Medical Transformer: Gated Axial-Attention for Medical Image Segmentation

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M.Patel
Johns Hopkins University, Baltimore, MD, USA
Rutgers, The State University of New Jersey, NJ, USA

Motivation

- 医学影像的数据较少，且标注需耗费大量时间, 无法像2D图像一样进行大量的预训练
- 卷积神经网络缺乏对映像中存在的远程依赖项进行建模的能力, 当分割的mask较大时，通过transformer学习与该mask相对应的像素之间的远程依存关系也有助于做出有效的预测

Contribution

- 提出了局部全局（LoGo）训练策略: 使用了浅层的全局分支和深层的局部分支来对医学图像的patch进行操作
- 针对医学图像数据少的问题, 提出了gated position-sensitive的轴向注意机制, 引入了四个Gate来控制对key，query和value的位置嵌入供应的信息量, 使得MedT可以应用于任何大小的任何数据集

网络结构

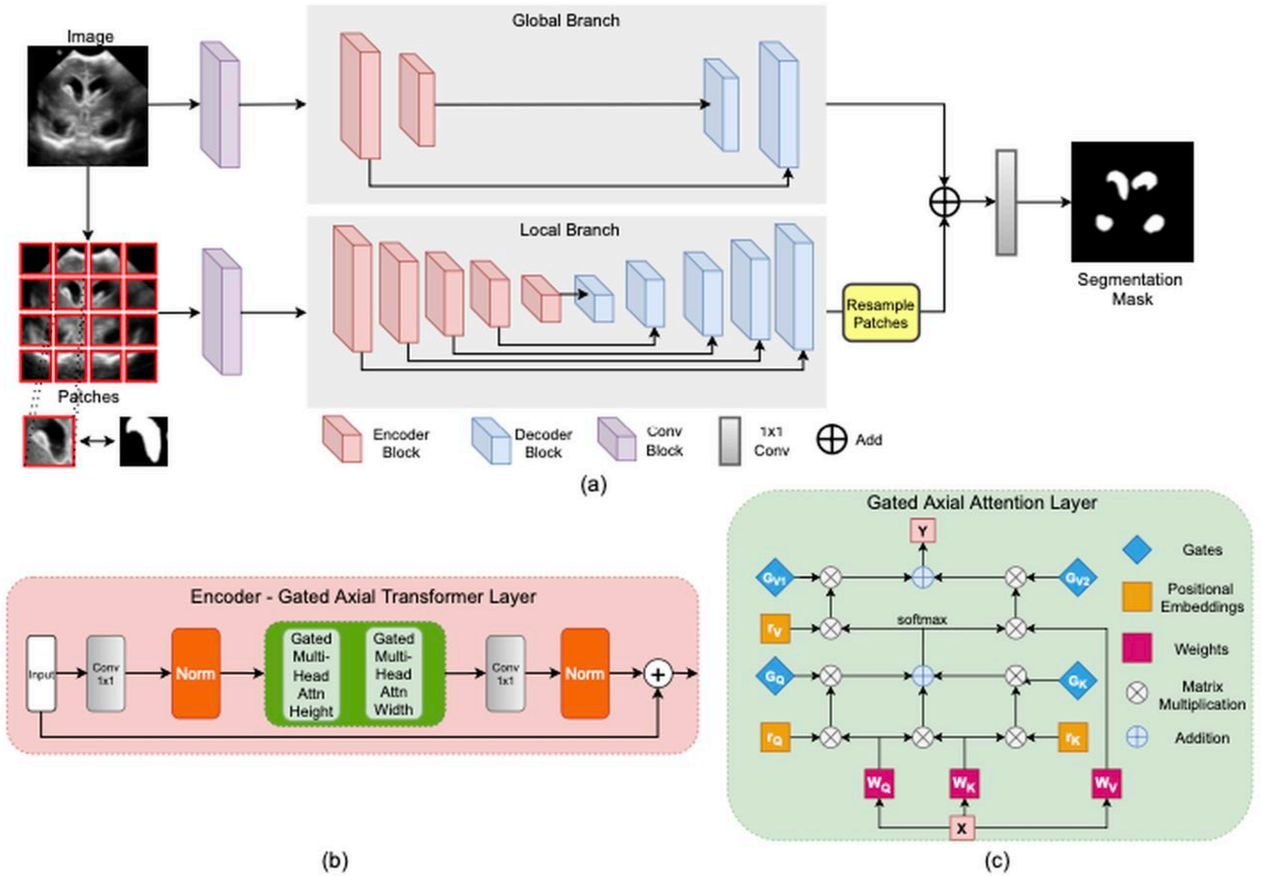


Fig. 2. (a) The main architecture diagram of MedT which uses LoGo strategy for training. (b) The gated axial transformer layer which is used in MedT. (c) Gated Axial Attention layer which is the basic building block of both height and width gated multi-head attention blocks found in the gated axial transformer layer.

细节: LoGo 训练策略

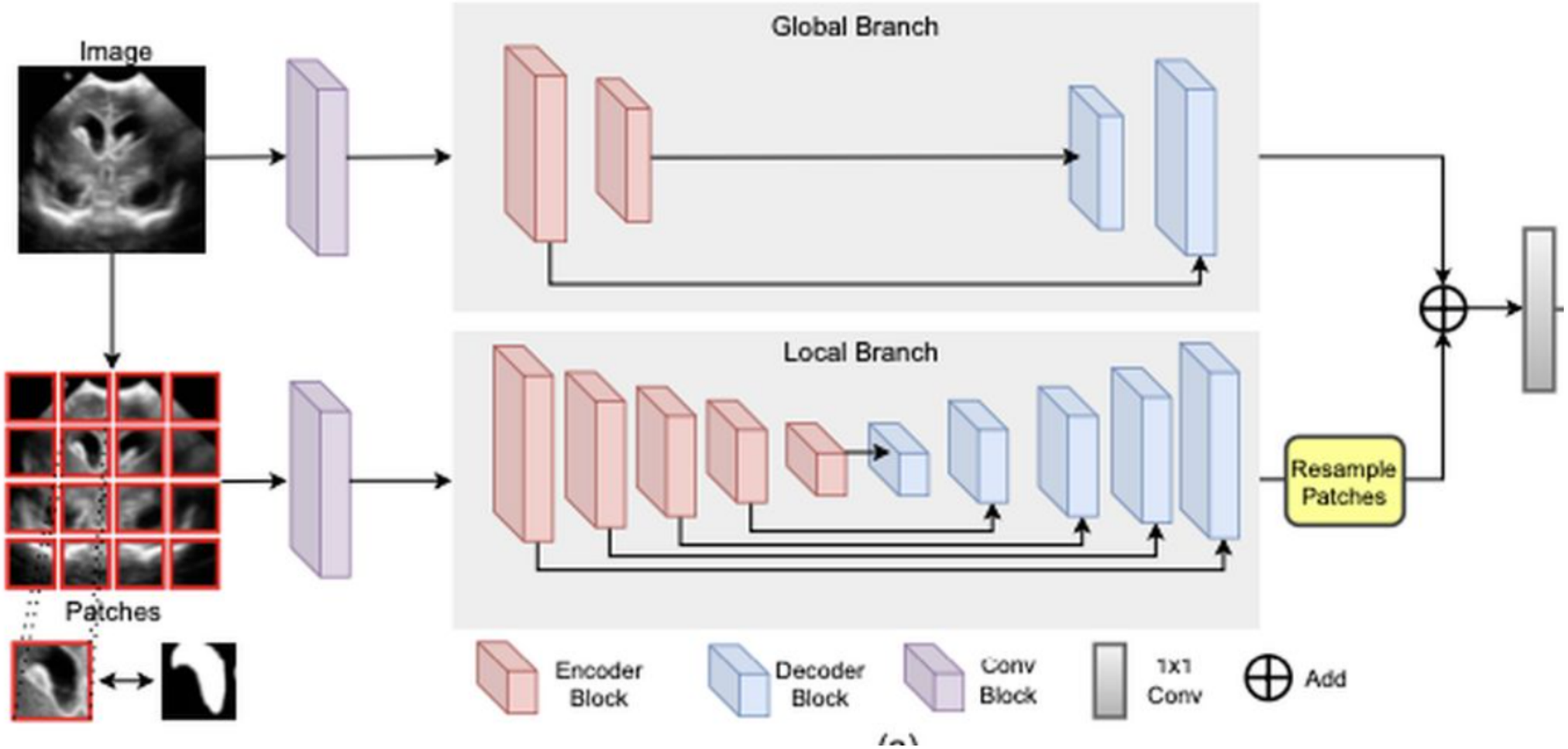
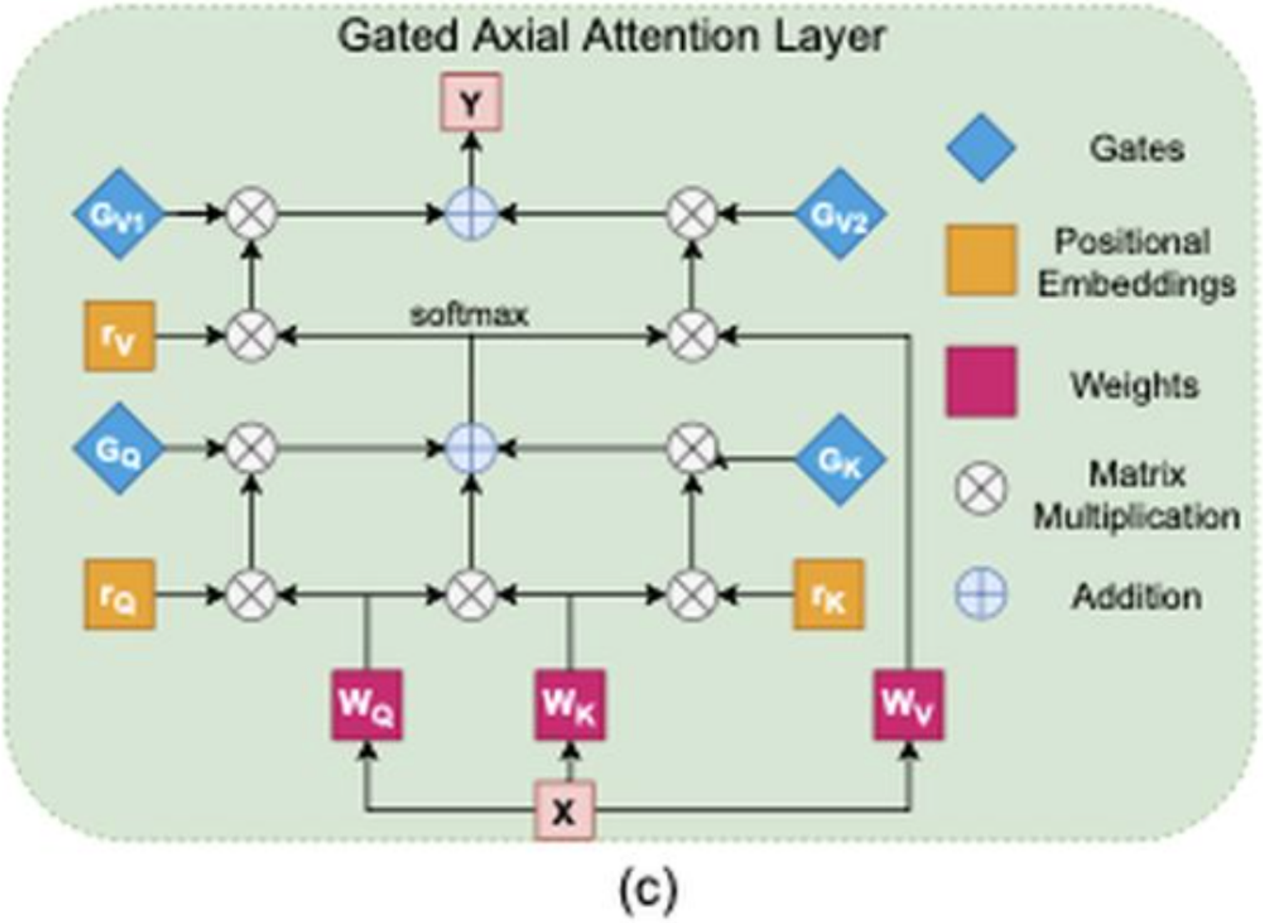


Image
Global Branch
Local Branch
Patches
Resample Patches
Encoder Block
Decoder Block
Conv Block
1x1 Conv
Add

细节: gated position-sensitive



(c)

$$y_{ij} = \sum_{w=1}^{W} \text{softmax}\left(q_{ij}^T k_{iw} + G_Q q_{ij}^T r_{iw}^q + G_K k_{iw}^T r_{iw}^k\right)\left(G_{V1} v_{iw} + G_{V2} r_{iw}^v\right), \quad (3)$$

动机:

1. **小规模数据集的位置偏差很难学习**，因此在编码长距离联系时并不总是准确的。

2. **相对位置编码不准确会引入噪音,** 从而降低性能

方法:

提出了一种改进的轴向注意block，该block可以控制位置偏差可在非局部上下文的编码中施加的影响。**GQ，GK，GV1，GV2**∈R是可学习的参数，该机制控制学习的相对位置编码对编码非局部上下文的影响。

实验结果

Table 1. Quantitative comparison of the proposed methods with convolutional and transformer based baselines in terms of F1 and IoU scores.

Type	Network	Brain US		GlaS		MoNuSeg	
		F1	IoU	F1	IoU	F1	IoU
Convolutional Baselines	FCN [1]	82.79	75.02	66.61	50.84	28.84	28.71
	U-Net [25]	85.37	79.31	77.78	65.34	79.43	65.99
	U-Net++ [44]	86.59	79.95	78.03	65.55	79.49	66.04
	Res-UNet [40]	87.50	79.61	78.83	65.95	79.49	66.07
Fully Attention Baseline	Axial Attention U-Net [37]	87.92	80.14	76.26	63.03	76.83	62.49
Proposed	Gated Axial Attn.	88.39	80.7	79.91	67.85	76.44	62.01
	LoGo	88.54	80.84	79.68	67.69	79.56	66.17
	MedT	**88.84**	**81.34**	**81.02**	**69.61**	**79.55**	**66.17**

Group-Free 3D Object Detection via Transformers

Ze Liu, Zheng Zhang, Yue Cao,Han Hu, Xin Tong
University of Science and Technology of China, Microsoft Research Asia

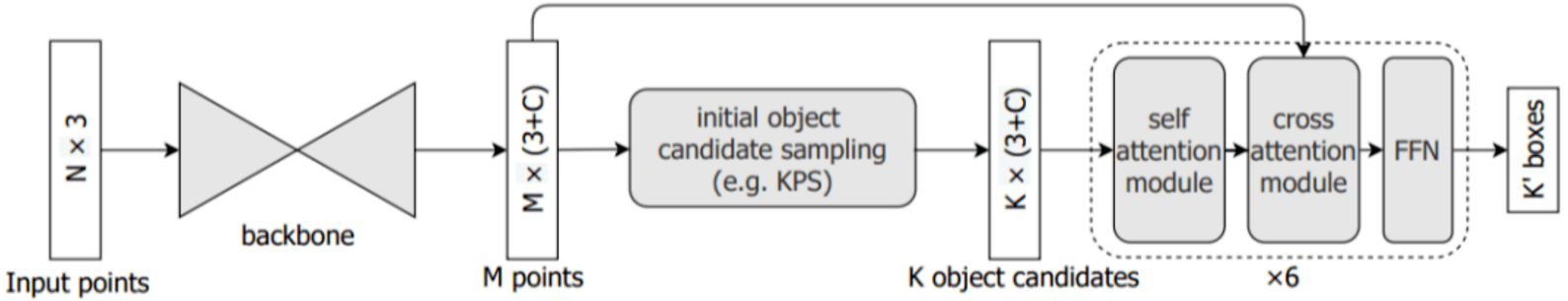
Motivation

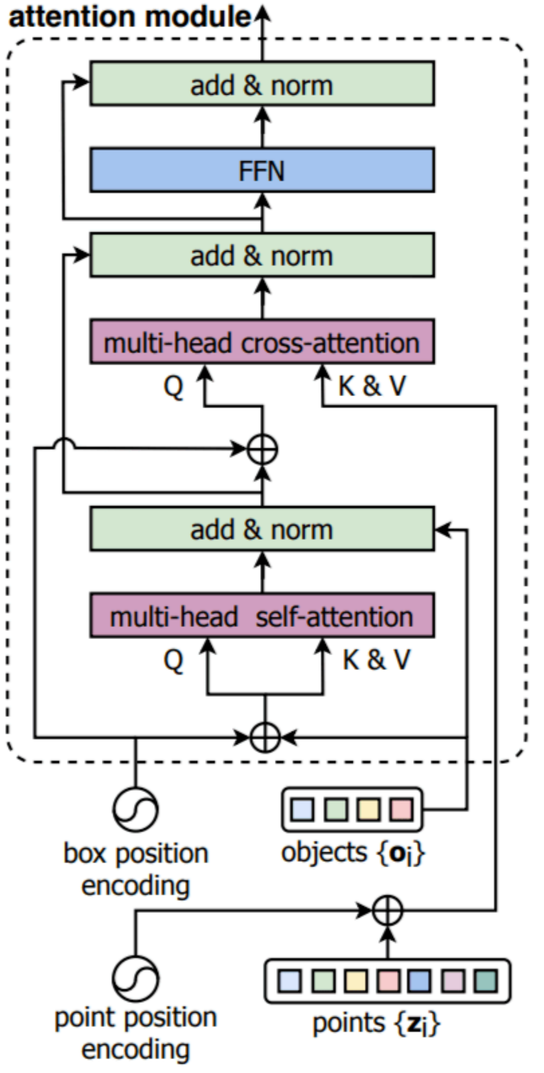
- 过去方法采用Bounding Box或Vote的形式进行聚合, 但真实场景的复杂和多样性往往会导致这些点聚合方法产生错误

Contribution

- 使用 Transformer中的注意力机制来利用点云中的所有点来**计算物体特征**，每个点对**物体的贡献通过网络训练而自动学习获得**
- 提出了**迭代式边界框预测（Iterative Box Prediction）**与多阶段预测整合（Ensemble Multi-stage Predictions），进一步提升了物体检测的准确度

网络结构


N × 3
Input points
backbone
M × (3+C)
M points
initial object candidate sampling (e.g. KPS)
K × (3+C)
K object candidates
self attention module
cross attention module
FFN
×6
K' boxes


attention module
add & norm
FFN
add & norm
multi-head cross-attention
Q
K & V
add & norm
multi-head self-attention
Q
K & V
box position encoding
objects {oi}
point position encoding
points {zi}

与DETR的区别

- **Position Encoding是迭代改进式的, 而不是固定的Sin或Cos函数**
- **利用PointNet预测的Object进行初始化, 而不是如传统Transformer随机初始化一些Query**

method	epoch	mAP@0.25	mAP@0.5
DETR	400	39.6	21.4
DETR+KPS	400	59.6	41.0
DETR+KPS+iter pred	400	59.9	42.9
DETR+KPS+iter pred	1200	61.8	45.2
Ours	400	**66.3**	**48.5**

Table 10. The comparison between DETR and our method on ScanNet V2. *KPS* represent *k-Closest Points Sampling*, *iter pred* represents iterative prediction.

实验结果

method	mAP@0.25	mAP@0.5
RoI-Pooing	65.1	44.4
Voting	64.2	44.1
Ours	**66.3**	**48.5**

Table 8. Comparison with grouping-based approaches.

method	backbone	mAP		frames/s
		0.25	0.5	
MLCVNet [31]	PointNet++	64.5	41.4	5.44
H3DNet [51]	4×PointNet++	67.2	48.1	3.76
Ours (L6, O256)	PointNet++	67.3	48.9	**6.71**
Ours (L12, O256)	PointNet++	67.2	49.7	5.70
Ours (L12, O256)	PointNet++w2×	68.8	52.1	5.23
Ours (L12, O512)	PointNet++w2×	**69.1**	**52.8**	5.17

Table 9. Comparison on realistic inference speed on ScanNet V2.