

基于轻量化 Transformer 的实时医学影像分割系统设计

刘浩悦

曲阜师范大学计算机学院 山东日照 276800

摘 要: 本文研究轻量化 Transformer 在医学影像分割领域的系统设计, 针对标准 Transformer 模型参数繁多、计算复杂度高的问题, 提出了稀疏注意力机制、模块化设计及参数共享与剪枝等轻量化改进方案。该系统在实时手术导航和远程医疗等临床场景中展现出广阔应用前景, 能够在资源受限设备上高效运行, 为精准医疗和普惠医疗的发展提供技术支持。

关键词: 轻量化 Transformer; 模型压缩; 实时推理; 边缘计算

引言

伴随深度学习技术持续突破, 医学影像分割承担起计算机辅助诊断的关键职责, 临床实践日益倚重这一手段, 传统医学影像分割方法主要依托卷积神经网络 (CNN), 曾取得一定成绩, 但在面对复杂解剖结构处理和捕捉长距离依赖时仍显不足, Transformer 依托强大全局建模能力, 近年来在医学影像分割领域展现亮眼表现, 可更精准区分器官边界与病变区域, 标准 Transformer 模型参数繁多且计算复杂度高。在临床实时诊断方面难以满足要求, 尤其在资源紧张的边缘设备上部署更显困难, 轻量化 Transformer 研究应势而生, 旨在保持模型性能同时显著降低计算负荷和资源消耗。

1 轻量化 Transformer 的理论基础

1.1 Transformer 的基本原理

Transformer 模型由 Vaswani 等人在 2017 年提出, 用于处理自然语言任务, 模型采用自注意力机制捕捉序列中任意位置元素间的依赖, Transformer 在医学影像分割中通过把图像划分成一系列 patch 并将 patch 当作序列进行输入, 利用自注意力机制构建像素之间的全局关联, Transformer 架构由编码器和解码器构成, 编码器堆叠多层自注意力模块和前馈神经网络, 每层搭配残差连接和层归一化, 自注意力机制通过计算查询、键和值之间的相似度生成加权上下文表示。在医学影像分割应用中 Transformer 展现独特优势, 全局感知能力帮助它捕获解剖结构间长距离依赖, 这一能力促使它准确识别复杂器官边界和病变区域, 自注意力机制使其自动锁定图像中的关键部位, 从而提升了细微病变的识别精度。

1.2 轻量化设计的理论支撑

轻量化 Transformer 设计的理论基础源自模型压缩与高

效注意力计算, 模型压缩采用剪枝技术去除网络中冗余参数, 剪枝技术可细分为结构化剪枝与非结构化剪枝, 结构化剪枝通过删除整个神经元或注意力头保持网络结构规整, 从而便于硬件加速, 非结构化剪枝针对各单个权重实施选择性删除以实现更高压缩率, 量化技术通过降低模型权重和激活值的精度将 32 位浮点数转换至 8 位甚至 1 至 2 位低精度表示以有效减少存储需求与计算量, 预训练的大型教师模型指导小型学生模型训练实现知识蒸馏, 小型学生模型因此获得更丰富特征表示即便参数规模大幅缩减仍能保持较高性能。

2 轻量化 Transformer 医学影像分割系统的设计

2.1 系统架构设计原则

设计轻量化 Transformer 医学影像分割系统时须遵循实时性、准确性和鲁棒性三大原则, 临床应用要求实时性, 系统必须加速模型推理满足临床工作流程, 特别在手术导航场景中要求帧率达到 30fps 以上保证视觉反馈流畅, 模型结构设计时需充分考虑并优化计算效率, 设计者需要平衡模型深度与宽度并采用高效的注意力计算方法, 医学影像分割注重准确性, 设计过程中仍要确保分割精度符合临床诊断标准, 压缩模型规模时必须保留关键特征提取能力, 特别在提取医学影像中细微病变区域和复杂解剖结构边界时要精准识别。

2.2 轻量化 Transformer 的改进方法

在医学影像分割任务中, 本研究针对标准 Transformer 模型计算复杂度高、参数量大的问题提出了一系列轻量化改进方案, 包括稀疏注意力机制、模块化设计以及参数共享和权重剪枝, 稀疏注意力机制作为降低 Transformer 计算复杂度的核心策略采用基于局部窗口的注意力计算方法将图像划分成多个重叠窗口, 每个像素仅在其所在窗口内计算注意

力关系以有效降低计算量，多尺度窗口注意力机制在不同层次运用不同窗口尺寸并在关键层引入跨窗口信息交换模块以保证全局信息传递。

2.3 医学影像分割流程

轻量化 Transformer 医学影像分割系统沿序执行数据预处理、模型推理和后处理三个关键环节，数据预处理阶段对输入医学影像实施标准化，包括执行强度归一化、将像素值缩放至 [-1, 1] 或 [0, 1] 以降低不同设备和扫描参数引起的差异，随后采用自适应直方图均衡化提升图像对比度以增强细微结构的可视性，接着利用非局部均值滤波等算法在保留边缘细节时抑制噪声，然后根据分割目标调整窗宽窗位以突出感兴趣区域，训练阶段同时引入随机旋转、缩放、翻转等数据增强技术扩充训练集以提升模型泛化能力。

在模型推理阶段，经过预处理的医学影像被分割成一系列互相重叠的 patch 后送入轻量化 Transformer 网络中，该网络首先使用浅层轻量卷积模块抽取局部特征，接着借助深层轻量 Transformer 模块捕捉全局依赖，最终经解码器生成分割掩码，系统为提升推理效率采用滑动窗口策略处理大尺寸医学影像，并利用模型量化及批处理技术加速计算，后处理阶段聚焦优化模型输出的原始分割结果，包括利用条件随机场 (CRF) 细化分割边界，采用形态学操作如开闭运算去除孤立噪点并填充小孔洞。通过连通域分析排除假阳性区域，同时依据解剖学先验进行结构修正以确保分割结果符合医学实际，整个流程设计既重视计算效率又兼顾临床实用性，使系统能够在资源受限环境中实现准确且实时的医学影像分割服务。

表一： 轻量化 Transformer 改进方法与性能对比

改进方法	具体技术	性能提升	应用优势
稀疏注意力机制	局部窗口注意力计算 多尺度窗口机制 动态稀疏注意力	计算复杂度从 $O(n^2)$ 降至 $O(n \cdot \log(n))$ 甚至 $O(n)$	保持全局信息感知能力显著降低计算量
模块化设计	浅层轻量级卷积 深层轻量化 Transformer 特征融合模块	提高局部特征提取效率 保留全局建模能力	兼顾局部与全局特征 适应医学影像特点
参数共享与剪枝	多头注意力参数共享 结构化剪枝 低秩分解	模型大小减少 75% 推理速度提升 3-4 倍	减少冗余计算 降低存储需求

3 系统性能评估与优化

3.1 模型性能指标

轻量化 Transformer 医学影像分割系统的性能评估需从分割精度、推理速度和资源占用三个维度展开全面考察，在分割精度方面本文选用多种指标进行量化分析，Dice 系数 (DSC) 用以计算预测分割与真实标注间的重叠情况，交并比 (IoU) 直接比较预测区域与真实区域的交集与并集，平均表面距离 (ASD) 和豪斯多夫距离 (HD) 则从几何角度检测分割边界的准确性，推理速度作为该轻量化模型的重要性能指标，主要通过每秒处理帧数 (FPS) 来衡量，单张图像的平均处理时间得以记录，从而反映实际应用中模型的响应速度。

3.2 实验设计与结果分析

为全面评估轻量化 Transformer 在医学影像分割任务中表现，本研究设计了对比实验，实验采用多个公开医学影像数据集进行验证，数据集包括脑肿瘤分割挑战赛数据集 (BraTS)、肺结节分析挑战赛数据集 (LUNA16)、多器官腹部 CT 数据集 (BTCV) 以及冠状动脉分割数据集 (CAMUS)，数据涉及 MRI、CT 和超声等多种成像模态，分割目标涵盖肿瘤、器官及血管，每个数据集按 7:1:2 比例划分为训练集、

验证集和测试集，划分过程中实现患者级别分离避免数据泄露。在对比实验中，本文将提出的轻量化 Transformer 模型与三种方法进行比较，包括传统 CNN 架构（例如 U-Net、ResUNet 和 DeepLab 系列）、标准 Transformer 架构（例如 UNETR、SwinUNETR 和 TransUNet）以及其他轻量化方法（例如 MobileNetV3 和 ShuffleNetV2），实验结果显示，轻量化 Transformer 在各数据集上均取得了令人满意的表现，在 BraTS 数据集中，轻量化 Transformer 达到平均 Dice 系数 0.85，数值比标准 Transformer 低 0.02，模型体积减少 75%，推理速度提升 3.5 倍，在 LUNA16 数据集中，轻量化模型实现 IoU 0.92，数值与最佳 CNN 模型相当，FPS 达到 35，满足实时处理要求，在 BTCV 多器官分割任务中，模型在肝脏、脾脏等大器官上表现突出，平均 Dice 系数达到 0.87，在 CAMUS 超声数据集中，轻量化 Transformer 抵抗噪声和伪影效果良好，边界定位精度优于传统 CNN 方法，实验结果证明该轻量化策略在显著降低计算复杂度的同时保持较高分割精度，为医学影像分割实时应用提供了一种可行解决方案。

3.3 优化方向

基于实验结果和实际应用需求，研究团队识别出轻量

化 Transformer 医学影像分割系统的关键优化方向, 首要是在硬件加速领域针对各计算平台特性进行调整, 在 GPU 平台上工程师实现了基于 CUDA 的自定义算子以优化稀疏注意力计算并提升并行效率, 在 TPU 环境中技术人员重构了模型架构以契合张量计算特性并充分利用矩阵乘法加速单

元, 针对边缘 AI 芯片开发了专用模型转换工具以确保轻量化 Transformer 能够充分发挥硬件算力, 实验数据显示硬件优化措施推动模型推理速度平均提升 40% 并在保持相同精度的前提下显著增强系统实时性能。

表二：不同数据集上的轻量化 Transformer 性能表现

数据集	成像模态	分割目标	轻量化 Transformer 性能	对比标准 Transformer	对比传统 CNN
BraTS	MRI	脑肿瘤	Dice 系数: 0.85 FPS: >30	Dice 系数: 0.87 模型大小: +75% 速度: -70%	Dice 系数: 0.83 边界精度较低
LUNA16	CT	肺结节	IoU: 0.92 FPS: 35	IoU: 0.93 不满足实时要求	IoU: 0.91 FPS: 28
BTCV	CT	多器官	平均 Dice 系数: 0.87 处理时间: <30ms	平均 Dice 系数: 0.89 处理时间: >100ms	平均 Dice 系数: 0.85 处理时间: <25ms
CAMUS	超声	心脏结构	边界定位精度高 抗噪性能好	边界定位精度最高计算资源需求大	边界定位精度较低 计算效率高

4 结论

本研究扩展了轻量化 Transformer 在医学影像领域的应用, 提出了针对医学影像特点的多种优化策略, 构建了轻量化深度学习模型的理论基础, 实践中该系统支持实时手术导航和远程医疗诊断, 缓解医疗资源分配不均, 提升基层医疗机构的诊断能力, 未来研究将探索轻量化 Transformer 和其他深度学习方法 (如图卷积网络和扩散模型) 的结合以提升分割性能, 考察其在多模态医学影像分割中的表现以利用不同成像模态间的互补优势, 并着力优化在资源受限设备上的部署, 通过开发专用边缘 AI 加速器和算法扩展系统应用范围以助力精准医疗和普惠医疗的发展。

参考文献:

- [1] 李雪, 石晋雪, 王会青, 等. 基于 MCSP 和 Swin Transformer 的转录因子结合位点预测模型 [J/OL]. 华东理工大学学报 (自然科学版), 1-12[2025-05-16].
- [2] 王骁崑, 邢树礼, 毛国君. 多尺度特征融合的轻量化 Transformer 医学图像分割研究 [J]. 中国生物医学工程学报, 2025, 44(02): 165-173.

- [3] 崔志军. 医学影像分割技术: 从模糊到精准的跨越 [J].

健康生活, 2025, (04): 1-2.

- [4] 高翔, 王凡, 胡小鹏. 基于轻量化 Transformer 的高效图像超分辨率算法研究 [J]. 大连理工大学学报, 2025, 65(02): 212-220.

- [5] 陈广秋, 刘枫铭, 段锦, 等. 基于轻量化 Transformer 的车道线检测方法 [J]. 华中科技大学学报 (自然科学版), 2025, 53(03): 117-126.

- [6] 李永辉, 赵耀, 加小红, 等. CNN 与 Transformer 协同的多模态边缘检测网络 [J/OL]. 计算机工程与应用, 1-12[2025-05-16].

- [7] 连家诚, 郝一帆, 张曦珊, 等. Lite-IJformer: 面向长序列 Transformer 的轻量化方法 [J]. 高技术通讯, 2025, 35(02): 167-174.

- [8] 赵小明, 石培炼, 王丹丹, 等. 分布感知均值教师网络的半监督医学影像分割 [J]. 中国图象图形学报, 2025, 30(02): 575-588.