2025 年 第 25 卷 第 8 期
2025, 25(8): 03333-07

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2401440
引用格式:张剑飞, 倪俊文. 基于多尺度上下文注意力的遥感图像语义分割[J]. 科学技术与工程, 2025, 25(8): 3333-3339.
Zhang Jianfei, Ni Junwen. Remote sensing image segmentation based on multi-scale context attention[J]. Science Technology and Engineering, 2025, 25(8): 3333-3339.

基于多尺度上下文注意力的遥感图像语义分割

张剑飞, 倪俊文
(黑龙江科技大学计算机与信息工程学院, 哈尔滨 150022)

摘 要 遥感图像的语义分割对农业生产、城市规划等领域有十分重要的作用,但受成像距离、光照、地物、环境等因素影响,遥感图像中存在目标语义信息模糊问题,导致在分割时存在不确定性。针对此问题,提出一种多尺度上下文注意力方法(multi-scale context attention,MSCA),其将金字塔池化方法与注意力方法相结合,可以更充分地利用上下文信息。同时该方法显著降低了注意力方法的计算量和内存占用。在 ISPRS Potsdam 数据集上进行了实验,实验结果表明:MSCA 方法在不显著增加内存开销,以及维持推理速度一致的情况下,对遥感图像中语义信息不明确的目标分类,具有更好的分割效果。

关键词 遥感图像; 语义分割; 注意力机制; 多尺度上下文
中图法分类号 TP751; **文献标志码** A

Remote Sensing Image Segmentation Based on Multi-scale Context Attention

ZHANG Jian-fei, NI Jun-wen
(School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China)

[**Abstract**] Semantic segmentation of remote sensing images plays a crucial role in agriculture production, urban planning, and other fields. However, due to factors like imaging distance, lighting conditions, objects, and environment, there is a problem of semantic ambiguity in remote sensing images, which leads to uncertainty in segmentation. A multi-scale context attention (MSCA) method that combined pyramid pooling with attention mechanisms to better utilize contextual information was proposed for this problem. Additionally, this method significantly reduced the computational complexity and memory usage of attention methods. Experimental results on the ISPRS Potsdam dataset demonstrate that the MSCA method achieves superior segmentation performance for target classification with ambiguous semantic information in remote sensing images while almost not increasing memory consumption and maintaining consistent inference speed.

[**Keywords**] remote sensing image; semantic segmentation; attention; multi-scale contextual information

遥感图像的语义分割是计算机视觉任务之一,通过对图像中的像素进行分类,从而分割出图像中的不同区域。在农业生产中,语义分割可以帮助监测作物生长、识别病虫害以及进行土地覆盖分类;在城市规划中,它有助于精确识别建筑物、道路和绿地,从而支持基础设施规划和城市扩展管理;在灾害评估方面,语义分割能够快速识别灾害区域,帮助制定救灾和恢复计划。但因为拍摄角度不同、天气变化、阴影方向和传感器硬件限制等的原因,使得分类目标存在着语义信息不明确,易导致分类错误的问题。

卷积神经网络(convolutional neural network, CNN)广泛应用于图像的语义分割任务的中的特征编码部分,但因为计算资源限制的原因卷积核通常都比较小,感受野只能覆盖图像中的一小部分,使得编码后的特征对于图像的上下文认知不足。针对这一问题,研究人员在解码端引入一些生成多尺度上下文信息的方法,如 PSPNet[1] 中的金字塔池化方法、DeepLabV3 + 中的空洞金字塔池化方法等[2-3]。这些方法都能够提取出多尺度上下文信息以辅助目标分类,但其采用的对特征插值上采样后在通道上连接的方式使得上下文信息只能在局部空间位置上融合,不能利用所有上下文信息。

同时 CNN 存在着难以学习长距离依赖关系的问题,非局部网络(non-local)[4] 引入基于查询键值对的空间注意力机制,这种机制会无视特征的空间

收稿日期: 2024-03-03; **修订日期**: 2024-12-15
基金项目: 国家自然科学基金(61803148);黑龙江省属高等学校基本科研业务费项目(2024-KYYWF-1099);黑龙江省哲学社会科学研究规划项目(23YSD245)
第一作者: 张剑飞(1978—),女,汉族,黑龙江哈尔滨人,博士,教授。研究方向:人工智能与模式识别。E-mail:zjfnefu2008@163.com。

距离并建立关系,让相关的空间特征可以相互响应。刘春容等[5]在残差网络中嵌入多尺度的 non-local 注意力机制,提升了遥感图像的分类性能。陈钊等[6]在 DeepLabV3 + 中嵌入 non-local 注意力机制,提升了对遥感图像中滑坡区域的识别能力。non-local 注意力机制的计算量和内存占用相比于卷积网络过大,ANN(asymmetric non-local neural network)[7]的 APNB(asymmetric pyramid non-local block)模块使用金字塔池化对键值对进行降维,并且复用键和值的线性层与金字塔池化模块,有效减小了计算资源消耗。

除 CNN 外,Vision Transformer[8]将自然语言领域的 Transformer[9]引入图像领域中,只使用了注意力机制和全连接层就完成了图像处理任务。Vision Transformer 通过将图像划分成 patch 来减小特征空间尺寸,因此计算资源消耗被控制;同时它还在注意力中引入多头机制,使得注意力可以在多个子空间中建立,提高了表达能力。但是 Vision Transformer中注意力机制的计算量和内存占用相比于卷积网络过大,也更难训练[10-12]。金字塔 Transformer(PVT)[13]、Shunted Transformer[14]、Segformer[15]从空间降维的角度出发,Swin Transformer[16-17]从限制注意力范围的角度出发,有效减少了计算量和内存占用。Shunted Transformer 中则将多头注意力中的不同头的 patch 划分设置成不一样的大小,使得注意力关系可以在不同尺度的 patch 之间形成。针对网络的空间分辨率小于原始输入图像问题,Li 等[18]将 Transformer 引入解码阶段,利用注意力机制来进行还原空间尺寸的工作。田雪伟等[19]对 SegFormer 进行改进,在解码时以级联的方式合并特征,使用门控注意力机制过滤掉影响图像中小目标和细节特征的高层语义特征,使用多局部通道注意力模块对通道进行注意力增强,实现了对遥感图像中小目标和边界的更好的语义分割效果。

总的来说,目前对遥感图像的语义分割存在以下问题:①一些方法虽然可以生成多尺度上下文信息来扩大 CNN 的感受野,从而提升模型对语义信息不明确目标的分割效果,但是其上采样插值后连接的特征融合方式使得上下文信息利用率不足;②基于查询键值对的注意力方法有全局的感受野,可以对所有特征建立依赖关系,从而生成语义信息更丰富的特征,但是计算量和内存占用过大。针对这些问题,提出一种多尺度上下文注意力方法(multi-scale context attention,MSCA)。MSCA 结合金字塔池化与基于查询键值对的注意力机制,对多尺度上下文信息进行特征融合,同时通过控制键值对数量有效降低计算复杂度和内存占用。研究表明,MSCA 在保持高效计算的同时能够适应遥感图像中的复杂场景与目标特性。研究成果可为遥感图像中语义信息不明确目标的精确分割提供新的解决方案,有助于提升语义分割在农业生产、城市规划及灾害评估等领域的应用效果。

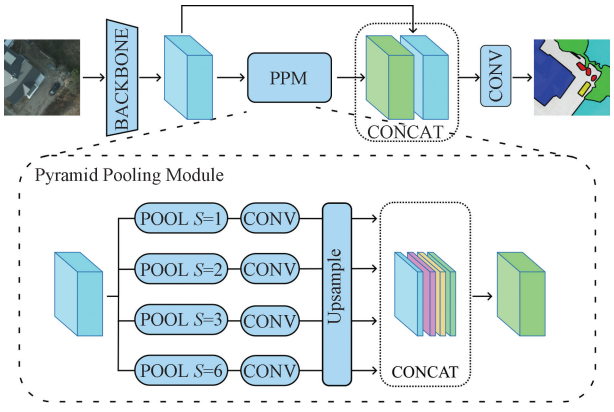
1 相关理论基础

1.1 金字塔场景解析网络(PSPNet)

金字塔场景解析网络(PSPNet)[1]的整体结构如图 1 所示,其属于编码-解码器结构,先使用主干网络(backbone)对图像编码以得到有高级语义信息的特征,再用后续模块作为解码器对这些特征进行解码,得到分割结果。

编码器部分的主干网络是 ResNet-d8,它的不同阶段(stage)生成的特征的空间尺寸是不一致的,越后面的阶段空间尺寸越小但通道维数量越大,会有更丰富的语义信息。ResNet-d8 有 4 个阶段特征输出,其中最后一阶段的特征的空间尺寸为原始图像的 1/8,通道维数量为 2 048,典型的 PSPNet 只使用该阶段的特征进行解码。

解码器部分使用了金字塔池化模块(pyramid pooling module,PPM)来生成含有多尺度上下文信息的特征,插值上采样后与原始特征在通道维上连接。典型的 PPM 有 4 个分支,每个分支都是池化(pool)加 Point-wise 卷积,其中池化分别将特征在空间尺寸上降维到 1 × 1、2 × 2、3 × 3 和 6 × 6(对应于图 1 中 POOL 模块的 S 值),而 Point-wise 将特征在通道上降维到 256。因为特征的空间尺寸被缩小了,所以又用了线性插值上采样恢复尺寸后再在通道维上连接。PPM 使得局部空间的特征拥有了多尺度的上下文信息,更精确地做出每个像素的分类结果。



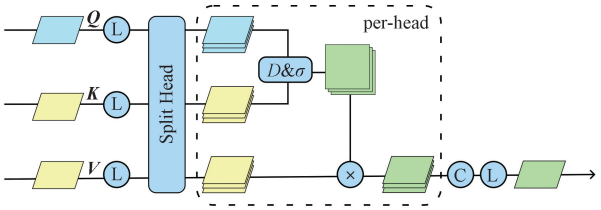
BACKBONE为主干网络;CONCAT为通道连接;CONV为卷积;POOL为平均池化;S为池化输出特征尺寸;Upsample为上采样;Pyramid Pooling Module为金字塔池化模块

图 1 PSPNet 模型结构图

Fig. 1 Structure chart of PSPNet

1.2　多头注意力机制 MHA

Vision Transformer[8] 中的多头注意力机制(MHA)结构如图 2 所示,它需要 3 组特征作为输入,分别用 $\boldsymbol{Q}$(查询)、$\boldsymbol{K}$(键)和 $\boldsymbol{V}$(值)表示。对它们应用线性变换后在通道维上划分成数个头部(split head),然后在每个头部中分别计算点积注意力,得到的结果在通道维上连接,再应用一个线性变换得到输出。点积注意力用点积(dot-product)计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的两两特征之间的相关性分数,再用 softmax归一化后与 $\boldsymbol{V}$ 的特征相乘,就此用 $\boldsymbol{Q}$ 查询 $\boldsymbol{K}$ 的方式形成了对 $\boldsymbol{V}$ 的注意力。这种注意力是无视空间距离的,单个特征对全局特征都有响应,可以建立空间上的长距离依赖关系。而多头的设计可以使得注意力机制在不同的子空间内建立,以形成复合的、表达能力更强的注意力关系。



L为线性层;Split Head为划分多头;$D\&\sigma$为点积和softmax;×为矩阵乘法;C为通道连接

图 2　多头注意力机制的结构图

Fig. 2　Structure chart of MHA

MHA 的计算公式为

$$\mathrm{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{cat}(H_1,H_2,\cdots,H_h)\,\boldsymbol{W}^{o} \quad (1)$$

$$H_i = \mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left[\frac{\boldsymbol{Q}\boldsymbol{W}_i^{Q}(\boldsymbol{K}\boldsymbol{W}_i^{K})^{\mathrm{T}}}{\sqrt{d_k}}\right]\boldsymbol{V}\boldsymbol{W}_i^{V} \quad (2)$$

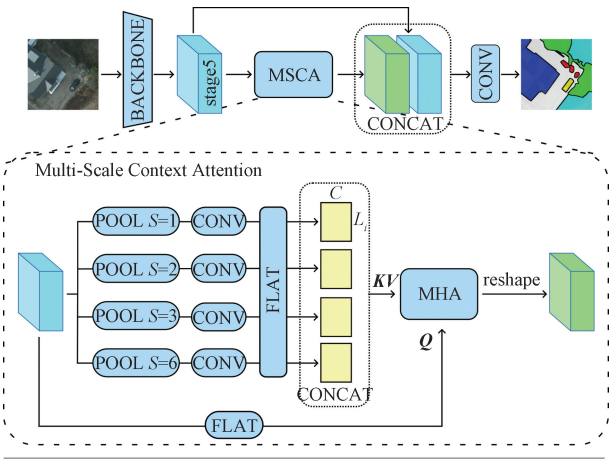
式中:H_i 为第 i 个头的输出结果;Attention 为计算注意力的函数;$\boldsymbol{W}_i^{Q}$、$\boldsymbol{W}_i^{K}$ 和 $\boldsymbol{W}_i^{V}$ 分别为第 i 个头中连接 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 的线性层权重;$\boldsymbol{W}^{o}$ 为最后的线性层权重;d_k 为 $\boldsymbol{K}$ 的维数,即键数量。

当 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 都是同一组特征时 MHA 计算的是特征对自身的注意力,也称为自注意力。这是最典型的情况,如 Vision Transformer 就是对特征进行 patch 划分后进行的自注意力计算。

2　改进 PSPNet 的分割方法

2.1　模型整体结构

如图 3 所示,所提出的模型以 PSPNet 为基础,将其中的 PPM 模块替换成 MSCA 模块,其余结构保持不变。MSCA 中每个分支使用池化和卷积提取出不同尺度的上下文特征,展平后在 L 维上连接作为注意力的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$;输入特征(局部特征)展平作为



BACKBONE为主干网络;stage 5为第5阶段的特征;CONCAT为特征连接;CONV为卷积;POOL为平均池化;S为池化输出特征尺寸;C为特征通道数;L_i为第i个分支的特征数量;FLAT为特征展平;reshape为特征尺寸变形

图 3　本文模型整体结构图

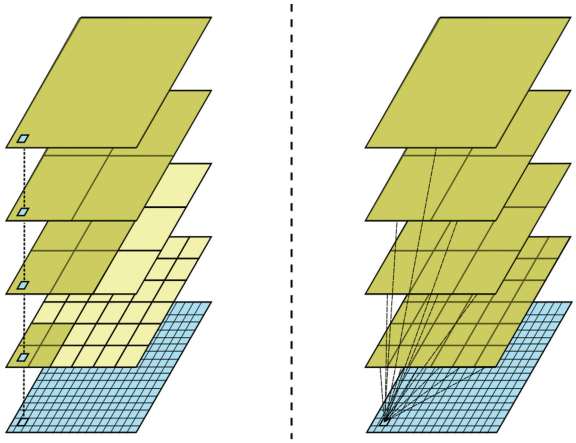
Fig. 3　Structure chart of MSCA model

$\boldsymbol{Q}$,使用 MHA 建立多头注意力;最后用 reshape 还原特征空间维度。

2.2　多尺度上下文注意力 MSCA

如图 4 所示,PSPNet 中的 PPM 模块在上下文特征利用率方面存在局限性。该模块采用插值上采样的方式恢复特征的空间尺寸,并通过在通道维度上的连接与原始特征进行融合。然而由于插值上采样方法仅利用了周边几个特征(如双线性插值仅利用周围 4 个特征;双三次插值仅利用 16 个特征),因此其上下文特征利用率不足。

基于 PPM 中上下文特征利用率不足的问题,提出多尺度上下文注意力方法(multi-scale context attention,MSCA)。MSCA 不会对多尺度上下文特征进行上采样,而是将原始特征作为查询 $\boldsymbol{Q}$,上下文特



(a) PPM　　(b) MSCA

蓝色为原始特征图;黄色为上下文特征图;MSCA 相比 PPM 可以利用所有上下文特征

图 4　PPM 与 MSCA 的特征融合方式对比

Fig. 4　Comparison of feature fusion between PPM and MSCA

征作为键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$,来进行注意力计算。注意力计算的输出特征的空间尺寸和 $\boldsymbol{Q}$ 相同,这种做法直接恢复了空间尺寸,而不需要插值上采样操作,并且得到的特征是融合了所有上下文特征的信息的,因此 MSCA 在复杂模糊、难以界定的目标上有更好的分割效果。

MSCA 的结构如图 3 所示。典型的 MSCA 与 PPM 一样有 4 个分支,并且各分支上的池化和卷积参数设置相同,于是生成了空间尺寸为 1×1、2×2、3×3 和 6×6,通道数为 256 的特征。但是后面并没有像 PPM 一样上采样后连接,而是展平后在 L 维上连接后作为 MHA 的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$。可表示为

$$C_i = \text{flat}\{\text{conv}[\text{pool}(X,S_i)]\} \tag{3}$$

$$C = \text{cat}(C_1,C_2,\cdots,C_n) \tag{4}$$

式中:X 为输入特征;C_i 为第 i 个分支输出的上下文特征;S_i 为第 i 个分支的平均池化输出特征的尺寸的边长;C 为在 L 维上连接后的上下文特征;pool 为平均池化操作;flat 为将维度为 $C\times H_i\times W_i$ 维的特征转换成 $C\times L_i$,其中 $H_i\times W_i = L_i$;conv 为 point-wise 卷积、层归一化(LayerNorm)、激活函数(ReLU)的复合操作;cat 表示将特征在 L 维上连接。

2.3 MSCA 与其他注意力方法对比

MHA 会计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 中两两特征之间的相关性以生成注意力图,而图像分割任务中的特征数量是非常多的,直接使用 MHA 计算特征之间的注意力会消耗大量的计算资源。计算资源指计算量 GFLOPs(giga floating-point operations per second)和内存占用。

Vision Transformer 将图像特征在空间上进行 patch 划分,以减少特征数量来避免这个问题。典型情况下每个 patch 包含 16×16 个特征,所以特征数量会减少到 1/256。但是这种方式会使得注意力只能在 patch 之间形成,精度相比于不进行 patch 划分下降很多。

PVT 提出一种只对 MHA 中的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 进行降维的方法,使用卷积或者池化来减少 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 的数量,但维持 $\boldsymbol{Q}$ 特征数量不变,有效减少了计算资源消耗。这种方法实际上建立了局部特征 $\boldsymbol{Q}$ 和单尺度上下文特征 V 之间的注意力关系,故称其为单尺度上下文注意力(single-scale context attention,SSCA)。

MSCA 相比于 SSCA,改进点在于使用了多个不同降维幅度的池化,生成多尺度上下文特征作为 $\boldsymbol{K}$ 和 $\boldsymbol{V}$,可以建立局部特征 $\boldsymbol{Q}$ 和多尺度上下文特征之间的注意力关系,比 SSCA 的单尺度注意力关系表达能力更强。

根据式(1)和式(2),设输入特征 X 的通道数为 2 048,空间尺寸为 64×64,共计 4 096 个特征。直接应用 MHA 方法中,使用 X 作为 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$;SSCA 方法中,使用 X 作为 $\boldsymbol{Q}$,对 X 进行池化得到的特征作为 $\boldsymbol{K}$ 和 $\boldsymbol{V}$,其空间尺寸为 16×16,共计 256 个特征;MSCA 方法中,使用 X 作为 $\boldsymbol{Q}$,使用多个池化的输出作为 $\boldsymbol{K}$ 和 $\boldsymbol{V}$,各池化输出尺寸分别为 1×1、2×2、3×3、6×6,$\boldsymbol{K}$、$\boldsymbol{V}$ 的特征数量为 50。同时,所有方法中的 MHA 的头数量都是 8,用 $\boldsymbol{W}_i^{Q}$ 和 $\boldsymbol{W}_i^{K}$ 将 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的通道数降维到 256,所以每头中特征通道数是 32;而 $\boldsymbol{W}_i^{V}$ 维持 V 的通道数不变为 2 048。

表 1 对比了它们消耗的计算资源。其中计算量只计算了乘法与加法,而忽略了激活和归一化等操作;内存占用只计算了注意力图的占用,数据类型为 32 位浮点数;表 1 中所用的 MHA 省略了式(1)中的 $\boldsymbol{W}^{o}$,即没有在连接多个头部结果之后再接一个线性层。

表 1 不同方法的 MHA 消耗的计算资源对比

Table 1 Comparison of computational resources consumed by MHA of different methods

方法	头数量	$\boldsymbol{Q}$ 数量	$\boldsymbol{K}$、$\boldsymbol{V}$ 数量	GFLOPs	内存占用/MB
直接应用 MHA	8	4 096	4 096	26.01	512.00
SSCA 的 MHA	8	4 096	256	1.15	32.00
MSCA 的 MHA	8	4 096	50	0.61	6.25

由表 1 可知,MSCA 中 MHA 的计算资源消耗最小,计算量不大,内存占用也不会影响到模型训练。MSCA 与 SSCA 对 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 降维的方法可以减少非常多的计算资源。

3 实验与分析

3.1 实验数据集

为了验证本文 MSCA 方法的有效性,使用其构建的模型在 ISPRS Potsdam 数据集上进行图像语义分割实验。

ISPRS Potsdam 数据集为城市区域的遥感影像语义分割数据集,以德国波茨坦地区为研究区域,包含 5 个前景分类:不透水表面(Imp surf)、建筑(Build)、低植被(Low veg)、树木(Tree)、汽车(Car),以及 1 个背景分类:杂乱物(Clutter),有 4 个光谱波段:红、绿、蓝和红外。本次实验对所有分类进行分割,使用红、绿和蓝共 3 个波段的图像,图像统一裁剪成 512×512 尺寸,在训练时使用随机等比缩放、裁剪、翻转、光度失真(随机改变图像的饱和度、亮度、色调)的数据增强方法。

3.2 实验平台与环境

实验所使用的深度学习框架为 MMSegmenta-

tion1.1.1 + Pytorch1.11 + cuda11.3。模型训练时使用4卡A5000,模型评估时使用V100单卡。使用Adamw[20]作为优化器,学习率设为0.000 06,解码器网络的学习率设为编码器的10倍,每卡批量大小为2,迭代10 000次,每1 000次设保存点,取最好的性能模型做对比。

模型的主干网络使用在ImageNet[21]上预训练的ResNet50-d8或者ResNet101-d8,其第3、第4阶段的输出特征的空间尺寸都是原始图像的1/8大小,通道数分别为1 024和2 048。将第4阶段的输出连接到解码器网络,而第3阶段的输出连接到一个FCN辅助头。模型解码部分中,池化后的卷积将通道数降成512;MHA的输入$\boldsymbol{Q}$和输出特征都为2 048维,$\boldsymbol{K}$和$\boldsymbol{V}$为512维,内部再用线性层将$\boldsymbol{K}$和$\boldsymbol{V}$降到256维,头数为8;模型最后的CONV为卷积核大小为3×3的标准卷积,输出维数是512。

3.3　评价指标

为了度量模型的分割精度和计算效率,使用以下评价指标。

(1)交并比(intersection over union,IoU)与平均交并比(mean intersection over union,mIoU),单位为百分比。IoU计算分割区域与真实区域之间的交集与并集的比例大小,来衡量单个类别的分割质量,而平均交并比mIoU就是各类别的交并比的平均值。IoU和mIoU越高,说明模型的分割更精确。

(2)训练时每卡内存占用Mem(memory),单位:MB。模型训练时相比推理时会使用更多的内存,这是因为模型的梯度反向传播算法为了速度在训练时保存了很多中间变量。显卡拥有比CPU更强的并行算力被广泛用于训练模型,但其显存非常有限,所以训练时的内存占用(显存占用)是一个重要的指标,决定了模型能否训练。所有Mem指标都是在每卡batch_size=2的环境下测得的。

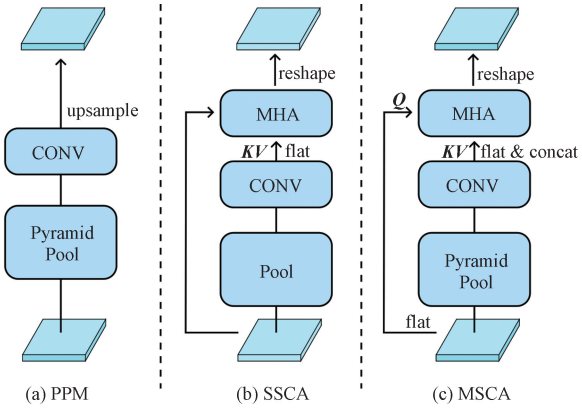
(3)推理时每秒帧数(frame per second,FPS)。FPS越高说明模型推理时单位时间可以处理更多图像。

3.4　消融实验

为进一步了解MSCA中多尺度上下文特征和注意力机制的有效性,以及它们的计算资源消耗情况,设计消融实验。消融实验使用ResNet-50-d8作用主干网络、ISPRS Potsdam作为数据集训练4个模型进行对比。这4个模型分别为记为baseline、PPM、SSCA和MSCA模型,它们都基于图1的PSPNet进行修改。baseline模型将PSPNet中的PPM去除;PPM模型就是PSPNet模型本身;SSCA模型将PSPNet中的PPM更换为SSCA方法实现的模块;MSCA即为本文模型。

PPM、SSCA和MSCA方法的细节结构如图5所示。

消融实验的结果如表2所示。可以看出,PPM、SSCA相比于baseline,mIoU提升了0.39%和0.67%,说明单独使用多尺度上下文或注意力机制都可以提升对图像分割的精度,但都较明显增加了训练时的内存占用,推理速度也有所降低;MSCA相比SSCA,mIoU提升了0.24%,说明注意力机制中使用多尺度上下文特征相比单尺度特征更有效,且几乎不会增加计算资源消耗;MSCA在所有方法中取得了最好的性能,说明多尺度上下文与注意力机制结合使用可以最有效提升模型性能。



upsample为上采样;CONV为卷积;Pyramid Pool为金字塔池化;reshape为特征尺寸变形;flat为特征展平;Pool为平均池化;concat为特征连接

图5　金字塔池化模块PPM、单尺度上下文注意力SSCA和多尺度上下文注意力MSCA的细节结构对比

Fig.5　Comparison of the detailed structures of PPM, SSCA and MSCA

表2　各方法的实验模型结果比较

Table 2　Experimental results of different methods

方法	多尺度上下文	注意力机制	mIoU/%	Mem/MB	FPS/(帧·s^{-1})
baseline	×	×	77.81	3 117	40.53
PPM	√	×	78.20	3 891	33.65
SSCA	×	√	78.48	4 075	33.07
MSCA	√	√	78.72	4 078	33.08

注:×表示未加对应模块;√表示添加对应模块。

3.5　实验结果与分析

为了测试本文模型的有效性,将本文模型、PSPNet以及DeepLabV3+在ISPRS Potsdam数据集上进行训练和测试。同时控制变量,各模型的主干网络均使用ResNet101-d8。实验结果如表3所示。可以看出,在计算资源消耗情况上,MSCA相比于PSPNet只在训练时多占用了186MB的内存,FPS只下降了

表 3 MSCA 和主流方法的实验结果

Table 3 Experimental results of MSCA and mainstream methods

模型	各分类精度 IoU/%						mIoU/%	Mem/MB	FPS/(帧·s^{-1})
	Imp surf	Build	Low veg	Tree	Car	Clutter			
PSPNet	87.44	94.03	76.64	79.33	93.07	41.24	78.62	5 890	21.74
DeepLabV3 +	87.40	93.82	76.60	79.28	93.07	42.00	78.70	6 047	20.04
MSCA	86.97	93.32	77.28	79.54	92.82	46.38	79.39	6 076	21.55

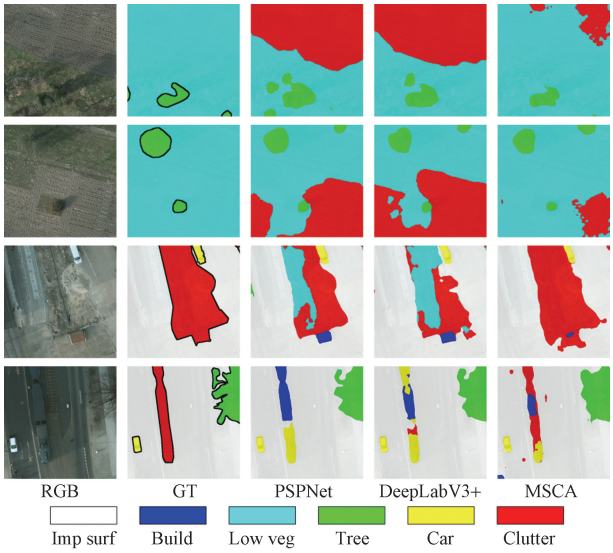


图 6 含有杂乱物目标的分割结果可视化对比

Fig. 6 Visual comparison of segmentation results with cluttered object targets

0.19。与 DeepLabV3 + 相比,内存占用几乎一致。这说明 MSCA 使用的注意力机制显著降低了计算量和内存占用,从而确保了整体模型在训练和推理过程中的高效表现。

在各分类精度 IoU 方面,MSCA 相比于其余模型,虽然在大部分目标上都拉不开差距,但对杂乱物目标(Clutter)的分割有明显的优势,其 IoU 相比于 PSPNet 提升了 5.14%,从而使得 mIoU 达到最高。

选取语义信息相对不明确的图像进行可视化对比,如图 6 所示。从最左列到最右列分别为输入图像、ground-truth、PSPNet 分割图、DeepLabV3 + 分割图和 MSCA 分割图。

图 6 的第 1、2 行中,PSPNet 和 DeepLabV3 + 将低植被错误地分类为杂乱物,而 MSCA 情况较好;第 3 行的施工路面属于杂乱物,被 PSPNet 和 DeepLabV3 + 错误分类成了低植被,但 MSCA 正确对其分类;第 4 行中的火车属于杂乱物,所有模型都有错误分类成汽车或者建筑的情况,MSCA 表现最好。

综上可知,杂乱物属于 Potsdam 数据集中最困难的目标,其边界信息不明显,纹理细节和别的目标很相似尤其是低植被。在这种情况下,仅依靠局部的纹理细节信息无法做出准确判断,必须参考周围的环境信息。多尺度上下文信息可以充当这种环境信息,但 PSPNet 和 DeepLabV3 + 都因为上下文信息利用率不高的原因导致分割错误。MSCA 能更有效地利用上下文信息,因此相比别的模型在杂乱物分类上优势非常明显,从而使得 mIoU 指标达到最高。

4 结论

提出一种结合使用多尺度上下文和注意力机制的方法 MSCA,其基于 PSPNet 的 PPM 进行改进,可以更有效地利用上下文特征来建立注意力关系。然后对比了 MSCA 与 Vision Transformer 和 PVT 中注意力方法的不同,分析了它们的计算资源消耗情况,证明了 MSCA 中注意力的计算量和内存占用得到有效控制。实验结果表明,在 ISPRS Potsdam 数据集上,所提出的 MSCA 方法相比于 PSPNet,在训练内存占用只增加 186 MB 以及推理 FPS 几乎一致的情况下,平均交并比提高了 0.77%,其中杂乱物目标的交并比提高了 5.14%。这说明在实际应用中,本文方法在无需显著增加计算量和内存占用的情况下,能够更为有效地区分语义信息模糊且容易误分类的复杂目标,同时提高所有分类目标的平均分割精度。本次研究不仅为遥感图像的分割任务提供了新的思路,也为未来其他需要高效上下文信息融合的图像任务提供了新的解决方案。

参 考 文 献

[1] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2881-2890.

[2] Liang-Chieh C, George P, Florian S, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv Preprint, 2017: DOI: 10.48550/arXiv. 1706.05587.

[3] 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). Berlin: Springer, 2018: 801-818.

[4] Wang X, Girshick R, Gupta A, et al. Non-local neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7794-7803.

[5] 刘春容, 宁芊, 雷印杰, 等. 改进残差神经网络在遥感图像分

类中的应用[J]. 科学技术与工程, 2021, 21(31): 13421-13429.
Liu Chunrong, Ning Qian, Lei Yinjie, et al. Application of improved residual network in sensing image classification[J]. Science Technology and Engineering, 2021, 21(31): 13421-13429.

[6] 陈钊, 鲁仕康, 覃章健, 等. SENet 优化的 Deeplabv3 + 滑坡识别[J]. 科学技术与工程, 2022, 22(33): 14635-14643.
Chen Zhao, Lu Shikang, Qin Zhangjian, et al. SENet-optimized Deeplabv3 + landslide detection in the Sichuan section of the Sichuan-Tibet Railway[J]. Science Technology and Engineering, 2022, 22(33): 14635-14643.

[7] 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. Piscataway: IEEE, 2019: 593-602.

[8] 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. Hoboken: John Wiley International Publishing Group, 2021: 1-22.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30: 1-11.

[10] 田永林, 王雨桐, 王建功, 等. 视觉 Transformer 研究的关键问题: 现状及展望[J]. 自动化学报, 2022, 48(4): 957-979.
Tian Yonglin, Wang Yutong, Wang Jiangong, et al. Key problems and progress of vision Transformers: the state of the art and prospects[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 48(4): 957-979.

[11] Touvron H, Cord M, Sablayrolles A, et al. Going deeper with image transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 32-42.

[12] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: a survey[J]. Computational Visual Media, 2022, 8(3): 331-368.

[13] 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. Piscataway: IEEE, 2021: 568-578.

[14] Ren S, Zhou D, He S, et al. Shunted self-attention *via* multi-scale token aggregation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10853-10862.

[15] Xie E, Wang W, Yu Z, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[16] 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. Piscataway: IEEE, 2021: 10012-10022.

[17] Liu Z, Hu H, Lin Y, et al. Swin Transformer v2: scaling up capacity and resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 12009-12019.

[18] Li Y, Cai W, Gao Y, et al. More than encoder: introducing transformer decoder to upsample[C]//2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway: IEEE, 2022: 1597-1602.

[19] 田雪伟, 汪佳丽, 陈明, 等. 改进 SegFormer 网络的遥感图像语义分割方法[J]. 计算机工程与应用, 2023, 59(8): 217-226.
Tian Xuewei, Wang Jiali, Chen Ming, et al. An improved SegFormer network based method for semantic segmentation of remote sensing images[J]. Computer Engineering and Applications, 2023, 59(8): 217-226.

[20] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv Preprint, 2019: DOI: 10.48550/arXiv.1711.05101.

[21] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.