, DNN）通常由若干具有可学习参数的线性变换层、非线性激活函数及归一化算子堆叠构成。设深度神经网络模型 $G$ 包含 $L$ 层，第 $l$ 层的可学习参数可以表示为权值矩

阵 $\boldsymbol{W}_l \in \mathbb{R}^{d_{l+1}\times d_l}$，输入激活为 $\boldsymbol{a}_l \in \mathbb{R}^{d_l\times 1}$，$d_l$为第$l$层激活维度，$d_{l+1}$ 为第 $l+1$ 层激活维度。全精度前向过程定义为：

$$\boldsymbol{a}_{l+1} = f(\boldsymbol{W}_l\boldsymbol{a}_l + \boldsymbol{b}_l)\,, \tag{1}$$

其中，$f(\cdot)$ 为非线性激活函数，$\boldsymbol{b}_l$ 为偏置项。在DNNs模型的训练和推理阶段，权值矩阵 $\boldsymbol{W}$ 与激活值 $\boldsymbol{a}$ 通常采用32位浮点数(FP32)表示，前向传播时会进行大规模的矩阵乘法运算，导致模型的存储与计算开销巨大。特别是在资源受限的边缘设备或嵌入式终端上，模型的高精度计算难以满足实时性与能耗要求[4-5]。

模型量化(model quantization)作为一种高效的模型压缩技术，通过降低权值 $\boldsymbol{W}$ 与激活值 $\boldsymbol{a}$ 的数值精度，将浮点数映射为低比特整数，从而减少计算资源占用并提升推理速度。为便于描述，本文以单个标量元素为例给出量化映射。设 $x\in\mathbb{R}$ 表示权重或激活张量中的一个实值元素，$\hat{x}\in\mathbb{Z}$ 表示其量化后的整数值，$x_q\in\mathbb{R}$ 表示反量化后的近似实值。设量化函数为 $q(\cdot)$，则量化与反量化(dequantization)过程可表示为：

$$\hat{x} = q(x),\quad x_q = q^{-1}(\hat{x}) \approx x\,。 \tag{2}$$

量化引入的误差 $\varepsilon = x - x_q$ 不可避免地影响模型性能。因此，模型量化的核心目标是在减少存储与计算复杂度的同时，尽可能降低量化误差[6]。

1.1.1 线性量化函数及其形式表示

在神经网络量化中，最常见的映射函数为线性量化(linear quantization)函数。如图3所示，该函数以等距划分连续实数区间，并将每个区间映射为一个固定整数。

线性量化函数定义为：

$$\hat{x} = Int\left(\frac{x}{s}\right),\quad x_q = s\hat{x}\,, \tag{3}$$

其中，$x$ 表示全精度的真实值，即原神经网络模型的权值或激活值；$s$ 为量化缩放系数，即缩放因子(Scale)，用于调整真实值的范围，控制量化步长；$Int(\cdot)$ 表示取整操作，通常采用四舍五入(round)或向下取整(Floor)策略；另外，$Int(\cdot)$ 可结合区间限制函数 $Clamp(\cdot)$ 实现对取整后的数值的截断操作，以保证取整后的数值在受限的整型数集合内。

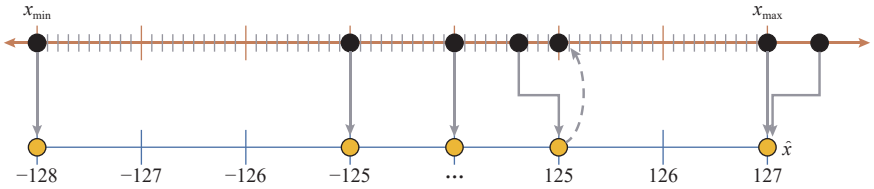



Fig. 3 Illustration of linear quantization

图3 线性量化示意图

整型数集合的范围由量化映射的目标位宽 $b$ 决定，对于 $b$ 位有符号整数，其取值范围为 $[-2^{b-1}+1, 2^{b-1}-1]$，无符号整数范围为 $[0, 2^b-1]$。

在神经网络模型中，量化后的整数值 $x$ 通常被送入硬件执行整数乘加运算。但在计算结果用于浮点操作(如非线性激活、梯度传播或误差计算)时，必须将整数结果映射回原始的数值范围，这一过程称为反量化。线性量化的反量化过程为：

$$x_q = s\times Int\left(\frac{x}{s}\right)\,。 \tag{4}$$

该过程引入量化误差：

$$\varepsilon = x - x_q\,。 \tag{5}$$

在神经网络的前向传播过程中，浅层的量化误差会在计算时不断扩大，最终导致神经网络模型任务精度的大幅下降，因此对量化缩放因子 $s$ 的决定策略是量化神经网络模型的关键。量化误差的大小取决于缩放因子 $s$ 与数据分布，为减少误差，通常通过以下方式确定：

$$s = \frac{p-n}{2^b-1}\,, \tag{6}$$

其中，$p$ 与 $n$ 为截断阈值(clipping thresholds)，表示真实值范围的上界与下界。需要注意的是，截断阈值 $p$、$n$ 并不等同于数据分布的真实最小值和最大值。由于神经网络权值与激活通常包含离群值，直接使用上下界会导致量化步长过大、信息丢失严重。实际中，截断阈值 $p$ 与 $n$ 通常根据数据统计特性或优化目标确定：

1) 统计法

依据数据分布统计特性设定截断区间，如基于分布的标准差或百分位，或通过KL散度等准则确定截断阈值[29]；

2) 最小化量化误差法

通过搜索并优化 $p$,$n$ 使得量化误差 $\varepsilon$ 最小[43]；

3）可学习阈值法

在 QAT 中，将 $p,n$ 设为可训练参数并通过反向传播联合优化[44]。

线性量化通过线性映射实现浮点到低位宽整数的转换，并保持与全精度计算形式的一致性，具有计算开销低、实现简单、硬件友好等优点。

### 1.1.2 对称与非对称线性量化

根据截断阈值的分布是否对称，线性量化可分对称量化（symmetric quantization）和非对称量化（asymmetric quantization）两类。

1）对称量化

对称量化假设数据分布以 0 为中心，满足 $p=-n$。设量化位宽为 $b$，其离散取值集合为：

$$\{-2^{b-1}+1,\cdots,-1,0,1,\cdots,2^{b-1}-1\}\times s。$$

对称量化计算处理过程简单且实现开销较低，通常要求被量化的数值分布也是对称分布的，例如服从高斯或拉普拉斯分布的权值矩阵。适用于输出存在负值的激活函数，以及权值分布接近零均值的卷积层。对称量化的映射方式如图 4（a）所示。

2）非对称量化

对于偏移分布或非零均值数据，非对称量化引入零点（zero-point）参数 $z$ 来修正映射范围。设量化位宽为 $b$，其离散取值集合可表示为：

$$\{0,s,2s,\cdots,(2^b-1)s\},$$

对应于 $b$ 位无符号整数的表示范围 $[0,2^b-1]$。

与对称量化不同，非对称量化的上下截断阈值 $p,n$ 无需满足零点对称性。当数据分布偏移较大时，直接采用固定区间会造成部分取值区间被浪费，降低数值表示效率。为此，参数 $z$ 被用于调整缩放后的整数范围，使得量化映射能更准确地覆盖输入数据分布。

非对称线性量化及其反量化过程可表示为：

$$\hat{x}=clip\left(round\left(\frac{x}{s}\right)+z,0,2^b-1\right),\quad x_q=s\times(\hat{x}-z),\tag{7}$$

其中，$s$ 为缩放因子；$z$ 为零点偏移，用于平移整数表示的中心。

非对称量化在处理激活值为非对称正分布（如 ReLU 输出）或权值矩阵均值偏移的层时量化误差更小，但量化与反量化过程会引入额外的计算开销。图 4 展示了对称量化与非对称量化的区别。

### 1.1.3 均匀与非均匀量化

上述线性量化方法均属于等距离散映射形式，即各量化级之间间隔相等，同一量化步长 $s$ 区间内的数值被映射为相同的离散值。这种形式称为均匀量化（uniform quantization），因结构规则、实现简单而被广泛应用于深度学习推理框架。

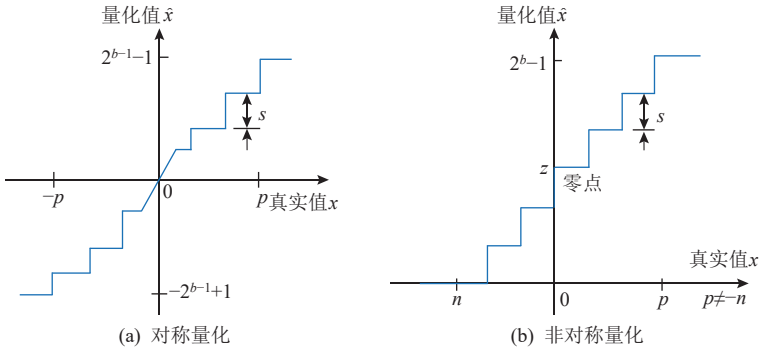
然而，在实际应用中，权重与激活值通常呈现非对称或长尾分布，等距划分难以充分匹配数据特性。为此，可采用非均匀量化（non-uniform quantization），根据数据分布自适应调整区间宽度，从而在实际场景下有效降低量化误差。

设真实值 $x$ 的量化位宽为 $b$，非均匀量化通过若干不等间隔的划分点集合 $\{t_1,t_2,\cdots,t_{2^b+1}\}$ 将实数域划分为多个子区间，并为每个区间指定一个对应的离散取值集合 $\{d_1,d_2,\cdots,d_{2^b}\}$。其量化函数可表示为：

$$q_{\text{non-uniform}}(x)=d_k,\quad \text{if } t_k\leqslant x<t_{k+1},\quad k=1,\cdots,2^b,\tag{8}$$

其中，$d_k$ 为可学习离散取值[45-46]或固定为聚类中心[47]。

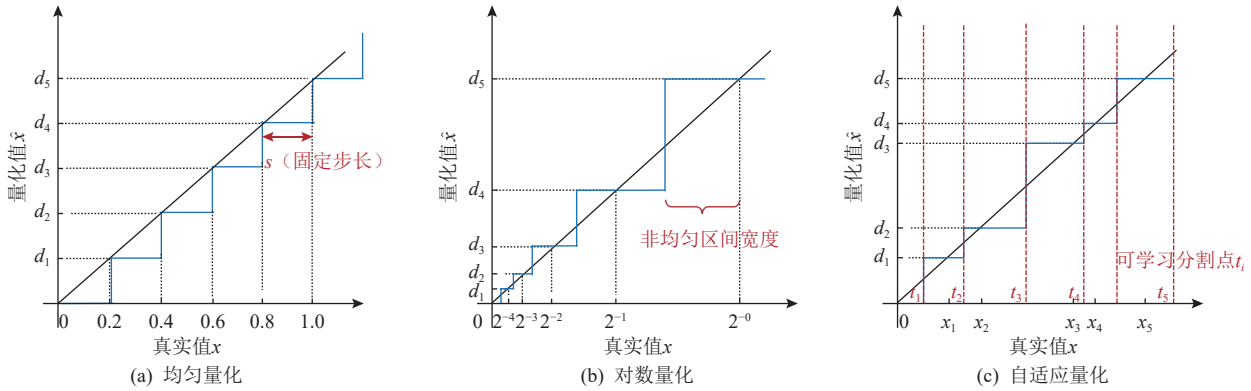
图 5 展示了均匀量化与非均匀量化的阶梯映射特征，其中均匀量化采用固定步长进行等距划分，而非均匀量化允许区间宽度随数值范围变化，通过非均匀区间划分增强了对小幅值区域的表示能力，如图 5（b）所示；自适应量化可通过可学习划分点进一步调整量化区间，以提高不同分布条件下的量化适配性，如图 5（c）所示。非均匀量化通过自适应调整区间宽度，能够更好地适应实际数据分布（如注意力




(a) 对称量化　　(b) 非对称量化

Fig. 4 Illustration of symmetric and asymmetric quantization mappings

图 4 对称与非对称量化映射示意图




(a) 均匀量化 (b) 对数量化 (c) 自适应量化

Fig. 5 Function images of uniform and non-uniform quantization

图 5 均匀与非均匀量化的函数图像

层的长尾分布或 LayerNorm 输出），从而在低比特（2~4 bit）条件下有效降低量化误差并保持较高的模型精度。

非均匀量化的核心问题是如何确定分割点与离散取值。现有工作主要从 3 个角度考虑：1）基于聚类与权值共享机制，通过聚类中心确定离散取值[47]；2）假设数据近似服从指数或对数分布，并采用等比间隔构造非均匀量化级，以更好地逼近钟形分布特征[48]；3）将非均匀映射视为多个局部均匀区间的组合，通过多点线性组合或分段近似优化实现整体量化函数的构建[49]。近年来，学习驱动量化进一步将分割点集合与离散取值视为可学习参数或将量化映射函数本身参数化为可学习量化器[45]，并在训练过程中通过反向传播自动优化，从而在模型精度保持与硬件部署效率之间实现更优权衡。

### 1.2 2 种主流量化范式

#### 1.2.1 PTQ

PTQ 是深度模型压缩中广泛应用的技术路线之一，其核心思想是在模型训练完成后，通过有限的校准样本或统计信息确定量化参数，实现从浮点表示到低比特整数表示的映射。该过程无需端到端重新训练，可在保持较高模型性能的同时显著降低计算与存储开销，从而成为边缘部署与高效推理的重要途径[6,9,13]。

相对于需要完整训练流程的量化策略，PTQ 无需在训练阶段显式模拟量化误差，具有部署成本低、训练负担小、实现流程简洁等优势，能够快速适配不同网络结构与硬件平台。然而，其在极低比特（≤4 bit）场景下易出现显著精度退化，主要原因在于离散量化误差在网络中的逐层传播与累积，以及参数空间近似与模型空间目标之间的不一致；进一步地，激活量化带来的分布偏移、跨层依赖和低比特模型平坦性不足也会放大这一问题。

PTQ 的优化目标可形式化为：在冻结预训练权重集合 $W^*$ 的条件下，寻找最优量化参数集合 $\mathcal{F}$，使量化模型输出尽可能接近全精度模型：

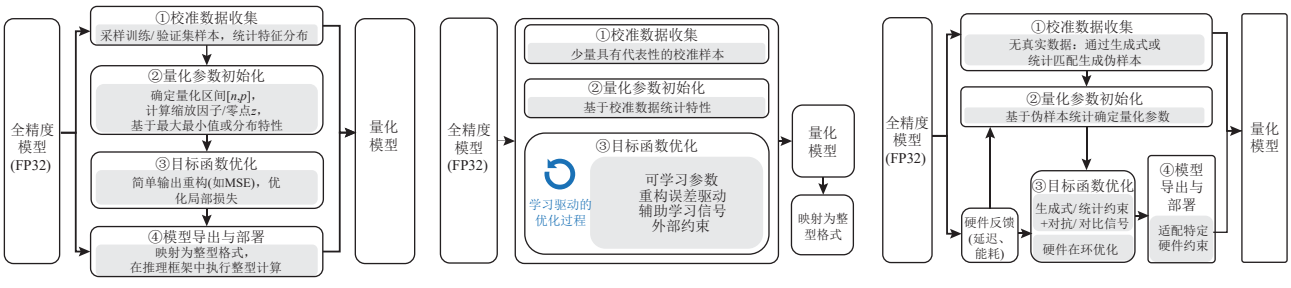
$$\min_{\mathcal{F}} \mathcal{L}_{\text{task}}\left(Q(G;\mathcal{F},W^*),G(W^*)\right), \tag{9}$$

其中，$\mathcal{L}_{\text{task}}$ 常采用逐层或块级重构损失 $\mathcal{L}_{\text{recon}}$[9]、预测分布差异[14]或对比损失 $\mathcal{L}_{\text{CL}}$[50]等。

如图 6 所示，PTQ 的典型流程主要包括 4 个阶段：

1）校准数据收集

在量化前，PTQ 需采集少量具有代表性的校准




(a) 传统统计式PTQ (b) 学习驱动PTQ (c) 无数据与硬件感知PTQ

Fig. 6 Illustration of typical pipelines for PTQ

图 6 PTQ 的典型流程示意图

样本，以统计网络各层权重与激活值的分布特征。校准数据一般来自训练集或验证集的一个小规模子集，规模通常为数百到数千个样本，用于估计各层特征的动态范围。当原始数据集不可获取或受隐私限制时，可通过生成式或统计匹配方式生成伪样本，如利用批归一化统计信息重建特征分布[51]，或通过伪标签过滤机制提高校准质量[52]。该步骤为后续参数初始化提供统计依据，是 PTQ 精度稳定性的基础环节。

2)量化参数初始化

根据校准数据的统计特性，首先确定各层权值与激活的截断区间 $[n, p]$ 或动态范围 $[x_{\min}, x_{\max}]$，并在此基础上计算缩放因子 $s=\dfrac{p-n}{2^b-1}$ 与零点 $z$。传统线性量化方法直接取权值或激活张量中的最小值与最大值作为截断边界[6]，而对称或非对称映射的选择则取决于数值分布特性。典型实现中，量化参数可按层、通道或张量级粒度独立设置，以进一步平衡模型精度与计算开销。

3)目标函数优化

在量化参数初始化之后，PTQ 通常需要进一步优化目标函数，以最大限度降低离散化带来的性能退化。这一阶段的核心任务是，在不重新训练模型的前提下，通过近似优化或输出重构恢复模型的特征表示与预测能力。早期研究通常将全局任务损失分解为逐层局部损失，以量化前后输出的均方误差或分布差异作为优化目标，从而在局部范围内近似全局性能[53]，但由于各层独立优化无法捕捉层间相关性，误差会逐层积累，导致模型整体精度显著下降。随着研究的深入，这一思路逐渐演化为基于输出重构的优化策略，不再仅依赖损失分解，而是直接最小化量化前后层或块级输出之间的重构误差，以显式减少量化扰动在网络中的传播，从而显著提升低比特量化下的稳定性[9,14,17]。

在此基础上，后续研究进一步引入了更精细的学习驱动机制，通过将量化步长、舍入阈值或位宽分配等参数转化为可优化变量[10,54]，引入连续松弛、二阶信息[9,15]或硬件资源约束[55]等策略，实现了从静态映射到自适应优化的跨越。此外，针对不同网络架构与硬件平台的特性，研究者还提出了基于预测分布对齐和系统感知校准的多样化方案，以增强模型在极低比特和跨设备部署场景下的鲁棒性[13-14,22]，这些优化策略的代表性实现与性能特征将在后续各节中分别展开讨论。

4)模型导出与部署

优化完成后，模型的权值和激活参数被映射为低比特整数格式，并在推理框架中以整型乘加方式执行计算，以实现更高效的部署与推理[6]。针对特定设备约束，研究者提出了多种硬件感知量化策略[22]，进一步引入硬件延迟、能耗等反馈信号，自动调整量化策略以获得更优的性能与能效平衡。

PTQ 部署成本低，适用于推理时延敏感且计算资源受限的场景。然而，由于优化空间受限于量化参数，PTQ 在极低比特条件下往往面临较大的精度损失。

### 1.2.2 QAT

在极低比特量化条件下，单纯依赖 PTQ 往往面临以下瓶颈：1)训练与部署阶段的分布失配问题较为突出。由于传统 PTQ 在训练过程中未显式建模量化噪声与饱和效应，当模型在推理阶段由实数计算切换至离散量化计算时，部分对数值变化高度敏感的模块容易出现数值不稳定。2)量化过程中涉及的截断与取整等离散算子本身不可导，使得量化误差难以通过基于梯度的优化机制进行有效修正，这一问题在极低比特设定下尤为显著。

QAT 通过在训练阶段显式引入与部署阶段一致的量化与反量化路径，并借助近似梯度方法跨越不可导操作，使模型在参数更新过程中能够直接感知离散化约束与量化噪声，从而逐步适应低比特表示。相较于 PTQ，QAT 在方法论上系统性缓解了低比特量化条件下的分布失配与误差不可回传 2 类核心问题。已有研究表明，QAT 在极低比特场景中通常能够获得更为稳定且显著的性能提升，尤其在对量化高度敏感的模型结构，如深层 Transformer、长链归一化路径以及深度可分离卷积网络中表现出更明显的优势[56-57]。

QAT 的优化目标可表示为：

$$\min_{W^*,\mathcal{F}} \mathcal{L}_{\text{task}}(Q(G;\mathcal{F},W^*))+\lambda\cdot\mathcal{R}(\mathcal{F}), \tag{10}$$

其中，$\mathcal{L}_{\text{task}}$ 为任务损失，$\mathcal{R}(\mathcal{F})$ 为量化参数的正则化项，可包括梯度缩放约束[11]、可学习裁剪约束[44]或蒸馏正则 $\mathcal{L}_{\text{distill}}$[58]等。

如图 7(a)所示，QAT 的典型流程主要包括 4 个阶段：

1)量化图初始化与后端对齐

在初始化阶段，需要依据目标推理后端的数值表示与算子约束，预先确定量化位宽、粒度和整数算术路径，并在训练图中插入与部署路径对应的伪量

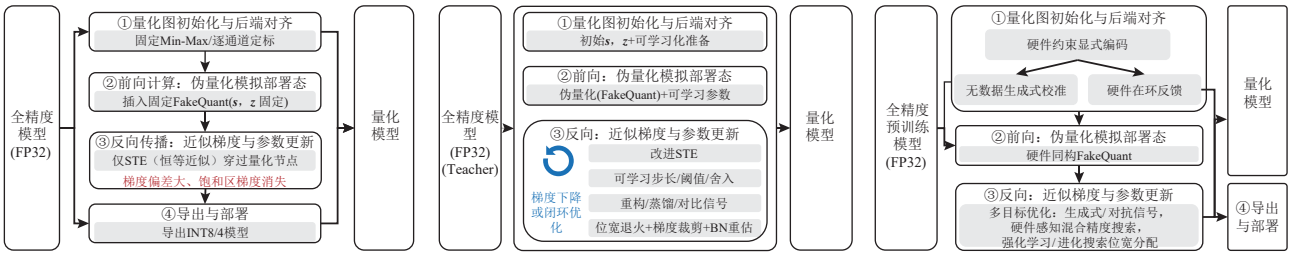



Fig. 7 Illustration of typical pipelines for QAT

图 7 QAT 的典型流程示意图

化节点。这样做的目的是在训练阶段尽可能模拟实际整数推理的数值行为，减少训练与部署之间的不一致。

量化参数的定标与校准对 QAT 的性能至关重要。静态定标方法通常依据训练数据的统计特性直接确定量化区间，而不依赖梯度优化；相比之下，学习式定标方法通过反向传播对量化参数进行联合优化，使其随数据分布与任务目标自适应收敛。实践中，权重量化通常采用逐通道定标以降低误差，而激活的定标粒度则根据算子敏感性在逐张量或分组级别进行选择。为进一步减小误差，常在网络首尾层或数值敏感路径中采用更高位宽，或结合高精度整型累加与重定标策略；当硬件对量化参数施加特定约束时，还需在训练过程中显式引入相应的约束条件或软惩罚项。

2）前向计算：伪量化模拟部署态

QAT 的核心在于将量化过程融入到训练的每一轮中，使得模型在训练过程中逐步适应低比特表示。前向传播在权重与激活路径执行伪量化，量化/反量化映射遵循式（7）的记号与含义：

$$\hat{x}=clip\left(round\left(\frac{x}{s}\right)+z,0,2^{b}-1\right),\quad x_q=s\times(\hat{x}-z)\ 。$$

量化过程通过四舍五入和零点调整映射到整数范围，再进行反量化恢复浮点数近似值 $\hat{x}$。对于注意力分数、归一化操作等数值敏感模块，通常采用更保守的位宽配置或混合精度旁路策略，以降低数值溢出或下溢的风险。

3）反向传播：近似梯度与参数更新

由于量化过程中的取整与截断操作不可导，若直接反向传播将导致梯度中断，因此 QAT 需要引入近似梯度机制以维持端到端可训练性。最常见的方法是采用直通估计器（straight-through estimator, STE）在反向传播中将量化函数的梯度近似为恒等函数，从而允许梯度通过量化节点进行传播，维持端到端可训练性。其标准形式为：

$$\frac{\partial\mathcal{L}}{\partial x}\approx\frac{\partial\mathcal{L}}{\partial\hat{x}}\cdot 1\left(q_{\min}\leqslant\frac{x}{s}+z\leqslant q_{\max}\right),\qquad(11)$$

其中，$\hat{x}$ 为反量化后的浮点近似值，$x$ 为原始浮点数值，$s$ 为缩放因子，$z$ 为零点，而 $1(\cdot)$ 为指示函数，表示仅在未饱和区传递梯度，当量化值在指定范围内时返回 1，否则为 0。

STE 属于有偏近似，在极低比特条件下可能引发梯度偏差放大或数值不稳定。为缓解上述问题，后续研究提出了多种改进策略，例如采用平滑替代函数近似取整算子的导数，或结合位宽退火、梯度裁剪、动量过滤、蒸馏增强与统计重估等稳定化机制。在此基础上，可进一步引入可学习步长、可学习阈值或渐进舍入策略，以降低边界区域的近似误差。梯度估计方法的选择对 QAT 的训练稳定性和最终精度具有关键影响。当前主流方法包括 STE、软量化、元梯度等，各方法在梯度偏差、方差、计算开销和适用场景上的特性对比如表 2 所示。

**Table 2 Feature Comparison of Different Gradient Estimation Methods**

**表 2 不同梯度估计方法特性对比**

| 梯度估计方法 | 代表工作 | 核心机制 | 梯度偏差 | 方差 | 计算开销 | 适用场景 |
|---|---|---|---|---|---|---|
| STE | 文献 [5, 57] | 饱和区梯度置零，非饱和区恒等传递 | 有偏 | 中等 | 低 | 通用 QAT |
| 软量化（soft quantization） | 文献 [19] | tanh/sigmoid 平滑替代舍入函数，温度控制逼近程度 | 低偏 | 低 | 中等 | 极低比特 |
| 元梯度（meta-gradient） | 文献 [59] | 超网络预测量化参数，双层优化显式计算超梯度 | 近无偏 | 低 | 高 | 生成/重建任务 |
| 梯度缩放 STE | 文献 [11] | STE 配合梯度缩放因子平衡参数更新幅度 | 有偏 | 较低 | 低 | 通用 QAT |
| 重参数化/软舍入 | 文献 [60] | 将离散量化决策连续化，交替优化缩小梯度差 | 低偏 | 低 | 中等 | 向量量化 |

4)导出与部署

当训练过程收敛后，模型进入导出与部署阶段。此时需导出整型权重及相应的量化参数，并将其映射至推理框架支持的整数计算路径。对于采用混合精度配置的模块，应保留训练时确定的位宽、量化粒度和缩放因子等配置，以减少训练模拟与实际推理之间的数值偏差，实现低比特模型的稳定部署。

相较于 PTQ，QAT 在训练过程中显式引入量化约束，使模型能够逐步适应低比特表示，因此具有更好的精度保持能力；但其代价是需要进行完整训练或较长时间的量化微调，计算开销与时间成本通常显著高于 PTQ。

### 1.3 学习驱动的量化

传统量化方法多依赖固定舍入规则与经验性参数设置，其量化流程具有较强的确定性，难以利用量化误差或精度反馈对策略进行动态调整。在 8 bit 等中等位宽条件下，该类方法通常能够实现较好的压缩与精度折中；但在 4 bit 及以下的极低比特场景中，量化误差显著放大，模型精度易出现明显退化，并表现出对校准数据敏感、跨架构泛化能力不足等问题。相比之下，学习驱动量化通过将量化过程建模为可优化问题，引入可学习机制与多样化监督信号，使量化参数与策略能够在训练或校准过程中自适应调整，从而在极低比特条件下提升精度保持能力与鲁棒性[61-63]；对于资源受限部署场景，还可进一步结合硬件反馈实现对不同模型结构与硬件平台的适配[21,55]。

学习驱动的量化可统一建模为以下优化问题：

$$\min_{\mathcal{F},W}\mathcal{L}_{\text{task}}(Q(G;\mathcal{F},W)),\ \text{s.t.}\ bitwidth(\mathcal{F})\leqslant B_{\text{budget}}。\tag{12}$$

设 $G$ 表示待量化的神经网络计算图或模型结构；$\mathcal{F}=\{q_1,q_2,\cdots,q_L\}$ 表示量化参数集合，包含各层权重与激活的量化函数或量化配置，其中 $q_l$ 表示第 $l$ 层的量化算子或量化策略，$L$ 为量化层数；$W$ 表示模型的全精度权重参数；$Q(G;\mathcal{F},W)$ 表示在给定模型结构 $G$、权重参数 $W$ 以及量化参数集合 $\mathcal{F}$ 条件下得到的量化模型；$\mathcal{L}_{\text{task}}(\cdot)$ 为任务损失函数；$bitwidth(\mathcal{F})$ 表示由量化配置 $\mathcal{F}$ 决定的整体比特宽度或平均量化位宽；$B_{\text{budget}}$ 表示部署场景下的比特预算约束。

该优化目标旨在在满足位宽预算的条件下，通过优化量化参数集合 $\mathcal{F}$ 最小化任务损失，从而获得性能最优的量化模型。

在实际部署场景中，数据可用性往往对量化优化过程构成重要限制。当原始训练数据部分可用或完全不可用时，目标函数难以通过标准的反向传播机制直接进行优化，尤其是在 PTQ 或数据受限场景下更为突出。针对这一问题，可将数据生成机制引入统一优化框架，在缺乏真实校准数据 $D$ 的情况下，通过生成器$G_{\text{gen}}$合成伪样本 $\tilde{x}$，并构建如下对抗式优化问题：

$$\min_{\mathcal{F}}\max_{G_{gen}}E_{\tilde{x}}[\mathcal{L}_{\text{task}}(Q(G;\mathcal{F},W),G)]。\tag{13}$$

该形式本质上构成生成器 $G_{\text{gen}}$ 与量化模型 $G$ 之间的对抗博弈，通过生成逼近真实分布的样本以缓解数据缺失带来的优化困难[61-64]。

另一方面，硬件资源约束同样会显著影响量化策略的设计。当部署平台对延迟、能耗或存储开销提出严格限制时，模型权重与激活的表示位宽需受到约束，从而改变优化目标的结构。为此，可在原有任务损失基础上引入硬件反馈函数 $\mathcal{H}(\mathcal{F})$(如延迟或功耗估计)，构建如下多目标优化问题：

$$\min_{\mathcal{F}}\mathcal{L}_{\text{task}}(Q(G;\mathcal{F},W))+\lambda_{\text{hw}}\cdot\mathcal{H}(\mathcal{F}),\tag{14}$$

其中 $\lambda_{\text{hw}}$ 用于权衡精度与硬件代价。该类问题通常通过强化学习[21]或可微搜索方法[55,65]进行求解，以实现性能与资源消耗之间的自适应平衡。

## 2 方法总览与分类

模型量化将高精度浮点参数映射为低位宽离散整数，能够降低模型推理的资源开销，实现大规模深度神经网络在算力网边端侧节点高效部署。针对传统静态策略在极低比特(≤4 bit)场景下面临的精度下降与适配性难题，学习驱动的量化通过将离散化过程重构为包含可学习参数的可微优化问题，利用重构误差最小化、辅助监督信号及系统级反馈机制，实现了量化策略对数据分布与硬件约束的自适应适配。本节将以此为核心，系统梳理该范式的关键机制与技术演进。

### 2.1 学习驱动量化方法的发展历程

学习驱动量化的发展历程可划分为 3 个阶段，如图 8 所示，早期探索阶段以二值网络、三值网络及增量量化等工作为代表，通过端到端训练验证了极低比特条件下模型可训练性的基本可行性[56,66-69]。技术成熟阶段逐渐形成 2 条并行的技术路线：一类方法通过引入可学习的量化参数，将量化步长、舍入策略及量化区间等关键因素纳入梯度优化或校准优化框架，实现量化映射的自适应调整[10-12,16,54]；另一类方法则聚焦于 PTQ 场景，在校准阶段利用重构误差最小化与二阶敏感度分析等机制，对量化参数进行数据

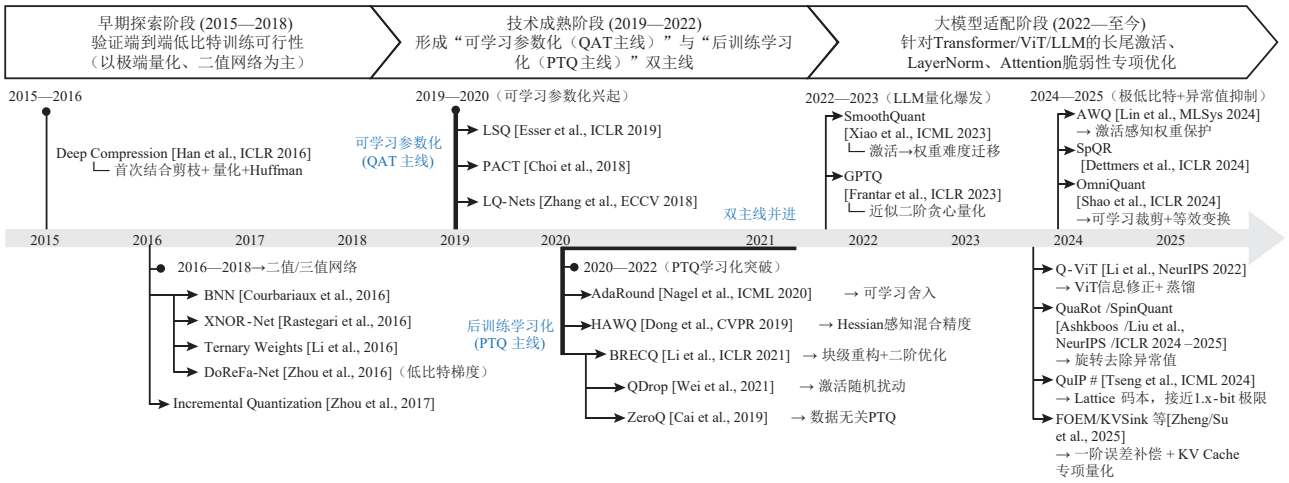



Fig. 8 Development process diagram of learning-driven model quantization methods

图 8 学习驱动模型量化方法的发展历程图

驱动的自适应估计[9,15,17]。大模型适配阶段面向 Transformer 与大模型中的激活离群值、层间敏感性及注意力分布失真等特异性挑战，发展出平滑迁移、可学习等效变换、近似二阶贪心搜索以及面向 ViT 注意力失真的专用校正机制等策略[24-26,70]，可实现大规模参数模型在异构算力节点上的高效精准部署。

## 2.2 学习驱动量化方法的分类框架

现有模型量化综述多从量化流程阶段或目标模型架构出发进行单一维度梳理，难以系统刻画学习驱动范式下不同方法在优化机理与适用场景上的差异。为此，本文对 PTQ 与 QAT 中具有代表性的学习驱动量化方法进行归纳，形成如图 9 所示的分类框架。该框架从学习机制介入量化过程的方式出发，重点考察优化对象、优化信号和作用阶段 3 个方面。其中，优化对象用于区分方法主要作用于量化函数参数、模型表征还是量化配置；优化信号用于刻画方法依赖任务损失、重构误差、教师监督、自监督信号或硬件反馈等信息；作用阶段用于说明相关机制主要应用于校准、训练或部署搜索过程。基于上述标准，本文将学习驱动量化方法归纳为参数可学习化、重构与近似优化、辅助学习信号和约束驱动方法 4 类：

1）参数可学习化

该类方法将量化步长、偏移、舍入策略及量化区间等量化函数的内部参数松弛为连续可微变量，通过梯度优化在训练或校准过程中自适应调整。该类方法的核心在于改变量化映射本身的结构，使其具备可优化自由度，学习信号直接来自任务损失或重构误差经由量化算子的梯度回传。

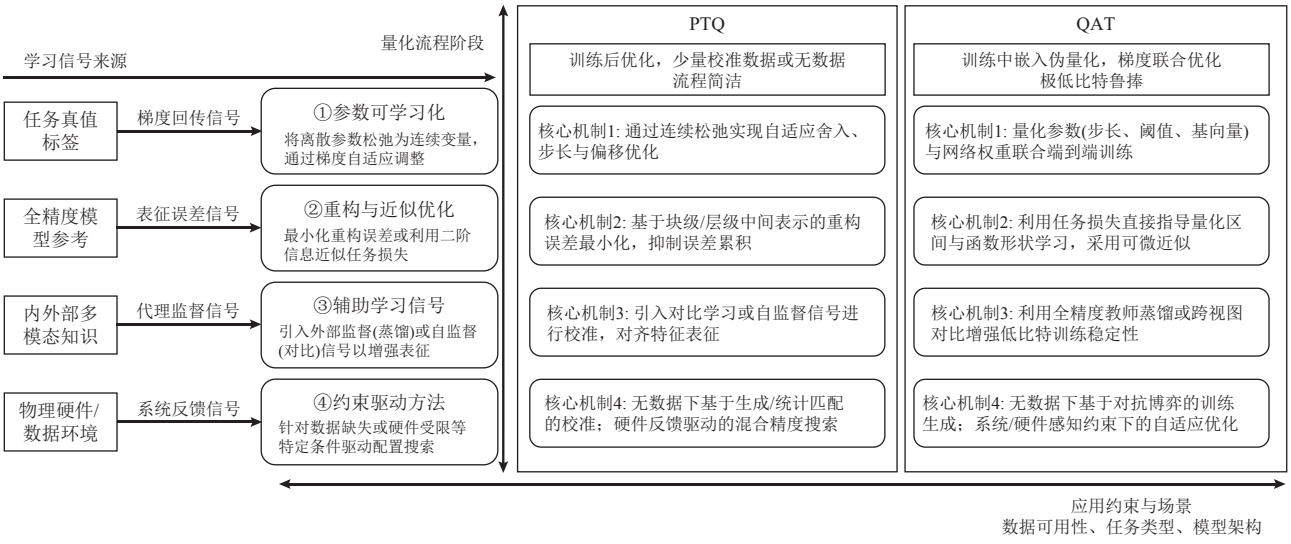



Fig. 9 Unified taxonomy framework of learning-driven quantization: A four-signal-source system under PTQ and QAT paradigms

图 9 学习驱动量化的统一分类框架：PTQ 与 QAT 范式下的四信号源体系

2)重构与近似优化

该类方法通过最小化量化模型与全精度模型在输出或中间表示上的重构误差来优化量化过程。常见的误差度量指标包括均方误差(mean squared error, MSE)、KL 散度(Kullback-Leibler divergence, KLD)和预测分布差异等。部分方法还引入 Hessian 或 Fisher 信息，用于估计局部量化扰动对模型损失的影响。该类方法主要用于缓解量化误差在层间传播带来的性能下降。

3)辅助学习信号

该类方法在主任务损失之外，引入额外监督信息来帮助量化模型保持表征能力。常见做法包括利用全精度模型进行知识蒸馏，构造正负样本进行对比学习，或设计自监督任务约束中间特征和输出分布。该类方法适合用于低比特训练不稳定或单一重构目标不足的场景。

4)约束驱动方法

该类方法面向数据受限或硬件受限等实际部署条件，将外部约束纳入量化优化过程。在无真实数据或样本极少的隐私受限场景下，系统利用生成模型合成代理校准数据，部分方法通过批归一化统计匹配来驱动量化参数寻优。在硬件资源受限条件下，算法将推理延迟、能耗及底层算子支持度纳入优化目标，结合强化学习或可微架构搜索机制，可以自动搜寻完全满足异构算力网节点特性的最优混合精度配置。

本框架围绕学习能力的引入方式，将 4 类策略分别对应于：1)学习量化参数(参数可学习化)；2)学习表征一致性(重构与近似优化)；3)学习内外部知识(辅助学习信号)；4)学习约束适配(约束驱动方法)，揭示了方法体系在优化对象与信号来源 2 个维度上的演进路径。对于同时融合多类机制的方法，本文按其主要贡献维度归入相应类别，并在行文中标注所涉及的其他机制，以保证框架的完整性与方法描述的准确性。基于该框架：第 3 节围绕 PTQ 流程，系统分析不同学习信号如何驱动校准、优化与部署阶段的性能提升；第 4 节在 QAT 场景下展开平行讨论，重点关注端到端训练中的梯度近似与稳定化机制；第 5 节则沿应用约束与架构特性，针对 CNN、ViT 与 LLM 三类模型，深入剖析学习驱动量化策略在不同计算单元中的具体实现形式及其面临的独特挑战。

表 3 基于上述框架，对代表性方法在量化范式、学习信号以及应用约束与目标场景等维度进行了逐项定位，并概括其关键技术要点。

## 3 学习驱动的 PTQ 方法

本节系统综述学习驱动的 PTQ 方法。PTQ 的基本概念与典型流程已在 1.2 节概述，本节直接聚焦于学习驱动机制如何在校准与优化阶段引入可学习参数、重构约束、辅助信号及特殊约束，围绕 4 类代表性方向展开技术分析。各类方法在优化机制、数据需求和适用场景上的差异汇总于表 4。

### 3.1 可学习参数的自适应校准

参数可学习化 PTQ 通过将离散映射重构为可微优化问题，旨在突破传统静态策略在极低比特场景下的性能瓶颈，以适应算力网边缘节点受限的存储与通信资源。传统 PTQ 依赖最近舍入等确定性规则，缺乏对网络层间异构权重分布的自适应能力，导致量化噪声在深层网络中逐层累积与放大，进而引发严重的精度下降。针对此局限，该范式将量化步长、舍入策略及截断阈值等关键参数转化为可优化变量，将原本不可导的离散量化过程转化为可近似优化的问题。在校准阶段，方法通常利用少量无标签样本最小化层输出、块输出或预测分布差异；在部署阶段，再将连续优化结果映射回离散量化空间，从而在不进行全量重训练的条件下改善低比特模型的精度保持能力。

可学习舍入是该方向中较具代表性的技术路线，如图 10 所示，传统最近舍入采用固定映射规则，误差不可优化；可学习舍入通过连续松弛将舍入决策转化为可梯度回传的变量，代表性方法在不同机制下各有侧重。早期工作 AdaRound[10] 将舍入方向建模为可学习变量，并利用二阶泰勒近似刻画舍入误差对任务损失的影响，再通过退火正则使软舍入结果逐步收敛到离散状态。该方法说明，PTQ 中的舍入误差并非只能由固定规则决定，而可以在校准数据上进行自适应调整。沿这一思路，FlexRound[54] 进一步从元素级尺度建模出发，联合优化全局步长与局部缩放因子，以增强对非均匀权重分布的适配能力；LLT[45] 则将量化映射表示为可微查找表，通过温度软化学习查表参数，适用于需要提高量化函数表达能力的场景。

除舍入和步长的局部优化外，面向层间耦合效应与算力网节点的动态部署需求，后续工作进一步考虑更大粒度上的参数耦合。AdaQuant[53] 在小规模校准集上联合更新权重与量化参数，并以层输出重

**Table 3 Taxonomy and Key Technical Highlights of Representative Methods**

**表 3 代表性方法的分类与关键技术要点**

| 范式 | 方法类别 | 学习信号 | 应用约束与目标场景 | 关键技术要点 |
|---|---|---|---|---|
| PTQ | 参数可学习化[10,45,71-73] | 可学习量化参数 | 数据充足；通用视觉任务；CNN/ViT | ① 从元素级或局部尺度出发联合学习步长、缩放与舍入参数；<br>② 通过连续松弛或可微查表建模离散量化映射；<br>③ 借助概率舍入、共享权重或位宽耦合提升低比特稳定性。 |
| | 重构/二阶为核心[9,17,23,74-76] | 重构与近似优化 | 数据充足或小样本；低/极低比特；CNN/ViT | ① 采用逐层或逐块重构，并结合 ADMM、Hessian/Fisher 等二阶信息抑制误差传播；<br>② 通过离散-连续联合优化与算子级分配提升量化配置质量；<br>③ 将激活量化转化为等效权重扰动进行统一分析；<br>④ 非均匀量化增强低比特表示能力。 |
| | 输出对齐与平坦化增强[13-14] | 重构与辅助信号协同 | 小样本；极低比特；CNN/ViT | ① 以预测差异或 KL 散度约束全精度与量化模型输出一致性；<br>② 通过随机激活丢弃等机制增强解的平坦性，提升跨分布鲁棒性。 |
| | 蒸馏或自监督信号驱动[50,77-82] | 辅助学习信号 | 数据充足、小样本或无数据；检测/点云/分类；CNN/ViT | ① 通过 logits 蒸馏或中间特征蒸馏维持任务相关表征一致性；<br>②以对比校准对齐全精度与量化表征，并在 ViT 中结合块级对比和语义合成增强鲁棒性；<br>③ 利用 BN 对齐、自监督或任务特定监督补充优化信号。 |
| | 无数据/生成/统计驱动[51-52,61,83-85] | 生成与统计约束驱动 | 无数据或隐私受限；通用分类；CNN | ① 基于 BN 统计匹配或代理分布恢复生成校准数据；<br>② 通过伪标签筛选、多样性增强等方法提升合成数据有效性；<br>③ 利用对抗或博弈机制暴露脆弱区域并改善泛化；<br>④ 在类条件生成或因果约束生成下提高代理数据的可用性。 |
| | 系统/硬件感知[15,22,55,86-87] | 系统与硬件约束驱动 | 硬件受限；边缘部署；CNN/Transformer | ① 基于 Hessian 敏感度或强化学习搜索分配混合精度位宽；<br>② 通过超网或 Pareto 前沿搜索快速生成多约束候选配置；<br>③ 结合通道分组、硬件兼容量化格式与整数推理路径优化部署效率。 |
| | ViT 专项[20,88-93] | 多信号联合优化 | 数据或硬件受限；低/极低比特；ViT | ① 采用分组可微量化、组内离散化或注意力结构感知策略，保持注意力秩与特征依赖关系；<br>② 通过训练稳定化机制缓解激活长尾和量化振荡；<br>③ 借助整数近似算子与硬件感知设计提升 ViT 推理效率。 |
| | LLM 专项[24-26,54,65,94-96] | 重构、参数优化与资源约束协同 | 小样本或免重训练；低比特推理；LLM/长上下文 | ① 在块级误差最小化框架下联合优化裁剪、舍入、缩放或等价变换参数；<br>② 通过激活平滑、分布校正与误差补偿提升低比特稳定性；<br>③ 在显存或算力约束下引入任意比特加速框架与全局位宽分配策略，实现实际推理加速。 |
| QAT | 参数可学习化[11-12,18,44,97-98] | 可学习量化参数 | 数据充足；通用视觉任务；CNN/ViT | ① 学习裁剪阈值、量化步长与偏移，使量化区间更贴合权重和激活分布；<br>② 通过可学习码本、基向量或量化级增强非均匀表达能力；<br>③ 部分方法根据输入分布动态调节步长或在线更新码本，提高训练适配性。 |
| | 重构/可微近似[16,19,46,99-100] | 重构与近似优化 | 数据充足；极低比特；CNN | ① 通过连续松弛、软量化等缓解离散量化带来的不可导问题；<br>② 结合二次幂或非均匀量化与可学习裁剪阈值，在保持硬件友好性的同时降低误差；<br>③ 采用滤波器级或神经元级渐进量化、改进梯度估计与梯度转发机制稳定训练。 |
| | 对比/教师信号驱动[58,101-102] | 辅助学习信号 | 数据充足；低比特感知任务；CNN/二值网络/3D 感知 | ① 通过对比学习或互信息最大化增强低比特表征的判别性；<br>② 在 3D 感知任务中利用视角引导蒸馏，对齐图像特征、BEV 特征与几何先验；<br>③ 借助 KL-only 或自监督知识蒸馏减少多损失权重调参与标签依赖。 |
| | 无数据/生成/对抗驱动[62-63,103-106] | 生成与对抗约束驱动 | 无数据；通用分类；CNN | ① 通过生成器建模、边界与困难样本构造提升合成样本质量；<br>② 采用零和博弈与平衡机制稳定生成-校准过程；<br>③ 部分方法使用 KL-only 与梯度增强策略改善极低比特 zero-shot quantization 的训练稳定性。 |
| | 系统/硬件感知（混合精度）[21,107-112] | 系统与硬件约束驱动 | 硬件受限；边缘部署；CNN/ViT | ① 通过硬件反馈或成本约束驱动混合精度搜索；<br>② 采用 BOPs、EBOPs 等资源指标建模片上开销与部署代价；<br>③ 借助背包优化或动态规划实现高效混合精度决策，部分方法支持实例自适应位宽控制。 |
| | LLM 专项[27,113-116] | 参数优化与辅助信号联合 | 极低显存；长上下文推理；LLM | ① 通过低秩适配与量化协同，结合量化网格感知、块级训练-端到端训练解耦或量化空间内联合优化，降低 QAT 开销；<br>② 部分方法引入数据驱动蒸馏增强机制；部分方法扩展至 KV cache 与权重/激活联合量化，缓解长上下文推理瓶颈。 |

构误差作为校准目标；LAPQ[71] 从损失曲面角度出发，将多层量化区间和步长纳入联合搜索，以缓解逐层独立校准带来的局部性问题。另有研究通过概率分布对舍入位置进行建模，并结合混合精度位宽分配，使权重可在多个候选量化格点之间选择[72]。这类方法在优化自由度上更高，但通常也带来更大的离线搜索或校准开销。

总体来看，参数可学习化 PTQ 的主要特点是以数据驱动的校准过程替代静态量化规则，使舍入、步长、量化区间或映射函数能够随模型分布进行调整。早期方法多关注舍入方向和步长等局部参数，优化流程相对简洁；后续方法逐步引入跨层联合搜索、概率建模和混合精度配置，以提升对复杂分布和资源约束的适配能力。表 5 从可学习参数类型、优化目标、松弛机制和稳定化策略等方面对相关方法进行比较。

### 3.2 重构误差驱动的优化方法

面向算力网边缘节点的隐私约束与低比特推理需求，传统基于统计量的静态校准策略难以抑制量

**Table 4 Comparison of Key Contributions of Representative Learning-Driven PTQ Methods**

**表 4 学习驱动 PTQ 代表性方法关键贡献对比**

| 方法类别 | 代表方法 | 核心方法论 | 优势 | 局限性 |
|---|---|---|---|---|
| 参数可学习化 | AdaRound[10] | 基于二阶泰勒近似的连续松弛舍入，配合退火正则化策略实现自适应舍入方向学习 | 无需全量数据；低比特精度保持好 | 逐层独立优化，未考虑层间耦合 |
| | FlexRound[54] | 元素级除法视角联合优化全局步长与局部缩放因子，大权重可探索更广离散值范围 | 适应非均匀权重分布；LLM 兼容 | 元素级参数学习带来更高的优化开销 |
| | LLT[45] | 将量化映射构建为可微查找表，通过温度软化学习最优查表参数 | 推理时查表高效；精度与效率兼顾 | 需额外存储查找表 |
| 重构与近似优化 | LAPQ[71] | 损失感知联合优化框架，通过二次插值在多层参数空间搜索最优量化区间 | 全局一致性好；2~4 bit 收敛稳定 | 无梯度优化计算开销大 |
| | L-DNQ[23] | 逐层 ADMM 优化，将量化形式化为带离散约束的二次优化问题 | 极少样本下稳定；理论保证 | 逐层独立、误差累积 |
| | BRECQ[9] | 块级二阶重构，以 Fisher 信息近似敏感度，对关键通道赋予更高权重 | 显式削弱跨层误差传播；4 bit 可用 | 块划分策略需人工设计 |
| | PD-Quant[14] | 以预测差异为全局校准指标，并通过分布校正缓解小校准集过拟合 | 任务目标一致；分布校正有效 | 依赖 BN 统计准确性 |
| | QDrop[13] | 在重构阶段随机跳过部分激活量化，以增强解的平坦性并提升极低比特泛化 | 2~4 bit 表现突出；泛化性强 | 引入额外随机性 |
| 辅助学习信号 | CL-Calib[50] | 对比学习校准，最大化 FP 与量化激活互信息实现分布对齐 | 无标签条件下保留语义信息任务对齐强； | 需构建正负样本对 |
| | LiDAR-PTQ[79] | 利用 FP 模型输出构造正样本监督，约束量化模型最终预测一致性 | 3D 检测适用 | 任务依赖强，通用性有限 |
| | QuanDCL[117] | 知识蒸馏与对比学习并行，约束 logits 与特征一致性 | 多信号协同；低比特有效 | 多损失权重调节复杂 |
| 约束驱动（数据缺失约束） | ZeroQ[51] | BN 统计匹配重建代理数据，优化蒸馏数据集逐层匹配统计量 | 完全无真实数据；混合精度支持 | 深层网络统计匹配收敛慢 |
| | ZAQ[61] | 对抗学习驱动生成器合成挑战样本，最大化量化与 FP 输出差异 | 暴露脆弱区域；鲁棒性提升 | 对抗博弈平衡困难 |
| | GDFQ[104] | 知识匹配生成器合成类条件样本，BN 统计与分类边界双重对齐 | 生成样本语义一致；开创性工作 | 生成器训练开销大 |
| 约束驱动（硬件资源约束） | HAWQ[15] | Hessian 敏感度分析指导混合精度分配，分块微调提高精度 | 理论指导位宽分配；自动化 | Hessian 计算开销大 |
| | HAQ[22] | 强化学习（RL）框架自动搜索最优位宽，硬件反馈驱动策略优化 | 硬件适配性强；端到端优化 | RL 训练不稳定 |
| | One-Shot MPS[55] | 一次性 Pareto 前沿搜索，可学习位宽预测模型线性生成配置 | 搜索效率高；无需再训练 | 超网训练成本 |

化误差的逐级累积，且在激活量化过程中由于截断边界与离散格点分布难以精确匹配真实数据分布，部分激活值可能集中饱和于区间端点或压缩至有限离散级别，导致有效比特利用率下降与中间特征表达退化，进而在深层网络中放大量化扰动。为缓解上述问题，重构误差驱动的方法依托少量无标注样本，在层级或块级尺度上对齐量化模型与全精度模型的中间表示。通过最小化重构误差及其二阶近似，优化过程能够将量化扰动限制在局部范围内，从而控制误差的跨层传播。

逐层重构是较早采用的优化方式。L-DNQ[23] 将每层量化表述为带离散约束的二次优化问题，并利用 ADMM 将连续更新与离散投影分开求解，从而在有限校准数据下完成逐层逼近。与逐层方法相比，BRECQ[9] 将重构单元扩展到网络块，在块内联合优化舍入与步长，并引入 Fisher 信息近似二阶敏感度，对量化扰动较敏感的通道赋予更高权重。该方法更适合处理残差连接和归一化结构带来的层间依赖，是重构类 PTQ 中较具代表性的工作。AdaQuant[53] 同样以输出重构为目标，联合校准权重与量化参数，并通过重新估计 批归一化统计缓解量化后的分布偏移，可视为逐层重构思路的扩展。

仅对齐中间表示并不一定保证最终预测一致，因此部分研究进一步将优化目标扩展到输出分布或解空间性质。PD-Quant[14] 以预测差异作为全局校准指标，通过约束量化模型与全精度模型的输出分布差异来搜索量化参数，并结合分布校正缓解小样本校准中的统计偏移；QDrop[13] 则在块级重构阶段随机跳过部分激活量化操作，将激活量化带来的扰动纳入校准过程，并从损失平坦性角度提高低比特 PTQ 的稳定性；后续研究进一步尝试将权重量化和激活量化放入统一分析框架中，分别结合块级重构、二阶近似和随机化校准解释不同扰动来源对最终误差的影响[17]。

当逐层或逐块校准难以刻画全网误差传递时，一些方法转向更大范围的联合优化。文献 [74] 将

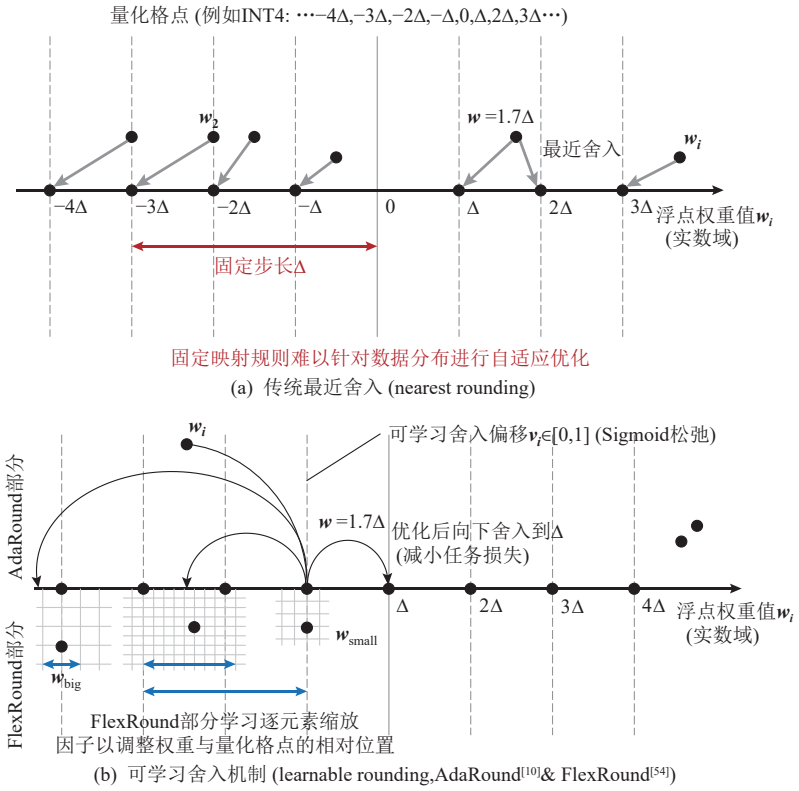



Fig. 10 Comparison of traditional nearest rounding and learnable rounding mechanisms

图 10 传统最近舍入与可学习舍入机制对比

**Table 5 Technical Detail Comparison of Learnable-Parameter PTQ Methods**

**表 5 参数可学习化 PTQ 方法技术细节对比**

| 方法 | 可学习参数 | 优化目标 | 松弛机制/可微建模 | 正则/稳定化机制 | 适用位宽 |
|---|---|---|---|---|---|
| AdaRound[10] | 舍入方向 | 逐层重构误差 | Sigmoid 连续松弛舍入 | 退火正则 | 4 bit |
| FlexRound[54] | 步长、逐元素/逐通道缩放因子 | 层/块输出重构误差 | 元素级除法重参数化 | L2 正则 | 3~4 bit |
| LLT[45] | 查表参数 | 任务损失 | 可微查找表；温度软化 | 温度退火 | 2~4 bit |
| LAPQ[71] | 量化区间、步长 | 全局损失曲面 | 损失感知搜索；二次插值 | — | 2~4 bit |
| AdaQuant[53] | 权重、步长 | 逐层/逐块重构 MSE | 联合连续优化 | 权重正则 | 4 bit |
| Attention Round[72] | 舍入概率分布 | 重构误差、混合精度分配目标 | 高斯衰减分布 | 有损编码长度约束 | 2~4 bit |

注: “—”表示未涉及或无明确机制。

PTQ 建模为离散与连续变量共同参与的端到端优化问题, 并利用温度退火的 ASoftmax 机制处理离散选择的梯度近似; 文献 [53] 顺序最小化各层输出的重构误差, 联合校准权重、步长与 批归一化统计, 并通过整数规划完成混合精度位宽分配, 在时延与能耗约束内取得可靠折中。另一些工作从算子或表示形式入手, 文献 [75] 通过码字分配与缩放因子的联合求解降低算子输出近似误差; 文献 [76] 采用非均匀或多级二进制表示, 并结合 soft-bit 退火优化码字与缩放参数。这类方法能够增强低比特表示能力, 但通常依赖特定量化形式或额外的离线优化过程。

总体来看, 重构误差驱动方法的差异主要体现在优化粒度和约束目标上。如图 11 所示, 逐层重构(独立逐层优化, 无法控制跨层误差传播, 误差逐级放大导致深层精度严重退化)实现简单, 但对层间依赖刻画有限; 块级重构(通过块内联合优化步长与舍入, 显式抑制误差跨块传播; 利用二阶敏感度加权保护敏感特征)能够更好地适配残差块等结构, 是当前低比特 PTQ 的重要技术路线; 输出分布对齐直接约束预测行为, 但通常依赖额外的统计校正; 全网络、算子级和非均匀表示方法提供了更大的优化空间, 同时也带来更高的搜索和部署复杂度。表 6 对相关方法的优化粒度、重构目标和误差控制方式进行了比较。

### 3.3 辅助信号驱动的量化策略

辅助信号驱动范式跳出局部数值重构, 通过引

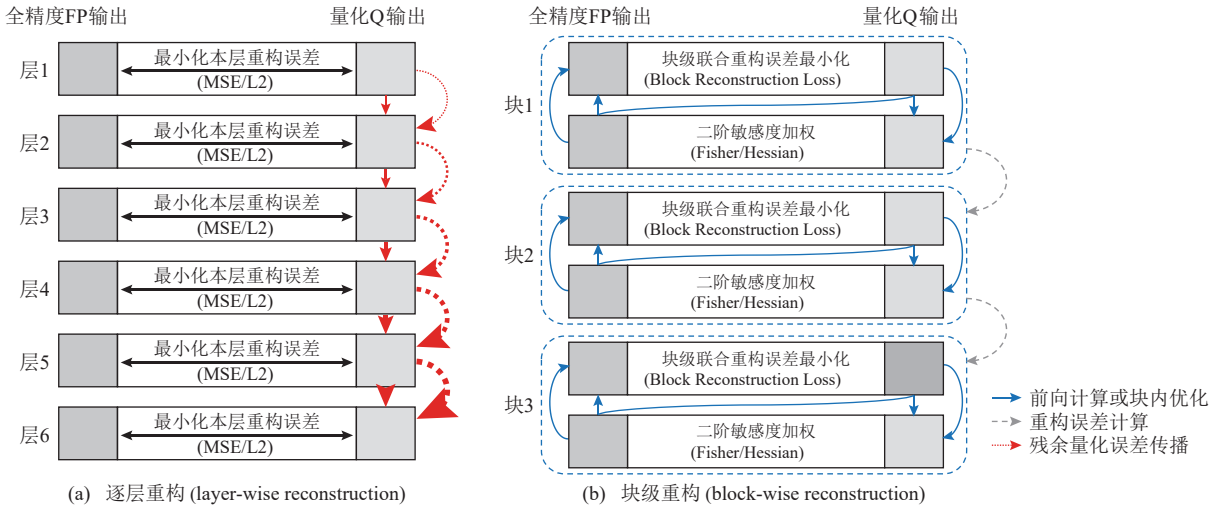



Fig. 11 Comparison of error propagation between layer-wise reconstruction and block-wise reconstruction

图 11 逐层重构与块级重构误差传播对比

**Table 6 Comparison of Optimization Granularity and Objectives in Reconstruction-Optimized PTQ Methods**

**表 6 重构优化 PTQ 方法的优化粒度与目标对比**

| 优化粒度 | 方法 | 重构目标 | 二阶信息 | 误差传播控制 | 理论保证 | 核心创新 | 局限 |
|---|---|---|---|---|---|---|---|
| 逐层重构 | L-DNQ[23] | 层输出 MSE | 离散二次优化 | 逐层级联重构，限制单层误差 | ADMM 收敛性分析 | ADMM 解耦连续更新与离散投影 | 忽略层间依赖 |
| | AdaQuant[53] | 层输出 MSE | | 顺序校准，联合 BN 统计校正 | | 权重与量化参数联合优化；BN 统计校正 | 局部最优；深层网络效果受限 |
| 块级重构 | BRECQ[9] | 块输出 MSE | Fisher 对角近似 | 块内联合优化，显式抑制跨层误差 | 二阶误差分析 | 在依赖建模与泛化之间取得块级折中 | 计算开销较大 |
| | QDrop[13] | 块输出 MSE | | 随机丢弃激活量化，平坦化解空间 | | 将激活量化显式纳入重构，提升平坦性 | 迭代次数较多 |
| 全局输出对齐 | PD-Quant[14] | 预测差异损失 | | 分布校正，直接约束输出一致性 | | 以全局预测差异替代局部特征重构 | 依赖分布校正效果 |
| 全网络联合 | NWQ[74] | 全局输出 MSE | 隐式二阶（ASoftmax） | 全网联合优化，显式建模层间依赖 | 收敛性与稳定性分析 | 网络级重构 + 退火离散优化 | 优化规模大、调参较复杂 |
| 算子级优化 | 文献 [75] | 算子输出近似误差 | | 码字分配与缩放联合最优 | 全局最优性保证 | 算子层全局最优码字求解模块 | 需与整体量化流程结合 |
| 表示增强 | Mr.BiQ[76] | 块输出重构误差 | | soft-bit 退火稳定离散化 | | 非均匀/多级二进制表示提升低比特表达力 | 非均匀量化硬件支持有限 |

入内外部监督信息来指导并约束量化过程。如表 7 所示，根据监督来源的差异，此类方法主要划分为教师监督与自监督信号两大类。图 12 展示了不同信号在网络中的作用位置：特征蒸馏与对比学习通常作用于网络中间层，旨在对齐全精度与量化模型的表征空间以保持语义一致性；Logits 蒸馏与任务监督则聚焦于输出层，直接约束模型在特定任务上的预测结果。在实际应用中，多层级信号的组合（如文献 [117] 同时采用特征蒸馏与对比学习）通常能够取得更优的量化效果，但也增加了损失权重调节的复杂性。

3.3.1 教师监督信号驱动

教师监督是辅助信号中较常见的一类。其基本思路是将全精度模型作为教师，通过输出分布、特征表示或任务预测引导量化模型，使其在低比特条件下尽量保持与原模型一致的行为，如图 13 所示。现有方法主要通过 2 类蒸馏机制实现上述目标，即基于输出的 Logits 蒸馏与基于中间表示的特征蒸馏。前者通过对齐量化模型与教师模型的预测分布，优化量化模型的决策边界，以减少量化引入的性能偏移；后者则通过引导量化模型学习教师模型中间层特征的分布结构，在保持计算效率的同时提升模型的表示能力与泛化性能。总体而言，该类方法通过引入蒸馏信号对量化过程进行任务感知的微调，使量化模型能够更好地适应低比特表示带来的约束，从而获得更接近全精度模型的性能表现。

在代表性方法中，PD-Quant[14] 以预测差异作为全局校准指标，通过约束量化模型与全精度模型输出分布的一致性来搜索量化参数，并结合分布校正

缓解小样本校准中的统计偏移。QuanDCL[117]进一步将 Logits 蒸馏、特征蒸馏与对比学习结合，在输出层和特征层同时施加约束，以减轻低比特量化带来的表征退化。面向具体任务的工作则多在这一思路上引入任务相关监督。例如，Reg-PTQ[77]利用分类与回归信号引导检测模型校准；LiDAR-PTQ[79]针对点云检测中的稀疏特征和前景样本不足问题，引入任务导向的正样本监督；2DQuant[118]面向图像重建任务，

**Table 7 Comparison of Teacher-Supervised and Self-Supervised Signals**

**表 7 教师监督与自监督信号对比**

| 信号类型 | 代表方法 | 监督来源 | 对齐层级 | 损失函数 | 优势 | 局限 |
|---|---|---|---|---|---|---|
| 预测差异监督 | PD-Quant[14] | 教师输出 | 输出层 | 预测差异损失/KL 散度 | 任务对齐直接 | 依赖教师与统计校正 |
| 任务监督 | Reg-PTQ[77] | 教师预测 | 任务输出层 | 局部重构损失/全局任务损失 | 任务级优化 | 任务特定设计 |
| | LiDAR-PTQ[79] | 教师正样本框/任务输出信号 | 任务输出层 | 任务导向的全局正样本损失 | 任务导向，保留几何语义 | 场景依赖强 |
| | 2DQuant[118] | 教师软标签/中间特征 | 输出层与特征层 | 区间初始化/蒸馏校准 | 适配重建任务，兼顾视觉效果 | 双阶段流程复杂 |
| 对抗蒸馏 | ZAQ[61] | 教师输出与中间通道关系 | 输出层与特征层 | 对抗样本差异最大化 | 零样本可用教师信号 | 对抗训练不稳定 |
| 对比学习 | CL-Calib[50] | 自身激活 | 激活/特征层 | 对比损失/互信息最大化目标 | 无标签下保持表征信息 | 负样本构造敏感 |
| 自监督 | FDDA[81] | 合成数据 BN 统计分布 | 分布/特征层 | 类内对齐/类间区分 | 统计对齐，适合数据受限 | 依赖 BN 统计与合成数据质量 |
| | SR-PTQ[80] | 自监督重建一致性 | 输出层与特征层 | 像素一致性损失 +特征校准损失 | 自监督保持细节一致 | 任务专用性强 |
| 蒸馏+对比联合监督 | QuanDCL[117] | 教师 logits、教师特征与对比关系 | 输出层与特征层 | logits/特征蒸馏损失+对比损失 | 恢复输出一致性与特征判别性 | 目标函数较多，权重敏感 |

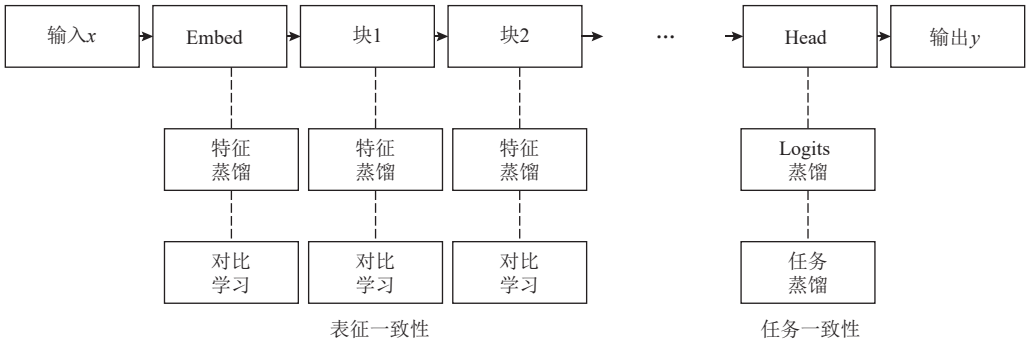



Fig. 12 Illustration of the functional hierarchy of auxiliary learning signals

图 12 辅助学习信号的作用层级示意图

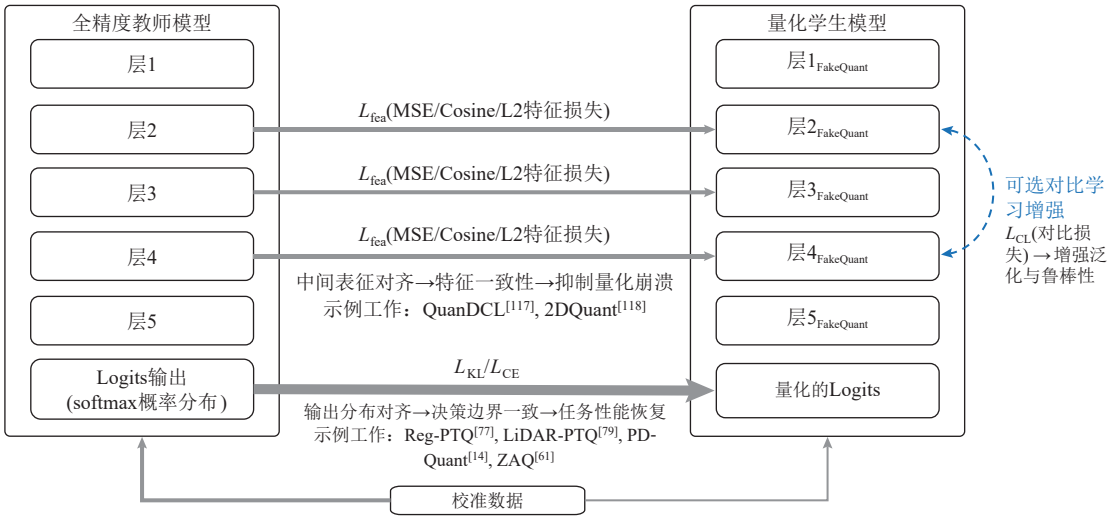



Fig. 13 Hierarchy of teacher supervision signals in quantized networks

图 13 教师监督信号在量化网络中的作用层级

通过区间初始化和蒸馏校准约束输出与中间特征一致性。另有工作在无真实数据条件下引入生成式或对抗式信号，如 ZAQ[61] 通过生成样本放大全精度模型与量化模型之间的输出或通道差异，再利用这些样本优化量化模型。

总体来看，教师监督信号驱动的量化方法借助全精度模型提供的输出、特征或任务级参考，将校准目标由局部数值重构进一步扩展到预测行为与任务性能的一致性。相关方法覆盖输出层、特征层和任务层等多个作用层级，既包括基于输出分布或中间表示的对齐，也包括面向检测和重建任务的任务感知校准。不过，此类方法通常依赖可获得的教师模型，并会带来额外的损失设计与优化复杂度。

3.3.2 自监督与对比学习信号驱动

面向隐私敏感与数据受限的边缘计算场景，传统教师监督 PTQ 依赖高质量全精度模型与标注数据而存在一定的局限。为此，自监督与对比学习信号提供了另一类校准思路：不直接依赖人工标签，而是利用样本自身的统计结构、量化前后的表征关系或合成数据中的分布约束，引导量化模型保持特征表示的一致性。具体而言，自监督方法通常通过重构、预测或统计匹配等代理目标约束量化误差；对比学习方法则构造全精度表征与量化表征之间的正样本关系，并利用不同样本或不同视图作为负样本，以减轻低比特量化造成的表征偏移，如图 14 所示。

基于自监督与对比学习信号驱动的 PTQ 方法在量化精度和鲁棒性方面取得了显著的进展。CL-Calib[50] 是该方向中较具代表性的工作。该方法将对比学习引入 PTQ 校准阶段，将同一样本的全精度激活与量化激活视为正样本对，将不同输入对应的激活视为负样本对，通过对比目标约束量化前后的激活分布一致性。与仅采用均方误差重构的校准方法相比，这类方法更关注特征空间中的相对关系，有助于在无标签条件下保持一定的语义结构。QuanDCL[117] 进一步将对比学习与蒸馏信号结合，在特征层和输出层同时约束量化模型，使对比信号作为教师监督的补充，用于减轻低比特条件下的表征退化。

在无数据或小样本量化场景中，部分工作将自监督信号与合成数据生成结合。文献 [78] 通过合成数据和对比学习任务构造校准信号，用于无数据视觉模型量化；FDDA[81] 利用合成数据的细粒度 BN 统计信息对齐类内与类间分布，以改善代理数据的校准质量；IntraQ[82] 则关注合成样本的类内差异，通过局部目标增强和边界距离约束提高样本多样性。面向图像超分辨率任务，SR-PTQ[80] 将自监督重建一致性引入校准过程，在粗粒度区间初始化后，进一步对齐量化模型与全精度模型的中间特征和输出结果。

总体来看，自监督与对比学习信号能够在缺少标签、教师监督不足或真实数据不可用时，为量化优化提供额外信息，帮助模型减少量化后的特征偏移。其优势在于对任务标签依赖较低，适合数据受限和隐私受限场景；局限在于，这类信号多作用于中间特征、统计分布或代理任务，对最终任务输出的约束相对间接。因此，在极低比特量化中，自监督或对比学习方法常与教师蒸馏、重构优化或生成式校准联合使用，但相应会增加损失设计和超参数调节的复杂度。

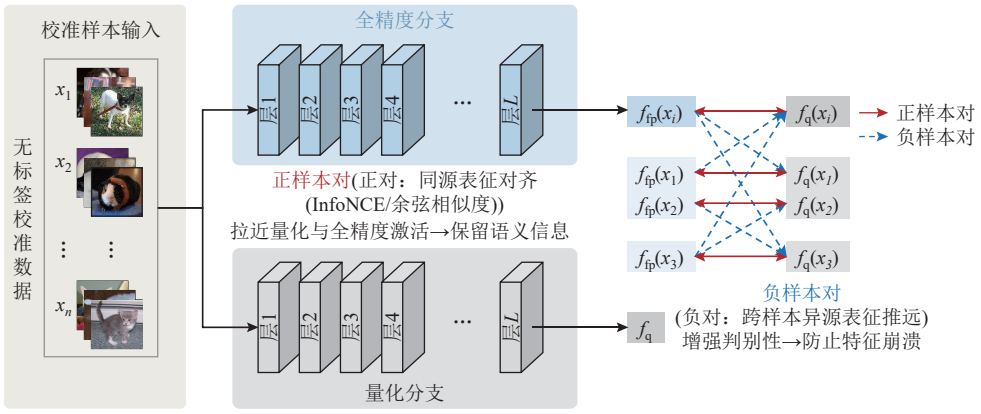



Fig. 14 Construction of positive and negative samples and signal flow of contrastive learning in PTQ

图 14 PTQ 中对比学习的正负样本构建与信号流向

## 3.4 特殊约束下的校准方案

3.4.1 数据受限：统计匹配与生成式校准

原始训练数据往往受到隐私保护、合规要求或产权限制而难以获取，这一问题在云边端协同场景中尤为突出，边端节点通常无法访问集中式训练数据，使得依赖真实校准数据的传统 PTQ 方法在实际

部署中存在局限。为应对这一问题，数据无关量化（data-free quantization, DFQ）近年来受到广泛关注，成为支持边端侧本地部署的关键技术路径。该范式旨在摆脱对原始训练数据的依赖，通过利用预训练模型中的统计先验或构造代理样本，为量化参数校准提供替代数据来源。现有方法大致可分为 2 类：一类基于 BatchNorm 统计、激活分布或伪标签筛选构造校准约束；另一类通过生成模型、对抗优化或条件约束合成代理数据，以补充真实样本缺失带来的信息不足。

统计匹配是 DFQ 中较早采用的思路。ZeroQ[51] 提出零样本量化框架，通过优化随机输入，使其在前向传播过程中匹配预训练模型各层的 BatchNorm 均值和方差，从而生成可用于校准的代理样本，并结合自动精度选择支持混合精度量化；DFQ-PLF[52] 进一步引入伪标签可靠性筛选，根据熵或自信息度量样本置信度，筛选更可靠的合成样本用于校准。该类方法实现相对简洁，但对模型中统计信息的可用性和合成样本质量较为敏感。

生成式方法进一步提高了代理数据的表达能力。GDFQ [104] 与 CGAN[84] 分别从生成器设计和类别条件约束 2 个方向增强合成样本质量：前者通过知识匹配生成器同时利用分类边界知识与 BatchNorm 统计生成更具判别性的代理数据，后者则借助零样本条件生成框架生成类一致的合成样本，并将其用于无数据知识蒸馏和量化优化。为缓解生成样本覆盖不足的问题，ZAQ[61] 通过放大全精度模型与量化模型在输出或中间通道关系上的差异，生成更具挑战性的样本；IntraQ[82] 和文献 [83] 则分别从类内异质性和特征多样性角度增强样本覆盖范围。近期的 Causal-DFQ[85] 尝试引入因果建模，通过内容与风格解耦构造代理样本，以在生成过程中保持任务相关内容的一致性。

综上所述，DFQ 的核心在于利用模型内部统计信息或合成代理数据替代真实校准集。统计匹配和伪标签筛选方法流程较轻，适合无数据但模型统计信息可用的场景；生成式和对抗式方法能够提供更丰富的样本分布，但通常需要额外的生成器训练或更复杂的优化过程；类条件、多样性增强和因果约束等方法进一步改善了代理数据质量，但也提高了实现成本。表 8 从生成机制、约束来源和主要局限等方面对代表性方法进行比较。

**Table 8　Comparison of Data Generation Strategies in Data-Free Quantization Methods**

**表 8　无数据量化方法的数据生成策略对比**

| 方法 | 范式 | 生成机制 | 统计约束 | 语义约束 | 多样性增强 | 生成质量 | 局限 |
|---|---|---|---|---|---|---|---|
| ZeroQ[51] | 统计驱动生成 | 输入优化 | BN 均值/方差 | | | 中等 | 强依赖 BN |
| DFQ-PLF[52] | 统计+伪监督 | 伪标签筛选 | BN/激活统计 | 高置信伪标签 | 熵/自信息筛选 | 较高 | 伪标签敏感 |
| GDFQ[104] | 条件 GAN 生成 | ACGAN | BN 统计 | 类条件先验 | 类覆盖更均衡 | 较高 | 需训练生成器 |
| ZAQ[61] | 对抗生成 | 对抗生成器 | 隐式对齐 | 差异最大化 | 难样本覆盖 | 高 | 对抗开销大 |
| IntraQ[82] | 多样性生成 | 输入优化 | BN 统计 | 软标签 | 类内异质性 | 较高 | 类内约束复杂 |
| 文献 [83] | 多样性生成 | 多层特征混合器 | 激活统计/裁剪范围适配 | 类间与类内多样性约束 | 模式坍塌缓解 | 高 | 机制较复杂，需训练生成器 |
| CGAN[84] | 条件生成式生成 | 零样本条件生成器 | BN 统计 | 类别条件约束 | 类条件样本生成 | 较高 | 依赖预训练分类器 |
| Causal-DFQ[85] | 因果约束生成 | 内容-风格解耦生成 | | 内容一致性约束（任务相关） | 风格干预扩大覆盖 | 高 | 因果建模与优化流程较复杂 |

### 3.4.2　硬件受限：敏感度分析与精度配置

在边缘计算与移动设备等资源受限场景中，量化方法不仅需要控制精度损失，还需考虑存储占用、推理延迟、能耗和算子支持等部署代价。仅依据数据分布或重构误差确定量化参数，往往难以反映不同硬件平台上的实际执行开销。因此，硬件感知量化将系统反馈或资源约束纳入量化决策过程，重点解决混合精度位宽分配和硬件适配问题。

基于敏感度分析的方法通常先评估不同层对量化扰动的敏感程度，再据此分配位宽。HAWQ[15] 代表了二阶敏感度分析驱动的典型方向。该方法利用层级 Hessian 谱刻画不同模块对量化扰动的敏感程度，并据此确定混合位宽的相对分配顺序与分阶段微调顺序，从而在较小搜索代价下完成更稳定的混合精度配置；HAQ[22] 将硬件反馈引入强化学习搜索过程，根据目标平台上的延迟、能耗等指标，为不同层选择量化位宽，从而使量化配置更贴近实际部署需求。One-Shot MPS[55] 则采用 one-shot 超网搜索框架，通过可学习的位宽概率模型一次性生成一组 Pareto 前沿候选架构，并在常数级搜索时间内完成不同压

缩率下的混合精度配置选择。与需要多次重启搜索的方案相比,该方法显著降低了大模型混合精度搜索的时间成本。面向 SAM(segment anything model)等具有重尾激活与显著通道差异的视觉基础模型,AHCPTQ[86] 提出了一种基于硬件感知的 PTQ 方法,利用混合量化和通道感知分组技术,在控制参数开销的同时缓解通道间分布差异,从而在 FPGA 和 ASIC 场景下实现更高效的低比特部署。

总体来看,硬件受限场景下的 PTQ 不再只追求量化误差最小化,而是需要在精度、延迟、能耗和存储开销之间进行权衡。二阶敏感度方法具有较好的可解释性,适合在较小搜索空间内进行混合精度配置;强化学习和 one-shot 搜索能够处理更大的设计空间,但通常需要额外的搜索或训练成本;面向 FPGA、ASIC 的方法则进一步将位宽分配扩展到通道分组和算子实现层面,使量化策略更贴近底层执行约束。表 9 对相关方法的反馈信号、搜索策略和目标硬件进行了比较。

**Table 9 Comparison of Hardware-Aware Quantization Methods**

**表 9 硬件感知量化方法对比**

| 方法 | 核心反馈信号 | 搜索策略 | 硬件代理指标 | 敏感度分析 | 配置粒度 | 目标硬件 |
|---|---|---|---|---|---|---|
| HAWQ[15] | 二阶敏感度 | 贪心搜索 | 模型压缩率 | Hessian 谱 | 层级位宽 | 通用 GPU |
| HAQ[22] | 硬件反馈 | 强化学习搜索 | 延迟/能耗/存储 | 直接硬件反馈 | 层级位宽 | 多平台 |
| One-Shot MPS[55] | 硬件代理指标 | 一次性搜索或位宽概率建模 | BOPs/延迟 | 位宽概率模型 | Pareto 候选配置 | 通用 |
| AHCPTQ[86] | 硬件资源约束 | 混合量化与通道感知分组 | 查找表/数字信号处理单元/寄存器开销 | 通道感知分组 | 通道或分组粒度 | FPGA/ASIC |

## 4 学习驱动的 QAT 方法

在 1~4 bit 等极低比特条件下,PTQ 面临训练-部署分布失配与误差不可回传的固有瓶颈,QAT 通过在训练阶段显式引入与部署阶段一致的量化与反量化路径,并借助近似梯度方法跨越不可导操作,使模型在参数更新过程中能够直接感知离散化约束与量化噪声,逐步适应低比特表示。本节围绕学习驱动的 QAT 展开分析,其基本原理与典型流程已在 1.2 节概述,本节重点分析学习驱动机制在 QAT 范式中的具体实现。各类方法的比较见表 10。

### 4.1 可学习参数的梯度优化训练

在 QAT 中,传统的量化方法通常依赖于固定的量化参数,这些参数在训练过程中未得到进一步的优化。然而,这些量化参数对低比特量化模型的性能具有重要影响。因此,近年来的研究提出将量化参数显式设为可学习变量,并通过端到端训练中的任务损失进行优化。这一方法使得量化过程不仅能够适应网络权重的变化,还能随着任务目标的优化自动调整,从而改善低比特条件下的精度退化。

在实际应用中,量化器的参数更新通常与网络权重同步进行,以确保模型在量化后的低比特表示下保持较高的任务性能。可学习裁剪和步长学习是该方向的代表性思路。PACT[44] 面向激活量化,将裁剪阈值设为可学习参数,并通过正则项约束其取值范围,使不同层能够根据激活分布学习合适的截断区间。LSQ[11] 进一步将量化步长设为可学习参数,并结合直通估计器和梯度缩放因子,平衡步长与网络权重的更新幅度,缓解低比特训练中的梯度尺度不匹配问题。现有可学习量化[12,44] 默认激活为无符号或对称有符号,Swish 等激活函数的负值部分会被钳制为零,导致信息丢失并限制了量化效果。为了应对这一问题,LSQ+[12] 在文献 [11] 的基础上引入可学习偏移量,构造非对称量化形式,以适配 Swish 等包含负值输出的激活函数,并通过初始化策略降低训练初期的不稳定性。

除基础量化参数外,部分方法进一步扩展了量化器结构或动态适配能力。LQ-Nets[18] 将量化级表示为可学习基向量与二值编码的组合,在训练中联合优化网络参数和量化器参数,从而支持权重和激活的多比特量化。ASQ[98] 则在激活端引入动态步长,使量化尺度能够随输入分布变化进行调整;其权重端采用非均匀量化形式,以更好匹配权重分布。L4Q[115] 结合 QAT 和低秩适配(low-rank adaptation, LoRA)思想,在量化感知微调过程中联合优化量化参数与低秩适配参数。其核心做法是在合并原始权重与 LoRA 更新后再执行量化,从而避免推理阶段的混合精度路径;同时通过局部计算并及时释放量化相关梯度,降低了训练阶段的显存开销。AdaQAT[119] 进一步将权重和激活位宽建模为可学习变量,并在目标函数中同时考虑任务损失与 BitOPs 等资源代价,

用于混合精度 QAT 的位宽配置。

综上所述，参数可学习化 QAT 的关键在于将量化过程从静态映射转化为可随任务目标联合优化的训练变量。表 11 对相关方法进行了系统对比，现有方法通常以步长、裁剪阈值和偏移量等基础量化参数为切入点，通过将其设为可学习变量，并结合 STE、梯度缩放或辅助模块与网络权重共同优化，在极低比特设置下实现对权重与激活分布的自适应调节。

**Table 10　Comparison of Key Contributions of Representative Learning-Driven QAT Methods**

**表 10　学习驱动 QAT 代表性方法关键贡献对比**

| 方法类别 | 代表方法 | 核心方法论 | 优势 | 局限性 |
|---|---|---|---|---|
| 参数可学习化 | PACT[44] | 首次引入可学习激活裁剪参数，配合 L2 正则化控制阈值大小，为每层提供自适应裁剪 | 激活裁剪范围可学习；分层量化策略灵活 | 仅针对激活量化；正则化超参需调节 |
| | LSQ[11] | 可学习量化步长、梯度缩放因子平衡步长与权重更新 | 训练稳定；实现简洁；广泛适用 | 对非对称激活分布的适配能力有限 |
| | LSQ+[12] | 在 LSQ 基础上加入可学习偏移，支持非对称量化，采用基于 MSE 的初始化提高训练稳定性 | 适合含负值激活的网络；较 LSQ 训练方差更低 | 三参数共优化易震荡，存储开销增加 |
| | LQ-Nets[18] | 将量化级表示为可学习基向量与二进制编码的组合，并联合训练网络参数与量化器 | 量化器可学习；适用权重和激活的任意比特量化 | 训练复杂度较高 |
| 重构/可微近似 | QIL[16] | 将量化器分解为变换器与离散器，并以任务损失直接学习量化区间 | 区间学习直接受任务损失驱动；兼顾剪枝与裁剪 | 区间参数化与训练过程相对复杂 |
| | DSQ[19] | 采用双曲正切近似构造可微软量化函数，并通过演化式训练逐步逼近标准量化 | 前后向一致性更好；有利于稳定训练 | 软硬切换需调参；温度调度需精心设计 |
| | APoT[46] | 提出加性二次幂非均匀量化，并通过重参数化裁剪函数改善裁剪阈值的梯度学习 | 硬件友好移位运算；适配钟形/长尾分布 | 查表开销；表示形式受限于二次幂 |
| | NbN[99] | 采用滤波器级和神经元级 2 层渐进量化，逐步量化权重与激活避免一次性扰动 | 细粒度控制；提升训练稳定性 | 渐进策略设计复杂 |
| | LG-LSQ[100] | 结合缩放模拟梯度（SSG）、反正切软舍入（ASR）、最小化离散化误差（MDE）3 项机制 | 3 种技术协同兼顾误差控制；边缘场景高精度 | 方法组件较多，联调成本较高 |
| 辅助学习信号 | SQAKD[58] | 统一量化器前向与反向动态，将量化误差与 KL 散度共同最小化，无需多重损失权重 | 跨架构平均增益；避免超参数调节；1~4 bit 稳定 | 依赖教师模型 |
| | CMIM[101] | 将二值激活与潜在实值激活视为同一样本的 2 种视角，通过对比学习最大化二者互信息 | 增强二值激活的判别表示；可迁移 | 需构造对比样本并引入额外训练开销 |
| | QD-BEV[102] | 渐进式 QAT 配合视图引导蒸馏，结合相机外参生成 BEV-mask 实现视图一致性分布对齐 | 3D 感知任务适用；视图一致性 | 任务特定设计 |
| 约束驱动（数据缺失约束） | HAST[106] | 利用样本难度度量优先合成并强化难样本训练，同时加入特征对齐约束提升零样本量化性能 | 强化对难样本的建模；增强量化模型鲁棒性 | 样本难度调控策略需要额外设计 |
| | Qimera[105] | 通过叠加潜在嵌入生成靠近决策边界的支持样本，并引入解耦映射层改善叠加嵌入的分布表达 | 强化边界样本生成；有利于提升无数据量化性能 | 嵌入与生成器设计较复杂 |
| | AIT[103] | 舍弃交叉熵仅用 KL 散度，引入梯度淹没机制动态调整各层梯度缩放因子 | 消除多目标冲突；极低比特稳定 | 较依赖教师模型提供的蒸馏信号 |
| | AdaDFQ[62] | 将 DFQ 建模为零和博弈，并通过适应性边界调节生成样本对量化网络的有效性 | 避免样本过难或过易 | 博弈式优化过程复杂 |
| | AdaSG[63] | 基于纳什博弈框架，用 Balance Gap 引导生成与校准过程达到平衡 | 理论支撑强；自适应停训 | 平衡判据与训练过程仍需验证和调节 |
| 约束驱动（硬件资源约束） | RL-MPQ[21] | 引入强化学习智能体，以硬件延迟配合精度作为奖励，在层级混合精度搜索空间中自动分配位宽 | 能够把硬件约束纳入位宽决策过程；适配多平台 | 强化学习搜索与量化重训练开销较高 |
| | HMQ[110] | 开发出硬件友好的混合精度量化块，将位宽纳入可学习变量，通过有限差分近似位宽梯度 | 同时考虑位宽与阈值；满足各种硬件约束 | 搜索空间大；位宽离散性优化困难 |
| | AdaQAT[119] | 采用可学习的实值松弛位宽，并在量化运算中离散化，从而在训练中自适应优化权重与激活位宽 | 支持混合精度；既适用于从头训练，也适用于微调 | 位宽优化仍存在离散-连续近似误差 |

**Table 11　Comparison of Representative Methods for Learnable-Parameter QAT**

**表 11　参数可学习化 QAT 代表性方法对比**

| 方法 | 可学习参数 | 梯度估计 | 核心创新 | 位宽支持 | 特点 |
|---|---|---|---|---|---|
| PACT[44] | 裁剪阈值 | STE | L2 正则化控制激活裁剪阈值 | 2~4 bit | 可学习激活裁剪 |
| LSQ[11] | 量化步长 | STE+梯度缩放 | 步长与权重联合优化 | 2~4 bit | 代表性步长学习方法 |
| LSQ+[12] | 步长+偏移 | STE | 支持可学习非对称量化 | 2~4 bit | 适配 Swish 等负值激活 |
| ASQ[98] | 动态步长 | STE | 基于输入特征动态调整步长 | 3~4 bit | 动态激活量化 |
| LQ-Nets[18] | 基向量+量化级 | 可微内积 | QEM 量化误差最小化 | 1~4 bit | 联合学习量化器 |
| L4Q[115] | LoRA+量化参数 | STE | 量化与低秩适配协同优化 | 3~4 bit | 降低训练显存开销 |
| AdaQAT[119] | 量化位宽 | 有限差分 | 任务+硬件双目标优化 | 自适应 | 自动位宽搜索 |

在此基础上,进一步扩展到量化器结构、动态步长或位宽优化的工作,增强了量化映射的灵活性,使模型能够更好地应对输入分布变化、极低比特离散化误差以及部署侧资源约束。尽管此类方法通常引入更高的训练开销,并对初始化和超参数设置较为敏感,但在低比特推理和资源受限部署场景中仍具有较高应用价值。

### 4.2 重构误差驱动的可微优化

在 QAT 场景中,早期方法往往更关注量化误差、分布拟合或固定量化配置,而在低比特条件下,这类策略未必能够直接对应最终任务性能。为此,后续研究逐步将优化重点转向 2 条互补路径:一类通过重构量化区间、量化函数或裁剪机制,使量化器本身可训练;另一类则从梯度近似误差出发,改进优化过程以缓解训练偏差与不稳定性。

在区间与函数重构方面,QIL[16] 直接以任务损失学习量化区间,将量化过程分解为变换器和离散器,并对量化区间的中心、尺度及区间内变换进行参数化建模,使裁剪边界能够由任务损失端到端学习。加性二次幂量化(additive powers-of-two, APoT)[46] 通过将量化级表示为多个二次幂项之和,在保持移位友好计算特性的同时,构造更符合钟形分布的非均匀量化级分布。相较于原始二次幂量化,APoT 在峰值区域分配了更合理的量化级,并缓解了分辨率僵化问题;同时,它还引入重参数化裁剪函数以改进阈值学习,并结合权重归一化提升训练稳定性。DSQ[19] 则利用双曲正切函数对阶梯量化进行平滑近似,并在训练过程中逐步由软量化过渡到硬量化,以减小前向离散化与反向梯度近似之间的不一致。

除量化函数本身的可学习化外,也有研究从梯度近似误差入手改进训练过程。文献 [60] 在向量量化网络中分析了 STE 可能带来的优化偏差,并通过仿射重参数化、交替优化和同步承诺损失等机制缓解码本更新滞后、梯度稀疏和索引塌缩问题。文献 [59] 则从元学习角度建模码本训练,引入超网络生成码本,并通过双层优化使码本更新更直接地响应任务目标。上述工作主要面向向量量化网络,但其对梯度近似偏差和量化器层级优化的分析也为 QAT 中的稳定训练提供了参考。

表 12 对比了重构与近似优化 QAT 方法的设计思路与特性。总体来看,这 2 类方法都试图将量化过程由对固定映射的被动适配,转向由任务目标牵引的可训练过程。前者通过量化区间、裁剪机制或量化函数的可学习化增强量化器表达能力,后者则通过重参数化、交替优化或双层优化等策略改善梯度近似质量与训练稳定性。已有研究表明,这些方法能够在低比特设置下缓解性能退化,但其实际效果仍与模型结构、位宽配置和训练策略密切相关。

**Table 12 Comparison of Reconstruction and Approximation-Optimization QAT Methods**

**表 12 重构与近似优化 QAT 方法对比**

| 方法 | 可微机制 | 优化目标 | 核心创新 | 优势 |
|---|---|---|---|---|
| QIL[16] | 区间参数化 | 任务损失 | 将量化器分解为变换器与离散器,区间端到端学习 | 区间即任务响应域,量化区间可随任务目标端到端更新,精度与可部署性兼顾 |
| DSQ[19] | 双曲正切平滑近似 | 任务损失 | 用可微平滑函数逼近阶梯量化,并通过演化式训练逐步逼近硬量化 | 平滑到离散的渐进过渡,低比特训练稳定 |
| APoT[46] | 重参数化裁剪 | 任务损失 | 非均匀加性幂量化,形状与硬件协同设计 | 硬件友好非均匀量化,表达能力更强 |
| STE 改良[60] | 仿射重参数化 | 减少梯度估计偏差 | 训练中引入交替优化与同步承诺损失,缓解分布失配与索引塌缩 | 梯度校正理论化,收敛更稳定 |
| MetaQuant[59] | 超网络生成码本 | 元学习目标 | 双层优化,显式学习量化器参数 | 梯度真实回传,训练期增强,推理零开销 |

### 4.3 辅助信号增强的稳定化策略

在低比特 QAT 中,量化会压缩中间表征并引入梯度近似误差,仅依赖任务损失训练时容易出现收敛不稳定或精度下降。为缓解这一问题,部分研究在训练阶段引入教师监督、自监督或对比学习等辅助信号,对量化模型的输出分布、中间特征或实例级表征关系进行约束。与增加推理分支的方法不同,这类辅助信号通常只作用于训练过程,部署时可移除教师模型或辅助模块,因此一般不会增加推理阶段的计算和存储开销。

教师监督是较常用的稳定化方式。Zhuang 等人[120] 提出在训练期为低比特网络接入全精度辅助模块,与低比特主干共享参数、联合优化,形成混合精度网络。该方法使主干参数既可通过量化路径更新,也可通过全精度辅助路径获得额外梯度信号,从而缓解低比特训练中的梯度传播困难。文献 [121] 进一步结合渐进量化、随机精度训练和知识蒸馏,通过由高位宽到低位宽的过渡,以及部分层或模块随机保持全精度的方式,降低低比特训练初期的优化难度。对于更复杂的多视角三维感知任务,QD-BEV[102]

在分阶段渐进式 QAT 的基础上，引入视图引导蒸馏，同时利用图像域与鸟瞰域特征，并借助相机外参生成的鸟瞰掩码融合 2 类蒸馏损失，从而提升多视角量化训练的稳定性与性能。

在自监督与对比增强方向，SQAKD[58] 提出自监督量化感知知识蒸馏框架，通过统一不同量化器的前向与反向动态，将量化训练表述为蒸馏散度与离散化误差的协同优化问题。该方法仅使用蒸馏散度作为训练目标，避免了多损失权重平衡，并在无标签条件下完成训练。CMIM[101] 则主要针对二值神经网络，利用二值激活与潜在实值激活在同一次前向传播中的共存关系，将二者视为同一样本的 2 种表征视角，并通过对比学习近似最大化它们之间的互信息，以增强极低比特模型的判别能力。

上述方法从不同维度指向同一目标：在低比特条件下引入额外的稳定化约束，使量化训练不再只依赖单一任务损失。表 13 总结了代表性方法中辅助信号的组合方式。教师监督通过软输出蒸馏或特征对齐，将优化约束扩展到输出分布和中间表征层面；自监督与对比学习则通过一致性约束和互信息增强稳定低比特模型的表征结构。从优化角度来看，不同方法在多损失组合、权重设置和统一目标构造上各有侧重：显式加权方法更灵活，但通常更依赖超参数；统一目标或免权重平衡设计则有助于降低调参负担。总体而言，辅助学习信号为低比特量化提供了有效补充，能够在不改变部署形态的前提下改善训练稳定性与最终性能。

### 4.4 特殊约束下的联合优化方案

#### 4.4.1 数据受限：生成对抗与博弈式训练

无数据量化在 PTQ 中主要关注校准样本的构造，而在 QAT 中还需依赖合成样本持续参与梯度更新，因此对样本质量、多样性和训练稳定性提出了更高要求。当真实数据不可用时，生成器、教师模型和量化模型之间需要共同优化，若生成样本过于简单或过度偏向对抗性，都会影响低比特模型的收敛。因此，无数据 QAT 的核心问题在于如何构造有效代理样本，并在生成过程与量化训练之间保持相对稳定的优化关系。

从生成式伪样本驱动的基础范式来看，GDFQ[104] 提出知识匹配生成器，以高斯噪声和随机标签为输入，通过类别边界约束和批归一化统计匹配生成伪样本，再利用这些样本蒸馏并微调量化模型。该方法表明，即使不访问真实数据，也可以从全精度模型中提取类别信息与分布统计，构造对量化训练有效的代理样本。这一工作为后续的生成式量化方法[62,105] 奠定了基础。在此基础上，Qimera[105] 进一步关注现有生成式方法对决策边界附近样本覆盖不足的问题，通过叠加潜在嵌入生成边界支撑样本，以增加决策边界附近样本；HAST[106] 则强调困难样本的重要性，通过样本重加权、扰动增强、特征对齐和输出蒸馏，提高合成样本对量化训练的约束作用。

另一类方法从生成器与量化模型的交互关系出发，将无数据 QAT 视为动态博弈过程。AdaSG[63] 将生成器和量化模型之间的优化建模为零和博弈，并利用 Balance Gap 判断训练过程的平衡状态。AdaDFQ[62] 则在此基础上进一步区分分歧样本与一致样本，并通过优化二者之间的边界间隔来调节样本适应性，从而缓解低比特下的欠拟合和较高位宽下的过拟合问题。与单纯追求更强对抗性的样本相比，这类方法更强调生成过程与量化过程之间的动态平衡，图 15 直观展示了生成器与量化模型之间的动态博弈过程。AIT[103] 发现传统“交叉熵损失（CE）+KL 散度”双损失在零样本量化蒸馏中梯度持续冲突，导致权重更新停滞；提出“All In the Teacher”策略：彻底丢弃 CE，仅保留 KL 散度单一目标，并设计梯度淹没机制动态缩放每层梯度，确保量化权重稳定更新，显著提升无数据低比特蒸馏效率与性能。

综上所述，无数据 QAT 通过在缺乏真实样本的条件下构造代理信号，实现了生成过程与量化优化的协同推进，为数据受限或隐私保护场景下的低比

**Table 13 Auxiliary Signal Combination Strategies in QAT**

**表 13 QAT 中辅助信号的组合策略**

| 方法 | 主损失 | 辅助信号 1 | 辅助信号 2 | 信号融合方式 | 权重调节 | 机制 |
|---|---|---|---|---|---|---|
| 标准 QAT | 交叉熵 | | | | | 仅依赖任务损失，低比特易失稳 |
| KD-QAT | 交叉熵 | 软输出蒸馏 | | 加权求和 | 手动设置 | 教师软标签缓解量化信息丢失 |
| SQAKD[58] | KL 散度 | 自监督蒸馏 | | 统一目标 | 自适应 | 蒸馏与量化误差协同最小化 |
| CMIM[101] | 交叉熵 | 对比学习 | | 加权求和 | 手动设置 | 最大化二值激活与实值激活的互信息，增强极低比特判别性 |
| QD-BEV[102] | 任务损失 | 图像域 KL 散度 | 鸟瞰域 KL 散度 | 视图引导融合 | 手动设置 | 结合相机外参与鸟瞰掩码实现多视角蒸馏 |

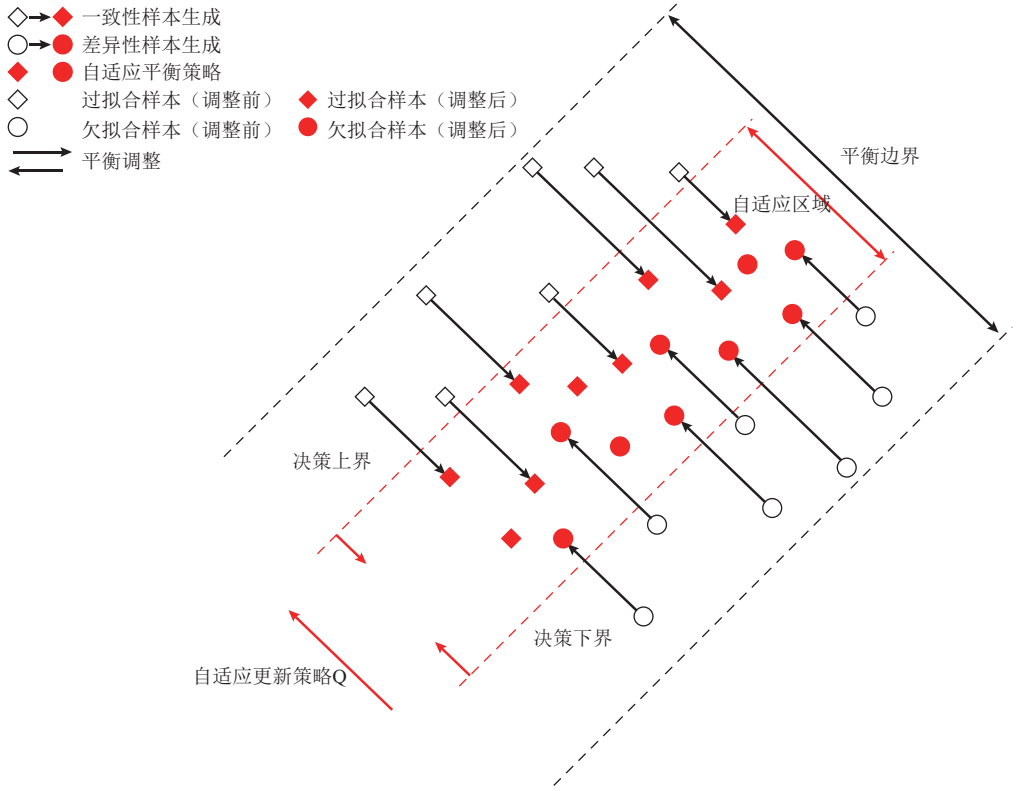



Fig. 15 Adaptive sample generation and balancing

图 15 自适应样本生成与平衡

特量化提供了切实可行的解决思路。表 14 从生成信号来源、生成机制、量化器更新方式以及协同优化策略等关键维度，对代表性方法进行了系统对比。现有方法中，一类研究依托噪声驱动或网络统计信号，通过对抗生成、边界样本合成或基于 KL 散度约束的伪样本构造，尝试覆盖模型的决策空间[103-105]，其优势在于对数据分布假设较弱、适用范围广，但对训练稳定性、停止条件及生成质量控制提出了更高要求；另一类研究则在生成阶段显式引入标签结构、决策边界信息或类别均衡约束，使合成样本在语义层面更贴近任务判别结构[62-63,106]，从而在量化训练过程中获得更稳定的收敛行为与更可靠的精度提升。这一范式成功将量化训练与真实数据解耦，为构建隐私安全、合规可控的分布式 AI 推理系统提供了切实可行的工程化路径。

#### 4.4.2 硬件受限：可微搜索与精度分配

硬件感知量化中利用延迟、能耗等系统反馈信号指导精度分配的基本思路，已在前述章节中有所涉及。在 QAT 中，混合精度配置需要同时考虑任务精度和硬件代价。与固定精度训练不同，位宽选择本身是离散变量，难以直接纳入梯度优化；同时，随着网络层数增加，候选位宽组合迅速扩大，使搜索效率和资源约束建模成为主要问题。为此，现有方法通常从 2 条路径展开：一类将位宽选择松弛为可微变量，并在训练过程中联合优化量化参数与硬件代价；另一类利用强化学习或真实硬件反馈，在较大设

**Table 14 Generation–Quantization Co-optimization Strategies in Data-Free QAT Methods**

**表 14 无数据 QAT 方法的生成-量化协同优化策略**

| 方法 | 生成信号来源 | 生成器机制 | 量化器更新 | 协同机制 | 停止准则 | 训练稳定性 |
|---|---|---|---|---|---|---|
| GDFQ[104] | 噪声+随机标签 | ACGAN 生成样本 | 交替更新 | BN 匹配+蒸馏 | 固定 epoch | 中等 |
| Qimera[105] | 噪声+叠加嵌入 | 边界支持样本生成 | 交替更新 | 决策边界对齐 | 固定 epoch | 较高 |
| HAST[106] | 噪声 | 难样本生成 | 交替更新 | 难度重加权+蒸馏 | 固定 epoch | 中等 |
| AdaDFQ | 噪声 | 自适应生成 | 博弈更新 | 纳什均衡 | Balance Gap | 高 |
| AdaSG[63] | 噪声 | 零和博弈生成 | 博弈更新 | BG 平稳性约束 | Balance Gap → 0 | 高 |
| AIT[103] | 统计信号 | KL-only 策略 | 仅蒸馏 | 梯度淹没 | 收敛判断 | 高 |

计空间中搜索满足资源约束的混合精度配置。

可微搜索方法通过连续松弛降低混合精度配置的搜索成本。HMQ[110]提出提出硬件友好的混合精度量化块，利用 Gumbel-Softmax 在候选集合中搜索位宽和量化阈值，并引入对称、均匀和 2 的幂阈值等硬件友好约束，使量化配置更易映射到实际计算单元。CGMQ[107]则从约束优化角度出发，通过门控变量控制权重和激活的位宽选择，并在给定 BOP 上界下优化量化模型，使位宽配置在训练过程中满足资源预算。与依赖外部搜索器的方法相比，这类方法能够将精度和资源约束放入同一训练框架中，但其效果依赖于连续松弛和资源代理指标的准确性。

在资源代理指标与更细粒度位宽优化方面，HGQ[108]针对 FPGA 平台的查找表（look-up table, LUT）或数字信号处理单元（digital signal processing block, DSP）资源约束，利用 EBOPs 指标替代人工超参，通过梯度下降自动完成关键权重的保留与冗余剪枝。ADQ[109]则结合硬件代价函数和敏感度分析，将有限比特预算更多分配给量化敏感层，以实现精度与计算预算之间的折中。对于难以显式建模的硬件约束，RL-based MPQ[21]采用强化学习策略，将硬件延迟、能耗或精度作为奖励信号，根据层结构和权重统计等状态决定位宽分配。该类方法能够利用真实或模拟硬件反馈，但通常需要较高的搜索成本。

总体来看，硬件受限的混合精度 QAT 主要围绕“可微松弛”和“硬件反馈搜索”展开。可微方法训练流程较紧凑，适合资源约束可用代理指标描述的场景；强化学习方法对黑盒硬件和非线性约束的适应性更强，但搜索过程更复杂。表 15 从搜索算法、约束形式、搜索粒度和反馈信号等方面比较了代表性方法。

**Table 15 Comparison of Bit-Width Search Methods for Mixed-Precision QAT Under Hardware Resource Constraints**

**表 15 硬件资源约束下混合精度 QAT 位宽搜索方法对比**

| 方法 | 搜索算法 | 可微性 | 资源约束 | 搜索粒度 | 硬件反馈信号 | 搜索开销 |
|---|---|---|---|---|---|---|
| HMQ[110] | Gumbel-Softmax | 可微 | 候选量化配置搜索 | 层级 | 硬件友好量化约束 | 1 个训练 |
| RL-MPQ[21] | 强化学习 | 不可微 | 硬约束（延迟/能耗） | 层级 | 模拟加速器速度反馈 | 多个训练 |
| CGMQ[107] | 门控变量+方向更新 | 可微 | 硬约束（BOP 上限） | 层级 | BOPs | 1 个训练 |
| ADQ[109] | 敏感度加权分配 | 可微 | 计算预算约束 | 层级/通道 | 计算预算/启发式敏感度 | 1 个训练 |
| HGQ[108] | 梯度下降 | 可微 | 硬约束（EBOPs） | 参数级 | EBOPs/片上资源估计 | 1 个训练 |

## 5 大模型与跨架构量化方法

本节围绕大模型与跨架构量化方法展开系统介绍。针对 CNN 架构，梳理 PTQ 与 QAT 在低比特场景中的应用演进，重点分析可学习量化参数、混合精度配置与渐进式量化等策略；针对 ViT 架构，围绕自注意力机制量化敏感性、LayerNorm 数值脆弱性及长尾激活分布等结构特异性问题，介绍信息修正、分布改造与纯整数推理等代表性方法；针对 LLM 架构，聚焦异常激活主导的长尾分布、KV 缓存压缩及极低比特部署等核心挑战，总结激活平滑、二阶敏感度引导校准与可学习等价变换等关键技术路径；最后从跨架构视角横向对比 3 类模型在核心计算单元、激活分布特性、量化敏感层及推荐安全位宽等维度的异同，揭示共性挑战与架构特异性规律，并展望统一量化框架的未来研究方向。

### 5.1 CNN 架构

CNN 作为深度学习领域中最为经典且应用广泛的模型架构之一，在计算机视觉等任务中取得了显著成效。然而，随着网络深度与参数规模的不断增长，其计算与存储开销不断上升，在嵌入式设备和资源受限平台上的部署压力也随之加大。

CNN 量化面临 3 类核心挑战：首层与尾层对量化扰动远比中间层敏感，层级敏感度差异显著；深度可分离卷积等轻量结构中的激活分布较为集中，截断区间难以通过简单统计获得；在极低比特（2~3 bit）条件下，离散化误差会显著放大梯度估计偏差，导致训练不稳定。针对这些挑战，学习驱动量化分别在 PTQ 与 QAT 两条路线下发展出一系列代表性方法，图 16 给出了 PTQ 与 QAT 的基本区别。

针对 CNN 的低比特部署，早期研究多围绕 PTQ 展开。该类方法通过将权重和激活由浮点表示映射为低位宽整数表示，在不重新训练的前提下降低模型规模与推理开销。文献 [55] 提出 One-Shot 混合精度搜索模型，通过可学习的位宽概率预测机制，在单次超网训练后即可高效生成多组帕累托前沿候选配置，从而降低混合精度搜索开销；文献 [122] 结合边缘图引导的双精度子模型与零跳计算，实现任务自适应混合精度部署；Edge-MPQ[112]面向边缘部署场景，

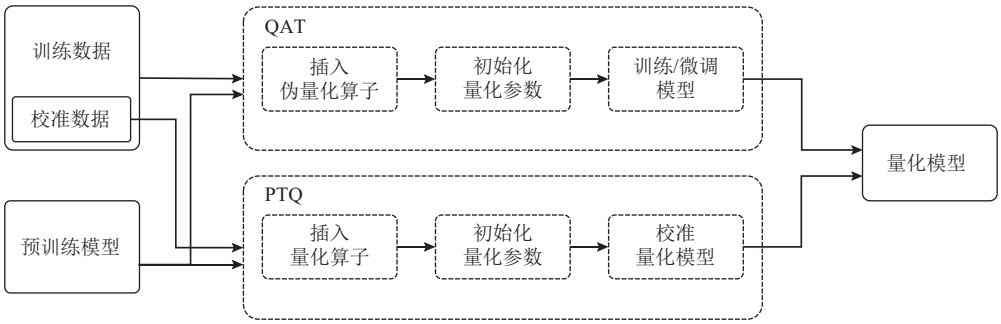



Fig. 16 Illustration of QAT and PTQ

图 16 QAT 与 PTQ 示意图

结合层敏感度度量与动态规划式搜索完成逐层混合精度分配, 并进一步引入训练必要性评估, 以判断在给定资源约束下是否需要启用 QAT 恢复精度; 文献 [123] 将量化器中的可学习尺度因子作为层级重要性指标, 并将混合精度搜索转化为一次性整数线性规划, 在 BitOps 或压缩率约束下实现高效位宽分配; 针对 CNN-Transformer 混合模型在边缘部署中的量化困难, HyQ[87] 提出了一种硬件友好的 PTQ 策略, 通过分布感知的缩放机制缓解不同模块间的量化失配问题, 并对 Transformer 中的 Softmax 运算进行线性近似, 从而降低计算复杂度并实现纯整数推理, 显著提升了混合模型在资源受限设备上的可部署性。

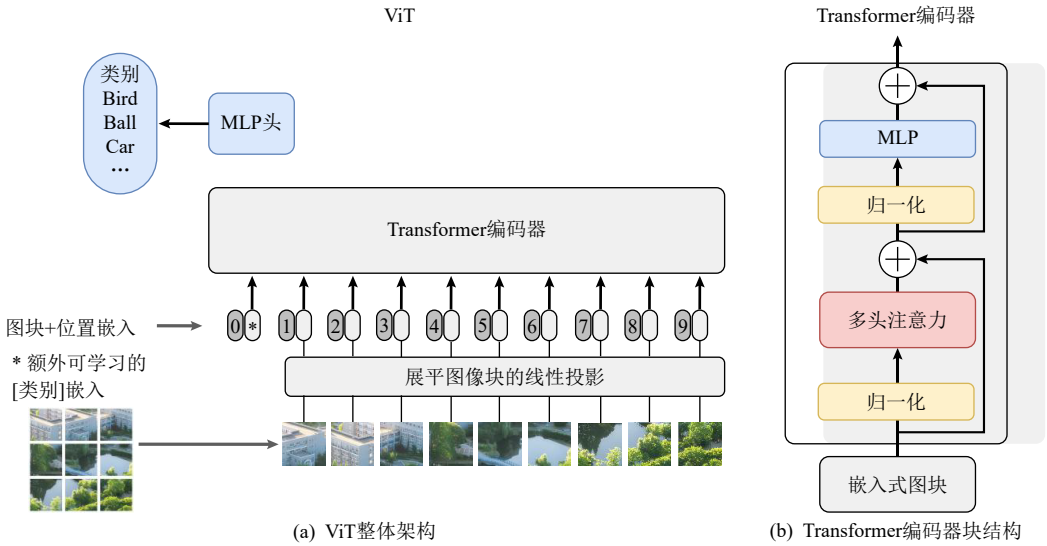
相较之下, QAT 在训练阶段显式引入量化过程, 使模型逐步适应低比特扰动, 因此在极低比特条件下通常更有利于精度恢复与稳定训练。LG-LSQ[100] 针对低比特训练中的梯度失真与离散化误差, 引入缩放模拟梯度、反正切软舍入和离散化误差最小化机制, 以改进线性对称量化下的训练稳定性; Sher 等人[99] 提出逐神经元量化方法, 并结合 STE 梯度转发, 以更细粒度的渐进式量化方式降低一次性量化带来的扰动; 文献 [111] 进一步引入 Bit-Controller 子网络, 根据样本复杂度动态调整位宽, 实现实例级自适应量化; 为缓解静态量化编码中固定码本带来的误差, 动态量化编码方法[97] 在训练过程中交替更新权重与码本, 使量化表示能够随权重分布变化自适应调整。

从整体上看, CNN 架构下的量化研究已从早期以 PTQ 为主的静态压缩手段, 逐步演进为以学习驱动方法为核心的优化范式。PTQ 路线通过智能化位宽分配与硬件感知校准实现高效部署, QAT 路线则通过梯度近似改进与渐进训练策略在极低比特条件下恢复精度, 二者共同突破了传统统一位宽假设, 为跨架构量化奠定了方法论基础。

### 5.2 ViT 架构

视觉 Transformer(vision Transformer, ViT)架构自提出以来, 凭借其全局建模能力以及在多类计算机视觉任务中的优异性能, 受到了广泛关注。但相较于 CNN, ViT 在量化时面临更突出的数值敏感性问题。其核心原因在于: 自注意力机制引入全局依赖关系, 局部量化误差可能在多层传播中被累积放大; LayerNorm 等算子对数值精度和分布稳定性要求较高, 在低比特设置下易引起表示偏移; ViT 中常见的长尾激活分布使少量极端值占据较大量化区间, 压缩了主体信息的有效表示范围。这些因素共同导致传统面向 CNN 设计的量化方法难以直接适用于 ViT, 如图 17 所示。围绕上述挑战, 学习驱动量化在 PTQ 与 QAT 两条路线下分别发展出专门的应对策略, 相关代表性方法如表 16 所示。

ViT 的 PTQ 研究面临 3 类主要难点: 传统 Hessian 敏感度近似在 ViT 结构中理论依据不充分; 激活的长尾分布与不对称特性使截断区间难以准确设定; 不同层与不同 patch 的量化敏感性差异显著, 统一位宽方案难以适配。针对敏感度建模缺陷, FIMA-Q[20] 重新审视了基于二阶信息的量化指导方法, 指出传统 Hessian 近似在 ViT 低比特量化中的理论缺陷, 并提出基于 Fisher 信息的高效敏感度度量, 以兼顾计算复杂度与建模精度。针对长尾激活与分布不对称, NoisyQuant[93] 在量化前向激活中加入固定均匀噪声偏置, 以主动调整分布形态, 并在层输出端通过去噪偏置进行补偿, 从而在较小额外开销下改善 PTQ 性能; AMP-ViT[89] 通过离线对称化对齐与 GELU 三区域自动缩放, 系统解决激活不对称与尺度失配, 无需任何额外推理开销。针对层间与 patch 间的异构敏感性, Mix-QViT[124] 引入可解释性驱动的层重要性分析, 将模型可解释性与量化敏感度统一到混合精度优化框架中, 并针对 LayerNorm 后激活的异常分布问题进行专门建模, 有效提升了超低比特量化下的稳定性与鲁棒性; PMQ[90] 提通过全局敏感度排序与自适应注意力收缩机制为不同 patch 分配差异化位宽;




(a) ViT整体架构 (b) Transformer编码器块结构

Fig. 17 Architecture of ViT

图 17 ViT 架构

**Table 16 Comparison of Key Challenges, Mechanisms, and Applicability of Representative ViT Quantization Methods**

**表 16 ViT 量化代表性方法的关键挑战、机制与适用性对比**

| 方法 | 类型 | 主要解决的 ViT 量化问题 | 核心方法论 | 优势 | 局限性 |
|---|---|---|---|---|---|
| Q-ViT[70] | QAT | 自注意力信息失真，蒸馏监督较粗 | 信息修正模块缓解量化后注意力表征失真；分布引导蒸馏细粒度注意力分布对齐 | 3~4 bit 精度保持好；理论支撑 | 额外模块开销 |
| FIMA-Q[20] | PTQ | Hessian 近似在 ViT 中失效 | Fisher 信息矩阵对角+低秩（DPLR）近似，纠正传统 Hessian 近似在 ViT 中的理论缺陷 | 理论严谨；标准量化器兼容性好 | 计算复杂度$O(a^k)$ |
| I-ViT[88] | QAT | Softmax 等非线性模块难以整数化 | 设计 Shiftmax、ShiftGELU 与整数 LayerNorm，使推理计算由整数与位移实现 | 硬件友好；适合整数推理部署 | 非线性近似误差累积 |
| NoisyQuant[93] | PTQ | 长尾激活导致量化不稳定 | 固定均匀噪声偏置主动改造激活分布，预计算去噪偏置消除推理影响 | 推理开销可忽略；长尾有效 | 噪声参数需调优 |
| AMP-ViT[89] | PTQ | 激活不对称，低比特参数难设定 | SymAlign 离线对称化对齐；AutoScale 后 GELU 三区域自动缩放 | 离线改造零推理开销；SQNR 指导 | 分区策略固定 |
| Mix-QViT[124] | PTQ | 层间敏感性差异大，异常值干扰明显 | LRP 层重要性评分与量化敏感度联合优化，CRL 裁剪通道重量化抑制异常值 | 可解释性较强；3~4 bit 稳定 | 层重要性计算开销 |
| StatsQ/CGA/QKR[92] | QAT | 量化训练振荡；Q/K 梯度耦合 | 统计权重量化（StatsQ）+置信度引导退火（CGA）+Q/K 重参数化（QKR）解耦梯度依赖 | 系统解决训练振荡问题 | 多机制组合复杂 |
| CLAMP-ViT[78] | PTQ | 无数据校准下样本语义不足 | Patch 级对比学习生成语义关联合成数据，层进化搜索逐层优化位宽与缩放因子 | 无数据混合精度 ViT 量化 | 两阶段迭代耗时 |
| PMQ[90] | PTQ | patch 与层间量化敏感性不均 | 基于全局敏感度排序与自适应注意力收缩机制，进行 patch 级混合精度分配 | 分配更细，量化更灵活 | 实现复杂度较高 |

CLAMP-ViT[78] 针对数据缺失场景以 patch 级对比学习生成语义关联的合成校准样本，结合逐层演化搜索联合确定各层量化参数，实现无数据混合精度量化。

对于 QAT 方向，Q-ViT[70] 指出，低比特 ViT 的主要瓶颈在于自注意力图的量化信息失真及蒸馏信号过于粗粒度。为此，该方法在前向过程中引入信息修正模块，在反向过程中采用分布引导蒸馏，以缓解全量化 ViT 中注意力分布偏移带来的性能退化。文献 [92] 则从训练稳定性角度分析 ViT 低比特 QAT 中的权重量化振荡问题，指出可学习缩放因子会加剧振荡，而自注意力中 Query 与 Key 的相互依赖又会进一步诱发共同振荡。为此，该方法分别引入统计权重量化、置信度引导退火和 Query-Key 重参数化，以缓解振荡并改善收敛稳定性。针对纯整数推理需求，I-ViT[88] 通过整数运算与位移操作实现全计算图的整数化，使整个推理计算图均可由整数运算与位移操作完成，并针对 Softmax、GELU 与 LayerNorm 设计了对应的整数近似方法，从而减少整数与浮点之间来回切换带来的额外开销，提升部署效率。文献 [91] 则通过容量感知的注意力保持与分组量化策略，显著改善了极低比特下 ViT 的性能退化问题。

总体来看，ViT 量化研究沿 PTQ 与 QAT 两条路线形成了互补的解决方案：PTQ 侧重通过精确的敏感度建模与分布改造降低校准误差，QAT 侧重通过稳定化训练机制与整数算子设计提升极低比特下的

训练与部署可靠性。学习驱动机制已成为 ViT 量化的重要发展方向，其核心作用不再局限于直接压缩参数，而是通过显式建模量化误差对表示空间、优化过程和部署路径的影响，提升低比特条件下的精度、稳定性与实用性。

### 5.3 LLM 架构

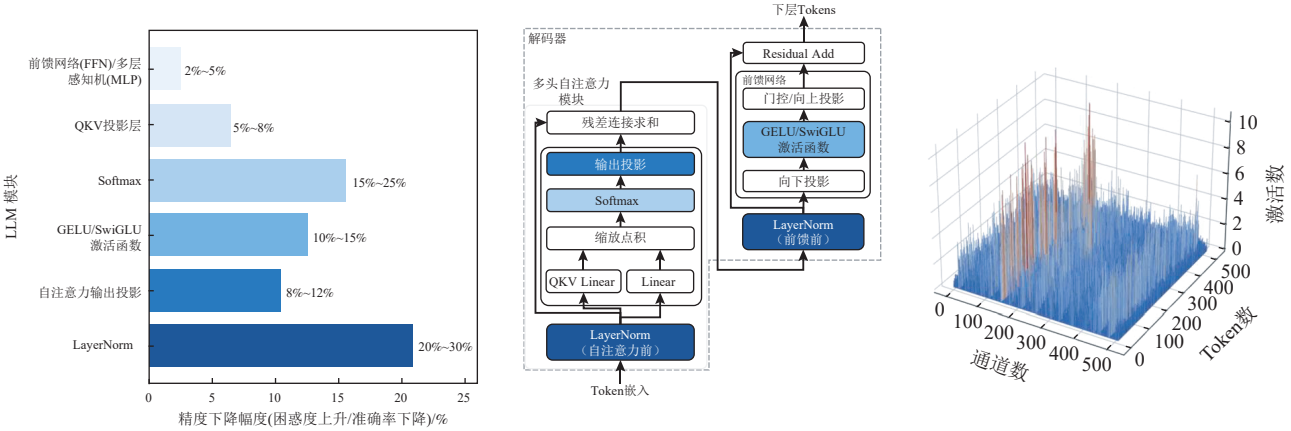
LLM 凭借大规模参数和较强的语义建模能力，已成为当前自然语言处理的重要基础模型。然而，与其性能优势相伴随的是在量化过程中所面临的显著挑战。相比于中小规模模型，LLM 由于参数规模巨大、激活分布呈现明显长尾特性，并高度依赖自注意力机制与归一化算子，其低比特量化问题表现得尤为复杂。

一方面，LLM 的模型参数量通常达到数十亿至数千亿规模，量化需要同时控制存储、计算和通信开销。另一方面，LLM 普遍存在明显的异常激活与长尾分布，如图 18 所示，自注意力、LayerNorm 及 KV 缓存等关键环节对数值扰动较为敏感，使得极低比特量化更容易出现性能退化。在云边端协同推理场景中，LLM 的边端侧量化部署还需应对长序列上下文在 KV 缓存上的显存压力，进一步加大了量化设计的难度。已有研究表明，LLM 量化的主要困难并不只来自权重量化本身，更与激活量化、异常值处理及不同结构模块对扰动的敏感性密切相关[26,126]。

近年来，基于学习机制的量化方法成为解决上述挑战的主要方向。研究者从 PTQ 和 QAT 两条路线提出了多种解决方案，通过引入近似二阶敏感度建模、激活平滑以及生成式校准等技术，这类方法能够在低比特量化条件下显著提升模型的鲁棒性与精度表现，同时有效降低计算与存储开销，为 LLM 的高效部署提供了可行路径。依据全文 4 信号源框架，LLM 量化方法同样可归入参数可学习化、重构与近似优化、辅助学习信号及约束驱动 4 类；由于 LLM 的异常激活与资源约束问题尤为突出，约束驱动与参数可学习化 2 类方法在此场景下得到了最为广泛的应用。

从 PTQ 角度来看，现有方法可沿 4 信号源框架分为 4 个方向。首先在参数可学习化方向，OmniQuant[24] 将可学习权重裁剪与可学习等价变换结合起来，在保持 PTQ 效率的同时提升了低比特性能；文献 [94] 则只调整 LayerNorm 参数，以较低代价修正量化后激活分布偏移；FlexRound[54] 通过逐元素除法实现可学习舍入机制，优化量化网格尺寸与权重缩放因子，允许大权重探索更广的离散值范围，在 2~4 bit 下有效保持精度。在重构与近似优化方向，针对自注意力机制和 LayerNorm 操作的量化问题，GPTQ[25] 通过引入近似二阶信息指导量化过程，结合 Hessian 矩阵进行误差补偿，并以固定列顺序量化、延迟批量更新和 Cholesky 分解预计算 3 项设计突破计算与内存瓶颈。SmoothQuant[26] 则针对激活异常值问题，通过离线等价变换将量化难度从激活迁移到权重，使权重和激活能够共同适配 INT8 计算。针对边缘部署下的显存与位宽预算约束，MXQ[96] 和 CMPQ[65] 分别从层级预算控制和通道级混合精度分配出发，更强调在给定显存或位宽预算下实现更灵活的内存与精度折中；ABQ-LLM[95] 进一步将任意位宽量化与推理内核协同设计结合起来，体现出面向实际硬件部署的工程导向。

从 QAT 角度，各方法同样可依据 4 信号源框架定位。在参数可学习化方向，LR-QAT[113] 结合低秩分




(a) 量化敏感度柱状图 (b) LLM Transformer Block热力示意图 (c) LLM 激活分布的长尾特性[125]

Fig. 18 Analysis of quantization-sensitive layers in LLM architecture and long-tail characteristics of activations

图 18 LLM 架构量化敏感层分析与激活长尾特性

解，通过引入低秩适配器和多尺度显存优化，显著缓解了 LLM 量化训练的显存开销。该方法冻结预训练权重，并在量化函数中融入可训练的低秩适配器，使训练完成后直接融合为整数权重矩阵，实现与 PTQ 相同的零推理开销。EfficientQAT[114] 采用两阶段解耦训练，首先通过块内全参数训练联合优化量化步长、零点与权重，扩展低比特量化的解空间。随后，在端到端量化参数微调阶段，冻结已量化权重，仅对步长进行微调，以极低内存成本捕捉跨块交互效应，解决了 LLM 量化训练中的内存和时间瓶颈。DL-QAT[116] 则将量化训练分解为组级幅度调整与量化空间内的低秩方向微调，在仅更新少量参数的条件下改善下游任务性能。在辅助学习信号与约束驱动协同方向，LLM-QAT[27] 采用无数据蒸馏方式生成训练样本，并同时量化权重、激活和 KV 缓存，说明 QAT 在长序列推理与缓存压缩中的潜力。

综上所述，LLM 量化需要同时处理参数规模、激活异常值、结构敏感性和系统资源约束。表 17 对相关方法进行了系统对比，概括了 LLM 量化在精度保持、资源消耗与部署灵活性之间的权衡关系。PTQ 方法强调低成本、快部署和预算可控，适合资源受限下的快速压缩与推理落地；QAT 方法则通过训练阶段的显式校正进一步提升低比特稳健性，更适合对精度要求较高，同时关注长上下文推理和缓存压缩的场景。未来面向 AI 算力网与云边端智能协同，LLM 量化还需要进一步与协同训练、协同推理、异构硬件预算约束和跨节点缓存管理等问题结合，形成从模型压缩到系统部署的一体化优化路径。

**Table 17 Comparison of Key Contributions of Representative LLM Quantization Methods**

**表 17 LLM 量化代表性方法关键贡献对比**

| 类型 | 框架定位 | 方法 | 核心方法论 | 优势 | 局限性 |
|---|---|---|---|---|---|
| PTQ | 重构与近似优化 | GPTQ[25] | 利用近似二阶信息进行逐行或逐块权重量化，并结合误差补偿降低累积误差 | 175B 模型 4 h 完成；3~4 bit 近无损 | 仅权重量化，激活仍保持高精度 |
| | | SmoothQuant[26] | 逐通道等价缩放，将激活量化难度离线迁移至权重 | 适合 8 bit 推理，硬件兼容性较好 | 平滑强度需要调节，低于 8 bit 需额外处理 |
| | 参数可学习化 | OmniQuant[24] | 结合可学习权重裁剪和可学习等价变换优化少量量化参数 | 2~4 bit 支持；部署友好、无额外算子 | 仍依赖少量校准数据，优化流程复杂 |
| | | FlexRound[54] | 基于元素级除法的可学习舍入，同时学习公共量化网格与权重个体缩放 | 对大权重更灵活，适合低比特权重量化 | 计算复杂度较高 |
| | | Norm Tweaking[94] | 仅微调 LayerNorm 参数，并以通道级分布约束对齐量化后激活分布 | 额外开销小，可作为现有 PTQ 插件 | 仅调整归一化层，能力边界较明确 |
| | 约束驱动（硬件资源约束） | MXQ[96] | 以层级量化配置和存储预算约束构造 MiLP，进行数据无关混合精度分配 | 可直接权衡内存与精度，适合边缘场景 | 配置求解流程较复杂 |
| | | CMPQ[65] | 基于激活分布的通道级混合精度分配，结合非均匀量化与异常值保护 | 支持任意位宽，更细粒度利用存储预算 | 策略设计和部署实现较复杂 |
| | | ABQ-LLM[95] | 分布校正 + bit balance + 任意位宽量化推理框架 | 任意位宽组合，兼顾精度与实际推理加速 | 依赖定制软件/内核支持，实现门槛高 |
| QAT | 辅助学习信号+数据缺失约束 | LLM-QAT[27] | 无数据蒸馏生成校准样本，同时量化权重、激活与 KV 缓存 | 不依赖原始训练数据，适合低比特与 KV 缓存压缩 | 仍需 QAT 训练成本，自生成数据存在分布偏移风险 |
| | 参数可学习化 | EfficientQAT[114] | 两阶段解耦训练：块级全参数优化 + 端到端步长微调 | 适合超大模型训练 | 两阶段流程复杂 |
| | | LR-QAT[113] | 低秩适配器嵌入量化函数内部，训练完成后融合为整数权重矩阵 | 零推理开销，显存效率高；LoRA 兼容 | 低秩近似仍可能带来表示受限 |
| | | DL-QAT[116] | 结合组级幅度调整和低秩方向微调 | 可减少训练参数量 | 依赖低秩假设，秩和分组设置影响效果 |

### 5.4 学习驱动量化的跨架构讨论

本节系统回顾了学习驱动量化方法在 CNN、ViT 与 LLM 三类主流架构中的研究进展。由于不同架构在核心算子、参数规模、激活分布及关键敏感模块等方面存在明显差异，量化方法通常不能直接平移，而需要结合具体结构特性进行适配。表 18 从核心计算单元、敏感层分布、激活特征及典型方法等维度，对 3 类主流架构的量化特性进行了对比。总体来看，尽管 CNN、ViT 与 LLM 在结构设计和计算范式上差异显著，但在低比特量化条件下仍面临若干共性难点，包括长尾分布或异常值压缩了有效量化分辨率，注意力、归一化等关键算子对量化扰动更为敏感，以及当位宽降至 2~4 bit 时传统 PTQ 方法往往出现明显性能退化。对于 LLM，进一步研究还表明[126]，其量化误差对不同模型家族、模型规模及数值区间并不一致，说明大模型量化问题具有更强的架构依赖性与数值敏感性。

为应对上述挑战，学习驱动量化方法逐步形成 4 条代表性技术路线。参数自适应方向[24-25,94,100] 以可学习量化参数替代启发式规则，实现对复杂分布的精

**Table 18 Comparison of Quantization Characteristics across CNN, ViT, and LLM Architectures**

**表 18 CNN/ViT/LLM 架构量化特性对比**

| 特性维度 | CNN | ViT | LLM |
|---|---|---|---|
| 核心计算单元 | 卷积层（Conv） | 多头自注意力（MHSA）+FFN | 多头自注意力（MHSA）+FFN |
| 参数规模 | 10~100×10$^6$ | 10×10$^6$~1×10$^9$ | 1×10$^9$~1 000×10$^9$ |
| 主要敏感层 | 首层/尾层 | LayerNorm 后/Attention $\boldsymbol{Q}$，$\boldsymbol{K}$ | Attention、LayerNorm 及异常值主导通道 |
| 激活分布特点 | 近似高斯/ReLU 截断 | 后 GELU 长尾、通道间变异明显 | 强长尾与持续异常值并存 |
| 量化核心挑战 | 层级敏感度差异；深度可分离卷积 | Query/Key 分布失配；Softmax 与归一化数值敏感 | 小值更敏感、大值更鲁棒，KV 缓存带来长序列显存压力 |
| 权重量化难度 | 中等 | 中等 | 较低（分布相对均匀） |
| 激活量化难度 | 中、低等 | 高（后 LayerNorm） | 极高（异常值主导） |
| 典型 PTQ 方法 | BRECQ[9]，QDrop[13] | FIMA-Q[20]，NoisyQuant[93] | GPTQ[25]，SmoothQuant[26] |
| 典型 QAT 方法 | LSQ[11]，DSQ[19] | Q-ViT[70]，I-ViT[88] | LLM-QAT[27]，EfficientQAT[114] |
| 推荐安全位宽 | W4A4 | W4A4/W4A8 | W4A16/W3A16 |
| 硬件部署重点 | Tensor Core INT8/INT4 | 整数位移近似非线性 | 显存优化、KV 缓存压缩与长序列推理 |

细适配。难度迁移方向[26,89,126]通过重分配权重与激活的数值尺度，将量化难度由激活侧迁移至权重侧，从整体上稳定量化过程。混合精度与硬件协同方向[65,90,95,112,124]基于敏感度分析在不同层级分配差异化位宽，并将算法设计与硬件执行约束联合优化。训练范式创新方向[27,78,113-114]则通过低秩适配与流程解耦降低显存开销，并探索无数据量化范式以拓展方法在数据受限场景下的适用范围。表 19 进一步归纳了典型量化技术路线在 CNN、ViT 与 LLM 架构中的适用程度及其跨架构迁移障碍。

总体来看，现有方法已在 4~8 bit 内实现了较好的模型压缩效果。然而，学习驱动量化技术要进一步走向大规模、通用化应用，仍有赖于在理论建模、算法设计与系统实现等多个层面实现更为根本性的突破：理论层面，现有方法对长尾分布与异常值的处理仍以经验策略为主，未来需构建分布特性与数值敏感度的统一理论框架，建立长尾分布、异常值与量化误差之间的可解析刻画，为位宽分配与量化策略提供可验证的数学依据；算法层面，极端低比特下均匀量化的表达受限问题尚未从根本上突破，非均匀量化机制与硬件原语的协同设计是重要突破方向；系统层面，算法创新与硬件效能之间的转化鸿沟依然显著，亟需推进算法与体系结构协同演化，将分布感知与动态位宽决策机制嵌入新型硬件架构，实现端到端原生低精度推理。

## 6 挑战与未来研究方向

学习驱动的模型量化通过引入自适应优化机制，打破了传统静态规则的局限，在极低比特场景下实现了精度与效率的更优权衡。然而，从经验规则向数据驱动优化的范式转变，导致底层优化机理、系统部署架构以及大规模模型适配等方面面临新的挑战。

**Table 19 Applicability Matrix of Cross-Architecture Quantization Methods**

**表 19 跨架构量化方法适用性矩阵**

| 方法类别 | 代表方法 | CNN | ViT | LLM | 核心迁移障碍 |
|---|---|---|---|---|---|
| 可学习步长 | LSQ[11] | ✔ | ○ | ○ | ViT/LLM 长尾激活与异常值需额外处理 |
| 块级重构 | BRECQ[9] | ✔ | ✔ | ○ | LLM 块划分、跨层依赖与显存开销需重新设计 |
| 激活平滑 | SmoothQuant[26] | ○ | ○ | ✔ | CNN 无持续异常值，ViT 还需处理非对称激活 |
| 对比校准 | CL-Calib[50] | ✔ | ✔ | ○ | LLM 序列长、层间状态复杂，对比校准代价高 |
| 二阶敏感度 | HAWQ[15] | ✔ | ○ | ✗ | LLM 二阶信息估计与存储开销过高 |
| 无数据生成 | ZeroQ[51] | ✔ | ○ | ✗ | ViT/LLM 缺乏可直接利用的 BN 统计 |
| 硬件感知搜索 | HAQ[22] | ✔ | ✔ | ○ | LLM 规模大、位宽组合多，搜索空间迅速膨胀 |

注：✔ 完全适用，○ 需适配修改，✗ 不适用或效果差。

### 6.1 基于误差解析的量化参数与学习信号优化

在学习驱动量化中，可学习参数的引入与多源辅助信号的叠加，使量化过程面临复杂的非凸优化问题与梯度退化挑战。通过刻画量化对模型精度影响的逐层传播机理，构建多学习参数协同优化以及多学习信号下的高效量化方案，是未来重要的研究方向。

1）量化误差传播的可解析建模。量化误差在网络中的跨层传播是制约量化模型精度上限的根本原因。多粒度重构方法对比表明，块级优化比逐层优

化更有利于控制误差累积，而全网输出对齐则进一步证明局部重构误差与最终任务损失之间存在非线性传播关系。然而，目前没有方法能够在给定单层量化扰动的条件下，解析扰动对下游层乃至最终输出的影响程度。因此，现有工作中混合精度位宽分配通常依赖经验性的 Hessian 对角近似或端到端搜索，而无法从理论上给出各层位宽的最优分配依据。未来研究可以从量化误差传播建模的角度出发，利用网络的局部线性近似建立逐层误差传播的递推模型，将每层量化扰动的期望幅值沿网络深度方向进行解析估计，为混合精度位宽搜索提供不依赖完整前向传播的轻量敏感度评分依据。

2）可学习参数的协同优化。传统量化中的步长、截断阈值等参数被松弛为连续变量后，它们在量化函数中表现出高度的非线性耦合。这些参数共同界定了量化映射的离散空间，但现有方法通常将这些量化参数作为相互独立的变量分别优化，往往会因忽略参数间的内在制约关系而陷入局部次优解，并使局部参数更新放大跨层误差。如何在量化参数之间建立显式约束，使它们在参数空间中协同收敛，是参数可学习化方向目前尚未系统解决的核心问题。以步长与舍入阈值的联合优化为例，可根据当前步长引起的层级误差变化，自适应地调整舍入阈值的优化目标，从梯度流的角度建立 2 类参数的显式耦合，而非依赖端到端训练隐式地寻找协调点。

3）多辅助信号协同训练中的梯度协调。准确的梯度信号是模型稳定优化的前提。然而，量化算子不可导带来的梯度近似偏差，会在深层网络中产生非线性的链式累积与放大，导致训练过程剧烈震荡且难以收敛。此外，为弥补极低比特信息损失而叠加的多源辅助信号，在梯度空间中常发生方向冲突与抵消，例如在极低比特 QAT 中同时引入任务损失、知识蒸馏与重构约束时，三者在量化器参数子空间上的梯度方向会在特定训练阶段发生反向冲突，导致参数更新震荡；而以单一蒸馏目标配合梯度动态缩放的方案，在跨架构场景中均表现出比多损失加权更稳定的收敛行为。由此可见，多损失冲突是低比特训练不稳定的重要来源，未来研究可以从梯度冲突实时感知的角度出发，以各损失分量在量化器参数子空间上的局部梯度余弦相似度为监控信号，当相似度低于某阈值时适当降低冲突分量的权重，从而在不显著增加计算负担的前提下动态调整各分量权重，解决多信号协同训练中静态权重配置引发的优化失稳困境。

### 6.2 学习驱动量化中精度与资源的权衡

模型量化需要在位宽、计算开销与模型精度之间进行权衡。降低位宽可以减少存储与计算开销，但也会引入量化误差并造成精度损失。学习驱动方法通过自适应优化在一定程度上解决这一权衡问题，但也带来了数据依赖、训练成本和部署效率等新的问题。

1）量化校准对外部数据的依赖。无数据量化方法通常利用批归一化统计量匹配或生成对抗网络生成的代理样本作为校准依据，在不访问真实训练数据的条件下优化量化参数。然而，基于统计匹配的方法在深层网络中收敛困难，基于生成模型的方法则通常需要额外训练生成器，并可能受到生成质量和优化稳定性的影响。事实上，预训练模型的权重本身已通过大规模训练将数据分布的结构信息编码于参数之中，批归一化的均值与方差直接记录了各层激活的一阶和二阶统计特征，这些信息无需任何真实数据即可获取，但现有方法仅将其用于约束外部生成器，而非直接驱动校准过程。未来研究可以从模型内部先验挖掘的角度出发，将批归一化统计量、权重分布、激活范围估计等内部分布信息从生成器的辅助约束提升为直接驱动校准过程的主信号，直接优化量化参数，从而减少乃至消除对外部生成器的依赖，为隐私敏感与数据受限场景提供更轻量高效的无数据量化方案。

2）QAT 计算资源分配效率。层级敏感度差异是 PTQ 中被广泛证实并支撑混合精度量化的重要依据。然而现有 QAT 方法大多对所有层均投入相同规模的迭代预算，高敏感层因训练资源被均摊而往往未能充分收敛，低敏感层则造成大量计算浪费。未来研究可以从敏感度感知训练调度的角度出发，将 PTQ 中已成熟的层级敏感度分析机制引入 QAT 的迭代预算分配，在微调初始阶段通过轻量前向扰动或局部重构误差估计各层敏感度，并据此动态调整训练预算、更新频率或辅助损失权重。对于高敏感层，可分配更多迭代次数或更强的蒸馏、重构约束；对于低敏感层，则可减少更新频率或提前冻结，从而使 QAT 的训练资源分配与层级量化敏感性相匹配，在保持精度的同时降低训练成本。

3）面向混合精度量化的软硬件协同优化。学习驱动的混合精度量化方法在理论层面可实现精度与压缩比间权衡，但实际部署性能往往与理论结果存在一定差距。其根本原因在于，当前推理硬件普遍针对固定位宽的规则运算进行优化，非均匀的混合

精度配置带来的不规则内存访问与算子切换开销,会抵消一部分理论加速收益。因此,未来研究应将硬件可执行性纳入学习驱动搜索过程,从算法与硬件约束联合建模的角度出发,在搜索空间构建阶段约束候选位宽组合,使其符合目标硬件支持的算子粒度、内存对齐方式和整数计算路径;也可将实际延迟、能耗或内核调用开销作为反馈信号,引导量化策略选择更易部署的位宽配置。通过建立量化算法与底层推理内核之间的显式映射,有助于推动混合精度量化从理论压缩到系统加速的有效转化。

### 6.3 跨模型架构的学习驱动量化统一框架

随着模型参数规模的飞速增长,LLM 与大规模 ViT 的量化已成为学习驱动量化领域重要的研究方向。跨架构量化的相关研究结果表明,CNN、ViT 和 LLM 在量化理论与方法上既存在共性规律,也表现出架构特异性的差异。因此,未来的研究方向包括层级敏感度的跨架构分析、激活长尾分布的架构感知处理,为跨架构的统一可学习量化框架以及面向 LLM 的低成本、高压缩比、高精度保持的量化奠定基础。

1)层级敏感度的跨架构分析。主流的神经网络架构均呈现出明显的层级敏感度不均,但高敏感层的分布位置随架构不同而存在系统性差异:对于 CNN,首尾层以及部分承担特征分辨率变化的卷积层往往对量化扰动更敏感;对于 ViT,LayerNorm 后的激活路径、注意力模块和 MLP 模块中的特征变换更容易影响后续表示;对于 LLM,注意力投影通道和前馈网络中特定通道是量化误差更容易被放大的结构。这一共性规律说明,敏感度感知的量化策略具有跨架构的普适性,但识别高敏感区域所依赖的分析工具需要针对各架构的核心计算单元定制。未来研究可以从跨架构敏感度统一建模的角度出发,为三类架构设计统一的轻量化敏感度评估方法。

2)激活长尾分布的架构感知处理。长尾激活分布在 3 类架构中均有出现,但其成因与表现形式具有一定差异:CNN 中激活的长尾主要来自卷积层输出在通道间的统计差异;ViT 中的长尾与 GELU 激活函数的非对称性密切相关,表现为块内的位置相关分布偏移;LLM 中则出现了持续性的极端通道离群值,即少数特定通道在不同输入样本下均表现出远超其他通道的激活幅值。对这 3 类成因的精确区分直接影响激活量化策略的选择,而适用于 LLM 的通道级等价缩放对 ViT 的 GELU 诱导偏移是否同样有效,当前尚无系统性验证与理论分析。未来研究可以从激活异常值成因分类的角度出发,针对 GELU 诱导的位置相关偏移设计对应的等价变换形式,明确通道级缩放策略的适用边界,为大模型量化中的激活处理提供架构感知的设计准则。

## 7 研究展望

随着人工智能特别是大模型技术的飞速演进,如何在资源受限的泛在计算环境中高效部署庞大的深度神经网络,已成为当前学术界与工业界面临的关键挑战。学习驱动的模型量化技术凭借其自适应优化的特性,有效突破了传统静态规则在极低比特场景下的精度瓶颈,正成为模型压缩领域的核心研究热点。本文以学习驱动的模型量化方法为主线,系统综述了该领域的研究进展与前沿动态。围绕 PTQ 与 QAT 两大范式,本文创新性地构建了基于 4 类学习信号源的统一分类框架,深度剖析了参数可学习化、重构与近似优化、辅助信号监督及约束驱动方法等核心策略。此外,本文横向对比了上述机制在 CNN、ViT 与 LLM 等主流架构中的应用特性与适配规律。研究表明,学习驱动量化不仅能显著缓解低比特条件下的精度退化与训练失稳问题,也为异构硬件上的模型高效部署提供了坚实的方法论指导。

尽管学习驱动量化展现出显著优势,但在实际落地中仍面临诸多具体的技术瓶颈:多参数与多信号的联合优化难以稳定控制、无数据校准与训练的算力代价过高,以及极低比特算法在现有硬件上的实际加速效果往往不及预期。展望未来,该领域亟待在以下 3 个切实方向寻求突破:一是理清量化误差的传播机理与参数间的耦合关系,设计更稳定的梯度调控机制,以解决极低比特下的训练震荡问题;二是降低量化过程对数据和算力的依赖,尝试挖掘模型内部先验信息来减少数据合成开销,并基于算子敏感度对网络不同层进行差异化优化,避免盲目的全局联合训练;三是提升大模型量化的硬件执行效率,在处理长尾异常值的同时探索量化友好的网络架构设计,推动算法与底层计算原语的深度融合。解决这些现实问题,将帮助模型量化技术大幅降低经验调参的试错成本,真正实现大规模神经网络在边端设备上的低延迟、低能耗部署。

**作者贡献声明**:黄馨怡主要完成文献调研、论文撰写及修改;屈志昊主要负责确定论文思路与整体

框架，以及整体论文修改；叶保留对论文选题、组织结构提供指导意见，并完成论文审阅工作；胡世红提供论文修改意见并参与研究展望章节修改；贾柠晖参与论文引言、背景知识、方法总览与分类等章节修改；钱琳提供指导与修改意见。

## 参 考 文 献


[1] Han Song, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C]//Proc of the Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2015: 1135−1143

[2] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint, arXiv: 1503.02531, 2015

[3] Denton E, Zaremba W, Bruna J, et al. Exploiting linear structure within convolutional networks for efficient evaluation[C]//Proc of the 27th Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2014: 1269−1277

[4] Han Song, Mao Huizi, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding[C]//Proc of the 4th Int Conf on Learning Representations. Washington DC: ICLR, 2016: 136−150

[5] Jacob B, Kligys S, Chen Bo, 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. Piscataway, NJ: IEEE, 2018: 2704−2713

[6] Krishnamoorthi R. Quantizing deep convolutional networks for efficient inference: A whitepaper[J]. arXiv preprint, arXiv: 1806.08342, 2018

[7] Banner R, Nahshan Y, Soudry D. Post training 4-bit quantization of convolutional networks for rapid-deployment[C]//Proc of the 33rd Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2019: 7950−7958

[8] Choukroun Y, Kravchik E, Yang F, et al. Low-bit quantization of neural networks for efficient inference[C]//Proc of the IEEE/CVF Int Conf on Computer Vision Workshops. Piscataway, NJ: IEEE, 2019: 3009−3018

[9] Li Yuhang, Gong Ruihao, Tan Xu, et al. BRECQ: Pushing the limit of post-training quantization by block reconstruction[C]//Proc of the 9th Int Conf on Learning Representations. Washington DC: ICLR, 2021: 18472−18488

[10] Nagel M, Amjad R A, van Baalen M, et al. Up or down? Adaptive rounding for post-training quantization[C]//Proc of the 37th Int Conf on Machine Learning. New York: ICML, 2020: 7197−7206

[11] Esser S K, McKinstry J L, Bablani D, et al. Learned step size quantization[C]//Proc of the 8th Int Conf on Learning Representations. Washington DC: ICLR, 2020: 6910−6921

[12] 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. Piscataway, NJ: IEEE, 2020: 2978−2985

[13] Wei Xiuying, Gong Ruihao, Li Yuhang, et al. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization[C]//Proc of the 10th Int Conf on Learning Representations. Washington DC: ICLR, 2022: 2831−2849

[14] Liu Jiawei, Niu Lin, Yuan Zhihang, et al. PD-Quant: Post-training quantization based on prediction difference metric[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 24427−24437

[15] Dong Zhen, Yao Zhewei, Gholami A, et al. HAWQ: Hessian aware quantization of neural networks with mixed-precision[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 293−302

[16] Jung S, Son C, Lee S, et al. Learning to quantize deep networks by optimizing quantization intervals with task loss[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4350−4359

[17] Gong Ruihao, Liu Xianglong, Li Yuhang, et al. Pushing the limit of post-training quantization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(7): 5556−5570

[18] Zhang Dongqing, Yang Jiaolong, Ye Dongqiangzi, et al. LQ-Nets: Learned quantization for highly accurate and compact deep neural networks[C]//LNCS 11212: Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 373−390

[19] Gong Ruihao, Liu Xianglong, Jiang Shenghu, et al. Differentiable soft quantization: Bridging full-precision and low-bit neural networks[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4851−4860

[20] Wu Zhuguanyu, Wang Shihe, Zhang Jiayi, et al. FIMA-Q: Post-training quantization for vision Transformers by Fisher information matrix approximation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2025: 14891−14900

[21] Wang Yuejiao, Ma Zhong, Yang Chaojie, et al. Reinforcement learning based quantization strategy optimal assignment algorithm for mixed precision[J]. Computers, Materials & Continua, 2024, 79(1): 819−836

[22] Wang Kuan, Liu Zhijian, Lin Yujun, et al. HAQ: Hardware-aware automated quantization with mixed precision[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8612−8620

[23] Chen Shangyu, Wang Wenya, Pan S J. Deep neural network quantization via layer-wise optimization using limited training data[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3329−3336

[24] Shao Wenqi, Chen Mengzhao, Zhang Zhaoyang, et al. OmniQuant: Omnidirectionally calibrated quantization for large language models[C]//Proc of the 12th Int Conf on Learning Representations. Washington DC: ICLR, 2024: 9680−9704

[25] Frantar E, Ashkboos S, Hoefler T, et al. GPTQ: Accurate post-training quantization for generative pre-trained transformers[J].

arXiv preprint, arXiv: 2210.17323, 2022

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

[27] Liu Zechun, Oguz B, Zhao Changsheng, et al. LLM-QAT: Data-free quantization aware training for large language models[C]//Findings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 467−484

[28] Rokh B, Azarpeyvand A, Khanteymoori A. A comprehensive survey on model quantization for deep neural networks in image classification[J]. ACM Transactions on Intelligent Systems and Technology, 2023, 14(6): 1−50

[29] Gholami A, Kim S, Dong Zhen, et al. A survey of quantization methods for efficient neural network inference[J]. arXiv preprint, arXiv: 2103.13630, 2021

[30] Qin Haotong, Gong Ruihao, Liu Xianglong, et al. Binary neural networks: A survey[J]. Pattern Recognition, 2020, 105: 107281

[31] Liu Kai, Zheng Qian, Tao Kaiwen, et al. Low-bit model quantization for deep neural networks: A survey[J]. arXiv preprint, arXiv: 2505.05530, 2025

[32] Xu Xiaohua, Zhou Changbing, Hu Zhongxu, et al. A survey of lightweight deep neural network model adaptation for edge intelligence[J]. Computer Science, 2024, 51(7): 257−271 (in Chinese)
(徐小华, 周长兵, 胡忠旭, 等. 轻量级深度神经网络模型适配边缘智能研究综述[J]. 计算机科学, 2024, 51(7): 257−271)

[33] Wang Wenxiao, Chen Wei, Luo Yicong, et al. Model compression and efficient inference for large language models: A survey[J]. arXiv preprint, arXiv: 2402.09748, 2024

[34] Liang Feng, Zhang Zhen, Lu Haifeng, et al. Communication-efficient large-scale distributed deep learning: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2021, 14(8): 1

[35] Ji Rongrong, Lin Shaohui, Chao Fei, et al. A survey of deep neural network compression and acceleration[J]. Journal of Computer Research and Development, 2018, 55(9): 1871−1888 (in Chinese)
(纪荣嵘, 林绍辉, 晁飞, 等. 深度神经网络压缩与加速综述[J]. 计算机研究与发展, 2018, 55(9): 1871−1888)

[36] Gao Yang, Cao Yangjie, Duan Pengsong. A survey of lightweight methods for neural network models[J]. Computer Science, 2024, 51(S1): 23−33 (in Chinese)
(高杨, 曹仰杰, 段鹏松. 神经网络模型轻量化方法综述[J]. 计算机科学, 2024, 51(S1): 23−33)

[37] Lei Jie, Gao Xin, Song Jie, et al. A survey of deep network model compression[J]. Journal of Software, 2017, 29(2): 251−266 (in Chinese)
(雷杰, 高鑫, 宋杰, 等. 深度网络模型压缩综述[J]. 软件学报, 2017, 29(2): 251−266)

[38] Gao Han, Tian Yulong, Xu Fengyuan, et al. A survey of deep learning model compression and acceleration[J]. Journal of Software, 2020, 32(1): 68−92 (in Chinese)
(高晗, 田育龙, 许封元, 等. 深度学习模型压缩与加速综述[J]. 软件学报, 2020, 32(1): 68−92)

[39] Zhou Zhifei, Li Hua, Feng Yixiong, et al. Research progress on lightweight deep convolutional neural network design[J]. Computer Engineering and Applications, 2024, 60(22): 1−17 (in Chinese)
(周志飞, 李华, 冯毅雄, 等. 轻量化深度卷积神经网络设计研究进展[J]. 计算机工程与应用, 2024, 60(22): 1−17)

[40] Wang Zeqin, Wen Ming, Xu Yuedong, et al. Communication compression techniques in distributed deep learning: A survey[J]. Journal of Systems Architecture, 2023, 142: 102927

[41] Xu Hang, Ho Chen-Yu, Abdelmoniem A M, et al. Compressed communication for distributed deep learning: Survey and quantitative evaluation[J]. arXiv preprint, arXiv: 2003.03092, 2020

[42] Guo Yunhui. A survey on methods and theories of quantized neural networks[J]. arXiv preprint, arXiv: 1808.04752, 2018

[43] Choi J, Chuang P I J, Wang Z, et al. Bridging the accuracy gap for 2-bit quantized neural networks (QNN)[J]. arXiv preprint, arXiv: 1807.06964, 2018

[44] Choi J, Wang Zhuo, Venkataramani S, et al. PACT: Parameterized clipping activation for quantized neural networks[J]. arXiv preprint, arXiv: 1805.06085, 2018

[45] Wang Longguang, Dong Xiaoyu, Wang Yingqian, et al. Learnable lookup table for neural network quantization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 12413−12423

[46] Li Yuhang, Dong Xin, Wang Wei. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks[C]//Proc of the 8th Int Conf on Learning Representations. Washington DC: ICLR, 2020: 9512−9526

[47] Wu Junru, Wang Yue, Wu Zhenyu, et al. Deep k-means: Re-training and parameter sharing with harder cluster assignments for compressing deep convolutions[C]//Proc of the 35th Int Conf on Machine Learning. New York: ICML, 2018: 5363−5372

[48] Miyashita D, Lee E H, Murmann B. Convolutional neural networks using logarithmic data representation[J]. arXiv preprint, arXiv: 1603.01025, 2016

[49] Fang Jun, Shafiee A, Abdel-Aziz H, et al. Post-training piecewise linear quantization for deep neural networks[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2020: 69−86

[50] Shang Yuzhang, Liu Gaowen, Kompella R R, et al. Enhancing post-training quantization calibration through contrastive learning[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15921−15930

[51] Cai Yaohui, Yao Zhewei, Dong Zhen, et al. ZeroQ: A novel zero shot quantization framework[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 13166−13175

[52] Fan Chunxiao, Wang Ziqi, Guo Dan, et al. Data-free quantization via pseudo-label filtering[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 5589−5598

[53] Hubara I, Nahshan Y, Hanani Y, et al. Accurate post training quantization with small calibration sets[C]//Proc of the 38th Int Conf on Machine Learning. New York: ICML, 2021: 4466−4475

[54] Lee J H, Kim J, Kwon S J, et al. FlexRound: Learnable rounding based on element-wise division for post-training quantization[C]//Proc of the 40th Int Conf on Machine Learning. New York: ICML, 2023: 18913−18939

[55] Koryakovskiy I, Yakovleva A, Buchnev V, et al. One-shot model for mixed-precision quantization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7939−7949

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

[57] Hubara I, Courbariaux M, Soudry D, et al. Quantized neural networks: Training neural networks with low precision weights and activations[J]. Journal of Machine Learning Research, 2017, 18(187): 1−30

[58] Zhao Kaiqi, Zhao Ming. Self-supervised quantization-aware knowledge distillation[C]//Proc of the Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2024: 4375−4383

[59] Chen Shangyu, Wang Wenya, Pan S J. MetaQuant: Learning to quantize by learning to penetrate non-differentiable quantization[C]//Proc of the 33rd Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2019: 3918−3928

[60] Huh M, Cheung B, Agrawal P, et al. Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks[C]//Proc of the 40th Int Conf on Machine Learning. New York: ICML, 2023: 14096−14113

[61] Liu Yuang, Zhang Wei, Wang Jun. Zero-shot adversarial quantization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1512−1521

[62] Qian Biao, Wang Yang, Hong Richang, et al. Adaptive data-free quantization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7960−7968

[63] Qian Biao, Wang Yang, Hong Richang, et al. Rethinking data-free quantization as a zero-sum game[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 9489−9497

[64] Kim M, Choi J, Lee J, et al. Zero-shot quantization: A comprehensive survey[C]//Proc of the 34th Int Joint Conf on Artificial Intelligence. Freiburg: IJCAI, 2025: 10508−10516

[65] Chen Zihan, Xie Bike, Li Jundong, et al. Channel-wise mixed-precision quantization for large language models[J]. arXiv preprint, arXiv: 2410.13056, 2024

[66] Zhu Chenzhuo, Han Song, Mao Huizi, et al. Trained ternary quantization[C]//Proc of the 5th Int Conf on Learning Representations. Washington DC: ICLR, 2017: 2106−2115

[67] Hubara I, Courbariaux M, Soudry D, et al. Binarized neural networks[C]//Proc of the 30th Annual Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2016: 4107−4115

[68] Zhou Aojun, Yao Anbang, Guo Yiwen, et al. Incremental network quantization: Towards lossless CNNs with low-precision weights[C]//Proc of the 5th Int Conf on Learning Representations. Washington DC: ICLR, 2017: 2160−2173

[69] Rastegari M, Ordonez V, Redmon J, et al. XNOR-Net: ImageNet classification using binary convolutional neural networks[C]//Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 525−542

[70] Li Yanjing, Xu Sheng, Zhang Baochang, et al. Q-ViT: Accurate and fully quantized low-bit vision transformer[C]//Proc of the 36th Annual Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2022: 34451−34463

[71] Nahshan Y, Chmiel B, Baskin C, et al. Loss aware post-training quantization[J]. Machine Learning, 2021, 110(11): 3245−3262

[72] Diao Huabin, Li Gongyan, Xu Shaoyun, et al. Attention round for post-training quantization[J]. Neurocomputing, 2024, 565: 127012

[73] Tang Chen, Meng Yuan, Jiang Jiacheng, et al. Retraining-free model quantization via one-shot weight-coupling learning[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15855−15865

[74] Wang Changbao, Zheng Dandan, Liu Yuanliu, et al. Leveraging inter-layer dependency for post-training quantization[C]//Proc of the 36th Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2022: 6666−6679

[75] Idelbayev Y, Molchanov P, Shen M, et al. Optimal quantization using scaled codebook[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 12095−12104

[76] Jeon Y, Lee C, Cho E, et al. Mr.BiQ: Post-training non-uniform quantization based on minimizing the reconstruction error[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 12319−12328

[77] Ding Yifu, Feng Weilun, Chen Chuyan, et al. Reg-PTQ: Regression-specialized post-training quantization for fully quantized object detector[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 16174−16184

[78] Ramachandran A, Kundu S, Krishna T. CLAMP-ViT: Contrastive data-free learning for adaptive post-training quantization of ViTs[C]//Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2024: 307−325

[79] Zhou Sifan, Li Liang, Zhang Xinyu, et al. LiDAR-PTQ: Post-training quantization for point cloud 3D object detection[C]//Proc of the 12th Int Conf on Learning Representations. Washington DC: ICLR, 2024: 14922−14936

[80] Tu Zhijun, Hu Jie, Chen Hanting, et al. Toward accurate post-training quantization for image super resolution[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 5856−5865

[81] Zhong Yunshan, Lin Mingbao, Chen Mengzhao, et al. Fine-grained data distribution alignment for post-training quantization[G]//LNCS 13671: Proc of the Eur Conf on Computer Vision. Berlin: Springer,

2022: 70−86

[82] Zhong Yunshan, Lin Mingbao, Nan Gongrui, et al. IntraQ: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 12329−12338

[83] Zhao Kai, Zhuang Zhihao, Zhang Miao, et al. Enhancing diversity for data-free quantization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2025: 20969−20978

[84] Choi Y, El-Khamy M, Lee J. Zero-shot learning of a conditional generative adversarial network for data-free network quantization[C]//Proc of the 28th IEEE Int Conf on Image Processing. Piscataway, NJ: IEEE, 2021: 3552−3556

[85] Shang Yuzhang, Xu Bingxin, Liu Gaowen, et al. Causal-DFQ: Causality guided data-free network quantization[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 17391−17400

[86] Zhang Wenlun, Ando S, Yoshioka K. AHCPTQ: Accurate and hardware-compatible post-training quantization for segment anything model[J]. arXiv preprint, arXiv: 2503.03088, 2025

[87] Kim N J, Lee J, Kim H. HyQ: Hardware-friendly post-training quantization for CNN-Transformer hybrid networks[C]//Proc of the 33rd Int Joint Conf on Artificial Intelligence. Freiburg: IJCAI, 2024: 4291−4299

[88] Li Zhikai, Gu Qingyi. I-ViT: Integer-only quantization for efficient vision Transformer inference[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 17065−17075

[89] Tai Yu-Shan, Wu An-Yeu. AMP-ViT: Optimizing vision transformer efficiency with adaptive mixed-precision post-training quantization[C]//Proc of the IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2025: 6828−6837

[90] Xiao Junrui, Li Zhikai, Yang Lianwei, et al. Patch-wise mixed-precision quantization of vision Transformer[C]//Proc of the Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2023: 1−7

[91] Wang Ziwei, Wang Changyuan, Xu Xiuwei, et al. Quantformer: Learning extremely low-precision vision transformers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(7): 8813−8826

[92] Liu S, Liu Zechun, Cheng K. Oscillation-free quantization for low-bit vision transformers[C]//Proc of the 40th Int Conf on Machine Learning. New York: ICML, 2023: 21813−21824

[93] Liu Yijiang, Yang Huanrui, Dong Zhen, et al. NoisyQuant: Noisy bias-enhanced post-training activation quantization for vision transformers[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 20321−20330

[94] Li Liang, Li Qingyuan, Zhang Bo, et al. Norm Tweaking: High-performance low-bit quantization of large language models[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2024: 18536−18544

[95] Zeng Chao, Liu Songwei, Xie Yusheng, et al. ABQ-LLM: Arbitrary-bit quantized inference acceleration for large language models[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 22299−22307

[96] Zhang Feng, Liu Yanbin, Li Weihua, et al. A mixed quantization approach for data-free quantization of LLMs[C]//Proc of Int Conf on Agents and Artificial Intelligence (ICAART). Setúbal, Portugal: SciTePress, 2025: 353−363

[97] Rao Chuan, Chen Liangying, Xu Ruyi, et al. A deep neural network compression method based on dynamic quantization coding[J]. Acta Automatica Sinica, 2019, 45(10): 1960−1968 (in Chinese)
(饶川, 陈靓影, 徐如意, 等. 一种基于动态量化编码的深度神经网络压缩方法[J]. 自动化学报, 2019, 45(10): 1960−1968)

[98] Zhou Wenqiang, Yu Zhendong, Liu Xinyu, et al. Precision neural network quantization via learnable adaptive modules[J]. arXiv preprint, arXiv: 2504.17263, 2025

[99] Sher A, Trusov A, Limonova E, et al. Neuron-by-neuron quantization for efficient low-bit QNN training[J]. Mathematics, 2023, 11(9): 2112

[100] Lin Shih-Ting, Li Zhaofang, Cheng Yu-Hsiang, et al. LG-LSQ: Learned gradient linear symmetric quantization for low-precision integer hardware[C]//Proc of the 5th IEEE Int Conf on Artificial Intelligence Circuits and Systems. Piscataway, NJ: IEEE, 2023: 1−5

[101] Shang Yuzhang, Xu Dan, Zong Ziliang, et al. Network binarization via contrastive learning[G]//LNCS 13671: Proc of the Eur Conf on Computer Vision. Berlin: Springer, 2022: 586−602

[102] Zhang Yifan, Dong Zhen, Yang Huanrui, et al. QD-BEV: Quantization-aware view-guided distillation for multi-view 3D object detection[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 3802−3812

[103] Choi K, Lee H, Hong D, et al. It's all in the teacher: Zero-shot quantization brought closer to the teacher[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 8301−8311

[104] Xu Shoukai, Li Haokun, Zhuang Bohan, et al. Generative low-bitwidth data free quantization[G]//LNCS 12357: Proc of the European Conf on Computer Vision. Berlin: Springer, 2020: 1−17

[105] Choi K, Hong D, Park N, et al. Qimera: Data-free quantization with synthetic boundary supporting samples[C]//Proc of the 35th Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2021: 14835−14847

[106] Li Huantong, Wu Xiangmiao, Lv Fanbing, et al. Hard sample matters a lot in zero-shot quantization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 24417−24426

[107] Van Baelen Q, Karsmakers P. Constraint guided model quantization of neural networks[J]. arXiv preprint, arXiv: 2409.20138, 2024

[108] Sun Chang, Årrestad T K, Loncar V, et al. Gradient-based automatic per-weight mixed precision quantization for neural networks on-chip[J]. arXiv preprint, arXiv: 2405.00645, 2024

[109] Jia Shaohang, Huang Zhiyong, Yu Zhi, et al. Adaptive distribution-aware quantization for mixed-precision neural networks[J]. arXiv

preprint, arXiv: 2510.19760, 2025

[110] Habi H V, Jennings R H, Netzer A. HMQ: Hardware friendly mixed precision quantization block for CNNs[G]//LNCS 12371: Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 448−463

[111] Liu Zhenhua, Wang Yunhe, Han Kai, et al. Instance-aware dynamic neural network quantization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 12424−12433

[112] Zhao Xiaotian, Xu Ruge, Gao Yimin, et al. Edge-MPQ: Layer-wise mixed-precision quantization with tightly integrated versatile inference units for edge computing[J]. IEEE Transactions on Computers, 2024, 73(11): 2504−2519

[113] Bondarenko Y, Del Chiaro R, Nagel M. Low-rank quantization-aware training for LLMs[J]. arXiv preprint, arXiv: 2406.06385, 2024

[114] Chen Mengzhao, Shao Wenqi, Xu Peng, et al. EfficientQAT: Efficient quantization-aware training for large language models[C]// Proc of the 63rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2025: 10081−10100

[115] Jeon H, Kim Y, Kim J. L4Q: Parameter efficient quantization-aware training on large language models via LoRA-wise LSQ[J]. arXiv preprint, arXiv: 2402.04902, 2024

[116] Ke Wenjing, Li Zhe, Li Dong, et al. DL-QAT: Weight-decomposed low-rank quantization-aware training for large language models[C]// Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 113−119

[117] Pei Zehua, Yao Xufeng, Zhao Wenqian, et al. Quantization via distillation and contrastive learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(12): 17164−17176

[118] Liu Kai, Qin Haotong, Guo Yong, et al. 2DQuant: Low-bit post-training quantization for image super-resolution[C]//Proc of the 38th Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2024, 37: 71068−71084

[119] Gernigon C, Filip S I, Sentieys O, et al. AdaQAT: Adaptive bit-width quantization-aware training[C]//Proc of the IEEE Int Conf on AI Circuits and Systems. Piscataway, NJ: IEEE, 2024: 442−446

[120] Zhuang Bohan, Liu Lingqiao, Tan Mingkui, et al. Training quantized neural networks with a full-precision auxiliary module[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 1485−1494

[121] Zhuang Bohan, Tan Mingkui, Liu Jing, et al. Effective training of convolutional neural networks with low-bitwidth weights and activations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6140−6152

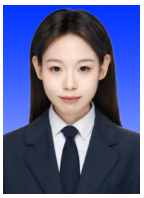
[122] Tian Luo, Wang Peng. An effective mixed-precision quantization method for joint image deblurring and edge detection[J]. Electronics, 2025, 14(9): 1767

[123] Tang Chen, Ouyang Kai, Wang Zhi, et al. Mixed-precision neural network quantization via learned layer-wise importance[G]//LNCS 13671: Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 259−275

[124] Ranjan N, Savakis A. Mix-QViT: Mixed-precision vision transformer quantization driven by layer importance and quantization sensitivity[J]. arXiv preprint, arXiv: 2501.06357, 2025

[125] Liu Yutong, Zhao Cairong, Hu Guosheng. A comprehensive evaluation on quantization techniques for large language models[J]. arXiv preprint, arXiv: 2507.17417, 2025

[126] Gong Zhuocheng, Liu Jiahao, Wang Jingang, et al. What makes quantization for large language model hard? An empirical study from the lens of perturbation[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 18082−18089



**Huang Xinyi**, born in 2002. Master candidate. Her main research interests include edge intelligence and model compression.

**黄馨怡**, 2002 年生。硕士研究生。主要研究方向为边缘智能、模型压缩。



**Qu Zhihao**, born in 1989. PhD, associate professor. His main research interests include edge computing and federated learning.

**屈志昊**, 1989 年生。博士，副教授。主要研究方向为边缘计算、联邦学习。



**Ye Baoliu**, born in 1976. PhD, professor. His main research interests include distributed systems, cloud computing, and edge intelligence.

**叶保留**, 1976 年生。博士，教授。主要研究方向为分布式系统、云计算、边缘智能。



**Hu Shihong**, born in 1993. PhD. Her main research interests include edge intelligence and resource scheduling.

**胡世红**, 1993 年生。博士。主要研究方向为边缘智能、资源调度。



**Jia Ninghui**, born in 1998. PhD candidate. His main research interests include federated learning and edge intelligence.

**贾柠晖**, 1998 年生。博士研究生。主要研究方向为联邦学习、边缘智能。



**Qian Lin**, born in 1984. PhD, senior engineer. His main research interest includes distributed computing and systems.

**钱　琳**, 1984 年生。博士，高级工程师。主要研究方向为分布式计算与系统。