第52卷 第2期 上海师范大学学报(自然科学版) Vol. 52, No. 2
2023年4月 Journal of Shanghai Normal University (Natural Sciences) Apr., 2023
DOI: 10.3969/J.ISSN.1000-5137.2023.02.003

融合注意力机制的PSPnet多目标语义分割

张 帅, 杨春夏*
(上海师范大学 信息与机电工程学院,上海 201418)

摘 要: 采用加强特征提取网络为MobileNetV2的融合多特征金字塔场景解析网络(PSPnet)来实现复杂场景下的图像语义分割.相对于深度残差网络ResNet50和MobileNetV1,引入了线性瓶颈结构和反向残差结构,利用金字塔池化模块(PPM)来处理不同层级的图像特征信息,并将其进行特征拼接,有效避免了不同分割尺寸下,子区域之间关键特征信息的缺失.在此基础上,引入注意力机制模块,结合通道注意力机制(CAM)和空间注意力机制(SAM),进一步提高分割精度.实验结果表明:该方法可以提高图像识别的准确率,并节省训练时间.

关键词: 语义分割; 金字塔池化模块(PPM); 注意力机制; 特征融合
中图分类号: TP 391.4 **文献标志码:** A **文章编号:** 1000-5137(2023)02-0170-06

Multi-objective semantic segmentation based on PSPnet with attention mechanisms

ZHANG Shuai, YANG Chunxia*
(College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China)

Abstract: A pyramid scene parsing network (PSPNet) with MobileNetV2 as enhanced feature extraction network was adopted to achieve semantic segmentation of images in complex scenes. Compared with the deep residual network ResNet50 and MobileNetV1, linear bottlenecks and inverted residuals were introduced and the pyramid pooling module (PPM) was used to process the image feature information of different layers and to feature stitching, which could avoid missing key feature information between sub-regions under different segmentation sizes effectively. On this basis, the attention mechanism module was introduced to further improve the segmentation accuracy by combining the channel attention mechanism (CAM) with the spatial attention mechanism(SAM). The experimental results verified that the method could achieve the expected goals, improve the accuracy of image recognition and reduce the training time.

Key words: semantic segmentation; pyramid pooling module (PPM); attention mechanisms; feature integration

收稿日期: 2022-12-23

作者简介: 张 帅(1995—), 男, 硕士研究生, 主要从事基于深度学习的目标检测方面的研究. E-mail: 892486535@qq.com

* **通信作者:** 杨春夏(1988—), 女, 副教授, 主要从事电磁场与微波技术方面的研究. E-mail: chunxiay@shnu.edu.cn

引用格式: 张帅, 杨春夏. 融合注意力机制的PSPnet多目标语义分割 [J]. 上海师范大学学报(自然科学版), 2023, 52(2):170-175.

Citation format: ZHANG S, YANG C X. Multi-objective semantic segmentation based on PSPnet with attention mechanisms [J]. Journal of Shanghai Normal University(Natural Sciences), 2023,52(2):170-175.

语义分割是对图像中某一目标部分的像素进行分类，并给出指定的目标分类标签.现实生活场景中，复杂的背景因素以及目标类型给语义分割过程增加了难度.因此，应尽可能地利用图像特征信息进行目标区域特征提取与特征融合，使目标分类结果更为准确.融合多特征金字塔场景解析网络PSPnet模型[1]中，利用金字塔池化模块(PPM)，对复杂场景下包含多重信息特征层的图像做了更为高效的深度分解[2].本文作者在提取图像各部分特征信息的基础上，预测了全局特征与部分特征之间的信息关联，解析了高维信息，同时引入了注意力机制，显著提升了复杂场景下图像语义分割的质量.

1　语义分割网络模型

1.1　融合注意力机制的PSPnet网络模型

在复杂场景下，图像的语义分割的任务量由低像素单一化目标向更高像素、更多目标类别进一步增加，因此需要更为精细、高效的图像特征提取与特征融合方法加以实现.传统的图像解析网络中，以全卷积网络(FCN)为基本模型，解析网络框架，在处理复杂场景中的语义分割问题时，经常出现像素分类错误的情况[3].在融合注意力机制的PSPnet模型中，PPM可以充分整合不同目标范围相互联系的信息要素，提升了模型的效率和准确度，同时引入注意力机制模块，对感兴趣的目标区域进行特征加强，并通过特征融合将3部分特征进行拼接，提高了模型的整体精度，如图1所示.

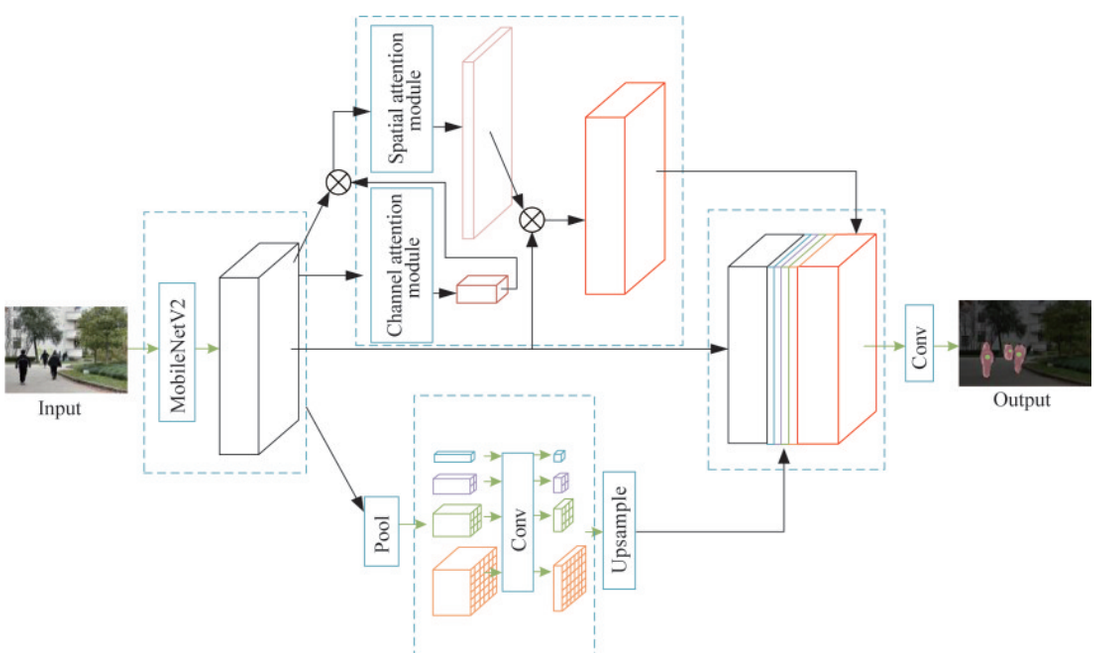

图1　融合注意力机制的PSPnet模型

1.2　加强特征提取网络

图1中，MobileNetV2为PSPnet的加强特征提取网络，是以深度方向可分离卷积为基础的一种轻量级深层神经网络，和标准卷积方式在操作上存在显著的不同之处.深度方向可分离卷积是通过改变标准卷积层数的方式，把标准卷积中的单层卷积扩展为两层卷积：在纵向卷积方面，使用单个卷积核对所有输入的通道执行滤波操作[4]；在点卷积方面，利用各通道之间特征的线性组合，形成包含更多信息的新特征，和标准卷积相比，减少了计算量，提高了计算精度.MobileNetV2在MobileNetV1的基础上做了相应改进，引入了反向残差结构和线性瓶颈结构.在反向残差结构中，在3×3卷积网络之前，利用1×1卷积对目标图像进行升维、扩张操作；在3×3卷积网络之后，再利用1×1卷积操作，对目标图像进行降维操作，以提升精确度，减少特征信息丢失.由于Relu函数在低维空间上会造成部分信息的丢失，引入线性瓶颈结构消除其对特征的破坏[5].

2 主干网络及其损失函数

网络中的注意力机制模块包括两个部分:通道注意力机制(CAM)和空间注意力机制(SAM),从特征深度和特征层两个作用域出发,求解注意力机制感兴趣的区域,如图1所示.CAM采用全局平均池化和最大池化融合的方式[6],将特征图的空间维度进行压缩,利用Sigmoid函数对结果进行处理,得到每个通道上的权重,并与输入特征层的通道逐个相乘,得到特征图.SAM是CAM的信息扩充,同样采用全局平均池化和最大池化融合的方式,对每个特征点的通道方向进行计算,将结果叠加起来,形成特征描述符,并进行卷积操作,依次对通道数进行调整,再利用Sigmoid函数得到每一个特征点的权重值,并与输入特征层的通道逐个相乘,得到特征图,用于后续的特征拼接操作.

PPM将经MobileNetV2处理的特征图重新按4种尺寸的特征进行融合[7],4种尺寸的网格数代表了不同层级的处理过程,最上层的处理最为粗略,通过全局池化的方式加以实现.对输入的特征图进行子区域划分,在划分后的每个子区域上进行池化操作,生成一个包含4种层级的模块,大小分别为1×1,2×2,3×3和6×6.经测试,选择平均池化的效果优于最大值池化[8].PPM中,4种层级的池化操作将生成大小不同的特征图.为了可以将上述特征进行融合,在每个层级上均进行卷积核为1×1的卷积操作.假设该层级的维数是n,利用特征图的语境降维到原始特征的$\frac{1}{n}$,再利用双线性插值对卷积得到的维数较低的特征图进行上采样操作[9],上采样后的特征图和原始特征图的尺寸一致,以便后期的特征拼接.图像处理中的坐标系如图2所示,在图像中插入一点$P(x,y)$,Q_{11},Q_{12},Q_{21}和Q_{22}是距离P点最近的4个像素点,双线性插值公式如下:

$$\begin{aligned} f(P)=&\frac{(x_2-x)(y_2-y)}{(x_2-x_1)(y_2-y_1)}f(Q_{11})+\frac{(x-x_1)(y_2-y)}{(x_2-x_1)(y_2-y_1)}f(Q_{21})+\\ &\frac{(x_2-x)(y-y_1)}{(x_2-x_1)(y_2-y_1)}f(Q_{12})+\frac{(x-x_1)(y-y_1)}{(x_2-x_1)(y_2-y_1)}f(Q_{22}) \end{aligned}. \tag{1}$$

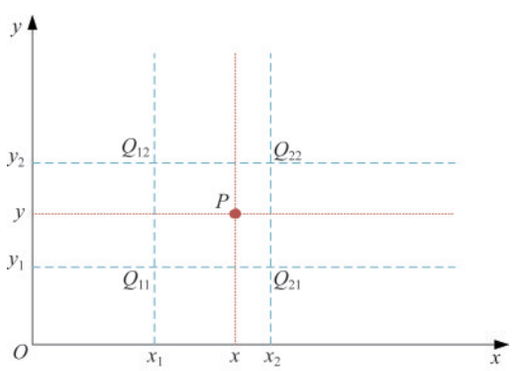

图2 双线性插值示例

PSPnet中,引入两种损失函数:交叉熵损失函数和Dice Loss函数,共同优化特征参数.交叉损失熵函数主要用于网络模型的分类模块,判断像素点的分类是否准确[10].在深度学习模型中,常用$p(x)$表示样本真实的分布情况,用$q(x)$表示模型预测的分布情况,采用相对熵来判断真实分布和预测分布之间的差异,相对熵

$$D_{\mathrm{KL}}(p\parallel q)=\sum_{i=1}^{n}p(x_i)\log_2\left(\frac{p(x_i)}{q(x_i)}\right), \tag{2}$$

其中,n代表像素点的数量.相对熵的值越小,真实分布和预测分布之间的差异越小.

信息熵H表示在预测结果生成之前,对可能产生的信息量的期望,用来衡量该特征图中像素点的不

确定性，信息熵的数值越大，表示信息不确定性越高，目标场景越复杂.其计算公式为：

$$H(p)=-\sum_{i=1}^{n}p(x_i)\log_2 p(x_i). \tag{3}$$

在Dice Loss函数中，Dice系数(D_{ice})用来计算真实样本和预测样本之间的相似度，取值范围：0~1.Dice系数数值越大，代表图像分割效果越好：

$$D_{ice}=\frac{2(X\cap Y)}{X\cup Y}, \tag{4}$$

其中，X代表真实值的集合；Y代表预测值的集合.

Dice Loss函数用于判断两个样本之间的相似程度，Dice Loss函数值D_{Loss}越小，则证明真实样本和预测样本之间相似度越高：

$$D_{Loss}=1-D_{ice}. \tag{5}$$

3　实验结果

本实验所用设备的CPU型号为Intel(R) Core(TM) i5-9300H，显卡为NVIDIA GeForce GTX 1660 Ti，深度学习的框架为Pytorch.在PSPnet中，训练文件格式为.voc.语义分割模型训练中，把数据集分为两个部分，第一部分是原图，如图3所示，为了方便训练，统一将图像格式调整为.jpg，对图像的尺寸大小无特殊要求，在图像输入端进行归一化操作.

图3　原始图像

第二部分是添加标签后的图像，如图4所示，标签采用8位彩色图.根据实际场景中的人或者物体用不同的颜色进行标注操作，原图所对应的数据维度为[3 024，4 032，3]，标签的数据维度为[3 024，4 032].标签中每个像素点的数值与目标类型相对应.本模型设定8种目标类型，对应标签中每个像素点的数值是0~7，0代表背景，1~7分别代表人、自行车、汽车、垃圾桶、摩托车、广告牌和消防栓.

图4　添加标签后的图像

图5~8为不同场景下的实验结果,圆点为模拟注意力机制的目标关注位置.

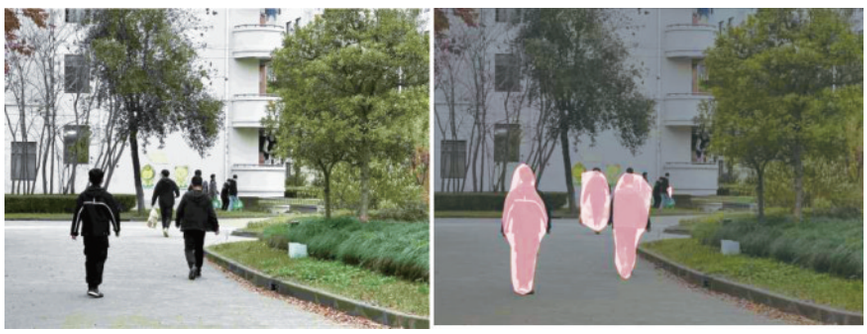

图5 情景1下的测试结果

图6 情景2下的测试结果

图7 情景3下的测试结果

图8 情景4下的测试结果

4　结　论

应用一种基于MobileNetV2的融合多特征PSPnet实现图像分割,网络中的线性瓶颈和反向残差结构可以消除激活函数Relu对特征信息的破坏,交叉熵损失函数和Dice Loss函数可以提高模型的准确率.引入注意力机制,将各部分得到的特征进行融合,实现了复杂场景下的图像语义分割.实验获得了预期效果,可以对图像中的目标对象进行有效的语义分割,正确识别对象类别.

参考文献:

[1] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE, 2017:6230-6239.

[2] LIU Y, WANG Y D, ZHANG C X, et al. Classification of benign and malignant pulmonary nodules by deep learning with anatomybased attention mechanism [J]. Chinese Journal of Medical Physics, 2022,39(11):1441-1447.

[3] LIU Y M, MA X, MEN C G. A hyerspectral remote sensing image classification method based on multi-spatial information [J]. Chinese Spaces Science and Technology, 2019,39(2):73-81.

[4] LI Y W, XU J J, LIU D X, et al. Field road scene recognition in hilly regions based on improved dilated convolutional networks [J]. Transactions of the Chinese Society of Agricultural Engineering, 2019,35(7):150-159.

[5] LIU S W, ZHANG Y Y, CAI T B, et al. An improved PSPnet model for semantic segmentation of UAV farmland images [J]. Journal of Irrigation and Drainage, 2022,41(4):101-108.

[6] JI S P, WEI S Q. Builiding extraction via convolutional neural networks from an open remote sensing building dataset [J]. Acta Geodaetica et Cartographica Sinica, 2019,48(4):448-459.

[7] SUN G L, WANG W G, DAI J F, et al. Mining cross-image semantics for weakly supervised semantic segmentation [C]// European Conference on Computer Vision. Glasgow: Springer, 2020:347-365.

[8] LIU Y F, ZHANG Q Z, WANG G H, et al. Building extraction in remote sensing imagery based on deep residuanl network [J]. Remote Sensing Information, 2020,35(2):59-64.

[9] YANG J H. Building extraction from high: resolution imagery based on FPN [J].Remote Sensing Information, 2021,36(4):133-141.

[10] WU H, ZHANG X C, SUN Y, et al. Building extraction in complex scenes based on the fusion of multi-feature improved PSPNet model [J]. Bulletin of Surveying and Mapping, 2021(6):21-27.

(责任编辑:包震宇,顾浩然)