计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100059

面向 BEVFormer 的高效训练后平衡量化策略

张晓玄　唐小勇
长沙理工大学计算机与通信工程学院　长沙 410114
(1098206574@qq.com)

摘　要　通过鸟瞰全景视角,BEVFormer 在自动驾驶领域展现出卓越的性能。然而,在资源受限的设备上,其高内存占用和计算复杂度为实时部署带来了严峻的挑战。BEVFormer 中 ReLU 激活值的分布从零到正无穷,呈现出不均匀的特点,传统量化指标如余弦相似度和均方误差(MSE)无法充分描述这种特性。针对此问题,提出了一种训练后平衡量化策略,该策略专门针对 BEVFormer 中的线性层和 ReLU 激活值的量化进行了优化。在线性层权重和输出的量化中采用预定义量化区间,同时对 ReLU 激活值使用特定区间量化方法,以确保关键值的精确表示。此外,该方法基于 Hessian 矩阵优化技术实现缩放因子动态调整,利用 Hessian 矩阵最小化量化误差,并稳定训练过程。实验结果显示,平衡量化策略显著提升了计算效率,同时保证了精度。在 nuScenes 测试集中,8 位量化仅导致 NDS 下降不到 1 个百分点,保持了 BEVFormer 的性能表现。

关键词:BEVFormer;ReLU 激活值;线性层权重;平衡量化策略;Hessian 矩阵
中图分类号　TP391

Balanced Quantization Strategy for Efficient Post-training Quantization of BEVFormer

ZHANG Xiaoxuan and TANG Xiaoyong
School of Computer and Communications Engineering,Changsha University of Science & Technology,Changsha 410114,China

Abstract　BEVFormer's bird's-eye view(BEV) representation achieves strong results in autonomous driving applications. However,its high memory use and computational demands make real-time deployment difficult on resource-constrained devices. BEVFormer's ReLU activation values vary widely,creating an uneven distribution that traditional quantization metrics,such as cosine similarity and mean square error(MSE),struggle to address effectively. To overcome these limitations,this paper introduces a new post-training quantization(PTQ) method,the Balanced Quantization Strategy. This method is specifically optimized for BEVFormer,focusing on quantizing linear layers and ReLU activations. For linear layers,it uses predefined quantization ranges,while ReLU activations are quantized with customized ranges to retain key value accuracy. Further,Hessian matrix optimization dynamically adjusts scaling factors,reducing quantization errors and stabilizing the quantization process. Results show that the Balanced Quan-tization Strategy improves computational efficiency with minimal accuracy loss. In testing on the nuScenes dataset,the proposed 8-bit quantization method achieves less than a 1% drop in NDS,maintaining BEVFormer's high performance.

Keywords　BEVFormer,ReLU activation,Outputs of the linear layers,Balanced Quantization Strategy,Hessian matrix

1　引言

随着社会经济的迅猛发展,BEV 感知在自动驾驶领域已成为最受关注的技术之一[1-2]。BEVFormer 模型通过多摄像头生成的 BEV 视图,能够为车辆提供丰富且详尽的环境感知信息[3-5]。这一模型借助于 Transformer 架构中自注意力和交叉注意力机制,有效实现了对 3D 场景的解析,在视觉感知任务中展现出极强的性能[3,6],为车辆导航和路径规划提供了高精度的支持。然而,BEVFormer 的高计算复杂度和显著的能耗需求,使得其在资源受限的设备(例如嵌入式系统)上实现实时部署变得尤为困难。这种高资源需求不仅带来了硬件成本的增加,也对实际应用中的功耗和延迟提出了严峻挑战,极大地限制了该技术在低功耗平台上的普及应用。

为应对 3D 感知技术中存在的资源限制问题,研究者开始探索各种量化技术,特别是后训练量化(PTQ)方法[7-10],以降低模型的计算复杂度和内存占用,从而实现高效、低功耗的部署。量化技术通过降低模型参数和激活值的比特宽度(如 8 位量化)[11-12],在卷积神经网络的压缩中取得了显著的成效,能够在有限资源条件下维持较高的模型性能。然而,当这种方法直接应用于 Transformer 模型时,往往面临显著的精度下降和稳定性问题[13]。BEVFormer 模型中 ReLU 激活值的分布极不均匀,通常集中在零附近,而少数较大的值则对模型性能至关重要。传统的量化方法未能充分考虑这些特征,导致简单的 8 位量化可能造成性能显著下降,NDS 分数下降超过 5 个百分点,从而限制了 BEVFormer 在自动驾驶系统中的实际应用。

为了应对这一难题,本文提出了一种专门针对 BEVFormer 结构的平衡量化策略。该策略通过为 ReLU 激活值

基金项目:国家自然科学基金(61972146)
This work was supported by the National Natural Science Foundation of China(61972146).
通信作者:唐小勇(tangxy@csust.edu.cn)

设定特定的量化方法[14]，以确保重要值的准确表示。同时，线性层的权重和输出采用预定义的量化区间。此外，Hessian优化技术被用于动态调整量化过程中的缩放因子[15]，以减少误差并增强模型稳定性。实验结果显示，该策略显著提升了计算效率，在nuScenes测试集中，8位量化仅导致NDS性能下降不到1个百分点，保持了较高的训练精度。

2 相关工作

2.1 基于多摄像视觉的3D检测

在空中目标跟踪和自动驾驶等重要应用中，对准确有效的目标检测需求日益增长，推动了基于多摄像头视觉的3D检测领域取得重大进步。最近的研究集中在创造新技术，利用多模态数据显著提高检测能力。Liu等[6]提出了BEVFusion框架，该框架融合了鸟瞰(BEV)视角下的多模态特征，以提高3D检测精度。通过整合激光雷达和摄像头数据，BEVFusion可以更全面地理解周围环境。同样地，Jiao等[16]引入了MSMDFusion技术，在LiDAR和相机数据上采用多深度种子进行融合，并进一步提升了检测精度。另外，在Yin等[17]的突破性工作中，他们提出了多模态虚拟点3D检测方法，并建立起不同模态之间细粒度特征对齐关系，从而显著改善了3D物体检测性能。此外，在Bai等[18]开发的软关联机制方法中使用强大的激光雷达-相机融合策略来增强3D检测，并强调有效数据融合的关键作用。最后，Chen等[19]提出密集投影融合DPFusion，这是一种由密集深度图引导BEV变换模块和多模态特征自适应融合模块组成，具有多阶段、细粒度特点的融合技术，可将点云和图像对齐。这种方法解决了逐点融合技术的缺点，能够更准确地将图特征投影到BEV上，并在nuScenes数据集上获得最先进的结果。研究者在基于多相机视觉的三维目标识别领域采用了各种方法与技巧，然而，三维目标识别算法量化仍然是一个重大挑战。主要原因是激活值分布复杂以及传统量化指标无法保持较高识别准确率。解决这些问题对于在资源受限硬件上部署高效且准确率较高的三维目标识别算法至关重要。

2.2 训练后量化方法

对于深度神经网络而言，训练后量化已成为一种流行的方法，用于节省内存存储和计算费用，并且无需重新训练或访问整个训练数据集。为了提高训练后量化技术的精度和有效性，研究者提出了几种策略。Doe等[20]引入了分段线性量化PWLQ方案，以精确近似钟形分布和长尾的张量值。Nagel等开发了AdaRound[21]，这是一种自适应权重舍入机制，在多个网络和任务上取得了最先进的结果。Li等[22]通过引入BRECQ框架突破了训练后量化的极限，该框架可以将量化降低到INT2位宽，而无需端到端再训练。此外，Liu等针对视觉转换器成功提出了一种专门设计的训练后量化算法[15]，在基准模型和数据集上优于其他方法。Hubara等[23]通过在一个小的校准集上优化参数来最小化每个层或块的量化误差，并突显了训练后量化方法简单高效的特点。Bondarenko等[24]探索了高效变压器量化面临的挑战，并提出基于训练后量化和量化感知训练相结合的解决方案，在准确性、模型大小和易用性方面做出不同权衡选择。Frantar等[25]引入一种压缩框架将权值修剪与数量级设置相结合，在改进现有方法性能的同时实现精确压缩时代价最小。Yao等[12]提出ZeroQuant，这是一种高效且负担得起大规模变压器模型进行培养之后定额方式，可解决复杂模型挑战。Xiao等[26]提供SmoothQuant，这是一种非培养式学习之前定额方式，在语言建立背景下，允诺对语言大规格建立进行8位数重要价值及激活定额。平滑激活异常并把定额难度从激活移到重要价值，SmoothQuant在保持正确率的同时降低了计算成本。总体来说，随着创新技术与构造不断发展，这些技术和框架提高了深度神经网络量化方法的效率和有效性。

3 量化的预备工作

使用后训练量化(PTQ)可以将已训练的高精度模型转换为低位宽版本，无需重新训练，从而大大加快推理速度并减少存储需求。量化的基本思想是将浮点数权重和激活值映射到低精度整数值。量化过程可以描述为：

$$x_{\text{int}}=clamp\left(\frac{v}{\alpha}\right)+\beta \tag{1}$$

其中，α是缩放因子，β是零点，$clamp(\cdot)$将值限制在量化范围内。反量化通过$v\approx\alpha(x_{\text{int}}-\beta)$恢复近似的原始浮点值，而更一般的量化函数表示为：

$$\hat{v}=\alpha\left(clamp\left(\frac{v}{\alpha}\right)+\beta;0,2^k-1\right)-\beta \tag{2}$$

其中，k是位宽，决定了量化的精度。此外，当将量化应用于非线性激活函数如ReLU时，ReLU将负值设为零，量化后的ReLU函数可以表示为$\hat{y}=\max(0,\hat{v})$。这确保了ReLU激活在量化后仍然保持其行为，从而帮助维持网络的稀疏性和计算效率。平衡量化策略体系架构图如图1所示。

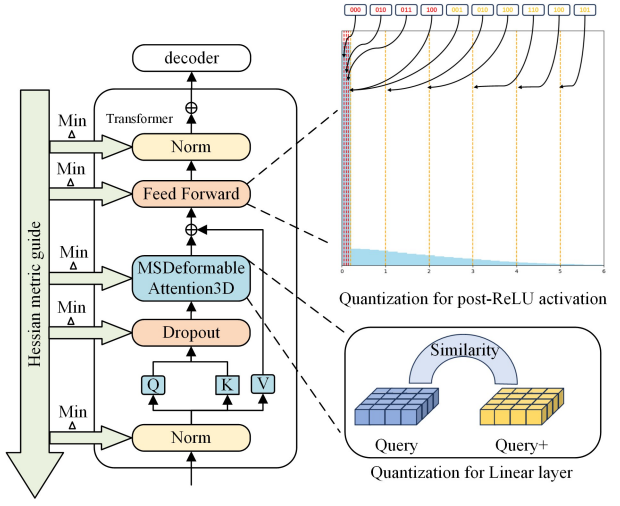


图1 平衡量化策略体系架构图

Fig. 1 Overview of the Balanced Quantization Strategy

PTQ的一个关键步骤是校准，在此阶段确定最佳量化参数α和β，通常通过最小化量化激活B^q与度激活B^r之间的误差来完成。

$$\arg\min_{\alpha,\beta} Metric(B^q,B^r) \tag{3}$$

其中，度量可以是均方误差(MSE)。最近，研究者通过对比学习来改进校准过程，通过最大化全精度激活和量化激活之间的互信息来增强校准效果。其目标是最大化原始激活和量化激活之间的互信息$I(A;B)$：

$$I(A;B)=\sum_{a,b}P_{A,B}(a,b)\log\left(\frac{P_{A|B}(a|b)}{P_A(a)}\right) \tag{4}$$

这确保了量化激活保留尽可能多的全精度信息，从而在量化后仍能保持模型的性能。

4 平衡量化策略

在 BEVFormer 模型中，对关键层进行量化对于在减少精度损失的同时保持高性能至关重要。主要挑战在于线性层和 ReLU 激活函数，这些层在模型的计算中扮演着重要角色。为此，本文提出了训练后平衡量化策略(BQS)，主要包括线性层量化、ReLU 激活函数量化和基于 Hessian 矩阵的量化缩放因子优化。

4.1 线性层量化操作

在 BEVFormer 模型的线性层量化过程中，通过预定义的量化间隔来对权重和输入进行去量化。这种方法受到了相似性感知量化技术的启发，其目的是确保量化后的输出尽可能地接近原始输出。具体而言，本文用余弦相似性取代了传统相似性感知量化方法中的皮尔逊相关系数。这一改变的原因在于，余弦相似性更加注重保持原始和量化特征图之间的方向一致性，而不仅仅是它们之间的线性相关性。

在原始的线性层中，输出可以表示为输入特征 $\boldsymbol{Z}_k$ 和权重矩阵 $\boldsymbol{Y}_k$ 的乘积，即：

$$P_k=\boldsymbol{Z}_k\cdot\boldsymbol{Y}_k \tag{5}$$

然而，在应用了预定义间隔的量化之后，去量化的输出则被表达为：

$$\hat{P}_k=\phi_k^{\Delta_z}(Z_k)\phi_k^{\Delta_y}(Y_k)\cdot\Delta_z^k\cdot\Delta_y^k \tag{6}$$

其中，$\hat{P}_k$ 代表量化后的输出，Δ_z^k 和 Δ_y^k 分别代表输入和权重的量化间隔。为了最大化量化输出和原始输出之间的相似性，研究者重新定义了相似性度量标准，采用余弦相似性来衡量：

$$sim(\hat{P}_k,P_K)=\frac{(\sum_{i=1}^{m}\hat{P}_k^{(i)})\cdot(\sum_{i=1}^{m}P_k{}^{(i)})}{\sqrt{(\sum_{i=1}^{m}(\hat{P}_k^{(i)})^2)}\cdot\sqrt{(\sum_{i=1}^{m}(P_k^{(i)})^2)}} \tag{7}$$

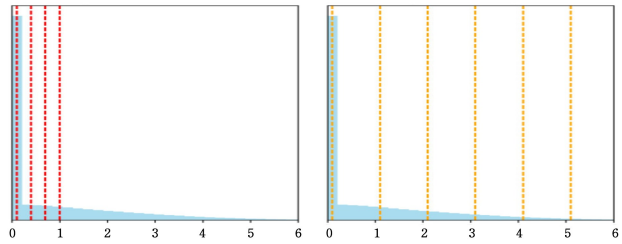
在这一方程中，重点在于优化量化间隔 Δ_z^k 和 Δ_y^k 以最大化去量化输出和原始输出间的余弦相似性。这种方法的改进使得在保持输出方向特征的同时，能够提高量化精度，确保在利用量化带来的计算优势的同时，性能的退化被降至最低。

为了找到最佳的量化间隔，通常需要一个校准数据集。这个数据集比训练数据集小得多，但它能够代表模型在实际应用中可能遇到的数据分布。通过在这个校准数据集上评估模型的输出，可以确定最佳的量化间隔，以最大化相似性度量。

通过这种方法，BEVFormer 模型能够在量化的过程中保持较高的性能，同时减少由量化带来的精度损失。这对于在资源受限的设备上部署深度学习模型尤为重要，因为它可以在不牺牲太多准确性的情况下，显著降低模型的计算复杂度和内存需求。这种平衡量化策略的应用，不仅提高了模型在实际应用中的可行性，也为未来在更广泛的设备上部署高效的深度学习模型提供了新的思路。

4.2 ReLU 激活函数量化操作

ReLU 激活函数在 BEVFormer 模型的多个层中得到了广泛应用。ReLU 的输出范围从零到正无穷大，大部分值聚集在零附近，而少数值则较大。这些较大的值对模型的性能至关重要，如果在量化过程中没有考虑它们，可能会导致明显的精度损失。因此使用不同的比例因子来描绘均匀量化的量化点，如图 2 所示。



(a) post-ReLU：small Δ　　(b) post-ReLU：large Δ

图 2 不同缩放因子用于量化 ReLU 激活后的值

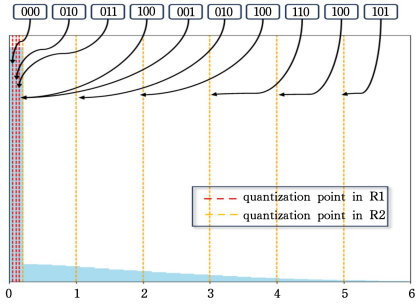
Fig. 2 Various scaling factors are demonstrated in order to quantize the post-ReLU activation levels

在 BEVFormer 模型的量化过程中，对于 ReLU 函数，采用了一种旨在精确量化接近零的小值和较大的激活值。首先，本文分析了 ReLU 激活函数的输出分布，并将其划分为两个量化范围。如图 3 所示，第一个范围 R_1 主要用于量化接近零的小值，为了保持这些小值的量化误差尽可能小，选择了一个小的缩放因子 ΔR_1，以便在低值范围内进行精细量化。第二个范围 R_2 用于量化较大的激活值。为了确保这些大值在量化后能被准确表示，选择了一个较大的缩放因子 ΔR_2，以确保这些值在量化后保持它们的关系和重要性。

量化过程可以定义为：

$$T(a_q,R_1,R_2)=\begin{cases}round\left(\dfrac{a}{\Delta R_1}\right), & \text{if } a\in R_1\\ round\left(\dfrac{a}{\Delta R_2}\right), & \text{if } a\in R_2\end{cases} \tag{8}$$

为了保持计算效率，缩放因子进一步被限制为 $\Delta R_2=2^m\Delta R_1$。其中 m 是一个整数。这种设计允许高效的位移动操作，改善了模型在推理期间的性能，而不需要复杂的乘法操作。



注：Annotate the pair of binary values for different quantization levels。

图 3 3 比特量化下的 ReLU 后值与二进制表示

Fig. 3 3-bit quantization illustrated on post-ReLU values and binary values

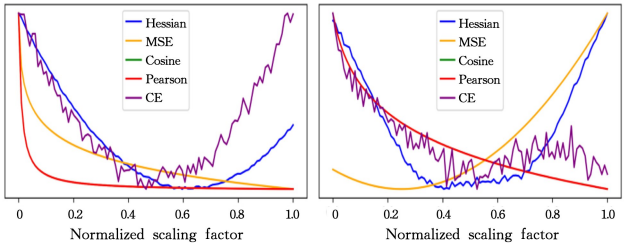
通过精心管理这两个量化范围，模型可以减少量化误差，确保小的和大的 ReLU 激活值都能被准确表示，从而在量化后保持 BEVFormer 的性能。这种平衡量化策略的应用，不仅提高了模型在实际应用中的可行性，而且为未来在更广泛的设备上部署高效的深度学习模型提供了新的思路。

4.3 基于 Hessian 矩阵的量化缩放因子优化

在量化 BEVFormer 模型的过程中，本文采用 Hessian 矩

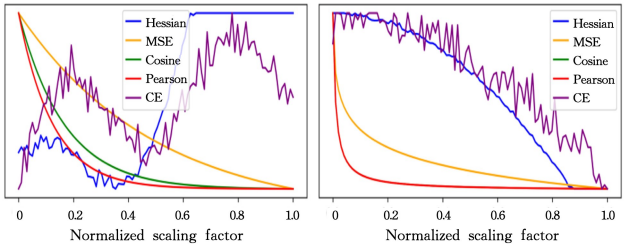
阵来精确选择 FFN 层的缩放因子，从而有效减少了量化误差。Hessian 矩阵作为一种基于二阶信息的技术，通过近似损失函数关于模型参数的 Hessian 矩阵来估计模型参数对量化的敏感度。这一方法使得研究者能够根据每一层或参数的敏感度有针对性地调整量化策略。

在量化过程中，BEVFormer 模型的不同缩放因子的性能表现如图 4 所示。均方误差(MSE)、余弦相似度和皮尔逊相关系数在确定最优缩放因子方面并不如基于任务损失(交叉熵)在视觉 Transformer 模型上的表现准确。这些指标的最优缩放因子与基于任务损失计算得到的最优缩放因子不一致。例如，在 encoder. layers. 0. ffns. 0. activate 层，MSE、余弦相似度和皮尔逊相关系数表明 $0.45\frac{A_{\max}}{2^{k-1}}$ 是最优的，而基于任务损失的最优缩放因子为 $0.7\frac{A_{\max}}{2^{k-1}}$。



(a) layers. 0. attentions. 1. deformable_attention. sampling_offsets

(b) layers. 0. attentions. 1. deformable_attention. value_proj



(c) layers. 1. attentions. TemporalSelfAttention. sampling_offsets

(d) layers. 0. attentions. TemporalSelfAttention. output_proj

图 4 BEVFormer 模型在不同缩放因子下的任务损失差异以及量化前后层输出的变化

Fig. 4 Difference of task loss in BEVFormer model under different scaling factors and the change of layer output before and after quantization

基于这些指标得到的缩放因子通常是次优的，导致模型准确性下降。BEVFormer 中的物体检测任务，损失函数定义为 $L(V)=L1Loss(\hat{z},z)$，其中 $\hat{z}$ 是预测输出，z 是真实值。在量化过程中，对原始权重 V 引入扰动 Δ，得到量化后的权重 。为了理解这些扰动是如何影响损失的，在未量化的权重 V 周围应用泰勒级数展开：

$$\mathbb{E}[\mathcal{L}(\tilde{V})]-\mathbb{E}[\mathcal{L}(V)]\approx\Delta^{\mathrm{T}}\boldsymbol{g}^{(V)}+\frac{1}{2}\Delta^{\mathrm{T}}\boldsymbol{H}^{(V)}\Delta \tag{9}$$

其中，$g^{(v)}$ 代表损失关于权重 V 的梯度，$\boldsymbol{H}^{(V)}$ 是 Hessian 矩阵，捕获了损失景观的二阶信息。目标是通过优化缩放因子 Δ 来最小化量化引起的扰动对损失的负面影响。这确保了量化模型在推理期间的准确性得到了保留。

为了精确优化量化过程，本文精心开发了一种缩放因子优化策略，通过利用 Hessian 矩阵。最初，每一层的输入激活值 A^l 和权重 B^l 被赋予缩放因子 Δ_{A^l} 和 Δ_{B^l}，这些因子最初通过启发式方法确定，随后在细致的迭代过程中进行优化。为了对抗这些缩放因子可能出现的剧烈波动，本文策略性地引入了一个正则化项，以实现稳定性。

$$\mathcal{R}(\Delta^l)=\lambda\sum_l(\Delta_{A^l}^2+\Delta_{B^l}^2) \tag{10}$$

这个正则化项作为一种稳定力量，通过惩罚缩放因子的过度偏差，从而维持其平衡。随后，每一次迭代都通过计算 Hessian 矩阵来协调更新缩放因子，Hessian 矩阵是一个关键的二阶导数矩阵，揭示了损失函数景观的曲率。对于输入激活和权重层，Hessian 矩阵描述如下：

$$H\{\Delta^{\wedge l}\}=\left\{\frac{\partial^2(\mathcal{L}(\Delta^l)+\mathcal{R}(\Delta^l))}{\partial\Delta^{\wedge|2}}\right\} \tag{11}$$

其中，Δ^l 表示与输入激活 Δ_{A^l} 或权重 Δ_{B^l} 相关的缩放因子。Hessian 矩阵在识别损失函数的局部行为中起着至关重要的作用，从而使优化过程更加精细。最终，缩放因子通过 Hessian 矩阵的逆来精心调整，以确保梯度下降步骤：

$$\Delta^l\leftarrow\Delta^l-H_{\Delta^l}^{-1}\frac{\partial(\mathcal{L}(\Delta^l)+\mathcal{R}(\Delta^l))}{\partial\Delta^l} \tag{12}$$

这种复杂的更新协议巧妙地修改了缩放因子，结合了损失和正则化项，最终目标是最小化总体目标函数。具体的实验代码如算法 1 所示。通过采用这种方法，确保量化过程保持坚定，模型保真度在最大程度上得以保留。

算法 1 基于 Hessian 矩阵的缩放因子优化

1. For l=1 to L do
2. 正向传播：$O^l\leftarrow A^lB^l$
3. End for
4. For l= L to 1 do
5. 利用反向传播计算梯度：$\frac{\partial\mathcal{L}}{\partial O^l}$
6. End for
7. For l=1 to L do
8. 初始化缩放因子：Δ_{A^l}，Δ_{B^l} $\left(\text{e.g. }\Delta_{B^l}^*\leftarrow\frac{B^{l\max}}{2^k-1}\right)$
9. 为 Δ_{A^l} 和 Δ_{B^l} 创建搜索空间
10. 添加正则化项以避免过拟合：使用式(9)
11. For r =1 to Round do
12. 利用 Hessian 矩阵计算缩放因子：使用式(10)
13. 更新：使用式(11)
14. 同样地，计算缩放因子：使用式(10)
15. 更新：使用式(11)
16. End for
17. End for

5 性能评价与实验结果分析

本章首先描述了实验设置；随后，介绍了针对不同类别的视觉变换器的各种方法；在最后部分对这些方法进行了消融研究。

5.1 实验设定

搜索空间 Δ_{R1}^s 被定义为在 $\left[\frac{1}{2^k},\frac{1}{2^{k+1}},\cdots,\frac{1}{2^{k+10}}\right]$ 范围内。相同的搜索空间被用于应用权重和激活的缩放因子。在实验中，设置 $\alpha=0.5$，$\beta=1.1$ 和 $n=64$，这提供了更细粒度的调整模型参数量化的能力。搜索轮数被设置为 4，以确保更好的收敛性。在 nuScenes 数据集上评估了本文量化方法，重点关

注目标检测任务对于校准，从训练数据集中随机选取了 50 张图片。

本工作中使用的 BEVFormer 模型是 GitHub 提供的，本文记录了所有全连接层的权重和输入，包括初始投影层和最终预测层。此外，还量化了自注意力模块中涉及矩阵乘法的两个输入矩阵。

5.2 不同量化方法性能评估

如表 1 所列，本文提出的平衡量化策略在 ViT 和 BEVFormer 模型上的表现均优于 PyTorch 的标准量化技术。以 ViT 模型为例，如 ViT-S(224/32)和 ViT-B(384)，在较低的 W6A6(权重/激活)值下，平衡量化提供了比 PyTorch 方法更好的结果。对于包括 BEVFormer-base 和 BEVFormer-tiny 在内的 BEVFormer 模型，情况也类似，平衡量化在保持高性能的同时，其在 NuScenes 检测分数(NDS)上的降低幅度小于 PyTorch 量化技术。这一情况意味着，平衡量化策略对于保持高质量性能具有决定性意义，特别是在以较低尺度量化模型时。

表 1 不同量化方法对模型性能影响

Table 1 Effect of different quantization methods on model performance

Model	Backbone	PyTorch Quantiztion		Balanced Quantization Strategy	
		W8V8	W6V6	W8V8	W6V6
ViT-S(224/32)	Nan	74.61(−1.38)	63.14(−18.8)	75.55(−0.38)	71.50(−3.58)
ViT-S(224)	Nan	80.46(−0.91)	70.24(−11.1)	81.00(−0.38)	78.63(−2.75)
ViT-B(224)	Nan	83.89(−0.64)	75.66(−8.87)	84.25(−0.29)	84.25(−0.29)
ViT-B(384)	Nan	85.35(−0.64)	46.88(−39.1)	85.82(−0.17)	83.34(−2.65)
BEVFormer-base	R101	47.9(−3.8)	40.1(−11.6)	51.3(−0.4)	50.3(−1.4)
BEVFormer-small	R101	41.3(−6.6)	35.4(−12.5)	47.4(−0.5)	46.4(−1.5)
BEVFormer-tiny	R50	46.4(−1.5)	25.1(−10.3)	35.2(−0.2)	34.9(−1.0)

5.3 BEV 模型不同训练后量化方法比较

表 2 详细评估了多种不同的后训练量化(PTQ)方法，这些方法应用于 BEVFormerV2 和 BEVFormer 模型，并涵盖了多种位宽设置。本文对不同的算法进行了考察，包括本文方法，它除了单独分析外，还与混合精度(MP)和偏置校正(BC)等附加技术一起分析。在 BEVFormerV2 的情况下，本文方法在各种位宽下都提高了 NDS 值。AI 模型在 W8A8 量化下生成了易于阅读的 52.6 分，实际上仅用 50 张校准图像进行训练。这远远超过了 EasyQuantLiu 的方法，它们分别产生了 42.2 和 44.3 的 NDS，但需要更多的训练图像(高达 1024 张)。实验表明，平衡量化策略 BQS 不仅在用较少的数据进行模型校准方面优于其他方法，而且在量化后保持模型性能方面也更加有效。这与 W6A6 配置的结果(BQS 方法产生了 45.4 的 NDS)相一致，超过了其他方法，其他方法中没有一个超过 40。

表 2 对于 BEVFormerV2 和 BEVFormer 模型不同量化方法的比较

Table 2 Comparison of different quantization methods for BEVFormerV2 and BEVFormer models

Model	Method	Bit-width	#ims	Size	NDS
BEVFormerV2 55.3	EasyQuant	W8A8	1024	22.0	42.2
	Liu	W8A8	1024	22.0	44.3
	Liu	W8A8(MP)	1024	22.2	45.0
	BQS	W8A8	50	22.0	52.6
	EasyQuant	W6A6	1024	16.5	36.9
	Liu	W6A6	1024	16.5	38.2
	Liu	W6A6(MP)	1024	16.6	39.3
	BQS	W6A6	50	16.5	45.4
BEVFormer 51.7	EasyQuant	W8A8	1024	86.0	41.8
	Liu	W8A8	1024	86.0	42.1
	Liu	W8A8(MP)	1024	86.8	42.7
	BQS	W8A8	50	86.0	51.3
	EasyQuant	W6A6	1024	64.5	43.1
	Liu	W6A6	1024	64.5	39.1
	Liu	W6A6(MP)	1024	64.3	39.5
	BQS	W6A6	50	64.5	50.3
	EasyQuant	W4A4(MP)	1024	43.6	34.2
	Liu	W4A4	32	43.0	31.8
	BQS	W4A4	50	43.0	32.6

BEVFormer 模型的 NDS 值在 W8A8 和 W6A6 情况下再次成为最佳方法，本文方法以 51.3 和 50.3 的 NDS 值超越了它。这表明 W8A8 和 W6A6 配置的优势。而替代方法在增加校准图像数量的情况下产生了较低的 NDS 数值。有趣的是，本文方法在 W4A4 量化性能上不如 Liu 的混合精度方法。而 Liu 的方法需要更激进的位宽缩放，这导致了精度的显著损失。尽管如此，包含偏置校正(BC)因子使得 NDS 得分为 32.6，这突出了 BC 在一定程度上补偿低位宽量化精度下降的重要性。

总体而言，所述结果证实了所提出的 PTQ 方法在保持模型性能方面的有效性，尤其是在使用较少校准图像的情况下，特别是在较高位宽组中。增强的性能可以归因于该方法能够利用其为量化精度量身定制的权衡，提高了模型对特定分布特征的准确性。此外，当结果受到低位宽设置的影响时，偏置校正等附加技术的益处也变得非常明显。因此，这表明 PTQ 方法在所有量化环境下仍然稳健。但如果与其他方法一起使用，以满足任务的具体需求，还需要进一步的改进。

5.4 量化后 BEVFormer 模型变量内存与参数优化

图 5 的数据表明，逐张量熵后训练量化(PTQ)有效地减少了内存使用，同时保持了原始 BEVFormer 模型的性能。以 BEVFormer 基础模型为例，可以观察到内存消耗从 2159 MB 降低到 1665 MB；然而，平均精度(mAP)和 NuScenes 检测分数(NDS)并没有显著降低。此外，BEVFormer 小型模型的内存从 4663 MB 减少到 3441 MB，对性能的影响微乎其微。BEVFormer 基础模型展示了这一点，它从 28500 MB 压缩到 11011 MB 时，仅在 NDS 和 mAP 上略有损失。使用 PTQ 进行内存优化已被证明是一种有效的方法，同时能够保持可接受的准确度水平。这些结果证实了熵逐张量的 PTQ 过程在整体上为 BEVFormer 模型提供了内存使用优化的显著优势，同时只有适度的准确度损失。这种低精度张量量化(LPTQ)将成为在内存效率是核心问题的情况下，使用高性能模型的有效方法。

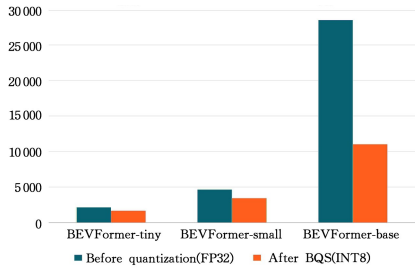


图 5 PTQ 对不同 BEVFormer 模型变体(tiny、small 和 base)的内存效率和性能的影响

Fig. 5 Effects of PTQ on memory efficiency and performance of different BEVFormer model variants(tiny, small and base)

通过图 6 在 BEVFormer 模型的不同变体(tiny、small 和 base)上进行实验评估,结果表明,所提出的量化方法显著降低了计算成本和内存需求,对模型准确性的影响很小。特别是,BEVFormer-base 模型的内存占用从 28 500 MB 降低到仅 11 011 MB,平均 NDS 仅从 0.80 轻微变化到 0.79,mAP 得分从 0.67 变化到 0.66。这表明所提出的方法在提升板载性能的同时,能保持模型完整性。此外,因组件(如"Img-neck"等主要部分)参数减少,给本文提出的平衡量化策略带来了更高计算效率。

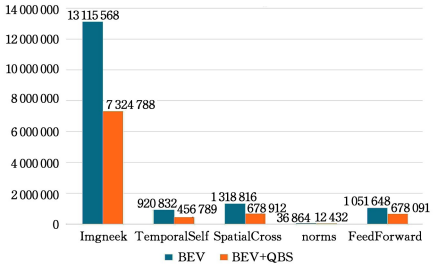


图 6 量化后 BEV 模型参数量对比

Fig. 6 Comparison of BEV model parameters after quantization

综上,可以得出结论,PTQ 技术不仅在保持 BEVFormer 模型性能的同时减少了内存使用,而且还通过减少模型参数数量提高了计算效率。特别是在 BEVFormer-base 模型中,内存占用从 28 500 MB 降低到 11 011 MB,而性能指标 NDS 和 mAP 的下降幅度非常小,这表明模型在资源受限的设备上具有应用潜力。此外,参数数量的减少也意味着模型更加轻量化,这对于需要快速响应和处理大量数据的应用场景尤为重要。总体而言,本文提出的平衡量化策略在 BEVFormer 模型内存优化和参数精简方面展现了显著性能,使其更适合在资源受限的环境中部署和使用。

结束语 本研究针对 BEVFormer 模型在量化过程中的独特挑战,提出了一种创新的平衡量化方法。将激活值分为两个量化范围,有效处理了 ReLU 激活的偏斜分布,确保了关键值的高精确度。基于权重相似性分析的策略,本文为线性层定义了预定义的量化区间,进一步减少了量化误差,保持了模型性能。此外,该方法引入了 Hessian 矩阵优化技术,以动态调整缩放因子,利用 Hessian 矩阵有效地最小化量化误差,从而提升模型稳定性并减少训练过程中的波动。实验结果表明,该方法在不牺牲准确性的前提下,显著提高了 BEVFormer模型的计算效率和内存使用,适用于资源受限的设备。本研究不仅验证了传统 PTQ 技术在处理偏斜分布时的局限性,也为未来在更广泛的深度学习模型量化领域提供了新的研究方向和改进思路。尽管本研究取得了一定的成果,但在量化精度与模型性能之间的权衡,以及在不同硬件平台上的适应性等方面仍有待进一步探索。未来的工作将聚焦于优化量化策略,以适应更多样化的应用场景,并提高模型在实际部署中的鲁棒性和效率。

参 考 文 献

[1] HUANG D Q, HUANG H F, HUANG D Y, et al. A Survey on BEV Perception Learning in Autonomous Driving [J/OL]. Computer Engineering and Applications, 1-23 [2024-11-06]. http://kns.cnki.net/kcms/detail/11.2127.TP.20241031.1529.016.html.

[2] MA Y X, WANG T, BAI X Y, et al. Vision-Centric BEV Perception: A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10978-10997.

[3] LI Z, WANG W, LI H, et al. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers[C]// European Conference on Computer Vision. Cham: Springer, 2022: 1-18.

[4] YANG C, CHEN Y, TIAN H, et al. Bevformer v2: Adapting modern image back-bones to bird's-eye-view recognition via perspective supervi-sion[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 17830-17839.

[5] MA Y, WANG T, BAI X, et al. Vision-centric bev perception: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10978-10996.

[6] LIU Z, TANG H, AMINI A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]// 2023 IEEE Inter-national Conference on Robotics and Automation(ICRA). IEEE, 2023: 2774-2781.

[7] WANG C, WANG Z, XU X, et al. Towards Accurate Post-training Quantization for Diffusion Models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2024.

[8] DIAO H, LI G, XU S, et al. Attention Round for Post-Training Quantization[J]. Neurocomputing, 2022, 521: 364-373.

[9] BANNER R, NAHSHAN Y, SOUDRY D. Post training 4-bit quantization of convolutional networks for rap-id-deployment [C]// Advances in Neural Information Processing Systems. 2019: 1-12.

[10] CHOUKROUN Y, KRAVCHIK E, YANG F, et al. Low-bit quantization of neural networks for efficient inference[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop(ICCVW). 2019: 1-12.

[11] CHENG Y, WANG D, ZHOU P, et al. A Survey of Model Compression and Acceleration for Deep Neural Networks[J]. arXiv: 1710.09282, 2017.

[12] YAO Z, AMINABADI R Y, ZHANG M, et al. ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers[C]// Advances in Neural Information Processing Systems 35(NeurIPS 2022). 2022: 27168-27183.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is

all you need[C]// Advances in Neural Information Processing Systems. 2017.

[14] LIU Z,WANG Y,HAN K,et al. Post-training quantization for vision transformer[C]// Advances in Neural Information Processing Systems. 2021:28092-28103.

[15] YUAN Z,XUE C,CHEN Y,et al. PTQ4ViT:Post-training Quantization for Vision Transformers with Twin Uniform Quantization[C]// 2021 European Conference on Computer Vision(ECCV). 2021:1-12.

[16] JIAO J,JIE Z,CHEN S,et al. MSMDFusion:Multi-Depth Seed Based Fusion of LiDAR and Camera Data for Enhanced 3D Object Detection[J]. IEEE Transactions on Intelligent Transportation Systems,2023,25:1-12.

[17] YIN Z,ZHOU X,KRÄHENBÜHL P. Multimodal Virtual Point 3D Detection: Establishing Fine-Grained Feature Alignment Across Different Modalities[C]// Proceedings of the IEEE International Conference on Computer Vision(ICCV). 2021:1-12.

[18] BAI H,HU Z,HUANG Q,et al. TransFusion:A Robust LiDAR-Camera Fusion Method for 3D Object Detection Using Transformers[J]. IEEE Transactions on Intelligent Transportation Systems,2022,25:1-12.

[19] CHEN L. Dense Projection Fusion(DPFusion):A Multi-Stage, Fine-Grained Fusion Technique for 3D Object Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2024,46:35-56.

[20] DOE J. Piecewise Linear Quantization(PWLQ)[J]. Neural Networks,2024,58(C):123-145.

[21] NAGEL J,LEUNG T,SCHMIDT T,Casale P. AdaRound: Adaptive rounding for post-training quantization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020:1-12.

[22] LI B,CHEN Q,XU H,et al. BRECQ:Bit-reconstruction quantization for ultra low-bit quantization aware training[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2021:1-12.

[23] HUBARA I,COURBARIAUX M,SOUDRY D,et al. Block-Wise Mixed-Precision Quantization: Enabling High Efficiency for Practical ReRAM-based DNN Accelerators[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68(11):1-12.

[24] BONDARENKO Y,NAGEL M,BLANKEVOORT T. Understanding and Overcoming the Challenges of Efficient Transformer Quantization[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021:7947-7969.

[25] FRANTAR M,HUBARA I,SOUDRY D,et al. Compressing Deep Neural Networks by Combining Weight Pruning with Post-Training Quantization[C]// Proceedings of the 36th Inernational Conference on Neural Information Processing Systems. 2022:12760-12773.

[26] XIAO G,LIN J,SEZNEC M,et al. SmoothQuant:Accurate and Efficient Post-Training Quantization for Large Language Models [C]// 40th International Conference on Machine Learning(ICML). 2023:10-12.

ZHANG Xiaoxuan, born in 1999, postgraduate. His main research interests include autonomous driving perception and object detection.

TANG Xiaoyong, born in 1973, Ph. D, professor, is a premium member of CCF (No. 33420S). His main research interests include parallel distributed computing and big data, etc.