文章编号：2095-2163(2023)10-0159-07 中图分类号：TP391 文献标志码：A

基于视觉 Transformer 的多级特征聚合图像语义分割方法

孔玲君[1]，郑斌军[2]
（1 上海出版印刷高等专科学校，上海 200093；2 上海理工大学 出版印刷与艺术设计学院，上海 200093）

摘 要：针对传统卷积神经网络在图像语义分割领域进行特征提取时未能充分利用上下文信息的问题，提出一种基于视觉 Transformer 的多级特征聚合图像语义分割方法。首先，将输入图像分割成一系列切片进行线性投影，并加入可学习的位置嵌入，获得编码输入序列；通过一个基于视觉 Transformer 的编码器，将图像编码为一系列补丁，从而在整个网络中建模全局上下文。Transformer 编码器可与一个简单的线性解码器组合来获得优秀的效果，通过多级特征聚合解码器能进一步提升性能。大量实验表明，所提出的方法能够有效建模全局上下文信息，以进行图像特征提取。实验在 3 个公开数据集（ADE20K（49.97% mIoU）、Pascal Context（55.43% mIoU）、Cityscapes（82.03% mIoU））的语义分割任务中达到了良好的分割精度。设计的消融实验结果也充分证明了所提方法的有效性，能够更好地运用在高精度的图像语义分割领域。

关键词：语义分割；自注意力机制；特征聚合；视觉 Transformer

Multi-level feature aggregation with vision transformer for semantic segmentation

KONG Lingjun[1], ZHENG Binjun[2]
(1 Shanghai Publishing and Printing College, Shanghai 200093, China;
2 School of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 Aiming at the problem that the traditional convolutional neural network cannot make full use of the context information when extracting features in the field of image semantic segmentation, a multi-level feature aggregation image semantic segmentation method based on Vision Transformer is proposed. First, the input image is divided into a series of slices, linear projection is performed, and a learnable position embedding is added to obtain the coded input sequence. Through a transformer-based encoder, the image is encoded into a series of patches so as to model the global context in the entire network. This encoder can be combined with a simple linear decoder to obtain excellent results, and the performance can be further improved through multi-level feature aggregation decoder. A large number of experiments show that the proposed method can effectively model the global context information for image feature extraction, and achieves good segmentation accuracy in the semantic segmentation tasks of three public datasets ADE20K (49.97% mIoU), Pascal Context (55.43% mIoU), and Cityscapes (82.03% mIoU). The ablation experiments fully prove the effectiveness of the proposed method, which can be better used in the field of high-precision image semantic segmentation.

【Key words】 semantic segmentation; self-attention mechanism; feature aggregation; vision transformer

0 引 言

语义分割是计算机视觉领域的一个重要的研究任务，具有广泛的应用，如自动驾驶、视频监控、增强现实、机器人技术等等。语义分割通过给图像的每个像素分配语义标签，进而为目标任务提供高级图像表示，如在自动驾驶场景中识别行人和车辆以进行规避。Long 等人[1]开创性地使用完全卷积网络（Full Convolutional Network，FCN）进行图像语义分割任务，并取得良好的效果，这激发了许多后续的工作，并成为语义分割的主要范式。

图像分类与语义分割有着密切的联系，许多先进的语义分割框架是在 ImageNet 上流行的图像分类体系结构的变种。因此，主干框架设计一直是语义分割的重要活跃领域。从早期的 VGG[2]到具有更深层、更强大的主干方法，主干网络的进步极大地推动了语义分割性能的提升。通过可学习的堆叠卷积，可以捕获语义丰富的信息。然而，卷积滤波器的

基金项目：上海市一流院校建设项目（ylyx2022-3）。

作者简介：孔玲君（1972-），女，博士，教授，硕士生导师，主要研究方向：图文信息处理与色彩再现、数字印刷及质量评价；郑斌军（1997-），男，硕士研究生，主要研究方向：数字图像处理、计算机视觉和深度学习。

通讯作者：孔玲君 Email：908641376@qq.com

收稿日期：2022-11-03

局部性质限制了对图像中的全局信息的分享,但这些信息对图像分割十分重要。为了避免这个问题,Fisher 等人[3]引入了扩张卷积,通过在内核上“膨胀”空洞来增加感受野;Chen 等人[4]更进一步地使用具有空洞卷积和空洞空间金字塔池化进行特征聚合,扩大卷积网络的感受野并获得多尺度的特征。

自 Transformer 网络在自然语言领域取得巨大成功后,研究人员开始尝试将 Transformer 网络引入视觉任务中,Dosovitskiy 等人[5]提出了用于图像分类的视觉 Transformer(Vision Transformer,VIT),按照 NLP 中的转换器设计,把原始图像分割成多个切片,展平成序列,输入到标准的 Transformer 网络中,最后使用全连接层对图片进行分类,在 ImageNet 上获得了令人印象深刻的性能表现。VIT 虽然拥有良好的性能,但是也存在一些不足,如:需要庞大的训练数据集;对于高分辨率图像,计算成本高等。为了突破上述局限,Hugo 等人[6]提出了一种基于蒸馏的训练策略 Deit,仅使用 120 万张图像就可实现高效训练,并取得良好的表现。Wang 等人[7]提出一种用于密集预测的金字塔视觉 Transformer(Pyramid Vision Transformer,PVT),可以显著减少计算量,并且在语义分割方面有很大的改进。然而,包括 Cswin[8]、Swin Transformer[9]等新的方法均着重考虑编码器设计部分,却忽略了解码器部分对进一步提升性能的贡献。

基于此,本文提出了一种基于视觉 Transformer 的多级特征聚合图像语义分割方法(Multilevel Feature Aggregation with Vision Transformer,MFAVT),将原始图像分割成切片后,使用线性切片嵌入作为 Transformer 网络编码器的输入序列;解码器将编码器生成的上下文词符序列上采样到逐像素类分数。关键思想是利用 Transformer 网络的感应特性,即较低层注意力倾向停留在局部,而高层的注意则高度非局部。通过聚合来自不同层的信息,解码器结合了来自局部和全局的注意,从而有效地提升分割精度,实现分割目标。

1 MFAVT

MFAVT 主要由编码器和解码器模块组成,模型结构如图 1 所示。在编码器部分,是将图像分块并投影到一系列嵌入位置,并使用 Transformer 网络进行编码;解码器部分,是将编码器的输出作为输入进行多层聚合,来预测分割掩膜。

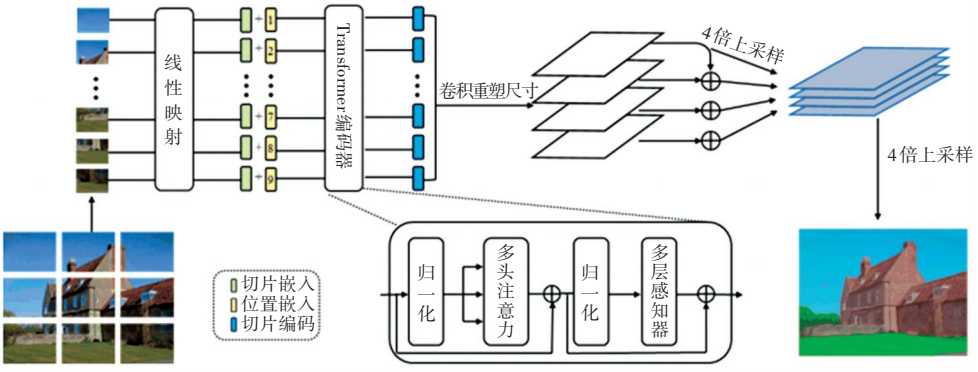


图 1 MFAVT 结构示意图

Fig. 1 The illustration of MFAVT

1.1 编码器

标准的 Transformer 网络编码器接收一维的序列词符作为输入,但二维图像和一维序列之间存在不匹配的问题,因此需要将二维图像重塑为一维序列。具体而言,将输入图像 $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ 分割成一系列切片 $x=[x_1,\cdots,x_N] \in \mathbb{R}^{N\times P^2\times C}$。其中,$(H,W)$ 是原始图像的分辨率,C 是图像的通道数,(P,P) 是每个图像切片的分辨率,$N=HW/P^2$ 是生成的切片数量,且是 transformer 有效序列输入长度。将每个切片展平为一个序列,使用线性投影函数将其映射到切片嵌入,得到图像 X 的一维切片嵌入序列 $\boldsymbol{x}_0=[\boldsymbol{E}x_1,\cdots,\boldsymbol{E}x_N] \in \mathbb{R}^{N\times D}$,其中 $\boldsymbol{E} \in \mathbb{R}^{D\times(P^2C)}$。为了对切片的空间信息进行编码,添加一个可学习的位置嵌入 $p=[p_1,\cdots,p_N] \in \mathbb{R}^{N\times D}$ 到序列切片中,以形成最终的输入序列 $g_0=x_0+p$。

以一维嵌入序列 g_0 作为输入,采用基于纯 transformer[10] 网络的编码器学习特征表示。Transformer 网络层由多头自注意力(Multi-head Self-attention, MSA)块和多层感知器(Multilayer Perception,MLP)块组成。在每个块之前使用层归一化(Layer Normalization, LN),在每个块之后添加残差链接,计算过程如式(1)所示。

$$\begin{aligned} c_{i-1} &= MSA(LN(g_{i-1})) + g_{\ell-1}, \\ g_i &= MLP(LN(c_{i-1})) + c_{i-1} \end{aligned} \tag{1}$$

其中，$i \in \{1, \cdots, L\}$。

MSA 由多个独立的 SA 操作组成，并投射其级联输出。自注意力层通过查询(Query)与键(Key)-值(Value)对之间的交互，实现信息的动态聚合。对输入序列，通过线性映射矩阵将其映射到 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$($\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} \in \mathbb{R}^{N\times D}$)3 个向量，计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 间的相似度，并对 $\boldsymbol{V}$ 进行加权处理。自注意力计算公式如式(2)所示：

$$SA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}\right)\boldsymbol{V} \tag{2}$$

Transformer 网络编码器将带位置信息的切片嵌入连续序列 $g_0 = [g_0, 1, \cdots, g_0, N]$，编码成一个供解码器使用的、带有丰富语义信息的序列 $g_L = [g_L, 1, \cdots, g_L, N]$。

1.2　解码器

解码器的目标是将切片编码序列 $g_L \in \mathbb{R}^{N\times D}$ 解码成分割图 $\boldsymbol{Seg} \in \mathbb{R}^{H\times W\times K}$。其中，$K$ 是类别数量。解码器来自编码器的切片级编码映射到切片级别类分数，通过双线性插值将这些切片级别的类分数向上采样到像素级别的分数。下面将描述一个线性解码器作为基线对比，以及介绍 MFAVT 解码器。

(1)线性解码器：首先使用了一个逐点线性层(1 × 1 卷积 + 同步批归一化 (ReLU) + 1 × 1 卷积)将 Transformer 网络特征 $g_L \in \mathbb{R}^{N\times D}$ 投影到切片类维度 $g_{bas} \in \mathbb{R}^{N\times K}$ (例如对 Pascal Context 数据集是 59)，然后将序列重整为二维特征图 $Seg_{bas} \in \mathbb{R}^{H/P\times W/P\times K}$ 并双线性上采样到原始图像大小 $\boldsymbol{Seg} \in \mathbb{R}^{H\times W\times K}$，最后在类维度上应用一个像素级交叉熵损失的分类层。当使用这种解码器时，称其为 Seg-Basic。

(2)MFAVT 解码器：采用多级特征融合的方式设计编码器，核心思想类似于特征金字塔网络。具体地说，将 Transformer 网络编码器的特征表示均匀分布在 4 层中，到达解码器；然后部署 4 个流，每个流聚焦于一个特定的选定层；在每个流中，将特征编码从 2D 特征 $\frac{HW}{PP}\times D$ 转换为 3D 特征 $\frac{H}{P}\times\frac{W}{P}\times D$。采用 3 层(卷积核大小为 1×1、3×3 和 3×3)网络，第一层和第三层分别将特征通道减半，第三层之后通过双线性运算将空间分辨率提升 4 倍，通过元素添加引入自上而下的聚合设计，来增强不同流之间的交互；按元素添加后，再使用一个 3×3 卷积；最后使用通道级联获得所有流的融合特征，通过 4 倍双线性上采样操作恢复图像到原始分辨率，形成最终的分割图。当使用这种解码器时，称其为 Seg-MFAVT。

2　实验结果与分析

2.1　数据集

实验在 3 个公开数据集上进行。其中，ADE20K[11]是最具挑战性的语义分割数据集之一，该训练集包含 20 210 幅图像，150 个语义类。验证集和测试集分别包含 2 000 和 3 352 幅图像。Pascal Context[12]数据集为整个场景提供像素级语义标签，包含 4 998(最常见的 59 个类和背景类)和 5 105 张用于训练和验证的图像。Cityscapes[13]数据集侧重于从汽车角度对城市街道场景进行语义理解。该数据集分为训练集、验证集和测试集，分别有 2 975、500 和1 525 张图像；注释包括 30 个类，其中 19 类用于语义分割任务；数据集的图像具有 2 048×1 024 的高分辨率，本文实验采用其中的精细标注图像数据集。

2.2　实验设置

2.2.1　实验环境

实验运行环境为 Win10 专业版操作系统，处理器为 Intel Core i9-9900k，内存 32 GB，图形处理卡为一张 Nvidia GeForce GTX1080 Ti(11 GB)，Cuda 版本为 10.2，数据处理使用 Python3.6 和 Matlab2020a。

2.2.2　数据增强

训练期间，遵循语义分割库 MMSegmentation[14]中的标准流程，使用比例因子(0.5、0.75、1.0、1.25、1.5、1.75)对图像执行多比例缩放以及随机的水平翻转。随机裁剪大图像，并将小图像填充到固定尺寸大小：ADE20K 为 512×512，Pascal Context 为 480×480，Cityscapes 为 768×768。辅助分割损失有助于模型训练，每个辅助损失头遵循 2 层网络，辅助损失和主损失头共同使用，此外在解码器和辅助损失头使用同步批归一化操作。

2.2.3　优化

使用标准的像素级交叉熵损失对语义分割任务的预训练模型进行微调，而无需重新平衡权重。使用随机梯度下降(SGD)[15]作为优化器，基本学习率 β_0，并将权重衰减设置为 0。采用“poly”学习率衰减 $\beta = \beta_0\left(1 - \frac{N_{\mathrm{iter}}}{N_{\mathrm{total}}}\right)^{0.9}$，其中 N_{iter} 和 N_{total} 表示当前迭代次数和总迭代次数。对于 ADE20K，其基本学习率 β_0 设置为 10^{-4}，并以 16 个批量进行 160 K 次迭代；Pascal Context，将 β_0 设置为 10^{-4}，并训练 160 K

迭代,批量大小为 16;Cityscapes,将 β_0 设置为 10^{-3},并以 8 的批量进行 160 K 迭代。

2.2.4 预训练

使用 VIT[5] 和 Deit[6](一种 VIT 的变体)提供的预训练权重,初始化模型中的所有 Transformer 网络层和输入线性投影层。将 Seg-MFAVT-Deit 表示为利用 Deit 中预训练模型的同时,使用 MFAVT 作为解码器。所有未经预训练的层均随机初始化。

2.2.5 推理

使用平均交并比(mean Intersection over Union,mIoU)作为语义分割性能的评估指标。实验报告了单尺度(Single Scale,SS)和多尺度(Multi Scale,MS)推理。对于多尺度推理,使用比例因子(0.5、0.75、1.0、1.25、1.5、1.75)对图像执行多比例缩放和随机水平翻转。测试采用滑动窗口(例如,Pascal 上下文为 480×480)。如果图像尺寸的短边长度小于滑动窗口,则在保持纵横比的同时,将短边长度调整为滑动窗口的大小(例如 480)。

2.3 消融实验

本节将在 Cityscapes 验证集上进行消融实验,评估了 Transformer 网络层大小、补丁大小、预训练集数据大小、模型性能、与 FCN 卷积网络的比较,并验证了不同的解码器。除非另有说明,否则使用 8 批次处理,80 K 迭代次数,并使用单尺度推断报告结果。表 1 中"R"代表随机初始化权重。

表 1 不同分割模型变体的性能比较

Tab. 1 Performance of different segmentation variants

Method	Pretraining	Backbone	Patch size	Params	mIoU
FCN	1K	Rsenet-101	-	68.61M	75.51
FCN	21K	Rsenet-101	-	68.61M	77.02
Seg-Basic	21K	T-base	32	88.67M	76.57
Seg-MFAVT	21K	T-base	32	91.73M	76.71
Seg-Basic	21K	T-base	16	88.67M	77.01
Seg-MFAVT	21K	T-base	16	91.73M	77.43
Seg-MFAVT	R	T-base	16	91.73M	44.14
Seg-Basic	21K	T-large	16	310.34M	78.35
Seg-MFAVT	21K	T-large	16	320.01M	78.92
Seg-Basic- Deit	1K	T-large	16	310.34M	79.28
Seg-MFAVT- Deit	1K	T-large	16	320.01M	**79.51**

观察表 1 中数据,可以得出如下结论:

(1)Seg-MFAVT-Deit 在所有的变体中取得了最佳的性能表现。

(2)使用 T-large 的变体优于 T-base 的对照物,这与实验预期一样,即 Transformer 网络层数加深会相对应的增强模型性能。如:Seg-MFAVT 使用的主干网络(Backbone)从 T-base 转换到 T-large,获得了 1.92%的提升。

(3)切片尺寸(patch size)是语义分割性能的关键因素,切片尺寸从 32 到 16,Seg-MFAVT 提高了 0.72%。可见,当图像用切片表示时,较大的切片尺寸会使模型获得有意义的全局分割,但是会产生较差的边界;而使用较小的切片尺寸会使图像边界更清晰。这一结果表明,减少切片尺寸是一个能够获得强大性能的改进来源,其不会引入任何参数,但是需要在更长的序列中计算注意力,从而增加计算时间和成本。

(4)预训练模型对于模型性能的表现至关重要。随机初始化权重的 Seg-MFAVT 只达到了 44.14% MIoU,显著低于其它变体。在 Imagenet-1K 上用 Deit 预先训练好的模型略优于在 Imagenet-21K 上用 VIT 预先训练出的模型。

(5)为了与 FCN 基线进行公平比较,使用分类任务,在 Imagenet-21K 和 1K 上对 Resnet101 进行预训练,然后在 Cityscapes 上采用预训练权重进行 FCN 训练。与在 Imagenet-1K 上的预训练变体相比,在 Imagenet-21K 上预训练的 FCN 基线得到了明显地改善。但是,本文方法在很大程度上优于 FCN 方法,体现了所提出的多层聚合策略方法的有效性,而不是更大的预训练数据。

2.4 对比分析

为了验证 MFAVT 的有效性与先进性,将 MFAVT 与一些对比方法在 Cityscapes、ADE20K 和 Pascal Context 数据集上进行性能比较。测试结果在表 2~表 4 中进行展示。在数据可视化中,为方便直观地展现分割效果,将分割结果图与原图像进行叠加并采用一定的透明化处理 ,以 DeeplabV3+分割结果代表其他方法作为锚定参照对象,与 MFAVT 分割结果进行突出化对比,结果如图 2~图 4 所示。

表 2 在 ADE20K 验证集上的性能表现

Tab. 2 Performance comparison on ADE20K validation set

Method	Pretraining	Backbone	mIoU
OCRNet[16]	1K	ResNet-101	45.61
CCNet[17]	1K	ResNet-101	45.51
DANet[18]	1K	ResNet-101	45.32
DRANet[19]	1K	ResNet-101	46.17
CPNet[20]	1K	ResNet-101	46.25
UperNet[21]	1K	ResNet-101	44.92
Deeplabv3+[22]	1K	ResNet-101	46.39
Seg-Basic(SS)	21K	VIT-L/16	47.89
Seg-Basic(MS)	21K	VIT-L/16	48.78
Seg-MFAVT(SS)	21K	VIT-L/16	48.01
Seg-MFAVT(MS)	21K	VIT-L/16	**49.97**
Seg-Basic-Deit(SS)	1K	VIT-B/16	46.41
Seg-Basic-Deit(MS)	1K	VIT-B/16	47.35
Seg-MFAVT-Deit(SS)	1K	VIT-B/16	46.53
Seg-MFAVT-Deit(MS)	1K	VIT-B/16	47.65

表 3　在 Pascal Context 验证集上的性能表现

Tab. 3　Performance comparison on Pascal Context validation set

Method	Pretraining	Backbone	mIoU
DANet[18]	1K	ResNet-101	52.59
APCNet[23]	1K	ResNet-101	54.70
SVCNet[24]	1K	ResNet-101	53.19
CFNet[25]	1K	ResNet-101	54.05
ACNet[26]	1K	ResNet-101	54.11
EMANet[27]	1K	ResNet-101	53.10
DeeplabV3+[22]	1K	ResNet-101	48.51
Seg-Basic(SS)	21K	VIT-L/16	52.35
Seg-Basic(MS)	21K	VIT-L/16	53.62
Seg-MFAVT(SS)	21K	VIT-L/16	54.16
Seg-MFAVT(MS)	21K	VIT-L/16	**55.43**
Seg-Basic-Deit(SS)	1K	VIT-B/16	52.02
Seg-Basic-Deit(MS)	1K	VIT-B/16	53.13
Seg-MFAVT-Deit(SS)	1K	VIT-B/16	53.74
Seg-MFAVT-Deit(MS)	1K	VIT-B/16	54.18

表 4　在 Cityscapes 验证集上的性能表现

Tab. 4　Performance comparison on Cityscapes validation set

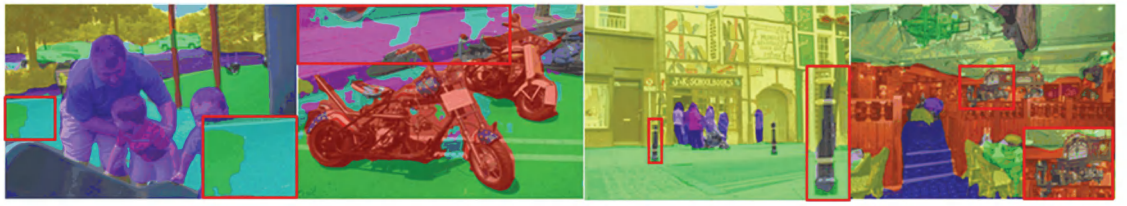
Method	Pretraining	Backbone	mIoU
DeeplabV3+[22]	1K	ResNet-101	79.32
CCNet[17]	1K	ResNet-101	80.20
PSPNet[28]	1K	ResNet-101	78.51
GCNet[29]	1K	ResNet-101	78.10
ANN[30]	1K	ResNet-101	79.90
ENcNet[31]	1K	ResNet-101	76.90
DNL[32]	1K	ResNet-101	80.50
Seg-Basic(SS)	21K	VIT-L/16	78.35
Seg-Basic(MS)	21K	VIT-L/16	81.21
Seg-MFAVT(SS)	21K	VIT-L/16	79.42
Seg-MFAVT(MS)	21K	VIT-L/16	**82.03**

(a) Deeplabv3+方法分割结果

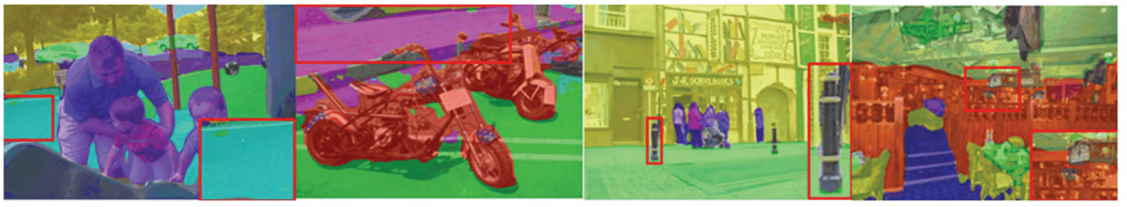
(b) MFAVT方法分割结果

图 2　在 ADE20K 上定性的可视化结果

Fig. 2　Qualitative visualization results on ADE20K



(a) Deeplabv3+方法分割结果

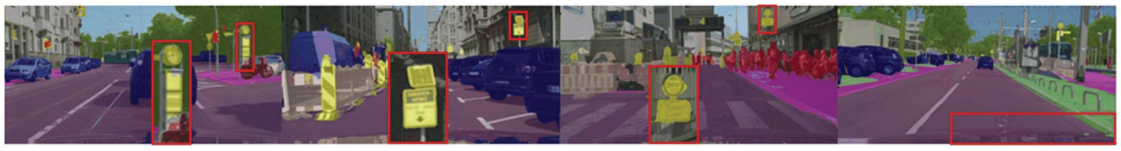


(b) MFAVT方法分割结果

图 3　在 Pascal Context 上定性的可视化结果

Fig. 3　Qualitative visualization results on Pascal Context

(a) Deeplabv3+方法分割结果



(b) MFAVT方法分割结果

图 4 在 Cityscapes 上定性的可视化结果

Fig. 4 Qualitative visualization results on Cityscapes

表 2 展示了在最具挑战性的 ADE20K 数据集上的结果,Seg-MFAVT 在单尺度推理下(SS),取得了 48.01%的 mIoU 分数,在多尺度推理(MS)下取得了最佳的 49.97%的 mIoU 分数,优于所有的卷积网络方法,比 DeeplabV3+的 mIoU 分数高出 3.58%。图 2 展示了在 ADE20K 上定性的可视化结果。

表 3 比较了在 Pascal Context 上的分割结果。在单尺度推理时,Seg-MFAVT 得到了 54.16%的 mIoU 分数,而在多尺度推理时获得了最佳的 55.43%mIoU 分数,超过了所有 FCN 方法。与最有竞争力的 APCNet 相比,mIoU 分数提高了 0.73%。图 3 展示了在 Pascal Context 上定性的可视化结果。

在 Cityscapes 验证集上的比较结果见表 4。Seg-MFAVT 在单尺度推理下取得了 79.42%的 mIoU 分数,而在多尺度推理下取得了令人印象深刻的 82.03%mIoU 分数。需要注意的是相比于一些方法在训练中采用全尺寸图像分辨率(2 048×1 024)输入,MFAVT 的图像输入尺寸为 768×768,训练过程有一定劣势,但最终的性能表现超过了其他有竞争力的方法。与 DeeplabV3+相比提高了 2.71%mIoU,与最有竞争力的 DNL 相比提高了 1.53%mIoU。图 4 展示了在 Cityscapes 上定性的可视化结果。

3 结束语

本文介绍了一种基于视觉 Transformer 的序列到序列的分割方法,为语义分割任务提供了一种新的视角。现有的基于 FCN 的方法通常使用扩张卷积和注意力模块来扩大感受野,与之相比,本文的编码器部分采用当下流行的视觉 Transformer 主干网络,对图像切片进行编码。基于视觉 Transformer 的编码器很好地建模了全局上下文信息,随着一组不同的复杂性的解码器设计,建立了强大的分割模型。简单的线性解码器就取得了非常好的效果,使用 MFAVT 进行解码进一步提高了性能。大量的实验表明,本文方法在 ADE20K、Pascal Context 和 Cityscapes 数据集测试上展示了最佳的性能表现。

参考文献

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[3] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[4] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[5] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale[C]//International Conference on Learning Representations. 2020:3031-3052.

[6] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers and distillation through attention [C]//International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[7] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 568-578.

[8] DONG X, BAO J, CHEN D, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12124-12134.

[9] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in neural information processing

systems. 2017：5998-6008.
[11] ZHOU B，ZHAO H，PUIG X，et al. Semantic understanding of scenes through the ade20k dataset[J]. International Journal of Computer Vision，2019，127(3)：302-321.
[12] MOTTAGHI R，CHEN X，LIU X，et al. The role of context for object detection and semantic segmentation in the wild[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2014：891-898.
[13] CORDTS M，OMRAN M，RAMOS S，et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016：3213-3223.
[14] MMSegmentation Contributors. MMSegmentation：OpenMMLab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation，2020.
[15] ROBBINS H，MONRO S. A stochastic approximation method[J]. The annals of mathematical statistics，1951：400-407.
[16] YUAN Y，CHEN X，WANG J. Object-contextual representations for semantic segmentation[C]//European conference on computer vision. Springer，Cham，2020：173-190.
[17] HUANG Z，WANG X，HUANG L，et al. Ccnet：Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019：603-612.
[18] FU J，LIU J，TIAN H，et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019：3146-3154.
[19] FU J，LIU J，JIANG J，et al. Scene segmentation with dual relation-aware attention network[J]. IEEE Transactions on Neural Networks and Learning Systems，2020，32(6)：2547-2560.
[20] YU C，WANG J，GAO C，et al. Context prior for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：12416-12425.
[21] XIAO T，LIU Y，ZHOU B，et al. Unified perceptual parsing for scene understanding[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018：418-434.
[22] CHEN L C，ZHU Y，PAPANDREOU G，et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018：801-818.
[23] HE J，DENG Z，ZHOU L，et al. Adaptive pyramid context network for semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：7519-7528.
[24] DING H，JIANG X，SHUAI B，et al. Semantic correlation promoted shape-variant context for segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：8885-8894.
[25] ZHANG H，ZHANG H，WANG C，et al. Co-occurrent features in semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：548-557.
[26] FU J，LIU J，WANG Y，et al. Adaptive context network for scene parsing[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：6748-6757.
[27] LI X，ZHONG Z，WU J，et al. Expectation-maximization attention networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：9167-9176.
[28] ZHAO H，SHI J，QI X，et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017：2881-2890.
[29] CAO Y，XU J，LIN S，et al. Gcnet：Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019：1971-1980.
[30] ZHU Z，XU M，BAI S，et al. Asymmetric non-local neural networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：593-602.
[31] ZHANG H，DANA K，SHI J，et al. Context encoding for semantic segmentation[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018：7151-7160.
[32] YIN M，YAO Z，CAO Y，et al. Disentangled non-local neural networks[C]//European Conference on Computer Vision. Springer，Cham，2020：191-207.