Computer Science and Application 计算机科学与应用, 2024, 14(7), 78-83
Published Online July 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.147165

Hans汉斯

基于CoTr分割网络的3D多器官CT图像分割

赵　威

浙江财经大学数据科学学院，浙江　杭州

收稿日期：2024年6月12日；录用日期：2024年7月17日；发布日期：2024年7月25日

摘　要

在医学图像分割领域U-Net已经成为了被应用最广泛的医学图像分割模型，许多有关医学图像分割的研究都用U-Net作为基线标准。以U-Net为基础的一系列变体分割模型也相继问世，其中包括CoTr，其为Convolutional neural network and a Transformer的简称。就如其名，CoTr是一个结合了卷积神经网络和Transformer，具有类似U-Net的U形结构的分割网络。CoTr构造卷积层以提取特征表示，并且构造有效的可变形Transformer (DeTrans)以对提取的特征图的长程依赖性进行建模。与平等对待所有关键位置的vanilla Transformer不同，DeTrans通过引入可变形的自注意机制，只关注一小部分关键位置。因此，DeTrans的计算和空间复杂性大大降低，使得处理多尺度和高分辨率特征图成为可能，而这些特征图通常对图像分割至关重要。CoTr模型在多模态腹部分割数据集(Amos数据集)上进行了广泛评估。结果表明，在3D多器官分割任务上，与其他基于CNN、基于Transformer和混合方法相比，CoTr带来了持续的性能改进。

关键词

U-Net，卷积神经网络，分割网络

3D Multi-Organ CT Images Segmentation Based on CoTr Segmentation Network

Wei Zhao

School of Data Science, Zhejiang University of Finance & Economics, Hangzhou Zhejiang

Received: Jun. 12th, 2024; accepted: Jul. 17th, 2024; published: Jul. 25th, 2024

Abstract

U-Net has become the most widely used medical image segmentation model in the field of medical image segmentation, and many studies related to medical image segmentation use U-Net as the

文章引用: 赵威. 基于 CoTr 分割网络的 3D 多器官 CT 图像分割[J]. 计算机科学与应用, 2024, 14(7): 78-83.
DOI: 10.12677/csa.2024.147165

baseline standard. A series of variant segmentation models based on U-Net have also emerged, including CoTr, which stands for Convolutional Neural Network and a Transformer. As its name suggests, CoTr is a segmentation network that combines convolutional neural networks and Transformers, with a U-Net like U-shaped structure. CoTr constructs convolutional layers to extract feature representations and constructs effective deformable Transformers (DeTrans) to model the long-range dependencies of the extracted feature maps. Unlike vanilla Transformers that treat all key positions equally, DeTrans introduces a deformable self attention mechanism and only focuses on a small portion of key positions. Therefore, the computational and spatial complexity of DeTrans is greatly reduced, making it possible to process multi-scale and high-resolution feature maps, which are usually crucial for image segmentation. The CoTr model has been extensively evaluated on the multimodal abdominal segmentation dataset (Amos dataset). The results indicate that CoTr brings continuous performance improvement in 3D multi organ segmentation tasks compared to other CNN based, Transformer based, and hybrid methods.

Keywords

U-Net, Convolutional Neural Network, Segmentation Network

1. 引言

图像分割是医学图像分析中的一个长程挑战。自 U-Net 引入以来，全卷积神经网络(CNNs)已成为解决这一任务的主要方法。尽管其应用普遍，但由于局部性和权重共享的诱导偏差，卷积神经网络(CNN)的感受野仍然有限，无法捕捉到长程依赖性。Transformer 是一种序列到序列的预测框架，由于其强大的长程建模能力，在机器翻译和自然语言处理方面有着良好的表现。Transformer 中的自注意机制可以根据输入内容动态调整感受野，因此在建模长程依赖性方面优于卷积运算。最近，Transformer 被认为是一种替代架构，并在许多计算机视觉任务上取得了有竞争力的性能，如图像识别、语义/实例分割、目标检测、和图像生成。一个典型的例子是视觉 Transformer (ViT)，它在识别任务上优于基于 ResNet 的卷积神经网络，但代价是需要使用大量数据进行训练。由于并不是所有任务都有庞大的训练数据可用，特别是在医学图像领域公开可用的图像数据及其稀少。因此最近的研究试图将卷积神经网络和 Transformer 组合成一个混合模型。

CoTr 框架有效地桥接了卷积神经网络和 Transformer，用于 3D 医学图像分割。CoTr 具有编码器－解码器结构。在编码器中，采用简洁的卷积神经网络结构来提取特征图，并使用 Transformer 来捕获长程依赖关系。Transformer 中还引入了可变形的自注意机制，这种注意力机制只关注一小部分关键采样点，从而显著降低了 Transformer 的计算和空间复杂性。因此，Transformer 可以处理卷积神经网络生成的多尺度特征图，并保留丰富的高分辨率信息用于分割。

相关研究

卷积神经网络(CNN)的感受野有限，无法捕捉到长程依赖性。许多研究都致力于扩大 CNN 的感受野，从而提高其上下文建模能力。余等人[1]提出了具有可调扩张率的萎缩卷积，其在语义分割中显示出优越的性能。彭等人[2]设计了大型内核来捕获丰富的全局上下文信息。赵等人[3]在多个特征尺度上采用金字

塔池来聚合多尺度全局信息。王等人[4]提出了非局部运算，该运算通常嵌入在编码器的末端，以捕获长程依赖性。

在将 CNN 和 Transformer 组合成一个混合模型的研究中。Carion 等人[5]使用 CNN 提取图像特征，并使用 Transformer 对提取的特征进行进一步处理。陈等人[6]设计了 TransUNet，其中 CNN 和 Transformer 以级联方式组合，以制作用于 2D 医学图像分割的强编码器。尽管 TransUNet 的设计很有趣，性能也很好，但由于其存在自注意力，优化该模型具有挑战性。首先，它需要非常长的训练时间来将注意力集中在显著位置，尤其是在 3D 场景中，最初注意力均匀地投射到每个像素。其次，由于其高计算复杂度，vanilla Transformer 很难处理多尺度和高分辨率的特征图，这在图像分割中起着关键作用。

2. 方法

CoTr 的结构如图 1 所示，它由用于特征提取的 CNN 编码器、用于长程依赖建模的可变形 Transformer 编码器(DeTrans 编码器)和用于分割的解码器组成。

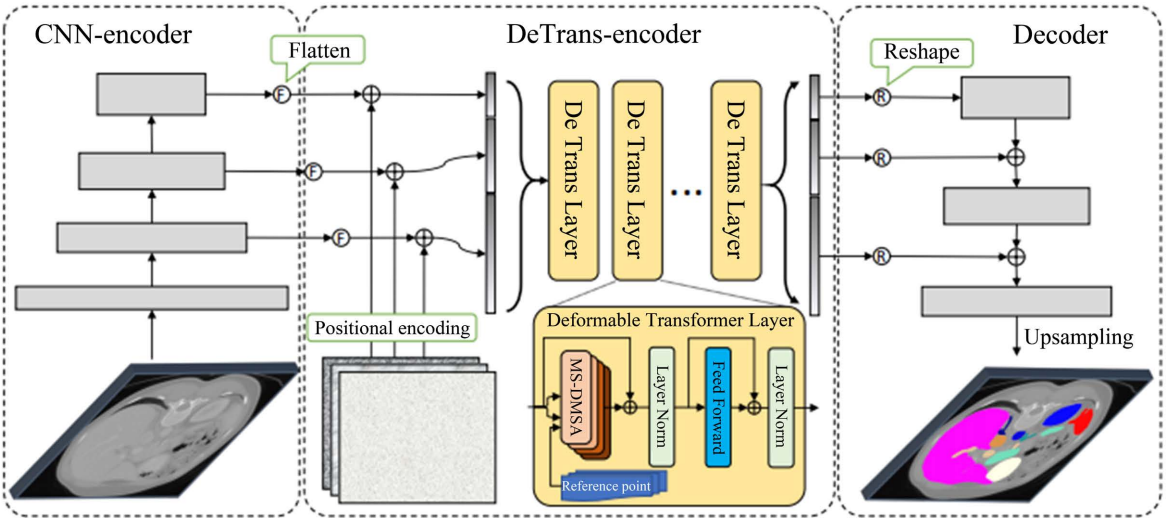

Figure 1. CoTr network structure diagram
图 1. CoTr 网络结构图

2.1. CNN 编码器

CNN 编码器由一个 Conv-IN-ReLU 块和三个 3D 残差块串联组成。其中 Conv 代表 3D 卷积层，IN 代表实例归一化，是分割算法中常用的归一化方法，ReLU 是激活函数。残差块是残差网络(Resnet)的基本组成单元，是现在最常用的一种卷积网络块。CNN 编码器的主要任务是对输入的图像进行特征提取和下采样即在缩小特征图尺寸的同时增加特征图的通道数量。CNN 编码器的整个流程由式(1)所示：

$$y = F_{\mathrm{Re}s}\left(f_{relu}\left(f_{in}\left(f_{conv}\left(x\right)\right)\right)\right)\#\tag{1}$$

其中 y 代表最终输出的特征图，x 代表输入的原始图像，f_{Conv} 代表 3D 卷积层，f_{in} 代表实例归一法，f_{relu} 代表 Relu 激活函数。

2.2. DeTrans 编码器

由于卷积运算的固有局部性，CNN 编码器无法有效地捕捉像素的长程依赖性。为此，我们提出了 DeTrans 编码器，该编码器引入了多尺度可变形自注意(MS-DMSA)机制，用于高效的长程上下文建模。

输入到序列的转换：考虑到 Transformer 以序列到序列的方式处理信息，首先将 CNN 编码器输出的特征图展平为一维序列。但是，平坦化特征的操作会丢失对图像分割至关重要的空间信息。为了解决这个问题，在展平的一维序列上加入 3D 位置嵌入(positional encoding)。

MS-DMSA 层：在 Transformer 的体系结构中，自注意力层会查看特征图中所有可能的位置，它具有收敛速度慢、计算复杂度高的缺点，难以处理多尺度特征。为了解决这一问题，MS-DMSA 层只关注参考位置周围的一小部分关键采样位置，而不是所有位置。MS-DMSA 层的流程如式(2)所示：

$$y_{MS-DMSA} = F_{liner}\left(Concat\left(head_1, head_2, \cdots\cdots, head_i\right)\right)\# \tag{2}$$

其中 $y_{MS-DMSA}$ 代表 MS-DMSA 层的最终输出，F_{liner} 代表线性映射层，对所有注意力头部的特征表示进行加权和聚合，Concat 代表拼接操作，将注意力头的输出在某个维度上拼接，$head_i$ 代表注意力头的输出，i 代表 Transformer 层中注意力头的数量，一般取 8 或 12。

DeTrans 层：DeTrans 层由 MS-DMSA 层和前馈网络组成，每个层之后进行层归一化。在每个子层中采用跳跃连接以避免梯度消失。DeTrans 编码器是通过重复堆叠 DeTrans 层来构建的。

2.3. 解码器

DeTrans 编码器的输出序列根据每个比例的大小重新整形为特征图然后输入到解码器。解码器是纯 CNN 架构，使用转置卷积将特征图逐步上采样到输入分辨率，然后使用 3D 残差块细化上采样的特征图。此外，还增加了编码器和解码器之间的跳过连接，以保留更多的低级别细节，从而更好地进行分割。

3. 实验

3.1. 数据集

多模态腹部分割数据集(AMOS)由深圳市大数据研究院、香港中文大学(深圳)、香港大学、中山大学等机构联合深圳市龙岗区人民医院、深圳市龙岗中心医院提出，是一个大规模，多样性的，收集自真实临床场景下的腹部多器官分割基准数据。AMOS 总计提供了 500 个 CT 与 100 个 MRI 扫描，每个扫描附带了 15 个腹部器官的体素级标注，是目前已知最全面的腹部分割基准数据集。同时，AMOS 的数据收集于多模态，多中心，多厂商，多阶段，多病种的病人，具有丰富的数据多样性，也更符合真实临床场景。图 2 展示了 AMOS 数据集中前十个最常见疾病和对应病变器官的分布数量，图 3 展示了每个器官类别的注释体素数量。本次实验使用了 AMOS 数据集中的 500 张 CT 影像，按 8:2 划分成训练集和测试集，对 15 个器官中的肝、脾、左肾和右肾四种器官进行分割。

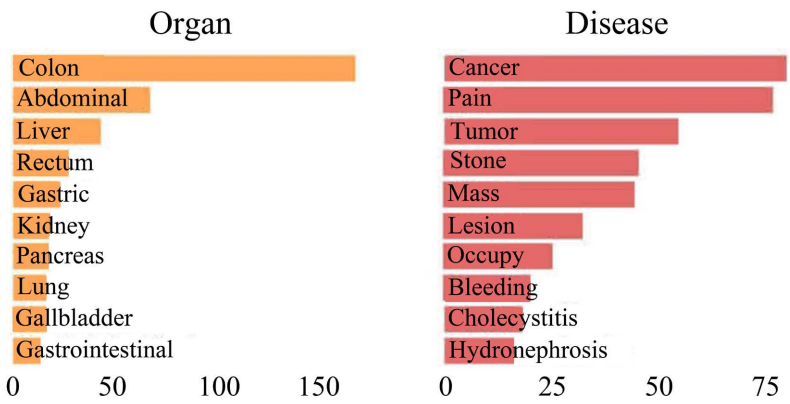

Figure 2. Top-ten most frequent diseases and diseased organs
图 2. 前十个最常见疾病和对应病变器官

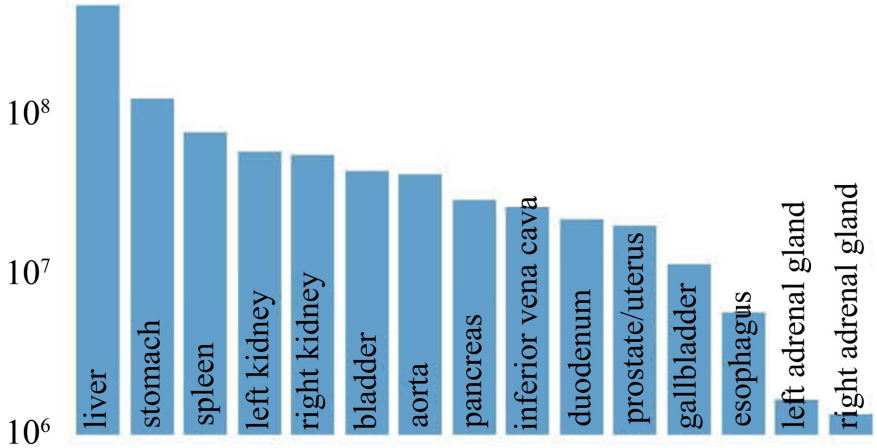

Figure 3. Number of annotated voxels per category
图 3. 每个类别的注释体素数量

3.2. 对比实验

本次实验在 U-Net 和 CoTr 模型上进行了评估，并对两种模型的四种器官的训练精度(Dice 精度)进行了比较，具体结果如表 1 所示。

Table 1. Comparison of segmentation results
表 1. 分割结果对比

模型	平均 Dice	器官 Dice			
		脾	肝	左肾	右肾
U-Net	89.06	91.09	91.61	86.36	87.18
CoTr	96.23	96.31	97.05	96.28	95.29

表 1 展示了 U-Net 和 CoTr 两种模型对四种器官的分割精度以及平均分割精度。可以看出 CoTr 在四种器官上的分割精度都要优于 U-Net 的分割精度，特别是在左肾和右肾这种比较小的目标上，CoTr 的分割精度要远远高于 U-Net 的结果。这可能是因为 Transformer 结构可以更好的捕捉上下文信息，使特征图具有更大的感受野，从而在小目标的分割上有更好的表现。

3.3. 消融实验

为了验证 CNN 编码器和 DeTrans 编码器的有效性，我们分别将 CoTr 模型与没有 CNN 编码器的 CoTr 模型，以及没有 DeTrans 编码器的 CoTr 模型的分割结果进行了比较，具体结果如表 2 所示。

Table 2. Results of ablation experiment
表 2. 消融实验结果

方法	平均 Dice	器官 Dice			
		脾	肝	左肾	右肾
CoTr without CNN 编码器	93.83	95.23	96.21	92.36	91.52
CoTr without DeTrans 编码器	93.71	94.96	95.47	92.65	91.75
CoTr	96.23	96.31	97.05	96.28	95.29

由表 2 可知与没有 CNN 编码器的 CoTr 模型相比，具有 CNN 编码器的 CoTr 模型在平均 Dice 和各器官 Dice 上的结果均处于领先地位。这可以证明 CNN 编码器的有效性，即在医学图像分割方面，混合了 CNN 结构的 Transformer 编码器比单纯的 Transformer 编码器具有更好的性能。具有 DeTrans 编码器的 CoTr 的分割性能也比不具有 DeTrans 编码器的 CoTr 的分割性能要好，这证明了 DeTrans 编码器的有效性，即混合了 Transformer 结构的 CNN 编码器比单纯的 CNN 编码器具有更强的能力来学习用于医学图像分割的有效表示，进而有助于更准确的分割。

4. 结论

在本次实验中使用了一种用于 3D 医学图像分割的 CNN 和 Transformer 的混合模型，即 CoTr。该模型具有可变形 Transformer (DeTrans)，该 Transformer 采用可变形自注意机制来降低对多尺度和高分辨率特征图的长程依赖性建模的计算和空间复杂性。在 AMOS 数据集上进行了比较实验。与基于 CNN 的 U-Net 模型相比，CoTr 的性能更加优越。除此之外还进行了消融实验，实验证明了与分别没有 CNN 编码器和 DeTrans 编码器的 CoTr 相比，原 CoTr 模型具有更好的分割效果。由对比试验和消融实验的结果可以证明结合 CNN 和 Transformer 可以提高模型的分割性能。CoTr 在保持低级别特征的细节和建模长程依赖性方面实现了平衡。

参考文献

[1] Yu, F. and Koltun, V. (2016) Multi-Scale Context Aggregation by Dilated Convolutions. *International Conference on Learning Representations* (*ICLR*), San Juan, 2-4 May 2016, 1-13.

[2] Peng, C., Zhang, X., Yu, G., Luo, G. and Sun, J. (2017). Large Kernel Matters—Improve Semantic Segmentation by Global Convolutional Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 4353-4361. https://doi.org/10.1109/cvpr.2017.189

[3] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017). Pyramid Scene Parsing Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2881-2890. https://doi.org/10.1109/cvpr.2017.660

[4] Wang, X., Girshick, R., Gupta, A. and He, K. (2018). Non-local Neural Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18-23 June 2018, 7794-7803. https://doi.org/10.1109/cvpr.2018.00813

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020) End-to-End Object Detection with Transformers. *Computer Vision—ECCV* 2020, Springer International Publishing, Cham, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[6] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. and Zhou, Y. (2021) Transunet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306