文章编号：2095-2163(2024)03-0076-05 中图分类号：TP391 文献标志码：A

基于 Transformer 的肺肿瘤三维 CT 图像分割

王伟桐[1]，玄 萍[1,2]
（1 黑龙江大学 计算机科学技术学院，哈尔滨 150080；2 汕头大学 工学院，广东 汕头 515063）

摘 要：基于信息学技术自动分割病人的肺部 CT 图像，有助于医生对于肺癌患者的早期诊断，提取和整合图像区域间的空间关联，对于提升肺肿瘤分割性能是十分重要的。本文提出了一个新的基于 Transformer 的分割模型，用于肺肿瘤三维 CT 图像分割、学习和整合此类关联。本文分别设计了带有混合多头图像区域节点注意力的 Transformer 模块和类别注意力模块，学习并融合了肺部 CT 图像的空间层面和通道层面的信息。将新的基于 Transformer 的分割模型同其他较为先进的模型进行了对比实验，实验结果表明新的模型在骰子系数、交并比和豪斯多夫距离等方面优于其他模型。

关键词：肺部 CT 图像；图像区域节点注意力；Transformer；类别注意力

Transformer-based segmentation of CT images of patients with lung tumors

WANG Weitong[1], XUAN Ping[1,2]
(1 College of Computer Science and Technology, Heilongjiang University, Harbin 150080, China;
2 College of Engineering, Shantou University, Shantou 515063, Guangdong, China)

Abstract: Automatic segmentation of CT images of patients with lung tumors based on informatics technology is helpful for promoting the early diagnosis of lung cancer patients. Extracting and integrating the spatial correlation among image regions is very important for improving the segmentation performance of lung tumors. The paper designs a Transformer module and a category attention module with a hybrid multi-head image region node attention, respectively, which learned and fused the spatial and channel-level information of the lung CT image. The paper designs the image region node attention mechanism and the category attention mechanism to learn and fuse the information of spatial level and that of channel level. The proposed method is compared with several segmentation methods for segmenting lung tumors. The experimental results show that the proposed segmentation model is superior to the compared segmentation models in *Dice*, *IoU* and *HD* distance.

Key words: CT image of the lungs; image area-wise node attention; Transformer; category attention

0 引 言

在过去几年里，肺癌是全世界癌症相关死亡的主要原因，占全部因癌症死亡人数的四分之一以上[1]。随着计算机视觉技术的不断发展，使用计算机辅助分割肺肿瘤的模型逐渐丰富，准确率也不断提高[2]。尽管如此，目前肺肿瘤的自动分割仍然是一项具有挑战性的任务，原因之一是肺部组织与肺肿瘤的语义和纹理特征相似度较高；另一个原因是不同病人的肿瘤在位置、形状上具有较大的差别。由于肺肿瘤和肺部其他组织的位置之间存在很强的联系，有效的建模肺部 CT 图像各区域间的相关性有助于肺肿瘤的检测和分割。

随着卷积神经网络(CNN)的出现，其衍生出的模型在自动化的医学图像分析任务中展现出了不错的性能，该类模型具备高表示能力、快速推理以及卷积核共享的特性，全卷积网络和 U-Net 就是该类模型的代表[3-4]。尽管这些模型具有较好的肺肿瘤分割能力，但当肿瘤的形状和大小等方面出现较大差异时，这些模型更多地依赖于多级级联 CNN 的方法[5]。这些模型固定了滤波器的大小，所以难以捕捉远程节点间的关系，进而导致在肿瘤大小不断变化的情况下很难实现精准分割，这也是目前卷积神经网络在图像处理方面的局限性[6]。

为了克服卷积神经网络的这一局限性，已有研究提出基于 CNN 特征建立自注意机制[7]。在深度

基金项目：国家自然科学基金（61972135）；黑龙江省自然科学基金项目（LH2019F049）；中国博士后科学基金（2019M650069）；黑龙江省博士后科研启动基金（BHLQ18104）。
作者简介：王伟桐（1996-），男，硕士研究生，主要研究方向：计算机视觉。
通讯作者：玄 萍（1979-），女，博士，教授，主要研究方向：医学图像处理与分析、深度学习。Email：xuanping@ hlju.edu.cn
收稿日期：2023-03-05

学习领域的其他方向，完全依赖于注意力机制的 Transformer 已经出现，这类模型应用于包括计算机视觉在内的其他方向，并取得了不错的效果[8-9]。与以往基于 CNN 的方法相比，Transformer 更擅长学习图像区域的全局上下文，被广泛地应用于视频处理以及其他场景的图像分割，均得到了相对较好的预测结果[10-12]。

因为不同患者的肿瘤大小不同、形状差异大、边界特点各异，所以从 CT 中自动分割肺肿瘤是一个具有挑战性的问题[13]。为此，Kim 等学者[14]提出了一种可用于肺肿瘤分割的从粗略到精细的神经结构分割模型，仍然使用卷积块来作为分割模型的构成元件，因此无法避免卷积的局限性。Oktay 等学者[15]提出了一种通过门控注意力来整合空间信息的 Attention U-Unet 分割模型，在多次竞赛中取得了不错的成绩。Fabian 等学者[16]建立了 nnU-Net 模型，并且优化了数据的预处理、网络结构、模型训练和分割结果的后处理。nnU-Net 模型在 2018 年的 Medical Segmentation Decathlon Challenge 比赛中排名第一，并在后续的其他比赛中也取得了不错的成绩[17]。然而，肺肿瘤三维 CT 图像的特点是左右 2 个区域的肺组织和肺肿瘤具有相似的纹理特征，上述模型均忽略了长距离对象间的空间依赖。

1　肺肿瘤三维 CT 图像分割模型

为了实现肺肿瘤的分割，本文提出了基于 Transformer 的肺肿瘤三维 CT 图像分割模型。首先，通过分割主干模块来学习上下文表示，以达到抓取图像的纹理和语义特征的目的；针对分割编码器提取到的特征，设计了一种带有混合多头自注意力的 Transformer 模块，该模块中的注意力结构带有独立的空间分支和通道分支，用来学习全局层面的内容表示；最后，为了自适应融合不同分支学习到的空间表示和通道表示，设计了类别注意力模块，并最终实现分割输出。

1.1　分割主干模块

本文选用 3D nnU-Net 作为分割主干模块来提取上下文表示，编码器和解码器分别由 6 个编码层和 6 个解码层构成。3×3×3 的卷积块是所有编码层的重要组成部分，Instance Normalization 和 Leaky Relu 是激活函数。在每个下采样阶段，使用跨步卷积代替池化层，获得更具代表性的上下文表示。建立每一个解码层时，选用转置卷积的方式来进行上采样。为了集成更多的细节特征，从编码层到解码层的跨层连接。用 $\boldsymbol{Z} \in R^{H\times W\times D\times C}$ 表示最后一个编码层的输出特征，H、W、D 分别表示高度、宽度和深度，通道数用 C 来表示，$\boldsymbol{Z}$ 是输入图像的上下文表示。

1.2　带有混合多头图像区域节点注意力的 Transformer 模块

带有混合多头图像区域节点注意力（如图 1 所示）的 Transformer 模块将主干网络最后一个编码层的输出特征 $\boldsymbol{Z} \in R^{H\times W\times D\times C}$ 进行通道层面的分割（Split），分别得到 $\boldsymbol{X}_1 \in R^{N\times \frac{c}{4}}$、$\boldsymbol{X}_2 \in R^{N\times \frac{c}{4}}$、$\boldsymbol{X}_3 \in R^{N\times \frac{c}{2}}$ 作为由上至下的 3 个分支的输入。$\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$ 经过通道和空间的学习得到的特征可分别表示为：

$$\boldsymbol{Y}_1 = \boldsymbol{W}_2[\boldsymbol{W}_1\boldsymbol{X}_1 + \boldsymbol{B}_1] + \boldsymbol{B}_2 \tag{1}$$

$$\boldsymbol{Y}_2 = DWConv(FC(\boldsymbol{X}_2)) \tag{2}$$

其中，$\boldsymbol{B}_1$ 和 $\boldsymbol{B}_2$ 是偏差向量，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 是可学习的权重矩阵。

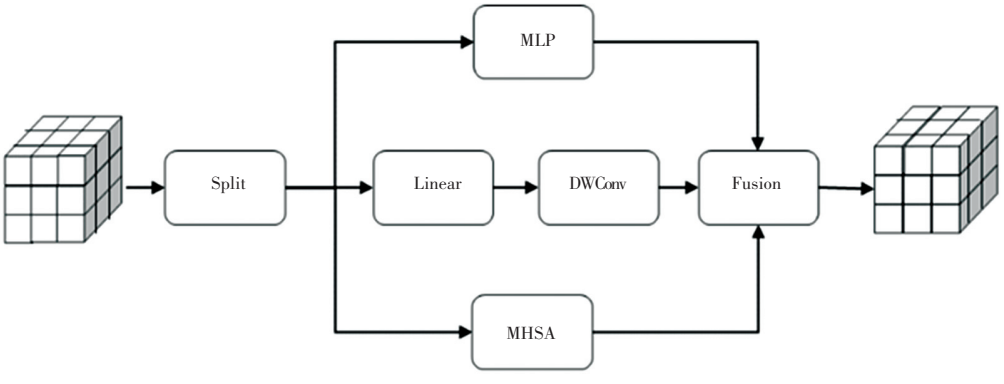

图 1　混合多头图像区域节点注意力

Fig. 1　Mixer multi-head image region node attention

此外,给定分割出的特征图 X_3,该模块将其重塑为 $\boldsymbol{X} \in R^{N\times C}$,$\boldsymbol{X}$ 中的第 j 个位置 $\boldsymbol{X}_j \in R^{1\times C}$ 对应输入图像的特定区域,其中 $N = H \times W$ 是图像区域节点的个数。随后将 $\boldsymbol{X}$ 与可学习的矩阵相乘,生成新的特征图 $\{\boldsymbol{Q}、\boldsymbol{K}\} \in R^{N\times C}$,对此可表示为:

$$\boldsymbol{Q} = \boldsymbol{W}_q\boldsymbol{X} \tag{3}$$

$$\boldsymbol{K} = \boldsymbol{W}_k\boldsymbol{X} \tag{4}$$

其中,$\boldsymbol{W}_q$、$\boldsymbol{W}_k$ 是权重矩阵。

在 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的转置之间执行矩阵乘法,目的是得到相似性矩阵 $\boldsymbol{A} \in R^{N\times N}$,见式(5):

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}} \tag{5}$$

2 个节点的特征越相似表示 2 个节点的语义之间的相关性越强。将特征 $\boldsymbol{X}$ 与另一个权重矩阵 $\boldsymbol{W}_v$ 相乘,生成特征投影 $\boldsymbol{V}$,可由式(6) 进行描述:

$$\boldsymbol{V} = \boldsymbol{W}_v\boldsymbol{X} \tag{6}$$

将相似性矩阵 $\boldsymbol{A}$ 经过 *softmax* 激活后的结果与 $\boldsymbol{V}$ 之间执行矩阵乘法,计算公式具体如下:

$$Y_{3j} = \frac{\exp\left(\frac{a_{ij}}{\sqrt{d_k}}\right)}{\sum_{i=1}^{N}\exp\left(\frac{a_{ij}}{\sqrt{d_k}}\right)}V_i \tag{7}$$

其中,d_k 是每个区域节点的特征数,a_{ij} 表示图像区域节点 i 与 j 的相似性,$i \in (1\cdots N)$、$j \in (1\cdots N)$ 分别表示矩阵 $\boldsymbol{A}$ 的行和列。

从式(7)可以推断第 j 个位置的结果特征 Y_{3j} 是所有位置的原始特征的加权和。因此,该特征具有全局上下文视图,并根据区域节点的相似性矩阵选择性地聚合上下文。为了避免学习过程中的偏差,本文将上述图像区域节点注意力推广为多头图像区域节点注意力。

1.2.1 类别注意力模块

因为混合多头图像区域节点注意力模块中不同分支学习到的多个信息对于肺肿瘤分割任务具有不同的重要性,所以本文建立了一个新的类别注意力模块(Fusion),以得到增强后的特征,通过对注意力得分的自适应加权,实现对来自不同分支的信息进行自适应融合。

第 h 个分支的信息分数为 S^h 的定义公式可写为:

$$s_i^h = [\tanh(\boldsymbol{W}_t x_i^h + \boldsymbol{b}_t)]^{\mathrm{T}}\boldsymbol{S}_t \tag{8}$$

其中,$\boldsymbol{W}_t$ 是可学习的权重矩阵;$\boldsymbol{b}_t$ 是偏差向量;$\boldsymbol{S}_t$ 用于捕获不同类别信息特征的上下文。

归一化注意力权重 α_i 的定义公式如下:

$$\alpha_i^h = \frac{\exp(s_i^h)}{\sum_{h=1}^{H}\exp(s_i^h)} \tag{9}$$

注意力增强后的节点特征向量 $\tilde{Y}_i$,其计算见式(10):

$$\tilde{Y}_i = \sum_{h=1}^{H} Y_i^h \otimes \alpha_i^h \tag{10}$$

其中,“$\otimes$”表示元素级乘积运算符。

最终,该模块将特征矩阵 $\tilde{\boldsymbol{X}}_1$ 和原始信息 Y_i 进行通道维度的拼接,并通过全连接层将其维度重新降回到原始的通道数量,得到输出特征 $\boldsymbol{Y}$。

1.2.2 混合多头图像区域节点注意力模块

混合多头图像区域节点注意力模块在 Transformer 中的位置如图 2 所示,带有混合多头图像区域节点注意力的 Transformer 模块的编码层中包含 M 层的混合多头图像区域节点注意力模块以及前馈层(FFD),因此第 m 层的编码图像表示 P_m 的运算需要用到:

$$P_m = FFD(LN(P_m^{'})) + P_m^{'} \tag{11}$$

$$P_m^{'} = MixerMHSA(LN(P_{m-1})) + P_{m-1} \tag{12}$$

其中,LN 表示层归一化运算符。

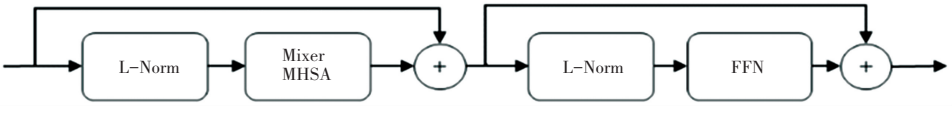

图 2 混合多头图像区域节点注意力模块在 Transformer 中的位置

Fig. 2 Position of Mixer MHSA in Transformer

经过混合多头自注意力的计算后,将 P 重塑成 $\boldsymbol{Z} \in R^{H\times W\times D\times C}$。

2 实验

2.1 相关数据集

本文使用 Medical Segmentation Decathlon Challenge 比赛中的 Lung Tumors dataset 数据集来评估模型的性能,将其中 64 例患者的肺部 CT 图像分为训练集和测试集两个部分,该数据集中每位患者的 CT 图像对应的真实标签仅有一项肺肿瘤标记,数据采自不同医院的多台扫描仪,各个病例的体素大小并不完全相同。本文将所有 CT 图像的体素重采样为

1.24×1.24×0.78 大小，并使用水平和垂直翻转、随机旋转、亮度和伽马噪声增强以及随即缩放等数据增强方法来扩充现有的数据集。64 个案例被随机选择 20%作为测试集，将其余的 56 个案例随机划分 10 份，随机选择其中的 9 份作为训练集，1 份用于验证。

2.2　评价指标

使用衡量空间体积关系的骰子系数（*Dice*）[18] 和交并比（*IoU*）[19]，以及形状相似度方面的豪斯多夫距离（*HD*）[20]，来评估肺肿瘤的分割性能。

肿瘤的骰子系数 $Dice_{tum}$ 在 0 和 1 范围内，值越大表示分割结果越好。可由式(13)来计算：

$$Dice_{tum}=\frac{2\mid P_{tum}\cap G_{tum}\mid}{\mid P_{tum}\mid+\mid G_{tum}\mid} \tag{13}$$

其中，P_{tum} 和 G_{tum} 分别表示分割结果和真实标签。

肿瘤的 IoU_{tum}，可由式(14) 来求值：

$$IoU_{tum}=\frac{\mid P_{tum}\cap G_{tum}\mid}{\mid P_{tum}\cup G_{tum}\mid} \tag{14}$$

真实标签(Ground Truth)的边界与分割的肿瘤边界之间的 *HD* 定义见式(15)：

$$HD_{tum}(P_{tum},G_{tum})=\max\{h(P_{tum},G_{tum}),h(G_{tum},P_{tum})\} \tag{15}$$

其中，$h(P_{tum},G_{tum})$ 表示 P_{tum} 与 G_{tum} 之间的距离，见式(16)：

$$h(P_{tum},G_{tum})=\max_{p\in P_{tum}}\min_{g\in G_{tum}}\|p-g\| \tag{16}$$

其中，p 和 g 分别来自 P_{tum} 和 G_{tum}。

研究可知，HD_{tum} 值越小，分割效果越好。

2.3　与其他方法的比较

为了进一步评估模型的性能，将本文的模型与其他先进的肺肿瘤分割模型 SCANS、3D U-Net、Attention U-Net、3D nnU-Net 、3D ResNet 进行了对比实验，实验结果见表 1。本文的模型在骰子系数和豪斯多夫距离等方面取得了领先于其他网络的分割效果，在肿瘤分割方面，本文模型的骰子系数达到了最高水平、为 0.778，比 SCANS 高 1.21%，比 Attention U-Net 提升了 2.25%，比 3D U-net 高4.43%，比 3D nnU-Net 以及 3D ResNet 分别提高了3.5%和2.5%；在交并比方面，模型的最佳交并比为 0.655 1，比 Attention U-Net、3D U-net、3D nnU-Net 和 3DResNet 分别高出 3.39%、5.35%、3.57%和2.96%。此外，本文的模型得到了最低的豪斯多夫距离值 63.641 7 mm，分别比其他方法优化了30.355 3 mm、2.857 1 mm、45.731 6 mm、和40.387 2 mm。

表 1　本文的模型和其他先进模型比较的结果

Table 1　Results comparison of the model in this paper with other advanced models

模型	骰子系数（$Dice_{tum}$）	交并比（IoU_{tum}）	豪斯多夫距离（HD_{tum}）/ mm
SCANS	0.765 9	-	-
3D U-Net	0.733 7	0.601 6	93.977 0
Attention U-Net	0.752 5	0.621 2	66.498 8
3D nnU-Net	0.743 0	0.610 4	109.013 3
3D ResNet	0.753 0	0.625 5	104.028 9
本文模型	0.778 0	0.655 1	63.641 7

3　结束语

本文提出了一种新的基于 Transformer 的分割模型提取和整合图像区域间的空间关联。该模型所包含的带有混合多头图像区域节点注意力的 Transformer 模块以及类别注意力模块，分别学习和融合了肺部 CT 图像的空间层面和通道层面的信息。实验结果表明，该模型具有从 CT 图像中分割肺肿瘤的潜力并为治疗提供帮助。

参考文献

[1] GANSLER T, GANZ P A, GRANT M, et al. Sixty years of CA: A cancer journal for clinicians[J]. CA: A Cancer Journal for Clinicians, 2010,60 (6): 345-350.

[2] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.

[3] CUI Hui, XU Yiyue, LI Wanlong, et al. Collaborative learning of cross-channel clinical attention for radiotherapy-related esophageal fistula prediction from CT[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020: 212-220.

[4] HSU J, CHIU W, YEUNG S. DarCNN: Domain adaptive region-based convolutional neural network for unsupervised instance segmentation in biomedical images[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 1003-1012.

[5] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431-3440.

[6] 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. Cambridge, UK: Springer, 2015:234-241.

[7] SCHLEMPER J, OKTAY O, SCHAAP M, et al. Attention gated

networks: Learning to leverage salient regions in medical images [J]. Medical Image Analysis, 2019, 53: 197-207.

[8] 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. Long Beach, USA: NIPS Foundation, 2017: 6000-6010.

[9] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2018: 4171-4186.

[10] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[C]// International Conference on Machine Learning. Stockholm: dblp, 2018: 4055-4064.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition atscale[J]. arXiv preprint arXiv: 2010.11929, 2020.

[12] MOBINY A, NGUYEN H V. Fast CapsNet for lung cancer screening [C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2018: 741 - 749.

[13] BERTASIUS G, WANG Heng, TORRESANI L. Is space-time attention all you need for video understanding? [C]// Proceedings of the International Conference on Machine Learning (ICML). IEEE, 2021, 2: 4.

[14] KIM S, KIM I, LIM S, et al. Scalable neural architecture search for 3d medical image segmentation [M]// SHEN D, et al. Medical image computing and computer aided intervention - MICCAI 2019. Lecture Notes in Computer Science. Cham: Springer, 2019, 11766: 220-228.

[15] OKTAY O, SCHLEMPER J, FOLGOC L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv: 1804.03999, 2018.

[16] FABIAN I, JAEGER P F, KOHL S A, et al. Maier-Hein. nnu-net: A self - configuring method for deep learning - based biomedical image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 18 (2): 203-211.

[17] ANTONELLI M, REINKE A, BAKAS S, et al. The medical segmentation decathlon [J]. Nature Communication, 2022, 13 (1): 1-13.

[18] ZOUK H, WARFIELD S K, BHARATHA A, et al. Statistical validation of image segmentation quality based on a spatial overlap index1: Scientific reports[J]. Academic Radiology, 2004, 11 (2): 178-189.

[19] CSURKA G, LARLUS D, PERRONNIN F, et al. What is a good evaluation measure for semantic segmentation[C]// Proceedings of the British Machine Vision Conference. Manchester, UK : BMVA Press , 2013, 27: 10-21.

[20] ÇIÇEK Ö, ABDULKADIR A, LIENKAMP S S, et al. 3D U-Net: Learning dense volumetric segmentation from sparse annotation [C]//International Conference on Medical Image Computing and Computer-assisted Intervention. Athens, Greece: Springer, 2016, 9901: 424-432.