基于 Ttransformer 交错卷积的医学图像分割

摘要：传统卷积神经网络凭借强大的特征学习能力，在医学图像分析中占主导地位。之后的研究进一步融入了捕获数据间长期依赖的 Transformer 模型，然而现有的研究对二者的融合工作不充分。本项目聚焦于 3D 医学图像分割，基于 nnFormer 框架探讨如何通过融合 CNN 和 Transformer 模型进行交错卷积优化医学图像分割的准确性。本项目完成了整个模型的训练以及针对脑瘤分割任务的推理工作，复现结果与原文结果相近，但仅有两项指标高于原文（原因之一可能是原文进行了 10 次重复实验取均值），考虑通过更换预训练模型和优化模型参数，进一步提升模型性能后，在 WT、TC、ET 三类子任务上的 8 个指标中有 4 项高于原文基准模型。

关键词：Transformer，CNN，医学图像分割

1 引言

1.1 研究背景

医学图像分割领域，传统的卷积神经网络 CNNs[1]扮演了至关重要的角色，它们能够提取输入图像特征。凭借强大的特征学习能力和模式识别能力，它能够在医学图像中自动识别和分割感兴趣的区域，如肿瘤、器官、血管等,同时它的局部感知和权重共享特性，使得它在医学图像分析中占据主导地位。目前基于 CNN 的网络架构已经较为成熟，具有代表性的是 U-Net 架构[2],它是专门为医学图像分割设计的 CNN 架构，由左半部分的编码器和右半部分的解码器组成，形似字母 U。编码器负责提取图像的特征，而解码器则用于恢复图像的分辨率，使输出与输入尺寸匹配。U-Net 的关键创新在于它使用跳跃连接（skip connections），将编码器的特征直接传递到解码器的对应层，这有助于保留图像的细粒度信息，对分割精度有显著提升。

后来一些研究者，开始融入了主要应用在序列数据捕获数据间长期依赖关系的 Transformer[3]模型。该模型最初是为了自然语言处理（NLP）任务设计的，如机器翻译。大家发现如图 1-1 所示 Transformer 通过多头自注意力机制，能够捕捉到图像之间任意两个位置上的关联，更能够处理非局部的交互，提供了比较好的理解全局上下文信息的效果。通过 Transformer 这种全局视角可以看到这些器官、病变组织在整个图像中的分布，探寻其中的关系，然后也去依靠卷积更细粒度地去捕获具体的局部特征差异。基于 Transformer 模型进行医学图像分割的架构有很多，包括 TransUNet[4]、VT-Unet 等等。TransUNet 是将 Transformer 应用于医学图像分割的早期尝试之一，它在编码器中使用 Transformer 结构，并将其与经典的 U-Net 解码器相结合，以产生高质量的分割结果。下图是一个用于 3D 医学图像肿瘤分割的 Transformer 模型，它结合了 Transformer 的优点和 U-Net 的结构，特别设计用于处理 3D 医学图像，提高分割的计算效率和精度。

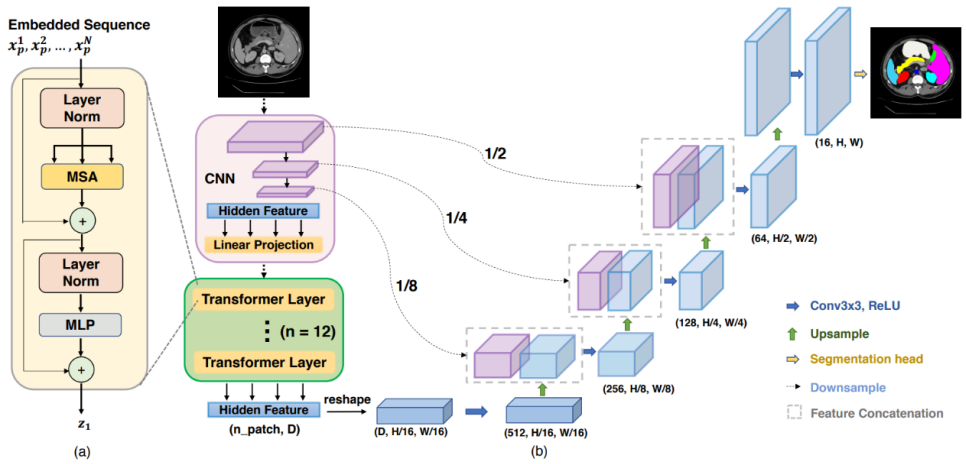

图 1-1 基于 Transformer 的医学图像分割架构

上述这些现有的一些网络结构，虽然做了卷积和 Transformer 的融合，但是研究者们要么是将卷积作为网络的主体，这可能会因一层或两层转换器不足以将长期依赖性与通常包含精确空间信息并提供分层概念的卷积表示纠缠在一起；要么就是使用 Transformer 作为分割模型的主干，丢失掉细节部分；即研究者们并没有探索适当地结合卷积和自注意力来构建最佳的医学分割网络。

1.2 项目描述

医学图像分割是医学影像分析中的一个重要环节，它涉及到从医学图像中识别并分离出特定的解剖结构或组织，这个过程对于疾病的诊断、治疗规划和手术指导具有重要的意义。医学图像分割主要目标就是确定 ROI，精确地描绘出图像中组织类型或病灶这些结构的边界。本项目是对 nnformer[5]论文复现的工作，本人完成了模型训练、推理以及优化改进工作。主要的几点贡献是：完成了整个模型的训练以及针对脑瘤分割任务的推理工作，复现结果与原文结果相近。并额外更换预训练模型以及调整模型参数进行了改进工作，改进结果比直接复现效果较好，在 WT、TC、ET 三类任务上的 8 个指标中有 4 项高于原文，而直接复现结果仅有两项高于原文。

2 模型架构与实现

2.1 Transformer 交错卷积

不同于 1.1 小节介绍的研究者们并未探索适当地结合卷积和自注意力来构建最佳的医学分割网络，nnFormer 采用了一种创新的卷积自注意力机制融合方法如图 2-1 所示来处理复杂的三维图像数据，旨在从不同尺度上捕捉图像的丰富信息。这一过程分为几个关键步骤，首先是通过卷积嵌入层进行特征提取，卷积嵌入层负责捕捉图像中的低层次特征，这些特征虽然简单，但具有高分辨率，能够精细地描绘出图像的基本组成元素，如边缘、纹理等，即在卷积嵌入层提取低层次但高分辨率的三维图像特征。然后为了进一步处理和理解这些低层次特征，引入了 Transformer 和卷积下采样层的交织使用。Transformer，以其强大的序列建模能力，被用来处理长距离依赖关系，确保即使是在大规模图像中，

也能捕捉到不同部分之间的关联性。而卷积下采样层则有助于逐步降低特征图的分辨率，从而提取更抽象、更高层次的概念。这种交织使用的方式，将长期依赖与各种尺度的高级和分层对象完全融在一起，使得模型能够同时处理局部细节和全局结构，实现了对图像特征的多层次理解。

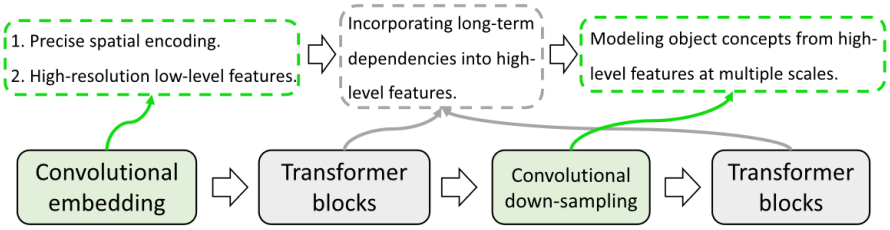

图 2-1 融合 Transformer 的交错卷积

2.2 模型整体架构

项目整体网络架构如图 2-2 所示，它保持与 U-Net 类似的 U 形，主要由编码器、瓶颈和解码器三部分组成。

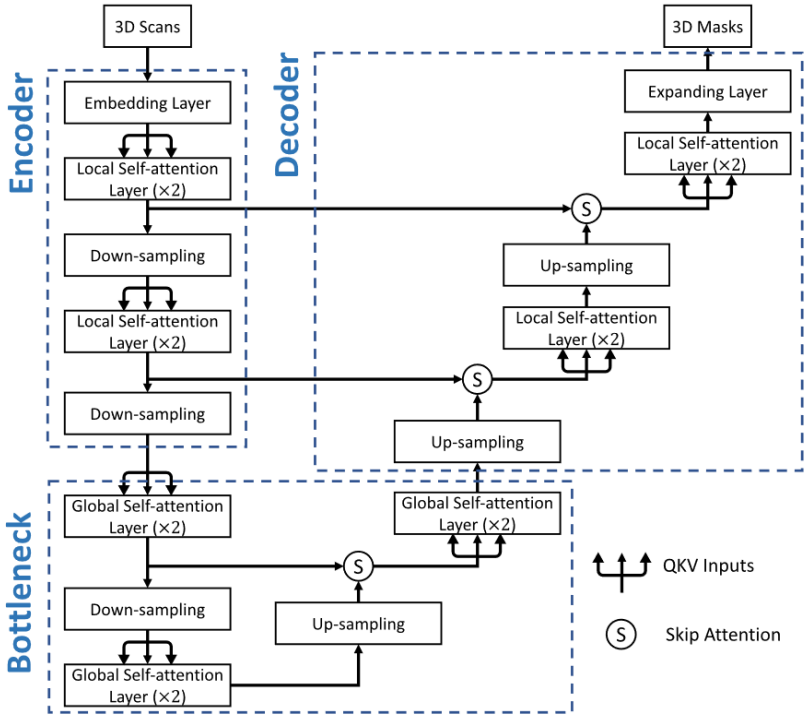

图 2-2 nnFormer 模型架构

其中编码器涉及一个嵌入层，两个局部变压器块（每个块包含两个连续的层）和两个下采样层。对称地，解码器分支包括两个变换器块、两个上采样层和用于进行掩模预测的最后一个补丁扩展层。此外，瓶颈包括一个下采样层、一个上采样层和三个全局变换器块，用于提供大的感受野来支持解码器。在编码器和解码器的相应特征金字塔之间添加跳跃连接，这有助于恢复预测中的细粒度细节。但是区别于常使用求和或串联操作的非典型跳跃连接不同，项目引入了跳跃注意力来弥合编码器和解码器之间的差距。

2.3 各模块实现

（1） Encoder 模块

编码器包括一个卷积嵌入层、和两个下采样层。目的是想去通过卷积下采样得到不同尺度的特征信息。还包括两个 3D 局部 Transformer 块（每个块包含两个连续的层）。目的是捕获细粒度的长期依赖。

Encoder 的输入是一个三维 Patch $X \in R^{H\times W\times D}$（通常从原始图像中随机裁剪），其中$H$、$W$、$D$分别表示每个输入扫描的高度、宽度和深度。在卷积嵌入层会将输入转化为高维张量$X_e \in R^{\frac{H}{4}\times\frac{W}{4}\times\frac{D}{2}\times C}$，其中$\frac{H}{4}\times\frac{W}{4}\times\frac{D}{2}$代表 Patch 的数量，$C$代表序列长度。在嵌入块中使用的卷积层应用了核大小为 3 的小卷积核，在每个卷积层之后（除了最后一个），附加一个 GELU 激活函数和一个 layer normalization 层，与大尺寸的内核相比，小的内核尺寸有助于降低计算的复杂性，同时提供同等大小的感受野。

在嵌入层之后，将高维张量X_e传递给如图 2-3 所示的 3D 局部 Transformer 块即 LV-MSA，并计算自注意力。假设 $X_{LV} \in R^{L\times C}$表示局部变压器块的输入，其中L表示输入第一维的大小，它取决于输入图像的大小和局部变换器块的层索引。C代表通道数。X_{LV}首先被重塑为$\hat{X}_{LV} \in R^{N_{LV}\times N_T\times C}$，其中$N_{LV}$是预定义的 3D 局部体积数量，$N_T = S_H \times S_w \times S_D$表示每个 volume 中的 patch 数量。如图 2-3 左侧所示，本人在每个块中进行两个连续的 transformer 层，其中第二层可以被视为第一层的移位版本即 SLV-MSA,计算过程可概括如下式（1）：

$$\hat{X}_{LV}^{l} = LV - MSA(Norm(X_{LV}^{l-1})) + X_{LV}^{l-1},$$

$$X_{LV}^{l} = MLP(Norm(\hat{X}_{LV}^{l})) + \hat{X}_{LV}^{l},$$

$$\hat{X}_{LV}^{l+1} = SLV - MSA(Norm(X_{LV}^{l})) + X_{LV}^{l},$$

$$X_{LV}^{l+1} = MLP\left(Norm\left(\hat{X}_{LV}^{l+1}\right)\right) + \hat{X}_{LV}^{l+1}. \quad (1)$$

其中l代表层索引,MLP 是多层感知器的缩写,每个三维局部体中的查询键值(QKV) 注意力通过公式(2)计算：

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad (2)$$

其中Q，K，V表示查询、键和值矩阵，B是相对位置编码。

Transformer 层之后，进入积下采样层，卷积下采样产生了层次化的表示，有助于在多个尺度上对物体概念进行建模，下采样可以大大降低 GPU 显存的消耗，多次下采样可以建立多尺度的特征金字塔结构。下采样层涉及跨步卷积运算，跨步在所有维度上都设置为 2，其中相对于特定维度的步长设置为 1。

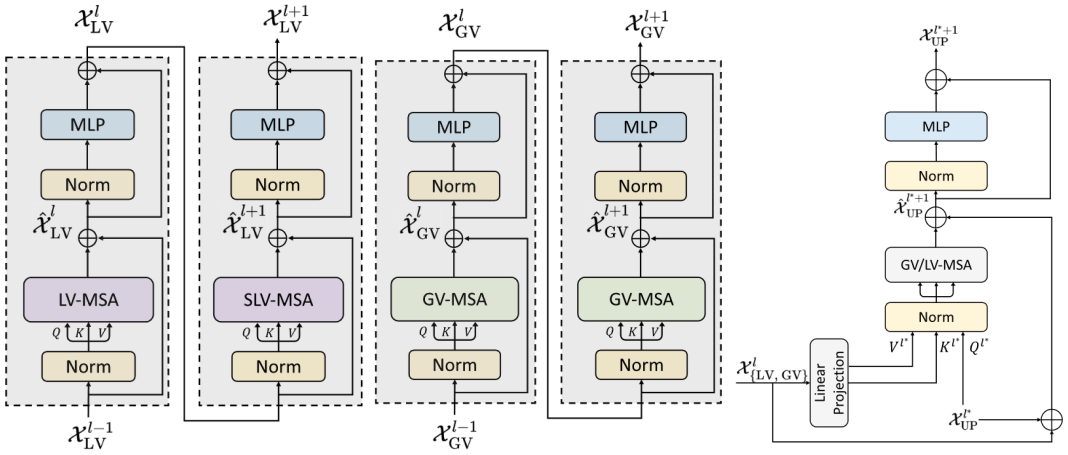

图 2-3 LV-MSA（左）GV-MSA（中）跳跃注意力（右）

（2） Bottleneck 模块

Bottleneck 层包括一个下采样层、一个上采样层和三个全局 Transformer 块，用于提供大的感受野以支持解码器即看的范围更大，捕获全局信息的长期依赖。

在瓶颈中，经过几个下采样层后，{H、W、D}已经变得小得多，使得它们的乘积，具有与{S_H、S_W、S_D}相似的大小，这为应用 GV-MSA 创造了条件，与 LV-MSA 相比，GV-MSA 能够提供更大的感受野，项目中使用的三个全局变压器块（即六个 GV-MSA 层）如图 2-3 中侧所示。

（3） Encoder 模块

与 Decoder 对称的是，解码器包括两个 3D 局部 Transformer 块，两个反卷积上采样层。目的是重建图像。然后在编码器和解码器的相应特征金字塔之间以对称的方式添加了跳跃注意力：将深层的特征与浅层的特征连接起来，以便在上采样阶段恢复丢失的细节特征。

在如图 2-3 右侧所示的跳跃注意力中，编码器第l个 Transformer 块的输出，即$X^l_{\{LV,GV\}}$经过线性投影被转换并分割成式（3）所示键矩阵K^{l*}和一个值矩阵V^{l*}：

$$K^{l*}, V^{l*} = LP\left(X^l_{\{LV,GV\}}\right) \tag{3}$$

其中LP代表线性投影，然后，可以在 Decoder 中对Q^{l*}, K^{l*}, V^{l*}进行 LV/GV-MSA，即下式（4）：

$$Attention(Q^{l*}, K^{l*}, V^{l*}) = softmax\left(\frac{Q^{l*}(K^{l*})^T}{\sqrt{d_k^{l*}}} + B^{l*}\right)V^{l*} \tag{4}$$

3 数据处理

3.1 数据集介绍及可视化

原文对三个数据集（任务）进行了实验：医学分割十项全能 (MSD)中的脑肿瘤分割任务[6]、Synapse 多器官分割和自动心脏诊断挑战赛 (ACDC) 。对于每个实验，作者使用相同的训练、验证、测试分组和不同的随机种子值重复十次，并报告其平均结果，还计算 p 值来证明 nnFormer 的重要性。本人的工作中仅针

对脑肿瘤分割任务数据集进行实验，使用与论文中相同的训练、验证、测试分组但不做十次重复实验，仅报告一次结果作为 nnFormer 的印证。

脑瘤分割任务使用的是 MRI 扫描图像：该任务由 484 个 MRI 图像组成，每个图像包含四个通道， FLAIR、T1w、T1gd 和 T2w。这些数据包含 2016 年和 2017 年脑肿瘤分割挑战中使用的数据的子集。对应的目标 ROI 是三个肿瘤亚区域。为了后文复现时与 UNETR[7]中报告的结果保持一致，在将 nnFormer 与基于 Transformer 的模型进行比较时，本人展示了如图 3-1 所示的目标 ROI 三个肿瘤子区域，即水肿（ED）、非增强型肿瘤（NET）和增强型肿瘤（ET）的实验结果。对于数据的分割，遵循 UNETR 的指示，其中训练/验证/测试集的比例分别为 80%、15%和 5%。

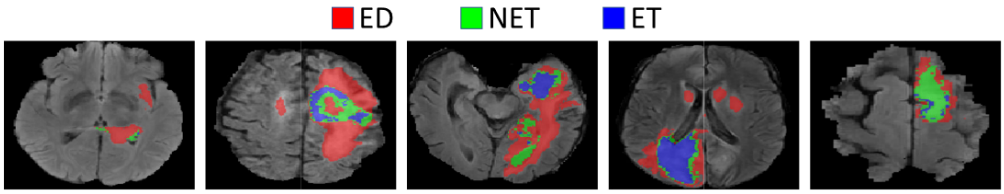

图 3-1 脑瘤分割任务数据集

3.2 数据预处理

数据预处理时，使用作者提供的源码，所有图像将首先重新采样到相同的目标间距，按照给定的顺序应用旋转、缩放、高斯噪声、高斯模糊、亮度和对比度调整、低分辨率模拟、伽马增强和镜像等增强功能。同时还在训练阶段添加深度监督,其中解码器的每个阶段输出都会传递到最终的扩展块，其中将应用交叉熵损失和骰子损失。在实践中，给定一个典型阶段的预测，对地面真实分割掩模进行下采样以匹配推理，数据预处理结果会在后文介绍。

4 论文复现与改进

4.1 复现工作

（1）环境配置

本文的复现工作是基于 Linux 系统、VScode、Python3.8（ubuntu）、PyTorch 2.0.0 、GPU RTXA6000(50GB)*2 和 CUDA 11.8 环境，并在终端中创建如题 4-1 所示虚拟环境 nnFormer。

```
# conda environments:
#
base                  *  /home/wqq/anaconda3
MST                      /home/wqq/anaconda3/envs/MST
nnFormer                 /home/wqq/anaconda3/envs/nnFormer
```

图 4-1 配置虚拟环境

（2）网络参数设计

对网络的学习率设置为 0.01，使用默认优化器 SGD 同时将动量设置为 0.99，权重衰减设置为 3×10^{-5}，训练 epoch 的数量为 1000，其中一个 epoch 包含 250 次迭代。 脑瘤分割任务上不同编码器阶段使用的多头自注意力的头数为[3,6,12,24]，卷积嵌入层 patch 大小为[4,4,4]。

（3）数据集设置

首先需要下载数据集[8]，之后创建名为 DATASET 文件夹，并在其中依次创建三个文件夹 nnFormer_preprocessed、nnFormer_raw、nnFormer_trained_models。在 nnFormer_raw 目录下依次创建目录。将下载的数据集放置在命名为 Task_03_tumor 的目录内，之后进入.bashrc 文件，添加环境变量并激活文件。

在数据集预处理阶段，首先对移动到 Task_03_tumor 目录中的数据集进行分割，并转换数据集，使其可以被 nnFormer 识别，在终端运行命令：nnFormer_convert_decathlon_task -i /自己目录路径/Task03_tumor。之后运行命令：nnFormer_plan_and_preprocess -t 3 进行数据集插值操作，完成效果如图 4-2 所示。

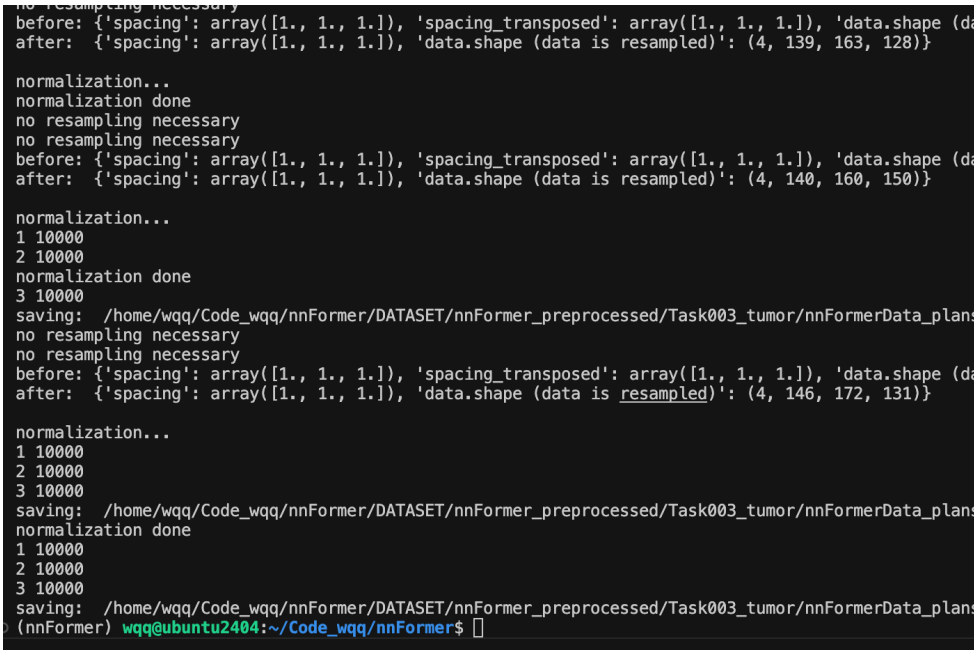

图 4-2 数据预处理

（4）模型训练

处理完成的数据集就可用于训练了，首先按照下图4-3修改tran_inference.sh中的代码，即修改目录路径，并注释掉推理部分代码。

```
if ${train}
then

    cd /home/wqq/Code_wqq/nnFormer/nnformer
    CUDA_VISIBLE_DEVICES=${cuda} nnFormer_train 3d_fullres nnFormerTrainerV2_${n
fi
```

图 4-3 训练命令修改

之后执行训练代码：bash train_inference.sh -c 0 -n nnformer_tumor -t 3。训练时长需要两天左右，单个 epoch 耗时三分钟左右如图 4-4 所示，我的环境总计运行了 50 多小时。

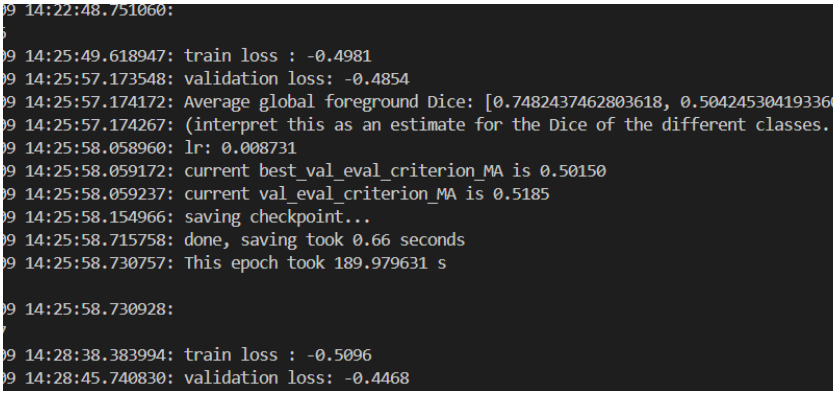

图 4-4 模型训练过程图

（5）复现结果

应用预训练结束的模型进行推理，在推理阶段按照下图 4-5 注释掉训练部分的代码，并修改推理部分代码执行推理命令：bash train_inference.sh -c 0 -n nnformer_tumor -t 3。

```
if ${predict}
then


    cd /home/wqq/Code_wqq/nnFormer/DATASET/nnFormer_raw/nnFormer_raw_data/Task00
    CUDA_VISIBLE_DEVICES=${cuda} nnFormer_predict -i imagesTs -o inferTs/${name}
    python inference_tumor.py ${name}
```

图 4-5 推理命令修改

推理结果的评价用到的指标是 HD95 和 DSC，前者更侧重于评估分割边界的最大偏差，值低比较好。后者关注分割结果和真实标签的整体一致性，得分高比较好。推理结果如图 4-6 所示，得到了 WT、ET、TC 所有类别的 HD95 分数和 DSC 分数。

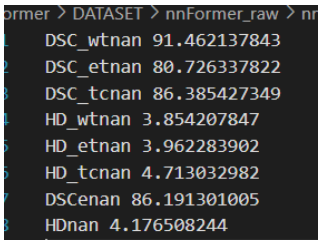

图 4-6 推理结果

4.2 改进优化工作

对 nnFormer 工作复现完成之后，考虑从以下四个方面进行实验改进优化。最终确定利用前者进行改进。

（1）使用医学预训练模型—加速模型收敛

在模型训练的初始阶段，利用医学图像领域预训练的模型参数或许能够加速模型收敛，减少训练所需的时间和资源。

（2）模型参数微调

对于 nnFormer 模型，考虑调整学习率、优化器参数、训练周期等超参数来

优化模型。例如，可以使用学习率衰减策略，随着训练的进行逐渐减小学习率，以避免梯度爆炸或过拟合现象。

（3）多模态数据融合

将多种成像模态的数据融合到模型中，可以提供更丰富的信息，有助于提高模型的分割精度。例如，同时使用 FLAIR、T1w、T1gd 和 T2w 等不同类型的 MRI 图像可以捕捉到脑肿瘤的不同属性，进而提高分割的准确性。

```
                deterministic, fp16)
self.max_num_epochs = 1000
self.initial_lr = 5e-4#降低学习率促进更稳定的训练过程
self.deep_supervision_scales = [1.0,0.5,0.25]#开启深度监督添加多尺度损失
self.ds_loss_weights = None
self.pin_memory = True
self.load_pretrain_weight=True#使用预训练权重加快模型收敛

self.load_plans_file()

if len(self.plans['plans_per_stage'])==2:
    Stage=1
else:
    Stage=0

self.crop_size=self.plans['plans_per_stage'][Stage]['patch_size']
self.input_channels=self.plans['num_modalities']
self.num_classes=self.plans['num_classes'] + 1
self.conv_op=nn.Conv3d

self.embedding_dim=96
self.depths=[2, 2, 2, 2]
self.num_heads=[4, 8, 16, 32]#增加头数以提高模型捕获复杂关系的能力
self.embedding_patch_size=[4,4,4]
self.window_size=[4,4,8,4]

self.deep_supervision=False
initialize(self, training=True, force_load_plans=False):
```

图 4-7 模型改进（参数调整及预训练模型）

（4）模型结构优化

尝试对 Transformer 和卷积层的交织方式做进一步的优化，如改变 Transformer 块的数量或卷积层的结构，探索更适合当前数据特性的网络架构。

由于想法三和四较为复杂，本人仅针对想法一和二进行了实现。选择了大型医学图像数据集预训练的模型作为初始化参数来帮助模型更快地学习到医学图像的特征表示，提高分割精度，更改了学习率，调整了多头注意力数量，具体参数调整如上图 4-7 所示。

4.3 效果分析

nnFormer 的工作将其与被认为是最强大的三维医学图像分割模型之一的 nnUNet 进行比较。可以看下如图 4-8 所示的结果。从特定类别的 HD95 结果来看，nnFormer 在 16 个类别中的 11 个方面优于 nnUNet。在特定类别的 DSC 中，nnFormer 在 16 个类别中的 9 个方面优于 nnUNet。因此，在 HD95 下，nnFormer 似乎更有优势，这意味着 nnFormer 可能更好地划定对象边界。

	Average		WT		ET		TC		ED		NET	
	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DS
[42]	4.60	81.87	**3.64**	**91.99**	4.06	80.97	4.91	85.35	4.26	**84.39**	6.14	66
rmer	**4.42**	**82.02**	3.80	91.26	**3.87**	**81.80**	**4.49**	**86.02**	**4.17**	83.76	**5.76**	**67**
	< 1e-2 (HD95), 8.8e-2 (DSC)											
	4.09	**82.65**	**3.43**	**92.33**	**3.69**	**82.26**	**4.17**	**86.14**	**3.92**	**84.95**	**5.23**	**67**

(a) Brain tumor segmentation

Average		Aotra		Gallbladder		Kidney (Left)		Kidney (Right)		Liver		Pancreas		Spleen		St
HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95
10.78	**86.99**	**5.91**	**93.01**	15.19	**71.77**	18.60	85.57	**6.44**	**88.18**	**1.62**	**97.23**	4.52	83.01	24.34	**91.86**	9.5
10.63	86.57	11.38	92.04	**11.55**	70.17	**18.09**	**86.57**	12.76	86.25	2.00	96.84	**3.72**	**83.35**	**16.92**	90.51	**8.5**
2e-2 (HD95), 7.7e-2 (DSC)																
7.70	**87.51**	**5.90**	**93.11**	**8.63**	**72.08**	18.42	86.20	8.56	87.76	1.63	97.20	**3.64**	**84.21**	**9.42**	**91.94**	**5.4**

(b) Multi-organ segmentation (Synapse)

Methods	Average		RV		Myo		LV	
	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑
nnUNet* [42]	1.15	91.61	1.31	90.24	1.06	89.24	1.09	95.36
Our nnFormer	**1.12**	**92.06**	**1.23**	**90.94**	**1.04**	**89.58**	1.09	**95.65**
P-values	2e-2 (HD95), < 1e-2 (DSC)							
nnAvg	**1.10**	**92.15**	**1.19**	**91.03**	1.04	**89.75**	**1.06**	**95.68**

(c) Automated cardiac diagnosis (ACDC)

图 4-8 nnFormer 在三项数据集对比实验

Methods	Average		WT		ET		TC	
	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑
SETR NUP [34]	13.78	63.7	14.419	69.7	11.72	54.4	15.19	66.9
SETR PUP [34]	14.01	63.8	15.245	69.6	11.76	54.9	15.023	67.0
SETR MLA [34]	13.49	63.9	15.503	69.8	10.24	55.4	14.72	66.5
TransUNet [11]	12.98	64.4	14.03	70.6	10.42	54.2	14.5	68.4
TransBTS [25]	9.65	69.6	10.03	77.9	9.97	57.4	8.95	73.5
CoTr w/o CNN encoder [24]	11.22	64.4	11.49	71.2	9.59	52.3	12.58	69.8
CoTr [24]	9.70	68.3	9.20	74.6	9.45	55.7	10.45	74.8
UNETR [33]	8.82	71.1	8.27	78.9	9.35	58.5	8.85	76.1
Swin UNETR* [35, 36]	6.43	84.4	6.61	89.4	7.38	80.2	5.29	83.5
Our nnFormer	**4.05**	**86.4**	**3.80**	**91.3**	**3.87**	**81.8**	**4.49**	**86.0**
P-values	< 1e-2 (HD95), < 1e-2 (DSC)							

图 4-9 原文脑瘤分割任务结果

本人复现结果如表 4-1 所示，在 DSC 得分上对于 WT、TC 的效果比文中报告的结果要好，但在 HD95 指标上，均未达到文中如图 4-9 效果。改进之后的推理得到的评分中，HD95 有两项优于原文，总体上得到了较好的效果（个别指标与原文偏差大的原因，个人认为是原文做了十次取平均的原因）。

表 4- 1 本人复现结果

Methods	WT		ET		TC		Average	
	HD95	DSC	HD95	DSC	HD95	DSC	HD95	DSC
个人复现	3.85	91.46	3.96	80.73	4.71	86.39	4.18	86.19
优化改进	3.76	91.28	3.92	81.02	4.81	86.15	4.16	86.15

5 总结

本项目聚焦于 3D 医学图像分割，基于 nnFormer 框架探讨通过融合 CNN 和 Transformer 模型进行交错卷积优化医学图像分割的准确性。模型架构设计了三个模块分别为编码器、瓶颈和解码器。编码器包括一个卷积嵌入层、和两个下采样层。目的是通过卷积下采样得到不同尺度的特征信息，还包括两个 3D 局部 Transformer 块，目的是捕获细粒度的长期依赖。Bottleneck 层包括一个下采样层、一个上采样层和三个全局 Transformer 块，用于提供大的感受野以支持解码器。与 Encoder 对称的是解码器包括两个 3D 局部 Transformer 块，两个反卷积上采样层。目的是重建图像。在编码器和解码器的相应特征金字塔之间以对称的方式添加了跳跃注意力：将深层的特征与浅层的特征连接起来，以便在上采样阶段恢复丢失的细节特征。

本项目完成了整个模型的训练以及针对脑瘤分割任务的推理工作，复现结果与原文结果相近。还通过更换预训练模型和优化模型参数，进一步提升了模型性能。实验结果显示，改进后的 nnFormer 在 WT、TC、ET 三类子任务上的多个评估指标上超越了原文基准模型（8 个指标中有 4 项高于原文），而直接复现结果仅有两项高于原文。

6 参考文献

[1] Alzubaidi L, Zhang J, Humaidi A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. Journal of big Data, 2021, 8: 1-74.

[2] O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Medical Image Computing and Computer-Assisted Intervention, pp.234- 241, 2015.

[3] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey[J]. ACM computing surveys (CSUR), 2022, 54(10s): 1-41.

[4] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, arXiv:2102.04306.

[5] Zhou, Hong-Yu, et al. "nnformer: Volumetric medical image segmentation via a 3d transformer." IEEE Transactions on Image Processing (2023).

[6] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, “DS-TransUNet: Dual Swin transformer U-Net for medical image segmentation,” 2021, arXiv:2106.06716.

[7] A. Hatamizadeh et al., “UNETR: Transformers for 3D medical image segmentation,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2022, pp. 1748–1758.

[8] https://drive.google.com/file/d/1A2IU8Sgea1h3fYLpYtFb2v7NYdMjvEhU/view