

端侧量化与多模态协同实现资源受限下的口腔全景片高效智能诊断

赵涛涛, 焦粤豪, 倪 铭, 罗友璐, 夏顺兴, 徐颖颖, 何亚婷

四川农业大学信息工程学院, 四川 雅安 625014

摘要:目的 提出一种以前期高精度算法 YOLOv11-TDSP 为基准的全场景端侧量化部署与多模态协同诊疗框架。方法 构建基于 KL 散度的数据感知型训练后量化(PTQ)机制,通过校准集锁定特征最佳分布区间,将 FP32 模型近乎无损压缩至 INT8 精度。建立跨平台异构编译流水线,针对 Intel Hybrid Architecture(大小核架构)、Rockchip RK3588 及 Ascend 310P 进行算子深度融合与指令集优化,实现软硬件极致解耦。构建面向资源受限边缘终端的“检诊一体化”协同推理架构,利用极致量化的检测模型释放关键算力,驱动本地部署的 Qwen2.5-3B 轻量化大模型,结合专家规则库实现从视觉检测到文本报告的端侧闭环。结果 INT8 量化后的模型权重仅为 6.1 MB(压缩率 57.3%),mAP@0.5 保持在 95.4%,精度损失仅为 0.4%。模型在标准办公终端与资源受限的移动诊疗设备上均表现出卓越的泛化性能。特别是在算力与功耗双重受限的移动巡诊场景(Intel i5-12450H)中,推理速度提升至 22.91 FPS,实现了 3.64 倍的加速比。系统在完全断网且仅有 8GB 内存的普通设备上,成功实现了毫秒级病灶检测与符合临床规范的报告生成。结论 该框架有效解决了高精度算法在低配硬件上的部署瓶颈,在确保患者隐私安全的前提下,实现了口腔全景片分析从“云端依赖”向“边缘普惠”的范式转变,为基层医疗机构的智能化升级提供了一种低成本、高可靠的工程化解决方案。

关键词:口腔全景片;YOLOv11-TDSP;边缘计算;隐私保护;本地大语言模型

Efficient dental panoramic radiographic diagnosis under resource constraints via edge quantization and multi-modal synergy

ZHAO Taotao, JIAO Yuehao, NI Ming, LUO Youlu, XIA Shunxing, XU Yingying, HE Yating

College of Information Engineering, Sichuan Agricultural University, Ya'an 625014, China

Abstract: **Objective** To propose a full-scenario framework for edge-side quantization deployment and multi-modal collaborative diagnosis and treatment based on the prior high-precision algorithm YOLOv11-TDSP. **Methods** A data-aware post-training quantization (PTQ) mechanism based on Kullback-Leibler divergence was constructed. The optimal distribution range of the features was determined using the calibration dataset for near-lossless compression of the FP32 model to INT8 precision. A cross-platform heterogeneous compilation pipeline was then established, with in-depth operator fusion and instruction set optimization for Intel Hybrid Architecture, Rockchip RK3588 and Ascend 310P to achieve extreme software-hardware decoupling. An integrated detection-diagnosis collaborative inference architecture for resource-constrained edge devices was constructed. The ultra-quantized detection model was used to free critical computing resources to drive the locally deployed lightweight large model Qwen2.5-3B, which, along with an expert rule base, allowed for an edge-side closed loop from visual detection to text report generation. **Results** Experiments showed that the INT8-quantized model had a weight of only 6.1 MB (57.3% compression ratio), with mAP@0.5 of 95.4% and only a 0.4% precision loss. The model demonstrated excellent generalization on both standard office terminals and resource-limited mobile diagnostic devices. In the mobile diagnostic scenario constrained by computing power and power consumption (Intel i5-12450H), the inference speed of the model reached 22.91 FPS with a 3.64 speedup. The system achieved millisecond-level lesion detection and clinically compliant report generation on devices with only 8 GB RAM in a fully offline environment. **Conclusion** This framework allows deployment of high-precision algorithms on low-end hardware and enables a paradigm shift of panoramic dental radiograph analysis from cloud dependence to edge inclusiveness while ensuring patient privacy, providing a low-cost, high-reliability engineering solution for intelligent upgrading of primary medical facilities.

Keywords: panoramic dental radiograph; YOLOv11-TDSP; edge computing; privacy protection; local large language model

世界卫生组织报告与流行病学调查显示,当前全球及我国面临的口腔健康防治形势依然严峻^[1,2]。口腔全景片作为基层初筛与治疗规划的首选影像学手段,由于成像固有的解剖结构重叠与畸变,其解读极度依赖医师的主观经验,在医疗资源匮乏的基层极易发生漏诊与误

诊^[3-5]。近年来,计算机视觉算法在医学影像辅助诊断中取得了显著进展^[6,7],例如基于 Mask R-CNN 的实例分割^[8]、基于 Transformer 的联合检测^[9]以及 YOLO 系列轻量化算法在复杂牙科场景中的广泛应用^[10,11],大幅提升了病灶检测的自动化水平。特别是在前期研究中,本团队提出了一种轻量化高精度异常牙检测模型 YOLOv11-TDSP^[12],该算法在受控实验环境下展现出了卓越的病灶识别精度与临床应用潜力。然而,尽管以 YOLOv11-TDSP 为代表的高精度医学 AI 模型在算法

收稿日期:2026-02-13

基金项目:四川省自然科学基金(2022NSFSC0172)

作者简介:赵涛涛,在读硕士研究生,E-mail: 1425067966@qq.com

通信作者:倪 铭,副教授,硕士生导师,E-mail: nm@sicau.edu.cn

指标上成果丰硕,但若要将其真正下沉至基层临床场景,仍面临三重严峻的“落地”壁垒,这构成了本研究开展的核心依据:其一,算力与成本壁垒。主流高精度检测模型极度依赖昂贵的国外高性能 GPU,基层卫生院受限于预算难以负担,而分布式边缘 AI 架构已被证明是解决海量医疗数据计算的有效途径^[13-15];其二,数据隐私与网络壁垒。医疗影像具有极高的隐私敏感性,严格限制了“数据上云”的模式,边缘计算能提供保护隐私的本地化部署方案,最小化患者健康信息泄露风险^[16-18];其三,临床交互体验壁垒。传统的计算机辅助诊断系统仅能输出生硬的检测框,缺乏生成结构化诊断报告的能力,难以满足医生“检诊一体化”的实际 workflow 需求,而大型语言模型(LLMs)及视觉-语言大模型(VLMs)在医疗健康领域已被证明具备解决复杂临床任务和生成结构化医学报告的巨大潜力^[19-23]。此外,在当前全球科技博弈背景下,过度依赖国外算力底座更存在供应链隐患。因此,如何摆脱高昂算力依赖,利用国产化硬件在边缘侧构建低成本、高精度的多模态诊疗闭环,成为了打通 AI 落地基层的破局关键。

面对上述多重约束,本研究立足于基层口腔医疗的实际需求与国家医疗装备自主可控的战略导向,创新性地提出了一套兼容存量通用硬件与国产异构算力的边缘智能诊疗框架。本研究的主要创新点与核心贡献体现在以下 3 个层面:首先,提出了一种基于数据分布校准的极低比特量化策略。以本团队前期提出的高精度算法 YOLOv11-TDSP^[12]为基准,本研究构建了基于 KL 散度的数据感知型训练后量化(PTQ)机制^[24, 25],旨在不重新训练的前提下将模型近乎无损压缩至 INT8 精度,从算法底层突破边缘设备的内存与计算瓶颈。其次,设计了面向国产 NPU 的异构编译与算子融合流水线。为解决现有 CUDA 架构模型难以直接适配国产硬件的难题,本框架针对华为昇腾(Ascend 310P)及瑞芯微(RK3588)等国产芯片进行了深度的算子重构与图优化,打通了核心算法全国产化替代的工程链路^[26]。最

后,构建了“视觉-语言”多模态本地协同诊疗范式。借助极致量化释放的终端硬件资源,本研究在无独立显卡的纯 CPU 基层存量设备上,成功实现了 Qwen2.5-3B 大语言模型的本地化部署^[27],创新性地打通了从“影像病灶检测”到“自然语言结构化报告生成”的端侧跨模态闭环。

1 材料和方法

1.1 材料

1.1.1 实验数据集 本研究以团队前期已发表的高精度检测算法 YOLOv11-TDSP^[12]为目标对象。需要特别指出的是,该视觉检测算法的深度学习网络训练与权重收敛过程已在其算法研发阶段完整完成,本研究的部署全流程无需对该模型进行任何形式的重新训练。文中构建的包含 100 张真实临床口腔全景片的数据集,其本质是专用于数据感知型训练后量化(PTQ)机制的典型影像校准集与独立量化性能评估集。该数据集旨在通过真实临床样本的激活值分布,计算最优量化缩放因子与零点,并验证模型在低功耗边缘端部署前后的推理一致性与泛化能力。样本涵盖了龋齿、阻生齿等 8 种典型的牙齿异常类型,各类别的具体样本分布如表 1 所示。

表 1 测试集牙齿异常类别分布

Tab.1 Distribution of dental anomaly categories in the test dataset

Species	Sample count
Decayed tooth	30
Impacted tooth	41
Ectopic Eruption	32
Root canal therapy	25
Dental crown	63
Dental implant root system	39
Dental implant	26
Dental filling	11

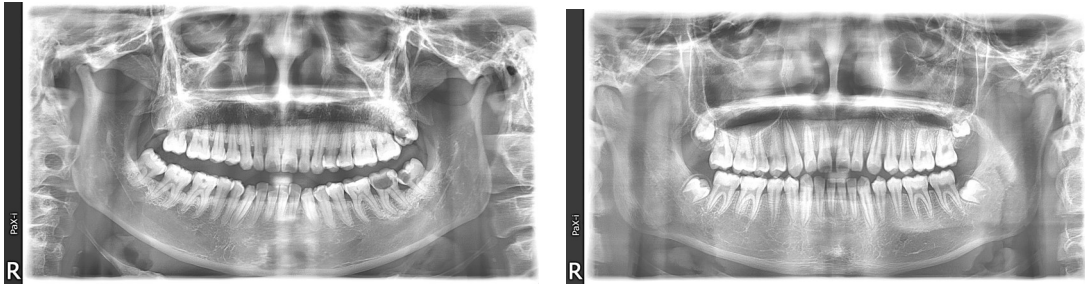


图 1 口腔全景片影像示例
Fig.1 Example of panoramic dental radiograph.

1.1.2 实验硬件与软件环境 在硬件环境方面,为实验具有代表性,本文对常见口腔诊所以及医院的设备进行调研,实验数据均参考各医院实际硬件配置。为了覆盖从三甲医院到基层乡镇卫生院的实际算力跨度,本研究选取了具有代表性的异构计算设备进行推理评估:包括高性能 X86 工作站(Intel i5-12600KF)、以及乡镇卫生院常见的陈旧低功耗设备(Intel i3-8100T)。在边缘计

算平台方面,为响应国产化与边缘智能趋势,本研究搭建了基于 Rockchip RK3588 和 Huawei Ascend 310P 的推理环境。其中,RK3588 端利用 RKNN-Toolkit2 进行模型转换与加速;Ascend 端则基于 CANN 架构与 ACL C++接口部署。详细的软硬件环境配置分别如(表2、3)。

表2 RK3588 边缘端实验环境配置

Tab.2 Experimental environment and software configuration (Rockchip RK3588)

Component	Software/Library	Version	Role in system
Hardware platform	Rockchip RK3588	NPU 6.0 TOPS	Edge inference device
Operating system	Linaro linux	Alip (Debian)	Host OS
Deep learning framework	PyTorch	2.2.0	Model training & pruning
Model conversion	RKNN-Toolkit2	2.3.2	Quantization & export
Inference engine	RKNN-Toolkit-Lite2	2.3.2	On-device inference
Intermediate format	ONNX	1.16.1	Model exchange
Computer vision	OpenCV-Python	4.11.0	Image preprocessing
Scientific computing	NumPy	1.26.4	Tensor operations

表3 昇腾边缘端实验环境配置

Tab.3 Experimental environment and software configuration (Huawei Ascend)

Component	Software/Library	Version	Role in system
Hardware platform	Huawei Ascend server	Ascend 310P (NPU)	Ai inference accelerator
Operating system	openEuler	22.03 (LTS-SP3) aarch64	Host operating system
Deeplearning framework	PyTorch	2.1.0/2.2.0 (compatible)	Model training & export
Development toolkit	Ascend CANN Toolkit	8.0.0	Npu driver & compiler stack
Model conversion	ATC (ascend tensor compiler)	8.0.0	ONNX to OM conversion
Inference engine	AscendCL (ACL)	8.0.0	On-device C++ inference
Intermediate format	ONNX	1.14.0+	Model exchange format
Video processing	FFmpeg & x264	4.1.3 / Latest	Video decoding & encoding
Computer vision	OpenCV	4.x (C++)	Image preprocessing
Compiler	GCC/G++	7.3.0+ (system default)	Sample code compilation

1.2 基于数据分布校准的量化策略

为解决通用 CPU 及边缘设备推理延迟高的问题,本文引入一种基于 KL 散度 (Kullback-Leibler Divergence)的数据感知型训练后量化(PTQ) 机制。

传统的线性量化易受激活值分布中的离群点 影响导致精度截断。为此,我们构建了包含 100 张典型影像的校准集。该策略的目标是寻找最优的缩放因子 S 和零点 Z,使量化后的 INT8 激活值分布 P_{int8} 与原始 FP32 分布 Q_{fp32} 之间的 KL 散度最小化。该机制通过截断长尾部分的离群点(Outliers),锁定特征最集中的最佳校准区间(图 2)。通过该校准过程,模型能够学习牙齿特征在数值域上的统计分布,从而在将权重位宽从

32-bit 压缩至 8-bit 的过程中,最大限度地抑制量化噪声,保留对微小病灶的特征提取能力。蓝色曲线表示原始 FP32 权重的概率密度分布,绿色阶梯线表示量化后的 INT8 能级。红色虚线展示了通过最小化 KL 散度搜索到的最优截断阈值,该策略有效去除了两端的离群点噪声,保留了核心特征信息。

1.3 量化相关公式

量化过程旨在建立浮点数 r 与整数 q 之间的映射关系。本研究采用非对称线性量化方案,其数学表达如下:

公式 1 (映射变换):

$$q = \text{clip}(\text{round}(\frac{r}{S} + Z), -128, 127)$$

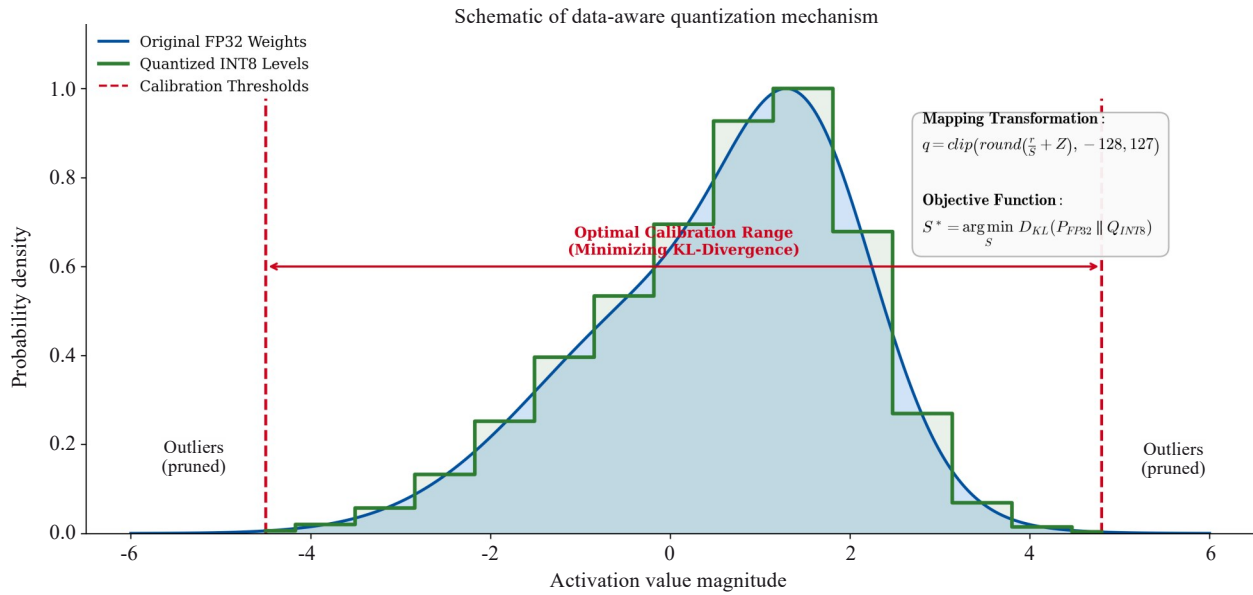


图2 基于 KL 散度的动态校准量化原理示意图

Fig.2 Schematic illustration of data-aware quantization mechanism based on KL divergence.

其中, S (Scale) 为缩放因子, Z (Zero-point) 为零点偏移量, $\text{clip}(\cdot)$ 函数用于将数值限制在 $q_{\min}$, $q_{\max}$ 范围内(对于 INT8, 范围为 $[-128, 127]$)。

公式2 (反量化):

$$r = S(q - Z)$$

公式3 (KL 散度优化目标): 为了确定最优的 S 和 Z , 我们通过最小化 FP32 分布 P 与 INT8 量化分布 Q 之间的 KL 散度来求解:

$$S^* = \arg \min_S D_{KL}(P_{FP32} \| Q_{INT8}) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

1.4 面向 RK3588 的边缘端模型转换

本研究构建了从训练端到边缘端 (RKNN) 的模型转换链路。

1.4.1 ONNX 导出与算子重构 利用 PyTorch (v2.2.0) 将训练完成的权重文件导出为 ONNX 格式。针对自定义单头分裂注意力 (SHSA) 模块中算子在标准 ONNX Opset 12 中的映射约束, 我们在导出阶段采用 Conv-Sigmoid-Mul 门控机制替换原有的 Self-Attention 算子 (图 3), 以构建适配边缘 NPU 算子库的计算图

1.4.2 NPU 量化编译 选用 RKNN-Toolkit2 (v2.3.2) 工具链, 对优化后的 ONNX 模型执行混合精量化, 并利用校准集计算量化参数, 最终生成适配 RK3588 NPU 的 .rknn 离线文件。

1.5 针对昇腾架构的算子映射与图优化

针对昇腾 AI 处理器的张量计算特性, 本研究构建了从 ONNX 到 OM 的异构编译流水线 (图 4)。

1.5.1 AIPP 硬件加速注入 在 ATC 编译阶段开启 FP16 模式并插入 AIPP 算子, 利用硬件单元处理色域转换 (YUV to RGB), 释放 CPU 算力。

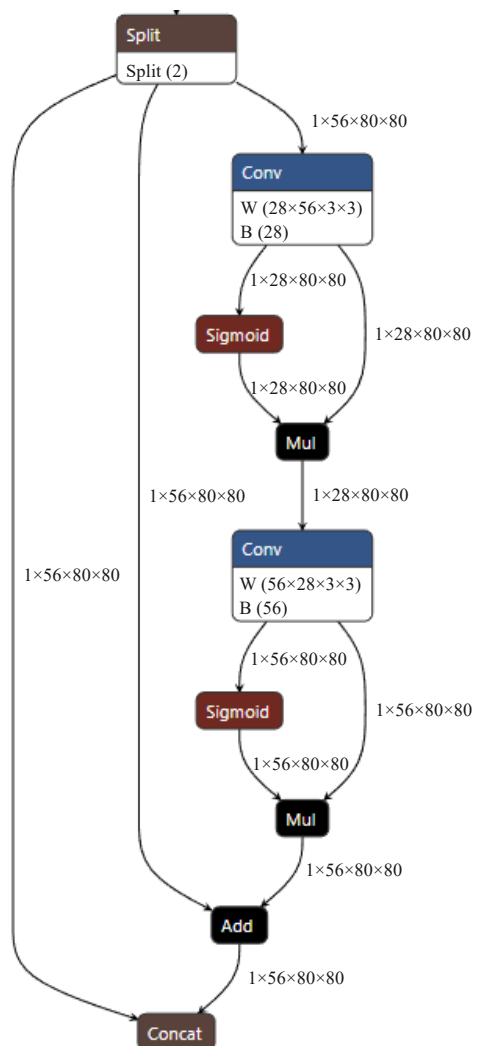


图3 适配边缘 NPU 的 SHSA 模块算子重构示意图

Fig.3 Operator restructuring diagram of the SHSA module adapted for Edge NPU.

1.5.2 算子深度融合 利用 ATC 工具的高级图融合策略,将 SHSA 模块中的连续节点映射为单一融合算子

(Fused Operator),减少内存搬运开销。

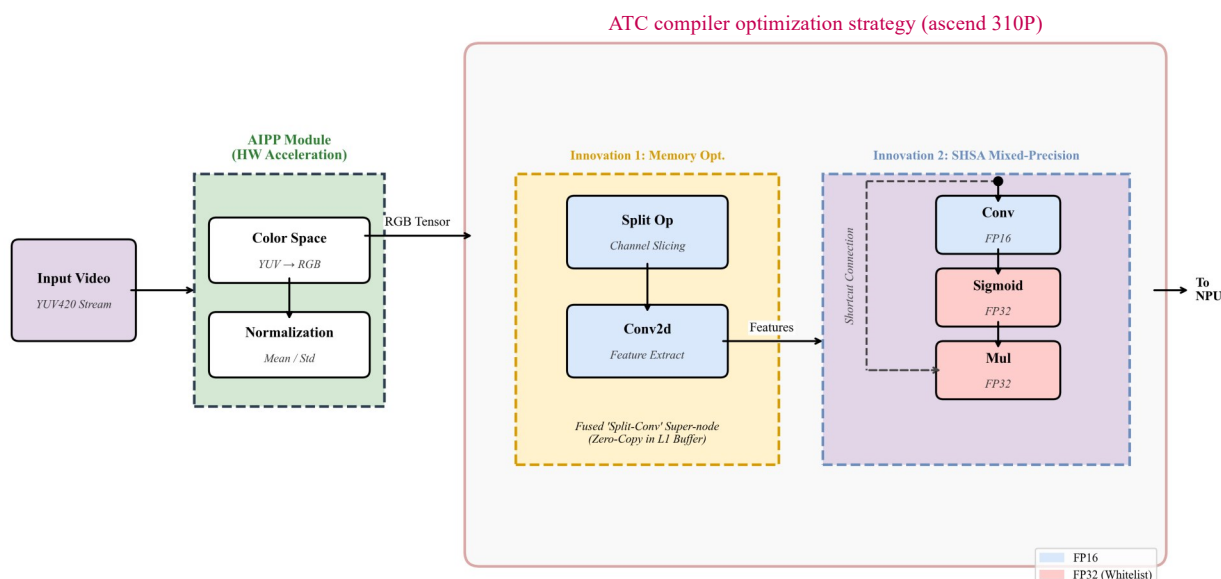


图4 YOLOv11-TDSP 异构编译流水线示意图

Fig.4 Optimized heterogeneous compilation pipeline for YOLOv11-TDSP.

1.6 本地大模型生成报告的盲评量化机制

为客观评估边缘端轻量化大模型(Qwen2.5-3B)生成的结构化医学报告的专业质量,本研究引入了临床双盲评价机制。从测试集中随机抽取 30 份具有代表性的全景片及其对应的AI生成报告。邀请 3 名具备口腔临床与影像诊断经验的医师组成独立评估小组。评估全程隐去报告的生成来源,评估医师需采用 5 分制李克特量表(1分=极差,5分=极好),从以下 3 个维度对报告进行量化打分:准确性:报告对影像学异常(如龋齿、阻生齿位置等)的描述是否与客观影像一致,无“幻觉”或误报;完整性:报告是否全面涵盖了检测模型识别出的所有关键病灶,有无遗漏;规范性:报告的格式、医学术语的使用以及处置建议的表述是否符合现行口腔临床规范。

2 结果

2.1 RK3588 与昇腾的计算图优化与编译效率分析

针对改进型 YOLOv11-TDSP 模型中自定义 SHSA 模块在异构 NPU 上的适配问题,本研究实施了分层级的计算图优化策略。

2.1.1 计算图拓扑精简 经 ONNXSlim 静态图优化工具处理后,模型的拓扑结构得到了显著精简。实验统计显示,优化过程成功剔除了计算图中 35% 的冗余节点... 并将所有动态形状算子完全转换为静态算子。回顾前文图 2 所示的重构结构,这种 NPU 友好的拓扑设计消除了非推理必要的计算分支,经测试实现了 100% 的算

子映射率,为后续的硬件加速奠定了基础。

2.1.2 RK3588 边缘端映射 在 Rockchip RK3588 平台上,重构后的 Conv-Sigmoid-Mul 门控分支成功通过了 RKNN 编译器的算子检查。实验结果表明,该模型在 RK3588 NPU 上实现了 100% 的算子映射率,未出现因算子不支持而回退至 CPU 执行(CPU Fallback)的现象,有效保证了边缘端的推理一致性。

2.1.3 昇腾达芬奇架构适配 针对华为昇腾(Ascend)平台,本研究结合达芬奇架构的数值特性进行了深度适配:数值稳定性优化:考虑到 Sigmoid 算子在 FP16 半精度模式下易产生数值截断误差,本策略遵循华为 CANN 开发文档的最佳实践,在量化过程中对 Transformer 类结构(含 SHSA Attention)锁定 FP32 精度计算。这一基准设计消除了量化噪声对注意力权重的干扰,确保了推理精度。算子深度融合:利用 ATC (Ascend Tensor Compiler) 的图融合技术,我们将 Split 与 Conv 操作构建为单一超节点。这种融合机制消除了中间张量的 DDR 读写操作,实现了片上缓存(On-chip Buffer)内的数据直接透传。硬件加速预处理:通过配置 AIPP(AI Pre-processing)模块,将 YUV 格式转换与 Normalization 操作从 CPU 剥离,完全下沉至硬件层面执行,进一步降低了端到端延迟。

2.2 口腔诊所 CPU 端多平台推理性能测试

在前一节中,我们验证了算法在国产 NPU 异构计算平台上的算子兼容性与部署可行性。然而,考虑到当前基层医疗机构(尤其是乡镇卫生院)尚未大规模普及

专用的 AI 硬件加速器,通用 CPU 依然是临床辅助诊断最主要的运行载体。因此,本节将评估重心聚焦于“通用 CPU 设备的普适性测试”,旨在验证量化模型能否在缺乏独立显卡的普通办公电脑上实现“低功耗”下的实时诊断。为了验证量化模型在真实临床环境中的通用性,本研究构建了覆盖“三甲医院影像中心”至“乡镇卫生院”的多层级硬件测试基准。表 2 展示了模型在不同 CPU 架构上的详细性能对比。

表 4 FP32 与 INT8 量化模型在不同临床硬件平台上的性能对比)
Tab.4 Performance comparison of FP32 and quantized INT8 models across various clinical scenarios

Deployment scenario	Hardware specification	Precision	Latency (ms) ↓	Throughput (FPS) ↑	Speedup
Research center (lab server)	Intel i7-12700	FP32	72.50	13.79	1.00
		INT8 (Ours)	38.15	26.21	1.9
Doctor's office	Intel i5-12400	FP32	98.6	10.14	1.00
		INT8 (Ours)	56.34	17.75	1.75
Mobile clinic	Intel i5-12450H	FP32	158.71	6.30	1.00
		INT8 (Ours)	43.66	22.91	3.64
Doctor's office	Intel i5-12600KF	FP32	79.04	12.65	1.00
		INT8 (Ours)	43.12	23.19	1.83
Mobile clinic	Intel i7-10750H	FP32	139.12	7.19	1.00
		INT8 (Ours)	92.02	10.87	1.51
Doctor's office	Intel i7-7700	FP32	125.00	8.00	1.00
		INT8 (Ours)	79.36	12.60	1.58
Rural clinic	Intel i5-6500	FP32	172.41	5.80	1.00
		INT8 (Ours)	106.38	9.40	1.62
Rural clinic	Intel i3-8100T	FP32	205.60	4.86	1.00
		INT8 (Ours)	135.80	7.36	1.51

实验结果表明,所提出的 INT8 量化策略实现了精度与速度的帕累托最优,在测试的硬件平台中,值得注意的是,在搭载 Intel i5-12450H 的移动诊疗终端上,量化模型实现了高达 3.64 倍的加速比(从 6.30 FPS 提升至 22.91 FPS)。INT8 量化模型均显著降低了推理延迟。这一显著提升的主要原因在于该处理器采用了 P-Core(性能核)与 E-Core(能效核)的混合架构。在基准 FP32 推理中,传统深度学习框架(如 PyTorch)在 Windows 环境下易出现线程调度次优,导致计算资源未被充分利用。而 OpenVINO 运行时通过针对 Intel Hybrid Architecture 的指令集优化(如 AVX-VNNI)及更高效的线程亲和性管理,不仅大幅降低了单帧延迟,更有效克服了移动端设备的功耗墙限制。这证明了本方案在移动巡诊、床旁推车等电池供电场景下具有不可替代的部署优势。

2.3 推理效果评估

实验结果表明,本文提出的基于 KL 散度的数据感知型量化策略在实现模型极致轻量化的同时,完整保留了临床级的诊断优越性。模型权重文件体积从 14.3 MB 大幅缩减至 6.10 MB,压缩率达 57.3%;然而,归一

化混淆矩阵(图 5)揭示,量化引入的细微误差主要局限于牙冠与填充物等高相似度类别的良性混淆,而针对龋齿等关键病灶的特征提取鲁棒性依然稳健量化前后的 P-R 曲线对比(图 6)进一步显示,INT8 模型在 mAP@0.5 指标上仅产生了<0.5%的微小波动,且在种植牙、阻生齿等具有刚性几何结构的类别上实现了极低损耗的高保真精度迁移,充分证明了该量化模型在边缘端部署中兼具高压缩比与高精度的帕累托最优特性。

2.4 本地化私有部署与多模态协同可行性分析

鉴于口腔医疗数据的高度敏感性及相关法规(如 HIPAA 或国内数据安全法)对患者隐私的保护要求,数据的“不出域、不脱网”已成为 AI 医疗落地的硬性约束。为此,本研究模拟了基层口腔医院最常见的标准办公环境(配置: Intel Core i5-12400 CPU, 8GB RAM, 无独立显卡)以及常用移动办公环境(Intel Core i5-12450H CPU, 8GB RAM),重点验证了量化模型在受限算力资源下的高吞吐批量推理能力及其与本地大语言模型协同运行的可行性。

最为关键的是,INT8 量化策略将 YOLOv11-TDSP 的显存/内存占用压缩至极致的 6.10 MB。这一极低的

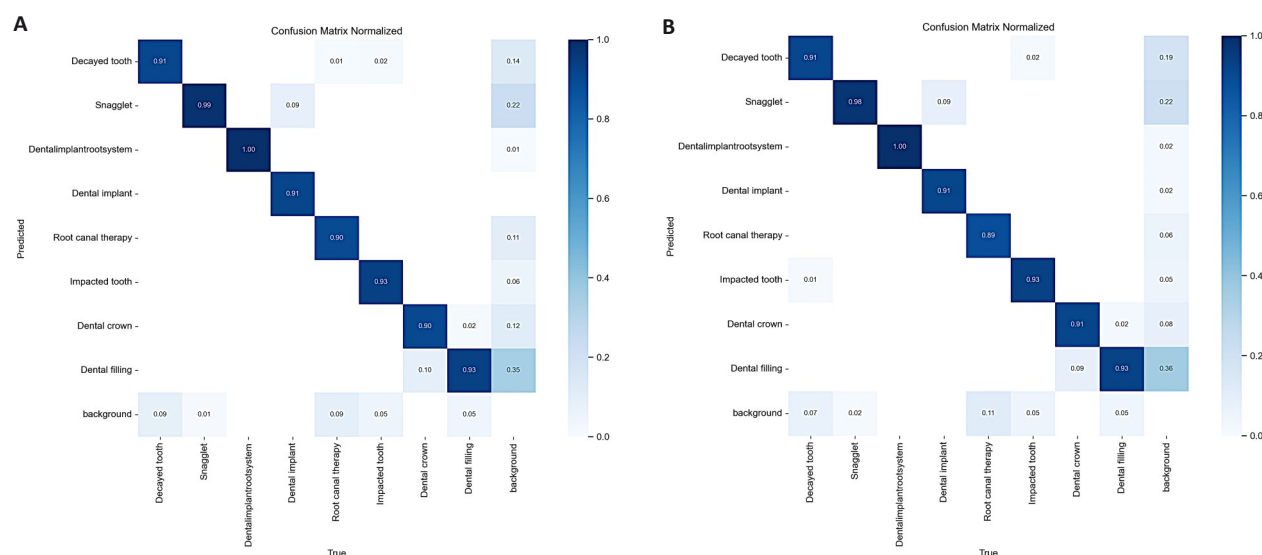


图5 混淆矩阵对比

Fig.5 Confusion matrix comparison. A: Confusion Matrix of YOLOv11-TDSP. B: Confusion matrix of quantized YOLOv11-TDSP.

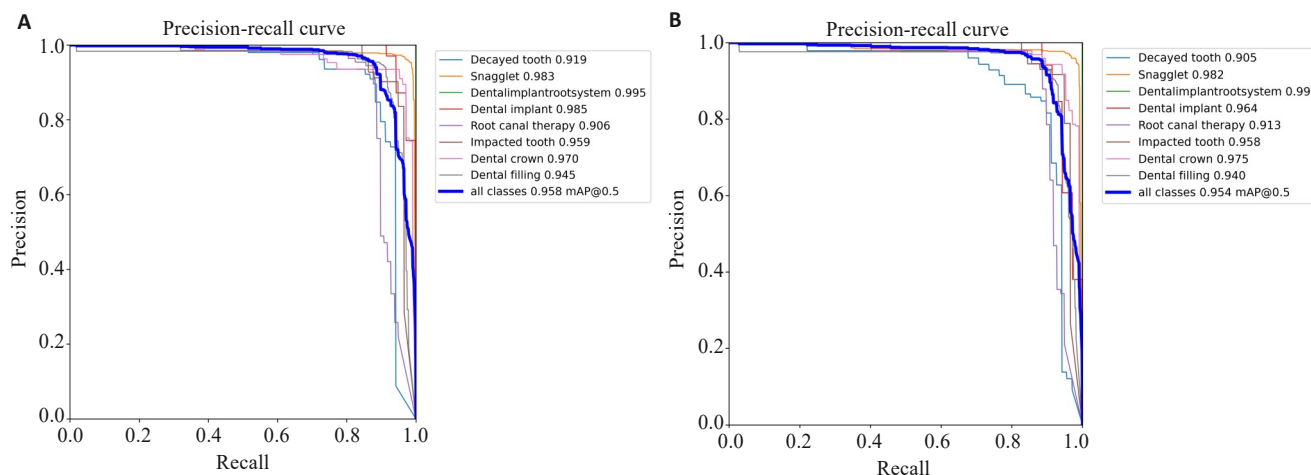


图6 YOLOv11-TDSP 算法量化前后 P-R 曲线及各类别 AP 值对比

Fig.6 Comparison of P-R curves and class-wise AP values for YOLOv11-TDSP before and after quantization. A: FP32. B: INT8.

资源开销成功突破了硬件瓶颈,为在仅有 8GB 内存的普通办公电脑上同时引入大语言模型创造了必要条件。

实验表明,量化检测模型通过释放关键的 CPU 算力与内存带宽,确保了系统能够流畅运行 Qwen2.5-3B 本地大模型。通过集成专家规则库并结合提示词工程,该系统在上述两类常用办公终端上均成功实现了从“视觉特征提取”到“文本诊断生成”的完全本地化闭环。如图 4 所示,该私有 AI 助手能够在完全断网的条件下,依据 OpenVINO 加速引擎的实时检测反馈,即时生成结构化的临床处置建议。这一结果有力证明了本研究提出的量化协同方案,通过拓展边缘设备的算力边界,使低功耗存量终端真正具备了运行多模态私有 AI 系统的能力。

由于篇幅限制,本文以图 7 的输出结果为例展示临床盲评的打分细节。在 5 分制评估中,该生成报告的完

整性获得满分(5.0 分),其结构完整,对检出的所有异常类别(龋齿、智齿)均进行了全面覆盖与闭环建议;规范性获得 4.1 分,其采用标准三段式影像报告模板,医学术语专业,处置建议符合临床常规诊疗逻辑且具备良好的合规性;准确性获得 4.0 分,影像学描述准确,但若增加具体牙位(如 38、48 等)的精准定位描述即可达满分。

2.5 大模型结构化报告临床盲评量化评估

基于 3 名临床医师对 30 份 AI 辅助诊断报告的双盲评分,系统生成的结构化文本在临床专业性上表现出了较高的可用性,具体量化评估结果(表 5)。

实验数据表明,系统在“完整性”维度得分最高(4.8 分),这得益于底层 INT8 检测模型(YOLOv11-TDSP^[12])提供的高精度先验边界框,确保了大模型在生成文本时极少遗漏关键病灶。在“准确性”和“规范性”方面,超过 80% 的报告获得了 4.1 分(良好)及以上评价,证实了 3B



图7 边缘端断网环境下私有 AI 助手的交互界面展示

Fig.7 Interface demonstration of the private AI assistant in an offline edge environment.

表5 边缘端轻量化大模型生成诊断报告的临床盲评量化结果 (5分制)

Tab. 5 Quantitative results of clinical blind review for diagnostic reports generated by the edge-side lightweight LLM (5-point scale)

Evaluation metrics	Mean±SD	Score≥4
Accuracy	4.2±0.6	86.0%
Completeness	4.8±0.5	93.0%
Standardization	4.1±0.7	80.0%

参数量的轻量化大模型在本地知识库与提示词工程的约束下,能够有效克服大模型常见的“幻觉”问题,生成的医学术语与处置建议基本符合基层口腔诊疗规范。尽管在面对部分复杂并发症时,系统偶尔出现术语表述生硬的现象(规范性评分略低),但整体而言,该端侧多模态协同框架已具备辅助基层医生“检诊一体化”的实用价值。

3 讨论

在数字化口腔医疗的推进过程中,尽管高精度的深度学习模型已在实验室环境下取得了卓越的检测性能,但其在真实临床场景的落地仍面临着算力资源受限、推理实时性不足以及患者隐私数据“不出域”的三重挑战。为有效打破这些壁垒,本文创新性地提出了一套面向全场景的“端侧量化与多模态协同诊疗框架”。该框架以 YOLOv11-TDSP 为核心,通过基于 KL 散度的数据感知量化策略与异构编译优化,成功将高算力需求的 AI 模型迁移至低功耗的边缘终端。这与孙召飞等^[28]提出的推动人工智能技术向口腔临床真实应用转化的工程化

方向高度一致。

从量化策略的精度与效率权衡来看,本研究取得了理想的帕累托最优(Pareto Optimality)。实验数据显示,量化后的模型权重仅为 6.1MB(压缩率高达 57.3%),且 mAP@0.5 依然保持在 95.4%。现有的口腔影像轻量化研究(如 Lin 等^[29])多倾向于通过大幅删减网络层数或通道数来降低算力消耗,但这往往以牺牲细粒度特征提取能力为代价。与此不同,本研究采用的数据感知型 PTQ 机制展现出独特的正交优势。正如杨春等^[30]以及国际底层量化研究^[31]指出的神经网络压缩趋势,本实验中的 P-R 曲线对比揭示,INT8 模型相较于 FP32 基准仅产生<0.5%的边际衰减(0.958 vs 0.954)。这证明了通过 KL 散度截断“长尾”离群点、锁定特征最佳分布区间的策略,能够有效抑制量化噪声对早期龋齿等模糊病灶特征的破坏,实现了极低损耗的高保真精度迁移。

在跨平台边缘推理效能方面,本研究实现了突破性进展。在受限功耗与散热的移动诊疗终端(Intel i5-12450H)上,基准 FP32 模型的推理延迟高达 158.7 ms,难以满足临床实时交互需求。经量化与指令集优化后,延迟骤降至 43.7ms,实现了高达 3.64 倍的加速比,吞吐量跃升至 22.91 FPS。当前学术界在探讨医疗边缘计算时(如 Rancea 等^[13]),通常倾向于依赖“云-边协同”架构来缓解终端的算力压力。而本研究的实测数据证实,通过深度挖掘底层通用硬件(如纯 CPU 的大小核架构)的整数计算潜能,完全有可能在单一边缘节点、无独立显卡设备上实现高吞吐量的实时推理。此外,模型在国产化硬件(RK3588 与 Ascend 310P)上实现的 100% 算子

映射率,进一步验证了该方案在软硬件解耦方面的巨大潜力,直接用实际工程数据呼应了李斌等^[32]关于加快高端医疗装备国产化替代的战略前景分析。

综合资源经济性与临床应用价值可知,本文构建的“视觉-语言”多模态本地化协同模式具有开创性意义。正如 Singhal 等^[19]在《Nature》上所证实,大型语言模型已具备编码复杂临床知识的能力;但林岚等^[33]亦明确指出,其在基层医疗的本地化部署面临着庞大算力成本与数据隐私的双重掣肘。本研究的核心创新在于“以空间换智能”:实测结果表明,利用极致量化的视觉检测模型(仅 6.1 MB 显存占用)成功打破了边缘设备的“内存墙”,为在仅有 8GB 内存的底层医用办公电脑上并行驱动 Qwen2.5-3B 大语言模型创造了极其关键的算力冗余。这种在完全断网环境下的纯本地运行模式,不仅从物理层面彻底杜绝了患者影像数据上云带来的隐私泄露风险^[16],更通过自动生成结构化诊断报告,有效填补了传统 CAD 系统“重检测、轻解释”的临床应用断层。此外,针对基层临床实践中多变的影像质量(如低质量或含金属伪影的输入)与容错需求,本级联系统在架构设计上具备高度的灵活性与可扩展性。底层的量化检测模型(YOLOv11-TDSP)在提取病灶特征时,原生具备输出各病灶检测置信度的能力。得益于视觉检测与语言生成的软硬件解耦设计,终端用户或医疗机构完全可以根据具体的业务容错需求与设备算力余量,通过自定义提示词工程与专家规则库,灵活设置置信度动态拦截机制。例如,可通过 Prompt 指令大模型:当接收到视觉端低于特定安全阈值的置信度时,自动在生成的结构化报告中触发“病灶特征模糊/图像质量欠佳,建议结合临床人工复核”的预警提示。这种高度可配置的跨模态级联策略,不仅赋予了系统极强的场景适应能力,也为边缘 AI 辅助诊断在复杂环境下的鲁棒性提供了可靠的技术预留机制。

综上所述,本文所提出的基于 YOLOv11-TDSP 的全场景量化部署方案,在口腔全景片智能分析任务中,展现出了高精度、低延迟与强隐私保护的显著优势。不过,该方法仍存在一定的局限性。一方面,虽然 3B 参数量的轻量化大模型能够生成规范的诊断报告,但在病种覆盖度上仍存在局限。由于本级联框架的多模态诊断高度依赖上游基准模型(YOLOv11-TDSP^[12])提供的检测框与置信度先验,而该视觉模型在前期训练阶段主要聚焦于牙齿单体病理(如龋齿、阻生齿等),尚未纳入牙周疾病(如牙槽骨吸收、骨缺损)及已补牙继发龋坏的影像特征。因此,在缺乏上游视觉输入的情况下,当前的大模型框架暂无法独立完成对上述并发症的多模态分析与识别。在面对这些超出视觉先验的罕见或复杂病理时,其推理深度与逻辑链条仍不及云端千亿级大模

型。另一方面,目前的量化策略主要针对 CNN 结构的视觉模型,针对 Transformer 架构大语言模型(LLMs)的边缘端极致量化与加速尚未在本文中深入展开。鉴于此,在未来的研究工作中,将着重探索端侧大模型的 4-bit 量化技术(如 GPTQ、AWQ)^[34,35]以及联邦学习机制^[36],在确保隐私的前提下利用多中心数据持续迭代模型的认知能力。此外,顺应国家医疗装备与 AI 算力国产化替代的战略大趋势,未来的研究重点将进一步聚焦于多模态大模型在国产 NPU 异构计算架构(如华为昇腾、瑞芯微等)上的深度适配。通过探索“算法-硬件”协同优化范式,全面释放国产 AI 芯片的底层算力潜能,从而为我国基层智慧医疗体系构建更加安全、自主、可控的软硬件一体化智能基座。

Declaration of interests: The authors declare no competing interests.

参考文献:

- [1] World Health Organization. Global oral health status report: towards universal health coverage for oral health by 2030[R]. Geneva: WHO, 2022.
- [2] 陆海霞,陶丹英,卢展民,等.第四次全国口腔健康流行病学调查-背景和方法[C]//2018年中华口腔医学会第十八次口腔预防医学学术年会论文集.西安,2018:23.
- [3] Ahmed N, Abbasi MS, Zuberi F, et al. Artificial intelligence techniques: analysis, application, and outcome in dentistry: a systematic review[J]. BioMed Res Int, 2021, 2021: 9751564.
- [4] Tuzoff DV, Tuzova LN, Bornstein MM, et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks[J]. Dentomaxillofac Radiol, 2019, 48(4): 20180051.
- [5] Ossowska A, Kusiak A, Świetlik D. Artificial intelligence in dentistry: narrative review[J]. Int J Environ Res Public Health, 2022, 19(6): 3449.
- [6] Zhao XT, Xu TK, Peng L, et al. Recognition and segmentation of teeth and mandibular nerve canals in panoramic dental X-rays by Mask RCNN[J]. Displays, 2023, 78: 102447.
- [7] Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision[J]. NPJ Digit Med, 2021, 4(1): 5.
- [8] Cui ZM, Li CJ, Wang WP. ToothNet: automatic tooth instance segmentation and identification from cone beam CT images[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2020: 6361-70.
- [9] Sheng C, Wang L, Huang ZH, et al. Transformer-based deep learning network for tooth segmentation on panoramic radiographs[J]. J Syst Sci Complex, 2023, 36(1): 257-72.
- [10] Ragab MG, Abdulkadir SJ, Qaid N, et al. Periodontitis bone loss detection in panoramic radiographs using modified YOLOv7[J]. PeerJ Comput Sci, 2025, 11: e3102.
- [11] 梁洪,邱定乾,丁世宇,等.基于改进YOLOv8的异常牙齿和修复体X射线影像检测[J].中国激光,2025,52(3): 0307106.
- [12] 赵涛涛,倪铭,夏顺兴,等.YOLOv11-TDSP:口腔全景片的轻量化高精度异常牙检测模型[J].南方医科大学学报,2025,45(8):

- 1791-9.
- [13] Rancea A, Anghel I, Cioara T. Edge computing in healthcare: innovations, opportunities, and challenges[J]. *Future Internet*, 2024, 16(9): 329.
- [14] Chen JS, Ran XK. Deep learning with edge computing: a review[J]. *Proc IEEE*, 2019, 107(8): 1655-74.
- [15] Shi WS, Cao J, Zhang Q, et al. Edge computing: vision and challenges[J]. *IEEE Internet Things J*, 2016, 3(5): 637-46.
- [16] Wang RJ, Lai JS, Zhang ZY, et al. Privacy-preserving federated learning for Internet of medical things under edge computing[J]. *IEEE J Biomed Health Inform*, 2023, 27(2): 854-65.
- [17] Rieke N, Hancox J, Li WQ, et al. The future of digital health with federated learning[J]. *npj Digit Med*, 2020, 3: 119.
- [18] Yang Q, Liu Y, Chen TJ, et al. Federated machine learning: concept and applications[J]. *ACM Trans Intell Syst Technol*, 2019, 10(2): 1-19.
- [19] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. *Nature*, 2023, 620(7972): 172-80.
- [20] Sha YY, Pan HX, Meng WY, et al. Contrastive knowledge-guided large language models for medical report generation[M]//*Medical Image Computing and Computer Assisted Intervention - MICCAI 2025*. Cham: Springer Nature Switzerland, 2025: 111-20.
- [21] Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence[J]. *Nature*, 2023, 616(7956): 259-65.
- [22] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine[J]. *Nat Med*, 2023, 29(8): 1930-40.
- [23] Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI[J]. *Nejm Ai*, 2024, 1(3): 14334.
- [24] Qu CY, Zhao R, Yu Y, et al. Post-training quantization for 3D medical image segmentation: a practical study on real inference engines[EB/OL]. 2025: arXiv: 2501.17343. <https://arxiv.org/abs/2501.17343>.
- [25] Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 2704-13.
- [26] Cai H, Gan C, Wang TZ, et al. Once-for-all: train one network and specialize it for efficient deployment[EB/OL]. 2019: arXiv: 1908.09791. <https://arxiv.org/abs/1908.09791>.
- [27] Yang A, Li AF, Yang BS, et al. Qwen3 technical report[EB/OL]. 2025: arXiv: 2505.09388. <https://arxiv.org/abs/2505.09388>
- [28] 孙召飞, 俞经虎, 朱行飞, 等. 基于改进YOLOv5s的口腔全景片牙齿病症识别算法[J]. *中国激光*, 2024, 51(15): 1507106.
- [29] Lin SY, Hao XJ, Liu Y, et al. Lightweight deep learning methods for panoramic dental X-ray image segmentation[J]. *Neural Comput Appl*, 2023, 35(11): 8295-306.
- [30] 杨春, 张睿尧, 黄洸, 等. 深度神经网络模型量化方法综述[J]. *工程科学学报*, 2023, 45(10): 1613-29.
- [31] Gholami A, Kim S, Dong Z, et al. A survey of quantization methods for efficient neural network inference[M]//*Low-power computer vision*. Chapman and Hall/CRC, 2022: 291-326.
- [32] 李斌, 汪黎君, 曹厚德, 等. 基于上海市场十年调查数据看医疗设备“中国制造”的提升[J]. *中国医疗设备*, 2018, 33(2): 23-6.
- [33] 林岚, 武雨桐. 大型语言模型在医疗领域的应用现状与展望[J]. *医疗卫生装备*, 2024, 45(8): 102-9.
- [34] Frantar E, Ashkboos S, Hoefler T, et al. GPTQ: Accurate post-training quantization for generative pre-trained transformers[C]//*International Conference on Learning Representations (ICLR)*. 2023.
- [35] Lin J, Tang J, Tang H, et al. AWQ: Activation-aware weight quantization for LLM compression and acceleration[C]//*Proceedings of Machine Learning and Systems (MLSys)*. 2024.
- [36] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//*Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2017: 1273-82.

(编辑: 经媛)