第50卷第8期 中国科学技术大学学报 Vol. 50, No. 8
2020年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2020
文章编号:0253-2778(2020)08-1170-11

MAEA-DeepLab:具有多特征注意力有效聚合模块的语义分割网络

赵 柳[1],陆 军[1,2],刘 杨[1]
(1.黑龙江大学计算机科学与技术学院,黑龙江哈尔滨 150080;
2.黑龙江省数据库与并行计算重点实验室(黑龙江大学),黑龙江哈尔滨 150080)

摘要:为了实现网络的低训练成本,在保持高精度的同时大大降低计算复杂性,提出了带有多特征注意力有效聚合模块(MAEA)的语义分割网络:MAEA-DeepLab.该编码器主网络采用了下采样16步幅的低分辨率特征映射,获得高级特征.解码器通过MAEA模块充分利用特征的空间注意力机制,有效聚合多特征,获得具有强大语义表示的高分辨率特征,有效地提高了解码器恢复重要细节信息的能力,实现了高精度分割.MAEA-DeepLab的Multiply-Adds只有DeepLabV3+架构的30.9%,即943.02 B,大大降低计算复杂性.架构不经过COCO数据集预训练,仅使用两张RTX 2080 ti GPU,在PASCAL VOC 2012数据集和CityScapes数据集的测试集上进行了语义分割基准测试,mIOU分数分别达到了87.5%和79.9%.实验结果表明,MAEA-DeepLab以低计算开销达到了很好的语义分割精度.

关键词:语义分割;编码器-解码器;MAEA-DeepLab;空间注意力
中图分类号:TP317.4 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.08.018
引用格式:赵柳,陆军,刘杨. MAEA-DeepLab:具有多特征注意力有效聚合模块的语义分割网络[J]. 中国科学技术大学学报,2020,50(8):1170-1180.
ZHAO Liu, LU Jun, LIU Yang. MAEA-DeepLab: A semantic segmentation network with multi-feature attention effective aggregation module[J]. Journal of University of Science and Technology of China, 2020,50(8):1170-1180.

MAEA-DeepLab: A semantic segmentation network with multi-feature attention effective aggregation module

ZHAO Liu[1], LU Jun[1,2], LIU Yang[1]
(1. *College of Computer Science and Technology, Heilongjiang University, Harbin* 150080, *China*;
2. *Key Laboratory of Database and Parallel Computing of Heilongjiang Province (Heilongjiang University), Harbin* 150080, *China*)

Abstract: To realize the low cost of network training, the computational complexity is greatly reduced while maintaining high precision. A semantic segmentation network with multi-feature attention effective aggregation module(MAEA) is proposed: MAEA-DeepLab. A 16 stride low-resolution feature map for down-sampling is adopted in the encoder's network backbone, and high-level features are obtained. The decoder makes full use of the feature's spatial attention mechanism through the MAEA module, effectively aggregates multiple features, and obtains high-resolution features with strong semantic representation. Then the ability of the decoder to recover important details is effectively improved, and high-precision segmentation is achieved. Multiply-adds in MAEA-DeepLab is 943.02B, only 30.9% of the DeepLabV3+ architecture, which greatly reduces the computational complexity. The architecture is not pre-training on the COCO dataset. It performs semantic semantic segmentation Benchmark tests on the test set of with PASCAL VOC 2012 dataset and CityScapes dataset with only two RTX 2080ti GPUs, and the mIOU scores reach 87.5% and 79.9%, respectively. The experimental results show that good semantic segmentation accuracy is achieved with low computational cost in MAEA-DeepLab.

Key words: semantic segmentation; encoder-decoder; MAEA-DeepLab; spatial attention

0 引言

图像语义分割[1-3]实现高精度的根本在于高分辨率和高级语义,但是二者在卷积神经网络中的设计是矛盾的:高级语义特征靠近输出端但分辨率低;高分辨率特征靠近输入端但语义级别低.高级

收稿日期:2020-07-11;**修回日期**:2020-08-04
作者简介:赵柳,男,1995年生,硕士生.研究方向:深度学习 计算机视觉.E-mail:2181411@s.hlju.edu.cn
通讯作者:陆军,博士/教授.E-mail:lujun111_lily@sina.com

特征(细小物体、边界)的分割结果保持了大的语义结构,但细小结构丢失严重;低级特征(图像的颜色、形状、纹理等)的分割结果保留了丰富的细节,但语义类别预测很差. 最新的编码器-解码器[4-8]架构主要从两个方面优化网络架构:

(Ⅰ)改进主网络结构,有效提取特征,减少不必要的信息丢失,尽可能降低主网络计算量和参数量;

(Ⅱ)改进解码器中上采样等优化方法,从而更高效地恢复高级特征的细节信息和空间信息.

在特征提取过程中,为了捕获特征的重要信息,LEDNet[9]采用特征的空间注意力机制设计了级联结构的解码器,实现了轻量级的编码器-解码器架构. 除了特征的空间注意力机制,SENet[10]中提出了一个 squeeze-and-excitation(SE)模块,充分利用了特征的通道注意力机制,用于学习不同通道上的特征映射权重,凸显重要的特征映射而忽略不重要特征映射. ECA-Net[11]则改进了 SE 模块,用一维卷积代替了 SE 模块中的全连接层,同时舍弃了 SE 模块中通道维数降维后再恢复的操作.

经典的 DeepLabV3+[12]中,编码器有效提取了高级特征,但是解码器过于简单. 解码器将主网络中的低级特征和编码器输出的高级特征直接融合. 这种简单融合低级特征的解码器并不能充分恢复一些重要细节信息,所以本文的工作重点是研究一种更有效的解码器. 由于卷积运算是跨通道的,直接将空间信息混合在一起提取特征信息. 本文的 MAEA 模块通过空间注意力机制对输入特征对应通道上的空间信息进行注意力提取,让网络学习到特征中"哪些信息重要"以及"这些信息在哪",然后通过逐点相加进行聚合. MAEA 通过学习要强调或抑制的信息,有效提升了网络内的信息流,大大提高了解码器性能. 本文的贡献如下:

(Ⅰ)提出了一种有效的端到端可训练的 MAEA 模块;

(Ⅱ)构建低计算复杂性的网络模型 MAEA-DeepLab;

(Ⅲ)MAEA-DeepLab 在 PASCAL VOC 2012 和 Cityscapes 语义基准上实现了高分割精度.

1　相关工作

1.1　早期架构

全卷积神经网络(FCN)[13]在 PASCAL VOC 2011 上取得了令人瞩目的成就. FCN 是以反卷积[14]上采样的形式替代了 CNN 最后的 Softmax 层,实现语义级图像分割. 为了保持特征的空间信息,单纯地将 CNN 的全连接层用卷积层替换,不足以细化粗糙的语义级标签预测. CRF-RNN[15]考虑了图像边缘等细节信息的问题,使用了条件随机场(CRF)做后处理,从而产生更加清晰的边界,实现了细粒度分割. FCN 和编码器-解码器结构如图 1 所示.

随着 U-Net[16]、SegNet[17]等语义分割架构的提出,编码器-解码器结构成为主流,最具代表性之一的便是 DeepLabV3+.

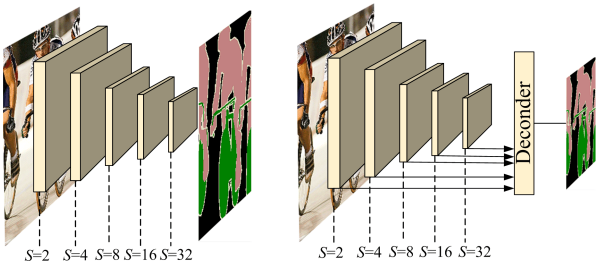

图 1　语义分割的两种主流架构

Fig. 1　Two mainstream architectures for semantic segmentation

1.2　解码器

Google 提出了经典 DeepLab 系列架构. 早期 DeepLab V1[18]、DeepLab V2[19]和 DeepLab V3[20]采用级联架构. 随后文献[21]提出空间金字塔池化(spatial pyramid pooling, SPP),并在 PSPNet[22]中取得了显著的性能提升. 文献[21]结合了经典的空洞卷积(atrous convolution)[23],在空间金字塔池化的基础上提出了空洞空间金字塔池化(ASPP),实现获取大感受野以及多尺度上下文特征映射. 空洞卷积的结构如图 2(a)所示,输入特征为 F_{in},$K=3$ 表示滤波器尺寸,通过 $r=2$ 的空洞率实现 $S=5$ 的感受野,卷积后得到输出特征 F_{out}.

空洞卷积通过设置不同的空洞率(atrous rate) r,在采样点之间填充 $r-1$ 个零,实现在不增加滤波器的参数量前提下,将感受野从 $K\times K$ 扩展到 $S\times S$,从而获得更大的感受野,S 的计算公式为

$$S=K+(K-1)(r-1) \tag{1}$$

对于输入特征 F_{in} 上的每个位置 i,滤波器为 W,经过空洞率为 r 的空洞卷积后,得到输出 F_{out}

$$F_{out}(i)=\sum_{K}F_{in}(i+r*K)W(K) \tag{2}$$

ASPP 通过 3 个空洞率得到不同尺寸的 3×3 空洞卷积,提取大感受野的多尺度特征. DeepLabV3+舍弃了 DeepLabV2 中使用 CRF 作后处理的操作,改用编码器-解码器架构,如图 2 所示.

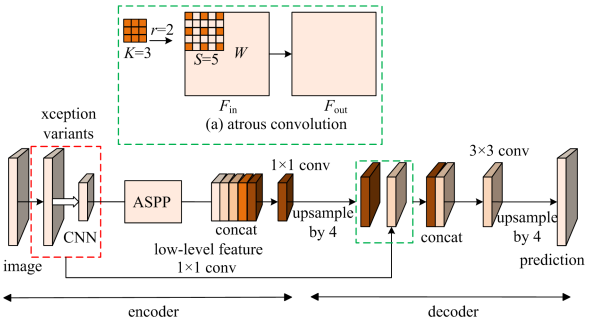

图 2　语 DeepLab V3+架构

Fig. 2　DeepLab V3+ architecture

DeepLab V3+在训练和评估过程中,为了保留更多的细节信息,牺牲了内存和速度. 主网络采用了下采样 8 步幅的大分辨率特征映射,所以 DeepLab V3+的计算开销非常大.

1.3　特征融合

DeepLab V3+的解码器在利用低级特征恢复高级特征丢失的细节信息时,只是将主网络中下采

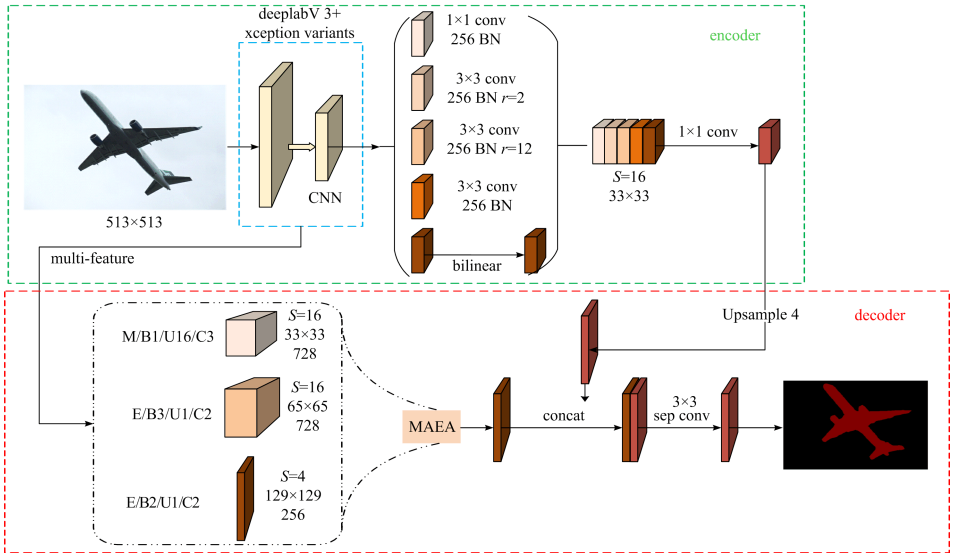

图 3 MAEA-DeepLab 架构

Fig. 3 The MAEA-DeepLab architecture

样 4 步幅的低级特征和编码器提取的高级特征直接融合，本文认为这是不充分的. 如何有效融合低级特征，逐步恢复高级特征的细节信息和空间信息，现有研究主要有以下两个方面：

（Ⅰ）FastFCN[24] 提出 JPU（joint pyramid upsampling）模块，即先将多特征双线性上采样到相同分辨率大小，再进行类似于 ASPP 的多尺度特征提取，最后直接进行 Concat 融合；

（Ⅱ）DUpsampling[25] 提出了一种新的 DUpsample 上采样方法，即先将多特征下采样到相同分辨率大小，再直接进行 Concat 融合，最后通过 DUpsample 上采样.

2 方法

2.1 MAEA-DeepLab

本文提出了多特征注意力有效聚合模块（MAEA），在 MAEA 模块的基础上进一步提出了 MAEA-DeepLab 架构，结构如图 3 所示. 图 3 中，$M(E)/Bx/Uy/Cz$ 表示中间流（输入流）的第 x 个块的第 y 个单元的第 z 个深度可分离卷积. MAEA-DeepLab 的编码器主网络采用了下采样 16 步幅的 Xception[26] 变体. 首先对输入图像进行尺寸裁剪后，经过 Xception 变体提取特征，再经过 ASPP 模块提取 5 种多尺度上下文特征，最后通过 1×1 卷积融合多尺度特征，输出最终提取到的高级特征. 解码器中，首先通过 MAEA 模块对主网络中 3 个不同阶段提取的低级特征进行空间注意力有效聚合，产生足够的语义级注意力，与高级特征融合时提供足够的细节信息. 再通过两个 3×3 深度可分离卷积逐步恢复特征空间信息，捕获清晰的目标边界，细化分割结果，实现密集图像预测.

2.2 编码器

MAEA-DeepLab 的编码器主网络为下采样 16 步幅的 Xception 变体，结构如图 4 所示.

Image 表示输入图像，分辨率为 513×513；Conv 为标准卷积；Spear Conv 为深度可分离卷积；

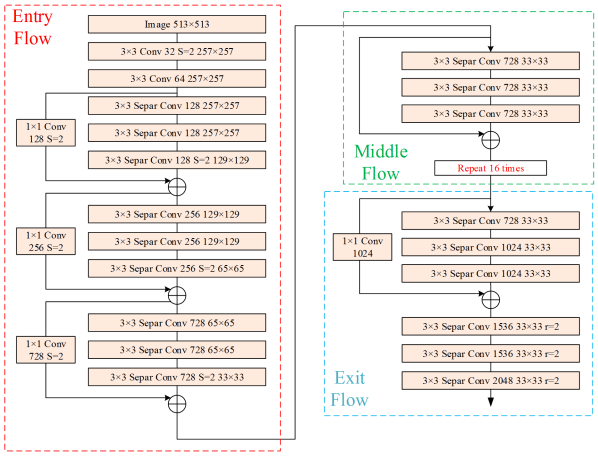

图 4 MAEA-DeepLab 的主网络结构

Fig. 4 MAEA-DeepLab's network backbone structure

S 表示下采样步幅. 输入图像经过主网络提取特征后，再经过 ASPP 提取多尺度上下文特征. ASPP 提取的 5 个多尺度特征直接在通道维上 Concat，再通过 1×1 卷积进行通道之间的信息交流以及空间信息的融合，并将通道维数降维到 256，输出最终提取到的高级特征.

2.3 解码器

不同于早期级联结构的 FCN 和 SegNet，二者将提取的高级特征直接上样采到初始输入图像分辨率大小，进行密集图像预测，并不能很好地恢复高级特征所丢失细节信息，所以精心设计的解码器就是要完成恢复细节信息的任务. MAEA-DeepLab 不仅保持详细的分辨率信息，而且还获得具有强大语义表示的特征.

2.3.1 MAEA

MAEA-DeepLab 解码器最重要的部分就是多特征注意力有效聚合模块. MAEA 通过空间注意力机制对多特征进行处理产生语义级注意，使模型注意特征图中感兴趣的特征空间区域，由低分辨率向

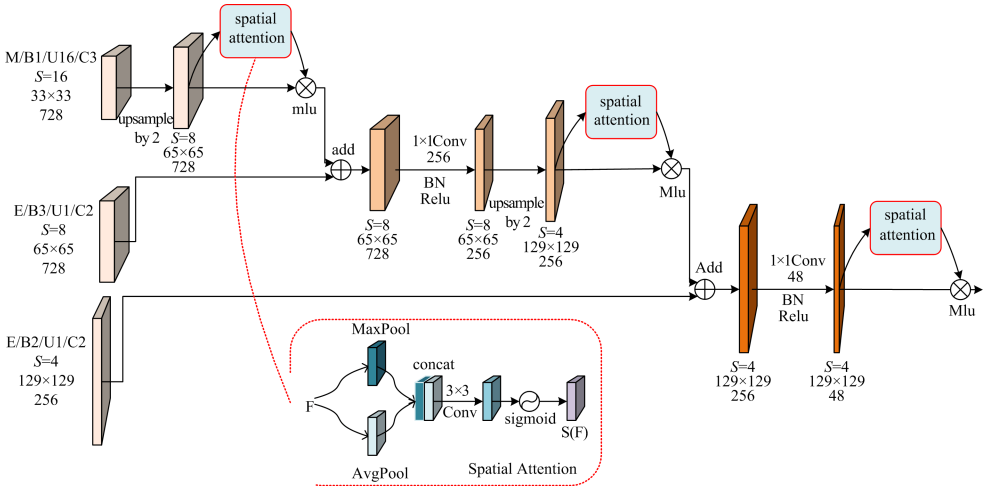

图 5　多特征注意力有效聚合模块

Fig. 5　Multi-feature attention efficient aggregation module

高分辨率学习相邻的多特征之间的语义流,逐级聚合多特征的语义级注意,凸显重要的特征域,获得具有强大语义表示的高分辨率特征,实现高精度分割. MAEA 结构如图 5 所示,S 表示下采样步幅;Add 表示特征逐元素相加操作;Mul 表示特征逐元素相乘加权操作.

在空间注意力模块中,为了计算特征的空间注意力,首先沿通道维经由平均池化和最大池化操作对特征映射的通道信息进行评估,生成通道上的平均池化特征 F_{avg} 和最大池化特征 $F_{\max}$,并将二者堆叠以生成一个有效的特征描述符;再通过 3×3 卷积融合特征并改变通道数,通过 Sigmoid 激活函数对特征的注意力权重进行归一化,生成空间注意图 $S(F)$,对要强调或抑制的位置进行编码. 沿着通道维应用池操作已被证明是突出信息区域的有效方法. S(F)的计算公式为

$$S(F)=\text{Sigmoid}(\text{Conv3}\times 3([F_{avg}(F);F_{\max}(F)])) \tag{3}$$

式中,F 为输入特征.

MAEA 模块从低分辨率特征向高分辨率特征逐次聚合两种特征. 首先,特征 M/B1/U16/C3 经过 2 步幅上采样,再通过空间注意力模块产生注意力权重,并加权到输入特征上,以进行自适应特征细化. 之后与特征 E/B3/U1/C2 进行 Add 聚合,再经过 1×1 卷积、批归一化、Relu 激活函数处理并改变通道维数. 得到的特征经过上述相同的操作处理后与特征 E/B2/U1/C2 进行聚合,再经过空间注意力模块处理后输出最终提取到的特征. MAEA 对多个低级特征的有效聚合,使得特征对应通道上的重要特征空间区域凸显. 这样 MAEA-DeepLab 的解码器实质上解耦了多个低级特征融合过程中的空间关系和通道关系,充分利用了特征空间注意力机制,有效地聚合了更多的重要细节信息,在最终的密集图像预测上实现了很好的分割精度.

2.3.2　Concat 和 Element-wise Add

现有的架构,如 ResNet[27],FPN[28] 等采用的是逐元素相加(Element-wise Add)来融合特征,而 DenseNet[29] 等采用的是 Concat 融合特征. Concat 方法进行多个不同分辨率融合时,只需将低分辨率特征直接上采样到高分辨率,然后在通道维上堆叠即可进行后续的卷积操作. 逐元素相加方法进行特征融合不但要保持多个特征的分辨率一致,还得保持通道数相同,Add 是在对应通道上进行特征图的逐元素相加,通道数保持不变,二者区别如图 6 所示.

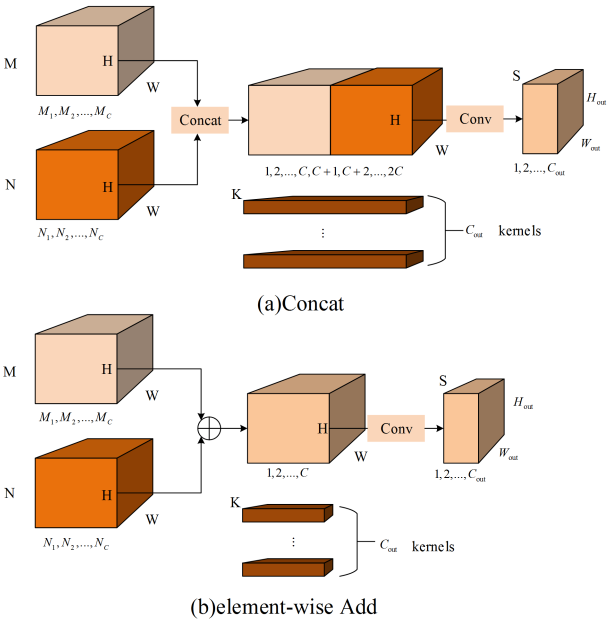

图 6　Concat 和 Element-wise Add 两种特征融合方法

Fig. 6　Two feature fusion methods, Concat and Element-wise Add

有两个输入特征 M、N,特征尺寸均为 $H\times W$,通道数均为 C,输入通道分别为 $M_1,M_2,\cdots,M_C$ 和 $N_1,N_2,\cdots,N_C$. 本文比较二者在单个输出通道上的区别以及参数量和计算量上的不同. 由于每个输出通道的卷积核是独立的,本文只研究单个通道的输出. Concat 的单个输出通道 S_{Concat} 的计算公式为

$$S_{\text{Concat}}=\sum_{i=1}^{C}\text{Conv}(M_i,K_i)+\sum_{i=1}^{C}\text{Conv}(N_i,K_{C+i}) \tag{4}$$

式中,Conv()为后续的卷积操作. Add 的单个输出通道 S_{Add} 的计算公式为

$$S_{\text{Add}}=\sum_{i=1}^{C}\text{Conv}((M_i+N_i),K_i)=\sum_{i=1}^{C}\text{Conv}(M_i,K_i)+\sum_{i=1}^{C}\text{Conv}(N_i,K_i) \tag{5}$$

本文可以理解 Concat 是通过增加通道数来保存增加的信息量. Add 等价于 Concat 之后对应通道共享同一个卷积核,Add 没有改变特征的通道维数,而是特征在对应通道的空间关系上增加信息量,这显然更有利于进行特征的空间注意力聚合;每次聚合后通过带有批归一化和 Relu 激活函数的 1×1 卷积进行特征的通道之间信息交流和空间信息的融合,并且凸显重要的空间信息.

Concat 和 Add 的参数量和计算量的区别主要集中在 Conv()操作,在考虑偏置的前提下,Concat 的参数量 P_{Concat} 和计算量 F_{Concat} 分别为

$$P_{\text{Concat}}=2C\times((K\times K)+1)\times C_{\text{out}} \tag{6}$$

$$F_{\text{Concat}}=2C\times K\times K\times H_{\text{out}}\times W_{\text{out}}\times C_{\text{out}} \tag{7}$$

Add 的参数量 P_{Add} 和计算量 F_{Add} 分别为

$$P_{\text{Add}}=C\times((K\times K)+1)\times C_{\text{out}} \tag{8}$$

$$F_{\text{Add}}=C\times K\times K\times H_{\text{out}}\times W_{\text{out}}\times C_{\text{out}} \tag{9}$$

式中,$P_{\text{Concat}}=2P_{\text{Add}}$,$F_{\text{Concat}}=2F_{\text{Add}}$,即 Concat 的参数量和计算量是 Add 的两倍. 分析可得出,使用逐元素相加方法进行多特征注意力有效聚合时,可以根据特征对应通道上的空间信息,利用空间注意力机制增加单通道上包含的空间信息,这有助于产生语义级注意,凸显重要的特征映射而忽视不重要的特征映射,并且 Add 的计算量和参数量只有 Concat 的一半,所以本文的 MAEA 模块采用了逐元素相加方法.

3 实验

本文在 PASCAL VOC 2012 数据集[30]和 CityScapes 数据集[31]的语义分割基准上评估了 MAEA-DeepLab. 对于这两个基准,本文根据当前类别的语义平均交并比(mIOU)衡量性能.

PASCAL VOC 2012 总共包含 20 个前景对象类和 1 个背景类. 数据集分成了 3 部分:训练集 1464 张图像,验证集 1449 张图像,测试集 1456 张图像. 本文还用文献[32]中提供的额外注释将原始数据集进行扩充,得到 10582 张图像的增强训练集(train_aug).

CityScapes 是语义城市场景解析的大规模基准. 它包含训练集 2975 张用图像,验证集 500 张图像,测试集 1525 张图像. 此外,它还提供约 19998 张弱注释图像(train_extra).

3.1 实施细节

MAEA-DeepLab 是在 Tensorflow[33]上实现的,超参数 train output stride 和 eval output stride 均为 16,decoder output stride 为 4. 学习率衰减采用“Poly”学习率策略,计算公式为

$$\text{LR}=\text{base\_learning\_rate}(\frac{1-\text{iter}}{\text{max\_iter}})^{\text{power}} \tag{10}$$

式中,power 参数设置为 0.9,初始学习率 base_learning 和总迭代次数 max_iter 根据具体实验进行设置. MAEA-DeepLab 训练采用 Nesterov 动量优化器,momentum=0.9,权重衰减设置为 4e-5,采用 pixel-wise cross-entropy 作为损失函数. 实验在两张 RTX 2080 Ti (GPU11G per GPU)上进行. 对于数据增强,采用随机缩放(从 0.5 到 2.0),然后左右翻转输入图像. 本文按照 DeepLab V3+中使用 ImageNet 数据集训练主网络的方法和训练细节对 MAEA-DeepLab 的主网络进行预训练,用所得到的预训练权重初始化主网络的参数.

3.1.1 PASCAL VOC 2012 实施细节

首先在 PASCAL VOC 2012 的增强训练集上进行预训练,batch size 应大于 12,最佳为 16,这样批归一化参数才能得到充分的训练. 预训练时,MAEA-DeepLab 的 batch size 设置为 16,受 GPU 显存限制,输入图像裁剪尺寸设置为 385×385,这对网络权重参数的训练会有轻微的误差,但足以训练网络的批归一化参数. 在增强训练集上预训练时衰减率设置为 0.9997,初始学习率为 0.01,训练总迭代次数为 300 k. 然后用所得到的预训练权重初始化网络的参数,冻结批归一化参数,在 PASCAL VOC 2012 数据集的训练验证集上进行微调(fine tuning),batch size 设置为 8,输入图像裁剪尺寸为 513×513,初始学习率为 1e-4,总迭代次数为 90 k.

3.1.2 CityScapes 实施细节

首先在 Cityscapes 的弱注释图像上进行预训练,batch size 设置为 16. 受 GPU 显存限制,输入图像裁剪尺寸设置为 385×385,预训练时衰减率设置为 0.9997,初始学习率为 0.005,训练总迭代次数为 250 k. 然后用所得到的预训练权重初始化网络的参数,冻结批归一化参数,在 CityScapes 的训练集上进行微调,batch size 设置为 4,输入图像裁剪尺寸为 769×769,初始学习率为 1e-3,总迭代次数为 90 k.

3.2 特征选择和聚合策略

MAEA 模块聚合主网络中不同的低级特征空间注意力,决定 MAEA 能否有效为高级特征提供足够的细节信息. 在主网络中,不同块中所提取的低级特征是有差异的,所以本文在 PASCAL VOC 2012 数据集上总共进行了 6 组对比实验,根据实验结果确定 MAEA 模块使用哪些低级特征进行聚合最佳,结果如表 1 所示.

对比实验采用 3.1.1 节的实施细节,输入图像尺寸裁剪为 513×513,特征的最高分辨率为 129×129,通道数分别为 128 和 256. 在 PASCAL VOC 2012 数据集的训练集上进行微调,在验证集上进行评估. 表 1 的实验结果表明,无论采取 Concat 还是 MAEA,靠近输出的特征组合分割精度均高于靠近输入的特征组合. 当采取 Concat 方法时,靠近输出的 3 组特征组合分割精度均高于靠近输入的 3 组特征组合,分割精度分别提升 0.46%、0.50%、

0.54%;当采取 MAEA 模块时,分割精度分别提升 0.50%、1.01%、1.00%.因为在主网络中,不同阶段提取到的分辨率特征包含的语义信息有明显差异.越靠近输出的特征包含的语义信息越抽象,也越有利于通过特征的空间注意力机制产生语义级注意,使得多特征的聚合更有效,充分恢复高级特征丢失的细节信息,因此,本文选取靠近输出的特征组合.

表 1　PASCAL 2012 验证集上的 mIOU%

Tab.1　The mIOU% on PASCAL 2012 Val set

used low-level features	Concat	MAEA
E/B2/U1/C2; M/B1/U16/C3	80.92	82.43
E/B2/U1/C2; E/B3/U1/C2; M/B1/U16/C3	81.01	83.28
E/B2/U1/C2; E/B3/U1/C2; M/B1/U8/C3; M/B1/U16/C3	81.12	83.35
E/B1/U1/C3; M/B1/U1/C1	80.38	81.93
E/B1/U1/C3; E/B2/U1/C3; E/B3/U1/C3	80.55	82.27
E/B1/U1/C3; E/B2/U1/C3 E/B3/U1/C3; M/B1/U8/C1	80.62	82.35

具体分析靠近输出的 3 组特征组合的实验结果.当采取 Concat 方法时,3 组特征分割精度并无明显差距,mIOU 分数相差在 0.2%之内,说明将多个低级特征直接在通道维上进行 Concat 过于简单,并不能充分聚合这些低级特征,导致高级特征恢复细节信息不充分.当采取 MAEA 模块时,E/B2/U1/C2;E/B3/U1/C2;M/B1/U16/C3 这 3 个特征组合的分割精度高于 E/B2/U1/C2;M/B1/U16/C3 这两个特征组合,mIOU 分数提升了 0.85%,表明 MAEA 模块充分聚合了这些低级特征,产生了丰富的语义级注意,凸显了重要的特征空间信息,为高级特征提供了足够的细节信息;但是 E/B2/U1/C2,E/B3/U1/C2,M/B1/U8/C3,M/B1/U16/C3 这 4 个特征组合的分割精度几乎与 3 个特征组合持平,mIOU 分数的提高(0.07%)可忽略不计.因为 M/B1/U8/C3,M/B1/U16/C3 这两个特征是在主网络中同一块的同一阶段提取的,二者包含的语义信息相似,所以增加同一阶段的特征数量只会带来不必要的参数量和计算量,并不能聚合更多语义信息.通过对比实验,确定了 MAEA 模块采用 E/B2/U1/C2,E/B3/U1/C2,M/B1/U16/C3 这 3 个特征组合.

本文还在 CityScapes 数据集上对该特征组合分别进行了 Concat 和 MAEA 的对比实验,结果如表 2 所示.表 2 的实验结果再次验证 MAEA 模块的有效性,分割精度明显优于简单的 Concat 方法.

3.3　评估策略

在评估 MAEA-DeepLab 时,为了探究最佳评估策略,研究是否添加输入图像左右翻转和不同的多尺度输入对 MAEA-DeepLab 分割精度的影响,本节进行了多组对比实验.在 PASCAL VOC 2012 数据集的训练集上进行微调,在验证集上进行评估,结果如表 3 所示.表 3 中,1 号实验为原始尺度评估,有无添加输入图像左右翻转 mIOU 分数均为 83.28%.这表明原始尺度评估时,添加输入图像左右翻转不能提升分割精度;而使用多尺度输入时,添加输入图像左右翻转均提升了分割精度.其中 2 号实验 mIOU 分数提高了 0.78%.多尺度输入采用了 0.2、0.25、0.3 三种不同的跨度.当添加输入图像左右翻转时,0.3 跨度的多尺度输入分割精度表现最差,其次是 0.2 跨度,0.25 跨度的多尺度输入表现最好.在 0.3 跨度的多尺度输入中,13 号实验表现最好,比 1 号实验 mIOU 分数提高了 0.86%.在 0.2 跨度的多尺度输入中,4 号实验表现最好,比 1 号实验 mIOU 分数提高了 1.11%.在 0.25 跨度的多尺度输入中,7、8、9 号实验表现最好,比 1 号实验 mIOU 分数提高了 1.32%.

表 2　Concat 和 MAEA 的 mIOU%

Tab.2　The mIOU% of Concat and MAEA

used low-level features	Concat	MAEA
E/B2/U1/C2 E/B3/U1/C2; M/B1/U16/C3	76.33	78.64

表 3　评估策略在 Pascal VOC 2012 上的 mIOU%

Tab.3　mIOU% of evaluation policy on Pascal VOC 2012

Number	Multi-Scale Inputs	Not Flip	Flip
1	[1.0]	83.28	83.28
2	[0.4,0.6,0.8,1.0,1.2,1.4,1.6]	82.91	83.69
3	[0.8,1.0,1.2,1.6]	83.81	84.35
4	[0.8,1.0,1.6]	83.95	84.39
5	[0.5,0.75,1.0,1.25,1.5,1.75]	84.27	84.56
6	[0.5,1.0,1.25,1.5,1.75]	84.19	84.56
7	[0.75,1.0,1.25,1.5,1.75]	84.32	84.60
8	[0.75,1.0,1.25,1.75]	84.32	84.60
9	[0.75,1.0,1.5,1.75]	84.33	84.60
10	[0.75,1.0,1.75]	84.21	84.51
11	[0.5,1.0,1.75]	83.89	84.27
12	[0.4,0.7,1.0,1.3,1.6,1.9]	83.15	83.78
13	[0.7,1.0,1.6,1.9]	83.78	84.14
14	[0.7,1.0,1.9]	83.14	83.77

本文具体分析表现最好的 0.25 跨度实验.在输入图像缩小尺度中,含 0.75 尺度的多尺度输入方式比含 0.5 尺度的多尺度输入方式分割精度高.在输入图像放大尺度中,含 1.5 尺度的多尺度输入方式和含 1.25 尺度的多尺度输入方式分割精度几乎一致,但是 1.5 尺度具有更大的计算量开销.综上所述,本文考虑精度和计算开销的平衡,选择 8 号实验评估策略.

本文还在 CityScapes 数据集上对该评估策略进行了对比实验,进一步验证该评估策略的有效性.由于 CityScapes 为大分辨率图像数据集,输出图像尺寸大,评估时受 GPU 显存限制,最大尺度不超过 1.9.含 1.9 尺度时,最多只能进行三个尺度的多尺度输入;含 1.75 尺度时,最多只能进行四个尺度的多尺度输入.结果如表 4 所示.表 4 的 CityScapes 实验结果再次表明 0.25 跨度的多尺度输入表现最好.在 0.25 跨度中,[0.75,1.25,1.75]

和[0.75,1.0,1.25,1.75]的多尺度输入方式分割精度几乎一致,所以采用计算量开销更小的[0.75,1.25,1.75]多尺度输入方式. MAEA-DeepLab 分别采用各自的最优评估策略,在 PASCAL VOC 2012 数据集的测试集上达到了 87.5% mIOU 的分割精度,在 CityScapes 数据集的测试集上达到了 79.9% mIOU 的分割精度.

表 4 评估策略在 CityScapes 上的 mIOU%

Tab. 4 mIOU% of evaluation policy on CityScapes

Multi-Scale Inputs	Not Flip	Flip
[1.0]	78.64	78.64
[0.8,1.0,1.6]	78.95	79.53
[0.75,1.25,1.75]	79.89	80.16
[0.75,1.0,1.25,1.75]	79.89	80.16
[0.7,1.0,1.9]	79.66	79.86

3.4 实验结果

3.4.1 计算量

Multiply-Adds 表示整个网络中所有特征权重 W 的相关乘法运算以及偏置 B 的相关加法运算之和,即计算量. 若输入特征的分辨率为 $H\times W$;C_{in} 表示输入通道数;卷积核的分辨率为 $K_H\times K_W$;C_{out} 表示输出通道数. 网络主要的计算量如下:标准卷积 M_{Conv} 和深度可分离卷积 M_{Sep} 的计算公式为

$$M_{\text{Conv}}=\{(K_H\times K_W\times C_{\text{in}})\times C_{\text{out}}+C_{\text{out}}\}\times(H\times W) \tag{11}$$

$$\begin{aligned}M_{\text{Sep}}&=\{(K_H\times K_W\times 1)\times C_{\text{in}}+C_{\text{in}}\}\times(H\times W)+\\&\{(1\times 1\times C_{\text{in}})\times C_{\text{out}}+C_{\text{out}}\}\times(H\times W)=\\&\{C_{\text{in}}(K_H\times K_W+C_{\text{out}}+1)+C_{\text{out}}\}\times(H\times W)\end{aligned} \tag{12}$$

全局平均池化 M_{avg} 和最大池化 M_{max} 的计算公式为

$$M_{\text{avg}}=C_{\text{int}}\times H\times W \tag{13}$$

$$M_{\text{max}}=K_H\times K_W\times C_{\text{out}}\times H\times W \tag{14}$$

以输入图像裁剪尺寸 $H\times W$ 为例,本文推理了 MAEA-DeepLab 和 DeepLabV3+主要的计算量差别. 不同于 DeepLabV3+的 8 步幅下采样提取$(H/8)\times(W/8)$的分辨率特征,MAEA-DeepLab 在主干网络的 E/B3/U1/C3 层进行了一次 2 步幅下采样,提取了$(H/16)\times(W/16)$的小分辨率特征. 二者在 E/B3/U1/C3 层及该层前的网络计算量 M_{before} 基本相同;E/B3/U1/C3 层后的网络计算量 M_{after},MAEA-DeepLab 在理论上降低了接近四分之三. 经过理论分析,二者主要的计算量 M_{MAEA} 和 $M_{\text{V3+}}$ 的表达式为

$$\begin{aligned}M_{\text{V3+}}&=M_{\text{before}}+M_{\text{after}}=\\&(\sum_{i_1=1}^{m_1}M_{\text{Conv}}+\sum_{i_2=1}^{m_2}M_{\text{Sep}})+\\&(\sum_{i_3=1}^{n_1}M_{\text{Conv}}+\sum_{i_4=1}^{n_2}M_{\text{Sep}}+\sum_{i_5=1}^{p_1}M_{\text{avg}})\end{aligned} \tag{15}$$

$$\begin{aligned}M_{\text{MAEA}}&=M_{\text{before}}+M_{\text{after}}=\\&(\sum_{i_1=1}^{m_1}M_{\text{Conv}}+\sum_{i_2=1}^{m_2}M_{\text{Sep}})+\\&(\sum_{i'_3=1}^{n'_1}M_{\text{Conv}}+\sum_{i'_4=1}^{n'_2}M_{\text{Sep}}+\sum_{i'_5=1}^{p'_1}M_{\text{avg}}+\sum_{i_6=1}^{p'_2}M_{\text{max}})\end{aligned} \tag{16}$$

M_{MAEA} 与 $M_{\text{V3+}}$ 的比值为

$$\begin{aligned}\frac{M_{\text{MAEA}}}{M_{\text{V3+}}}&=\\&\frac{54187(H\times W)+5179100(H\times W)}{54187(H\times W)+18315175(H\times W)}\times 100\%=\\&\frac{5233287(H\times W)}{18369362(H\times W)}\times 100\%\approx 28.5\%\end{aligned} \tag{17}$$

经过理论分析,MAEA-DeepLab 的主要计算量只有 DeepLabV3+主要计算量的 28.5%. 为了计算网络整体精准的计算量,只需通过 Tensorflow 实现的算法 3.1 计算对应模型的. pb 文件,根据计算图中连接结构的“索引”,寻找节点对应的具体运算,最终获得整个网络的 Multiply-Adds.

算法 3.1 计算 Multiply-Adds 的算法

输入:模型文件 frozen_inference_graph. pb;

输出:模型计算量 Multiply_Adds.

```
Multiply_Adds=0;/*初始化*/
With gfile.FastGFile(模型文件路径,'rb') as f:;/*实现对图片的读取*/
graph_def = tf.GraphDef();/*Tensorflow 数据结构*/
graph = tf.get_default_graph();/*获取当前默认计算图*/
graph_def.ParseFromString(f.read());/*加载模型*/
for op in graph.get_operations();/*for 获取所有卷积操作 op*/
if (op.type = ="Conv2D" or op.type = ="DepthwiseConv2dNative");/*if 判断卷积类型*/
Multiply_Adds = Multiply_Adds + op.outputs[0].shape[1] * op.outputs[0].shape[2] * prod(op.inputs[1].shape);
Print(Multiply_Adds);/*输出计算量*/
```

通过算法 1 的计算可得,MAEA-DeepLab 和 DeepLab V3+精准的计算量 Multiply-Adds 的值分别为 943.02 B 和 3055.35 B,二者比值为

$$\frac{M'_{\text{MAEA}}}{M'_{\text{V3+}}}=\frac{943.02\text{B}}{3055.35\text{B}}\times 100\%\approx 30.9\% \tag{18}$$

实验结果与本文的理论推理基本相符,存在的差距主要由于理论只推算了网络主要计算量,而过于细小烦琐的计算量如 Relu 和 Sigmoid 激活函数、批归一化和 Dropout 全连接层等没有计算在内. 总之,理论与实验均表明 MAEA-DeepLab 的计算量有大幅降低.

3.4.2 结果分析

本文的研究目的是在低训练成本的前提下,提出低计算复杂性的高分割精度语义分割架构. 本文在 PASCAL VOC 2012 的训练集上分别对 DeepLab V3+和 MAEA-DeepLab 进行训练时间的

对比. 使用两张 GPU,输入图像裁剪尺寸均采用 513×513,train_batch_size 均设置为 2,总迭代次数均为 90 k,学习率等其他超参数保持一致,结果如表 5 所示. 实验结果表明,MAEA-DeepLab 的训练速度比 DeepLab V3+快了一倍.

表 5　训练时间的对比

Tab. 5　Comparison of training time

Method	training time/h
DeepLab V3+	10.8
MAEA-DeepLab	5.4

本文在 PASCAL VOC 2012 和 CityScapes 的测试集上进行最优性能的详细对比. MAEA-DeepLab 不经过 COCO 数据集预训练,使用本文最优策略进行训练评估;DeepLab V3+使用原论文中最优策略进行训练评估,结果如表 6 所示.

表 6　在测试集上的对比

Tab. 6　Comparison on the test set

method	DeepLab V3+	MAEA-DeepLab
Output stride	8	16
COCO	√	
PASCAL VOC 2012	87.8%mIOU	87.5%mIOU
CityScapes	82.1%mIOU	79.9%mIOU

实验结果表明,MAEA-DeepLab 只用了 DeepLab V3+的 30.9%的计算量就达到几乎相同的语义分割精度,即以低计算复杂性实现了高精度语义分割. MAEA-DeepLab 在 PASCAL VOC 2012 和 CityScapes 的测试集上分别实现了 87.5%和 79.9%的 mIOU 分数.

不同于 PASCAL VOC 2012 的小分辨率图像数据集,CityScapes 为 1024×2048 的大分辨率图像数据集. 受 GPU 显存的限制,在输入图像裁剪时,如果采用 DeepLab V3+中 769×769 的大尺寸裁剪,每张 GPU 只能设置 batch size 为 2. 在 CityScapes 数据集的 train_extra 上预训练网络权重和批归一化参数时,batch size 设置为 16 最佳,所以本文将输入图像裁剪为 385×385,这是造成精度损失的根本原因,与 MAEA-DeepLab 性能无关. 本文的目的是在低训练成本和低计算复杂性的条件下,实现高精度语义分割架构. MAEA-DeepLab 只用了 DeepLab V3+的 30.9%的 Multiply-Adds,在测试集上的 mIOU 分数只是略低于具有高计算复杂性的 DeepLabV3+,但是依然优于 LEDNet 等先进的低计算复杂性网络.

MAEA-DeepLab 和 DeepLabV3+在 PASCAL VOC 2012 数据集的验证集上可视化结果如图 7 所示.

在部分类别的语义分割结果中,MAEA-DeepLab 明显优于 DeepLab V3+. 图 7 实验结果中,第 3 行的栏杆,DeepLab V3+错误地判别为马的语义类别;第 4 行坠落在湖中的汽车,DeepLab V3+错误地将汽车的语义类别判别为船;第 5 行的马,

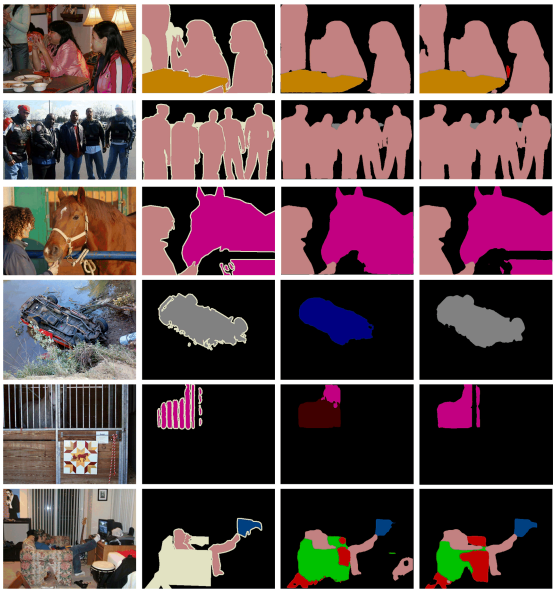

图 7　在 PASCAL VOC 2012 的验证集上可视化结果

Fig. 7　Visualization results on val set of PASCAL VOC 2012

DeepLab V3+只正确地分割出马的头的语义类别,马的身体的语义类别错误地判别为猫的语义类别. 在 DeepLab V3+没有正确分割的模式上,MAEA-DeepLab 可以正确分割出语义类别. 这表明本文所提出的 MAEA-DeepLab 是先进的. 最后一行为复杂的室内场景语义类别,MAEA-DeepLab 和 DeepLab V3+均存在失败之处,但是总体上前者表现比后者好.

由于 MAEA-DeepLab 未使用 COCO 数据集进行预训练,所以本文对多个未使用 COCO 预训练的已有架构进行了对比,结果如表 7 所示,MAEA-DeepLab 在 PASCAL VOC 2012 数据集上的性能表现远远优于现有的架构.

表 7　未进行 COCO 预训练的模型在测试集上的表现

Tab. 7　Performance of models without COCO pre-training on test sets

method	mIOU(%)
ResNet-38[34]	82.5
PSPNet[22]	82.6
DFN[35]	82.7
EncNet[36]	82.9
DUpsampling[25]	85.3
MAEA-DeepLab	87.5

本文分别在 PASCAL VOC 2012 和 CityScapes 的验证集上对 MAEA-DeepLab 进行了大量的可视化,结果如图 8 和图 9 所示. MAEA-DeepLab 整体上的分割结果是比较完美的. 当然,MAEA-DeepLab 也存在一些不足之处,比如图 8 中最后一组图像的分割结果显示,MAEA-DeepLab 将

靠近椅子的桌子那部分的语义类别错误地分类为椅子的语义类别,图 9 中 wall 类别的分割表现略显不足,有待提高.

为了进一步验证本文提出的 MAEA-DeepLab 先进性,分别在 PASCAL VOC 2012 数据集和 CityScapes 数据集上与多个已有的架构进行了比较,结果如表 8 和表 9 所示. 由表 8 和表 9 可以看出,MAEA-DeepLab 在 CityScapes 的 19 个类别和 PASCAL VOC 2012 的 20 个类别的平均 IOU 分数是最高的,分别达到 87.5%和 79.9%. 在详细类别中,一部分类别的 IOU 分数虽略低于其他架构,另一部分类别的 IOU 分数 MAEA-DeepLab 是最高的,尤其是 Wall 类别比 TKCN 提高了 6.9%,Chair 类别比 CFNet 提高了 8.5%.

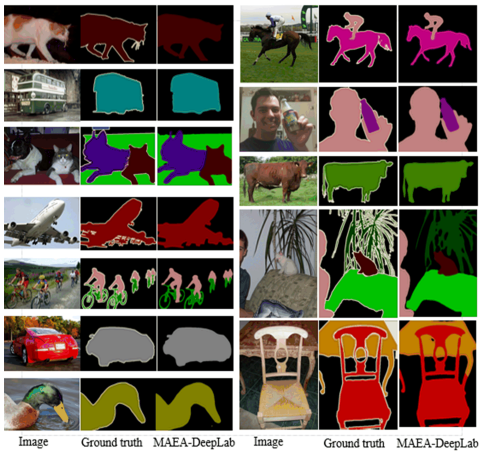

图 8 在 PASCAL VOC 2012 验证集上的可视化

Fig. 8 Visualization on the val set of the PASCAL VOC 2012

表 8 Cityscapes 的测试集上的 19 个类别的分割结果 IOU(%)和平均 IOU(%)Classes

Tab. 8 Segmentation results IOU (%) of 19 categories on CityScapes test set and mean IOU (%) Classes

method	SegNet[17]	ICNet[37]	LEDNet[9]	RefineNet[38]	FC-HarDNet[39]	TKCN[40]	MAEA-DeepLab
Classes	57.0	69.5	70.6	73.6	75.9	79.5	79.9
Road	96.4	97.1	98.1	98.2	98.5	98.4	98.6
Sidewalk	73.2	79.2	79.5	83.3	85.5	85.8	86.5
Building	84.0	89.7	91.6	91.3	92.5	93.0	93.1
Wall	28.4	43.2	47.7	47.8	49.0	51.7	58.6
Fence	29.0	48.9	49.9	50.4	54.4	61.7	62.8
Pole	35.7	61.5	62.8	56.1	64.0	67.6	66.7
T-Light	39.8	60.4	61.3	66.9	71.5	75.8	72.6
T-Sign	45.1	63.4	72.8	71.3	75.6	80.0	77.9
Vegetation	87.0	91.5	92.6	92.3	93.0	93.6	93.5
Terrain	63.8	68.3	61.2	70.3	70.6	72.7	72.1
Sky	91.8	93.5	94.9	94.8	95.4	95.4	95.1
Person	62.8	74.6	76.2	80.9	84.5	86.9	85.7
Rider	42.8	56.1	53.7	63.3	67.4	70.9	68.8
Car	89.3	92.6	90.9	94.5	95.7	95.9	95.8
Truck	38.1	51.3	64.4	64.6	67.7	64.5	73.5
Bus	43.1	72.7	64.0	76.1	79.0	86.9	88.1
Train	44.1	51.3	52.7	64.3	63.6	81.8	84.0
Motorcycle	35.8	53.6	44.4	62.2	60.7	69.6	69.2
Bicycle	51.9	70.5	71.6	70.0	72.7	77.6	75.0

表 9 Pascal VOC 2012 的测试集上的 20 个类别的分割结果 IOU(%)和平均 IOU(%)Mean

Tab. 9 Segmentation results IOU (%) of 20 categories on Pascal VOC 2012 test set and mean IOU (%) Mean

method	DeepLab V2[19]	PSPNet[22]	Auto-D-L[41]	DeepLab V3[20]	EncNet[36]	CFNet[42]	MAEA-DeepLab
Mean	79.7	85.4	85.6	85.7	85.9	87.2	87.5
Aeroplane	92.6	95.8	96.5	96.4	95.3	96.7	97.1
Bicycle	60.4	72.7	77.3	76.6	76.9	79.7	80.1
Bird	91.6	95.0	94.8	92.7	94.2	94.3	96.8
Boat	63.4	78.9	74.1	77.8	80.2	78.4	77.0
Bottle	76.3	84.4	84.0	87.6	85.3	83.0	87.4
Bus	95.0	94.7	97.1	96.7	96.5	97.7	97.2
Car	88.4	92.0	88.7	90.2	90.8	91.6	93.2
Cat	92.6	95.7	94.5	95.4	96.3	96.7	95.5
Chair	32.7	43.1	53.5	47.5	47.9	50.1	58.6

续表 9

method	DeepLab V2[19]	PSPNet[22]	Auto-D-L[41]	DeepLab V3[20]	EncNet[36]	CFNet[42]	MAEA-DeepLab
Cow	88.5	91.0	91.6	93.4	93.9	95.3	94.6
Diningtable	67.6	80.3	79.2	76.3	80.0	79.6	77.2
Dog	89.6	91.3	88.4	91.4	92.4	93.6	92.1
Horse	92.1	96.3	94.2	97.2	96.6	97.2	96.9
Motorbike	87.0	92.3	90.2	91.0	90.5	94.2	93.5
Person	87.4	90.1	91.2	92.1	91.5	91.7	92.8
Pottedplant	63.3	71.5	75.1	71.3	70.9	78.4	72.8
Sheep	88.3	94.4	90.1	90.9	93.6	95.4	92.8
Sofa	60.0	66.9	70.7	68.9	66.5	69.6	72.8
Train	86.8	88.8	89.1	90.8	87.7	90.0	90.4
Tvmonitor	74.5	82.0	79.7	79.3	80.8	81.4	81.8

4　结论

经过大量的研究分析和实验验证,本文提出的多特征注意力有效聚合模块(MAEA)具有重要意义,可以有效恢复高级特征丢失的细节信息,进而实现一种高精度、低计算开销的语义分割架构 MAEA-DeepLab. MAEA-DeepLab 仅使用两张 GPU,不经过 COCO 数据集预训练,Multiply-Adds 为 943.02 B,分别在 PASCAL VOC 2012 和 CityScapes 数据集的测试集上实现了 87.5% 和 79.9%的 mIOU 分数. 同时在多个复杂的语义类别上表现优于 DeepLab V3+架构,Multiply-Adds 为 DeepLab V3+的 30.9%,这表明 MAEA-DeepLab 是先进的. MAEA 模块有效利用特征空间注意力机制,逐步聚合 3 组不同阶段提取的低级特征,产生语义级注意,凸显重要的特征,从而更有效地为高级特征恢复空间信息时提供足够的细节信息. MAEA-DeepLab 的编码器中未使用特征的注意力机制,在未来的工作中,考虑使用特征的通道注意力机制以及更高效的轻量级主网络去改进 MAEA-DeepLab,实现更优秀的架构.

参考文献(References)

[1] MOTTAGHI R, CHEN X, LIU X, et al. The role of context for object detection and semantic segmentation in the wild[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014:891-898.

[2] CAESAR H, UIJLINGS J, FERRARI V. Coco-stuff: Thing and stuff classes in context[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1209-1218.

[3] ZHOU B, ZHAO H, PUIG X, et al. Scene parsing through ADE20K dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017:5122-5130.

[4] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation [J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[5] PENG C, ZHANG X, YU G, et al. Large kernel matters-improve semantic segmentation by global convolutional network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017.

[6] FU J, LIU J, WANG Y, et al. Stacked deconvolutional network for semantic segmentation [J]. IEEE Transactions on Image Processing, 2017:99.

[7] YU C, WANG J, PENG C, et al. Learning a discriminative feature network for semantic segmentation [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[8] ZHANG Z, ZHANG X, PENG C, et al. Exfuse: Enhancing feature fusion for semantic segmentation [J]. European Conference on Computer Vision, 2018.

[9] WANG Y, ZHOU Q, LIU J, et al. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation [C]// 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019.

[10] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018.

[11] WANG Q, WU B, ZHU P, et al. Eca-net: Efficient channel attention for deep convolutional neural networks [C]// 2020 IEEE/CVF Conference on Computer Vision & Pattern Recognition. IEEE, 2020.

[12] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European Conference on Computer Vision [C]// Computer Vision - ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11211. Springer, Cham, 2018.

[13] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.

[14] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation [C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2016.

[15] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, et al. Conditional random fields as recurrent neural networks [C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015.

[16] RONNEBERGER O, FISCHER P, BROX T. U-net:

Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham,2015.

[17] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[18] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2016-06-07)[2020-06-11]. https://arxiv. org/abs/1412.7062v4.

[19] 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.

[20] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation [EB/OL]. (2017-12-05) [2020-06-11]. https://arxiv. org/pdf/1706.05587.

[21] LAZEBNI S, SCHMID C, PONCE J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2006.

[22] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [J]. IEEE Computer Society, 2016.

[23] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions [C]// Proceedings of the International Conference on Learning Representations. IEEE, 2015.

[24] WU H, ZHANG J, HUANG K, et al. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation [EB/OL]. (2019-03-28)[2020-06-11]. https://arxiv. org/pdf/1903.11816. pdf.

[25] TIAN Z, HE T, SHEN C, et al. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[26] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[27] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016:770-778.

[28] LIN T Y, DOLL AR P, GIRSHICK R, et al. Feature pyramid networks for object detection [J]. IEEE Computer Society, 2017.

[29] HUANG G, LIU Z, MAATEN L, et al. Densely connected convolutional networks [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017:2261-2269.

[30] EVERINGHAM M, ESLAMI S A, GOOL L V, et al. The pascal visual object classes challenge: A retrospective [J]. Springer, 2015, 111(1): 98-136.

[31] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[32] HARIHARAN B, ARBEL AEZ P, BOURDEV L, et al. Semantic contours from inverse detectors[C]// IEEE International Conference on Computer Vision. IEEE, 2011.

[33] ABADI M, AGARWAL A, BARHAM P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems [EB/OL]. (2016-03-16) [2020-06-11]. https://arxiv. org/pdf/1603.04467.

[34] WU Z, SHEN C, HENGEL A. Wider or deeper: Revisiting the resnet model for visual recognition [J]. Elsevier, 2019: 119-133.

[35] YU C, WANG J, PENG C, et al. Learning a discriminative feature network for semantic segmentation [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[36] ZHANG H, DANA K, SHI J, et al. Context encoding for semantic segmentation [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[37] ZHAO H S, QI X J, SHEN X Y, et al. ICNET for real-time semantic segmentation on high-resolution images[C]//Computer Vision - ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11207. Springer, Cham, 2018.

[38] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[39] CHAO P, KAO C Y, RUAN Y S, et al. Hardnet: A low memory traffic network [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2020.

[40] WU T, TANG S, ZHANG R, et al. Tree-structured kronecker convolutional network for semantic segmentation [C]// 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019.

[41] LIU C, CHEN L C, SCHROFF F, et al. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 82-92.

[42] VALMADRE J, BERTINETTO L, HENRIQUES J, et al. End-to-end representation learning for correlation filter based tracking[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 5000-5008.