

# Medical Transformer: Gated Axial-Attention for Medical Image Segmentation

摘要

过去十年研究发现,卷积体系结构缺乏对图像中的远程依赖关系的理解。基Transformer的体系结构利用自我注意机制,编码长期依赖关系,并具有极富表现力的表示。作者提出了一种门控轴向注意模型,通过在自我注意模块中引入额外的控制机制来扩展现有的体系结构,减少了计算复杂度。为了对模型进行有效的医学图像训练,又提出了一种局部 - 全局训练策略 (LOGO),进一步提高了模型的性能。具体地说,即整个图像和 patches 进行操作,分别学习全局特征和局部特征。

关键词: transformer; Medical Image Segmentation; Gated Axial-Attention;

## 1 引言

目前用卷积神经网络来做图像分割虽然也能取得一些好效果,然而卷积神经网络无法对长依赖进行建模。而transformer以能够建模长依赖著称,在大量数据集训练的条件下取得了比卷积神经网络更好的准确率。

文章提出了用transformer来做医学图像分割。要解决的问题是, transformer在图像任务上相比卷积神经网络需要更大的数据集来训练,计算量巨大,而医学图像处理的一个难题就是数据不足,数据集不够大。

文章主要贡献是两点,一是提出了一种适用于较小数据集的门控位置敏感轴向注意机制,一个是引入了有效提高 Transformer 性能,的局部-全局 (LOGO) 训练方法。

## 2 相关工作

### 2.1 CNN

随着深度卷积神经网络 (ConvNets) 在计算机视觉中的普及, ConvNets被迅速应用于医学图像分割。像U-Net、V-Net、3D U-Net、Res-UNet、Dense-UNet、YNet、U-Net++、KiU-Net和U-Net3+这样的网络被专门用于对各种医学成像模式执行图像和体积分割。这些方法在许多困难的数据集上取得了令人印象深刻的性能,证明了ConvNets在从医学扫描中学习区分器官或病变特征方面的有效性。

convnet是目前提出的大多数图像分割方法的基本构造块。但是,它们缺乏对映像中存在的长期依赖关系建模的能力。更准确地说,在ConvNets中,每个卷积核只关注整个图像中的局部像素子集,并迫使网络关注局部模式,而不是全局上下文。已经有一些工作专注于使用图像金字塔、空洞卷积和注意机制对convnet的远程依赖性进行建模。然而,可以注意到,对于建模长期依赖关系,仍然有改进的余地。

### 2.2 Transformer

在许多自然语言处理（NLP）应用程序中，Transformer已经证明能够对长程依赖项进行编码。这是由于自我注意机制发现了给定顺序输入之间的依赖性。随着Transformer在NLP应用中的普及，最近Transformer已被应用于计算机视觉应用。关于分割任务的Transformer，Axial Deeplab使用了轴向注意模块，该模块将2D自我注意分解为两个1D自我注意，并引入了位置敏感轴向注意分段设计。在Segmentation Transformer（SETR）中，使用变压器作为编码器，输入一系列图像块，并使用ConvNet作为解码器，从而形成强大的分割模型。

### 3 本文方法

#### 3.1 本文方法概述

如图1所示，MedT两个分支机构：一个Global分支和一个Local分支。这两分支的输入是从初始卷积块提取的特征图。在MedT的全局分支中，我们有2个编码器块和2个解码器块。在本地分支中，我们有5个编码器块和5个解码器块。经残缺采样后输出分割掩码。

encoder部分如图2所示，在传统transformer的结构上将多头注意机制改成。两个轴向（纵横两个方向的）注意力机制，然后在此基础增加门控单元控制位置编码的影响，如图3所示。

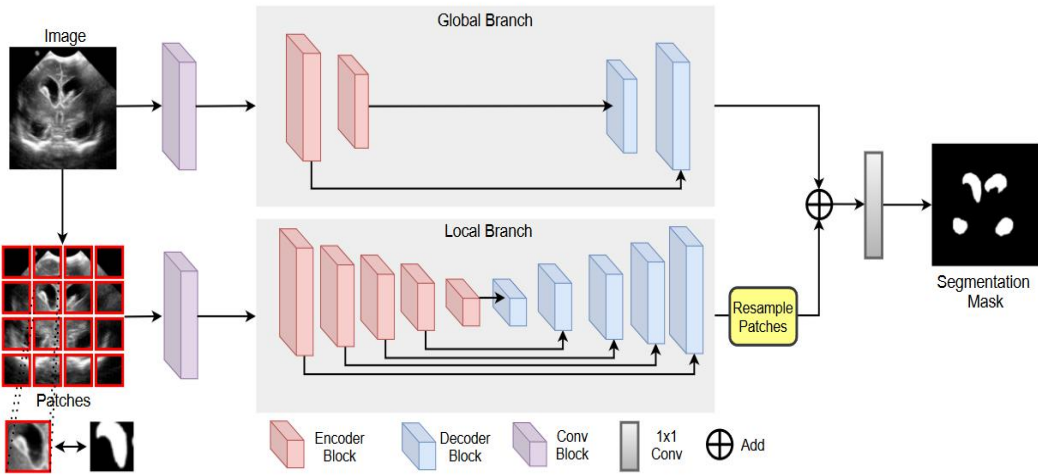


图 1 采用LoGo策略进行训练的MedT主要架构图

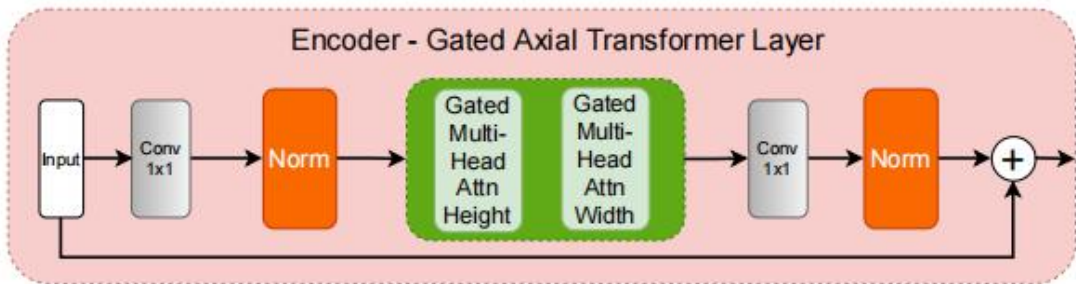


图 2 门控轴向transformer层

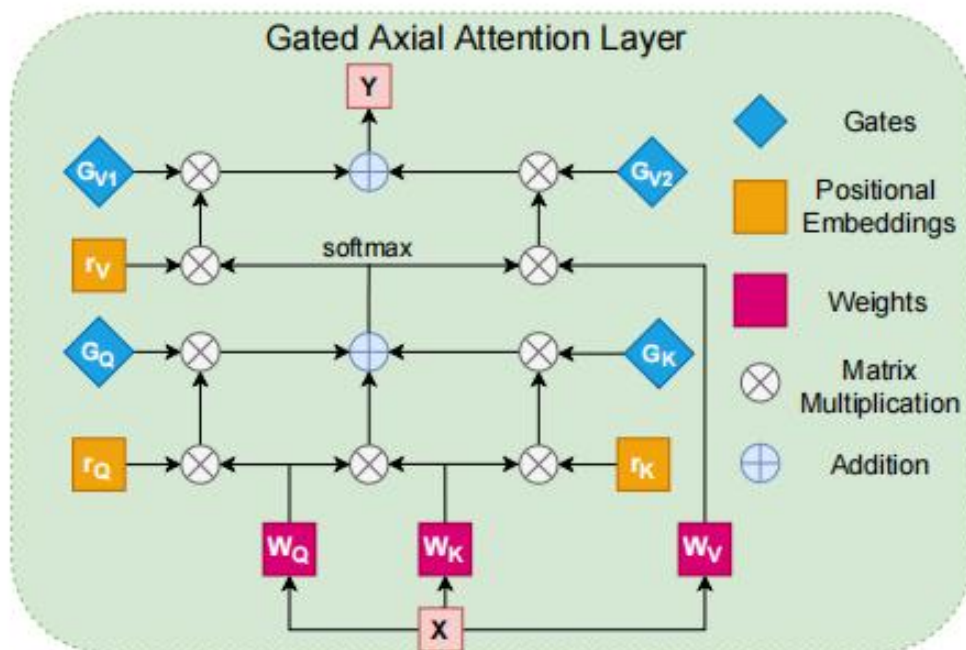


图 3 门控轴向注意层的基本组成部分

### 3.2 Gated Axial-Attention

ViT提出的时候，transformer的每个token会对所有的每个token都计算注意力，所以是 $(hw)^2$ 次计算，这样复杂度较高。Self-Attention公式：

$$y_{ij} = \sum_{h=1}^H \sum_{w=1}^W \text{softmax} (q_{ij}^T k_{hw}) v_{hw}$$

为了解决这种情况，将传统的自注意力模块分为宽度上以及高度上的两个注意力模块，称为 axial-attention，大大减小了计算复杂度。

还有，transformer其实是不具有位置表达能力的，为了添加位置表达能力，需要加一个 position embedding，就是用一个onehot的位置向量，经过一个全连接的embedding，产生位置编码，这个全连接是可训练的。

加上轴向注意力和多个位置编码的trick后现在变成这样（w方向的注意力，h方向和这个类似）：

$$y_{ij} = \sum_{w=1}^W \text{softmax} \left( \underbrace{q_{ij}^T k_{iw}}_{\text{横轴}} + q_{ij}^T \underbrace{r_{iw}^q}_{\text{相对位置编码}} + k_{iw}^T \underbrace{r_{iw}^k}_{\text{相对位置编码}} \right) (v_{iw} + \underbrace{r_{iw}^v}_{\text{相对位置编码}})$$

然而上述的trick需要大量数据集进行训练，小量的数据不足以训练QKV的三个position embedding，而医学数据集多数情况下就是少量的。在这种情况下，不准确的position embedding会给网络准确率带来负面影响，为此文章提出了门控单元用来控制这个影响的程度，修改上述公式如下：

$$y_{ij} = \sum_{w=1}^W \text{softmax} (q_{ij}^T k_{iw} + G_Q q_{ij}^T r_{iw}^q + G_K k_{iw}^T r_{iw}^k) (G_{V1} v_{iw} + G_{V2} r_{iw}^v)$$

这里三个G都是可学习的参数，当数据集不足以使得网络预测准确的position embedding时，网络的G会小一点，反之会大一点，因此起到一个所谓的“Gated”的作用。这是文章第一个贡献点。

### 3.3 Local-Global Training

transformer做图像分割可以用patch-wise的方式去做，也就是说把一张完整图片切割成多个patch，每个patch和这个patch对应的mask作为一个样本，用来训练transformer，这样又快，然而问题在于，一张图片的一个病灶可能比一个patch大，这样的话这个patch看起来就会很奇怪，因为被病灶充满了。

于是文章关于两个branch的做法是：global branch不做特殊处理，就是整张完整特征图，进行两次transformer block后送进decoder；而local branch切分成4x4个patch，每个patch单独送transformer block 前向传播，patch和patch之间没有任何联系，最后再把这4x4个patch的feature map拼回去。用深的网络来处理local信息，用浅的网络来处理global信息。这就是文章的第二个贡献点。

## 4 复现细节

### 4.1 与已有开源代码对比

在原有的开源代码基础上添加了对不同数据集像素大小输入的代码，探究了不同像素大小对实验结果的影响。512×512的数据集像素大小由于硬件资源有限，运行不了。同时需要注意，GLAS 腺体分割(显微)数据集是彩色图像需要经过二值处理得到灰度图。

### 4.2 实验环境搭建

#### 4.2.1 依赖安装

```
pip install -r
torch>=1.4.0
torchvision>=0.5.0
scikit-learn==0.23.2
scipy==1.5.3
```

#### 4.2.2 数据集格式准备

Train Folder-----

img----

0001.png

0002.png

.....

labelcol---

0001.png

0002.png

.....

4.2.3 训练/测试指令

```
python train.py --train_dataset "enter train directory" --val_dataset "enter validation
directory" --direc 'path for results to be saved' --batch_size 4 --epoch 400 --save_freq 10
--modelname "gatedaxialunet" --learning_rate 0.001 --imgsize 128 --gray "no"
```

```
python test.py --loaddirec "./saved_model_path/model_name.pth" --val_dataset "test
dataset directory" --direc 'path for results to be saved' --batch_size 1 --modelname
"gatedaxialunet" --imgsize 128 --gray "no"
```

5 实验结果分析

5.1 数据集

原论文实验所用数据集如表1所示，其中Na代表未知。有此可见，医学图像数据集的数量
的确较小，只有一百多甚至连一百也没有。由于Brain US 未公开，所以复现只复现了后
两个公开数据集的结果，然后在MoNuSeg数据集是实验了不同像素大小的结果。

表 1 MedT所使用的数据集

数据集名称	Brain US脑解剖分 割(超声)	GLAS腺体分割(显微 )	MoNuSeg多器官细胞 核分割(显微)
是否公开	未公开	公开	公开
数量（train，test）	Na	（85,80）	（30，14）
像素大小	Na	128×128	128×128 256×256

5.2 实验结果

表 2 MedT实验结果

实验对比	GLAS腺体分割(显微)		MoNuSeg多器官细胞核分割(显微)	
是否公开	F1	IoU	F1	IoU
原论文	81.02	69.61	66.04（128） 74.55（256）	51.24（128） 60.80（256）
我的	82.30	67.25	66.66（128） 75.78（256）	50.18（128） 60.82（256）

实验结果如表2所示。其中在GLAS腺体分割数据集上，F1（分类准确度标准）比原论文高了1.28，IoU（分割准确度）比原论文低了2.36；在128像素大小的MoNuSeg多器官细胞核分割数据集上，F1（分类准确度标准）比原论文高了0.52，IoU（分割准确度）比原论文低了1.06；在256像素大小的MoNuSeg多器官细胞核分割数据集上，F1（分类准确度标准）比原论文高了1.23，IoU（分割准确度）比原论文高了0.02；可能与硬件资源变动有关，复现结果还在合理的范围内。

实验分割效果如图4所示，虽然存在小细节错误，但直观感觉分割效果还是不错的。

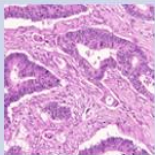


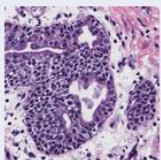
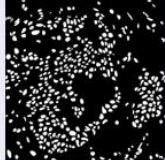
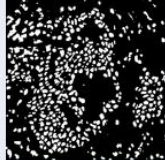
	原始图像	预测分割	真实标签
GlaS			
MoNuSeg			

图 4 MeT预测分割与真正标签比较

## 6 总结与展望

Gated Axial-Attention是一个比较通用的特征提取模块，不光医学图像分割任务可以用，其他分割、目标检测等任务都可以用。以后可以加在自己的项目中，看是否会有提高效果。局部-全局（LOGO）训练策略也是一个比较通用的策略，以后可以尝试用下。经过这次复现，感觉自己还要看和复现更多的论文，积累框架和方法等。

## 参考文献

[1]Valanarasu, Jeya Maria Jose et al. “Medical Transformer: Gated Axial-Attention for Medical Image Segmentation.” International Conference on Medical Image Computing and Computer-Assisted Intervention (2021).