Ladder Fine-tuning approach for SAM integrating complementary network

Shurong Chai[1], Rahul Kumar Jain[1], Shiyu Teng[1], Jiaqing Liu[1], Yinhao Li[1], Tomoko Tateyama[2] and *Yen-wei Chen[1]

最近，引入了基于计算机视觉领域的各种任务的基础模型。这些模型，比如Segment Anything Model (SAM)，是使用大规模数据集进行训练的通用模型。目前，正在进行的研究聚焦于探索如何有效地利用这些通用模型应用于特定领域，比如医学影像。

然而，在医学影像领域，由于隐私问题和其他因素导致缺乏训练样本，这对将这些通用模型应用于医学图像分割任务构成了重大挑战。

为了解决这个问题，有效地微调这些模型对于确保它们的最佳利用至关重要。在本研究中提出结合一个互补的卷积神经网络（CNN）和标准的SAM网络进行医学图像分割。为了减轻对大型基础模型进行精细调整的负担并实现成本高效的训练方案，本文仅集中于对额外的CNN网络和SAM解码器部分进行微调。这种策略显著减少了训练时间，并在公开可用的数据集上取得了竞争性的结果。

代码：https://github.com/11yxk/SAM-LST

1、简介

医学图像分割在医疗保健领域中起着至关重要的作用。它旨在使用各种医学成像模态（如X射线、CT扫描、MRI扫描或超声图像）对肝脏、脑部和病变等各种人体器官进行分割。因此，它在诊断、治疗计划和治疗后监测方面对临床医生有很大帮助。

在过去的十年中，卷积神经网络（CNN）在计算机视觉任务中变得流行起来。最近，Long等人提出了全卷积网络（FCN）。这种方法通过用卷积层替换全连接层，使得能够处理任意大小的输入图像并生成分割结果。U-Net是由Ronneberger等人开发的用于医学图像分割的最广泛使用的架构。它包括一个编码器和一个解码器，并且在相应层之间有Shortcut以保留重要的特征。编码器路径对输入图像进行下采样，同时捕捉High-Level特征。而解码器路径则对特征图进行上采样以预测分割结果。Zhou等人通过引入嵌套的Shortcut方案扩展了U-Net架构，这允许捕捉多尺度的上下文信息和更好地集成不同Level的特征。Chen等人提出了Deeplab系列模型，其中包括空洞卷积操作和全连接的条件随机场的概念。

最近，Transformer已经被引入到计算机视觉（CV）领域，它最初是为自然语言处理（NLP）而设计的。与传统的CNN架构相比，Transformer可以捕捉到远程依赖关系。多索维茨基等提出了基于自注意力机制的图像分类视觉Transformer（ViT）。随后，Chen等人提出了使用ViT进行分割任务的TransUNet。TransUNet联合利用CNN和ViT从输入图像中获取局部和全局上下文特征。Tang等人提出了采用ViT模型作为特征提取的主要编码器的Swin UNETR。Zhou等人提出了一个纯Transformer框架，它在编码器和解码器部分都使用ViT。Cao等人提出了采用双Transformer架构进行分割任务的双-unet。

目前，基础模型在自然语言处理领域已经证明了其能力。目前，SAM被引入用于各种计算机视觉任务。在SAM中，使用基础模型的prompt学习的概念可以在看不见的图像上执行多个任务。它允许通过有效的prompt工程将zero-shot转移到各种任务中。虽然将SAM模型直接应用于特定领域的任务，如医学图像分割，通常不能产生令人满意的性能。尽管SAM使用的是超过1100万张图像和10亿张GT Mask，但由于医学图像与真实图像相比的独特特征，其在医学图像分割中的应用带来了挑战。此外，医疗数据的缺乏也是调整SAM的一个主要问题。因此，对医学图像数据集的SAM进行有效的微调是非常必要的。

如今，已经引入了各种微调方法来优化在不同领域中的SAM。一些方法对SAM网络进行基于适配器的微调。然而，这些基于适配器的方法通常需要大量的训练模型的工作和资源成本。与以往的研究不同，本文的工作通过在SAM架构中结合一个额外的CNN作为补充编码器，引入了一种新颖的方法。

本文的方法受到了用于Transformer的Ladder-Side Tuning (LST)网络的启发。本文提出的方法能够灵活地集成额外的网络，同时避免在整个大模型（即SAM编码器）上进行反向传播，从而

加快了训练速度并降低了资源成本。额外的CNN网络可以根据具体任务需求轻松替换为其他设计，包括基于Transformer的设计。

本文将预训练的ResNet18作为额外的网络进行了融合。在训练过程中，只有额外的CNN和解码器部分的参数进行微调，而保持原始的SAM编码器参数不变。

本文的贡献可以总结如下：

1. 本文提出了在医学图像分割任务中，结合额外的CNN进行SAM微调的方法；
2. 所提出的方法在设计额外网络时提供了灵活性，同时通过避免在整个模型上进行反向传播来最大程度地减少资源成本；
3. 在一个公开可用的多器官分割数据集上，本文的方法在不使用任何prompt的情况下取得了与最先进方法相媲美的结果。

2、相关工作

2.1 医学图像分割

精确可靠的医学图像分割对于辅助医学诊断至关重要。在过去的几年中，已经提出了许多分割方法。特别是基于CNN的网络在这个任务中取得了显著的成功。最近，一些基于Transformer的High-Level网络也被提出来，在这个任务中取得了新的里程碑。尽管在医学图像分割方面取得了显著进展，但由于数据有限和需要临床专家对数据进行注释的要求等因素，它仍然是一个具有挑战性的任务。这些因素通常导致模型的泛化能力较差。

2.2 基础模型

基础模型是指在广泛数据上训练并可适应各种下游任务的模型。这种范式通常包含一些其他技术，比如自监督学习、迁移学习和prompt学习。一个基础模型的例子是生成式预训练 Transformer (GPT) 系列，这些模型在来自各种来源的大量文本数据上进行了预训练。这些模型在自然语言处理 (NLP) 的进展中做出了重大贡献。

具体而言，GPT-3是其中一个参数达到1750亿的大型语言模型 (LLM)，可以应用于广泛的任务，包括翻译、问答和完形填空等。另一个值得注意的工作是对比语言图像预训练 (CLIP)，它使用了一组大规模的带有图像和相应文本描述的数据集。CLIP能够根据给定的文本prompt有效地检索图像，这在图像分类和图像生成等领域有许多应用。这些基础模型已经取得了最先进的性能。它们在各个领域的发展方向广阔。

2.3 Parameter-Efficient Fine-Tuning

尽管基础模型取得了显著的成就，但它们仍然面临一些限制，比如需要大量标记数据进行训练和庞大的计算资源需求，这归因于它们巨大的参数数量。

为了降低巨大的计算成本，引入了参数高效微调（PEFT）的方法，通过训练现有模型中的一小部分参数或在架构中训练新添加的参数。Houlsby等人提出在原始基础模型中添加一个小的子网络，称为“适配器”。Lester等人提出在原始模型输入之前添加一个可训练的张量。Sung等人引入了一种新颖的阶梯侧调整（LST）范式，仅对原始模型旁边嵌入的一个小型Transformer网络进行微调。

在这种架构设计中，只更新新添加网络的参数以节省计算成本。Ben-Zaken等人建议仅训练原始网络的偏置，这也是一种简单而有效的方法。总体而言，基于PEFT的方法对GPU友好，在有限的计算资源下可以应用基础模型于各种下游任务中。

3、本文方法

3.1 Segment Anything Model

Segment Anything Model（SAM）是基础模型在分割任务中的首次尝试。

SAM由3个组件构成：

- image编码器
- prompt编码器
- mask解码器

image编码器采用了MAE 预训练的ViT网络来提取图像特征。

prompt编码器支持4种类型的prompt输入：点、框、文本和Mask。点和框通过位置编码进行嵌入，而文本则使用CLIP中的文本编码器进行嵌入。Mask则使用卷积操作进行嵌入。

Mask解码器以轻量级的方式将图像嵌入和prompt嵌入进行映射。这两种嵌入类型通过交叉注意力模块进行交互，其中一个嵌入作为query向量，而另一个嵌入作为key向量和value向量。最后，使用转置卷积来上采样特征。Mask解码器具有生成多个结果的能力，因为提供的prompt可能存在歧义。默认输出数量设置为三个。

值得一提的是，image编码器只需对每个输入图像提取一次图像特征。之后，轻量级的prompt编码器和Mask解码器可以根据不同的输入prompt在Web浏览器中与用户实时交互。

SAM使用超过1100万张图像和10亿个Mask进行训练。实验结果表明其出色的零样本迁移能力。正如其名称所暗示的，该模型几乎可以分割任何东西，甚至是以前未见过的情况（未知的测试样本）。

3.2 Ladder-Side Tuning with SAM

在这里，本文描述了本文提出的方案和集成网络。概述如图1所示。为了有效地应用于医学图像分割任务的SAM模型，本文建议添加一个轻量级的侧网络，同时避免通过整个SAM模型进行反向传播。

本文只更新SAM解码器和集成CNN网络的参数，以在医学数据集上微调SAM。形式上，给定输入样本$x \in R^{H \times W \times C}$，其中$H$ 、 W 、 C表示高度、宽度和通道数。输入图像首先同时输入到SAM image编码器和CNN编码器中：

CNN编码器的架构如图2所示。设计遵循ResNet18，其中包括Shortcut连接。然而，为了生成与SAM编码器相同大小的特征图，网络被修改为仅使用13层，而不是全部使用18层（ResNet18包含18个卷积层）。为了确定从SAM和集成CNN中提取的特征的重要性，本文提出在组合这两个提取的特征图时加入一个可学习的门（权重参数）α：

3.3 损失函数

本文使用交叉熵损失和Dice损失的组合来对网络进行微调。

其中λ是一个超参数，根据本文的实验将其值设置为0.8。

4、实验

4.1 数据集

本文使用Synapse数据集进行评估，这是MICCAI 2015多图谱腹部标记挑战的一个公开可用的多器官分割数据集。它包括30个腹部CT扫描。根据之前的工作，总共使用18个案例进行训练，12个案例用于测试。本文以Dice相似系数（DSC）和95% Hausdorff距离（HD95）的指标报告在8个腹部器官（即主动脉、胆囊、脾脏、左肾、右肾、肝脏、胰腺、胃）上的结果。

4.2 实施细节

输入图像的分辨率设置为224×224。本文使用随机旋转和翻转操作进行数据增强。本文使用ViT-B SAM模型作为基础Backbone模型。本文不微调SAM编码器和prompt编码器，而只微调SAM解码器的“输出上采样”部分，以避免过拟合。集成的CNN编码器在PyTorch Torchvision库提供的ImageNet上进行了预训练。该框架使用批量大小为24的Adam优化器进行200个epochs训练。学习率设置为0.001。在前250个迭代中采用了warmup策略。实验使用了2个RTX 3090显卡进行。

4.3 实验结果

表I报告了实验结果，并与其他最先进的方法进行了比较。本文提出的方法实现了79.45%的DSC和35.35mm的HD95分数。本文还观察到可学习权重参数的值为0.44。本文的方法在超过大多数最先进方法的同时取得了竞争性的得分。

图3展示了一些分割结果。然而，集成CNN编码器和可学习权重参数的设计可以进行修改，以分析和评估所提出方法的性能。本文相信利用Transformer或其他有效的网络设计将会获得更高的性能。在未来，本文将探索更先进的设计选择，以达到最佳结果。

4.4 消融实验

本文进行了消融实验来评估将CNN编码器与SAM编码器集成的有效性。

表II表明，SAM模型在没有对医学图像进行微调的情况下，仅实现了1.73%的Dice分数。需要注意的是，在此训练和评估过程中没有使用prompt，因此由于直接应用了通用模型，得分较

低。通过对整个SAM应用微调方法，准确性提高到58.97的Dice分数。同样，当将CNN编码器与SAM解码器模块相结合时，性能保持在78.05的Dice分数。这凸显了有效微调方法的必要性。

然而，通过将CNN编码器与SAM网络集成并利用可学习的门（权重参数），准确性显著提高至79.45的Dice分数。此外，本文还观察到训练时间显著减少，与其他微调方法相比，减少了约30%到40%。本文提出的方法在资源利用方面更具成本效益。

5、参考

[1].Ladder Fine-tuning approach for SAM integrating complementary network.

6、推荐阅读

南科大提出ORCTrack | 解决DeepSORT等跟踪方法的遮挡问题，即插即用真的很香

CSUNet | 完美缝合Transformer和CNN，性能达到UNet家族的巅峰！

InstructionGPT-4 | 200个数据集微调，源于MiniGPT-4又高于MiniGPT-4

AI视界引擎
「AI视界引擎」公众号不仅致力于分享AI视觉与大语言模型的前沿科技，还将成为连接业...
30篇原创内容

公众号

点击上方卡片，关注**「AI视界引擎」**公众号

收录于合集 #大模型 4

上一篇

下一篇

阅读 1261