ISSN: 3106-1540（ Print ）
3106-1559（ Online ）
Modern Development Research 现代发展研究
Vol.2, No.3, March 2026

E-mail:xdfzyjzz@163.com
http://www.hkxdcbs.com
Tel:0371-63705055

面向边缘计算的自适应量化与 BN 折叠视觉感知框架研究

胡雨　张泽民　梁冰锐 *
广西机电职业技术学院人工智能技术学院

摘　要：深度卷积神经网络本身的高参数量和计算开销，与嵌入式设备在存储、功耗、算力方面的严格约束之间长期存在结构性矛盾，制约了边缘智能的落地进程。模型量化是当前缓解这一矛盾的主流技术路径之一，在降低内存占用和加快推理速度方面已有较多工程验证，但位宽压缩到 4bit 及以下时，精度退化和训练收敛困难的问题仍未得到根本性解决。本文提出并实现了面向边缘部署的轻量化系统 NeuroEdge-Quant，采用 LSQ-BN 算法，将单 CNN 前向融合机制与权重分布感知的步长自适应初始化相结合，在 NVIDIA Jetson 与 Rockchip NPU 两类平台上完成系统验证。ImageNet 基准测试表明：系统在 W4A8（4-bit 权重、8-bit 激活）配置下，ResNet-50 模型体积压缩至原始的 1/8，推理延迟降低超过 60%，Top-1 精度损失仅 1.5%；与标准 PTQ 及传统 QAT 基线相比，精度损失分别降低 6.2 和 9.2 个百分点，在严格能效约束下实现高精度低比特推理部署。

关键词：边缘计算；模型量化；LSQ-BN 算法；BN 折叠；卷积神经网络；实时推理；嵌入式系统
DOI：10.65976/3106-1540.2026.03.018

Research on Adaptive Quantization and BN Folding Visual Perception Framework for edge computing

Hu Yu　Zhang Zemin　Liang Bingrui*
Artificial Intelligence Technology College of Guangxi Mechanical and Electrical Vocational and Technical College

Abstract: The high parameter count and computational overhead of deep convolutional neural networks have long been structurally contradictory to the strict constraints on storage, power consumption, and computing power of embedded devices, which have hindered the implementation process of edge intelligence. Model quantization is currently one of the mainstream technological paths to alleviate this contradiction. There have been many engineering validations in reducing memory usage and accelerating inference speed. However, when the bit width is compressed to 4 bits or less, the problems of accuracy degradation and difficulty in training convergence have not been fundamentally solved. This article proposes and implements a lightweight system called NeuroEdge Quant for edge deployment. The LSQ-BN algorithm is used to combine the forward fusion mechanism of a single CNN with the step size adaptive initialization of weight distribution perception. The system is validated on two platforms, NVIDIA Jetson and Rockchip NPU. The ImageNet benchmark test shows that under the W4A8 (4-bit weight, 8-bit activation) configuration, the ResNet-50 model volume is compressed to 1/8 of the original, the inference delay is reduced by more than 60%, and the Top-1 accuracy loss is only 1.5%; Compared with standard PTQ and traditional QAT baselines, the accuracy loss is reduced by 6.2 and 9.2 percentage points, respectively, achieving high-precision low bit inference deployment under strict energy efficiency constraints.

Keywords: edge computing; Model quantification; LSQ-BN algorithm; BN folding; Convolutional neural network; Real time reasoning; embedded system

引言

（一）研究背景

计算机视觉推理任务正在经历从云端向边缘侧迁移的明显趋势，这一转变的动力来自时延、带宽和隐私等多重现实约束。Gartner 等机构的预测数据显示，到 2025 年，企业产生的数据将有超过 75% 在云数据中心之外完成处理 [1]。边缘推理将计算资源置于数据采集点附近，有助于将端到端延迟控制在毫秒级，对自动驾驶、无人机巡检和工业流水线检测等对时延极为敏感的场景具有实际意义。

然而边缘侧的硬件条件相当有限。NVIDIA Jetson Nano 和 Rockchip RK3588 等工业级嵌入式平台的算力普遍在 0.5 ~ 6 TOPS 范围内，内存带宽与功耗预算均远低于数据中心 GPU。主流深度学习模型的参数量动

ISSN: 3106-1540（ Print ）
3106-1559（ Online ）
Modern Development Research 现代发展研究
Vol.2, No.3, March 2026

辄以亿计，且通常以32位浮点（FP32）格式存储与运算。在上述平台直接运行FP32模型，一方面会因计算量超出硬件承受范围而造成明显的推理延时，另一方面频繁的片外内存访问也会带来不可忽视的功耗代价，直接影响电池驱动设备的续航[2]。

（二）核心挑战与研究空白

模型量化是应对边缘算力与资源错配的主流技术路径，但在工程落地层面面临三项结构性挑战。

首先是低比特精度退化问题。位宽降至4 bit或以下时，传统PTQ方法难以有效处理权重分布中的离群值，量化误差扩大，模型精度往往出现超过10%的断崖式下降。

其次是训练与推理结构不匹配的问题。量化感知训练中，Batch Normalization层在训练阶段被单独保留，而在推理阶段通常会被折叠进卷积层以节省计算开销。这种前后结构上的不一致，会使训练期间学习到的量化参数在实际部署时偏离最优，进而引发精度回退现象。

最后是步长初始化敏感问题。以LSQ为代表的可学习量化方法对量化步长的初始值较为敏感，设置不当容易引发梯度消失或梯度爆炸，使训练在波动中难以收敛，往往需要反复手动调整参数才能勉强稳定。

（三）本文主要贡献

针对以上问题，本文以聂慧等人提出的LSQ-BN算法[3]为核心量化方案，构建了一套端到端的边缘模型部署系统NeuroEdge-Quant，主要工作包括以下几个方面：

（1）系统架构设计：面向软硬协同部署需求，设计了包含模型解析、LSQ-BN量化训练、图优化和异构硬件编译四个环节的完整工作流，支持从PyTorch/ONNX模型到边缘可执行文件的自动转换。

（2）算法工程化实现：详细说明了LSQ-BN算法中单CNN BN折叠与自适应步长初始化的具体实现方式，并在低功耗嵌入式平台上验证了其可行性。

（3）多场景应用验证：选取无人机输电线路巡检和工业质检两个实际场景，评估了系统在W4A8（4-bit权重、8-bit激活）配置下的推理速度和精度表现，证明了该方案在能效约束条件下的工程适用性。

一、相关工作与文献综述

（一）神经网络量化技术

神经网络量化方法通常划分为后训练量化（PTQ）和量化感知训练（QAT）两大类。AdaRound、BRECQ等PTQ方案无需重新训练，仅依赖少量校准数据完成量化参数的确定，在8-bit量化条件下表现尚可，但当位宽降至4 bit时，精度下降往往难以接受。QAT则在训练过程中引入模拟量化操作，借助反向传播对权重进行微调，通常能取得更好的低比特精度。LSQ（Learned Step Size Quantization）[4]将量化步长s作为可微参数纳入优化，是该领域颇具代表性的工作。但LSQ在处理BN层时存在固有缺陷，且其收敛质量对初始化方案较为依赖。

（二）Batch Normalization折叠方案

BN折叠（Batch Normalization Folding）是推理阶段常规的图优化操作[5]。在传统QAT框架中，Google提出了双CNN结构以近似模拟推理时的BN融合行为，即在训练图中保留两条卷积分支，以维持与折叠后结果的数学等价性。这一方法虽在理论层面解决了等价性问题，但代价是计算图复杂度和训练显存开销的明显上升，且在部分边缘推理引擎中存在映射困难的问题。本文采用的LSQ-BN算法改为基于单CNN的折叠构造方案，在训练前向传播中直接完成参数融合，从源头消除了训练与推理阶段的结构性差异。

（三）边缘推理优化研究现状

面向边缘部署的模型压缩研究已积累了剪枝、知识蒸馏、量化等多种技术路径[2]。在硬件适配方面，NVIDIA TensorRT和Rockchip RKNN-Toolkit2等厂商工具链为量化模型的编译部署提供了工程基础，但两者的量化参数输入接口不同，给跨平台统一量化方案的实现带来挑战。本文在工具链层面设计了统一的参数导出接口，实现了同一量化训练结果在两类硬件平台上的无缝部署。

二、LSQ-BN算法核心机制与数学原理

本节对NeuroEdge-Quant所依托的核心算法机制进行系统梳理，重点阐明其在低比特量化条件下保持精度的理论依据。

（一）量化函数与直通估计器

将全精度数据v（权重或激活）量化为b-bit整数，并反量化为$\hat{v}$的过程定义为：

$$\bar{v}=\text{roundclip}(v/s,QN,QP),\hat{v}=\bar{v}\times s$$

其中，s是可学习的量化步长，QN、QP分别为量化区间的下界和上界（对于无符号4-bit，QN=0，QP=15）。

由于round函数在数值上不连续，无法直接对s进行梯度传播，故引入直通估计器（STE）。LSQ-BN在此基础上做了改进：其梯度表达式显式保留了量化误差项，使s的更新方向更贴近使量化噪声最小的最优解：

$\partial\hat{v}/\partial s=-v/s+\text{round}(v/s)$，当$QN<v/s<QP$时；

$\partial\hat{v}/\partial s=QN$，当$v/s\leq QN$时；

$\partial\hat{v}/\partial s=QP$，当$v/s\geq QP$时。

（二）单 CNN 构造的 BN 折叠机制

单 CNN 构造的 BN 折叠机制是系统实现高效推理的关键。在推理阶段，BN 层的线性变换参数通常被吸收到卷积层的权重 w 和偏置 b 中：

$$w\_fused=\gamma/\sqrt{(\sigma^2+\varepsilon)}\cdot w,\ b\_fused=\beta+\gamma\cdot(b-\mu)/\sqrt{(\sigma^2+\varepsilon)}$$

其中 μ、σ 是 BN 的均值和方差，γ、β 是仿射参数。

传统 QAT 保留独立的 BN 层参与训练，使得损失函数在反向传播时感知不到融合后权重的量化误差，从而形成训练与部署之间的系统偏差。LSQ-BN 的做法是直接在前向传播中执行融合操作，具体步骤如下。

（1）统计计算：计算当前 Batch 的 μ_B、σ_B。

（2）在线融合：根据上述公式动态生成 w_fused。

（3）伪量化：对 w_fused 执行量化操作，得到 $\hat{w}$_fused。

（4）卷积运算：使用 $\hat{w}$_fused 进行卷积。

以上步骤可以使得网络在训练阶段直接接触到的是融合并量化后的等效权重，而非原始的浮点权重加独立 BN 参数的组合。这意味着训练时的损失梯度反映的是与实际推理结构完全一致的量化误差，权重的更新方向因此更贴近部署条件。传统方法中由于训练拓扑与推理拓扑不一致而产生的量化参数偏移，在这种构造下得到了规避，有助于缩小训练精度与部署精度之间的差距。

（三）自适应量化因子初始化

LSQ-BN 针对初始化敏感性问题，借助预训练权重的分布特征进行自适应设定。深度网络的权重分布通常近似高斯分布，但常伴有长尾离群值：若以最大绝对值作为初始步长，多数权重会被压缩至接近零的区间；若步长设定过小，则大量权重将被直接截断丢失。

为此，本系统采用双端截断策略，即对各层权重绝对值排序后，剔除头尾各 α%（通常取 α=2.5）的离群值，再取截断后序列的最大值 V_trunc_max，按公式 s_init=2·V_trunc_max/(QP-QN) 计算初始步长。实验表明，该策略可使模型在首个 Epoch 内迅速进入稳定收敛区间，有效规避训练发散。

三、NeuroEdge-Quant 系统架构设计

在上述算法基础上，本文实现了边缘 AI 部署平台。平台的设计目标是消除算法模型与边缘硬件之间的适配壁垒，使模型轻量化流程尽可能自动化、可复用。

（一）总体架构

系统由三层组成：模型优化层（Model Optimization Layer）、边缘编译层（Edge Compilation Layer）和端侧运行层（Runtime Execution Layer）。各核心模块功能如表 1 所示。

（二）关键模块设计

1.Auto-QAT Studio

Auto-QAT Studio 是基于 PyTorch 的服务端 Python 工具包，其核心组件包括三部分：算子融合引擎通过图匹配（Graph Matching）自动定位 Conv-BN-ReLU 拓扑结构，并将其整体替换为自定义的 LSQ_BN_Conv2d 算子；初始化向导在浮点模型加载完成后，自动执行双端截断算法，为各层生成初始步长 s；微调控制器负责驱动 QAT 流程，实时监测步长 s 的梯度变化，一旦检测到剧烈震荡，即触发学习率衰减以稳定训练进程。

2.PolyMorph Compiler

该模块将训练完成的“伪量化”模型转换为可直接在目标硬件上运行的二进制推理文件。编译流程首先从量化感知训练后的模型中提取融合后的整数权重 $\bar{w}$ 及学习所得步长 s，作为后续部署的核心参数。针对 NVIDIA Jetson 系列，系统通过 TensorRT API 构建含显式量化节点（Q/DQ）的 ONNX 计算图；对于支持 INT4 Tensor Core 的 Orin 架构，编译器会将两个 4-bit 权重打包至单字节存储，以降低显存占用并提升吞吐效率。针对 Rockchip NPU，则生成与 RKNN-Toolkit2 兼容的量化配置文件（quant_config），将 LSQ 习得的步长 s 直接写入其中，从而绕过 RKNN 默认的 PTQ 校准环节。

3. 实时视觉感知应用逻辑

在应用层，系统构建了通用的视觉处理管道以支撑边缘端实时检测需求。图像数据经 V4L2 接口从机载摄像头采集，保证低延迟的稳定输入；图像缩放与归一化等预处理步骤由边缘设备上的 ISP 或 GPU（如 CUDA 核）承担，避免 CPU 成为预处理瓶颈。预处理完成后，数据送入 Lite-Runner 进行 4-bit 与 8-bit 混合精度推理；推理结束后，非极大值抑制（NMS）等后处理操作在 CPU 上异步执行，与前端推理流水线并行，有效提升整体处理帧率与资源利用率。

四、应用验证与实验结果分析

以下以无人机输电线路绝缘子缺陷检测为典型验

表 1　NeuroEdge-Quant 系统核心模块说明

模块	功能描述	关键技术支撑
Auto-QAT Studio	模型摄入、算子重构、量化训练	单 CNN BN 折叠、自适应初始化
PolyMorph Compiler	异构硬件指令映射、位宽优化	INT4/INT8 混合精度映射
Lite-Runner	边缘端推理引擎、内存管理	零拷贝机制、流水线并行

ISSN: 3106-1540（ Print ）
3106-1559（ Online ）
Modern Development Research 现代发展研究
Vol.2, No.3, March 2026

证场景，并在标准数据集 ImageNet 上对 LSQ-BN 算法进行复现，结合真实边缘硬件推理性能测试，对 NeuroEdge-Quant 系统进行综合评估。

（一）无人机巡检场景应用验证

1. 场景需求分析

该系统的约束条件主要体现在三方面：实时性方面，无人机巡检飞行速度为 5 ~ 10 m/s，图像采集到缺陷识别的端到端延迟须控制在 20 ms 以内；功耗方面，机载计算平台（Jetson Orin Nano）依赖电池供电，整个检测模块的运行功耗不得超过 10 W；精度方面，Top-1 准确率需超过 74%，以满足实际巡检中准确定位各类缺陷的需求。

2. 系统配置与优化

主干网络选用 ResNet-50，配合 YOLO 检测头进行缺陷识别，以在精度与计算效率之间取得平衡。权重采用 4-bit LSQ-BN 量化，将 ResNet-50 的权重体积从 98 MB 压缩至约 12.5 MB，可完整装入片上缓存，大幅减少 DRAM 访问所带来的功耗开销。激活值保留 8-bit 精度，以在动态变化较大时维持足够的数值表示范围。部署流程为：先用 Auto-QAT Studio 对预训练 YOLO 模型进行约 20 个 Epoch 的量化感知训练，再经 PolyMorph Compiler 导出为 TensorRT 推理引擎。

（二）实验环境与设置

数据集：ImageNet（ILSVRC2012）；模型：MobileNet-v2（轻量级代表）、ResNet-50（高性能代表）；硬件平台：训练使用 NVIDIA RTX 3090，推理使用 NVIDIA Jetson Orin Nano（8GB）及 Rockchip RK3588；对比基准：FP32 Baseline、Google PTQ、Standard QAT、Original LSQ。

（三）精度对比分析

表 2 展示了不同量化配置下的 Top-1 准确率对比。

从表 2 数据可得出两点主要结论：其一，4-bit 鲁棒性方面，在 W4A8 这一极端配置下，基于 LSQ-BN 的 NeuroEdge 系统表现稳健，ResNet-50 的精度损失仅为 1.5%；相比之下，标准 QAT 出现了明显的模型崩溃（精度下降 10.7%），而本系统的自适应初始化策略有效规避了训练初期的梯度失效问题。其二，轻量级网络适配方面，MobileNet-v2 结构紧凑，本身量化难度较大，系统仍将精度损失压制在 2.7%，优于 Google PTQ（5.3%），体现了可学习步长在捕捉小容量网络特征上的优势。

（四）训练收敛性与效率

在收敛行为方面，采用自适应初始化的 LSQ-BN 算法对 MobileNet-v2 进行 4-bit 训练时，首个 Epoch 的训练 Loss 即由 2.22 降至 1.44，此后持续平稳下降。而原始 LSQ 算法使用随机初始化时，Loss 长期在 6.9 附近震荡，难以下降。LSQ-BN 的 QAT 流程通常 20 个 Epoch 左右即可收敛，传统 QAT 则往往需要 100 至 200 个 Epoch。这意味着本系统可显著压缩量化模型的训练周期，对工业场景中的快速迭代更新具有实际意义。

（五）推理延迟与功耗

在 Jetson Orin Nano 上对 ResNet-50 进行实测：延迟（Latency）方面，FP32 模型约为 30 ms/ 帧，W4A8 模型在 INT4 Tensor Core 加速下降至 8 ms/ 帧，吞吐量提升约 3.75 倍；功耗方面，单次推理能耗从 15 mJ 降至 6 mJ，节能效果较为显著，主要得益于内存访问量的大幅减少。

五、结论与研究不足

（一）主要结论

本文设计并实现了面向边缘计算的视觉感知系统 NeuroEdge-Quant。通过深度整合 LSQ-BN 算法的单 CNN BN 折叠与自适应量化因子初始化技术，系统在工程层面有效解决了边缘 AI 部署中低比特精度下降与训练难以收敛两个长期存在的问题。

单 CNNBN 折叠的核心优势在边缘推理引擎在

表 2　ResNet-50 与 MobileNet-v2 在不同量化策略下的性能对比

模型架构	量化方法	精度配置 (W/A)	Top-1 准确率	精度损失 (vs FP32)	压缩倍率
ResNet-50	FP32(Baseline)	32/32	75.6%	–	1 ×
	NeuroEdge(LSQ-BN)	8/8	75.2%	0.4%	4 ×
	NeuroEdge(LSQ-BN)	4/8	74.1%	1.5%	8 ×
	Standard PTQ	4/8	67.9%	7.7%	8 ×
	Standard QAT	4/8	64.9%	10.7%	8 ×
	Original LSQ	4/8	73.8%	1.8%	8 ×
MobileNet-v2	FP32(Baseline)	32/32	72.5%	–	1 ×
	NeuroEdge(LSQ-BN)	8/8	72.3%	0.2%	4 ×
	NeuroEdge(LSQ-BN)	4/8	69.8%	2.7%	8 ×
	Standard PTQ	4/8	67.2%	5.3%	8 ×

执行优化时通常会对 Conv-BN- 激活函数进行算子融合，以减少内核调用开销。若训练阶段沿用 Google 的双路分支结构，训练拓扑与实际推理拓扑并不等价，量化参数（尤其是尺度因子 Scale）在部署时便可能失去最优性。LSQ-BN 的单 CNN 折叠方案在训练阶段即模拟了算子融合的实际行为，使训练中所见的等效权重与推理中真正使用的整数权重保持一致，这是 NeuroEdge-Quant 能够将边缘端精度损失控制在 1% 以内的核心原因。

实验数据表明，系统在 W4A8 配置下将模型体积压缩至原始的 1/8，ResNet50 的 Top1 准确率可以保持在 74.1%，与 FP32 基线的差距仅为 1.5%。在无人机巡检和工业质检等实际应用场景中，该方案可以在明显降低硬件成本与功耗的同时，保证其实时处理能力。因此，本文既验证了 LSQ-BN 算法的实用价值，也为工业界提供了一套可落地、可复用的边缘 AI 部署方案。

（二）研究不足与未来工作

本系统在标准 CNN 架构上取得了较好的结果，但仍存在以下局限性：（1）Transformer 类模型（如 ViT、Swin Transformer）的适配仍存在局限。这类架构中 LayerNorm 和 Softmax 的数值分布与卷积层差异较大，直接套用线性量化往往难以奏效。（2）当前系统的训练收敛曲线分析仅基于 Loss 指标，缺乏对量化步长动态变化的可视化分析。（3）跨平台部署的自动化程度有待进一步提升。后续工作拟将 LSQ-BN 算法扩展至非线性量化与混合精度搜索（Mixed Precision Search），以满足边缘侧大语言模型（Edge LLM）日益增长的部署需求。

参考文献：

[1]Gartner.What Edge Computing Means for Infrastructure and Operations Leaders[R].Gartner Research,2018.

[2]Han S,Mao H,Dally W J.Deep compression:Compressing deep neural networks with pruning,trained quantization and huffman coding[C]//ICLR,2016.

[3] 聂慧 , 李康顺 , 苏洋 . 一种量化因子自适应学习量化训练算法 [J]. 系统仿真学报 ,2022,34(07):1639-1650.

[4]Esser S K,et al.Learned Step Size Quantization[C]//ICLR,2019.

[5]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(CVPR).Salt Lake City:IEEE,2018:2704-2713.