计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(9):178−183 [doi: 10.15888/j.cnki.csa.007541]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 Deeplab V3+网络的语义分割①

席一帆, 孙乐乐, 何立明, 吕　悦

(长安大学 信息工程学院, 西安 710064)
通讯作者: 席一帆, E-mail: 15735169750@163.com

摘　要: 深度学习的语义分割在计算机视觉领域中有非常广阔的发展前景, 但许多分割效果较好网络模型占用内存大和处理单张图片耗时长. 针对这个问题, 把 Deeplab V3+模型的骨干网 (ResNet101) 的瓶颈单元设计为 1D 非瓶颈单元, 且对空洞空间金字塔池化模块 (Atrous Spatial Pyramid Pooling, ASPP) 的卷积层进行分解. 该算法能大幅度降低 Deeplab V3+网络的参数量, 提高网络推理速度. 基于 PASCAL VOC 2012 数据集进行对比实验, 实验结果显示改进网络模型拥有更快的处理速度和更优的分割效果, 且消耗更少的内存.

关键词: 语义分割; Deeplab V3+模型; 骨干网 (ResNet101); 1D 非瓶颈单元; 空洞空间金字塔池化 (ASPP)

引用格式: 席一帆,孙乐乐,何立明,吕悦.基于改进 Deeplab V3+网络的语义分割.计算机系统应用,2020,29(9):178–183. http://www.c-s-a.org.cn/1003-3254/7541.html

Semantic Segmentation Based on Improved Deeplab V3+ Network

XI Yi-Fan, SUN Le-Le, HE Li-Ming, LYU Yue

(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: Semantic segmentation of deep learning has a very broad development prospect in the field of computer vision, but many network models with better segmentation effects take up a lot of memory and take a long time to process a single picture. In response to this problem, we replace the bottleneck unit of the Deeplab V3+ model backbone network (ResNet101) with a 1D non-bottleneck unit, and decompose the convolutional layer of the Atrous Spatial Pyramid Pooling (ASPP) module. The algorithm can greatly reduce the parameter amount of Deeplab V3+ network and accelerate the speed of network inference. Based on the PASCAL VOC 2012 dataset, the experimental results show that the improved network model has faster speed and better segmentation, and takes up less memory space.

Key words: semantic segmentation; Deeplab V3+ model; backbone network (ResNet101); 1D non-bottleneck unit; Atrous Spatial Pyramid Pooling (ASPP)

1　引言

图像分割是计算机视觉领域的重要分支, 在无人驾驶, 医学图像, 3D 重建等场景用广泛的应用. 传统的图像分割算法利用图像的颜色、纹理、形状等低级语义信息进行分割, 缺失像素的对比度, 方向度等中级语义[1–4]; 聚类是利用像素中级语义进行分割, 但缺少像素之间的实体类别之间的高级语义; 深度学习算法学习图像中的高级语义. 深度学习的语义分割是对图像像素进行逐个分类, 解析图像的深层次语义信息. Shelhamer 等[5] 提出 FCN (全卷积网络) 可以对任意大小的图片进行处理, 同时还引入跳级连接使低级语义信息和高级语义信息的融合, 反卷积上采样恢复图像分辨率, 但在细节上分割效果不好; Ronneberger 等[6] 提出一种基于编码解码器架构的 U-net, 编码器对图像进行深层次的特征提取, 生成高级语义信息, 解码器利用跳级连接的思想, 对不同分辨率特征图进行通道融合产生较好的分割效果.

① 收稿时间: 2019-12-12; 修改时间: 2020-02-08; 采用时间: 2020-03-11; csa 在线出版时间: 2020-09-04

Vigay Badrinaryanan 等[7,8] 提出 SegNet, 该网络架构与 U-net 类似, 不同的是 SegNet 上采样利用编码器池化操作的下标去恢复图像分辨率, 加速网络的推理, 且占用更少的内存. Zhao 等[9] 提出 PSPnet 利用空间金子塔模块以不同的感受野提取全局特征, 融合上下文信息进行上采样得到预测结果. Lin 等[10] 提出 Refinet, 充分利用下采样的特征图, 利用长范围残差链接的思想, 将粗糙的高层语义特征和细粒度的底层特征进行融合, 通过 Renfinet block 将特征图进行逐层融合生成分割图像. 谷歌提出一系列 Deeplab 模型[11–14], 其中 Deeplab V3+的分割效果最优, 但该模型在处理速度和模型容量上并不占优势, 本文依据 Deeplab V3+模型提出一种优化算法, 对骨干网残差单元重新设计, 对 ASPP 模块进行优化, 且在公开数据集进行对比实验, 改进后的模型在准确度和精度提高的情况下, 进一步提高网络的处理速度, 优化该模型的内存消耗.

2 方法与网络

2.1 Deeplab V3+网络概述

Deeplab V3+网络模型主要基于编码解码器结构, 如图 1 所示. 该模型的编码器架构由骨干网 ResNet101 和 ASPP 模块组成, 骨干网提取图像特征生成高级语义特征图, ASPP 模块利用骨干网得到的高级语义特征图进行多尺度采样, 生成多尺度的特征图, 在编码器尾部将多尺度的高级语义特征图在通道维度上进行组合, 通过 1×1 的卷积进行通道降维. 解码器部分将骨干网的低级语义特征通过 1×1 卷积进行通道降维, 保持与高级语义特征图串联在一起时的比重, 增强网络学习能力. 再用 3×3 的卷积提取特征, 编码器尾部进行上采样, 产生最终的语义分割图.

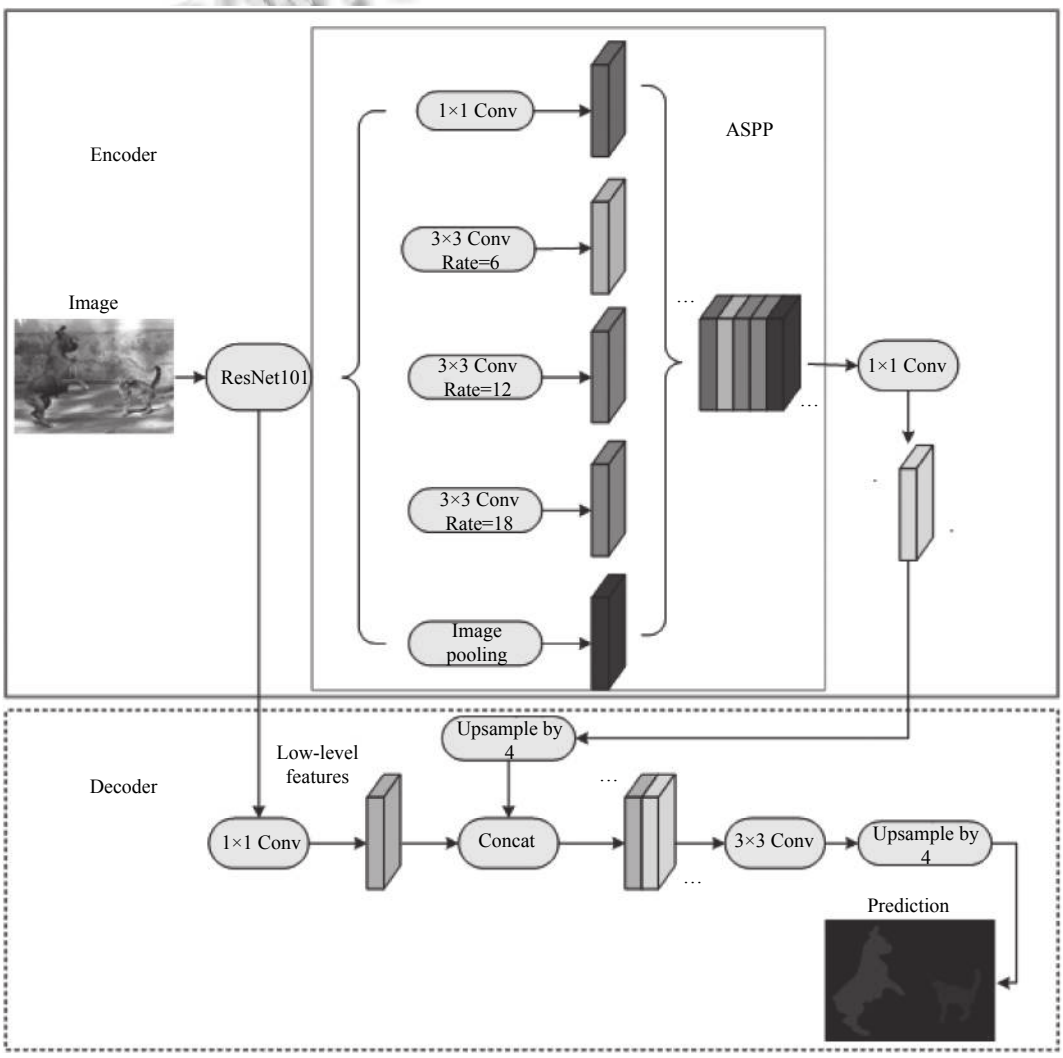
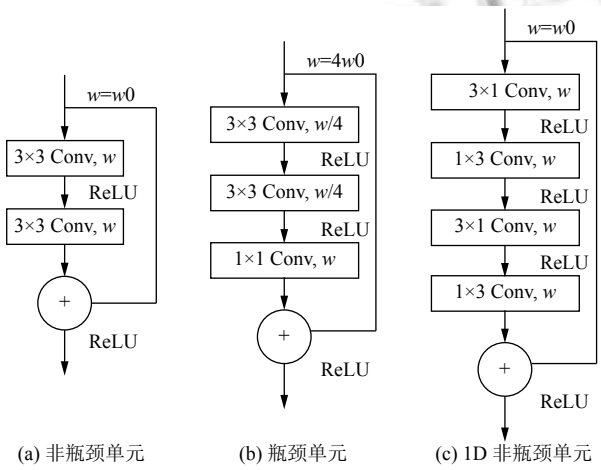


图 1 Deeplab V3+模型图

2.2 骨干网的改进

骨干网 ResNet101 利用基于瓶颈设计的残差块作为基本单元, 组成 101 层的残差网络. 如图 2(b) 骨干网由通道数为 w_0 的残差块组成, w_0 的组合为 (64,128, 256,512), 这 4 类瓶颈残差单元的数目分别为 (3,4,23, 3), 加上网络前端的 7×7 的卷积和最后 1×1 卷积层共 101 层. 瓶颈单元拥有更少的参数, 可以训练更深层次的网络, 而非瓶颈单元 (如图 2(a)) 随着深度增加, 可以获得更高的准确率, 结合瓶颈单元和非瓶颈单元的优点, 重新设计残差单元. 文献 [15,16] 已证明二维卷积能被分解成一系列一维卷积的组合. 依据文献 [17] 在卷积层松弛秩为 1 约束的条件下, 卷积层 f^i 可以重新写成:



(a) 非瓶颈单元 (b) 瓶颈单元 (c) 1D 非瓶颈单元

图 2 骨干网

$$f^i = \sum_{k=1}^{K} \sigma_k^i v_k^i \left(\bar{h}_k^i\right)^{\mathrm{T}} \tag{1}$$

其中, v_i^k和$\left(\bar{h}_k^i\right)^{\mathrm{T}}$是长度为 d 的向量, σ_k^i是权重标量, 且 K 为 f^i 的秩. 基于该表达式 Alvarez and Petersson 等[14] 提出每一个卷积层都可以分解为 1D 卷积和一个非线性$\varphi(\cdot)$的操作, 分解层在输入为 a_*^0 的条件下, 第 i 个输出 a_i^l 表达式为:

$$a_i^l = \varphi\left(b_i^h + \sum_{l=1}^{L} \bar{h}_{il}^{\mathrm{T}} * \left[\varphi\left(b_l^v + \sum_{c=1}^{C} \bar{v}_{lc} * a_c^0\right)\right]\right) \tag{2}$$

式中, L 为卷积层的数目, $\phi(\cdot)$为 ReLU. 将骨干网的瓶颈单元替换为 1D 非瓶颈单元 (如图 2(c)). 在 3×3 卷积输入特征图通道数相同的条件下, 1D 非瓶颈单元能减少 33% 非瓶颈单元的参数和 29% 的瓶颈单元参数. (假如 c 为 3×3 卷积输出通道数, 则 3×3 常规卷积参数量为 w_0×3×3×c, 2D 分解后的参数量为 w_0×3×1×c+w_0×1×3×c, 分解后能减少约 33% 权重参数; 1D 非瓶颈单元总参数量 $12w_0^2$, 瓶颈单元的总参数量 $17w_0^2$, 非瓶颈单元的总参数为 $18w_0^2$) 分解 2D 卷积后, 增加 ReLU 非线性操作, 能增强 1D 非瓶颈单元的学习能力. 因此 1D 非瓶颈单元拥有非瓶颈单元的准确率高的和瓶颈单元参数少, 易训练深层网络的优点.

2.3 ASPP 模块的改进

ASPP 模块主要是对骨干网的特征图进行多尺度语义信息提取. 由于 ASPP 模块中 3×3 卷积会学到一些冗余信息, 参数数量多, 因此会在训练中耗费很长时间. 常规卷积已被证明会计算许多重叠的冗余信息. 依据骨干网改进的方法, 将 ASPP 中 3×3 的空洞卷积进行 2D 分解 (如图 3 所示), 将其分解成 3×1 和 1×3 的卷积, 保持其空洞率. 该改进的 ASPP 模块卷积参数量比常规卷积的参数量要少 33%, 在速度上比 3×3 卷积快, 能够提取到重要的语义信息, 有效的减少该模块计算量.

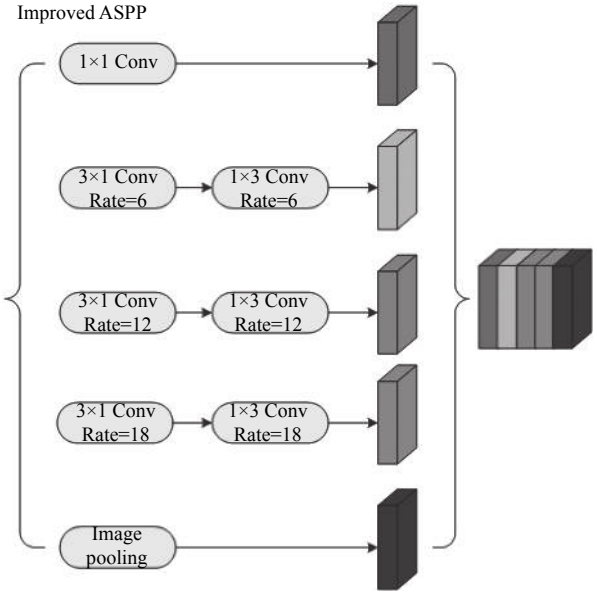


图 3 改进的 ASPP 模块

3 实验与分析

3.1 实验环境

实验运行环境 Win10 操作系统, 工作站 GPU 型号为: NVIDIA GeForce GTX 1070 (8 GB 显存), 基于 Tensorflow 深度学习框架, 本文利用 Deeplab V3+原文的 tensorflow 官方源码, 并对其进行改进, 进行对比实验.

3.2 实验训练与结果分析

实验用的是 PASCAL-VOC2012 增强版数据集, 训练集 10 582 张, 验证集 1449 张, 该数据集包括 20 个类别. 本实验将图片分辨率缩放至 513×513 像素, 由于真实标签和预测结果是灰度图, 为了显示分割效果采用 RGB 彩色图显示. 训练网络前, 将图像转化为 Tfrecord 文件, 便于高效读取数据.

本实验将基于的 1D 非瓶颈单元的骨干网在 Imagenet 数据集上进行预训练, 再将其预训练权重加载到改进的模型中. 利用上述数据集进行训练, 超参数设置如表 1 所示.

表 1 训练参数

参数	值
Base learning rate	0.007
End learning rate	0.000 001
Power	0.9
Batch size	5
Weight decay	0.0005
Max iteration	150 000
Batch normal decay	0.9997
Momentum	0.9

学习率采用多项式衰减, 当迭代次数超过 Max iteration 次, 学习率为 End learning rate. 采用动量梯度下降法去优化损失函数, 总共迭代 71 epochs, 如图 4 所示, 总共迭代 150 307 次, 每迭代一次大约耗时 7 s. 总损失 (总损失包括交叉熵损失、权重正则化损失) 在大约 12 万次左右开始收敛, 选取总损失最低的模型作为测试模型. 改进模型在训练集上的 *MIoU* 为 89.9%, 像素的平均准确率 97.3%.

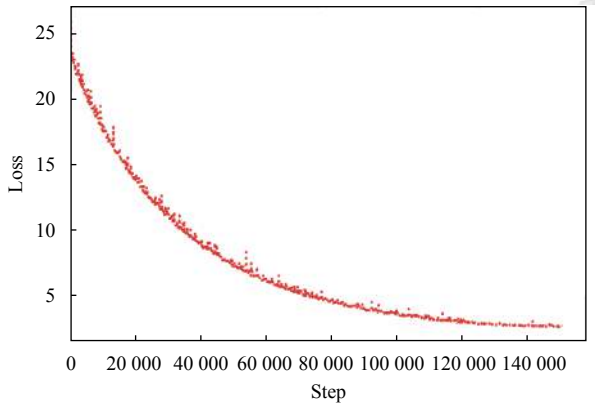
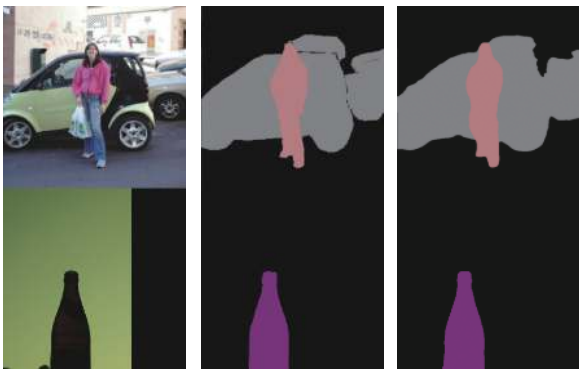


图 4 总损失函数图

图 5 所示, 改进后的模型在拥有多个类别对象的图像上, 有良好的分割结果, 尤其是在第一幅图将车与人两个类别的边界处分割效果较好.



(a) 原图 (b) 真实标签 (c) 预测

图 5 改进后模型在验证集分割结果

3.3 实验对比

语义分割有 4 种评价指标, 分别为像素精度 (*PA*), 均像素精度 (*MPA*), 均交并比 (*MIoU*), 频权交并比 (*FWIoU*). 假设有 *K*+1 个类, p_{ij} 表示被属于第 i 类但预测为第 j 类的像素数目, 即 p_{ii} 为真正的像素数量 (TP), p_{ij} 为假负的像素数量 (FN), p_{ji} 为假正像素数量 (FP).

PA: 为被分类正确的像素占总像素数目的比例:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{3}$$

MPA: 计算每个类被正确分类的像素比例, 再取平均:

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{4}$$

MIoU: 真实标签与预测标签的交集比上它们的并集, 计算每个类的 IoU, 再取平均:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{5}$$

FWIoU: 在 IoU 的基础上将每个类出现的频率作为权重:

$$FWIoU = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} * \sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{6}$$

为了便于对比, 实验将 *MPA*, *MIoU* 作为原模型与改进后的模型衡量标准.

表 2 说明, 改进后的模型在均像素精度上比原模型高 0.78%, 且在 *MIoU* 上比原模型高 0.63%, 因此改进模型拥有更准确和可靠的分割结果. 表 3 可以得出, 改进后的模型在设备上所占内存大小和单张图片处理速度上, 明显优于原模型, 其中在单张图片的运行时间上, 改进后的模型速度提高约 9.44%, 且模型容量减少了 19.6%. 主要由于对骨干网和 ASPP 模块的卷积层进行改进, 去掉冗余的权值, 参数量变少.

表 2 Deeplab V3+与 Modified Deeplab V3+的均像素精度和均交并比比较 (%)

模型	*MPA*	*MIoU*
Deeplab V3+	94.58	78.85
Modified Deeplab V3+	95.36	79.48

表 3 Deeplab V3+与 Modified Deeplab V3+在单张图片处理时间与模型大小的比较

模型	单张图片处理时间(ms)	模型大小 (MB)
Deeplab V3+	191.7	473.7
Modified Deeplab V3+	173.6	380.9

图 6 所示总损失函数, Deeplab V3+和 Modified Deeplab V3+模型的损失函数收敛速度几乎一样, 原模型 Total loss 最终收敛到 1.91, 而改进后模型 Total loss 收敛到 1.73, 且改进模型的损失函数摆动幅度小更稳定, 训练时间比原模型短 3.5 小时.

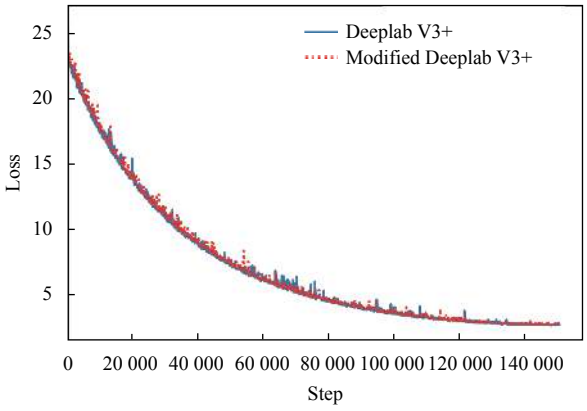


图 6 总损失函数

图 7 中圆圈标记出的图像区域, Modified Deeplab V3+的分割结果更精细. 例如: 第一幅图 Modified Deeplab V3+将椅子的空当分割出来, 而原模型未分割出, 且原模型将窗户误分类为显示屏; 第二幅图改进模型将飞机机翼准确分割出, 原模型未分割出机翼; 第三幅图改进模型能将车顶的人的跳跃姿态和车下的人准确分割, 原模型对车顶的人分割结果模糊, 且车下的人未被分割出; 第四幅图改进模型准确将椅子分割; 最后一幅图改进模型在马的腿部分割效果比原模型要完整. 明显可以看出改进模型分割效果更好, 且误分类少. 主要归因于 Modified Deeplab V3+的 1D 非瓶颈单元提高了图像分类的准确度, 且 ASPP 模块卷积分解后, 引入非线性操作, 增强网络学习能力, 有助于减少误分类, 同时在分解的卷积上再引入空洞卷积, 进一步扩大感受野, 提高网络在图像边缘分割的精细度.

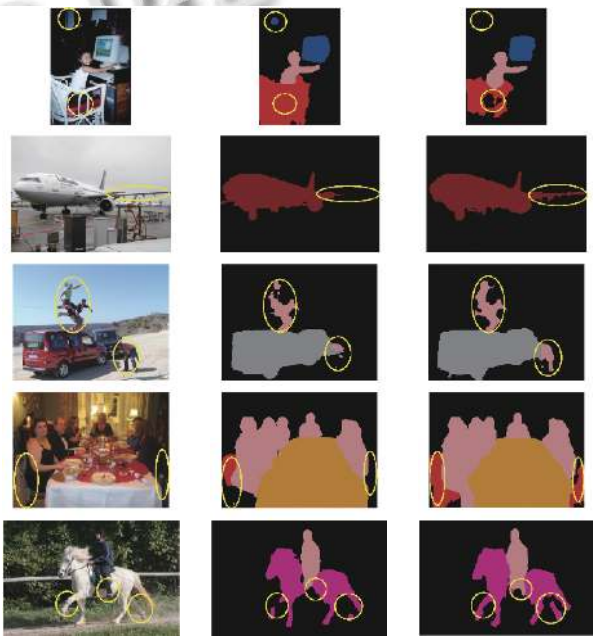


图 7 Deeplab V3+与 Modified DeeplabV3+测试集分割结果对比

4 结论

本文提出了一种基于卷积分解优化 Deeplab V3+网络的算法, 该算法主要利用 2D 卷积分解减少参数冗余, 提高处理速度, 同时引入非线性操作, 增强模型学习能力. 本文利用该算法重新设计 Deeplab V3+模型骨干网的残差单元, 使其既拥有非瓶颈单元的准确度, 又有瓶颈单元参数少, 易训练深层网络的优点; 同时又对 ASPP 模块也进行优化, 加速网络的推理速度, 减少其训练和处理时间. 实验结果证明 Modified Deeplab V3+与原模型相比在提高均像素精度的同时, 明显提升均

交并比, 且网络处理速度提高 9.44%, 优化网络模型的内存消耗. 测试集的结果表明, Modified Deeplab V3+在图像细节处分割结果更精确. 进一步的工作是探究如何控制感受野的大小, 提高模型对小目标分割的精确度.

参考文献

1 计梦予, 裘肖明, 于治楼. 基于深度学习的语义分割方法综述. 信息技术与信息化, 2017, (10): 137–140. [doi: 10.3969/j.issn.1672-9528.2017.10.037]

2 肖朝霞, 陈胜. 图像语义分割问题研究综述. 软件导刊, 2018, 17(8): 6–8, 12.

3 Arbeláez P, Hariharan B, Gu CH, *et al*. Semantic segmentation using regions and parts. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA. 2012. 3378–3385. [doi: 10.1109/CVPR.2012.6248077]

4 Lu ZW, Fu ZY, Xiang T, *et al*. Learning from weak and noisy labels for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(3): 486–500. [doi: 10.1109/TPAMI.2016.2552172]

5 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

6 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich, Germany. 2015. 234–241.

7 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495. [doi: 10.1109/TPAMI.2016.2644615]

8 de Oliveira Junior LA, Medeiros HR, Macêdo D, *et al*. SegNetRes-CRF: A deep convolutional encoder-decoder architecture for semantic image segmentation. Proceedings of 2018 International Joint Conference on Neural Networks. Rio de Janeiro, Brazil. 2018. 1–6.

9 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6230–6239. [doi: 10.1109/CVPR.2017.660]

10 Lin GS, Milan A, Shen CH, *et al*. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 5168–5177.

11 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA. 2014. 357–361.

12 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

13 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.

14 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, *et al*, eds. Computer Vision (ECCV 2018). Cham: Springer, 2018. 833–851.

15 Alvarez J, Petersson L. DecomposeMe: Simplifying ConvNets for end-to-end learning. arXiv: 1606.05426, 2016.

16 Na T, Mukhopadhyay S. Speeding up convolutional neural network training with dynamic precision scaling and flexible multiplier-accumulator. Proceedings of 2016 International Symposium on Low Power Electronics and Design. San Francisco, CA, USA. 2016. 58–63.

17 Sironi A, Tekin B, Rigamonti R, *et al*. Learning separable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 94–106. [doi: 10.1109/TPAMI.2014.2343229]