 DOI: https://doi.org/10.12349/mit.v3i2.11058

# Lightweight AI Empowers Real-Time Multi-Source Data Analysis for Industrial Internet Edge Terminals

**Qi Chen Han Li Lin Xie Yuan Huang Nian Li**

Mianyang Vocational and Technical College, Mianyang, Sichuan, 621000, China

**Abstract**

Industrial internet edge terminals face computational constraints in real-time processing of multi-source heterogeneous data. This paper proposes a real-time multi-source data analysis method based on lightweight AI models, integrating model compression, knowledge distillation, and hardware-aware quantization techniques to achieve efficient data fusion and parsing on resource-constrained devices. The approach reduces end-to-end latency and enhances applicability in industrial scenarios, providing theoretical and technical support for intelligent edge computing in the industrial internet.

**Keywords**

lightweight AI; industrial internet; edge computing; multi-source data analysis; real-time processing

# 轻量化 AI 赋能工业互联网边缘终端多源数据实时解析研究

陈琪 李晗 谢林 黄元 李念

绵阳职业技术学院，中国 · 四川 绵阳 621000

**摘 要**

工业互联网边缘终端要面对的是多源异构数据实时解析的算力约束问题。本文提出一种基于轻量级AI模型的边缘终端多源数据实时解析方法，综合使用模型压缩、知识蒸馏和硬件感知量化技术，在资源受限的设备上完成高效的多源数据融合和解析，减小端到端延迟，提高工业场景的适用性，为工业互联网边缘智能提供理论和技术支持。

**关键词**

轻量化AI；工业互联网；边缘计算；多源数据解析；实时处理

## 1 绪论

工业互联网迅速发展促使边缘终端向着智能化的方向不断前进。伴随着传感器密度不断上升，现场总线 (Modbus/OPC-UA)、工业视觉和振动传感等各种各样的数据在边缘侧以毫秒级的速度涌现出来，传统的把全部的数据都传送到云端之后再进行处理的方式，已经不能满足工业控制闭环对于实时性的高要求了。按照 IEC/IEEE 时间敏感网络标准，闭环控制的典型容忍时延为 10ms 左右，而云端往返时延一般在几百毫秒以上，相差很大 [1]。边缘终端设备内存一般只有百兆字节级别，浮点运算能力也很弱，通用的深度学习模型参数量常常超过百万，不能在这些平台上进行高效的推理。轻量化 AI 技术用结构化剪枝、量化感知训练、知识蒸馏等方法，在精度可以接受的情况下大大减小模型体积，给以上场景提供了一种可行的部署方式。本文以工业互联网边缘终端多源数据实时解析为研究对象，提出三类压缩策略融合的轻量化模型设计方法，并搭建起数据接入、轻量推理、应用输出三层流水线解析架构，在 ARM Cortex-A 级别边缘平台上对它进行测试。全文共分六章，第二章对多源数据解析的主要难题进行梳理，第三章提出轻量化模型的设计和优化策略，第四章给出实时解析框架的架构，第五章给出实验验证结果及分析，第六章总结研究贡献并展望未来研究方向。

## 2 边缘终端多源数据解析挑战

边缘终端多源数据解析存在三大主要问题。第一，多源异构性难题。工业现场有 Modbus RTU/TCP、OPC-UA、EtherNet/IP 等现场总线协议，并且还存在图像、声音、振动等非结构化感知数据流 [2]。不同的数据源时间戳精度从微秒到毫秒不等，采样率从 10Hz 到 10kHz 跨越了三个数量级，直接融合很容易产生时序错位和尺度不一致的误差。多协议解析层要维持各个协议的状态机，依靠准确的时钟同步（PTP 或者 IEEE 1588）来对齐多路的数据帧，工程复杂度很高。第二，资源约束问题。主流工业边缘网关搭载 ARM Cortex-A53 至 A72 处理器，DRAM 容量通常为 512 MB 至

【作者简介】陈琪（1997–），女，中国四川自贡人，硕士，助教，从事工业互联网、人工智能研究。

2 GB，无独立 GPU；部分 IoT 终端甚至仅有 Cortex-M 系列 MCU，SRAM 不超过数百 KB。ResNet-50 等标准模型参数量约为 25M（约 100MB FP32），在上述设备上推理一帧需要几十到几百毫秒，远远超过工业实时性的要求，直接部署是不可能的。第三，实时性约束问题。工业控制回路对于端到端时延非常敏感，过程控制一般要求在 50ms 到 200ms 之间，运动控制要严格到 1ms 到 10ms 之间。任何解析框架都必须把 AI 推理时延纳入控制周期预算中，并且要对突发的数据峰值不造成调度抖动，目前的通用推理框架并没有对这种实时工况进行深入的优化，需要专门的设计。

## 3 轻量化 AI 模型设计与优化

本文针对以上资源和实时性的挑战，提出三条互补的轻量化 AI 优化策略，并通过硬件感知进行部署来实现端到端的落地。结构化剪枝不同于非结构化剪枝，结构化剪枝会直接去掉整个卷积核通道或者注意力头，使得剪枝后的网络在没有稀疏加速库支持的 ARM 平台下也能得到真实的推理加速 [3]。本文用基于 Taylor 展开的重要性评分对各个层的通道进行评分，根据全局阈值剪除低重要性通道，再经过少轮次微调恢复精度。对 MobileNetV2 骨干施加 40% 通道剪枝后，FLOPs 降低约 55%，Top-1 精度损失不超过 1.2%。在量化感知训练中，把权重和激活值从 FP32 量化成 INT8 可以使得模型内存占用减少 4 倍，并且使用 ARM NEON SIMD 指令集可以获得 1.5 到 3 倍的推理速度。本文在训练阶段加入伪量化节点来模拟量化误差，使用端到端反向传播消除量化造成的精度损失，最后将得到的 INT8 模型导出为 TFLite 或者 MNN 格式，适配硬件推理引擎 [4]。在知识蒸馏方面，用大规模工业数据集上预训练的全量教师模型（ResNet-50 规模）作为监督信号，最小化学生模型输出 logits 和中间特征层响应的 KL 散度，把教师的暗知识迁移到轻量骨干，比单独训练可以提高 1 到 3 个百分点的解析精度。硬件感知部署上对 Cortex-A 系列的 NEON 向量单元和 L1/L2 缓存拓扑进行了算子融合（BN 折叠、ReLU 融合）、内存布局（NHWC 格式对齐）等方面的定制优化，同时还加入了 TFLite Delegate 和 MNN 后端，使同一个推理图可以在 CPU、NPU 等异构硬件单元之间自适应地分配任务，即实现负载均衡又达到节能的目的。这四项措施一起使用，把模型参数量压缩到原始 ResNet-50 的 5% 左右，但是仍然能保持工业解析任务所需要的精度水平。

## 4 多源数据实时解析框架

本文所设计的多源数据实时解析框架采用三层流水线结构，数据接入层、轻量推理层和应用输出层，三层之间用高性能事件总线（内存映射环形队列）异步解耦，屏蔽各层处理速率差异，消除阻塞依赖，保证整体系统低延迟、高吞吐。数据接入层对接各种工业协议驱动，把它们统一抽象成带有时间戳的数据帧（TLV 格式）。层内包含协议解码器池、PTP 时钟同步模块和缺失值补全子模块，时钟同步精度达到亚毫秒级，保证跨源数据帧的对齐，缺失值用历史滑动窗口的线性插值或者均值填充来避免由于传感器短暂失联造成的推理崩溃，协议解码器池可以按需实例化，支持 Modbus、OPC-UA 和自定义私有协议热插拔扩展。轻量推理层用固定步长滑动窗口（默认 256 个采样点）从事件总线上获取对齐数据帧，并根据事先设定好的触发策略（定时触发和事件驱动两种模式均可配置），将压缩后的 AI 模型部署到推理服务器上进行推理。推理引擎采用 INT8 量化模型，在 ARM Cortex-A53 平台上单次前向传播的时延小于 8ms，可以满足过程控制周期的要求。跨模态特征融合在推理层前端使用拼接 - 投影（Concat-Linear）结构把时序特征向量和视觉特征嵌入一起送入后面分类或者回归头，防止因为独立推理而产生的语义割裂。应用输出层把推理结果转换成工控指令、告警事件或者数字孪生状态更新，然后用 OPC-UA Publisher 或者 MQTT 客户端上报到上层系统。端到端（传感器采集、边缘推理、指令下发）平均时延实测小于 15ms，符合典型过程控制的实时性要求。框架整体用微服务形式容器化部署，支持在同一硬件节点上多个实例并行工作，实现多产线数据的并发解析 [5]。

## 5 实验验证与结果分析

为验证本文方法的有效性，在配备 NXP i.MX 8M Plus（ARM Cortex-A53 四核 1.8 GHz，2 GB LPDDR4）的工业边缘网关平台上搭建实验环境，操作系统为 Linux 5.10（PREEMPT-RT 实时补丁）。数据集使用公开的 CWRU 轴承振动故障数据集和模拟 Modbus/OPC-UA 混合协议采集数据，CWRU 数据集包含正常和 10 类故障共 11 类样本，用于故障分类任务；混合协议数据由仿真产线以 100ms 周期产生，用于多源对齐和数值解析任务的延迟测量。实验设置三个基线，（1）云端全量模型，ResNet-50 在服务器端推理，然后通过局域网将结果传回；（2）边缘原始模型，ResNet-50 FP32 直接部署到目标平台上；（3）传统规则解析，用专家定义的阈值和决策树做故障判断。本文的方法（MobileNetV2 骨干、40% 通道剪枝、INT8 QAT、知识蒸馏）和上述基线在解析准确率、推理延迟、内存占用这三个指标上做对比。从实验结果可以看出，本文方法对于 CWRU 故障分类任务的准确率为 97.3%，比传统的规则方法高出了 8.1 个百分点，和云端全量模型的差距只有 0.9 个百分点；单次推理延迟降低到 6.8ms，比边缘原始 ResNet-50 模型（大约 82ms）快了大约 12 倍；内存占用只有 47MB，比原来的模型要小得多。综上所述，本文方法在精度和效率之间取得较好的折中，端到端时延满足典型过程控制的实时性要求，有实际工业部署的可行性。

表 1 主流边缘 AI 模型压缩方法对比

| 方法 | 参数压缩率 | 精度损失 | 推理延迟变化 | 适用场景 |
| --- | --- | --- | --- | --- |
| 结构化剪枝 | 30%–60% | ＜ 2% | 延迟降低 20%–55% | 通用 CNN，无需稀疏加速库 |
| INT8 量化（QAT） | 内存减少约 75% | ＜ 1% | 延迟降低 40%–70% | ARM NEON/NPU 硬件加速 |
| 知识蒸馏 | 取决于学生规模 | 精度提升 1%–3% | 与学生架构规模相关 | 标注数据充足场景 |
| 本文（剪枝 +QAT+ 蒸馏） | 综合约 80% | ＜ 1.5% | 延迟降低约 92% | 工业边缘终端受限场景 |

## 6 结论与展望

本文针对工业互联网边缘终端多源数据实时解析算力不足的问题，提出用结构化剪枝、量化感知训练、知识蒸馏三者结合来设计轻量化 AI 模型的方法，并提出三层流水线的实时解析架构，即数据接入、轻量级推理、应用输出。根据实验结果可知，本文提出的方法在 ARM Cortex-A53 平台上的 97.3% 的故障分类准确率、6.8ms 的单次推理时延、47MB 的内存占用、端到端时延小于 15ms 等性能指标都优于传统的规则方法和直接使用全量模型，有很好的工程可行性、实用性。未来研究会沿着两条线发展，第一是用联邦学习框架实现多终端协同模型自适应增量更新，在保证工业数据隐私和安全合规的情况下不断改善解析模型在新工况下的泛化性能；第二是将方法应用到融合声发射、热成像、振动信号等更多样化的多模态工业场景中，并且尝试使用神经网络架构搜索的方式进行端到端的轻量化设计，从而大大减少模型压缩调优所花费的人力物力，促进边缘 AI 解析技术向更加广泛的工业互联网应用场景渗透。

## 参考文献

[1] 李响, 王健, 陈博. 工业互联网边缘计算技术综述. 计算机学报, 2022, 45(3): 612-638.

[2] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding[J]. International Conference on Learning Representations (ICLR), 2016.

[3] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.

[4] 张宇, 刘洋, 孙磊. 工业多源异构数据融合方法综述. 自动化学报, 2023, 49(1): 45-62.

[5] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. NIPS Deep Learning and Representation Learning Workshop, 2015.