

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-15

论文引用格式: Huo Ying, Xie Xinyi, Chen Jingjing, Feng Wenqing. Weight-Activation Co-Adaptive Quantization for RT-DETR[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260161. (霍瑛, 谢心怡, 陈晶晶, 封文清. 面向RT-DETR的权重激活协同自适应量化方法[J/OL]. 中国图象图形学报, XXXX:1-15. DOI: 10.11834/jig.260161.) [DOI:10.11834/jig.260161]

面向RT-DETR的权重激活协同自适应量化方法

霍瑛*, 谢心怡, 陈晶晶, 封文清

南京工程学院计算机工程学院, 南京 211167

摘要: 目的 针对实时检测Transformer(real-time detection transformer, RT-DETR)在极低比特后训练量化下因权重与激活联合离散化而导致检测精度显著退化的问题, 研究一种无需重新训练、适用于边缘部署的低比特量化方法。方法 提出权重激活协同自适应量化(weight-activation co-adaptive quantization, WCAQ)框架。该方法构建端到端、无需重新训练的后训练量化流程: 权重侧采用分组伪量化以抑制静态离散误差, 激活侧将量化置于算子输出端并引入样本自适应尺度调节, 结合敏感分支保护和轻量级逐层重构, 对双侧量化引起的联合误差进行补偿。结果 在COCO 2017(Common Objects in Context)上, WCAQ在8位和6位量化下平均精度(average precision, AP)分别达到46.2%和44.8%, 与全精度基线高度相当; 在4位极低比特下仍保持28.0%的AP, 较最优方法QRT-DETR提高5.4%, 克服了传统分离量化导致的精度坍塌。整个量化流程仅需19分钟, 较QRT-DETR的110分钟降低约82.7%。结论 WCAQ在无需重新训练的前提下实现了接近无损的8位量化和稳定的6位量化, 并显著改善了4位量化下RT-DETR的部署可用性, 为检测Transformer的边缘低比特部署提供了高效可靠的后训练量化方案。本文方法已发布于ScienceDB, 访问链接为: <https://doi.org/10.57760/sciencedb.j00240.00092>。

关键词: 实时目标检测; RT-DETR; 模型量化; 后训练量化; Transformer

Weight-Activation Co-Adaptive Quantization for RT-DETR

Huo Ying*, Xie Xinyi, Chen Jingjing, Feng Wenqing

Department of Computer Engineering, Nanjing Institute of Technology, Nanjing 211167, China

Abstract: **Objective** Post-training quantization (PTQ) is attractive for real-time object detection because it can be applied after model training and does not require task-level retraining. Yet this advantage is difficult to maintain for detection Transformers. In RT-DETR, multi-scale feature fusion, query interaction, and box regression are tightly coupled, so quantization noise introduced at one stage is easily amplified by later computations. Existing post-training pipelines usually quantize weights and activations as if they were largely separable sources of error: weights are discretized by static parameter statistics, activations are calibrated by fixed distribution summaries, and the final degradation is treated as an unavoidable consequence of low-bit deployment. This assumption is especially problematic for RT-DETR. Under ultra-low-bit settings, weight-side discretization shifts operator responses, activation-side discretization further distorts already perturbed features, and the compounded error is propagated through encoder-decoder interaction and prediction heads. The result is a

收稿日期: 2026-03-30; 修回日期: 2026-09-21

* 通信作者: 霍瑛 huoy@njit.edu.cn

基金项目: 江苏省高等学校自然科学研究项目(No. 23KJB520011); 南京工程学院校级科研基金项目(No. YKJ202215); 2023年江苏高校“青蓝工程”项目基金项目。

Supported by: Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 23KJB520011); Scientific Research Foundation of Nanjing Institute of Technology of China (No. YKJ202215); the Funding project of “Qing Lan Project” in Jiangsu Province (2023)

©中国图象图形学报版权所有

sharp collapse in detection accuracy, particularly at 4 bit precision. This study therefore aims to develop a training-free and deployment-oriented quantization framework for RT-DETR that suppresses coupled weight-activation error accumulation while preserving efficiency, stability, and usable low-bit detection performance. **Method** The proposed WCAQ framework is implemented as a complete post-training workflow rather than as an isolated quantization scheme. Starting from a pre-trained floating-point RT-DETR model, the pipeline performs structure fusion, quantized module replacement, calibration, fake quantization enablement, optional layer-wise reconstruction, and final validation, without entering a training loop. On the weight side, WCAQ adopts grouped pseudo-quantization. In the current implementation, convolutional and linear weights are reshaped and quantized group by group, using fixed-size local groups to reduce the sensitivity of low-bit mapping to outliers and heterogeneous parameter distributions. This design suppresses static discretization error more effectively than a single global scale. On the activation side, WCAQ uses a sample-aware adaptive quantizer based on adaptive step-size quantization (ASQ). The effective activation scale is obtained from a learnable base scale multiplied by a sample-dependent modulation factor generated from current feature statistics, and this factor is explicitly bounded to maintain numerical stability. The activation path is also structured carefully: fake quantization is placed at the output of convolutional and linear operators rather than at the input, so the core mapping operates on less disturbed features before discretization is applied. In addition, the framework protects highly sensitive branches by excluding modules related to sampling offsets, attention weights, score heads, and bounding-box prediction heads from aggressive quantization when necessary. After dual-side quantization is established, WCAQ applies lightweight layer-wise reconstruction using floating-point layer outputs as teacher targets. In the present implementation, reconstruction is performed with 200 iterations and a small reconstruction batch size, so recovery remains substantially lighter than retraining-based correction. Overall, the method realizes three explicit forms of cooperation: grouped weight quantization and dynamic activation scaling cooperate functionally, output-side activation quantization and reconstruction cooperate along the error propagation path, and main-path quantization cooperates structurally with sensitive-branch protection. **Result** Experiments were conducted on COCO 2017 using RT-DETR with a ResNet-18 backbone (RT-DETR-R18). The full-precision baseline achieved 46.40% AP. Under WCAQ, the detector reached 46.2% AP at 8 bit, 44.8% AP at 6 bit, and 28.0 % AP at 4 bit. These results support three main observations. First, the 8 bit result is nearly lossless relative to the floating-point baseline, showing that the proposed pipeline introduces minimal distortion when representational capacity remains sufficient. Second, at 6 bit precision the detector still maintains AP above 44%, indicating that the framework remains robust in a practically meaningful low-bit regime rather than only at near-floating-point settings. Third, although performance declines substantially at 4 bit precision, the detector preserves usable object detection capability instead of collapsing into severe numerical instability. Mechanism-oriented comparisons further explain the source of these gains. Most notably, output-side activation quantization yields a dramatic advantage over input-side activation quantization under 4 bit conditions, with an absolute AP gain of more than 24 points in the current experiment set, showing that quantization location is a dominant stability factor for RT-DETR. The reconstruction stage also restores part of the low-bit accuracy by realigning quantized layer outputs with floating-point teacher responses, confirming that explicit recovery of coupled error is necessary after dual-side discretization has already been introduced. Moreover, WCAQ reduces the quantization time from 110 min to 19 min, corresponding to an 82.7% reduction, without task-level retraining. These accuracy and efficiency gains are achieved within a training-free deployment pipeline. WCAQ does not rely on full fine-tuning, large-scale quantization-aware retraining, or expensive end-to-end optimization over the entire detector; instead, it recovers performance through calibration-aware module design, adaptive activation scaling, selective structural protection, and lightweight layer-wise correction. **Conclusion** This study presents a training-free quantization framework for RT-DETR that directly addresses the coupled nature of weight-side and activation-side error in detection Transformers. Instead of assuming that static calibration alone is sufficient, WCAQ integrates grouped weight pseudo-quantization, sample-aware adaptive activation scaling, output-side activation quantization, sensitive-branch protection, and lightweight layer-wise reconstruction into a unified post-training pipeline. Consequently, WCAQ preserves near-floating-point accuracy at 8 bit, remains highly competitive at 6 bit, and retains usable detection capability at 4 bit, all without task-level retraining. Methodologically, the results show that low-bit RT-DETR compression should be formulated as a coupled error control problem rather than as two independent quantization subproblems. Practi-

cally, the framework demonstrates that resource-constrained deployment of detection Transformers can be pursued through carefully structured PTQ instead of immediately resorting to retraining-heavy solutions. The current implementation also highlights three transferable design principles: activation handling in ultra-low-bit settings should remain distribution-aware, quantization should be delayed until after core operator responses whenever possible, and highly sensitive branches should be protected selectively rather than compressed indiscriminately. Future work may extend the same cooperative logic to larger RT-DETR backbones, refine attention-specific low-bit treatment, and evaluate the framework under stricter integer-only deployment constraints and broader edge inference scenarios. The method presented in this paper is available in ScienceDB at <https://doi.org/10.57760/sciencedb.j00240.00092>.

Key words: real-time object detection; RT-DETR; model quantization; post-training quantization; Transformer

论文引用格式: Huo Ying, Xie Xinyi, Chen Jingjing, Feng Wenqing. Weight-Activation Co-Adaptive Quantization for RT-DETR. Journal of Image and Graphics. DOI: 10.11834/jig.260161. (霍瑛, 谢心怡, 陈晶晶, 封文清. 面向 RT-DETR 的权重激活协同自适应量化方法. 中国图象图形学报. DOI: 10.11834/jig.260161.)

0 引言

RT-DETR (real-time detection transformer, RT-DETR) 作为一种端到端实时目标检测模型, 通过高效混合编码器和不确定性最小查询选择机制, 在检测精度与推理速度之间取得优异平衡 (Zhao 等, 2024)。随后提出的 RT-DETRv2 和 RT-DETRv3 引入了 bag-of-freebies 和分层密集正样本监督等策略, 进一步提升了模型性能 (Lv 等, 2024; Wang 等, 2025)。凭借较好的实时性与精度表现, RT-DETR 系列在自动驾驶 (Hu 等, 2025)、视频监控 (Yang L 等, 2024)、工业质检 (Liu 等, 2025) 和持刀危险行为检测 (金涛等, 2025) 等场景中展现出较强应用潜力。

然而在实际部署环境中, 边缘设备的显存容量、访存带宽和计算资源往往远比实验环境受限。以采用 50 层残差网络 (residual network, ResNet) 作为主干的 RT-DETR-R50 为例, 其单次前向推理的浮点运算次数 (floating-point operations, FLOPs) 约为 136×10^9 (Zhao 等, 2024)。在手机终端、车载平台和嵌入式微控制器等资源受限场景中, 如此庞大的计算与存储开销已成为模型部署的核心瓶颈。为在不改变基本推理算子形式的前提下降低模型资源需求, 后训练量化 (post-training quantization, PTQ) 成为适配边缘部署的重要技术路径 (杨春 等, 2023)。目前已

有研究表明, 混合精度量化和混合截断量化等策略能够在一定程度上缓解低比特条件下的精度退化问题 (黄钰喆 等, 2021; 李奕铎 等, 2022)。面向轻量卷积神经网络的研究也进一步印证了训练后量化在资源受限场景下的较高部署价值 (杨杰 等, 2025)。此外, 利用合成数据校准的无数据量化方法也已应用于深度伪造检测, 为原始训练数据不可用时的模型压缩提供了可行途径 (卓文琦 等, 2023)。

尽管量化技术能显著压缩模型, 现有通用量化方法难以直接适配 RT-DETR。端到端检测 Transformer 内部的查询机制对量化噪声极为敏感, 低比特条件下量化带来的信息失真会大幅削弱模型的目标表征与特征匹配能力 (Wang 等, 2024; Xu 等, 2023)。更为严峻的是, 传统后训练量化方法通常将权重量化与激活量化分离处理, 导致校准阶段与实际部署阶段存在统计环境不一致的问题。解码和回归等敏感路径上的误差会因此在向前传播过程中被持续放大并累积 (Bondarenko 等, 2021; Wei 等, 2023)。OPTQ (Frantar 等, 2023) 与 SmoothQuant (Xiao 等, 2023) 分别从权重重构和量化难度迁移的角度推进了低比特量化研究, 但它们对于端到端检测 Transformer 中权重侧与激活侧联合误差的统一建模与协同补偿机制仍显不足。

针对上述局限, 本文提出一种面向 RT-DETR 的权重激活协同自适应量化框架 (weight-activation co-adaptive quantization, WCAQ)。与传统分离量化方式不同, WCAQ 在统一的后训练量化框架下同时建立权重与激活的协同关系。具体方案以融合后的浮点模型为基线, 对卷积层和线性层进行统一的量化封装。在权重侧, 采用受 AWQ (Lin 等, 2024) 启发的分组式伪量化策略, 精细化削弱参数离散化带来的结构性误差。在激活侧, 引入受 ASQ (Zhou 等, 2023) 启发的逐层校准策略, 通过引入可学习的校准参数, 有效缓解了低比特下激活分布偏移带来的精度损失。实验结果表明, WCAQ 在保持推理速度的同时, 显著提升了 RT-DETR 在低比特下的检测精度, 尤其在复杂场景下的性能提升更为明显。本文的贡献总结如下:

2026)启发的样本自适应量化机制,根据当前特征统计分布动态调节量化尺度,并将激活量化置于算子输出端,避免输入特征的过早失真。为进一步补偿联合量化产生的残余偏差,框架结合逐层重构机制,缩小量化层与浮点层输出之间的特征偏差。同时,对查询与回归相关的高敏感模块实施选择性保护,保障极低比特下的数值稳定性。本文的主要贡献概括如下。

1)针对 RT-DETR 在极低比特 PTQ 中存在的权重误差与激活误差耦合放大问题,提出 WCAQ 框架,首次从统一 PTQ 流程的角度对检测 Transformer 的双侧量化误差进行系统建模与协同控制,突破了已有研究将权重侧和激活侧分离、难以稳定支撑极低比特部署的局限。

2)面向检测 Transformer 中权重统计相对稳定而激活分布强动态的结构特征,构建了分组权重量化与动态激活适配相结合的异构协同量化机制,并将激活量化后置于算子输出端,降低输入端过早离散化对核心映射的破坏。

3)设计了由输出端激活量化、逐层重构和敏感分支保护共同组成的低成本补偿路径,在无需任务级重新训练的前提下,将量化耗时从 110 分钟降至 19 分钟,并在 4 位宽下将平均精度 (average precision, AP) 提升至 28.0%,显著改善了 RT-DETR 的极低比特部署可用性。

1 相关研究工作

端到端检测 Transformer 引入了集合匹配机制,削弱了传统检测流程对候选框生成和非极大值抑制的依赖 (Carion 等, 2020)。Deformable DETR (Zhu 等, 2021)通过稀疏采样注意力有效缓解了收敛速度和小目标检测上的不足,RT-DETR 系列 (Lv 等, 2024; Wang 等, 2025; Zhao 等, 2024)则进一步验证了端到端 Transformer 检测器在实时场景中的可行性,并通过高效混合编码器与查询初始化策略实现了速度与精度之间的灵活权衡。然而,RT-DETR 在跨尺度特征交互、查询匹配以及边界框回归等环节对数值扰动较为敏感,这使得其在低比特部署时,尤其在权重与激活同时量化的条件下,更容易出现明显的精度退化。

围绕后训练量化,现有研究主要聚焦于权重侧

误差控制、激活侧分布校正以及检测器结构适配三个维度。在权重侧误差控制方面,BRECQ 通过块级重构减小了局部量化误差 (Li 等, 2021)。OPTQ (Frantar 等, 2023)则利用近似二阶信息实现高效权重量化,CFWS (Yang D 等, 2024)进一步针对重参数化网络中的权重分布不平衡问题提出 coarse-to-fine 分解策略,以改善低比特条件下的权重表示精度。上述方法为抑制参数离散化误差提供了重要基础,但其关注重点总体仍集中在权重空间本身,对于检测模型中激活分布随输入变化而剧烈波动的特性考虑相对不足。

相较于权重量化,激活量化通常更具挑战性,其根源在于激活响应应具有更强的动态性与重尾分布特征,少量离群值即可显著拉大量化范围,从而在低比特条件下加剧截断误差与舍入误差。针对这一问题,SmoothQuant (Xiao 等, 2023)通过等价变换将部分量化难度由激活侧迁移至权重侧,以平滑异常激活。Q-YOLO (Wang 等, 2023)则面向实时目标检测,重点缓解 YOLO 中激活分布失衡导致的低比特性能退化。MSQuant (Jiang 等, 2025)进一步针对目标检测中的双侧量化难题提出了迁移尺度搜索策略,有效平衡了权重与激活之间的量化难度。与此同时,ASQ & POST (Zhou 等, 2026)从自适应与非均匀量化的角度给出新的思路,其中 ASQ 通过依据输入激活统计动态调节量化步长来增强对分布变化的适应能力,POST 则利用非均匀量化网格提升权重表示精度。该工作表明,自适应激活建模对于低比特性能保持具有重要意义,但其主要面向通用视觉网络,并未直接针对 RT-DETR 这类端到端检测 Transformer 的后训练双侧量化问题展开研究。

综上所述,现有方法虽从多个角度提升了低比特量化性能,但仍面临两大局限:一是大量方法服务于 YOLO 系列、重参数化卷积网络或大语言模型等特定架构,难以直接迁移到 RT-DETR 框架中;二是多数方法偏重于权重或激活的单维优化,缺乏对统一量化路径中双侧误差协同建模与误差补偿的机制。本文正是针对这一空白,提出 WCAQ 框架,在统一 PTQ 路径下实现权重与激活的协同量化与误差补偿。

2 方法

现有方法普遍缺乏对端到端检测 Transformer 中权重与激活协同建模的机制。本节详细阐述 WCAQ 框架,该框架在保持原有网络结构不变的前提下,通过构建统一的后训练量化路径,有效缓解了低比特下的联合误差累积现象。下文先概述整体架构,随后按执行逻辑逐一解析激活感知分组权重量化、样本级动态激活适配以及耦合协同重构这三个核心执行阶段。

2.1 预备知识

2.1.1 RT-DETR 架构特性与部署挑战

RT-DETR 是一种高效的端到端实时目标检测模型,其网络拓扑主要由特征提取主干网络(backbone)、高效混合编码器(efficient hybrid encoder)以及带有辅助预测头的 Transformer 解码器(decoder)三部分构成。该架构通过不确定性最小化的查询选择机制,摒弃了传统目标检测中耗时的非极大值抑制(non-maximum suppression, NMS)后处理步骤。然而,RT-DETR 在实现高精度的同时,其编码器和解码器内部的多头注意力机制和前馈神经网络(feed-forward network, FFN)包含密集的矩阵乘法运算。在处理多尺度高分辨率特征时,这些模块会产生庞大的浮点计算量和显存访问开销。这种架构特性使得 RT-DETR 在算力和存储极其受限的边缘设备上

难以直接部署,迫切需要引入低比特模型量化技术进行压缩。

2.1.2 神经网络后训练量化基础

PTQ 旨在无需大量重新训练的前提下,将模型中连续的高精度浮点权重与激活值映射为离散的低精度定点数值。标准的均匀仿射量化过程主要由量化尺度 s 和零点 z_0 定义。给定任意浮点张量 $\mathbf{x}$ 和目标量化位宽 b ,其前向量化映射过程可形式化表示为:

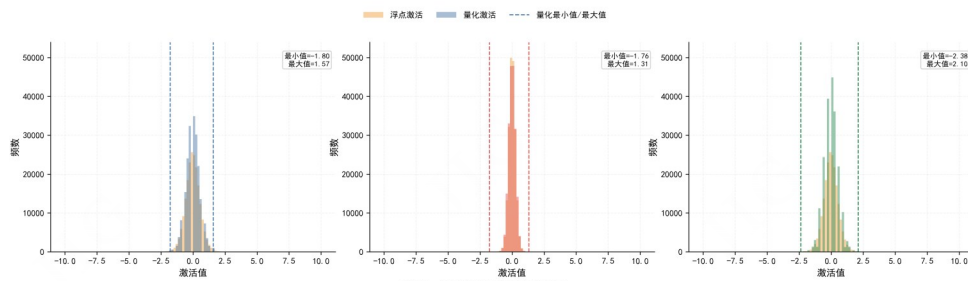
$$\mathbf{x}_q = \text{clamp}\left(\left\lfloor \frac{\mathbf{x}}{s} + 0.5 \right\rfloor + z_0, q_{\min}, q_{\max}\right) \quad (1)$$

式中, $\lfloor \cdot \rfloor$ 表示向下取整操作, $\text{clamp}(\cdot)$ 将结果限制在整数区间 $[q_{\min}, q_{\max}]$ 内,对 b 位量化位宽, $q_{\min} = 0$, $q_{\max} = 2^b - 1$ 。对应的反量化过程为:

$$\hat{\mathbf{x}} = s(\mathbf{x}_q - z_0) \quad (2)$$

式中 $\mathbf{x}_q$ 为量化后的整数张量, $\hat{\mathbf{x}}$ 为反量化后的近似浮点张量。

传统 PTQ 方法通常分别对权重和激活进行静态校准,并依据最小-最大值(minimum-maximum, MinMax)、百分位数或均方误差(mean squared error, MSE)等准则估计量化区间。该策略在卷积分类网络中尚可奏效,但在 RT-DETR 这类端到端检测 Transformer 中往往面临更大的困难:中间激活分布随输入内容、特征尺度和查询状态剧烈变化,且极低比特下权重量化误差与激活量化误差沿前向链路持续累积,最终导致检测性能显著下降。



(a)静态输出端量化 (b)输入端激活量化 (c)WCAQ输出端动态量化

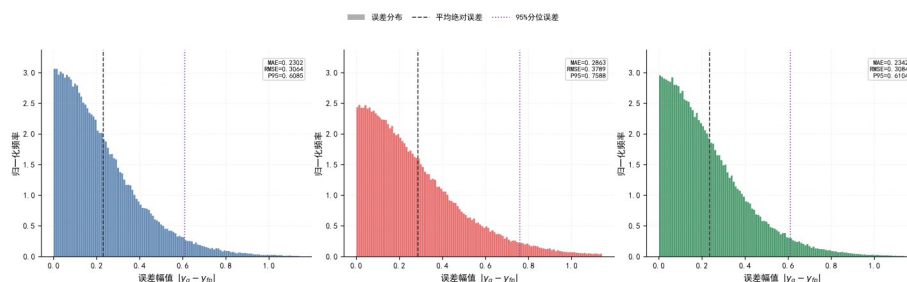
((a) static output-side quantization; (b) input-side activation quantization; (c) WCAQ dynamic output-side quantization)

图1 RT-DETR 代表性层激活分布重叠对比图

Fig. 1 Comparison of activation distribution overlap in a representative RT-DETR Layer

为直观展示不同激活量化路径的影响,图1给出了 RT-DETR 解码器第一层跨注意力值映射层在三种量化方式下的代表性激活分布重叠情况,其中横坐标为激活值,纵坐标为落入对应激活区间的特

征元素频数。三种方案的核心差异在于:“静态输出端量化”使用固定全局阈值,量化放在算子输出端;“输入端量化”同样使用固定阈值,但量化移到算子输入端;“WCAQ 输出端动态量化”则是本文方案,量



(a)静态输出端量化误差 (b)输入端激活量化误差 (c)WCAQ输出端动态量化误差

((a) static output-side quantization error; (b) input-side activation quantization error; (c) WCAQ dynamic output-side quantization error)

图2 RT-DETR代表性层量化误差分布对比图

Fig. 2 Comparison of Quantization Error Distributions in a Representative RT-DETR Layer

化尺度随当前样本特征统计动态调节,且量化位置保持在算子输出端。图1显示静态输出端量化后激活的主要取值区间约为 $[-1.80, 1.57]$,输入端量化收缩至 $[-1.76, 1.31]$,而WCAQ输出端动态量化扩展至 $[-2.38, 2.10]$,有效覆盖区间分别扩大约32.9%和45.9%。这说明WCAQ通过样本相关的动态尺度调节,在保持离散表示的同时更充分地覆盖有效响应区间。进一步观察三组分布的中心位置,WCAQ输出端动态量化后的激活仍主要集中在浮点分布高密度区域附近,未出现明显偏移。这表明激活量化的关键不仅在于是否量化,更在于量化尺度能否随当前特征响应自适应变化,以及量化操作是否放置在更合适的路径位置。

图2进一步比较了同一代表性层在三种量化方案下相对于浮点输出的误差分布,其中横坐标为量化输出与浮点输出的绝对误差,纵坐标为归一化频率,用于表征误差在各区间的相对分布。输入端激活量化的误差分布整体更宽,且长尾更明显,其平均绝对误差(mean absolute error, MAE)、均方根误差(root mean square error, RMSE)和95%分位误差(95th percentile, P95)分别达到0.2863、0.3789和0.7588,显著高于静态输出端量化的0.2302、0.3064和0.6085和WCAQ动态量化0.2342、0.3084和0.6104。WCAQ相比输入端量化在MAE、RMSE和P95上分别下降约18.2%、18.6%和19.6%,已基本恢复到静态输出端量化的误差水平。这说明输入端过早离散化会使误差在卷积或线性映射内部被进一步放大,而WCAQ通过输出端量化与动态尺度适配的组合,在保留更充分响应区间的同时未引入额外显著误差。基于上述分析,权重量化

与激活量化必须在统一框架下进行协同处理,从误差来源、传播路径和补偿机制三个层面统一设计。图3给出了WCAQ的整体框架。

2.2 整体架构概述

WCAQ从统一PTQ路径出发,将分组权重量化、样本级动态激活尺度、输出端量化路径、逐层重构补偿和敏感分支保护纳入同一框架,以协同控制极低比特条件下的联合量化误差。

根据图3的整体框架,该方法保持检测器原有主干网络、编码器、解码器和预测头结构不变,在浮点模型基础上构建统一的PTQ路径,使模型由浮点推理平滑过渡到低比特部署。整体流程包括激活感知分组权重量化、样本级动态激活适配和耦合协同重构三个阶段。量化模型由浮点模型复制生成,并通过卷积层和线性层的统一替换完成量化结构构建。

对于任意可量化算子,WCAQ的前向过程可表示为

$$\hat{\mathbf{W}} = Q_w(\mathbf{W}), \mathbf{r} = f(\mathbf{x}; \hat{\mathbf{W}}), \hat{\mathbf{y}} = Q_A(\mathbf{r}) \quad (3)$$

式中, $\mathbf{W}$ 为浮点权重, $\hat{\mathbf{W}}$ 为低比特权重, $\mathbf{x}$ 为输入特征, $\mathbf{r}$ 为中间响应, $\hat{\mathbf{y}}$ 为量化输出, $Q_w(\cdot)$ 为权重量化算子, $f(\cdot)$ 为卷积或线性映射, $Q_A(\cdot)$ 为激活量化算子。式(3)说明WCAQ并非在算子输入端同时对权重与激活进行粗粒度离散化,而是先在权重侧构造稳定的低比特表示,再对算子输出执行激活量化,从而使权重侧静态误差控制与激活侧动态尺度约束在统一前向路径中发生耦合。

结合图1和图2可以进一步理解该设计的必要性。图1说明输出端动态量化比输入端量化更有利

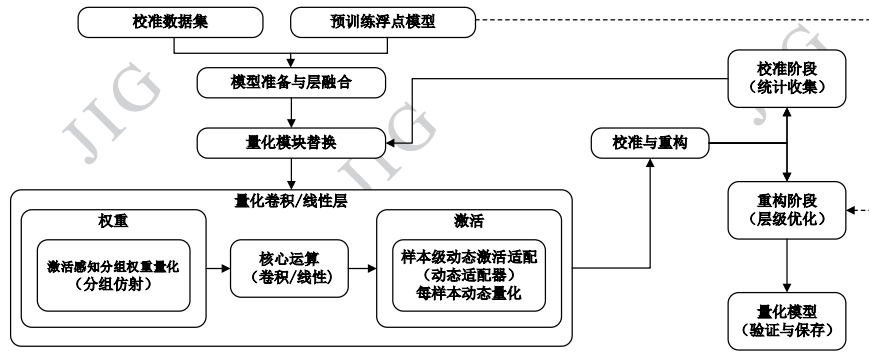


图3 WCAQ模型框架图

Fig. 3 The overall framework of the proposed WCAQ model

于保留有效响应区间,图2说明输入端量化会引入更大的平均误差和更长的误差尾部。因此WCAQ将激活量化后置到算子输出端,并通过样本级动态尺度调节机制使量化尺度与当前响应分布相匹配,以降低前向误差被过早放大的风险。

在完成双侧量化后,框架进一步引入逐层重构以补偿联合误差在深层网络中的持续累积。图4展示了逐层重构前后量化误差分布的变化,重构后误差分布整体向低误差区域收缩,MAE与高分位误差均得到抑制。

综上,WCAQ的协同主要体现在三个层面:其一,权重侧分组量化与激活侧动态尺度适配之间的功能协同;其二,输出端激活量化与逐层重构之间的路径协同;其三,主体量化与敏感分支保护之间的结构协同。三者共同构成了WCAQ在极低比特条件下保持检测鲁棒性的架构基础。

2.3 WCAQ算法

WCAQ构成面向RT-DETR的三阶段统一量化框架:阶段I通过激活感知分组权重量化构建稳定的低比特参数表示,阶段II通过样本级动态激活适配建立自适应量化响应,阶段III通过耦合协同重构补偿联合离散化误差。

2.3.1 阶段I:激活感知分组权重量化

阶段I的目标是构建稳定的低比特权重表示。RT-DETR中卷积层和线性层参数规模大、通道间分布差异显著,单一全局量化尺度易造成小数值组过度压缩、大数值组截断严重。为此,WCAQ采用分组伪量化策略,将权重在通道维度上划分为若干独立组,每组独立估计量化尺度与零点,有效隔离了不同分布区间之间的干扰,使量化后的权重表示更接近原始浮点分布,为后续激活量化提供更稳健的参数

基础。

对第 g 个权重分组 $\mathbf{W}^{(g)}$,其量化和反量化过程如下:

$$\mathbf{W}_q^{(g)} = \text{clamp} \left(\left\lfloor \frac{\mathbf{W}^{(g)}}{s^{(g)}} + 0.5 \right\rfloor + z_0^{(g)}, q_{\min}^w, q_{\max}^w \right) \quad (4)$$

$$\hat{\mathbf{W}}^{(g)} = s^{(g)} (\mathbf{W}_q^{(g)} - z_0^{(g)}) \quad (5)$$

式中, $s^{(g)}$ 和 $z_0^{(g)}$ 为第 g 组的权重量化尺度和零点, $[q_{\min}^w, q_{\max}^w]$ 为量化区间。对于 b_w 位宽非对称仿射权重量化,有 $q_{\min}^w = 0, q_{\max}^w = 2^{b_w} - 1$ 。该策略兼顾了压缩率与表示能力,适用于检测模型中的大规模卷积与线性映射。

2.3.2 阶段II:样本级动态激活适配

阶段II的目标是建立与当前输入样本相匹配的激活量化尺度。与权重统计相对稳定不同,RT-DETR中间激活随图像内容、跨尺度融合和查询状态显著波动,固定量化尺度难以稳定适配。图1显示静态输出端量化和输入端量化的有效覆盖区间分别为3.37和3.07,均小于浮点分布的原始跨度,而WCAQ输出端动态量化将其扩展至4.48,提升了约32.9%~45.9%。图2进一步表明,输入端量化引入的MAE(0.2863)和P95误差(0.7588)均显著高于输出端方案,说明量化位置与尺度适配方式直接影响误差传播。

为此,WCAQ将激活量化置于算子输出端,并依据当前样本响应动态调节量化步长。设中间响应为 $\mathbf{r}$,则激活侧的有效步长定义为

$$s_{\text{eff}} = s_0 \cdot \beta(\mathbf{r}) \quad (6)$$

式中, s_0 为基础激活量化尺度, $\beta(\mathbf{r})$ 为由当前样本响应统计生成的自适应调节因子。相应地,激活量化过程可表示为

$$\hat{y} = s_{\text{eff}} \cdot \text{clamp}\left(\left\lfloor \frac{r}{s_{\text{eff}}} + 0.5 \right\rfloor, q_{\min}^a, q_{\max}^a\right) \quad (7)$$

式中,对于 b_a 位宽对称激活量化,有 $q_{\min}^a = -2^{b_a-1}$,
 $q_{\max}^a = 2^{b_a-1} - 1$,零点固定为0。

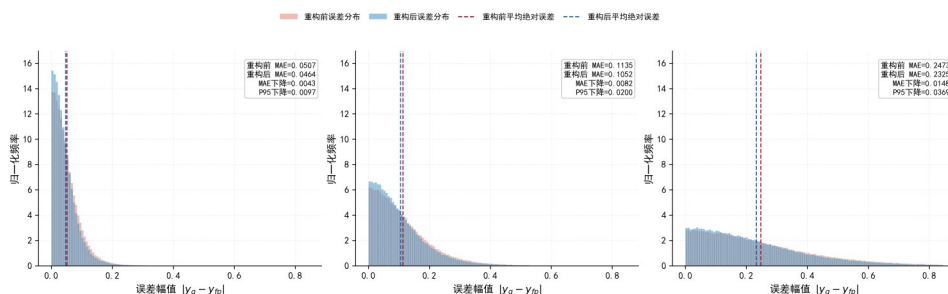
由此,激活量化步长能够随当前样本的输出分布动态调整,从而缓解权重量化引起的响应偏移。当前实现中,自适应因子由当前响应统计生成,并与基础步长共同决定实际量化尺度。

2.3.3 阶段III:耦合协同重构

阶段III的目标是补偿双侧量化后的联合误差。权重量化误差与激活量化误差并非简单相加,而是沿前向传播相互耦合并持续累积,即使前两阶段已

进行针对性控制,深层网络中仍可能残留不可忽略的特征偏移。为此,WCAQ在完成双侧量化后,以浮点模型对应层输出为教师信号执行逐层重构,对量化层输出偏差进行恢复。

图4展示了代表性层在不同位宽下重构前后的误差分布变化,其中横坐标为重构前后量化输出与浮点输出的绝对误差,纵坐标为归一化频率,用于比较重构前后的误差分布。从左到右三个子图可以看到,随着位宽降低,重构前后的差异逐渐更加明显。具体而言,在8位量化下,重构前后的MAE下降0.0043,P95由0.1343降至0.1246;在6位量化下,MAE下降0.0082,P95由0.2994降至0.2794;4位量化MAE下降0.0148,P95由0.6296降至0.5927。



(a)8位量化 (b)6位量化 (c)4位量化

((a) 8-bit quantization; (b) 6-bit quantization; (c) 4-bit quantization)

图4 逐层重构对低比特量化误差的抑制作用图

Fig. 4 Effect of Layer-wise Reconstruction on Suppressing Low-Bit Quantization Errors

结合相对降幅分析,逐层重构在三种位宽下均能稳定降低量化误差:MAE相对下降约8.5%、7.2%和6.0%,P95相对下降约7.3%、6.7%和5.7%。尤其是在低比特场景中,重构后误差分布整体向低误差区域收缩,说明其不仅降低了平均偏差,也抑制了误差尾部的异常扩张。这表明逐层重构并非可有可无的附加步骤,而是在权重量化误差与激活量化误差已经发生耦合后,对层输出进行针对性恢复的重要补偿机制。

在完成前两阶段后,权重量化与激活量化的联合离散化误差仍会在编码器、解码器和预测分支中逐层传播。WCAQ引入耦合协同重构,以浮点模型为教师,低比特模型为学生,通过最小化对应层输出差异恢复量化后的特征表达。设 l 为层索引,浮点教师层输出与量化层输出分别记为 $h_{i,FP}^{(l)}$ 和 $h_{i,Q}^{(l)}$,则重构目标可写为

$$\mathcal{L}_{\text{rec}} = \sum_{l \in L_q} \frac{1}{N} \sum_{i=1}^N \|h_{i,Q}^{(l)} - h_{i,FP}^{(l)}\|_2^2 \quad (8)$$

式中,重构损失为 $\mathcal{L}_{\text{rec}}$, L_q 为参与重构的量化层集合, N 为参与当前层重构损失计算的样本数, i 为样本索引, $h_{i,Q}^{(l)}$ 和 $h_{i,FP}^{(l)}$ 分别表示第 i 个样本在量化层和对应浮点层的输出, $\|\cdot\|_2$ 表示欧氏范数。

该约束通过局部特征一致性抑制联合量化误差的累积扩散。当前重构策略并不将权重侧与激活侧割裂处理,而是围绕整个量化单元展开协同补偿。只要量化层中仍保留可学习的调节参数,该层即可参与逐层重构。因此第三阶段本质上是对权重低比特表示与激活动态适配共同造成的特征偏差进行统一恢复。

算法的伪代码如下所示。

本文所称协同并非简单地同时使用权重量化与激活量化,而是指三种显式耦合关系。一是权重侧静态分组表达与激活侧动态尺度适配之间的功能协

算法 1: WCAQ 总体量化流程

输入: 浮点 RT-DETR 模型 M_{FP} , 校准数据集 D_{cal} 输出: 量化后的 RT-DETR 模型 M_Q

- 1: 对 M_{FP} 进行层融合, 并复制得到 M_Q
- 2: 将 M_Q 中可量化层替换为量化层, 敏感层保持浮点形式
- 3: 使用 D_{cal} 对 M_Q 进行校准, 初始化激活量化参数
- 4: 启用 M_Q 中各量化层的伪量化
- 5: for 每一个量化层 l do
- 6: 对权重执行分组量化, 得到 $\hat{W}^{(l)}$
- 7: 计算激活量化前的中间响应 $r^{(l)}$, 根据 $r^{(l)}$ 生成动态调节因子 $\beta(r^{(l)})$, 形成有效激活量化尺度 $s_{eff}^{(l)}$, 并对 $r^{(l)}$ 执行激活量化, 得到量化层输出 $\hat{y}^{(l)}$
- 8: end for
- 9: if 启用重构 then
- 10: 逐层对齐量化层输出与对应浮点层输出, 执行协同重构
- 11: end if
- 12: 返回 M_Q

算法 2: 量化层的前向传播

输入: 输入特征 $x^{(l)}$, 浮点权重 $W^{(l)}$, 基础激活尺度 $s_0^{(l)}$ 输出: 量化层输出 $\hat{y}^{(l)}$

- 1: 对 $W^{(l)}$ 进行分组伪量化, 得到 $\hat{W}^{(l)}$
- 2: 根据 $x^{(l)}$ 和 $\hat{W}^{(l)}$ 计算中间响应 $r^{(l)}$
- 3: 由当前响应生成动态调节因子 $\beta(r^{(l)})$
- 4: 构造有效激活量化尺度 $s_{eff}^{(l)} = s_0^{(l)} \cdot \beta(r^{(l)})$
- 5: 使用 $s_{eff}^{(l)}$ 对 $r^{(l)}$ 执行激活量化, 得到 $\hat{y}^{(l)} = Q_A(r^{(l)}; s_{eff}^{(l)})$
- 6: 返回 $\hat{y}^{(l)}$

同,前者负责降低基础离散误差,后者负责吸收样本分布波动;二是输出端激活量化与逐层重构之间的路径协同,前者减少算子前输入失真,后者进一步修正层输出偏差;三是固定量化权重与可学习激活适配参数之间的优化协同,前者提供稳定约束,后者承担低成本补偿。这三类协同共同构成了 WCAQ 在极低比特条件下保持检测性能的关键机制。

3 实验结果与分析

本节首先介绍数据集与实验实现细节,随后通过消融实验分析 WCAQ 各组成模块对整体性能贡献,并将所提方法与现有主流量化策略进行对比。最后在低比特量化条件下进一步验证该方法的鲁棒性与有效性。

3.1 数据集与实验实现细节

数据集与评价指标:采用 COCO 2017(Common

Objects in Context)(Lin 等,2014)对所提方法进行评价。评价指标为标准平均精度均值 AP,不同交并比(intersection over union, IoU)阈值下的检测精度 AP_{50} 和 AP_{75} ,以及针对不同目标尺度的检测结果,即小/中/大目标对应的 AP_s 、 AP_m 和 AP_l 。

基线模型:选取官方实现的 RT-DETR(Zhao 等,2024)的 32 位浮点(32-bit floating point, FP32)模型作为基线。各表中的 FP 表示浮点(floating point, FP)基线,其权重和激活均采用 32 位浮点表示。

实现细节:权重侧分组大小 64,激活侧使用样本自适应量化。校准集为从 COCO 训练集随机抽取的 100 张图像。在层级协同重构中,学习批次大小为 4,每层迭代 200 次,学习率为 4×10^{-5} 。

3.2 与现有量化方法的性能比较

为验证 WCAQ 在 RT-DETR 极低比特量化场景下的有效性,本文首先在相同的以 ResNet-18 为主干的 RT-DETR 模型(RT-DETR-R18)、相同数据集与评价指标下,将 WCAQ 与现有量化方法进行直接比较:

MinMax 量化器(Jacob 等,2018):通过记录所有校准批次中激活值的全局最大值和最小值,静态地确定量化边界。

指数移动平均(exponential moving average, EMA)量化器(Jacob 等,2018):采用 EMA 策略动态更新量化范围。在本文实验中,平滑因子设定为 0.9。

EMAQuantile 量化器:将基于分位数(quantile)的校准方法与动态 EMA 平滑相结合。其截断边界由包含 2048 个面元(bin)的直方图结合 0.99999 的阈值计算得出,并通过 EMA 策略(平滑因子:0.9)进行动态调整。

MSE 量化器:通过反馈机制不断迭代以最小化量化误差,从而确定最优量化参数。该优化过程共包含 80 次迭代,步长衰减系数设为 0.01。

QRT-DETR(Huo 等,2026):在两阶段重构框架内结合使用了 EMA 与 MSE 量化器。其重构过程采用自适应矩估计(adaptive moment estimation, Adam)优化器,在余弦退火学习率调度策略的控制下,以 4×10^{-5} 的初始学习率进行 20000 次迭代,从而最小化 MSE 损失。

为确保对比的公平性,除特别说明外,所有对比方法均采用了两阶段重构感知策略(Reconstruction-

aware strategy),该策略在恢复量化后检测精度方面的有效性已得到广泛认可(Huo 等,2026)。唯一的例外是 QRT-DETR*,该模型在评估时未引入此重构步骤。不同位宽下的详细性能对比结果如表 1 所示。

表 1 不同量化方法在 RT-DETR 上的 AP 对比 (%)
Table 1 AP comparison of various quantization methods on RT-DETR (%)

量化方法	量化位宽 8/ 8	量化位宽 6/ 6	量化位宽 4/ 4
MinMax	45.9	2.5	0.0
EMA	46.1	2.9	0.0
EMAQuantile	46.2	31.7	0.0
MSE	46.2	43.6	22.2
QRT-DETR*	46.0	34.6	0.8
QRT-DETR	46.3	44.0	22.6
WCAQ	46.2	44.8	28.0

注:加粗数据表示同一位宽下各比较方法的最高 AP;“*”表示未执行两阶段重构。

由表 1 可知,不同方法在 8 位位宽条件下均能较好地保持 RT-DETR 的检测精度,WCAQ 取得 46.2% 的 AP,表明量化误差在较高位宽下尚未对模型性能产生明显影响。当位宽降至 6 位时,各方法之间开始出现性能差异,WCAQ 的 AP 为 44.8%,分别高于 MSE 和 QRT-DETR 的 43.6% 和 44.0%。当位宽进一步降至 4 位时,MSE 和 QRT-DETR 的 AP 分别下降至 22.2% 和 22.6%,而 WCAQ 仍取得 28.0% 的 AP,较二者分别提高 5.8% 和 5.4%。由此可见,随着量化位宽降低,WCAQ 相对现有方法的性能优势逐渐扩大,尤其在量化误差更易累积的 4 位位宽条件下表现更为明显。

上述结果表明,高位宽量化产生的数值扰动相对有限,因而不同量化策略之间的性能差异较小。在极低比特条件下,权重与激活的离散误差不断增大,并在网络前向传播过程中相互耦合,使仅依赖静态量化尺度或局部重构的方法难以维持检测精度。WCAQ 通过分组权重量化降低组内权重离散误差,根据样本特征分布调整激活量化尺度,并利用逐层重构补偿双侧量化后的联合误差,使权重侧误差控制、激活侧分布适配和层间误差补偿形成连续的处

理过程。因此,WCAQ 在 6 位位宽条件下较 QRT-DETR 提高 0.8%,而在 4 位位宽条件下将这一优势扩大至 5.4%,说明所提方法对极低比特量化引起的误差耦合具有更强的抑制能力。

在检测精度得到改善的同时,WCAQ 也降低了量化重构过程的时间开销。QRT-DETR 采用约 20000 次迭代的两阶段重构,耗时约 110 分钟。WCAQ 将每个量化单元的重构迭代次数设置为 200 次,整个量化重构过程耗时约 19 分钟。与 QRT-DETR 相比,WCAQ 的耗时减少约 82.7%,同时在 4 位位宽条件下获得 5.4% 的 AP 提升。这说明 WCAQ 并非通过增加优化迭代次数换取检测精度,而是利用权重、激活与重构过程之间的协同设计,在较低计算开销下改善极低比特量化性能,从而提高了方法在快速量化和实际部署场景中的适用性。

为进一步考察所提方法在现有目标检测量化研究中的相对表现,本文汇总了不同检测模型上的代表性量化结果,如表 2 所示。

表 2 显示,YOLOv5s 在 4 位下 Q-YOLO 仅余 14.0% 的 AP,而 RT-DETR 在 WCAQ 下仍保持 28.0%,这表明 WCAQ 在极低比特下具备更强的检测精度保持能力。

3.3 消融实验

为验证 WCAQ 各组成模块的必要性,本文围绕三个问题开展消融实验:一是权重与激活联合量化是否会引入显著的联合误差;二是逐层重构在中低比特条件下是否能够有效补偿联合误差;三是激活量化为何应置于算子输出端而非输入端。相关实验均基于 RT-DETR 基线模型,在统一量化封装、相同校准流程和一致超参数设置下完成。

3.3.1 权重与激活联合量化的必要性分析

为分析 RT-DETR 对不同量化扰动来源的敏感性,本文设置 FP32、仅权重量化(weight-only quantization, W-Only)、仅激活量化(activation-only quantization, A-Only)和 WCAQ 四组对比方案。其中,FP32 为全精度模型,W-Only 为仅量化权重,A-Only 为仅量化激活,WCAQ 为同时对权重与激活进行量化,并在此基础上执行逐层重构。实验结果见表 3。

从表 3 可见,在 8 位位宽条件下,仅权重量化、仅激活量化和联合量化之间差异较小,说明高位宽下量化误差尚未显著累积。但当位宽下降至 4 位位宽时,W-Only、A-Only 和 WCAQ 的 AP 分别为 39.5%、

表 2 不同模型上各量化方法的 AP 对比 (%)
Table 2 AP comparison of quantization methods on different models (%)

量化模型	量化方法	量化位宽	AP
YOLOv5s	FP	32/32	37.4
	MinMax(Jiang 等, 2025)	8/8	26.6
	Percentile(Jiang 等, 2025)	8/8	25.9
	MSQuant(Jiang 等, 2025)	8/8	36.8
	MSQuant(Jiang 等, 2025)	6/6	30.6
	Q-YOLO(Wang 等, 2023)	6/6	35.9
	Q-YOLO(Wang 等, 2023)	4/4	14.0
RT-DETR	FP	32/32	46.4
	WCAQ	8/8	46.2
	WCAQ	6/6	44.8
	WCAQ	4/4	28.0

注:加粗数据表示本文方法的结果。

表 3 不同量化策略在 RT-DETR 模型上的性能比较
Table 3 Performance comparison of different quantization strategies on RT-DETR

量化方法	量化位宽	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
FP	32/32	46.4	63.7	50.3	28.4	49.7	62.9
W-Only	8/8	46.3	63.7	50.2	28.3	49.6	62.7
A-Only	8/8	46.3	63.5	50.2	28.5	49.4	62.7
WCAQ	8/8	46.2	63.5	50.2	28.2	49.4	62.4
W-Only	6/6	45.5	62.8	49.2	28.0	48.9	61.5
A-Only	6/6	45.5	62.6	49.3	27.3	48.5	62.0
WCAQ	6/6	44.8	61.7	48.4	27.4	47.8	60.6
W-Only	4/4	39.5	56.0	42.5	23.1	42.5	54.8
A-Only	4/4	35.5	50.0	38.5	18.6	38.2	50.8
WCAQ	4/4	28.0	40.8	30.2	15.1	30.4	39.4

35.5% 和 28.0%。这表明,在极低比特条件下,权重量化误差与激活量化误差并非相互独立,而会在前向传播过程中发生耦合并持续累积。

其原因可由下式说明。设第 l 层浮点输出为

$$\mathbf{y}^{(l)} = \mathbf{W}^{(l)} \mathbf{x}^{(l)} \tag{9}$$

若仅进行权重量化,则有

$$\hat{\mathbf{y}}^{(l)} = (\mathbf{W}^{(l)} + \Delta \mathbf{W}^{(l)}) \mathbf{x}^{(l)} = \mathbf{y}^{(l)} + \Delta \mathbf{W}^{(l)} \mathbf{x}^{(l)} \tag{10}$$

若仅进行激活量化,则有

$$\hat{\mathbf{y}}^{(l)} = Q_A(\mathbf{y}^{(l)}) = \mathbf{y}^{(l)} + \Delta \mathbf{a}^{(l)} \tag{11}$$

而在联合量化条件下,

$$\hat{\mathbf{y}}^{(l)} = Q_A((\mathbf{W}^{(l)} + \Delta \mathbf{W}^{(l)}) \mathbf{x}^{(l)}) \tag{12}$$

式中, $\mathbf{x}^{(l)}$ 为第 l 层输入, $\mathbf{W}^{(l)}$ 为该层浮点权重, $\mathbf{y}^{(l)}$ 为对应浮点输出, $\Delta \mathbf{W}^{(l)}$ 为权重量化引入的权重扰动, $\Delta \mathbf{a}^{(l)}$ 为激活量化引入的输出扰动。

此时激活量化作用的对象已不再是原始浮点输出 $\mathbf{y}^{(l)}$, 而是受到权重量化扰动后的中间响应, 因此联合量化误差不仅包含权重侧误差和激活侧误差, 还包含由特征分布变化引起的交互项。结合表 3 的结果可以看出, WCAQ 所要解决的核心问题并不是单侧量化本身, 而是极低比特条件下双侧误差的耦合放大。

为了更加直观地揭示上述双侧误差耦合与累积现象的底层根源, 图 5 和图 6 进一步展示了 RT-DETR 模型在不同量化位宽下权重与激活的均值差异分布对比情况, 结合这两张图可以清晰地观察到, 相较于权重侧较小的误差波动区间, 激活特征不仅对量化扰动更为敏感, 且随着位宽从 8 位压缩至 4 位, 其误差分布呈现出极其显著的重尾效应与剧烈的数值偏移。这一底层数据分布特性直接印证了公式推导中关于误差交互放大的结论。这也进一步凸显了本研究在激活侧引入基于样本统计的自适应动态尺度调节机制, 以及后续执行逐层重构补偿的必要性。

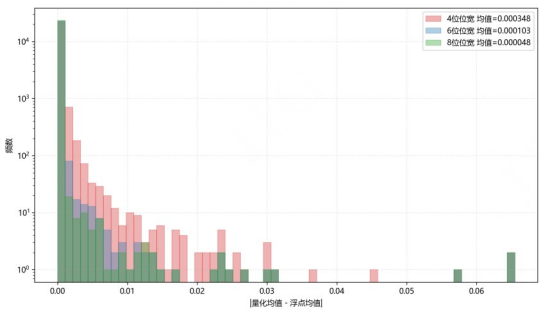


图5 量化前后权重均值差异分布对比图
Fig. 5 Comparison of mean difference distributions for weights before and after quantization

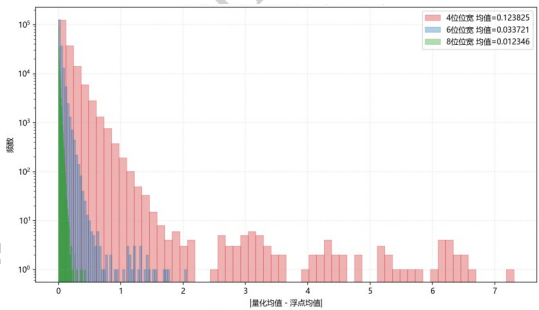


图6 量化前后激活均值差异分布对比图
Fig. 6 Comparison of mean difference distributions for activations before and after quantization

3.3.2 逐层重构的有效性分析

为验证逐层重构的必要性,本文在保持权重量化与激活量化配置一致的前提下,分别比较不同位宽下引入重构前后的检测性能,结果见表4。

表4表明,随着位宽降低,重构带来的增益持续增大。在4位位宽条件下,若关闭重构,AP将由28.0%降至25.1%,说明双侧量化后的中间特征偏移已成为性能退化的重要来源。

上述结果表明,重构并非可有可无的附加步骤,而是WCAQ在中低比特条件下保持检测性能的重要组成部分。在8位位宽条件下,由于量化误差本身较小,权重侧处理与激活侧动态适配已能够较好维持特征分布,因此重构的额外补偿空间有限。但当位宽降低至6位位宽和4位位宽时,权重量化误差与激活量化误差在前向传播过程中持续累积,导致量化层输出逐渐偏离浮点模型对应层的表示,最终反映为检测精度下降。

该过程并不改变原始网络结构,而是在完成权重量化和激活量化后,对联合误差进行逐层补偿。因此,逐层重构并非可有可无的附加步骤,而是

表4 不同位宽下是否引入重构对WCAQ检测性能的影响
Table 4 Impact of reconstruction on the detection performance of WCAQ under different bit-widths

是否 重构	量化 位宽	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
FP	32/32	46.4	63.7	50.3	28.4	49.7	62.9
Yes	8/8	46.2	63.5	50.2	28.2	49.4	62.4
None	8/8	46.1	63.4	50.1	27.9	49.3	62.4
Yes	6/6	44.8	61.7	48.4	27.4	47.8	60.6
None	6/6	44.3	61.1	48.0	26.7	47.0	60.6
Yes	4/4	28.0	40.8	30.2	15.1	30.4	39.4
None	4/4	25.1	37.2	27.0	13.5	28.0	36.4

WCAQ在中低比特条件下保持检测性能的重要组成部分。

3.3.3 激活量化位置有效性分析

为分析激活量化插入位置对模型性能的影响,本文在保持权重量化方式、激活量化方式、校准过程和重构过程一致的前提下,仅改变激活量化在网络中的插入位置,分别比较输入端量化与输出端量化在不同位宽下的检测性能,结果见表5。

表5 不同激活量化位置对WCAQ在RT-DETR上检测性能的影响

激活量 化位置	量化 位宽	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
FP	32/32	46.4	63.7	50.3	28.4	49.7	62.9
Output	8/8	46.2	63.5	50.2	28.2	49.4	62.4
Input	8/8	40.8	56.9	44.2	21.6	45.7	58.3
Output	6/6	44.8	61.7	48.4	27.4	47.8	60.6
Input	6/6	16.7	25.9	17.9	8.4	22.1	24.8
Output	4/4	28.0	40.8	30.2	15.1	30.4	39.4
Input	4/4	0.3	0.6	0.3	0.0	0.4	0.5

从表5可以看出,输出端量化在各位宽下均显著优于输入端量化,且位宽越低差异越明显。尤其在4位位宽条件下,输入端量化的AP仅为0.3%,而输出端量化仍可保持28.0%的AP,说明激活量化位置对RT-DETR的低比特稳定性具有决定性影响。

上述结果表明,在本方法中,将激活量化置于输
© 中国图象图形学报版权所有

出端是必要的。其原因在于,输出端量化对应的是

$$\hat{\mathbf{y}}_{\text{out}} = Q_A^{\text{out}}(\hat{\mathbf{W}}\mathbf{x} + \mathbf{c}) \quad (13)$$

即先利用量化权重完成线性变换,再对该层输出特征进行离散化,为简化表达,以下省略层索引。 Q_A^{out} 和 Q_A^{in} 分别表示作用于当前层输出和输入的激活伪量化算子, $\hat{\mathbf{y}}_{\text{out}}$ 和 $\hat{\mathbf{y}}_{\text{in}}$ 为对应的层输出, $\hat{\mathbf{W}}$ 为经量化和反量化得到的权重, $\mathbf{c}$ 为当前层的偏置向量;而输入端量化对应的是

$$\hat{\mathbf{y}}_{\text{in}} = \hat{\mathbf{W}}Q_A^{\text{in}}(\mathbf{x}) + \mathbf{c} \quad (14)$$

二者虽然都属于权激联合量化,但误差传播方式并不相同。对于输入端量化,激活离散化会直接作用于该层输入特征,输入端激活量化在卷积或线性运算之前引入离散化误差,该误差随后经过当前层权重映射,并可能在特定方向上被放大,其输出偏差可写为

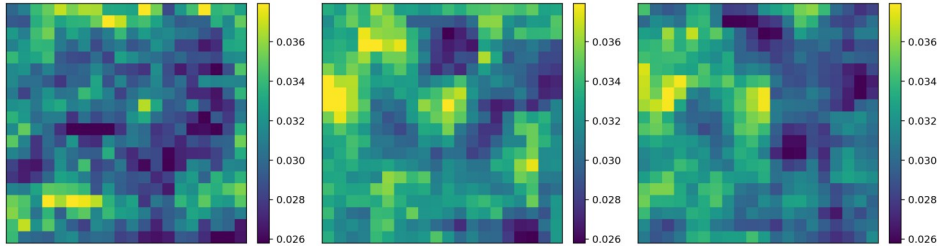
$$\Delta \mathbf{y}_{\text{in}} = \hat{\mathbf{y}}_{\text{in}} - (\hat{\mathbf{W}}\mathbf{x} + \mathbf{c}) = \hat{\mathbf{W}}(Q_A^{\text{in}}(\mathbf{x}) - \mathbf{x}) \quad (15)$$

相比之下,输出端量化能够保留该层输入特征的浮点表达,使核心映射过程在更完整的特征分布上完成,从而更有利于与后续动态尺度调节和逐层重构相配合。基于这一结果,本文最终采用输出端激活量化作为 WCAQ 的标准实现路径。

3.4 可视化分析

为进一步分析 WCAQ 在不同位宽条件下的特征保持能力,本文对中间特征层的响应分布进行了可视化,如图 7 所示。总体来看,随着量化位宽由 8 位位宽逐步降低至 4 位位宽,特征图中的响应强度与局部细节出现了一定程度的衰减,这是低比特量化带来的正常现象。然而,从空间分布上看,4 位位宽条件下的主要高响应区域并未完全消失,而是仍然与 6 位位宽和 8 位位宽条件下保持较为一致的结构位置关系。这说明 WCAQ 在极低比特条件下虽然对特征表达造成了一定压缩,但并未破坏网络对关键语义区域的基本表征能力。

这一现象表明 WCAQ 的低比特退化呈现出相对平稳的渐进特征,而非突发性的特征崩塌。换言之,在位宽持续降低的过程中,模型的中间表示虽然逐渐变得更加平滑,但仍能够维持对目标相关区域的有效响应。因此,即使在 4 位位宽条件下, WCAQ 仍保留了一定的检测判别基础。这也从可视化角度说明,所提出方法不仅能够显著压缩模型表示位宽,降低存储开销,而且在极低比特部署场景下仍具有一定的特征保持能力,从而体现出量化压缩与精度保持之间较好的平衡。



(a) 4 位位宽 (b) 6 位位宽 (c) 8 位位宽

((a) 4 bit; (b) 6 bit; (c) 8 bit)

图 7 WCAQ 在不同位宽下中间特征层响应分布的可视化结果

Fig. 7 Visualization of response distributions in intermediate feature layers across different bit-widths using WCAQ

4 结 语

本文面向 RT-DETR 在低比特部署中的联合量化误差累积问题,提出了一种权重激活协同自适应量化框架 WCAQ。与仅关注单侧量化误差的后训练量化方法不同, WCAQ 通过分组权重量化、动态激活适配以及逐层协同重构的联合设计,对双侧误差传播路径进行了统一约束与补偿,从而在无需重新训

练的条件下显著提升了 RT-DETR 在极低比特条件下的检测鲁棒性。研究表明,通过结合激活感知的分组权重量化、样本级动态激活适配以及耦合协同重构机制, WCAQ 能够在无需重新训练的条件下,有效抑制极低位宽下的量化噪声传播并保持目标表征能力。本研究深化了对检测 Transformer 中权重激活协同误差耦合机制的理论理解,并为复杂视觉模型的高效边缘部署提供了切实的工程方案。然而,当前框架对极敏感模块仍保留了浮点计算,尚未实现

全网络的纯整型部署,且量化效果对校准数据集分布具一定依赖性。未来研究将探索更细粒度的敏感路径保护机制与自注意力低比特策略,并重点将该协同量化范式拓展至无人机视觉感知等高动态实际应用场景中。WCAQ在模型压缩效率与检测精度之间取得了关键平衡,为实时边缘智能的广泛应用提供了重要支撑。

参考文献(References)

- Bondarenko Y, Nagel M and Blankevoort T. 2021. Understanding and overcoming the challenges of efficient transformer quantization//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics: 7947-7969 [DOI: 10.18653/v1/2021.emnlp-main.627]
- Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Computer Vision - ECCV 2020. Cham, Switzerland: Springer International Publishing: 213-229 [DOI:10.1007/978-3-030-58452-8_13]
- Frantar E, Ashkboos S, Hoefler T and Alistarh D. 2023. OPTQ: accurate quantization for generative pre-trained transformers//The Eleventh International Conference on Learning Representations[C/OL]. [2026-09-09]. <https://openreview.net/forum?id=tcbBPnfwS>
- Hu J, Zheng J H, Wan W W, Zhou Y Q and Huang Z K. 2025. RT-DETR-EVD: an emergency vehicle detection method based on improved RT-DETR. *Sensors*, 25 (11): 3327 [DOI: 10.3390/s25113327]
- Huang Z Z, Du H M and Chang L B. 2021. Mixed-clipping quantization for convolutional neural networks. *Journal of Computer-Aided Design & Computer Graphics*, 33(4): 553-559 (黄钰喆, 杜慧敏, 常立博. 2021. 卷积神经网络混合截断量化. *计算机辅助设计与图形学学报*, 33(4): 553-559) [DOI: 10.3724/SP.J.1089.2021.18509]
- Huo Y, Wu T L, Shen Y X, Li X M, Tao Z and Yang D W. 2026. QRT-DETR: post-training quantization for real-time detection transformer. *Neurocomputing*, 661: 131957 [DOI: 10.1016/j.neucom.2025.131957]
- Jacob B, Kligys S, Chen B, Zhu M L, Tang M, Howard A, et al. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society: 2704-2713 [DOI: 10.1109/CVPR.2018.00286]
- Jiang Z S, Li C, Qu T, He C and Wang D W. 2025. MSQuant: efficient post-training quantization for object detection via migration scale search. *Electronics*, 14 (3): 504 [DOI: 10.3390/electronics14030504]
- Jin T and Hu P Y. 2025. HK-DETR: improved knife-holding dangerous behavior detection algorithm based on RT-DETR. *Journal of Image and Graphics*, 30(4): 1027-1040 (金涛, 胡配雨. 2025. 改进实时目标检测 Transformer 的持刀危险行为检测算法. *中国图象图形学报*, 30(4): 1027-1040) [DOI:10.11834/jig.240295]
- Li Y D, Guo Z B, Liu K and Sun X Y. 2022. Mixed-precision quantization for neural networks based on error limit (invited). *Infrared and Laser Engineering*, 51(4): 20220166 (李奕铎, 郭子博, 刘凯, 孙逍遥. 2022. 基于误差限制的神经网络混合精度量化方法(特邀). *红外与激光工程*, 51(4): 20220166 [DOI: 10.3788/IRLA20220166]
- Li Y H, Gong R H, Tan X, Yang Y, Hu P, Zhang Q, et al. 2021. BRECQ: pushing the limit of post-training quantization by block reconstruction//International Conference on Learning Representations [C/OL]. [2026-09-09]. <https://openreview.net/forum?id=POWv6hDd9XH>
- Lin J, Tang J M, Tang H T, Yang S, Chen W M, Wang W C, et al. 2024. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration//Proceedings of Machine Learning and Systems[C/OL], 6: 87-100 [2026-09-09]. https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html
- Lin T Y, Maire M, Belongie S J, Hays J, Perona P, Ramanan D, et al. 2014. Microsoft COCO: common objects in context//Computer Vision - ECCV 2014. Cham, Switzerland: Springer: 740-755 [DOI:10.1007/978-3-319-10602-1_48]
- Liu G Y, Yang D R, Ye J, Lu H J, Wang Z and Zhao Y. 2025. A real-time welding defect detection framework based on RT-DETR deep neural network. *Advanced Engineering Informatics*, 65: 103318 [DOI:10.1016/j.aei.2025.103318]
- Lv W Y, Zhao Y A, Chang Q Y, Huang K, Wang G Z and Liu Y. 2024. RT-DETRv2: improved baseline with bag-of-freebies for real-time detection transformer[EB/OL]. [2026-03-15]. <https://arxiv.org/abs/2407.17140>
- Wang M Z, Sun H X, Shi J, Liu X H, Cao X B, Zhang L P, et al. 2023. Q-YOLO: efficient inference for real-time object detection//Pattern Recognition. Cham, Switzerland: Springer Nature Switzerland: 307-321 [DOI:10.1007/978-3-031-47665-5_25]
- Wang R Q, Sun H X, Yang L L, Lin S H, Liu C J, Gao Y, et al. 2024. AQ-DETR: low-bit quantized detection transformer with auxiliary queries//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press: 15598-15606 [DOI: 10.1609/aaai.v38i14.29487]
- Wang S, Xia C L, Lv F and Shi Y F. 2025. RT-DETRv3: real-time end-to-end object detection with hierarchical dense positive supervision//Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE: 1628-1636 [DOI:10.1109/WACV61041.2025.00166]

- Wei X Y, Gong R H, Li Y H, Liu X L and Yu F W. 2023. QDrop: randomly dropping quantization for extremely low-bit post-training quantization[EB/OL]. [2026-03-15].
<https://arxiv.org/abs/2203.05740>
- Xiao G X, Lin J, Seznec M, Wu H, Demouth J and Han S. 2023. SmoothQuant: accurate and efficient post-training quantization for large language models//Proceedings of the 40th International Conference on Machine Learning [C/OL]. PMLR, 202: 38087-38099 [2026-09-09]. <https://proceedings.mlr.press/v202/xiao23c.html>
- Xu S, Li Y J, Lin M B, Gao P, Guo G D, Lü J H, et al. 2023. Q-DETR: an efficient low-bit quantized detection transformer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA: IEEE Computer Society: 3842-3851 [DOI: 10.1109/CVPR52729.2023.00374]
- Yang C, Zhang R Y, Huang L, Ti S T, Lin J H, Dong Z W, et al. 2023. A survey of quantization methods for deep neural networks. Chinese Journal of Engineering, 45(10): 1613-1629 (杨春, 张睿尧, 黄珑, 遑书童, 林金辉, 董志伟, 等. 2023. 深度神经网络模型量化方法综述. 工程科学学报, 45(10): 1613-1629 [DOI: 10.13374/j.issn2095-9389.2022.12.27.004])
- Yang D W, He N, Hu X, Yuan Z H, Yu J Y, Xu C, et al. 2024. Post-training quantization for re-parameterization via coarse & fine weight splitting. Journal of Systems Architecture, 147: 103065 [DOI:10.1016/j.sysarc.2024.103065]
- Yang J, Li C, Hu Q H, Chen X D, Wang Y P and Liu X J. 2025. A post-training quantization method for lightweight CNNs. Journal of Graphics, 46(4): 709-718 (杨杰, 李琮, 胡庆浩, 陈显达, 王云鹏, 刘晓晶. 2025. 面向轻量卷积神经网络的训练后量化方法. 图学学报, 46(4): 709-718 [DOI: 10.11996/JG.j.2095-302X.2025040709])
- Yang L Y, Cheng Y H, Xu F, Li B N and Li X X. 2024. Real-time smoke detection in surveillance videos using an enhanced RT-DETR framework with triplet attention and HS-FPN. Fire, 7(11): 387 [DOI:10.3390/fire7110387]
- Zhao Y A, Lv W Y, Xu S L, Wei J M, Wang G Z, Dang Q Q, et al. 2024. DETRs beat YOLOs on real-time object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society: 16965-16974 [DOI:10.1109/CVPR52733.2024.01605]
- Zhou W Q, Yu Z D, Liu X Y, Yang J M, Xiao R, Wang T, et al. 2026. ASQ & POST: a synergistic framework for adaptive and non-uniform quantization. Neurocomputing, 666: 132332 [DOI: 10.1016/j.neucom.2025.132332]
- Zhu X Z, Su W J, Lu L W, Li B, Wang X G and Dai J F. 2021. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL]. [2026-03-15].
<https://arxiv.org/abs/2010.04159>
- Zhuo W Q, Li D Z, Wang W and Dong J. 2023. Data-free model compression for light-weight DeepFake detection. Journal of Image and Graphics, 28(3): 820-835 (卓文琦, 李东泽, 王伟, 董晶. 2023. 面向轻量级深度伪造检测的无数据模型压缩. 中国图象图形学报, 28(3): 820-835 [DOI:10.11834/jig.220559])

作者简介

第一/

霍瑛,女,副教授,主要研究方向为深度学习、神经网络量化、智能计算。E-mail: huoy@njit.edu.cn