计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010258] [CSTR: 32024.14.csa.010258]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

Fisher 信息引导的混合精度后训练量化算法①

郝泽颖, 王以桢, 王彦哲, 殷保群

(中国科学技术大学 信息科学技术学院, 合肥 230026)
通信作者: 殷保群, E-mail: bqyin@ustc.edu.cn

摘　要: 随着深度神经网络在计算机视觉领域的广泛应用, 后训练量化已成为其在资源受限设备上高效部署的关键技术. 针对现有混合精度后训练量化方法在低比特量化时计算最优位宽耗时较长, 以及由于校准样本有限而导致测试精度下降的问题, 提出了一种基于 Fisher 信息引导的混合精度后训练量化算法. 首先从 Kullback-Leibler 散度出发, 使用 Fisher 信息矩阵的平均迹快速计算各层的量化敏感度, 并构建融合量化误差与敏感度的扰动代价矩阵, 使用整数线性规划求解最优位宽分配问题. 此外, 在后训练过程中提出一种特征分布泛化的逐块重建方法, 使用全精度模型批归一化层中的统计信息重构输入特征, 增加重建过程中特征分布的多样性, 同时结合随机均匀特征混合机制丰富量化误差, 缓解有限校准数据条件下的泛化能力不足问题. 在 ImageNet 数据集上对多种主流卷积神经网络架构的实验结果表明, 所提方法能够在较短时间内完成混合精度位宽搜索, 并有效提升低比特量化模型的准确率.

关键词: 深度神经网络; 计算机视觉; 混合精度; 后训练量化; 特征分布泛化

引用格式: 郝泽颖,王以桢,王彦哲,殷保群.Fisher 信息引导的混合精度后训练量化算法.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10258.html

Fisher Information-guided Mixed-precision Post-training Quantization Algorithm

HAO Ze-Ying, WANG Yi-Zhen, WANG Yan-Zhe, YIN Bao-Qun

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: With the widespread adoption of deep neural networks (DNNs) in computer vision, post-training quantization has become a key technique for efficient deployment on resource-constrained devices. However, existing mixed-precision post-training quantization methods suffer from two major limitations: the long time required to compute optimal bit-width configurations for low-bit quantization, and the degradation in test accuracy caused by limited calibration samples. To address these issues, this study proposes a Fisher information-guided mixed-precision post-training quantization algorithm. Starting from the Kullback-Leibler divergence, the quantization sensitivity of each layer is efficiently estimated using the average trace of the Fisher information matrix, and a perturbation cost matrix that integrates quantization error and sensitivity is then constructed. The optimal bit-width allocation problem is then solved via integer linear programming. Furthermore, during post-training, a block-wise reconstruction method with feature distribution generalization is proposed, which uses the statistical information from batch normalization layers in the full-precision model to reconstruct input features, thus increasing the diversity of feature distributions during reconstruction. In addition, a random uniform feature-mixing mechanism is incorporated to diversify quantization errors, alleviating the insufficient generalization capability under limited calibration data. Experimental results on multiple mainstream convolutional neural network architectures on the ImageNet dataset demonstrate that the proposed method can complete mixed-precision bit-width search within a relatively short time and effectively improve the accuracy of low-bit quantized models.

① 收稿时间: 2026-02-09; 修改时间: 2026-03-02, 2026-03-16; 采用时间: 2026-03-20; csa 在线出版时间: 2026-08-21

Key words: deep neural network (DNN); computer vision; mixed precision; post-training quantization; feature distribution generalization

1 引言

深度神经网络 (deep neural network, DNN) 在人工智能的特定任务中展现出卓越的性能, 在计算机视觉中的图像分类[1]、目标检测[2]与图像分割[3]等领域均有着广泛应用. 然而, 更好的模型性能通常伴随着来自更深更宽网络的大量资源消耗, 显著的内存消耗和计算复杂性限制了高精度神经网络在端侧设备上的部署. 为了将深度神经网络快速部署到端侧设备上落地应用, 以资源约束场景下模型部署为目标的神经网络压缩与加速研究领域逐渐受到研究者们越来越多的关注. 模型量化是一种被广泛应用的模型压缩方法, 旨在通过降低数值精度来应对深度神经网络的存储与计算瓶颈. 该方法的核心机制在于将原有的高精度浮点权重映射为低位宽的定点数值, 从而在保持网络拓扑结构完整性的前提下实现模型体积的压缩. 当网络中的权重与中间激活值均被量化后, 底层运算逻辑可由复杂的浮点矩阵乘法转换为高效的整数算术运算. 这种计算模式的转换极大地减轻了内存带宽压力, 并显著降低模型在嵌入式终端设备上的推理延迟.

随着推理硬件的发展, 可变位宽的算术运算成为可能, 给网络量化带来了更大的灵活性, 混合精度量化由此进入研究者的视角, 通过将不同的网络层量化为不同的位宽, 从而在压缩比和精度之间实现更好的权衡. 量化感知微调 (quantization-aware training, QAT) 虽然能最大限度保留量化后模型的能力, 但是需要访问完整数据集重新训练模型, 消耗大量的计算资源. 而后训练量化 (post-training quantization, PTQ) 仅需要少量未标记数据即可对预训练模型进行量化, 适合模型的快速部署, 因此本文主要在混合精度后训练量化上展开研究. Dong 等人[4]认为在混合精度量化中 Hessian 矩阵的迹能整体反映该层参数在所有方向上的曲率, 能更真实地评估该层对量化误差的敏感程度, 因此使用 Hessian 矩阵的平均迹作为层敏感度. 但隐式计算 Hessian 矩阵特征值平均值的无矩阵 Hutchinson 算法对于每个网络层仍然需要 50 次迭代, 非常耗时. 如何高效地寻找到每层最优位宽分配是混合精度量化分配的难点. 同时在混合精度后训练量化过程中, 极低比特量化会导致模型泛化能力显著下降, 如何在少量的校准数据上减少模型的精度下降是另一个难点.

为了解决上述有限样本场景中的后训练量化挑战, 本文提出了一种基于 Fisher 信息引导的混合精度后训练量化 (Fisher information-guided mixed precision post-training quantization, FMPTQ) 方法. FMPTQ 在混合精度后训练量化场景下引入 Fisher 信息进行快速层敏感度建模, 并提出一种面向极低比特量化的特征分布泛化重建框架. 本文主要贡献包括: 1) 从 Kullback-Leibler (*KL*) 散度出发, 使用 Fisher 信息矩阵的平均迹快速计算各层敏感度. 随后构建结合量化误差与敏感度的扰动代价矩阵, 使用整数线性规划求解量化位宽分配问题. 2) 在后训练量化过程中提出了特征分布泛化的逐块重建方法, 使用全精度预训练模型批归一化层中的信息重构重建过程中的输入特征, 增加特征分布的多样性. 同时引入了随机均匀特征混合机制来丰富量化误差, 缓解重建过程中有限样本导致的泛化能力差的问题. 3) 通过在多种卷积神经网络架构上进行实验, 验证了本文所提方法在低比特后训练量化中的优越性.

2 相关工作

2.1 后训练量化

后训练量化是在模型训练完成后, 利用少量校准数据对模型参数进行量化, 无需重新训练, 相较于需要访问完整数据集的量化感知微调, 更适用于训练资源受限且要求快速部署的场景. 受限于可用校准样本数量, 现有研究通常关注如何缓解低比特量化带来的性能退化. Nagel 等人[5]在量化过程中引入可学习的舍入参数以优化权重量化误差. Li 等人[6]考虑了块内部层与层之间的相互影响, 在优化阶段最小化全精度模型块输出与量化模型块输出之间的误差. Wei 等人[7]为了缓解极少校准样本条件下的过拟合问题, 在 PTQ 的重构优化阶段对激活量化操作进行随机丢弃. 在前向传播的过程中, 对部分神经元的激活值不进行量化, 而是使用全精度激活替代. Jeon 等人[8]解耦量化步长与整数基座, 并结合块级重建和可学习的量化步长, 使得量化在低比特量化下依然能够保留高精度. 目前的后训练量化方法在极低比特量化下仍存在精度显著下降的

问题.

2.2 混合精度量化

卷积神经网络的不同层对量化的敏感度存在显著差异, 将所有权重与激活统一量化为相同精度往往会造成明显的性能退化. 为在模型压缩率与精度之间取得更优权衡, 研究者提出了混合精度量化策略, 即为敏感层分配较高位宽, 而对不敏感层采用较低位宽. 混合精度量化中的位宽分配方法通常可分为两类: 基于搜索的方法和基于敏感度分析的方法. 基于搜索的混合精度量化方法通常是使用强化学习自动探索网络各层的量化比特宽分配[9,10]. 这些方法擅长于管理硬件设备固有的不可微反馈. 然而, 它们的特点是往往搜索时间较长, 不利于模型的快速部署.

基于敏感度的混合精度量化方法通常通过构建代理度量来评估各层对量化扰动的影响, 并据此指导层级比特宽度的分配. 现有研究从不同角度探索了混合精度量化的敏感度建模与自动化搜索策略. Cai 等人[11]利用 Kullback-Leibler (*KL*) 散度作为代理度量, 并结合 Pareto 前沿优化的方法实现自动化位宽选择. Ma 等人[12]将正交性约束引入神经网络参数空间, 在多种资源约束条件下搜索最优比特配置. Dong 等人[13]使用每层的 Hessian 矩阵最大特征值作为敏感性指标, 但仍然需要手动选择混合精度设置. Dong 等人[4]在前面的方法上提出改进, 使用 Hessian 矩阵的平均迹作为更稳定的层级敏感度指标, 并引入 Pareto 前沿优化自动选择量化位宽. Huang 等人[14]进一步联合 Hessian 信息与参数规模构建综合敏感度度量, 以降低 Pareto 前沿搜索的计算开销. 尽管上述方法在混合精度量化方面取得了显著进展, 但在低比特后训练量化场景中, 层级敏感度的估计仍然需要较高的计算成本, 因此如何高效而稳定地评估各层敏感度, 并据此完成混合精度位宽分配, 仍然是后训练量化领域需要解决的问题.

3 算法设计与实现

针对前文分析的混合精度量化位宽计算效率低, 且在后训练量化过程中由于有限校准集而出现精度显著降低的问题, 本文针对性地提出了一种基于 Fisher 信息引导的混合精度量化位宽搜索方法 FMPTQ, 并在后训练过程中引入随机均匀混合策略与特征泛化方法提升量化模型在极少数据集上的性能表现. 算法的整体流程框架如图 1 所示. 本节将对所提算法的整体流程以及关键组件的设计进行详细介绍.

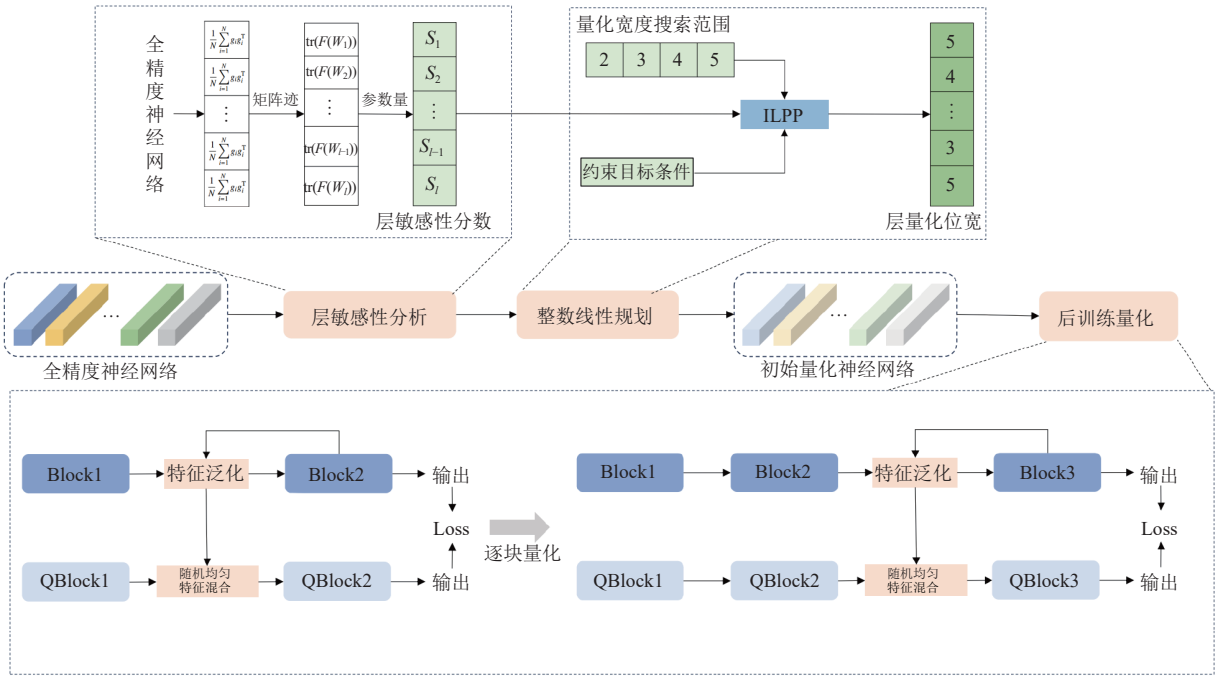


图 1 FMPTQ 整体量化流程

3.1 基于 Fisher 信息引导的层敏感度评估

量化的核心目标是以低比特精度表示模型权重, 同时使量化前后模型输出中的扰动最小[11]. 尽管量化会在各个网络层中引入一定的误差, 但我们更关注这

些误差对最终损失函数的整体影响, 而非单独考察每一层的局部扰动, 因为总体损失的变化更能直接反映量化后模型性能的退化程度. 在深度神经网络中, 不同层参数对输出分布及最终任务性能的贡献存在显著差异; 若对所有层采用统一的量化精度, 不可避免地会导致关键层的信息损失过大, 而在冗余层上则造成比特资源的浪费. 因此, 我们需要对模型中不同层的量化敏感性进行定量评估, 从而为后续自适应比特宽度分配策略提供理论依据.

本文从统计建模视角出发, 将神经网络视为由参数W所参数化的条件概率模型$p(y\mid x,W)$, 并将量化误差视为参数空间中的扰动. 对于第l层权重矩阵W_l, 引入零均值高斯扰动$W_l'=W_l+\varepsilon$, 其中$\varepsilon\sim N(0,\sigma^2 I)$, I为单位矩阵. 这种扰动模拟了训练期间的自然权重波动, 使我们能够分析权重的变化如何影响模型的输出. 设$L(W)$表示预训练后模型的损失函数. 对W_l在扰动ε下进行二阶泰勒展开, 如式 (1) 所示:

$$L(W_l+\varepsilon)\approx L(W_l)+\varepsilon^{\mathrm{T}}g^{(W_l)}+\frac{1}{2}\varepsilon^{\mathrm{T}}H^{(W_l)}\varepsilon \tag{1}$$

其中, $g^{(W_l)}$、$H^{(W_l)}$分别表示损失函数关于第l层参数的一阶梯度和 Hessian 矩阵. 由于模型已在训练集上收敛, 梯度项在期望意义下满足$\mathbb{E}[g^{(W_l)}]\approx 0$, 因此量化扰动引起的损失变化主要由二阶项 Hessian 矩阵和量化误差主导. 然而, 过往的工作在神经网络中直接计算 Hessian 矩阵的迹需要消耗大量计算资源[4,13,14]. 为了避免对 Hessian 矩阵迹的直接计算, 我们引入 KL 散度来刻画参数扰动前后模型输出分布的变化如式 (2) 所示:

$$\begin{aligned}&KL(p(y\mid x,W_l)\,\|\,p(y\mid x,W_l+\varepsilon))\\&=\int p(y\mid x,W_l)\log\frac{p(y\mid x,W_l)}{p(y\mid x,W_l+\varepsilon)}\mathrm{d}y=\frac{1}{2}\varepsilon^{\mathrm{T}}F(W_l)\varepsilon\end{aligned} \tag{2}$$

其中, $F(W_l)$为 Fisher 信息矩阵, 定义如式 (3) 所示:

$$F(W_l)=\mathbb{E}_{(x,y)\sim D}\left[\nabla_{W_l}\log p(y\mid x,W_l)\nabla_{W_l}\log p(y\mid x,W_l)^{\mathrm{T}}\right] \tag{3}$$

已有研究表明, 在极大似然或交叉熵训练条件下, Fisher 信息矩阵可作为 Hessian 矩阵的等价近似[15,16]. 因此我们可以使用 Fisher 信息矩阵来近似 Hessian 矩阵, 如式 (4) 所示:

$$\Delta L=\mathbb{E}[L(W_l+\varepsilon)-L(W_l)]\approx\frac{1}{2}\varepsilon^{\mathrm{T}}F(W_l)\varepsilon \tag{4}$$

其中, ε为零均值扰动, 对其取期望可得$\mathbb{E}[\varepsilon^{\mathrm{T}}F(W_l)\varepsilon]=\mathrm{tr}(F(W_l)\Sigma)$, 其中$\Sigma=\mathbb{E}[\varepsilon\varepsilon^{\mathrm{T}}]$为扰动协方差. 因此, 若进一步采用常见的各向同性假设$\Sigma\approx\sigma^2 I$, 可得$\mathbb{E}[\Delta L]\propto\mathrm{tr}(F(W_l))$. Fisher 信息矩阵刻画了模型输出分布对参数扰动的敏感程度, 其迹反映了各参数方向局部曲率强度的总体累积, 因此可作为衡量层量化敏感性的重要指标. 这一思路与 HAWQ-V2[4]、HMQAT[14]等二阶敏感度量化方法在理论框架上是一致的. 在实际计算中, Fisher 信息矩阵可由样本梯度的外积平均进行估计, 如式 (5) 所示:

$$F(W_l)=\frac{1}{N}\sum_{i=1}^{N}g_i g_i^{\mathrm{T}} \tag{5}$$

其中, N表示样本数量, g_i为$\nabla_{W_l}\log p(y_i\mid x_i,W)$. 进一步地, 由于矩阵的迹仅由对角元素之和决定, 采用$\mathrm{tr}(F(W_l))$作为层敏感度度量时, 无需显式构造和存储完整的 Fisher 信息矩阵, 只需计算各参数对应的梯度平方项即可, 从而显著降低计算与存储开销. 考虑到不同层的参数规模存在差异, 若直接使用 Fisher 信息矩阵的迹进行比较, 容易使参数量较大的层获得偏大的敏感度估计. 因此, 本文进一步采用平均迹作为第l层的量化敏感性度量, 如式 (6) 所示:

$$S_l=\frac{1}{|W_l|}\mathrm{tr}(F(W_l)) \tag{6}$$

其中, $|W_l|$表示第l层中包含的权重的总数量. S_l数值越大, 说明第l层对扰动越敏感, 在量化过程中应分配更高比特宽度以抑制性能退化, 反之S_l数值越小, 在量化过程中应分配更低比特宽度. 图 2 中展示了在 ImageNet 数据集上 ResNet-50 中各层与其对应敏感度分数的关系.

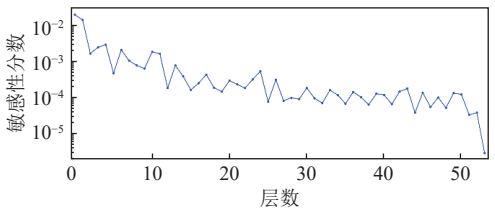
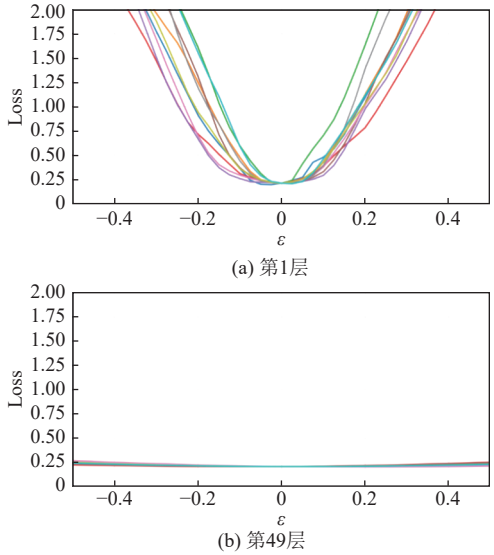


图 2 ResNet-50 各层敏感度评分

图 3 中展示了在 ResNet-50 第 1 层与第 49 层, 在 10 个随机方向上扰动预训练的权重后 Loss 的变化. 从图 3 中可以看出第 1 层的敏感性分数相较于第 49 层更高, 在引入扰动后 Loss 的变化更加明显, 因此在后续量化过程中应分配更大的量化宽度以减少模型性能的损失.



(a) 第1层

(b) 第49层

图 3 ResNet-50 中第 1 层和第 49 层 Loss 随扰动的变化

3.2 基于整数线性规划的混合精度量化宽度分配

在混合精度量化中, 一个核心挑战是如何在不损失模型准确度的情况下根据不同层的敏感性来为每层分配合适的量化宽度. 第 3.1 节提出的基于 Fisher 信息矩阵的平均迹提供了一个量化比特宽度的相对排序, 但是它不能直接提供不同层精确的量化比特宽度. 因为在图 2 中可以看出 ResNet-50 的第 1、2 层和第 51、52 层相差几个数量级, 因此即使我们知道第 2 层比第 52 层更敏感, 仍无法分配具体的量化宽度. 为了解决上述问题, 我们将混合精度量化问题转化为一个约束优化问题, 我们的目标是在满足给定的目标模型体积约束下, 最小化由量化引起的总预期损失变化. 由于该问题的解空间涉及所有层在备选位宽集合上的选择, 本质上由一系列布尔决策参数构成, 因此我们选择整数线性规划 (integer linear programming problem, ILPP) 来求解该问题.

首先, 我们将式 (6) 中定义的敏感度S_l与量化误差相结合, 构建扰动代价矩阵. 定义$M_{l,j}$为第l层在选择候选位宽集合$K=\{b_1,\cdots,b_J\}$中的第j个位宽时的预期损失变化如式 (7) 所示:

$$M_{l,j}=S_l\cdot\|Q_{b_j}(W_l^*)-W_l^*\|_2^2 \tag{7}$$

其中, W_l^*为原始权重, $Q_{b_j}(\cdot)$为使用b_j位宽时的量化函数. 该公式利用 Fisher 平均迹加权来量化噪声的L_2范数, 从而统一了不同层的度量尺度. 基于此, 我们构建式 (8) 作为 ILPP 优化目标.

$$\begin{cases}\min\limits_{C_{l,j}}\sum\limits_{l=1}^{L}\sum\limits_{j=1}^{J}M_{l,j}C_{l,j}\\ \text{s.t.}\begin{cases}\sum\limits_{j=1}^{J}C_{l,j}=1,\ \forall l\in\{1,2,\cdots,L\}\\ \sum\limits_{l=1}^{L}\sum\limits_{j=1}^{J}P_{l,j}C_{l,j}\leqslant P_{\text{target}}\\ C_{l,j}\in\{0,1\}\end{cases}\end{cases} \tag{8}$$

其中, $C_{l,j}$是二值决策变量, 当$C_{l,j}=1$时表示第l层被分配了第j种位宽; $P_{l,j}$表示第l层在采用第j种位宽时的参数体积. P_{target}是设定的目标总模型大小. 通过求解式 (8), 我们可以在全局模型体积约束下, 自动化地搜索出最优的位宽分配策略. 图 4 中展示了位宽为{2, 3, 4}时 ResNet-50 中第l层与$M_{l,j}$的对应关系, 在使用 ILPP 分配位宽时倾向于给更大的$M_{l,j}$更大的位宽.

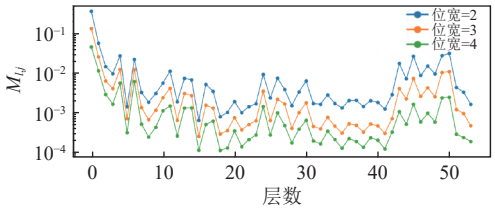


图 4 不同位宽下 ResNet-50 中第l层与$M_{l,j}$的对应关系

3.3 基于特征分布泛化的模型后训练量化

过往的以 AdaRound[5]为代表的算法在后训练量化过程中使用逐层重建方式, 在一定程度上缓解了量化带来的损失, 但未能充分考虑网络内部层与层之间的强耦合关系和误差传递的累积效应. 在低比特宽度量化场景下, 这种误差往往在网络深层逐渐放大, 从而显著影响最终性能. BRECQ[6]提出将多个相邻层组合为一个整体的块, 在块内同时进行量化参数的重构优化, 使得误差在块内部得到协调, 从而更好地捕捉内部层间的交互信息和误差传播规律. 因此在后训练量化过程中, 我们参考 BRECQ 的思想以块为基础将量化模型与其全精度对应模型之间的输出对齐, 以在量化后恢复模型精度. 具体而言, 对模型第k个块进行后训练量化时, 局部距离损失函数定义为量化块输出与浮点块输出的ℓ_2-norm, 具体计算如式 (9) 所示, 其中F_k表示第k个全精度块的输出, $\hat{F}_k$表示第k个量化块的输出.

$$L_{\text{rec}}=\left|F_k-\hat{F}_k\right|_2^2 \tag{9}$$

在权重量化过程中本文采用自适应舍入策略, 权

重的量化过程如式 (10) 所示:

$$\bar{x} = clip\left(\left\lfloor \frac{x}{s_w} \right\rfloor + r_w, q_{\min}, q_{\max}\right) \tag{10}$$

其中, x为全精度权重, $\bar{x}$为量化后权重. s_w为权重量化步长, r_w是决定每个权重值向上或向下取整的量化取整函数, 取值范围为 0 或 1. 而$q_{\min}$和$q_{\max}$则分别代表了量化整数空间的下界和上界, 这两个参数由该层分配到的量化位宽决定.

为了减少权重量化过程中因舍入决策不确定性带来的性能波动, 我们在量化优化目标中引入了正则化项. 令每个权重的量化取整函数r_w由 Rectified Sigmoid 函数从连续参数映射而来. 正则化项的设计旨在引导r_w远离中间值 0.5, 从而促使权重向 0 和 1 两端分布, 接近确定性的硬舍入结果. 我们定义正则化项如式 (11) 所示:

$$L_{\text{round}} = \lambda_r \sum_w (1 - |2r_w - 1|^{\alpha}) \tag{11}$$

其中, λ_r为正则项的权重系数, r_w表示第w个权重对应的软舍入值, α为随训练进程动态调整的温度指数, 在优化的初期, α较大, 正则化较弱. 允许r_w在 0–1 之间自由适应, 以寻找能够最小化局部距离损失的最优解. 在优化的后期, α逐渐减小, 强制r_w收敛到 0 或 1, 从而得到合法的二进制量化权重. 权重逐块重建过程中的损失函数的总损失由局部损失和正则项的加权和组成, 如式 (12) 所示:

$$L_w = L_{\text{rec}} + L_{\text{round}} \tag{12}$$

通过引入这些损失函数及正则化项, 逐块量化方法能够在量化过程中有效地减少误差并保持模型的精度, 同时通过正则项确保量化权重的稳定性, 从而进一步提升量化后的模型性能.

激活在前向传播中使用式 (13) 所示的量化方式, 其中a为全精度激活, $\bar{a}$为量化后激活. S_a为激活量化步长, 决定了将连续浮点数值映射到离散整数空间的精细程度.

$$\bar{a} = clip\left(\left\lfloor \frac{a}{S_a} \right\rfloor, q_{\min}, q_{\max}\right) \tag{13}$$

激活量化参数的更新在逐块重建中使用量化感知微调 LSQ[17]的思想, 考虑到后训练量化只能使用少量无标签校准集, 使用局部损失替代 LSQ 中的任务损失. 我们使用反向传播来优化激活量化步长参数S_a, 通过 STE 计算舍入函数的梯度, 梯度计算如式 (14) 所示, 其中$\hat{a}$为反量化后的激活.

$$\frac{\partial L_{\text{rec}}}{\partial S_a} = \begin{cases} \dfrac{\partial L_{\text{rec}}}{\partial \hat{a}} q_{\max}, & \dfrac{a}{S_a} \geqslant q_{\max} \\ \dfrac{\partial L_{\text{rec}}}{\partial \hat{a}} \left(\left\lfloor \dfrac{a}{S_a} \right\rfloor - \dfrac{a}{S_a} \right), & q_{\min} < \dfrac{a}{S_a} < q_{\max} \\ \dfrac{\partial L_{\text{rec}}}{\partial \hat{a}} q_{\min}, & \dfrac{a}{S_a} \leqslant q_{\min} \end{cases} \tag{14}$$

尽管混合精度量化搜索策略可以为每层权重实现更有利的位宽分配, 但由于量化噪声的逐层累积效应, 低位宽量化仍会导致显著的输出偏差. 利用少量校准数据对量化参数进行微调已被证明能有效恢复模型精度, 但这一过程受限于只能使用少量无标签数据, 在逐块量化校准过程中易在校准集出现泛化能力不足的现象[7]. 为了解决上述问题, 我们从两个角度提升逐块重建过程中输入特征的多样性.

首先是在重建过程中引入随机均匀特征混合机制 (random uniform feature mixing, RUFM). QDrop[7]提出了在模型 PTQ 的逐块重建过程中对每个量化块的输入激活特征引入随机丢弃策略, 即量化过程中以概率p随机丢弃量化块输入激活特征, 由全精度激活特征替代丢弃部分, 如式 (15) 所示:

$$\tilde{a} = \begin{cases} a, & \text{概率为} p \\ \bar{a}, & \text{概率为} 1-p \end{cases} \tag{15}$$

我们认为 QDrop 的随机丢弃 (random drop, RD) 机制虽然在单一方向上减少了过拟合的风险, 但只提供了全精度和量化精度两种选择, 限制了量化误差的多样性, 因此我们提出了随机均匀特征混合机制, 如式 (16) 所示.

$$\tilde{a} = a + \beta \odot (\bar{a} - a) \tag{16}$$

该机制将二值门控的随机替换扩展为连续区间上的随机插值, 从而在逐块重建阶段为量化块提供更丰富、更平滑的输入扰动分布. 具体而言, β是从均匀分布$U(0,1)$采样, 使得每个激活特征都能够以不同强度注入量化误差$(\bar{a}-a)$. 因此, $\tilde{a}$可被视为在全精度激活a与量化激活$\bar{a}$之间的随机路径采样点, 其分布覆盖了二者连线所张成的连续特征子空间, 而不再局限于端点集合$\{a,\bar{a}\}$, 该策略采用连续随机权重以增强扰动多样性, 使得重建过程在多方向上更加平滑稳定.

其次, 我们对逐块重建过程中的输入特征引入基于批归一化先验的特征重构 (feature reconstruction

based on batch normalization prior, FR). 由于在后训练量化过程中我们仅能访问极少量的无标签数据, 导致量化过程中各层的浮点激活输出存在显著的表征偏差. 仅基于这些校准样本得出的最优量化参数, 往往难以适配全局的训练集分布, 我们认为这是导致量化模型在测试集上性能较差的原因之一. 针对上述问题我们利用神经网络自身性质, 提出了一种基于批归一化(batch normalization, BN) 先验的特征重构方法, 使用全精度模型 BN 层中训练集统计信息重构校准集的输入特征. 模型的 BN 层在训练过程中对每个小批量数据的层内激活计算均值与方差, 并通过滑动平均将这些批统计量累积为稳定的运行均值μ与运行方差σ. 由于这些统计量刻画了训练数据特征激活分布的低阶矩信息. 区别于零样本量化方法 ZeroQ[11]中使用 BN 统计量生成训练样本, 后训练量化算法 BBL[18]在逐块重建过程中对量化块中的 BN 参数进行更新, 并对逐块重建过程中的输入数据进行块级的数据增强. 我们使用预训练模型中的 BN 层统计信息构成模型内部对训练数据分布特征的压缩表征, 用于重构与训练分布一致的输入特征, 而非从零生成训练数据或对 BN 中的参数进行微调.

具体而言, 我们利用 BN 层中记录的滑动均值和滑动方差来构建一个一致性目标. 通过该目标, 我们对全精度模型块输入的激活特征进行微调, 强制其输出的一阶矩均值与二阶矩方差与原始训练集的统计分布接近. 通过这种重构操作, 使量化器能够在一个更接近真实分布的特征空间中进行参数寻优, 增强模型的泛化性. 对于第k个模型块包含的N个 BN 层, 我们可以计算其输入的统计特征, 即均值$\hat{\mu}_k^n$和方差$\hat{\sigma}_k^n$, 其中$n\in\{1,2,\cdots,N\}$. 全精度模型 BN 层中存储的原始训练集统计先验记为μ_k^n和σ_k^n. 特征重构的约束目标如式 (17) 所示:

$$\min_{F_{k-1}^{\mathrm{FR}}} \| F_{k-1}-F_{k-1}^{\mathrm{FR}} \|_2^2+\sum_{n=1}^{N}\lambda_f\left(\| \hat{\mu}_k^n-\mu_k^n \|_2^2+\| \hat{\sigma}_k^n-\sigma_k^n \|_2^2\right) \tag{17}$$

其中, F_{k-1}为重构前的全精度激活输入, F_{k-1}^{FR}为重构后的激活输入. λ_f为平衡系数, 用于平衡统计误差损失与局部损失. 式 (17) 第 1 项用于约束特征偏移的幅度, 防止在特征重构的过程中丢失原始图像的信息. 第 2 项旨在最小化校准数据与原始分布的 BN 统计特征距离, 即让$\hat{\mu}_k^n$和$\hat{\sigma}_k^n$逼近μ_k^n和σ_k^n. 所提特征重构方法与随机均匀特征混合机制可以结合使用, 提高量化过程中的输入特征的泛化性. 图 5 展示了结合随机均匀特征混合机制和基于批归一化先验的特征重构的逐块后训练量化流程, 特征泛化后的结果作用于全精度块的输入与量化块的输入, 而随机均匀混合作用于量化块的输入.

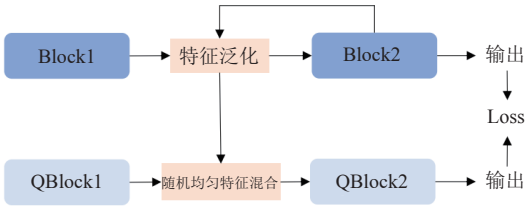


图 5 后训练量化过程示意图

4 实验结果与分析

4.1 数据集

ImageNet[19]数据集是计算机视觉领域最具影响力的图像分类数据集, 是评估卷积神经网络量化的核心基准, 包含 1 000 个物体类别和超过 128 万张训练图片, 我们从数据集中随机选择 1 024 张图像作为校准样本数据集, 量化后模型的性能在 ImageNet 验证集上进行评估, 该验证集包含 50 000 张图像. 我们主要对量化后的模型在图像分类任务进行评估, 没有应用额外的数据增强技术, 确保评估仅关注量化的效果.

4.2 实验细节

实验均基于 PyTorch 框架实现, 其中 ResNet-18 与 RegNetX-600MF 在 NVIDIA GeForce RTX 2080Ti GPU 上进行实验, 其余模型在 NVIDIA GeForce RTX 3090 GPU 上进行实验. 由于现有量化方法在量化位宽大于 4 时均能维持较少的精度损失, 因此本文主要关注平均量化位宽小于等于 4 的低位量化, 量化位宽搜索范围为{2, 3, 4, 5}. 激活量化参数的学习率为 4E–5, 取整量化参数学习率为 3E–3, 特征重构的学习率为 1E–3, 逐块迭代次数为 20 000.

4.3 实验结果

本节我们将所提 MPTQ 算法与多种主流的后训练算法在 ResNet-18[20]、ResNet-50[20]、MobileNetV2[21]、RegNetX-600MF[22]、RegNetX-3.2GF[22]、MNasNet-2.0[23]上进行性能对比, 在多种量化宽度配置下进行广泛的实验, 实验结果如表 1 所示. 在与其他算法的对比实验中我们主要比较了平均权重量化位宽、激活量化位宽、Top-1 准确率. 平均量化宽度通过式 (18) 计算, 其中b_l为模型中第l层的量化宽度.

$$\overline{b}=\frac{\sum_{l=1}^{L} b_l \cdot |W_l|}{\sum_{l=1}^{L} |W_l|} \tag{18}$$

在表 1 的对比结果中, “基线”表示没有量化的全精度模型, “W/A”表示权重量化宽度与激活量化宽度, “MP”表示混合精度量化. 对比结果中 QDrop、OMPQ 和 FIMA-Q 的结果使用开源代码在相同实验条件下复现得到. 从表 1 中可以看 FMPTQ 相较于其他主流后训练量化方法取得了更加优异的性能. 对于 ResNet-18 和 ResNet-50, 在平均量化位宽 W/A 为 4/4 时量化后的模型精度仍能维持全精度基线模型的 98.15% 和 98.30% 的相对性能准确率, 并将权重位宽和激活位宽压缩了近 8 倍. 在 W/A 为 4/2 时, 这种涉及 2 bits 的极低比特环境下, FMPTQ 仅使用 1 024 个校准样本, 在 ResNet-50 上仍能维持 65.96% 的准确率, 同为混合精度量化方法的 OMPQ 在极低比特下由于校准数据集过少导致泛化能力不足, 准确率仅为 54.11%, 这是因为我们所提出的随机均匀特征混合机制与特征重构能够提高输入特征的多样性, 有效地缓解了 Top-1 准确率的降低. FIMA-Q[24]是针对 ViT 均匀 PTQ 方法, 针对 ViT 具有全局 Token 相关性提出了 DPLR-FIM 损失, 使用低秩 FIM 结合对角线 FIM 近似法, 提出了一种量化局部重建损失函数, 我们将该方法在 CNN 上进行复现. 由于 FIMA-Q 官方实现主要面向 ViT, 量化算子与模型结构与 CNN 存在差异, 为保证对比公平, 我们将 FIMA-Q 的 DPLR-FIM 损失迁移到与本文一致的 CNN 后训练量化重建框架与量化算子设置中, FIMA-Q[24]中参数学习率, 逐块迭代轮次与第 4.2 节中 FMPTQ 的设置相同, DPLR-FIM 损失中的相关参数则与官方代码中的参数相同. 由于 CNN 结构区别于 ViT, Fisher 信息能量主要集中于对角线上, FIMA-Q 在 CNN 后训练量化上并没有显著的性能提升.

表 1 FMPTQ 与各种后训练量化算法的 Top-1 准确率性能对比 (%)

方法	W/A	ResNet-18	ResNet-50	MobileNetV2	RegNetX-600MF	RegNetX-3.2GF	MNasNet-2.0
基线	32/32	70.99	76.66	72.63	73.52	78.46	76.52
QDrop[7]	4/4	69.17	75.15	68.07	70.91	76.4	72.81
Genie-M+QDrop[8]	4/4	69.35	75.21	68.65	71.13	76.75	73.37
MRECG[25]	4/4	67.69	73.96	66.55	69.22	—	—
Q-limit[26]	4/4	69.17	75.02	68.46	71.02	76.47	73.31
BBL+QDrop[18]	4/4	—	—	68.13	70.74	—	—
FIMA-Q[24]	4/4	69.06	74.76	68.00	—	—	—
ZeroQ[11]	4MP/4	69.05	72.67	63.31	—	—	—
OMPQ[12]	4MP/4	69.57	75.14	65.01	—	—	—
PLMQ[27]	4MP/4MP	69.36	73.25	65.72	—	—	—
FMPTQ (ours)	4MP/4	**69.68**	**75.36**	**68.73**	**71.51**	**76.89**	**73.63**
MRECG[25]	3/3	60.01	66.82	48.87	57.75	—	—
Q-limit[26]	3/3	65.71	71.33	56.83	65.15	72.02	65.03
Genie-M+QDrop[8]	3/3	66.16	71.61	57.54	65.68	72.72	64.80
BBL+QDrop[18]	3/3	—	—	56.33	64.58	—	—
FIMA-Q[24]	3/3	65.27	71.32	55.78	—	—	—
FMPTQ (ours)	3MP/3	**67.04**	**72.86**	**57.9**	**66.62**	**72.94**	**66.62**
QDrop[7]	4/2	57.76	63.26	17.3	49.73	62.00	34.12
OMPQ[12]	4MP/2	49.75	54.11	0.41	—	—	—
FMPTQ (ours)	4MP/2	**61.3**	**65.96**	**23.21**	**52.00**	**62.92**	**42.99**

注: 加粗数据表示最佳结果

4.4 消融实验

4.4.1 校准样本数量对 Fisher 信息矩阵平均迹的影响

为了研究校准数据集中不同样本数量对敏感性分数的影响, 我们测试了 ResNet-18 的 Layer2.0.conv1 和 Layer3.0.conv2 层 Fisher 平均迹随 0–3 000 个校准样本数量的变化, 如图 6 所示, 可以看出所展示的两层在使用 1 000 个样本以后 Fisher 平均迹均在极小范围内变化, 因此使用 1 024 个校准数据样本集即可计算出各层稳

定的敏感性分数.

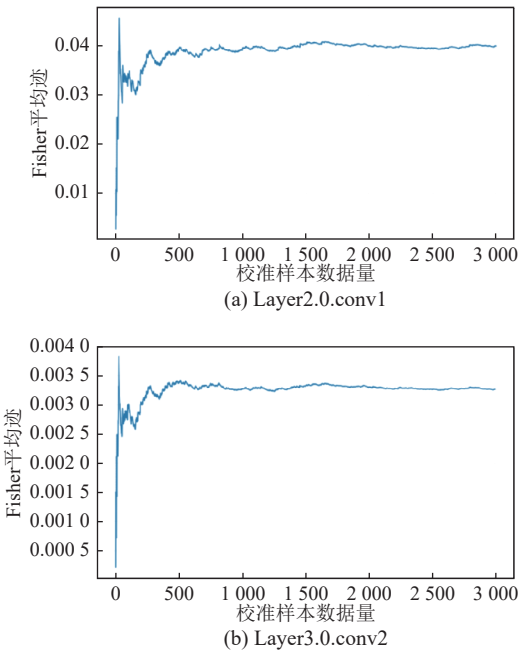


图 6 ResNet-18 不同层的 Fisher 平均迹随数据量变化

4.4.2 随机均匀特征混合机制有效性验证

为了证明所提随机均匀特征混合机制的优越性, 在平均混合量化宽度 W/A 为 3/3 时, 在 ResNet-18、ResNet-50 和 RegNetX-600MF 上对随机均匀特征混合机制 (RUFM) 与 QDrop 的随机丢弃机制 (RD) 进行消融实验, 随机丢弃概率p设置为 0.5, 对比结果如表 2 所示. 观察表 2 中数据, 可以发现 RUFM 策略在所有测试模型上的表现均优于 RD 策略, 在 RegNetX-600MF 上使用 RUFM 的后训练量化算法比使用 RD 的后训练量化算法 Top-1 准确率高 0.35%. 这是因为 QDrop 的 RD 本质上是一种离散的二值选择, 其仅能在全精度激活特征a与量化激活特征$\bar{a}$这两个端点之间进行切换. 尽管这种方式在一定程度上引入了随机性以防止过拟合, 但它限制了重建过程中样本的多样性. 相比之下, RUFM 策略通过引入服从均匀分布的系数进行激活特征混合, 增加了量化误差的多样性, 使得模型在重建阶段能够探索到更优的解.

表 2 RUFM 与 RD 在不同模型上的对比实验 (%)

模型	ResNet-18	ResNet-50	RegNetX-600MF
RUFM	67.04	72.86	66.61
RD	66.83	72.51	66.34

4.4.3 量化框架中各组件有效性验证

为了证明所提基于 Fisher 引导的低比特混合量化方法中各组件的有效性, 我们在 ImageNet 数据上在平均量化宽度配置 W/A 为 3/3 时, 对 ResNet-18、 ResNet-50 和 RegNetX-600MF 进行了一系列消融实验. 所有量化模型使用相同的实验配置, 实验结果如图 7 所示, 其中 RUFM 表示随机均匀特征混合, FR 表示特征重建, MP 表示混合精度策略.

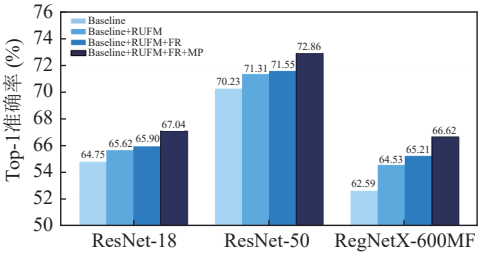


图 7 不同组件的消融实验结果

从图 7 中我们可以看出没有使用任何组件的基线模型均在评测集上表现出最差的 Top-1 准确率. 当引入 RUFM 策略后量化后模型性能有了大幅提升. 这一显著的性能增益主要归功于 RUFM 策略在后训练量化中引入了多样性的量化误差, 使得优化算法能够探索更平滑的损失曲面, 从而寻找到泛化能力更强的量化步. 在 RUFM 的基础上进一步引入 FR 模块后, 各模型的准确率再次得到了稳步提升, 这一结果证明了使用全精度模型对输入激活特征进行重构的必要性. 最后, 我们在上述优化的基础上引入了基于 Fisher 信息引导的混合精度策略, 即完整的 FMPTQ 方法. 相比于前面统一位宽的配置, 在低比特量化下各量化模型准确率均获得显著提升. 这一结果证明了混合精度量化宽度分配在低比特量化中的重要性. 统一位宽量化忽略了网络内部不同层对扰动的容忍度差异, 强制对高敏感层进行低比特压缩会导致不可逆的信息损失. 随着组件的不断叠加, 量化模型的准确率呈现递增趋势, 有力地证明了各组件在 FMPTQ 算法中的不可或缺性, 共同提升了低比特混合精度后训练量化算法的性能上限.

4.4.4 正则化参数消融实验

表 3 中展示了在 ResNet-18 和 RegNetX-600MF 上平均量化宽度 W/A 为 3/3 时, 采用不同的正则化权重系数λ_r进行消融实验的结果. λ_r用于限制正则化约束的强度, 该参数为 0 时代表不使用正则化约束. 从表 3 中数据可以看出, 应用正则化约束后的模型准确率相较于不应用正则化约束的模型准确率出现了明显提高, 这种差距表明正则化项在减少舍入不确定性、提高量化后模型的准确性方面起到了关键作用. 而使用正则

化约束后, 不同λ_r量化后的模型准确率之间变化范围较小, 证明所提算法对λ_r具有较好的鲁棒性, 能够在不同的正则化强度下获得相对稳定的性能.

表 3 不同正则化参数λ_r对量化模型性能的影响 (%)

模型	0	0.01	0.05	0.1	0.5
ResNet-18	65.85	67.05	67.01	66.92	67.00
RegNetX-600MF	64.87	66.61	66.53	66.59	66.49

4.5 量化位宽分配对比实验

为了验证本文的混合精度量化位宽搜索方法的优越性, 使用相同的后训练量化方案, 只测试不同量化位宽搜索方案带来的性能影响. 我们使用 BRECQ 的后训练量化方案, 在 ResNet-18 模型上使用 8 位激活量化宽度, 测试了不同模型约束大小下, OMPQ[12]、BRECQ[6]、PLMQ[27]和 FMPTQ 分配的权重量化位宽对模型的性能的影响, 结果如图 8 所示.

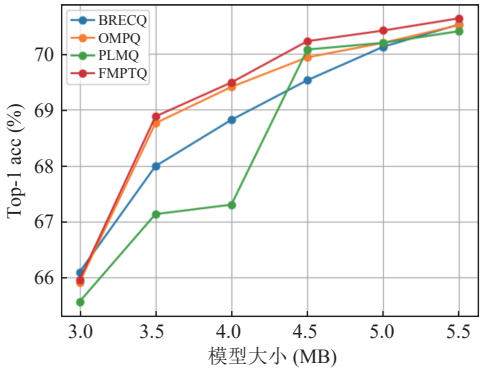


图 8 相同后训练量化算法下不同混合精度量化方案的比较

PLMQ 与 FMPTQ 都引入 Fisher 信息来刻画层级敏感度, PLMQ 中的 Fisher 信息更多用于提供分组依据, 结合聚类与位宽递减原则指导位宽调整的启发式决策, 而 FMPTQ 不将敏感度仅作为排序信号, 而是进一步将 Fisher 敏感度与 L2 量化误差联合建模为扰动代价矩阵, 使敏感度从启发式指标转化为可优化目标中的显式代价项. 在模型大小≤4M 后, PLMQ 的性能出现显著下降, 因为 PLMQ 的聚类算法不可避免地忽略了组内各层的敏感度差异, 容易陷入局部最优, 而 FMPTQ 直接在全局预算约束下求解单层的最优量化位宽组合, 因此在多个模型大小上取得更稳定的精度与压缩比的平衡.

4.6 混合精度量化位宽分配方法计算效率

在本节中对所提量化位宽分配方法的计算效率进行实验, 包含 Fisher 信息矩阵平均迹与整数线性规划分配量化位宽的总体计算时间. 表 4 展示了不同模型在单张 GPU 上的量化位宽分配总运行时间, 以及计算过程中消耗的内存. 可以看出随着模型尺寸的增加, 计算时间也相应增加, 但在卷积神经网络中最长计算时间在 53 s 左右, 这表明计算成本相对较低.

表 4 不同模型量化位宽搜索方法效率计算

模型	运行时间 (s)	内存使用 (MB)
MobileNetV2	19	5450
ResNet50	25	6778
MNasNet-2.0	34	6772
RegNetX-3.2GF	53	7190

5 结论与展望

针对过往深度神经网络混合精度量化宽度分配过程复杂, 且后训练过程中由于数据量极少导致在测试集上泛化能力不足的问题, 本文提出了一种基于 Fisher 信息引导的低比特神经网络混合精度后训练量化方法. 首先使用 Fisher 信息矩阵替代 Hessian 矩阵的计算, 结合层参数量计算平均迹作为层敏感性度量. 随后结合敏感度量和量化误差利用整数线性规划寻找量化宽度的最优分配. 最后在混合精度量化模型的逐块量化过程中引入随机均匀特征混合机制与特征重构方法提高量化模型在低比特量化上的泛化性, 并在多种神经网络架构上验证所提算法的有效性. 未来将适配更广泛的任务场景. 本文的压缩场景为图像分类任务, 人工智能领域还有很多其他任务场景如目标检测、语义分割和超分辨率重建等, 未来的研究工作会尝试将所提量化方法使用在更多领域中.

参考文献

1 Woo S, Debnath S, Hu RH, *et al*. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 16133–16142.

2 Cheng TH, Song L, Ge YX, *et al*. YOLO-world: Real-time open-vocabulary object detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 16901–16911.

3 Xu JC, Xiong ZX, Bhattacharyya SP. PIDNet: A real-time semantic segmentation network inspired by PID controllers. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 19529–19539.

4 Dong Z, Yao ZW, Arfeen D, *et al*. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks.

Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1555.

5 Nagel M, Amjad RA, van Baalen M, *et al*. Up or down? adaptive rounding for post-training quantization. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 7197–7206.

6 Li YH, Gong RH, Tan X, *et al*. BRECQ: Pushing the limit of post-training quantization by block reconstruction. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

7 Wei XY, Gong RH, Li YH, *et al*. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

8 Jeon Y, Lee C, Kim HY. Genie: Show me the data for quantization. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 12064–12073.

9 Wang K, Liu ZJ, Lin YJ, *et al*. HAQ: Hardware-aware automated quantization with mixed precision. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8612–8620.

10 Elthakeb AT, Pilligundla P, Mireshghallah F, *et al*. ReLeQ: A reinforcement learning approach for automatic deep quantization of neural networks. IEEE Micro, 2020, 40(5): 37–45. [doi: 10.1109/MM.2020.3009475]

11 Cai YH, Yao ZW, Dong Z, *et al*. ZeroQ: A novel zero shot quantization framework. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13166–13175.

12 Ma YX, Jin TS, Zheng XW, *et al*. OMPQ: Orthogonal mixed precision quantization. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 9029–9037.

13 Dong Z, Yao ZW, Gholami A, *et al*. HAWQ: Hessian aware quantization of neural networks with mixed-precision. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 293–302.

14 Huang ZY, Han X, Yu Z, *et al*. Hessian-based mixed-precision quantization with transition aware training for neural networks. Neural Networks, 2025, 182: 106910. [doi: 10.1016/j.neunet.2024.106910]

15 Martens J. New insights and perspectives on the natural gradient method. The Journal of Machine Learning Research, 2020, 21(1): 146.

16 Kunstner F, Hennig P, Balles L. Limitations of the empirical Fisher approximation for natural gradient descent. Proceedings of the 2019 International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 4158–4169.

17 Esser SK, McKinstry JL, Bablani D, *et al*. Learned step size quantization. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.

18 杨杰, 李琮, 胡庆浩, 等. 面向轻量卷积神经网络的训练后量化方法. 图学学报, 2025, 46(4): 709–718.

19 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255.

20 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

21 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520.

22 Radosavovic I, Kosaraju RP, Girshick R, *et al*. Designing network design spaces. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10425–10433.

23 Tan MX, Chen B, Pang RM, *et al*. MNasNet: Platform-aware neural architecture search for mobile. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2815–2823.

24 Wu ZGY, Wang SH, Zhang JY, *et al*. FIMA-Q: Post-training quantization for vision Transformers by Fisher information matrix approximation. Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025. 14891–14900.

25 Ma YX, Li HX, Zheng XW, *et al*. Solving oscillation problem in post-training quantization through a theoretical perspective. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 7950–7959.

26 Gong RH, Liu XL, Li YH, *et al*. Pushing the limit of post-training quantization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(7): 5556–5570. [doi: 10.1109/TPAMI.2025.3554523]

27 Feng C, Zhou GX, Wu X, *et al*. PLMQ: Piecewise linear mixed-precision quantization for deep neural networks. Neurocomputing, 2025, 651: 131000. [doi: 10.1016/j.neucom.2025.131000]

(校对责编: 张重毅)