http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2025.0411

面向无人机平台的CNN-Transformer混合模型训练后量化方法

赵呈功[1,2]，胡纪远[1,3]，杨星宇[4]，陈磊[5]，姜宏旭[3,*]

(1. 北京航空航天大学杭州创新研究院，杭州 310056；2. 北京航空航天大学 软件学院，北京 100191；
3. 北京航空航天大学 计算机学院，北京 100191；4. 内蒙古大学 计算机学院（软件学院），呼和浩特 010021；
5. 杭州萧山国际机场飞行区中心，杭州 311241)

摘 要：近年，无人机在各类低空生产、服务场景作用愈发重要。由于边缘计算平台的存储、计算资源受限，视觉模型的边端侧运行性能短板成为算法应用于以上场景的技术瓶颈。针对卷积神经网络(CNN)-Transformer混合模型激活值长尾分布导致量化精度严重退化问题，提出高效训练后量化方法。该方法通过激活值噪声补偿(ANC)和自适应难度迁移抑制异常值对量化精度的影响，在提升模型量化精度同时提高其推理速度。实验结果表明：所提方法在典型模型的8 bit量化场景下，精度损失小于1%，量化后推理速度最高提升超过200%。所提方法显著提升了边端侧设备的视觉模型推理性能，为低空无人机平台上的模型部署提供了技术支撑。

关 键 词：训练后量化；模型压缩；推理加速；异常值感知；边缘计算
中图分类号：V221+.3; TB553
文献标志码：A **文章编号**：1001-5965(2026)09-3066-09

近年来，随着低空经济的快速发展，无人机在低空物流、空中巡检、农林监测、应急通信等场景中得到广泛应用。这些任务往往对视觉感知系统提出了“高实时性、低功耗、轻量化”的严格要求：无人机等低空平台通常搭载片上系统(system-on-chip, SoC)或嵌入式AI芯片，计算资源和电池续航十分有限，却需在复杂环境下保持对目标的高精度识别和快速响应。视觉Transformer(vision Transformer, ViT)等基于自注意力机制的深度学习模型在图像分类、目标检测和语义分割等计算机视觉任务中取得了卓越性能，但其庞大的参数量和高计算开销导致在嵌入式平台上推理时出现内存占用大、功耗高、延迟长等问题，难以满足无人机对实时推理的苛刻需求。

为兼顾高精度与高效率，研究者提出了融合卷积神经网络(convolutional neural network, CNN)和视觉Transformer优势的混合架构，通过卷积层负责局部特征提取并与轻量化自注意力模块结合，实现了更优的性能，在保持全局信息建模能力的同时，有效降低了计算复杂度[1]。这类模型已在多种轻量级设计中得到验证，显示出在移动端设备上兼顾准确率与速度的潜力。即便如此，这些模型在实际部署时仍面临高计算量和较大内存占用的瓶颈。

量化技术通过将网络参数和激活值从高精度浮点数映射为低位宽整数表示，将浮点运算转换为整数运算，有效降低了模型存储与计算开销，已经成为压缩与加速模型的主流手段之一。传统针对标准Transformer架构的量化方法(如FQ-ViT、PTQ4ViT等)在通用视觉任务上取得了显著成果，但由于混合架构中深度可分离卷积产生的激活分布往往具备长尾特性，这些方法难以保证在混合模型上实现低误差量化。此外，许多硬件加速方案需要对原始

收稿日期：2025-06-27；**录用日期**：2025-09-25；**网络出版时间**：2025-09-30 13:44
网络出版地址：link.cnki.net/urlid/11.2625.V.20250930.0919.001
*通信作者. E-mail：jianghx@buaa.edu.cn

引用格式：*赵呈功，胡纪远，杨星宇，等. 面向无人机平台的CNN-Transformer混合模型训练后量化方法[J]. 北京航空航天大学学报，2026，52(9)：3066-3074. Zhao C G，Hu J Y，Yang X Y，et al. A post-training quantization method for CNN-Transformer hybrid models on UAV platforms[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52(9)：3066-3074 (in Chinese).*

网络结构进行插桩或引入特定算子, 从而增加了部署复杂性, 并对无人机算力与功耗提出挑战。因此, 如何在不引入额外算子、保持网络结构不变的前提下, 实现混合视觉模型的高效量化, 并在嵌入式平台上兼顾实时性与精度, 成为亟待解决的关键课题。

针对上述挑战, 本文提出了一种面向混合视觉模型的异常值感知的量化 (outlier-aware quantization, OAQ) 方法。该方法基于对激活值分布特性和权重量化难度的深入分析, 设计了 2 种互补优化策略: ①在推理阶段引入激活值噪声补偿 (activation noise compensation, ANC), 通过向激活值添加自适应噪声, 使深度可分离卷积输出的长尾分布得到平滑, 降低极端值对量化区间的拉拽效应; ②通过数学等效变换, 将激活端高动态范围的量化难度转移至权重量化, 并结合分层缩放与不同粒度的量化方案, 对权重进行优化, 从而简化激活端的量化过程, 并避免插入额外算子。这 2 项策略在不改变模型结构的前提下, 有效降低了推理时的计算复杂度和内存开销, 显著提高了模型在嵌入式平台上的实时推理能力, 同时保持了量化后的准确性。本文主要创新如下:

1) 从无人机边缘平台资源受限的角度出发, 提出了针对混合架构激活分布长尾特性的噪声补偿机制, 实现了对量化边界的自适应优化, 显著提高了复杂低空环境下的目标检测鲁棒性;

2) 通过数学等效变换将激活端的高动态范围量化压力转移至权重端, 并采用分层缩放策略, 减少了激活端量化误差, 无需修改网络结构即可在嵌入式设备上高效部署;

3) 本文在主流的 CNN-Transformer 混合架构模型上进行了大量实验。结果表明, 在 INT8 量化条件下, 本文 OAQ 方法在图像分类任务上实现了接近无损的预测精度, 量化后精度下降小于 1%;

1　相关工作

1.1　模型架构发展

无人机等边缘感知设备对视觉模型的实时性和资源消耗提出了更高要求。深度学习模型的架构演变正是为满足这类应用需求, 从最早的卷积神经网络到近年来兴起的视觉 Transformer 模型。最初, 卷积神经网络凭借局部特征提取的能力在计算机视觉领域取得了显著的成功[2-4], 如 AlexNet[2]、VGG[4]、Resnet[5] 等。但 CNN 在长程依赖处理上存在局限[6], 这促使了对新架构的探索。于是, Swin Transformer[6]、ViT[7]、DeiT[8] 等模型应运而生。这些模型通过自注意力机制取代了卷积层, 使得模型能够更有效地建模整个图像的全局依赖关系[9]。与传统的卷积神经网络不同, ViT 将输入图像分割成固定大小的 patch 序列, 然后, 通过多层 Transformer 编码器来捕捉全局依赖关系。这种设计不仅简化了模型结构, 还显著增强了对长距离特征关联的理解能力, 并在大规模数据集上训练时展示了超越传统 CNN 的性能, 取得了领先成果[10-11]。这些 Transformer 模型, 凭借其强大的全局信息建模能力, 已经成为深度学习中的一种主流方法。

Transformer 模型的计算复杂度高及参数规模大。于是研究人员探索了结合 CNN 与 Transformer 架构的方法, 旨在利用 CNN 的高效特征提取能力及 Transformer 处理长程依赖的优势, 从而提出了多种轻量化设计方案, 提升其在移动设备上的部署效率。例如, MobileViT[12] 通过融合局部卷积与全局自注意力机制, 实现了高效分类, 特别适合移动设备; EdgeNeXT[13] 通过引入分裂深度转置注意力编码器, 实现了在较低计算需求下优异性能表现; EfficientFormer[14] 通过重新设计 ViT 架构, 采用维度一致的设计范式和延迟驱动的瘦身方法, 显著提升了推理速度。CNN-Transformer 混合架构通过结合局部和全局信息的优势, 在特定应用场景中提供了良好的平衡。然而, 即便在混合架构下, 深度可分离卷积和自注意力模块的组合仍会带来较高的计算开销与内存占用, 制约了无人机等低空平台的长续航与实时性能。由此可见, 针对上述混合视觉模型的进一步压缩与加速, 成为低空经济环境中模型部署的关键技术挑战。

1.2　训练后量化

无人机或其他低空平台对视觉模型的部署环境往往资源受限, 模型推理必须兼顾实时性与功耗控制。量化技术通过将网络参数和激活值从浮点数表示映射为低位宽整数表示, 可显著降低模型的存储需求与计算开销, 是边缘设备实现高效推理的重要手段。量化可分为量化感知训练 (quantization-aware training, QAT)[15-16] 和后训练量化 (post-training quantization, PTQ)[17-19]。QAT 可在量化后保持较高的准确率, 但需要访问完整的训练数据集, 并进行额外的训练。PTQ 直接将预训练神经网络的高精度权重和激活值压缩到低精度值来加速推理, 而不需要重新训练, 比 QAT 更加灵活与高效。

近年来, 深度学习模型的量化研究从传统的卷积神经网络逐步扩展到 Transformer 模型, 推动了 PTQ 在不同架构中的广泛应用。例如, Jacob 等[20] 通过整数算术推理并优化训练过程, 提升了设备延

迟与准确性之间的平衡。Nagel 等[21]提出的改进 AdaRound 权值取整机制，通过优化权重量化方法，在 ResNet 模型中实现了更高效的 4 位权重量化。对于 Transformer 模型量化，Liu 等[17]通过引入排名损失保持自注意力机制的顺序，并采用混合精度量化方案，优化视觉 Transformer 的存储和计算成本。随后，PTQ4ViT[19]提出了一种双重均匀量化方法，通过后训练量化有效减少了视觉 Transformer 中的量化误差，同时提高了推理效率。

然而，面对 CNN-Transformer 混合架构的 ViT 时，现有 PTQ 技术在低空场景下的应用仍存在挑战。鉴于此，Lee 等[22]提出一种用于混合视觉 Transformer 的后训练量化方法 Q-HyViT，针对混合 ViT 中的高动态范围、零点溢出、不同的归一化类型和参数较少等问题，通过桥接块重构和误差最小化，自动选择量化尺度因子、粒度及量化方案，显著减少了量化误差。Kim 等[23]提出了一种面向硬件友好的 CNN-Transformer 混合网络训练后量化方法 HyQ，通过引入量化感知的分布缩放来解决卷积层中的通道间方差导致的离群值问题，并在 Transformer 模块中用整数线性近似代替了计算量大的 softmax 函数。虽然 Q-HyViT[22]可以避免量化 CNN-Transformer 混合模型出现精度严重退化的问题，但是混合量化对推理加速是非常不友好的，需要有专用的内核支持。HyQ[23]使用整数线性近似代替标准的 softmax 函数，减少了计算量，但是在目前主流的推理框架中，整数线性近似的 softmax 函数支持尚不完善，导致实际应用时需要额外的定制化实现。

2 本文 OAQ 方法

模型量化旨在将模型的权重和激活从浮点数转换为低精度的整数表示，从而减少模型的存储和计算需求。量化器一般使用均匀对称量化，因为大多数硬件平台都很好地支持均匀量化，其将输入激活值 $\boldsymbol{X}$ 映射到 b 位整数值 $Q(\boldsymbol{X}|b)$ 上：

$$Q(\boldsymbol{X}|b)=\mathrm{clip}\left(\lfloor \boldsymbol{X}/s \rfloor+z,0,2^b-1\right) \tag{1}$$

式中：clip 为裁剪函数，其作用是将结果限制在区间 $[0,2^b-1]$ 内；s 为比例因子，z 为零点，均为量化所需要的参数，通常由 $\boldsymbol{X}$ 中元素的下限 l 和 $\boldsymbol{X}$ 中元素的上限 u 确定，可以表示为

$$\begin{cases} l=\min(X_i) \\ u=\max(X_i) \end{cases} \tag{2}$$

$$s=(u-l)/(2^b-1) \tag{3}$$

$$z=\mathrm{clip}\left(\lfloor -l/s \rfloor,0,2^b-1\right) \tag{4}$$

在边缘计算中，高效的模型设计至关重要，因此，很多轻量化模型在设计时采用了深度可分离卷积来减少计算复杂度。深度可分离卷积在参数量和计算量方面具有显著优势，但是由深度可分离卷积在处理每个输入通道时，独立计算每个通道的特征图。这导致了通道之间的相互依赖性较弱，通道间数据的方差较大，激活分布存在长尾特性。EfficientFormer-L1 中间阶段和最后阶段全连接层的卷积层的输入激活的 95% 的激活都集中在 0 附近，但是数据范围分别为 [−0.17, 19.57] 和 [−55, 22.34]，具有非常明显的长尾分布。这种极端的激活经过式 (1) 的计算后，会将大部分的数据映射为 0，影响量化后模型的整体性能。在本节中，介绍一种两阶段的训练后量化方法，无需手工设计参数，可实现在精度退化很少的情况下量化模型。为实现这一目标，在 2.1 节和 2.2 节设计了 2 种针对异常值的量化方法，包括：ANC 使得激活值的分布更加平坦；自适应难度迁移 (adaptive difficulty migration, ADM) 通过将量化难度迁移到权重，使得激活值更易量化。如图 1 所示，其中，$\boldsymbol{Y}_{\mathrm{fp32}}$ 为浮点模型的参考输出；$\hat{\boldsymbol{Y}}_{\mathrm{wqaq}}$ 为融合 ANC/ADM 参数后的量化模型输出；$L(\cdot)$ 表示损失函数，用于衡量量化输出与浮点输出之间的差异，并据此更新 ANC/ADM 参数。

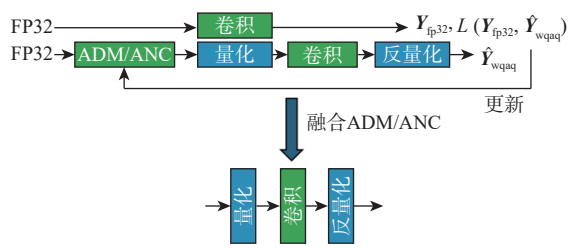


图 1 ANC/ADM 的优化过程

Fig. 1 Optimization process of ANC/ADM

2.1 激活值噪声补偿

激活值噪声补偿的核心思想是在对视觉 Transformer 的激活值进行量化之前，向激活值中添加噪声补偿。这种噪声补偿能够改变原始激活值的分布，使其更加平坦，减少长尾效应。本文将噪声应用于输入激活值 $\boldsymbol{X}$ 和权重 $\boldsymbol{W}\in\mathbf{R}^{C_{\mathrm{out}}\times C_{\mathrm{in}}\times K_{\mathrm{h}}\times K_{\mathrm{w}}}$，$C_{\mathrm{in}}$ 为输入通道数；C_{out} 为输出通道数；K_{h} 与 K_{w} 分别为卷积核的高度和宽度。使用噪声补偿的卷积层操作表示如下：

$$\boldsymbol{Y}=\boldsymbol{W}\boldsymbol{X}+\boldsymbol{B}=\boldsymbol{W}(\boldsymbol{X}-\boldsymbol{\beta})+(\boldsymbol{B}+\boldsymbol{W}\boldsymbol{\beta}) \tag{5}$$

式中：$\boldsymbol{Y}$ 为输出；偏置 $\boldsymbol{B}\in\mathbf{R}^{1\times C_{\mathrm{out}}}$；噪声 $\boldsymbol{\beta}\in\mathbf{R}^{1\times C_{\mathrm{in}}}$。在量化之前，需要初始化噪声补偿。噪声补偿的初始化需要结合激活值的分布，并且噪声补偿不宜过大，因为过大的噪声补偿会使激活值波动剧烈，导

致量化后模型精度退化严重的问题。因此, 设计了如下初始化噪声补偿的方式:

$$\beta_i = \mu_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij} \tag{6}$$

式中: x_{ij}为第 i 个通道的第 j 个激活值; N 为激活值的数量; β_i为第 i 个通道噪声补偿。该噪声补偿 β_i通过优化过程进行调整, 从而使激活值的分布更加适合量化, 减少极端值的影响。为优化噪声补偿 $\boldsymbol{\beta}$, 本文定义了一个损失函数, 该函数包括量化误差和激活值分布平滑 2 个方面。其中, 量化误差损失度量了加噪声模型与未加噪声模型之间的量化误差。具体地, 令 $Q(\cdot)$为量化操作, 则量化误差损失函数 $L_{\text{quant}}(\boldsymbol{\beta})$被定义为

$$L_{\text{quant}}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{j=1}^{N} |Q\left(Y_{\text{noisy}}^{(j)}\right) - Q\left(Y_{\text{orig}}^{(j)}\right)|^2 \tag{7}$$

式中: $\boldsymbol{Y}_{\text{noisy}}$和 $\boldsymbol{Y}_{\text{orig}}$分别为卷积层带噪声和未带噪声的输出。此外, 为避免激活值分布的极端波动, 通过计算激活值的动态范围和方差来平滑其分布, 则平滑损失 $L_{\text{smooth}}(\boldsymbol{\beta})$被定义为

$$L_{\text{smooth}}(\boldsymbol{\beta}) = \lambda_1 |X_{i,\max} - X_{i,\min}| + \lambda_2 \sigma_i^2 \tag{8}$$

式中: λ_1和 λ_2为超参数, 用来控制动态范围和方差的权重; $|X_{i,\max} - X_{i,\min}|$为激活值的动态范围, σ_i^2为激活值的方差。最终的综合损失函数为

$$L_{\text{total}}(\boldsymbol{\beta}) = L_{\text{quant}}(\boldsymbol{\beta}) + L_{\text{smooth}}(\boldsymbol{\beta}) \tag{9}$$

噪声补偿仅在量化前进行优化, 在获得最佳噪声补偿后, 本文将噪声补偿与激活值结合以此来调整每个通道的激活值分布。在校准阶段, 使用损失函数 (9) 进行优化以获取最佳的 ANC。在校准完成后, 将 ANC 合并到量化器中, 并对该卷积层的偏置进行更改。此外, 在推理阶段, 噪声补偿作为固定参数被应用于计算, 不再进行动态更新, 这使得推理过程的计算效率保持高效。

2.2　自适应难度迁移

激活值添加噪声补偿后, 其分布整体变得更加平坦。然而, 仍然存在部分通道的激活值相较于其他通道依然偏大, 这种不均匀性可能在量化过程中导致精度损失。此外, 权重的分布通常更均匀且数值范围较小, 具备更高的量化友好性。由此本文提出自适应难度迁移, 不需要人为设置难度迁移的因子。其核心思想是在量化之前, 使用迁移因子将激活值的量化难度迁移到权重中, 使激活值容易量化。本文将迁移因子应用于输入激活 $\boldsymbol{X}$ 和权重 $\boldsymbol{W}$。使用迁移因子的卷积层操作表示如下:

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X} + \boldsymbol{B} = \boldsymbol{W}\boldsymbol{\delta}(\boldsymbol{X}/\boldsymbol{\delta}) + \boldsymbol{B} \tag{10}$$

式中: 难度迁移因子 $\boldsymbol{\delta} \in \mathbf{R}^{1\times C_{\text{in}}}$。确定最佳的难度迁移因子 $\boldsymbol{\delta}$是整个量化过程中至关重要的一步。为使得难度迁移因子在减少激活值量化难度的同时不过多增加权重的量化难度, 本文设计了一个有效的优化策略, 以在训练过程中自动调节 $\boldsymbol{\delta}$, 使其适应不同通道的激活值分布特性。首先, 本文使用激活值的标准差初始化难度迁移因子。对于每个通道, 初始化迁移因子为该通道激活值的标准差 σ_{X_i}:

$$\delta_i = \sigma_{X_i} = \frac{1}{N}\sum_{j=1}^{N}(X_j - \mu_{X_i})^2 \tag{11}$$

为确定最佳的迁移因子, 本文采用基于量化误差的优化方法。在训练过程中, 本文不断调整迁移因子的值, 使得量化误差最小化。量化误差可以通过均方误差 (mean squared error, MSE) 来度量:

$$L_{\text{quant}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{Y}}_i - \boldsymbol{Y}_i)^2 \tag{12}$$

式中: N 表示校准集中样本数, $\hat{\boldsymbol{Y}}_i$为量化后的输出, $\boldsymbol{Y}_i$为真实输出。通过优化损失函数 L_{quant}, 迁移因子 $\boldsymbol{\delta}$会被调整到最佳值, 从而确保量化过程中的误差最小化。

与 2.1 节中操作类似, $\boldsymbol{\delta}$一旦确定后, 在推理过程中是固定的。得益于 2.1 节对于激活值的处理, 本文可以更快地获得最佳的 $\boldsymbol{\delta}$。在校准阶段, 使用平均绝对误差 (mean absolute error, MAE) 进行优化以获取最佳的 ADM。在校准完成后, 将 ADM 合并到量化器中, 并调整卷积层的权重。此外, 与 FQ-ViT[18] 和 Q-HyViT[22] 不同的是, OAQ 采用了均匀量化方法, 因此, 不需要依赖专门的硬件, 可以直接在主流的边缘设备上运行, 极大地提高了其实际应用的可行性。

3　实验结果与分析

3.1　数据集和评价指标

本文选择的是用于分类任务的 imagenet-1k 数据集, 有 1.2×10^6 个训练图像和 1 000 个类的 50 000 个验证图像。此外, 为获得最佳的噪声补偿和难度迁移因子, 本文从训练集中挑选出 1 000 个未标记的图像作为 PTQ 方法的校准集。该校准集可确保量化参数得到有效的微调, 从而提高量化模型的性能。本文对预训练的 CNN-Transformer 混合架构的视觉模型执行统一的 8 位 PTQ。本文量化的模型的来源包括两类, 一类是从 timm 库提供的 Mobile-ViT[12] 和 MobileViTv2[24], 另一类是开源代码仓库中的 EfficientFormer[14]。本文使用官方提供的浮点模型的 Top-1 准确率作为基线。本文也量化了全连接

层，包括第1个投影层和最后1个预测层。因为模型中并不是每一层的激活值都有异常激活值，所以本文只为出现异常值的层添加了噪声补偿及量化难度迁移。

3.2 实验平台及参数设置

为确保本文实验结果的准确性与可复现性，所有实验均在统一的软硬件环境下进行。实验所采用的计算平台操作系统为Ubuntu22.04.4LTS，处理器为Intel Core i9-13900K，搭载NVIDIA GeForce RTX 4090显卡，具备强大的并行计算能力和显存资源。在软件环境方面，实验基于Python 3.10.15编写，并使用PyTorch 2.5.0深度学习框架，配套CUDA 12.4作为GPU加速后端。实验中使用的是完整的校准集，学习率为0.001；对于常规线性模块采用层级量化，对于LayerNorm等非线性模块采用通道级量化。上述环境配置为获取自适应噪声与难度迁移因子提供了稳定高效的支持，确保了实验结果的可靠性与一致性。

3.3 性能分析

为全面评估本文OAQ方法在模型量化中的性能优势，本文与目前最先进的几种量化方法进行了对比，包括FQ-ViT[18]、PTQ4ViT[19]、Q-HyViT[22]、HyQ[23]、RepQ-ViT[25]、EasyQuant[26]。通过对比分析，展示了本文OAQ方法在多种CNN-Transformer混合模型上的量化效果，并进一步证明了其在精度方面的优越性。

本文选择了多个常见的模型架构进行比较测试，涵盖了从轻量级到较大体积的多种网络结构，如MobileViT系列模型和EfficientFormer模型。如表1～表3所示，在8 bit量化过程中，本文OAQ方法的表现明显优于当前领先的量化方法，尤其是在MobileViT-s模型上，精度仅退化了0.31%，这一结果相较于其他量化方法具有显著的优势。例如，RepQ-ViT在该模型上表现较差，精度下降幅度较大，而本文OAQ方法能够有效抑制量化误差，显著降低精度退化。

表1 MobileViT模型量化后精度

Table 1 Accuracy of MobileViT after quantization

方法	权重/激活值	精度Top-1/%		
		MobileViT-xxs	MobileViT-xs	MobileViT-s
全精度	32/32	69.00	74.80	78.40
FQ-ViT[18]	8/8	66.46	68.28	77.67
PTQ4ViT[19]	8/8	37.75	65.52	68.19
RepQ-ViT[24]	8/8	1.85	41.96	50.01
Q-HyViT[22]	8/8	67.20	73.89	77.72
HyQ[23]	8/8	68.15	73.99	77.93
EasyQuant[25]	8/8	36.13	73.16	74.21
本文OAQ方法	8/8	68.65	74.35	78.09

表2 MobileViTv2模型量化后精度

Table 2 Accuracy of MobileViTv2 after quantization

方法	权重/激活值	精度Top-1/%		
		MobileViTv2-50	MobileViTv2-75	MobileViTv2-100
全精度	32/32	70.20	75.60	78.10
FQ-ViT[18]	8/8	67.66	69.56	77.15
PTQ4ViT[19]	8/8	39.39	65.54	51.02
RepQ-ViT[24]	8/8	26.60	55.52	40.85
Q-HyViT[22]	8/8	69.89	75.29	77.63
HyQ[23]	8/8	69.16	74.47	76.63
EasyQuant[25]	8/8	66.80	62.91	69.34
本文OAQ方法	8/8	69.97	75.31	77.86

值得注意的是，随着模型规模的增大，本文OAQ方法在保持高精度恢复方面表现得更加稳健。在MobileViT-xs模型和MobileViT-s模型上，本文OAQ方法的量化精度不仅与主流动态量化方案不相上下，而且在实际应用中能够将性能损失进一步压缩至极小范围。本文OAQ方法未引入额外算子，且自适应噪声补偿与量化难度迁移在训练阶段参数一旦确定后即保持不变，因此，在低算力芯片上也能稳定部署而不增加额外计算开销。通过这些实验，可以清楚地看到，本文OAQ方法在量化大规模模型时具有更强的鲁棒性，并且能够处理更复杂的网络架构。

在 MobileViTv2 和 EfficientFormer 等模型上, 本文 OAQ 方法在量化精度方面表现优异。如表 2 和表 3 所示, 本文 OAQ 方法在各模型上均获得最佳 Top-1 精确, 尤其在结构更复杂的 EfficientFormer 模型中, 有效抑制了异常激活值带来的精度损失。同时, 如图 2 所示, 经 INT8 量化后, 这些模型在 Jetson AGX Xavier 上推理速度显著提升, 帧率整体高于未量化模型, 且模型规模越大加速越明显, 验证了本文 OAQ 方法在边缘设备上的高效性与适用性。

3.4　消融实验

为验证本文 OAQ 方法中 ANC 和 ADM 的作用及其协同效果, 本文对 MobileViT 系列模型进行了全面的消融实验。通过逐步移除这些模块并观察量化精度的变化, 本文分析了 ANC 和 ADM 各自的贡献及其协同作用对量化性能的提升效果。图 3 展示了各组件的重要性及其在不同模型中的性能

表 3　EfficientFormer 模型量化后精度

Table 3　Accuracy of EfficientFormer after quantization

方法	权重/激活值	Top-1精度/%		
		EfficientFormer-L1	EfficientFormer-L3	EfficientFormer-L7
全精度	32/32	80.20	82.40	83.30
FQ-ViT[18]	8/8	66.63	81.85	82.47
PTQ4ViT[19]	8/8	79.14	80.31	81.56
RepQ-ViT[24]	8/8	79.67	81.98	82.47
HyQ[23]	8/8	78.55	82.26	82.66
EasyQuant[25]	8/8	78.42	81.03	82.27
本文OAQ方法	8/8	80.09	82.14	83.15

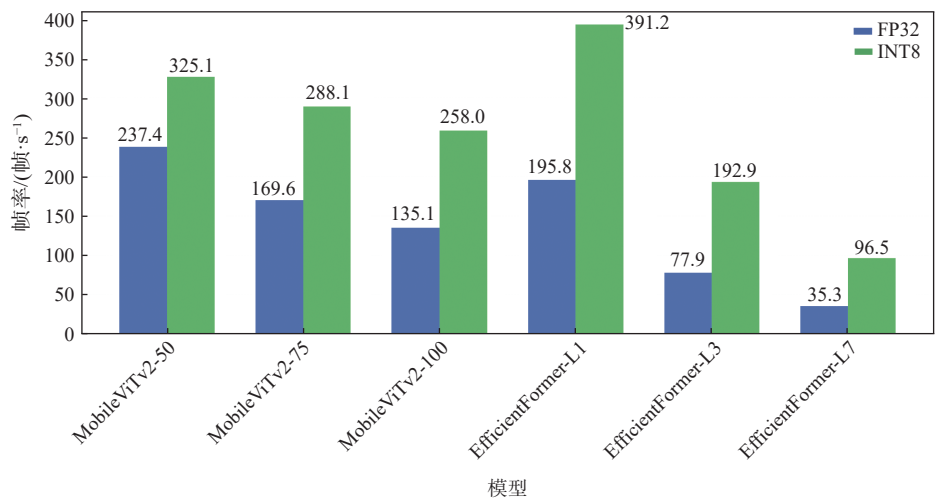


图 2　比较量化前后各模型的推理速度

Fig. 2　Comparison of inference speed before and after quantization across different models

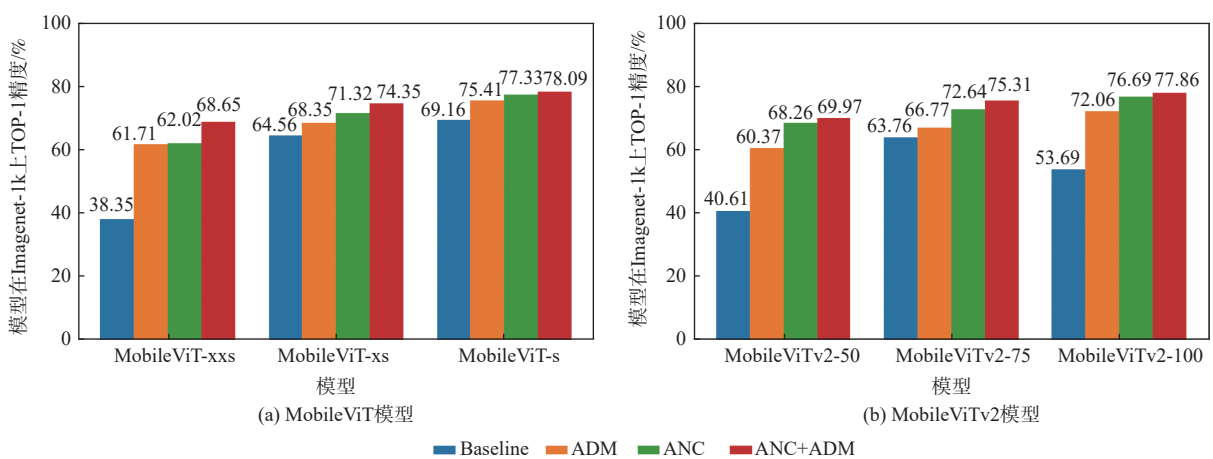


图 3　ANC 和 ADM 在 MobileViT 模型和 MobileViTv2 模型上的消融实验结果

Fig. 3　Ablation results of ANC and ADM on MobileViT and MobileViTv2 model

表现，其中，Baseline 表示在未引入 ANC 与 ADM 机制时，量化后模型所达到的精度。

移除 ANC 的实验分析。当移除 ANC 后，MobileViT 系列模型的量化精度出现了一定程度的下降。例如，在 MobileViT-s 模型上，Top-1 精度下降了 2.68%，在 MobileViT-xs 模型和 MobileViT-xxs 模型上分别下降了 6.00% 和 6.94%。这表明，虽然 ADM 依然可以对激活值的难度迁移进行优化，但没有 ANC 提供的激活值平坦化作用，模型在处理异常值时会受到更大的影响，导致量化误差增加。

移除 ADM 的实验分析。相比于移除 ANC，移除 ADM 对模型量化精度的影响更加显著。例如，在 MobileViT-s 模型上，Top-1 精度下降了 0.76%，在 MobileViT-xs 模型和 MobileViT-xxs 模型上分别下降了 6.63% 和 3.03%。这表明，ADM 在量化中效应更大，尤其是在处理复杂特征分布和高动态范围的激活值时，其优化效果尤为显著。没有 ADM 时，即使 ANC 对激活值分布进行了平滑处理，仍然会造成较大的量化误差。尤其是在 MobileViT-xs 模型和 MobileViT-xxs 模型这样的较小模型中，激活值分布本身已经相对复杂，移除 ADM 后精度下降幅度更大，进一步凸显了 ADM 的重要性。

ANC 和 ADM 的协同作用。实验结果表明，ANC 和 ADM 的协同作用显著增强了量化的整体表现。具体而言，ANC 的引入不仅抑制了激活值中的异常值，还通过动态范围的收敛和平坦化，降低了激活值分布对量化的干扰，为 ADM 学习到最优的难度迁移因子奠定了基础。如图 3 所示，在包含 ANC 和 ADM 这 2 个组件时，MobileViT 模型和 MobileViTv2 模型的量化精度相比移除 ANC 或 ADM 时都有所提升，充分验证了二者的互补性和协同优化能力。

总之，ANC 的异常值平滑作用为 ADM 提供了更加平稳的激活值分布，从而使 ADM 能够更高效地学习难度迁移因子。而 ADM 则进一步优化了难度较高的特征分布，将 ANC 的改进效果最大化。两者的联合优化显著增强了本文 OAQ 方法在量化过程中对激活值异常的鲁棒性。

3.5 激活值分布优化

激活分布中的异常值会放大动态范围，从而加剧量化误差，因此，通过抑制异常激活、使分布更加平坦，是降低量化难度的一种有效策略[27]。为验证激活噪声补偿和自适应难度迁移在优化激活值分布、减小异常值对量化精度影响方面的有效性，本文对 EfficientFormer 模型中不同阶段的激活值分布进行了分析。如图 4 所示，图中的激活值分布分别对应于 EfficientFormer 模型中间阶段全连接层和最后阶段全连接层的卷积输入激活值。

首先，图 4 中的第 1 行展示了 EfficientFormer 模型中间阶段全连接层激活值分布的变化。如图 4(a)所示，原始激活值呈现出显著的长尾分布，存在较多的异常值，这些异常值会严重影响量化精度。如图 4(b)所示，在引入噪声补偿后，激活值的分布明显趋于平坦，极端值的数量显著减少，激活值的范围更加均匀。这表明，激活噪声补偿通过对激活值添加噪声补偿，成功降低了激活值的方差，并有效减小了异常值的影响。如图 4(c)所示，进一步引入自适应难度迁移后，激活值的分布变得更加集中，且没有明显的异常值或长尾效应。图 4(d)～图 4(f)展示了 EfficientFormer 最后阶段全连接层的激活值分布，其优化趋势与中间阶段相似，进一步

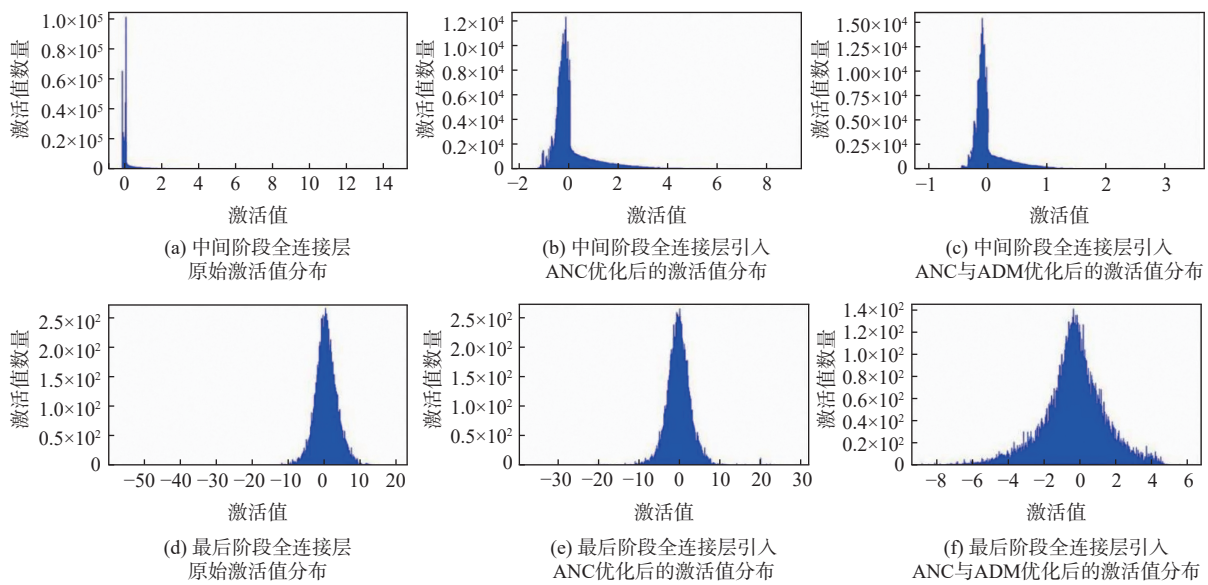


图 4 Efficientformer 中间和最后阶段全连接层卷积的激活值分布变化

Fig. 4 Change in activation value distribution of fully connected layer convolutions in middle and final stages of Efficientformer

证明了激活噪声补偿和自适应难度迁移策略在不同网络深度上的鲁棒性和有效性。

如图 4(d)所示, 最后阶段原始的激活值分布呈现出宽泛且低矮的形态。虽然分布形态相对对称, 但峰度较低, 表明激活值分散度高, 存在一定数量的离群点造成动态范围的扩大, 为量化带来了挑战。如图 4(e)所示, 在应用激活噪声补偿后, 激活值分布得到了显著的集中。其分布范围大幅收缩, 峰值高度显著增加, 且形态变得更加陡峭。这一结果有力地证明了激活噪声补偿通过降低激活值的离群值, 有效地将分散的分布平坦化并收敛到零值附近, 从而为后续的低比特量化提供了更为紧凑的量化区间。最后, 如图 4(f)所示, 在进一步集成自适应难度迁移策略后, 激活值分布达到了最优的集中程度。分布的范围进一步被压缩, 峰值高度达到最高, 分布形态高度集中且对称。自适应难度迁移通过将激活值量化的难度转移到权重参数上, 允许激活值本身向更利于量化的分布变化, 从而实现了对激活值分布的彻底优化, 使得异常值基本消除。这种高度集中的分布形态极大地减小了量化误差的潜在风险, 为实现高精度低比特量化奠定了基础。

激活值噪声补偿和自适应难度迁移在优化激活值分布、减少异常值方面发挥了关键作用。激活值噪声补偿通过降低激活值的方差和消除异常值, 使激活值分布更加平坦; 而自适应难度迁移则将量化难度从激活值迁移到权重中, 进一步简化了激活值的量化难度, 有效保障了模型的量化性能。实验结果表明, 这 2 种方法的结合成功解决了输入激活值分布异常的问题, 有利于减少量化误差, 并为高效部署模型提供了强有力的激活值分布优化工具。

4　结　论

1) 本文 OAQ 方法通过激活值噪声补偿和量化难度迁移抑制异常值以降低量化误差, 实现 CNN-Transformer 混合模型在边端侧计算平台的高精度、快速推理。

2) 多项基于典型模型的实验表明, 本文 OAQ 方法显著提升了量化后的推理精度并保持了量化模型的高效推理特性, 为在无人机等计算资源受限、对推理效率存在较高要求的设备上部署模型提供了高精度、高效率的量化方案, 而无需引入额外算子支持。

参考文献（References）

[1] Dai Z, Liu H, Le Q V, et al. Coatnet: marrying convolution and attention for all data sizes[J]. Advances in neural information processing systems, 2021, 34: 3965-3977.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[3] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//Proceedings of the Computer Vision-ECCV. Berlin: Springer, 2014: 818-833.

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2025-06-20]. https://arxiv.org/abs/1409.1556.

[5] Shafiq M, Gu Z Q. Deep residual learning for image recognition: a survey[J]. Applied Sciences, 2022, 12(18): 8972.

[6] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 9992-10002.

[7] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-12-22)[2025-06-20]. https://arxiv.org/abs/2010.11929.

[8] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[EB/OL]. (2020-12-23)[2025-06-20]. https://arxiv.org/abs/2012.12877.

[9] Selva J, Johansen A S, Escalera S, et al. Video transformers: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12922-12943.

[10] Hashiguchi R, Tamaki T. Vision transformer with cross-attention by temporal shift for efficient action recognition[EB/OL]. (2022-04-01)[2025-06-21]. https://arxiv.org/abs/2204.00452.

[11] Manzari O N, Ahmadabadi H, Kashiani H, et al. MedViT: a robust vision transformer for generalized medical image classification[J]. Computers in Biology and Medicine, 2023, 157: 106791.

[12] Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer[EB/OL]. (2021-10-05)[2025-06-21]. https://arxiv.org/abs/2110.02178.

[13] Maaz M, Shaker A, Cholakkal H, et al. EdgeNeXT: efficiently amalgamated CNN-transformer architecture forMobile vision applications[C]//Proceedings of the Computer Vision-ECCV Workshops. Berilin: Springer, 2023: 3-20.

[14] Evangelidis G, Hu J, Li Y Y, et al. EfficientFormer: vision transformers at MobileNet speed[C]//Proceedings of the Advances in Neural Information Processing Systems 35. San Diego: Neural Information Processing Systems Foundation, Inc., 2022: 12934-12949.

[15] Zhang D Q, Yang J L, Ye D, et al. LQ-nets: learned quantization for highly accurate and compact deep neural networks[C]//Proceedings of the Computer Vision-ECCV. Berlin: Springer, 2018: 373-390.

[16] Dong P Y, Lu L, Lyu C, et al. PackQViT: faster sub-8-bit vision transformers via full and packed quantization on the mobile[C]//Proceedings of the Advances in Neural Information Processing Systems 36. San Diego: Neural Information Processing Systems Foundation, Inc., 2023: 9015-9028.

[17] Liu Z H, Wang Y H, Han K, et al. Post-training quantization for vision transformer[EB/OL]. (2021-06-27)[2025-06-21]. https://arxiv.org/abs/2106.14156.

[18] Lin Y, Zhang T Y, Sun P Q, et al. FQ-ViT: post-training quantization for fully quantized vision transformer[EB/OL]. (2021-11-27)

[2025-06-21]. https://arxiv.org/abs/2111.13824.

[19] Yuan Z H, Xue C H, Chen Y Q, et al. PTQ4ViT: post-training quantization for vision transformers with twin uniform quantization[C]//Proceedings of the Computer Vision-ECCV. Berlin: Springer, 2022: 191-207.

[20] Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2704-2713.

[21] Nagel M, Ali Amjad R, Van Baalen M, et al. Up or down adaptive rounding for post-training quantization[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 7197-7206.

[22] Lee J, Kwon Y, Park S, et al. Q-HyViT: post-training quantization of hybrid vision transformers with bridge block reconstruction for IoT systems[J]. IEEE Internet of Things Journal, 2024, 11(22): 36384-36396.

[23] Kim N J, Lee J, Kim H. HyQ: hardware-friendly post-training quantization for CNN-transformer hybrid networks[C]//Proceedings of the 33th International Joint Conference on Artificial Intelligence . New York: ACM, 2024: 4291-4299.

[24] Mehta S, Rastegari M. Separable self-attention for mobile vision transformers[EB/OL]. (2022-06-06)[2025-06-21]. https://arxiv.org/abs/2206.02680.

[25] Li Z K, Xiao J R, Yang L W, et al. RepQ-ViT: scale reparameterization for post-training quantization of vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2024: 17181-17190.

[26] Wu D, Tang Q, Zhao Y L, et al. EasyQuant: post-training quantization via scale optimization[EB/OL]. (2020-06-30)[2025-06-21]. https://arxiv.org/abs/2006.16669.

[27] Chen J M, Chao Y H, Wang Y J, et al. QuantTune: optimizing model quantization with adaptive outlier-driven fine tuning[EB/OL]. (2024-03-11)[2025-06-21]. https://arxiv.org/abs/2403.06497.

A post-training quantization method for CNN-Transformer hybrid models on UAV platforms

ZHAO Chenggong[1, 2], HU Jiyuan[1, 3], YANG Xingyu[4], CHEN Lei[5], JIANG Hongxu[3, *]

(1. Hangzhou Innovation Institute of Beihang University, Hangzhou 310056, China;
2. College of Software, Beihang University, Beijing 100191, China;
3. College of Computer Science, Beihang University, Beijing 100191, China;
4. College of Computer Science (College of Software), Inner Mongolia University, Hohhot 010021, China;
5. Hangzhou Xiaoshan International Airport, Hangzhou 311241, China)

Abstract: In recent years, drones have played increasingly important roles in various low-altitude production and service scenarios. Due to the limited storage and computing resources of edge computing platforms, the deployment performance of vision models becomes a bottleneck for various algorithm applications. The long-tail distribution of activation values in convolutional neural network (CNN)-Transformer hybrid models causes substantial quantization accuracy deterioration, an efficient post-training quantization method is proposed, which this study addresses by suggesting an effective post-training quantization technique. This method uses activation noise compensation (ANC) and adaptive difficulty migration to suppress the impact of outliers on quantization accuracy and improve inference efficiency. According to experimental results, the proposed method achieves less than 1% accuracy loss on typical models under 8-bit quantization, and it improves inference time after quantization by at most 200%. In summary, the proposed method significantly enhances vision model inference performance on edge devices and supports model deployment on low-altitude drone platforms.

Keywords: post-training quantization; model compression; inference acceleration; outlier-aware; edge computing

Received: 2025-06-27; **Accepted**: 2025-09-25; **Published Online**: 2025-09-30 13:44

URL: link.cnki.net/urlid/11.2625.V.20250930.0919.001

* **Corresponding author**. E-mail: jianghx@buaa.edu.cn