面向边缘计算环境的深度学习模型压缩方法研究

赵家鹏 [1]
ZHAO Jiapeng

摘 要 针对边缘计算环境下深度学习模型参数规模大、存储开销高及推理效率受限等问题，文章通过构建模型冗余表征与压缩收益评估机制，提出了一种融合结构化动态剪枝、混合精度量化、知识蒸馏增强及多目标协同优化的协同压缩方法，并基于 ImageNet-1K 和工业视觉数据集开展验证。结果表明，模型参数量由 25.6×10^6 降至 2.7×10^6，压缩率达到 89.3%，推理时延由 63.4 ms 降至 26.2 ms，单位推理能耗降低 52.4%，Top-1 精度仅下降 1.1%。研究表明，该方法能够有效降低边缘部署资源消耗，提高模型推理效率与运行稳定性。

关键词 边缘计算；深度学习；模型压缩；混合精度量化；知识蒸馏

doi：10.3969/j.issn.1672-9528.2026.07.010

0 引言

随着人工智能技术的快速发展，深度学习模型逐步由云端向终端迁移，并广泛应用于工业视觉、智能医疗、智慧安防等场景 [1]。边缘计算的实际应用受计算资源、存储空间、能耗水平及部署环境等因素影响，对模型轻量化和实时响应能力提出了更高要求 [2]。卷积神经网络和 Transformer 等模型虽具备较强特征表征能力，但参数规模和计算复杂度持续增长。以 ResNet50、YOLOv8-S 为代表的模型参数量已达千万级，对存储带宽和计算资源提出较高要求。受限于边缘设备算力、内存及能耗条件，大规模模型直接部署往往面临推理时延高、资源占用大等问题，难以满足实时智能应用需求 [3]。

模型压缩通过削减冗余参数、降低存储开销和计算负载，成为提升边缘部署效率的重要技术路径。现有研究主要集中于网络剪枝、低比特量化、知识蒸馏及低秩分解等方法，但仍存在压缩策略协同性不足和部署适应性有限等问题。为此，本文构建模型冗余表征与压缩收益评估机制，提出融合结构化动态剪枝、混合精度量化、知识蒸馏增强及多目标协同优化的模型压缩方法，并基于 ImageNet-1K 和工业视觉数据集开展验证，分析压缩模型的边缘部署性能与应用效果。

1 模型压缩机制构建

1.1 边缘资源约束分析

边缘计算节点普遍采用 ARM Cortex-A72、RK3588、Jetson Nano 及 Ascend Atlas 等低功耗异构平台，算力 TOPS 通常处于 0.5~10 区间，内存容量 2~8 GB，持续功耗受限于 5~15 W。以 YOLOv8-S、ResNet50 等典型网络为例，参数规模分别约 11.2×10^6 和 25.6×10^6，FP32 存储需求达到 44.8 MB 和 102.4 MB，单次推理计算量 GFLOPs 超过 28.6 和 4.1。不同边缘平台资源配置及推理性能如表 1 所示。

表 1 边缘计算平台资源参数对比

平台型号	算力 TOPS	内存容量 /GB	功耗 /W	ResNet50 推理时延 /ms
ARM Cortex-A72	0.8	2	5	186.5
Jetson Nano	0.9	4	10	98.4
RK3588	6	8	12	63.4
Ascend Atlas 200I DK	8	8	15	41.7

边缘设备算力与存储资源受限，即使采用 RK3588 和 Ascend Atlas 等高性能平台，复杂卷积网络仍存在推理时延较高的问题 [4]。在 640×640 输入条件下，模型访存开销占总延迟 45% 以上，缓存失效率达到 18%~27%。Roofline 分析表明，边缘处理器计算密度普遍低于 8 FLOPs/Byte，性能瓶颈主要来源于内存带宽限制。实测中，ResNet50 在 RK3588 平台上的平均推理时延为 63.4 ms，峰值内存占用 1.21 GB，单位推理能耗 0.68 J，难以满足 20 ms 级实时响应需求，因此有必要开展模型压缩优化以降低计算与存储开销。

1.2 模型冗余特征表征

深度神经网络存在显著参数冗余现象，研究表明卷积神经网络中 30%~70% 的权重对最终预测结果贡献度较低 [5]。

1. 昆明卫生职业学院 云南昆明 650101

为定量评价网络结构重要性，采用 L_1 范数、Taylor 展开敏感度、Hessian 谱特征及 BatchNorm 缩放因子值构建通道重要度评价模型：

$$I_c = \frac{1}{N}\sum_{i=1}^{N} |W_i| \cdot S_a \cdot G_f \tag{1}$$

式中：I_c 为第 c 个输出通道的重要度指标，W_i 为通道内第 i 个权重参数，S_a 为特征图激活稀疏率，G_f 为梯度贡献系数，N 为通道参数数量。I_c 数值越小，表明该通道对模型输出贡献越低，可优先纳入压缩候选集合。

基于式（1）计算通道重要度后，对网络冗余特征进行量化分析。结果表明，ResNet50 第 3 阶段约 42.7% 的输出通道平均激活率低于 0.15；MobileNetV2 深度可分离卷积层参数贡献度标准差达到 0.38，呈现明显冗余分布。SVD 分析发现，部分卷积层有效秩仅占原始维度的 35%~50%，具有较大的低秩压缩空间。综合通道重要度、激活稀疏率、梯度敏感度及矩阵秩特征，构建参数重要性评价矩阵，为后续剪枝、量化及知识蒸馏提供依据。

2 协同压缩方法设计

2.1 结构化动态剪枝

基于 1.2 节构建的参数重要性评价矩阵，采用通道级结构化动态剪枝策略对卷积层输出特征进行稀疏化处理。剪枝对象覆盖卷积核、输出通道及残差分支，利用 BatchNorm 缩放因子、通道重要度指标及 Taylor 敏感度联合计算剪枝评分：

$$P_c = \lambda_1 I_c + \lambda_2 \gamma_c + \lambda_3 T_c \tag{2}$$

式中：P_c 为第 c 个输出通道的剪枝评分，I_c 为通道重要度指标，γ_c 为 BatchNorm 层缩放因子，T_c 为 Taylor 敏感度系数，λ_1、λ_2、λ_3 为权重参数，且满足 $\lambda_1+\lambda_2+\lambda_3=1$。$P_c$ 值越小，表示对应通道对模型输出贡献越低，优先纳入剪枝候选集合。

基于通道重要度、推理时延、模型压缩率及精度变化等指标，对剪枝方案进行综合评估。实验表明，当模型压缩率控制在 75% 以内时，Top-1 精度下降通常不超过 1.5%；压缩率超过 90% 后，精度衰减明显加剧。因此，剪枝过程中需兼顾模型轻量化与特征表达能力保持，为后续混合精度量化和知识蒸馏策略提供高质量压缩基础。

2.2 混合精度量化

针对不同网络层量化敏感度差异，构建基于 Hessian 谱特征的混合精度量化框架 [6]。首先统计各层权重分布及激活值动态范围，采用 KL 散度最小化方法确定量化阈值；其次依据层敏感度分配不同位宽，高敏感层采用 INT8 量化，中敏感层采用 INT6 量化，低敏感层采用 INT4 量化。不同网络层混合精度配置方案如表 2 所示。

表 2 不同网络层混合精度量化配置

网络层类型	位宽配置 /bit	量化方式	量化敏感度
输入层	8	INT8	高
浅层卷积层	8	INT8	高
中间卷积层	6	INT6	中
深层卷积层	4	INT4	低
全连接层	8	INT8	高

量化过程引入量化感知训练机制，卷积层权重量化尺度因子按通道独立更新，激活值采用动态量化处理，残差连接及输出层保留较高位宽以降低误差累积。实验表明，模型存储需求由 102.4 MB 降至 11.8 MB，压缩率达到 88.5%；RK3588 平台推理时延由 63.4 ms 降至 28.7 ms，缓存命中率提升 16.3%，内存访问次数降低 42.8%。量化后 Top-1 精度仅下降 0.8%，实现了模型精度与部署效率的良好平衡。

2.3 知识蒸馏增强

采用教师—学生双网络蒸馏架构增强压缩模型特征学习能力。教师网络选用 ResNet101，参数规模 44.5×10^6；学生网络采用经过剪枝与量化后的轻量化 ResNet50。蒸馏过程同时约束输出层软标签、特征映射层及注意力响应矩阵，构建分类损失、蒸馏损失和特征对齐损失联合优化目标。温度系数 T 设定为 4，蒸馏损失权重 λ 取 0.7。针对中间层特征，引入注意力迁移机制对通道响应分布进行约束，使学生网络保持教师网络的判别特征表达能力 [7]。实验结果显示，在参数规模减小 73.4% 的条件下，Top-1 精度由 76.2% 恢复至 77.5%，较未蒸馏模型提升 2.1 个百分点，有效缓解压缩过程中产生的性能退化问题。

2.4 多目标协同优化

为协调模型压缩率、推理时延、能耗水平与精度保持之间的约束关系，构建基于 NSGA-II 的多目标协同优化框架，以结构化剪枝率、量化位宽配置及蒸馏权重为决策变量，通过非支配排序与拥挤度距离机制搜索 Pareto 最优解集 [8]。种群规模设为 80，迭代 100 代，交叉概率 0.9，变异概率 0.1，并结合边缘设备实测数据动态更新适应度。结果表明，当剪枝率为 72%、INT4/INT6/INT8 配置比例为 35%/40%/25%、蒸馏权重为 0.7 时，模型压缩率达到 89.3%，推理时延降低 58.6%，单位推理能耗降低 52.4%，Top-1 精度仅下降 1.1%，实现了资源占用、运行效率与模型精度的协同优化。

3 实验验证与性能评价

3.1 数据采集方案

实验采用 ImageNet-1K 公开数据集构建训练与测试样本

库，包含 1 000 个类别、128.1 万张训练图像及 5 万张验证图像。为验证边缘场景适应性，补充采集工业视觉目标识别数据集，共计 3.6 万张图像，覆盖设备表计识别、零部件检测及缺陷分类等任务。训练集、验证集及测试集按照 8:1:1 划分，输入分辨率统一调整为 224 px×224 px。数据增强采用随机裁剪、颜色扰动、Mixup 及 Mosaic 策略。模型训练阶段同步采集参数量、FLOPs、显存占用及训练损失；部署阶段实时记录推理时延、CPU 占用率、NPU 利用率、内存占用、缓存命中率及单位推理能耗 [9]。所有性能指标连续采样 1 000 次取均值，以降低瞬时负载波动对实验结果的影响。

3.2 实验平台配置

模型训练平台采用 Intel Xeon Gold 6338 处理器、512 GB DDR4 内存及 NVIDIA RTX 4090 GPU，显存容量 24 GB；软件环境为 Ubuntu 22.04、Python 3.10、PyTorch 2.1 及 CUDA 12.1。边缘部署平台选用 RK3588 开发板，配置 4×Cortex-A76 与 4×Cortex-A55 异构 CPU，集成 6 TOPS NPU，内存容量 8 GB。模型推理采用 ONNX Runtime 与 RKNN Toolkit 联合部署，量化模式设置为 INT4/INT6/INT8 混合精度。训练批次大小设为 256，优化器采用 SGD，初始学习率 0.1，动量因子 0.9，权重衰减系数 1×10^{-4}，训练轮数 200 Epoch。推理测试阶段固定输入尺寸 224×224，连续运行 24 h 评估模型稳定性与资源利用特征。

3.3 压缩效果分析

基于 ResNet50 构建原始模型、动态剪枝模型、混合量化模型及协同压缩模型开展对比实验。重点考察参数规模、模型体积、计算量、识别精度及推理时延等关键指标，不同压缩方案性能对比结果如表 3 所示。

表 3 不同压缩方案性能比较

模型	参数量 /10^6	模型体积 /MB	FLOPs /10^9	Top-1 精度 /%	推理时延 /ms
原始 ResNet50	25.6	102.4	4.1	78.6	63.4
动态剪枝模型	6.8	27.2	1.2	77.4	41.6
剪枝 + 量化模型	6.8	11.8	1.2	76.8	28.7
协同压缩模型	2.7	10.4	0.9	77.5	26.2

动态剪枝将模型参数量由 25.6×10^6 降至 6.8×10^6，FLOPs 降低 70.7%；引入混合精度量化后，模型体积进一步压缩至 11.8 MB，存储开销降低 88.5%。协同压缩模型参数量仅为 2.7×10^6，总压缩率达到 89.3%，模型体积缩减至 10.4 MB，Top-1 精度仍保持 77.5%，较原始模型仅下降 1.1 个百分点。与此同时，推理时延由 63.4 ms 降至 26.2 ms，降幅达 58.7%，表明该协同压缩方法能够在保持模型精度的同时显著降低计算与存储开销，具备良好的边缘部署适用性。

3.4 边缘部署评价

为验证模型在实际边缘环境中的部署性能，在 RK3588 平台开展连续运行测试。重点考察推理时延、推理帧率、内存占用、能耗水平及硬件资源利用率等指标，边缘部署性能对比结果如表 4 所示。

表 4 边缘部署性能对比数据

模型	推理时延 /ms	推理帧率 /(帧 ·s^{-1})	峰值内存占用 /GB	单位推理能耗 /J	CPU 平均负载 /%	NPU 利用率 /%	目标识别准确率 /%
原始模型	63.4	15.8	1.21	0.68	78.4	53.2	95.9
协同压缩模型	26.2	38.2	0.34	0.32	41.7	82.5	96.7
变化幅度	-58.7%	141.8%	-71.9%	-52.9%	-46.8%	55.1%	0.8%

注：原始模型为未压缩 ResNet50；协同压缩模型为经过结构化动态剪枝、混合精度量化、知识蒸馏增强及多目标协同优化后的部署模型。

协同压缩模型推理时延由 63.4 ms 降至 26.2 ms，降幅达 58.7%，推理帧率由 15.8 帧 /s 提升至 38.2 帧 /s。峰值内存占用由 1.21 GB 降至 0.34 GB，单位推理能耗由 0.68 J 降至 0.32 J，分别降低 71.9% 和 52.9%；CPU 平均负载下降 46.8%，NPU 利用率提升至 82.5%。工业视觉测试集识别准确率达到 96.7%，较原始模型提高 0.8 个百分点。连续运行 24 h 期间，系统功耗稳定在 8.7 W 以内，设备温度低于 62 ℃，未出现性能衰减或运行异常，表明所提出方法具有良好的边缘部署稳定性与工程应用价值。

4 结论

面向边缘计算环境资源受限特点，本文构建了融合冗余特征分析、动态剪枝、混合精度量化、知识蒸馏及多目标协同优化的模型压缩方法。实验结果表明，所提出压缩框架在保持模型识别精度的同时显著降低了参数规模、存储开销和计算负载，实现 89.3% 的模型压缩率、58.6% 的推理时延降幅及 52.4% 的能耗降幅。边缘部署测试显示，模型推理帧率提升至 38.2 帧 /s ，内存占用和 CPU 负载明显降低，连续运行稳定性良好。研究结果验证了协同压缩策略在边缘智能场景中的工程适用性，为深度学习模型高效部署与边缘计算资源优化利用提供了参考。

参考文献：

[1] 付晓丽 . 一种支持边缘计算的轻量化深度学习推理系统设计 [J]. 信息记录材料 ,2025,26(8):62-64.

基于 AI 语义分析的客户服务对话系统设计

兰一杰 [1]
LAN Yijie

摘 要 面向多轮任务型语义交互场景，文章构建了一套基于人工智能语义分析的客户服务对话系统。研究对象涵盖自然语言输入、上下文状态与企业级多源知识内容。系统由语义理解、对话状态管理、知识融合与响应决策四个模块组成。语义理解模块采用双通道分类结构，结合双向编码器表示（bidirectional encoder representations from transformers, BERT）完成意图识别与槽位抽取；状态管理模块引入门控循环单元（gated recurrent unit, GRU）实现状态表示更新；知识融合模块联合密向量检索与图神经网络（graph neural network, GNN）完成语义匹配与知识推理；响应决策模块通过动作分类与语言生成系统响应。系统在 SMP2020-ECDT、MultiWOZ 2.2 及自建语料上进行评估，采用准确率、F_1 值、BLEU 值及接口性能作为指标。结果表明各模块在识别精度、知识召回与响应质量方面性能良好，具备工程部署可行性。

关键词 人工智能；语义理解；对话系统；多轮交互；对话状态建模；知识融合

doi：10.3969/j.issn.1672-9528.2026.07.011

0 引言

面向任务的客户服务对话系统是自然语言处理与人机交互融合的重要方向，广泛应用于智能客服、信息查询等场景。随着深度学习与预训练语言模型的发展，对话系统由规则驱动逐步转向语义建模为核心的多模块交互架构，对语义理解精度、上下文保持与知识调用能力提出更高要求。近年来，多轮状态建模、知识调度与响应生成成为研究热点，系统性能与工程可部署性同步提升。

李啸天等 [1] 基于本体构建与状态结构建模，提出融合知识蒸馏的对话状态跟踪方法，提升状态预测稳定性与泛化能力。杜维等 [2] 设计模态敏感注意力机制的多模态对话模型，增强了异构信息融合下的语义一致性。蔡辰等 [3] 聚焦适老化交互，构建基于意图迁移链的语义调节模型，提升特殊人群使用体验。熊文剑等 [4] 结合云平台应用，探讨预训练模型的参数优化策略，验证轻量网络在任务型系统中的部署效率。

在此基础上，本文归纳当前对话系统在多轮状态建模、知识融合调度与响应生成一致性方面的瓶颈，构建一套融合语义理解、状态建模、知识推理与响应控制的客户服务对话

1. 南宁市勘测设计院集团有限公司 广西南宁 530000

[2] 庞顺顺 . 边缘计算场景下系统集成的轻量化优化算法设计 [J]. 产品可靠性报告 , 2025(7): 213-214.
[3] 郝舒羽 . 基于边缘计算的智能物联网环境监测系统研究 [J]. 油气田环境保护 , 2026, 36(1): 50-53.
[4] 裴志刚 , 钟天成 , 范江鹏 , 等 . 基于轻量化 AI 模型配网缺陷实时识别与边缘计算部署 [J]. 机械制造与自动化 , 2026, 55(1): 1-6.
[5] 陈智伟 . 面向边缘计算的轻量化人工智能模型研究 [J]. 互联网周刊 , 2025(22): 26-28.
[6] 董浩 , 李劭辉 , 阚诺文 , 等 . 基于深度压缩感知的联合信源信道编码方法研究 [J]. 电子学报 , 2025, 53 (7): 2178-2192.
[7] 廖四龙 , 耿卓强 . 基于边缘计算的多模态数据处理与智能决策系统研究 [J]. 电脑编程技巧与维护 , 2026(5): 85-87.
[8] 杨雄 , 卓禀航 , 吴志诚 , 等 . 边缘计算中轻量化人工智能算法优化方法研究 [J]. 贵州大学学报 (自然科学版), 2026, 43(3): 20-25.
[9] 董甲东 , 桑飞虎 , 郭庆虎 , 等 . 基于深度学习的目标检测算法轻量化研究综述 [J]. 计算机科学与探索 , 2025, 19(8): 2057-2084.

【作者简介】
赵家鹏（1984—），男，云南昆明人，本科，讲师，研究方向：计算机教学。

（收稿日期：2026-06-18 修回日期：2026-07-16）