文章编号:1671-6833(2024)02-0027-06

基于卷积和可变形注意力的脑胶质瘤图像分割

高宇飞[1,2], 马自行[1], 徐 静[3], 赵国桦[4], 石 磊[1,2]

(1.郑州大学 网络空间安全学院, 河南 郑州 450002;2.嵩山实验室, 河南 郑州 450052;3.郑州大学 计算机与人工智能学院, 河南 郑州 450001;4.郑州大学第一附属医院, 河南 郑州 450003)

摘 要:对于脑胶质瘤图像分割这类密集预测的医学影像分割任务,局部和全局依赖关系都是不可或缺的,针对卷积神经网络缺乏建立全局依赖关系的能力,且自注意力机制在局部细节上捕捉能力不足等问题,提出了基于卷积和可变形注意力的脑胶质瘤图像分割方法。设计了卷积和可变形注意力 Transformer 的串行组合模块,其中卷积用于提取局部特征,紧随其后的可变形注意力 Transformer 用于捕捉全局依赖关系,建立不同分辨率下局部和全局依赖关系。作为一种 CNN-Transformer 混合架构,所提方法不需要任何预训练即可实现精准的脑胶质瘤图像分割。实验结果表明:所提方法在 BraTS2020 脑胶质图像分割数据集上平均 Dice 系数和平均 95%豪斯多夫距离分别为 83.56%和 11.30 mm,达到了与其他脑胶质瘤图像分割方法相当的分割精度,同时降低了至少 50%的计算开销,有效提升了脑胶质瘤图像分割的效率。

关键词:深度学习; 脑胶质瘤图像分割; 卷积神经网络; Transformer; 自注意力机制

中图分类号:O244;TP391.41 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2023.05.007

脑胶质瘤是一种具有高发病率和高致死率的原发性脑肿瘤,对人体的健康造成极大的危害。对脑胶质瘤核磁共振图像的分割可以帮助医生观察和分析脑胶质瘤的外部形态,从而进行诊断治疗。近年来,随着深度学习的发展,以 U-Net[1] 为主的全卷积神经网络在医学影像分割任务中占据主导地位。U-Net 通过构建具有跳跃连接的非对称编码器-解码器结构,达到了良好的医学影像分割效果。其中,编码器由一系列卷积层和下采样层组成,用于提取深层次特征;解码器将深层次特征逐步进行上采样,另外,解码器通过跳跃连接与编码器不同尺寸的特征进行融合,以补充下采样过程中带来的空间信息丢失。此后,基于 U-Net 的一系列网络也在医学图像分割领域得到应用,如 Res-UNet[2]、V-Net[3] 等。

然而,缺乏长距离依赖关系捕捉能力使得卷积神经网络并不能满足工业界的分割精度要求。尽管一些工作采用了空洞卷积[4-5] 来克服这一缺陷,但仍然存在局限性。近年来,Transformer[6] 在计算机视觉领域取得突破性成就,开始被应用于医学影像分割领域,如 TransUNet[7]、AA-TransUNet[8]、TransBTS[9] 等。TransUNet[7] 首次探索了 Transformer 在医学图像分割领域的可行性,其总体架构遵循 U-Net 的设计,利用 Transformer 将来自卷积神经网络的特征图编码为提取全局上下文的输入序列。同样,TransBTS[9] 将 Transformer 用于编码器末端进行全局信息建模,实现了良好的脑胶质瘤图像分割效果。但以上方法并未考虑大尺寸特征图下的长距离依赖关系。另外,其采用的 Transformer 自注意力机制耗费内存且计算量大。

为了解决自注意力机制内存和计算量消耗过大的问题,研究者们设计了不同的稀疏自注意力机制。PVT[10] 通过空间减少注意力(spatial-reduction attention)降低计算量,Swin Transformer[11] 则是限制在一个窗口中计算自注意力,以此降低计算量。很快,这 2 种方法也被应用于医学影像分割任务中,如 Swin-UNet[12]、UTNet[13] 等。但是,这 2 种方法中自注意

收稿日期:2023-04-18;**修订日期**:2023-06-19

基金项目:国家自然科学基金资助项目(62006210);河南省重大公益专项(201300210500);郑州大学高层次人才科研启动基金(32340306)

通信作者:石磊(1967—),男,河南郑州人,郑州大学教授,博士,博士生导师,主要从事云计算与大数据、网络与分布式计算、服务计算、人工智能等方面的研究,E-mail:shilei@zzu.edu.cn。

引用本文:高宇飞,马自行,徐静,等. 基于卷积和可变形注意力的脑胶质瘤图像分割[J]. 郑州大学学报(工学版),2024,45(2):27-32. (GAO Y F, MA Z X, XU J, et al. Brain glioma image segmentation based on convolution and deformable attention[J]. Journal of Zhengzhou University(Engineering Science),2024,45(2):27-32.)

力机制的设计可能会丢失关键信息,限制自注意力机制建立长距离关系依赖的能力。最近,具有可变形注意力的 Transformer[14] 通过设计一种名为可变形注意力的稀疏自注意力机制来缓解这一缺陷,并取得了更好的效果。

此外,Transformer 的自注意力机制缺乏局部上下文信息提取能力,为了解决上述问题,CoAtNet[15]、CMT[16] 等将卷积引入 Transformer 模型中,增强视觉 Transformer 的局部性,从而获得了更优的性能,这也验证了 CNN 与 Transformer 混合方法的有效性。

受上述研究的启发,本文提出一种基于 CNN-Transformer 混合的脑胶质瘤图像分割方法(Med-CaDA)。不同于 TransUNet、TransBTS 仅仅将 Transformer 应用于小尺寸特征图,本文采用了稀疏自注意力机制,并将其应用于各个尺寸特征图中提取全局上下文信息,建立不同分辨率下局部和全局的依赖关系。此外,将卷积的瓶颈残差模块和可变形注意力 Transformer 组成串行模块,命名为 CaDA 块,该模块既保留了卷积局部上下文提取的优势,又借助了 Transformer 全局信息捕捉的能力。

1 基于卷积和可变形注意力的脑胶质瘤图像分割方法

图 1 为 Med-CaDA 整体架构和 CaDA 块。Med-CaDA 的整体架构如图 1(a)所示,其遵循 U-Net 的编码器-解码器架构设计,由 6 个部分组成:嵌入层、CaDA 块、下采样层、上采样层、扩展层和跳跃连接。输入为 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times D\times K}$,其中 H、W、D 和 K 分别表示空间分辨率的高度和宽度、切片深度和模态数量。

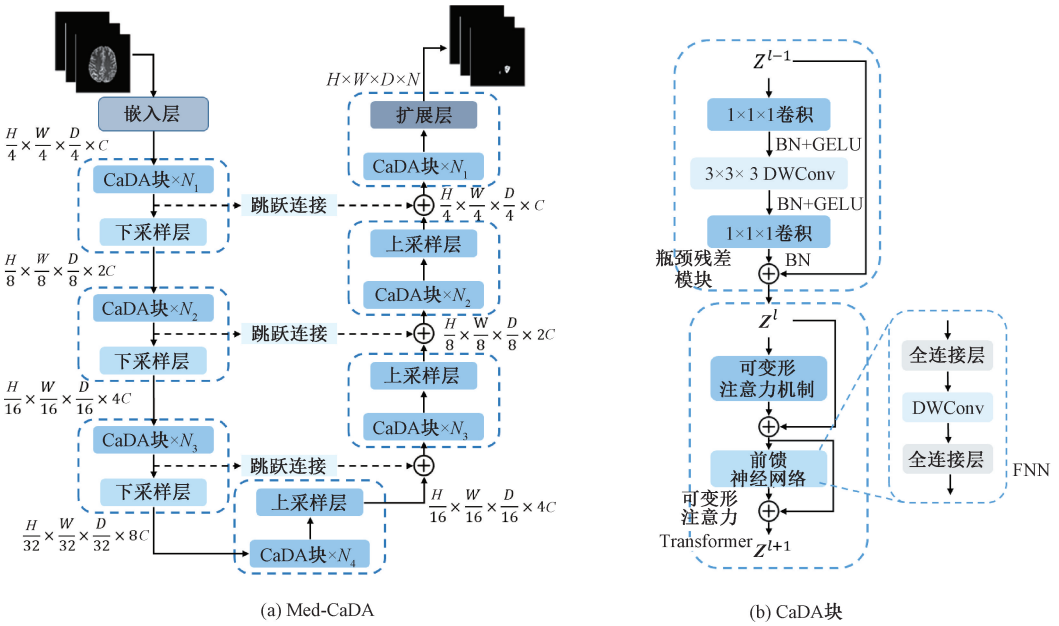

图 1 Med-CaDA 整体架构和 CaDA 块

Figure 1 Overall architecture of Med-CaDA and CaDA block

1.1 嵌入层和扩展层

嵌入层和扩展层如图 2(a)、2(c)所示,图中 GELU 表示激活函数,LN 表示 LayerNorm 标准化。嵌入层由 2 个步长为 2 的 3×3×3 卷积组成,逐步将输入图像 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times D\times K}$ 映射为特征向量 $\boldsymbol{X}_e\in\mathbf{R}^{\frac{H}{4}\times\frac{W}{4}\times\frac{D}{4}\times C}$,用于更精确地编码图像的像素级空间信息。

与嵌入层相对应,扩展层是一个步长为 4 的 4×4×4 卷积,负责将高维度张量还原回输入图像尺寸 $\boldsymbol{Y}\in\mathbf{R}^{H\times W\times D\times N}$,其中 N 表示分割类别数量。

1.2 下采样层和上采样层

下采样层和上采样层是构建编码器-解码器分层架构的关键,如图 2(b)、2(d)所示,下采样层采用步长为 2 的 2×2×2 卷积,逐步将特征图编码为尺

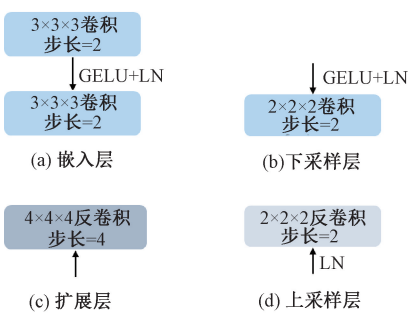

图 2 嵌入层、扩展层、下采样层和上采样层示意图

Figure 2 Schematic diagram of embedding, expanding, down-sampling and up-sampling layer

寸更小、维度更高的深层特征。上采样层与下采样层对应,是一个步长为 2 的 2×2×2 反卷积,用于将深层特征图尺寸加倍,维度减半。

1.3　CaDA 块

卷积和自注意力机制分别擅长局部上下文信息提取和长距离依赖捕捉,两者对于脑胶质瘤图像分割这类密集预测任务都至关重要。因此,本文设计了由卷积的瓶颈残差模块和可变形注意力 Transformer 串行组成的 CaDA 块,如图 1(b)所示,图中 BN、GELU、DWConv 分别表示 BatchNorm 标准化、激活函数和深度可分离卷积。

(1)瓶颈残差模块。图 1(b)上半部分是瓶颈残差模块,依次由 1×1×1 卷积、3×3×3 深度可分离卷积、1×1×1 卷积构成,采用深度可分离卷积可大幅度降低计算量和参数量。另外,不同于 MobileNetV2[17] 中 2 个 1×1×1 卷积用于先升维后降维,本文则先降维再升维,从而在计算 3×3×3 卷积时进一步降低计算量。计算过程可以表述为

$$\boldsymbol{Z}^{l}=\mathrm{Bottleneck}(\boldsymbol{Z}^{l-1})+\boldsymbol{Z}^{l-1};\tag{1}$$

$$\mathrm{Bottleneck}(\boldsymbol{X})=\mathrm{Conv}(\mathrm{DWConv}(\mathrm{Conv}(\boldsymbol{X})))。\tag{2}$$

式中:Conv(·)和 DWConv(·)分别表示 1×1×1 卷积和深度可分离卷积。

(2)可变形注意力 Transformer。图 1(b)下半部分显示的是可变形注意力 Transformer,由可变形注意力机制、前馈神经网络 FNN 和残差连接组成。受 DAT[14] 启发,本文实现了三维数据下的可变形注意力机制,如图 3 所示。可变形注意力机制的键值向量和值向量是在原图上采样特征投影得到的,这些采样特征由查询向量通过一个偏置学习网络学习的采样点经过双线性插值得到。具体实现过程如下。

假设给定输入 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times D\times C}$,首先对输入投影产生查询向量 $\boldsymbol{Q}$,同时根据输入 $\boldsymbol{X}$ 的尺寸生成一个大小为 $\boldsymbol{P}\in\mathbf{R}^{\frac{H}{F}\times\frac{W}{F}\times\frac{D}{F}\times 3}$ 的均匀网格点(F 为预定义参数),称之为参考点,参考点的值是线性间隔的三维坐标 $\left\{(0,0,0),\cdots,\left(\frac{H}{F}-1,\frac{W}{F}-1,\frac{D}{F}-1\right)\right\}$。然后对

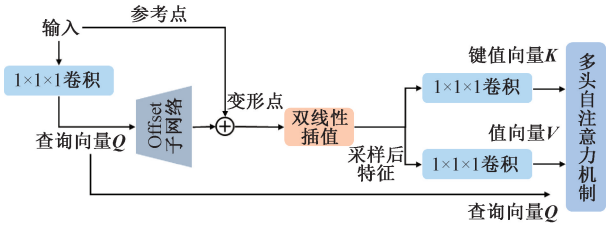

图 3　可变形注意力机制示意图

Figure 3　Schematic diagram of deformable attention

参考点的值归一化到[-1,1],那么三维坐标将变成{(-1,-1,-1),…,(1,1,1)}。随后,将查询向量 $\boldsymbol{Q}$ 输入到偏置学习子网络 Offset(·)以生成每个参考点的偏移量 $\Delta\boldsymbol{P}$(为了防止偏移量过大,采用可变参数 s 和 tanh(·)去控制),然后通过参考点和偏移量 $\Delta\boldsymbol{P}$ 相加得到变形点,根据变形点所对应的坐标位置在输入上采用双线性插值采样得到采样后的特征图 $\boldsymbol{X}_{\mathrm{z}}$,随后投影产生键值向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。最后,对得到的 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 计算多头自注意力输出结果。可变形多头注意力机制(deformable multi-head attention,DMHA)计算公式为

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_{\mathrm{q}};\tag{3}$$

$$\Delta\boldsymbol{P}=s\cdot\tanh(\mathrm{Offset}(\boldsymbol{Q}));\tag{4}$$

$$\boldsymbol{X}_{\mathrm{z}}=\mathrm{BI}(\boldsymbol{X},\boldsymbol{P}+\Delta\boldsymbol{P});\tag{5}$$

$$\boldsymbol{K}=\boldsymbol{X}_{\mathrm{z}}\boldsymbol{W}_{\mathrm{k}};\tag{6}$$

$$\boldsymbol{V}=\boldsymbol{X}_{\mathrm{z}}\boldsymbol{W}_{\mathrm{v}};\tag{7}$$

$$\boldsymbol{Z}^{(m)}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}^{(m)}\boldsymbol{K}^{(m)\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}^{(m)},m=1,2,\cdots,h;\tag{8}$$

$$\mathrm{DMHA}(\boldsymbol{Z})=\mathrm{Concat}(\boldsymbol{Z}^{(1)},\boldsymbol{Z}^{(2)},\cdots,\boldsymbol{Z}^{(m)})\boldsymbol{W}_{\mathrm{o}}。\tag{9}$$

式中:Offset(·)是依次由 1×1×1 卷积、深度可分离卷积和 1×1×1 卷积组成的子网络;$\Delta\boldsymbol{P}\in\mathbf{R}^{\frac{H}{F}\times\frac{W}{F}\times\frac{D}{F}\times 3}$ 表示参考点的偏移量;BI(·)表示双线性插值;$\boldsymbol{W}_{\mathrm{q}}\in\boldsymbol{R}^{C\times C}$、$\boldsymbol{W}_{\mathrm{k}}\in\mathbf{R}^{C\times C}$、$\boldsymbol{W}_{\mathrm{v}}\in\mathbf{R}^{C\times C}$ 分别表示查询向量 $\boldsymbol{Q}\in\boldsymbol{R}^{N\times C}$、键值向量 $\boldsymbol{K}\in\mathbf{R}^{M\times C}$ 和值向量 $\boldsymbol{V}\in\mathbf{R}^{M\times C}$ 的投影矩阵,其中 $N=H\times W\times D$,$M=H/F\times W/F\times D/F$;$\boldsymbol{Q}^{(m)}\in\boldsymbol{R}^{N\times d}$,$\boldsymbol{K}^{(m)}$、$\boldsymbol{V}^{(m)}\in\mathbf{R}^{M\times d}$ 分别代表第 m 个头部下的 3 个向量;$d=\frac{C}{h}$ 代表每一个头部的维度,h 为头部数量;$\boldsymbol{Z}^{(m)}\in\mathbf{R}^{N\times d}$ 为第 m 个头部的注意力输出;$\boldsymbol{W}_{\mathrm{o}}\in\mathbf{R}^{C\times C}$ 是可变形多头注意力机制的输出投影矩阵。

另外,前馈神经网络(feedforward neural network,FNN)由 2 个全连接层和 1 个深度可分离卷积组成,如图 1(b)所示。将深度可分离卷积引入前馈网络中可以为 Transformer 模块增加局部性。

最终,结合可变形多头注意力、前馈神经网络和残差连接,可变形注意力 Transformer 计算公式可以表示为

$$\widehat{\boldsymbol{Z}^{l}}=\mathrm{DMHA}(\mathrm{LN}(\boldsymbol{Z}^{l}))+\boldsymbol{Z}^{l};\tag{10}$$

$$\boldsymbol{Z}^{l+1}=\mathrm{FNN}(\mathrm{LN}(\widehat{\boldsymbol{Z}^{l}}))+\widehat{\boldsymbol{Z}^{l}}。\tag{11}$$

式中:DMHA(·)表示可变形多头注意力机制;LN(·)表示 LayerNorm 标准化;FNN(·)表示前馈神经网络。

2 实验与分析

2.1 数据集构成

本文采用 BraTS2020 脑胶质瘤图像分割数据集,训练集和验证集分别由 369 个和 125 个 3 维 MRI 组成,每个 MRI 包括 4 种模态:T1、T1ce、Flair 和 T2。需要分割的 3 个类别分别为整个肿瘤区域(whole tumor,WT)、肿瘤核心区域(tumor core,TC)以及活动肿瘤区域(enhance tumor,ET)。本文在训练集上进行训练,在验证集上进行测试,并将验证集上的预测标签上传到 BraTS2020 比赛官网以得到分割结果。

2.2 实验环境和参数设置

实验采用的编程语言为 Python 3.8,深度学习框架为 Pytorch 1.7.1,使用的显卡为 2 张 Tesla T4,显存一共为 32 GB。实验中,设置 Med-CaDA 模型不同阶段预定义参数如表 1 所示,预定义参数包括每阶段的分辨率、循环次数 N、通道数 C、可变形注意力的预定义参数 F 和 s、多头注意力机制的头数 h。在训练过程中,沿用了 TransBTS[9] 的随机裁剪、随机镜像翻转和随机强度偏移 3 种数据增强策略。采用 Adam 优化器,学习率设置为 1×10^{-4},*batch size* 为 4,训练轮数为 800。另外,采用余弦学习率衰减策略控制学习率大小,便于模型收敛;采用 L2 正则化去缓解模型过拟合问题(权重衰减设置为 1×10^{-5})。

表 1 不同阶段预定义参数设置

Table 1 Predefined parameters setting at different stages

阶段	分辨率	相关系数
阶段一	32×32×32	$N=1,C=96,F=4,s=-1,h=3$
阶段二	16×16×16	$N=1,C=192,F=2,s=-1,h=6$
阶段三	8×8×8	$N=2,C=384,F=1,s=2,h=12$
阶段四	4×4×4	$N=1,C=768,F=1,s=2,h=24$

2.3 评估指标

在脑胶质瘤图像分割实验中采用了浮点运算次数衡量模型的复杂度。采用 Dice 系数和 95%豪斯多夫距离 2 个评估指标衡量 2 个点集合间的相似程度,Dice 系数对集合内部填充比较敏感,95%豪斯多夫距离对边界比较敏感。可以表示为

$$\mathrm{Dice}(A,B)=\frac{2\mid A\cap B\mid}{\mid A\mid+\mid B\mid};\tag{12}$$

$$\mathrm{HD}(A,B)=\max(h(A,B),h(B,A));\tag{13}$$

$$h(A,B)=\max_{a\in A}\min_{b\in B}\parallel a-b\parallel;\tag{14}$$

$$h(B,A)=\max_{b\in B}\min_{a\in A}\parallel b-a\parallel。\tag{15}$$

式中:$|A\cap B|$ 表示 A、B 间交集的元素个数;$|A|$ 和 $|B|$ 分别表示 A 和 B 的元素个数;$h(A,B)$ 的实际意义为计算集合 B 到集合 A 每个点距离最近的距离并排序,然后选择距离中的最大值。

2.4 实验结果分析

2.4.1 推理实验

损失函数用来估量模型的预测值与真实值不一致的程度,损失函数越小,模型效果就越好。训练过程中损失函数变化曲线如图 4 所示。由图 4 可知,损失值随着训练轮数的增加逐步减少,逐渐趋近于 0,说明预测值越来越接近于真实值,模型的性能越来越好,进而说明将该模型用于 BraTS2020 数据集的分割是有效的。

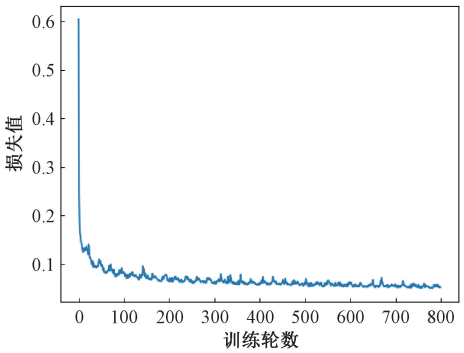

图 4 损失函数变化曲线

Figure 4 Curve of loss function

另外,为了验证所采用的可变形注意力机制的有效性,采用了几种不同的稀疏自注意力机制并进行实验,包括空间减少注意力[10](SRA)、窗口注意力[11](WA)以及可变形注意力(DA)。实验结果如表 2 所示。由表 2 可知,采用可变形注意力在 Dice 系数指标上完全优于其他的注意力机制,在 95%豪斯多夫距离指标上 ET 和 TC 的效果也优于其他注意力机制。由此表明,可变形注意力以数据依赖的方式选择键值向量和值向量,更有助于建立长距离依赖关系。

表 2 不同稀疏注意力机制在 BraTS2020 数据集中的结果

Table 2 Results of the BraTS2020 dataset under different sparse self-attention

注意力机制	Dice 系数/%			95%豪斯多夫距离/mm		
	ET	WT	TC	ET	WT	TC
SRA[10]	76.06	89.71	80.42	21.37	4.96	18.69
WA[11]	76.70	89.60	82.65	24.09	5.12	11.75
DA	77.87	90.05	82.76	19.08	5.81	9.03

2.4.2 对比实验

为了验证 Med-CaDA 在脑胶质瘤分割的有效性,在 BraTS2020 脑胶质瘤图像分割数据集上进行对比实验。选择了 3 种基于卷积的经典方法、2 种

基于卷积的先进方法以及 2 种引入 Transformer 的先进方法进行对比,分别为 3D U-Net[18]、V-Net[3]、3D Res-UNet[19]、MMTSN[20]、MDNet[21]、TransU-Net[7]及 TransBTS[9],对比实验结果如表 3 所示。

表 3　不同方法的对比实验结果

Table 3　Comparative experimental results of different methods

方法	复杂度/GFLOPs		Dice 系数/%				95%豪斯多夫距离/mm			
	单个样本	单个切片	ET	WT	TC	平均值	ET	WT	TC	平均值
3D U-Net[18]	1 669.53	13.04	68.76	84.11	79.06	77.31	50.98	13.37	13.61	25.99
V-Net[3]	749.29	5.85	61.79	84.63	75.26	73.89	47.7	20.41	12.18	26.76
3D Res-UNet[19]	407.37	3.18	71.63	82.46	76.47	76.85	37.42	12.34	13.11	20.96
MMTSN[20]	—	—	76.37	88.23	80.12	81.57	21.39	6.68	**6.49**	11.52
MDNet[21]	—	—	77.17	**90.55**	82.67	83.46	27.04	4.99	8.63	13.55
TransUNet[7]	1 205.76	9.42	78.42	89.46	78.37	82.08	**12.85**	5.97	12.84	**10.55**
TransBTS[9]	333.09	2.60	**78.73**	90.09	81.73	83.52	17.95	**4.96**	9.77	10.89
Med-CaDA	**103.72**	**0.81**	77.87	90.05	**82.76**	**83.56**	19.08	5.81	9.02	11.30

由表 3 可以看出,Med-CaDA 在 ET、WT、TC 3 个分割指标中,取得的 Dice 系数及其平均值分别为 77.87%、90.05%、82.76% 和 83.56%;95% 豪斯多夫距离及其平均值分别为 19.08、5.81、9.02 及 11.30 mm。与 3 个经典的方法相比,Med-CaDA 在 2 个评价指标上均有显著的提升。与 2 种基于卷积的先进方法和 2 种引入 Transformer 的先进方法相比,Med-CaDA 在 Dice 系数指标上虽然只在 TC 上超过其他方法,但是在平均水平上会高于其他方法。同时,Med-CaDA 的单个样本和单个切片下复杂度下降了 50%~90%。在 95%豪斯多夫距离上虽然未达到最佳,但在平均水平上与最好的方法仅仅相差 0.75 mm。所以,Med-CaDA 在保证脑胶质瘤图像分割精度的同时,大幅度提高了分割效率。

3　结论

本文提出了一种基于卷积和可变形注意力的脑胶质瘤图像分割方法,该方法继承了卷积建模局部上下文信息的优势,还利用了 Transformer 学习全局语义相关性,这种 CNN-Transformer 混合架构可以在没有任何预训练的情况下实现医学影像的精准分割。在 BraTS2020 数据集上的实验结果表明,与其他方法相比,本文提出的模型在保证分割精度的同时降低了至少 50%的计算开销,有效提升了脑胶质瘤图像的分割效率。所以,在脑胶质瘤图像这类医学影像分割任务中,采取稀疏的方法降低参数量同样可以达到良好的分割效果。

参考文献:

[1] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[2] XIAO X, LIAN S, LUO Z M, et al. Weighted Res-UNet for high-quality retina vessel segmentation[C]//2018 9th International Conference on Information Technology in Medicine and Education (ITME). Piscataway: IEEE, 2018: 327-331.

[3] MILLETARI F, NAVAB N, AHMADI S A. V-Net: fully convolutional neural networks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV). Piscataway: IEEE, 2016: 565-571.

[4] FENG S L, ZHAO H M, SHI F, et al. CPFNet: context pyramid fusion network for medical image segmentation [J]. IEEE Transactions on Medical Imaging, 2020, 39 (10): 3008-3018.

[5] XIA H Y, MA M J, LI H S, et al. MC-Net: multi-scale context-attention network for medical CT image segmentation[J]. Applied Intelligence, 2022, 52(2): 1508-1519.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[7] CHEN J N, LU Y Y, YU Q H, et al. AA-TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. (2021-02-08)[2023-01-12]. https://doi.org/10.48550/arXiv.2102.04306.

[8] YANG Y, MEHRKANOON S. AA-TransUNet: attention augmented transunet for nowcasting tasks[C]//2022 International Joint Conference on Neural Networks (IJCNN). Piscataway:IEEE, 2022: 1-8.

[9] WANG W X, CHEN C, DING M, et al. TransBTS: multimodal brain tumor segmentation using transformer [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Spring-

er, 2021: 109-119.
[10] WANG W H, XIE E Z, LI X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 548-558.
[11] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway:IEEE, 2022: 9992-10002.
[12] CAO H, WANG Y, CHEN J, et al. Swin-UNet: Unet-like pure transformer for medical image segmentation [EB/OL]. (2021-05-12)[2023-01-12]. https://doi.org/10.48550/arXiv.2105.05537.
[13] GAO Y, ZHOU M, METAXAS D N. UTNet: a hybrid transformer architecture for medical image segmentation [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021: 61-71.
[14] XIA Z F, PAN X R, SONG S J, et al. Vision Transformer with deformable attention[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE, 2022: 4784-4793.
[15] DAI Z H, LIU H X, LE Q V, et al. CoAtNet: Marrying convolution and attention for all data sizes[J]. Advances in Neural Information Processing Systems, 2021, 34: 3965-3977.
[16] GUO J Y, HAN K, WU H, et al. CMT: convolutional neural networks meet vision transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE, 2022: 12165-12175.
[17] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2018: 4510-4520.
[18] ÇIÇEK Ö, ABDULKADIR A, LIENKAMP S S, et al. 3D U-net: learning dense volumetric segmentation from sparse annotation[M]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016. Cham: Springer, 2016: 424-432.
[19] YU W, FANG B, LIU Y Q, et al. Liver vessels segmentation based on 3D residual U-NET[C]//2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2019: 250-254.
[20] LIU C Y, DING W B, LI L, et al. Brain tumor segmentation network using attention-based fusion and spatial relationship constraint[C]//International MICCAI Brainlesion Workshop. Cham:Springer, 2021: 219-229.
[21] VU M H, NYHOLM T, LÖFSTEDT T. Multi-decoder networks with multi-denoising inputs for tumor segmentation[C]//International MICCAI Brainlesion Workshop. Cham: Springer, 2021: 412-423.

Brain Glioma Image Segmentation Based on Convolution and Deformable Attention

GAO Yufei[1,2], MA Zixing[1], XU Jing[3], ZHAO Guohua[4], SHI Lei[1,2]

(1. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China; 2. Songshan Laboratory, Zhengzhou 450052, China; 3. School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China; 4. The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450003, China)

Abstract: For medical image segmentation tasks such as glioma image segmentation with dense prediction, both local and global dependencies were indispensable. To address the problems that convolutional neural networks lacked the ability to establish global dependencies and the self-attention mechanism had insufficient ability to capture local details, a convolutional and deformable attention-based method for glioma image segmentation was proposed. A serial combination module of convolution and deformable attention Transformer was designed, in which convolution was used to extract local features and the immediately following deformable attention. Transformer was used to capture global dependencies to the establishment of local and global dependencies at different resolutions. As a hybrid CNN-Transformer architecture, the method could achieve accurate brain glioma image segmentation without any pre-training. Experiments showed that the average dice score and the average 95% Hausdorff distance on the BraTS2020 glioma image segmentation dataset were 83.56% and 11.30 mm, respectively, achieving comparable segmentation accuracy compared with other methods, while reducing the computational overhead by at least 50% and effectively improving the efficiency of glioma image segmentation.

Keywords: deep learning; brain glioma image segmentation; CNN; Transformer; self-attention mechanism