DOI:10.11721/cqnuj20230608

融合金字塔池化和注意力机制的轻量语义分割方法*

廖恒锋,魏 延,杜韩宇
(重庆师范大学 计算机与信息科学学院,重庆 401331)

摘要:语义分割被广泛应用于医学图像分割、无人驾驶、遥感图像分割等计算机视觉任务中,而目前语义分割方法通常所需的计算量和参数量庞大,难以在算力和硬件存储有限的嵌入式平台部署。针对这一问题,从网络的参数量、计算量、性能等 3 个方面综合考虑,设计了 1 种轻量化语义分割方法。以轻量化网络 MobileNetV2 为主干,使用深度可分离卷积对模型进行压缩,分为高低语义 2 条路径向前推导。为了保证网络性能,高语义路径通过融合金字塔池化与双重注意力模块来获取准确的上下文信息;低语义路径通过多尺度拼接与类似于注意力机制的信息传递模块来获取清晰的分割边界;最后拼接 2 条路径获取分割结果。在 PASCAL VOC 2012 数据集上的实验中,与主流网络模型相比,该模型的网络参数量仅为 PSPNet 参数量的 4.9%,DeeplabV3+的 4.2%;浮点计算量仅为 PSPNet 浮点计算量的 6.7%,DeeplabV3+的 4.8%;平均交并比略低于 PSPNet 与 DeeplabV3+。所提模型在保证网络性能的同时实现了轻量化。

关键词:语义分割;轻量化;深度可分离卷积;空间金字塔池化;注意力机制
中图分类号:TP391 **文献标志码**:A **文章编号**:1672-6693(2023)06-0095-12

计算机视觉领域中语义分割是一项重要且具有挑战性的任务。层数较深的卷积神经网络在这些具有挑战性的视觉任务中有着优异的性能,例如 VGG-Nets[1]、ResNet[2]、DenseNet[3]等众多网络模型采用加深网络和增强卷积功能的方式提升网络准确率。但这样的网络结构也往往伴随着高昂的计算成本,远远超出了嵌入式平台的算力。而轻量化网络有更加精细的网络设计,可以有效降低计算成本,对语义分割有着不小影响。

轻量化卷积神经网络[4]是基于原有的卷积神经网络进行改进的,最早的轻量化网络 SqueezeNet[5]于 2016 年公开,达到了近似 AlexNet 的效果,但参数量仅为 AlexNet 的 2%。SqueezeNet 仍然使用的普通卷积,而在之后的 MobileNet[6-7]系列利用了更高效的深度可分离卷积并进一步加速了卷积神经网络运算。ShuffleNet[8]提出了通道混洗来完成通道之间的信息融合,之后 GhostNet[9]通过对特征图进行简单的线性运算和特征连接,以更少的参数生成更多的特征图。轻量化卷积神经网络具有结构轻便、计算简单、可移植性强等优点。

使用轻量化卷积神经网络降低模型参数量和计算量的同时,如何保证网络分割的精确性也是挑战。FCN[10]开启了全卷积、端到端的语义分割之路,其中问题也很明显:缺乏精细的结果、上采样的结果模糊以及忽略了图像中的细节以及像素之间的联系。针对这些问题,PSPNet[11]通过并联的不同大小的池化层来获取更大感受野,从而提高获取上下文信息的能力。DeepLab[12-15]、ESPNet[16]等通过串行空洞卷积来融合不同膨胀率的特征,有效还原了特征图的空间细节,并且扩大了网络的感受野。U-Net[17]通过编码-解码的结构,将高水平的语义信息与低水平位置信息进行融合,帮助还原图像的空间维度和像素的位置信息。

注意力机制在改善深度卷积神经网络性能方面发挥了不小的作用。SENet[18]巧妙地构建特征通道之间的联系,通过训练得到每个特征通道的重要程度。ECANet[19]也是通道注意力机制,改进了 SENet 通道注意力机制的计算消耗。相较于只关注通道的注意力机制,CBAM[20]在原有的通道注意力机制上引入了空间注意力机制,从而使网络性能进一步提升。有学者在 DANet[21]网络中所提出了双重注意力机制(dual attention),其中提出的位置注意力机制与通道注意力机制类似于自注意力机制,通过在高语义特征上建立丰富的上下文联系,明显改善了网络分割效果。在网络中合理使用注意力机制,较小的计算量和参数量增加会获得更好的性能。

* **收稿日期**:2022-10-12 **修回日期**:2023-05-12 **网络出版时间**:2023-06-26T08:52
资助项目:重庆市技术创新与应用发展重点项目(cstc2019jscx-mbdxX0061)
第一作者简介:廖恒锋,男,研究领域为计算机视觉语义分割,E-mail:879273352@qq.com;**通信作者**:魏延,男,教授,博士,E-mail:942503422@qq.com
网络出版地址:https://link.cnki.net/urlid/50.1165.N.20230625.1217.012

基于上述工作,本文设计了1种轻量的语义分割模型,较好地兼顾了网络的参数量、计算、性能等3个方面。网络的主干使用轻量化网络MobileNetV2[6],并使用深度可分离卷积替换普通卷积,从而减少整个网络的参数量和计算量;利用金字塔池化模块(pyramid pooling module)[11]进行空间金字塔池化并与改进后的轻量空洞空间金字塔池化(lightweight atrous spatial pyramid poding,L-ASPP)构成双分支来增强模型的鲁棒性,再与双重注意力机制相结合,提出1种高效获取上下文信息的并行双分支模块,在轻量化的同时保证模型的精度;设计了1种轻量化的信息传递模块,将高语义路径上的语义信息传递给含有更多细节的低语义特征图;最后将经过优化的高语义特征图和低语义特征图进行拼接,还原成输入图片大小进行预测。新模型和当前流行的语义分割模型相比具备3个优点:1) 占用内存小,整个网络的参数量仅为2.31×10^6,能够更好地进行部署;2) 计算量低,浮点计算量仅为7.989 G FLOPs;3) 性能良好,在公共数据集PASCAL VOC 2012上平均交并比达到73.75%。

1 深度可分离卷积与特征金字塔池化

1.1 深度可分离卷积

轻量级网络类似于MobileNet系列使用深度可分离卷积将通道域和空间域分开处理,大幅度降低模型所需的计算量和产数量。深度可分离卷积主要分为类似于分组卷积的深度卷积和大小为1×1的逐点卷积,如图1所示。

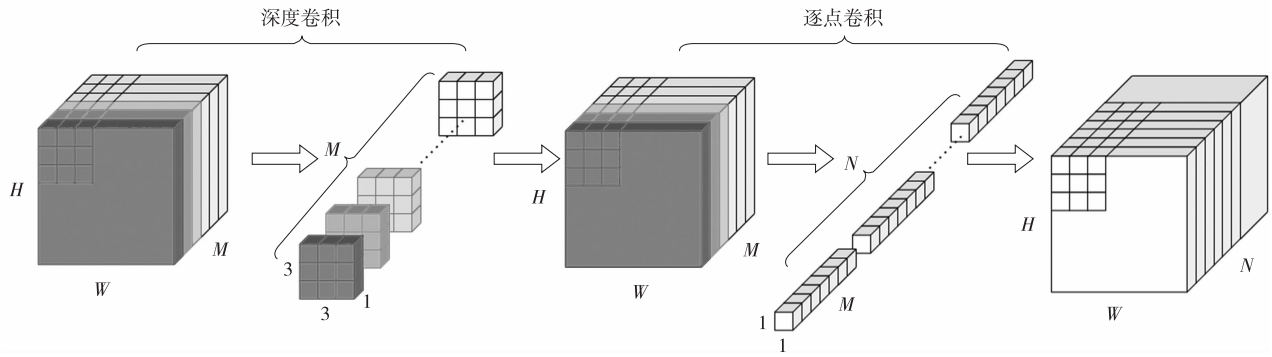


图1 深度可分离卷积

Fig. 1 Depthwise separable convolution

二维平面进行深度卷积操作,1个卷积核对应1个通道。设样本输入为$H\times W\times M$,其中H与W分别为图像的高和宽,M为图像通道数。选用M个$3\times3\times1$的卷积核对每个通道进行卷积操作,得到图像高、宽与通道数完全相同的特征图。然后使用逐点卷积对通道数进行调整,卷积核的个数为输出通道数N,最终得到$H\times W\times N$大小的特征图。

对应的计算量为:

$$H\times W\times3\times3\times M+H\times W\times1\times1\times M\times N。$$

普通卷积对应的计算量为:

$$H\times W\times3\times3\times N\times M。$$

深度可分离卷积的计算量与普通卷积的比值为:

$$\frac{1}{N}+\frac{1}{3^2}。$$

由上述公式可知,使用的卷积核大小为3×3时,深度可分离卷积只有普通卷积计算消耗的$\frac{1}{9}$,极大加快了网络的运算速度。

1.2 特征金字塔池化

在语义分割任务中,网络能够获取更大的感受野标志着能够获取图像中更多的上下文信息。虽然理论上网络深层所获得的感受野大小远超原图尺寸,但是实际感受野远比理论感受野要小,尤其对于深层网络来说,所以需要巧妙地设计来获取有效的全局上下文信息。

1.2.1　空洞空间金字塔池化(atrous spatial pyramid pooling，ASPP)

ASPP 是 DeepLab 系列的代表作，是捕获上下文信息的一种有效方式，它的结构如图 2 所示。

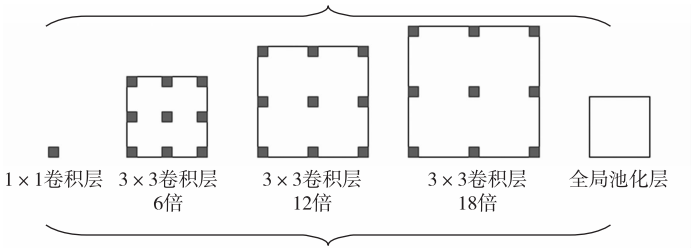


图 2　ASPP 结构

Fig. 2　The structure of ASPP

ASPP 主要是由 1 个 1×1 的卷积层，3 个不同膨胀率的 3×3 的卷积层提取图像多尺度信息以及 1 个全局池化层共 5 个分支组成。其中全局平均池化的输出尺寸为 1×1，通过 1×1 的卷积改变通道数和上采样恢复成输入的尺寸，最后 5 个分支进行拼接融合。ASPP 优化了使用空洞卷积时，由于网格效应导致的局部信息丢失和远距离信息缺少相关性的问题，可以在不使用池化层的前提下获取不同尺度特征信息。

1.2.2　金字塔池化模块

PSPNet 中提出的金字塔池化模块也是一种控制感受野，捕获上下文信息的有效方式，结构如图 3 所示。

将特征图按比例分为 4 个部分，然后通过不同大小的池化来形成不同区域的信息表达(图 3 括号中最上面的部分表示全局尺度的池化，向下依次为 2×2、3×3、6×6)。池化后调整通道数，再将得到的特征上采样至输入特征图的大小拼接在一起。

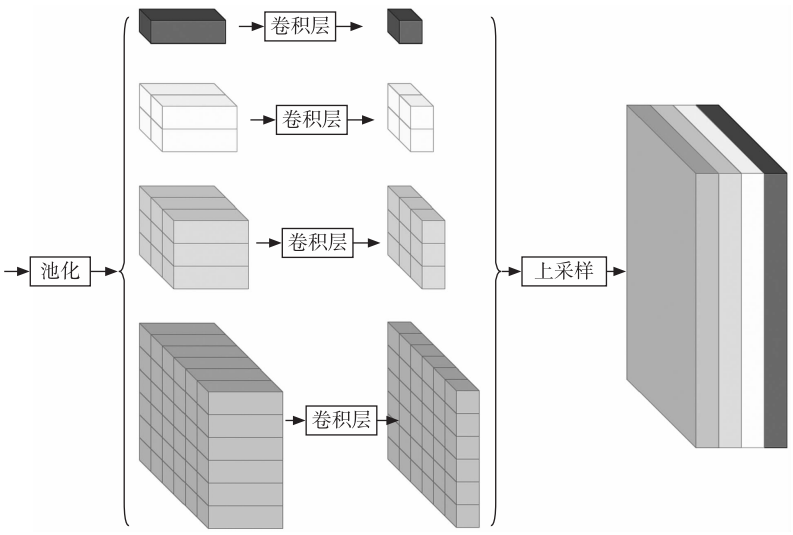


图 3　PPM 结构

Fig. 3　The structure of pyramid pooling module

分析上述结构，可以看出尺度信息来自于不同层次的操作；上下文信息也通过各种尺寸的池化操作组合出更多的感受野区域；通过拼接后 1×1 卷积调整通道数的同时融合了不同池化后通道信息，使得局部和全局的特征表达相互融合。

2　融合金字塔池化和注意力机制的轻量语义分割方法

2.1　整体框架

主干选用轻量化网络 MobileNetV2，整个模型分为高语义与低语义 2 条路径进行，整个网络均采用深度可分离卷积代替常规卷积，见图 4。

1) 高语义路径选取的 MobileNetV2 中最深的、下采样 16 倍后的特征图进行处理,通过双重注意力金字塔池化模块后,作为最终高语义路径的特征图。其中,双重注意力金字塔模块中包含金字塔池化模块、改进后的 L-ASPP 轻量化金字塔池化模块与双重注意力机制。

2) 低语义路径选取 MobileNetV2 处理过程中下采样 4 倍和 8 倍后的特征图,然后进行双线性插值上采样拼接,使用 1×1 的卷积核调整通道数至 24。然后接收信息传递模块的信息相结合,作为最终低语义路径的特征图。

将高语义特征图经过双线性插值上采样,与低语义特征图进行拼接,最后通过 2 个大小为 3×3 的深度可分离卷积进行特征提取送入分割头进行预测。

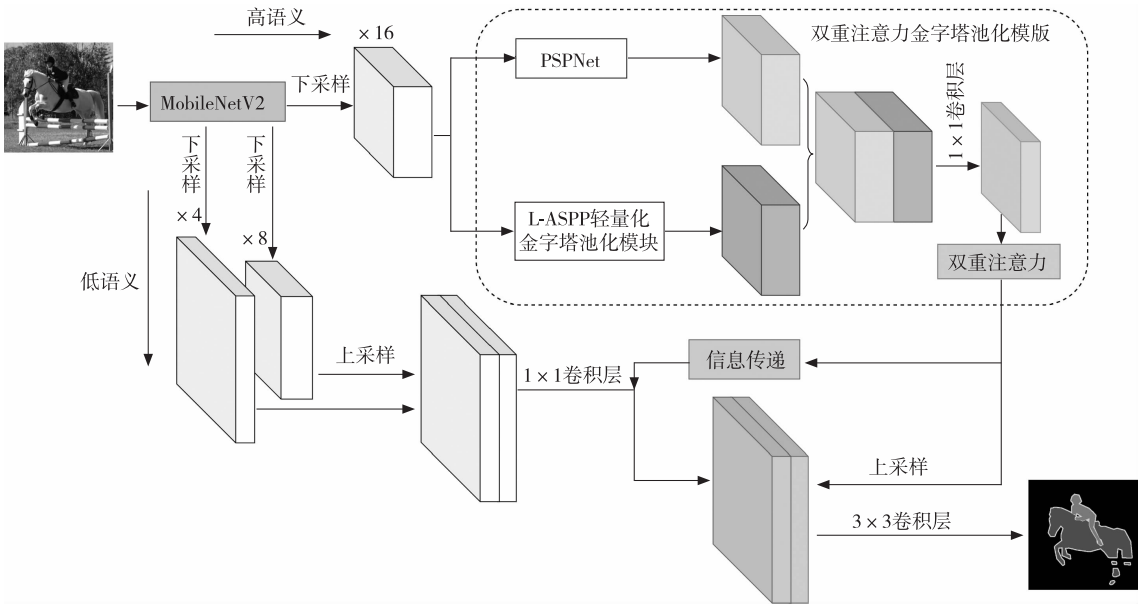


图 4 整体框架

Fig. 4 Overall framework

2.2 双重注意力金字塔池化模块

双重注意力金字塔池化模块包括双金字塔池化模块与双重注意力机制(图 4 虚线框住部分),其中双金字塔池化模块是增大特征图的感受野来获取更多的全局上下文信息,而双重注意力机制是为了挑选所获得的全局上下文信息,让特征图注意到更多重要的上下文信息。

2.2.1 双金字塔池化

双重金字塔池化是金字塔池化模块和改进后的 L-ASPP 模块所并联的结构。在 MobileNetV2 中加入的 ASPP 模块,模型分割精度上提升非常明显,但代价是计算量的增加也特别明显,所以在原 ASPP 模块中进行改进,通过以下 3 点,可以在实现了模型压缩的同时,精度也有所提高。

1) ASPP 模块的输出通道为 256,输出通道数改变为 128,以裁剪通道的方式降低网络计算量。

2) 去除原 ASPP 模块中的全局池化,并联金字塔池化模块,减弱化通道数减少对整个模型精度的影响,并且多分支的结构让模型的鲁棒性更强。

3) 将 ASPP 中的空洞卷积替换为深度可分离空洞卷积,进一步降低网络计算量。深度可分离空洞卷积由深度空洞卷积和逐点卷积组成。将改进后的 L-ASPP 结构如图 5 所示。

2.2.2 双重注意力机制

通道注意力机制与位置注意力机制并联得到双重注意力机制,进一步强化通过双重金字塔模块后的特征信息表达,使特征图更加注意一些重要的上下文信息。图 6 为位置注意力机制,图 7 为通道注意力机制。

1) 位置注意力机制的具体实现方式。

位置注意力机制可以捕捉任意 2 个位置之间的上下文信息。所有的位置信息,两两之间都有 1 个权重 γ,这个 γ 的值由 2 个位置之间的相似性来决定,可以通过矩阵的乘法使像素间产生联系。

对于位置注意力的实现，首先将特征图 $\boldsymbol{A}(C\times H\times W)$ 输入到卷积模块中，生成 $\boldsymbol{B}(C\times H\times W)$ 和 $\boldsymbol{C}(C\times H\times W)$，将 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 转化为成 $(C\times N)$ 维度，其中 $H\times W$ 就是像素点的个数。

随后将矩阵 $\boldsymbol{B}$ 转置后和矩阵 $\boldsymbol{C}$ 相乘，得到的结果输入到 softmax 中，得到位置注意力图 $\boldsymbol{S}$。其中，2 个位置相似度越高，S_{ji} 的值就越大，计算 S_{ji} 公式为：

$$s_{ji}=\frac{\exp(B_i\cdot C_j)}{\sum_{i=1}^{N}\exp(B_i\cdot C_j)}。$$

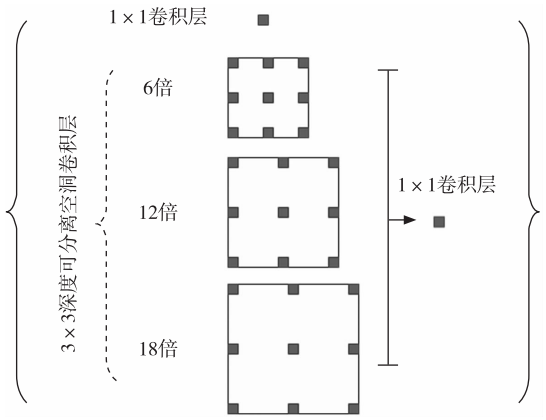


图 5　L-ASPP 结构

Fig. 5　The structure of L-ASPP

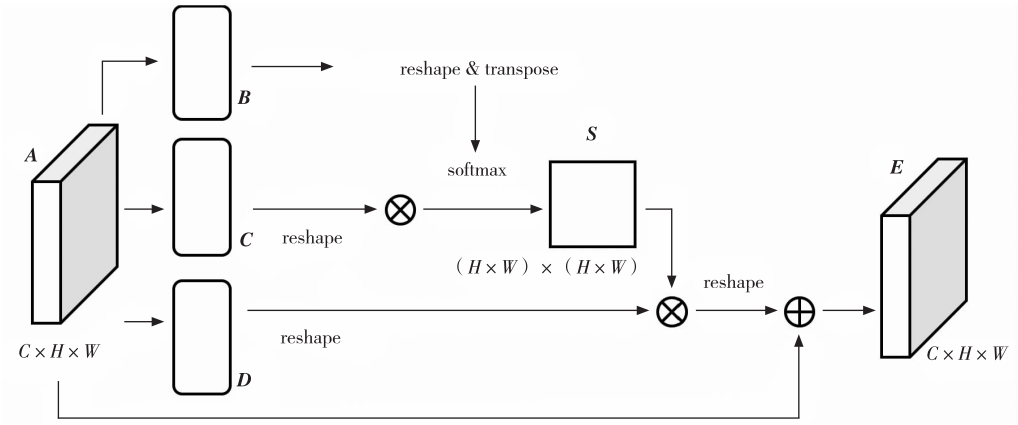


图 6　位置注意力机制

Fig. 6　Position attention module

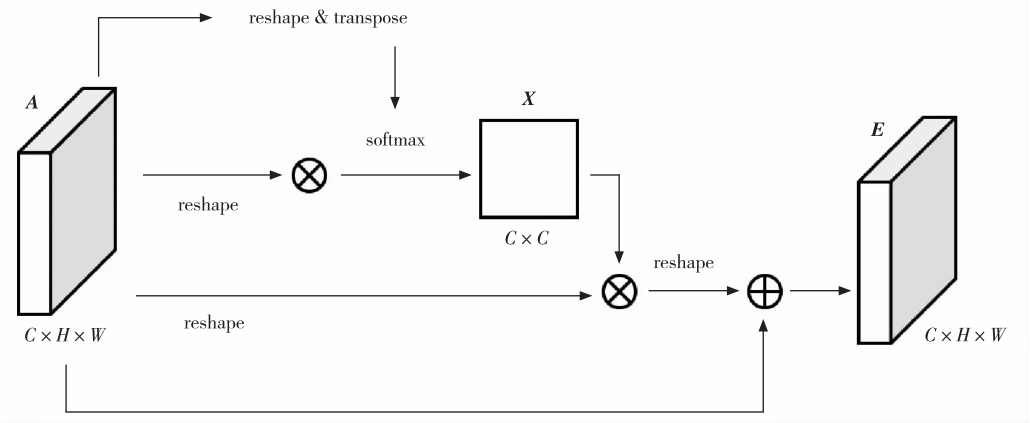


图 7　通道注意力机制

Fig. 7　Channel attention module

同样,$\boldsymbol{A}$ 输入到另一个卷积层生成新的特征映射 $\boldsymbol{D}(C\times H\times W)$,reshape 成 $C\times N$ 后与上述的空间注意力图 $\boldsymbol{S}$ 的转置相乘,这样就得到了 $C\times N$ 大小的矩阵,再将这个矩阵 reshape 成原来的 $C\times H\times W$ 大小,再乘以系数 α,最后与输入的特征图 $\boldsymbol{A}$ 相加,实现了空间注意力机制。参数 α 的值是通过学习不断更新的,初始化为 0。特征 E 计算公式为:

$$E_j=\alpha\sum_{i=1}^{N}(s_{ji}D_i)+A_j。$$

2) 通道注意力机制的具体实现方式。

通道注意力机制可以捕捉通道维度上的上下文信息。增强重要特征通道的权重可以有效地提高分割效果。通过计算 1 个权重因子,对每个通道进行加权,突出重要的通道从而增强特征表示。

通道注意力机制的实现与位置注意力机制类似,特征图 $\boldsymbol{A}(C\times H\times W)$ 转置成 $C\times N$ 的矩阵,分别经过矩阵乘法、softmax 函数得到注意力图 $\boldsymbol{X}(C\times C)$,计算公式为:

$$x_{ji}=\frac{\exp(A_i\cdot A_j)}{\sum_{i=1}^{C}\exp(A_i\cdot A_j)}$$

最后注意力图 $\boldsymbol{X}$ 与转变成 $C\times N$ 的矩阵 $\boldsymbol{A}$ 进行矩阵乘法,输出 $(C\times N)$ 转变成 $C\times H\times W$ 和原始特征图 $\boldsymbol{A}$ 进行加权,β 是一个可学习参数,初始化为 0,则上述过程的计算公式为:

$$E_j=\beta\sum_{i=1}^{C}(x_{ji}A_i)+A_j$$

2.3 语义信息传递模块

多尺度特征拼接能还原图片更多的细节信息,将下采样 4 倍和下采样 8 倍通过双线性插值上采样进行拼接,作为低语义分支加以使用。为了充分利用低语义特征图高分辨率含有更多细节的特点,参照 Coordinate Attention[22] 的思路,提出了类似于注意力机制的语义信息传递模块,利用高语义特征含有高语义信息的特点,通过精确的位置信息对通道关系和长期依赖性进行编码,将语义信息传递给高分辨率低语义的特征进行语义补充,从而使整个网络在细节信息上分割能力更强。语义信息传递模块结构如图 8 所示。

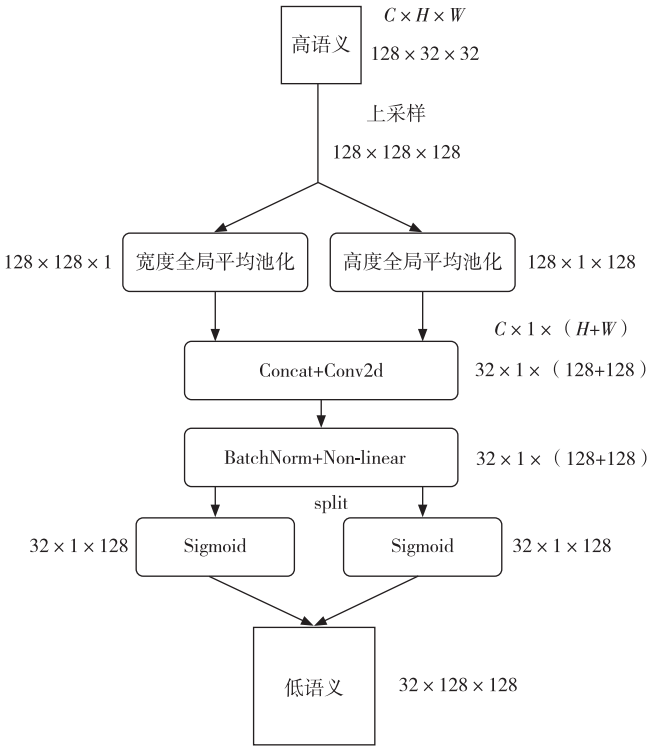


图 8 语义信息传递模块

Fig. 8 Semantic information transfer module

具体实现：首先将高语义特征图双线性插值上采样为 128×128 大小与低语义特征图保持一致。为了获取高语义图像宽度和高度上的信息，先将输入特征图分为宽度和高度 2 个方向分别进行全局平均池化，然后获取到宽度和高度上的特征图，计算公式为：

$$z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i\leqslant W}x_c(h,i),$$

$$z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j\leqslant H}x_c(j,w)。$$

随后将宽度和高度 2 个方向的特征图拼接在一起，送入卷积核为 1×1 的卷积模块在通道上进行处理，之后将维度降低为低语义特征图的相同通道数 32，接着将特征图 $\boldsymbol{F}_1$ 经过批量归一化处理送入非线性激活函数得到 32×1×(128+128)的特征图 $\boldsymbol{f}$，计算公式为：

$$f=\delta(F_1([z^h,z^w]))。$$

通过 split 操作得到特征 $\boldsymbol{f}^{\mathrm{h}}$ 与 $\boldsymbol{f}^{\mathrm{w}}$ 分别送入 Sigmoid 激活函数，得到特征图在高度和宽度上的注意力权重 $\boldsymbol{g}^{\mathrm{h}}$ 和在宽度方向的注意力权重 $\boldsymbol{g}^{\mathrm{w}}$，公式如下所示：

$$\boldsymbol{g}^{\mathrm{h}}=\sigma(\boldsymbol{f}^{\mathrm{h}}),$$

$$\boldsymbol{g}^{\mathrm{w}}=\sigma(\boldsymbol{f}^{\mathrm{w}})。$$

最后通过乘法加权计算，最终在宽度和高度方向上将得到带有高语义信息传递给低语义特征图上，计算公式为：

$$y_c(i,j)=x_c(i,j)\times g_c^{\mathrm{h}}(i)\times g_c^{\mathrm{w}}(j)。$$

3　实验与结果分析

3.1　实验设置

数据集选用语义分割常用的 PASCAL VOC2012 扩充版，对模型进行了大量实现和分析，它包含 20 个对象类别和 1 个背景类别，其中训练集扩充至 10 582 幅图像，验证集 1 449 幅图像，测试集 1 456 幅图像。系统环境为 Windows 11；CPU 为 Intel i5-12400F；内存大小为 16 GB；GPU 为 GeForce RTX3060，显存大小为 12 GB。基于开源的 PyTorch 1.11.0；CUDA 11.3.1；CUDNN 8.2.0.53 深度学习框架实现。

评价指标：实验通过计算量与参数量来描述网络计算复杂度。在语义分割中通常使用平均交并比(mean intersection over union，MIoU)作为衡量精度的标准，记为 σ_{MIoU}。计算所有类别交集和并集之比的平均值来评估网络的性能。以 PASCAL VOC 2012 数据集为例，其中包含 21 个类别，分别对每个类别求 IoU，令 n 表示类别，$n+1$ 表示加上背景类，i 表示正式值，j 表示预测值，p_{ij} 表示将 i 预测为 j，则某一类别的 MIoU 可按如下公式计算：

$$\sigma_{\mathrm{MIoU}}=\frac{1}{n+1}\sum_{i=0}^{n}\frac{p_{ii}}{\sum_{i=0}^{n}p_{ij}+\sum_{i=0}^{n}p_{ij}-p_{ii}}。$$

数据增强：在训练时，设定裁剪尺寸为 512×512，对图像进行缩放并且进行长和宽的扭曲；将图像多余的部分加上灰条；翻转图像；高斯模糊。

参数设置：主干网络 MobileNetV2 使用在 ImageNet 上预训练权重初始化模型参数，这样能减少模型的训练时间，加快模型收敛速度。随机梯度 SGD 优化器和动量法(动量参数设置为 0.9)结合进行优化，学习率下降方式遵循余弦曲线的方式下降，最大学习率为 0.007，最小学习率为 0.000 07，以及权值衰减率 0.000 1，可防止模型过拟合。损失函数采用交叉熵损失函数。考虑到目前环境计算资源的问题，设定输出最大下采样倍率为 16 和批次大小为 8，训练次数为 200 轮次。

3.2　实验结果

本文模型与其他模型在 PASCAL VOC 2012 数据集上的实验结果如表 1 所示。

由表 1 可得，与 FCN 和 BiSeNet 相比，本文模型从 3 个方面都优于这 2 个模型，大小与 PSPNet 替换主干网络为 MobieNetV2 的模型大小接近，MIoU 的值相比高出 5.28%。而与替换主干网络为 MobieNetV2 的

DeeplabV3+,模型更小的同时,MIoU的值却高出2.03%。与HrnetV2_w18相比较,MIoU的值相差不大,但模型大小不足HrnetV2_w18的$\frac{1}{4}$。与较大的PSPNet+ResNe50和DeeplabV3++ Xception模型相比,MIoU的值分别相差5.27%与3.22%,但模型大小只有它们的5%左右。总之,本文模型很好地兼顾了网络的参数量、计算量与性能,实现了轻量且高效的语义分割。

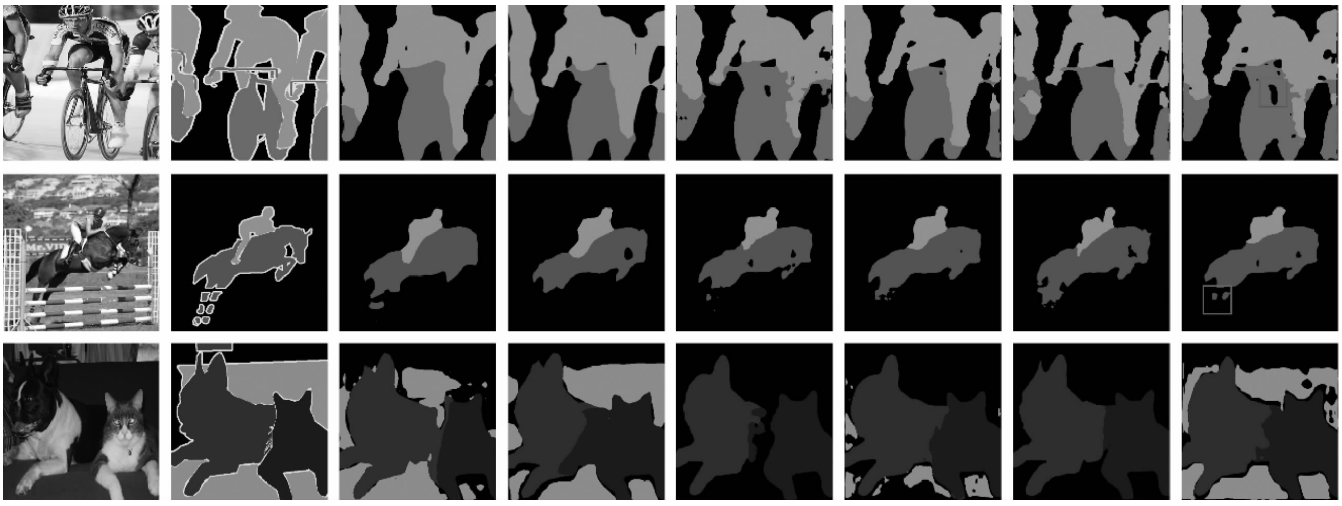
表1 与其他模型在PASCAL VOC 2012数据集上的对比结果

Tab.1 Comparison results with other models on the PASCAL VOC 2012 dataset

模型	主干网络	参数量	计算量	σ_{MIoU}/%
FCN-8s[10]	VGG16	134.68×10^6	321.343	62.48
BiSeNet[23]	ResNet18	49.43×10^6	13.651	65.34
PSPNet[11]	MobileNetV2	2.38×10^6	**5.873**	68.47
PSPNet[11]	ResNe50	46.72×10^6	118.128	**79.02**
DeeplabV3+[14]	MobileNetV2	5.82×10^6	52.846	71.72
DeeplabV3+[14]	Xception	54.71×10^6	166.342	76.97
HrnetV2_w18[24]		9.64×10^6	37.187	73.22
本文模型	MobileNetV2	$\mathbf{2.31\times10^6}$	7.989	73.75

注:加黑的数据表示模型对比参数量、计算量、MIoU中最优的值。

在PASCAL VOC 2012数据集上随机抽取3张图像(图9a)进行预测实验,为了增大模型分割难度,选择复杂的类别更多的图像进行分割,MIoU值相近的模型分割结果见图9c~h。从结果中可以看出,第1幅图像,本文模型(图9h)在自行车和人的的分割细节上与真实标签(图9b)相近;第2幅图像,相较于其他模型,本文模型(图9h)成功还原了部分马腿细节,最接近真实标签(图9b);第3幅图像,本文模型(图9h)也成功识别到沙发,与图9c、e、f和g相比效果更加突出,略逊色于图9d。综上所述,本文模型无论是对上下文信息的捕获,还是对图像细小部分分割都有着不错的效果。



a 原图 b 真实标签 c PSPNet+MobileNetV2 d PSPNet+ResNe50 e DeeplabV3++MobileNetV2 f DeeplabV3++Xception g HrnetV2_w18 h 本文模型

图9 不同模型的分割结果

Fig.9 Segmentation results of different models

3.3 消融实验

3.3.1 高低语义双分支架构实验

对模型采取高低语义双分支架构进行验证,由于MobileNetV2模型的特点,下采样16倍的过程中会产生很

多有价值的特征图，分别为第 2(16，256×256)、4(24，128×128)、7(32，64×64)和 14(160，32×32)层，括号中第 1 个数字为特征图通道数，相乘的 2 个数表示卷积核大小。选择第 2、4、7 层进行多尺度拼接进行实验，因为第 14 层特征图与高语义特征图的大小均为 32×32，所以舍去这一层。本实验上采样方式均为双线性插值，拼接方式为高语义特征上采样加第 7 层，再上采样加上第 4 层，再上采样加上第 2 层表示为 $H+7+4+2$。因为怀疑第 2 层通过的卷积处理太少，语义信息不够丰富，所以去除第 2 层特征图，表示为 $H+7+4$。高低语义双分支的拼接方式是先将第 4 层与第 7 层拼接，再与高语义拼接表示为 $4+7+H$。并在 PASCAL VOC2012 验证集上进行了对比试验，结果如表 2 所示。

表 2　高低语义双分支实验

Tab. 2　High and low semantic dual-branch experiments

模型	σ_{MIoU}/%
$H+7+4+2$	70.16
$H+7+4$	72.46
$4+7+H$	**72.72**

注：加黑的数据表示最优的 MIoU 值。

由表 2 可得结论：第 2 层特征图对整个模型的精度有影响，同时还会增大计算量。与类似于 U-Net 的拼接方式($H+7+4$)相比，使用高低语义双分支的拼接方式精度更高。

3.3.2　模块有效性验证

双重注意力金字塔池化模块中金字塔池化模块部分和原 PSPNet 相同，所以替换 PSPNet 的主干网络为 MobileNetV2 作为基准模型 Baseline Model (BM)。验证双重注意力金字塔池化模块以及语义信息传递模块的有效性实验结果如表 3 所示，其中：M+PPM+ASPP 表示基准模型并联未处理的 ASPP 模块；上文提到的改进后的 L-ASPP 模块表示为 L-ASPP；双重注意力机制表示为 DA；高低语义结合表示为 Mul；语义信息传递模块表示为 T。

表 3　消融实验结果

Tab. 3　Ablation experiment results

模型	深度可分离卷积	参数量	计算量	σ_{MIoU}/%
BM	否	2.377×10^6	**5.873**	68.47
M+PPM+ASPP	否	5.061×10^6	18.730	72.03
M+PPM+L-ASPP	是	$\mathbf{2.242\times10^6}$	6.464	70.08
M+PPM+L-ASPP+DA	是	2.263×10^6	6.589	71.16
M+PPM+L-ASPP+DA+Mul	是	2.306×10^6	7.978	72.72
M+PPM+L-ASPP+DA+Mul+T	是	2.310×10^6	7.989	**73.75**

注：加黑的数据表示消融实验结果中各项对应最优值。

通过表 3 可知：L-ASPP 相较于 ASPP 在参数量和计算量上有很大幅度降低，但牺牲了接近 2%的精度。双重注意力机制在模型中以较少的参数量和计算量的增加换来了精度提升 1.06%。高低语义结合方式使得模型精度再次提升至 72.72%，此时已经超过 M+PPM+ASPP 的精度，但参数量和计算量只有它的一半不到。最后在此基础上，语义信息传递模块(T)以较小开销换来了精度提高 1.03%；表 3 中 M+PPM+L-ASPP+DA+Mul+T 为本文模型，合理使用深度可分离卷积压缩模型，最终模型仅 2.310×10^6 参数量，不足 8GFOLPs 的计算量，MIoU 提升到 73.75%，与基准模型 BM 相比提高 5.28%。

表 4 报告了基准模型，引入双重注意力金字塔池化模块后，以及引入语义信息传递模块后在 PASCAL VOC 2012 数据集上测试 21 个类的交并比(intersection over union，IoU)的对比结果。

表4 PASCAL VOC 2012验证集上各类别的IoU

Tab.4 IoU of each category on the PASCAL VOC 2012 validation set

类别	BM	M+PPM+L-ASPP+DA+Mul	M+PPM+L-ASPP+DA+Mul+T
background	0.92	0.93	0.93
aero plane	0.78	0.85	0.85
bicycle	0.38	0.41	0.41
bird	0.76	0.84	**0.86**
boat	0.59	0.61	0.61
bottle	0.67	0.73	**0.75**
bus	0.88	0.93	0.93
car	0.81	0.85	0.85
cat	0.85	0.88	**0.89**
chair	0.36	**0.38**	0.36
cow	0.69	0.80	**0.84**
dining table	0.49	0.56	0.56
dog	0.72	0.82	0.82
horse	0.75	0.76	**0.81**
motorbike	0.78	**0.81**	0.80
person	0.78	0.81	0.81
potted plant	0.50	0.51	**0.54**
sheep	0.82	0.84	**0.86**
sofa	0.41	0.48	**0.49**
train	0.81	0.81	**0.82**
monitor	0.63	0.67	**0.69**

注:加黑的数据表示最优的IoU值。

4 结语

在MobileNetV2作为主干网络与深度可分离卷积的加持下,模型有着较低的参数量和计算量;双重注意力金字塔池化模块捕获了广泛的上下文信息;高低语义双路径让模型在小尺度目标上的分割效果更佳,同时也加强了模型的空间信息细节的保留;语义信息传递模块强化了低语义路径上特征图的语义信息,进一步增强了模型在小尺度目标的分割效果。实验结果表明,本文模型占用内存少,计算效率高且分割效果良好,在计算效率和分割精度之间达到了较好平衡,也满足了移动和嵌入式设备上的要求。后续将从网络的主干网络和整体架构继续进行优化,以及进一步研究激活函数对网络的影响。

参考文献:

[1] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2022-10-12]. https://arxiv.org/pdf/1409.1556v6.pdf.

[2] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition,June 26-July 1,2016,Las Vegas,Nevada. Piscataway:IEEE,2016:770-778.

[3] HUANG G,LIU Z,Van Der MAATEN L,et al. Densely connected convolutional networks[C]//30th IEEE Conference on Computer Vision and Pattern Recognition,July 21-26,2017,Honolulu,Hawaii. Piscataway:IEEE,2017:2261-2269.

[4] 马金林，张裕，马自萍，等. 轻量化神经网络卷积设计研究进展[J]. 计算机科学与探索，2022，16(3)：512-528.
MA J L，ZHANG Y，MA Z P，et al. Research progress of lightweight neural network convolution design[J]. Journal of Frontiers of Computer Science and Technology，2022，16(3)：512-528.

[5] IANDOLA F N，HAN S，MOSKEWICZ M W，et al. SqueezeNet：AlexNet-level accuracy with 50x fewer parameters and＜0.5 MB model size[EB/OL]. (2016-11-04)[2022-10-12]. http://www.arxiv.org/abs/1602.07360

[6] HOWARD A G，ZHU M l，CHEN B，et al. Mobilenets：efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17)[2022-10-12]. https://arxiv.org/pdf/1704.04861.pdf.

[7] SANDLER M，HOWARD A，ZHU M L，et al. Mobilenetv2：Inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 18-23，2018，Salt Lake City，UT，USA. Piscataway：IEEE，2018：4510-4520.

[8] ZHANG X Y，ZHOU X Y，LIN M X，et al. Shufflenet：an extremely efficient convolutional neural network for mobile devices [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 18-23，Salt Lake City，UT，USA. Piscataway：IEEE，2018：6848-6856.

[9] HAN K，WANG Y H，TIAN Q，et al. Ghostnet：more features from cheap operations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 13-19，2020，Seattle，WA，USA. 2020：1580-1589.

[10] SHELHAMER E，LONG J，DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(4)：640-651.

[11] ZHAO H S，SHI J P，QI X J，et al. Pyramid scene parsing network[C]//30th IEEE Conference on Computer Vision and Pattern Recognition，July 21-26，2017，Honolulu，Hawaii. Piscataway：IEEE，2017：2881-2890.

[12] 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.

[13] CHEN L C，PAPANDREOU G，SCHROFF F，et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-11-05)[2022-10-12]. https://arxiv.org/pdf/1706.05587.pdf.

[14] CHEN L C，ZHU Y K，PAPANDREOU G，et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//FERRARI V，HEBERT M，SMINCHISESCU C，et al. Lecture Notes in Computer Science. Cham：Springer，2018，11211：801-818.

[15] 何雪东，宣士斌，王款，等. 融合累积分布函数和通道注意力机制的 DeepLabV3＋图像分割算法[J]. 计算机应用，2023，43(3)：936-942.
HE X D，XUAN S B，WANG K，et al. DeepLabV3＋ image segmentation algorithm fusing cumulative distribution function and channel attention mechanism[J]. Journal of Computer Applications，2023，43(3)：936-942.

[16] MEHTA S，RASTEGARI M，CASPI A，et al. Espnet：Efficient spatial pyramid of dilated convolutions for semantic segmentation[C]//15th European Conference on Computer Vision，September 8-14，2018，Munich，Germany. Berlin：Springer-Verlag，2018：552-568.

[17] RONNEBERGER O，FISCHER P，BROX T. U-net：Convolutional networks for biomedical image segmentation[C]//18th International Conference on Medical Image Computing and Computer-Assisted Intervention，October 5-9，2015，Munich，Germany. Cham：Springer，2015：234-241.

[18] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 18-23，2018，Salt Lake City，UT，USA. Piscataway：IEEE，2018：7132-7141.

[19] WANG Q L，WU B G，ZHU P F，et al. ECA-Net：efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 13-19，2020，Seattle，WA，USA. 2020：11531-11539.

[20] WOO S，PARK J，LEE J Y，et al. CBAM：convolutional block attention module[C]//15th European Conference on Computer Vision，September 8-14，2018，Munich，Germany. Berlin：Springer-Verlag，2018，11211：3-19. https://doi.org/10.1007/978-3-030-01234-2_1.

[21] FU J，LIU J，TIAN H J，et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 15-20，2019，Long Beach，CA，USA. 2019：3146-3154.

[22] HOU Q B，ZHOU D Q，FENG J S. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 20-25，2021，Nashville，TN，USA. 2021：13713-13722.

[23] YU C Q, WANG J B, PENG C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]//15th European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin: Springer-Verlag, 2018: 325-341.

[24] SUN K, ZHAO Y, JIANG B R, et al. High-resolution representations for labeling pixels and regions[EB/OL]. (2019-04-09) [2022-10-12]. https://arxiv.org/pdf/1904.04514.pdf.

Lightweight Semantic Segmentation Method Fusing Pyramid Pooling and Attention Mechanisms

LIAO Hengfeng, WEI Yan, DU Hanyu

(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: Semantic segmentation is widely used in medical image segmentation, unmanned driving, remote sensing image segmentation and other computer vision tasks. In order to solve the problem of deploying embedded platforms with limited computing power and hardware storage, a lightweight semantic segmentation model is proposed by considering three aspects of network parameters, calculation and performance. The model takes the lightweight network MobileNetV2 as the backbone, depthwise separable convolution is applied to compress the model, which is divided into two paths of high and low semantic features for derivation. High-semantic features can obtain accurate contextual information through the dual attention pyramid pooling module. Low-semantic features can obtain clearer segmentation boundary by multi-scale feature stitching and high semantic information transmission. Finally, high and low semantic features are fused to obtain the segmentation results. In the experiments on PASCAL VOC 2012 dataset, compared with the mainstream network model, the number of network parameters of model is 2.31×10^6, which is only 4.9% of PSPNet and 4.2% of DeeplabV3+. The number of floating point computing is 7.989GFLOPs, only 6.7% of PSPNet's floating point computing and 4.8% of DeeplabV3+. The mean intersection over union is 73.75%, slightly lower than PSPNet and DeeplabV3+. It achieves a better balance between computational efficiency and segmentation accuracy.

Keywords: semantic segmentation; lightweight; depthwise separable convolution; spatial pyramid pooling; attention mechanism

(责任编辑 黄 颖)