文章编号：2095-2163(2021)06-0181-04 中图分类号：TP391.4 文献标志码：A

# 基于 DeepLab V3+改进的图像语义分割模型

徐志凡，杜洪波，韩承霖，李恒岳，祁 新，林凯迪，黎 诗
（沈阳工业大学 理学院，沈阳 110870）

**摘 要**：针对 DeepLab V3+模型的解码器部分对于特征图的多尺度连接不够充分，会使最终的语义分割图的分割精细度较低的问题，本文在 DeepLab V3+模型的编码器部分增加输出一个中级语义特征图，在解码器部分对所得的特征图进行了 concat 融合，进而提高了分割精度。在公开数据集上进行验证，实验结果表明改进的 DeepLab V3+模型的平均交并比相比于原模型提高了 0.76%。

**关键词**：语义分割；DeepLab V3+模型；解码器；平均交并比

## Improved based on DeepLab V3+ network

XU Zhifan, DU Hongbo, HAN Chenglin, LI Hengyue, QI Xin, LIN Kaidi, LI Shi
(School of Science, Shenyang University of Technology, Shenyang 110870, China)

**【Abstract】** To solve the problem that the decoder part of Deeplab V3+ model does not have sufficient multi-scale connection to the feature graph, which will lead to low segmentation precision of the final semantic segmentation graph, this paper adds an intermediate semantic feature graph to the encoder part of Deeplab V3+ model, and carries out concat fusion on the obtained feature graph in the decoder part. Furthermore, the segmentation accuracy is improved. Experimental results on open datasets show that the average intersection ratio of the modified Deeplab V3+ model is 0.76% higher than that of the original model.

**【Key words】** semantic segmentation; deepLab V3+ model; decoder; average intersection and ratio

## 0 引 言

图像语义分割在计算机视觉领域起着重要的作用，在虚拟现实[1-2]、医学影像[3-4]、人机交互[5-6]等领域有着越来越普遍的应用。

深度学习[7]与传统语义分割算法的结合，使图像语义分割精度得到极大的提升。全卷积网络(FCN)[8]为最初与深度学习结合的网络，其是传统卷积神经网络(CNN)的扩展，为减少计算量 FCN 将 CNN 中的全连接层转化为卷积层。但 FCN 产生的分割图较为粗略。SegNet[9]为了提高效果，复制了最大池化指数，引入更多的跳跃连接。这些语义分割模型在空间分辨率方面有着明显的缺陷，于是 RefineNet[10]利用残差连接的思想，降低了内存使用量，提高了模块间的特征融合。由于基于 FCN 的很多架构都未引入充分的全局信息，PSPNet[11]提出了一个金字塔池化模块，充分利用了局部信息与全局信息，使得最后的分割结果更加精确。综上所述可以发现，提高图像语义分割的精确度是目前的主要研究方向和热点。

Google 团队自 2015 起提出了一系列 DeepLab 模型[12-15]，在语义分割领域有着重要作用。虽然其中的 DeepLab V3+模型的分割效果最优，但其在解码器部分对于特征图的多尺度连接不够充分，使最终的语义分割图的分割精细度尚有提高的空间。本文据此提出了一种基于 DeepLab V3+改进的模型，优化了编码器与解码器部分，在公开数据集上进行验证，结果表明 MIoU 相较于原模型有所提高。

## 1 DeepLab V3+模型与改进

### 1.1 DeepLab V3+基础模型

DeepLab V3+网络模型为编解码结构。编码器部分的基础网络 ResNet101 提取图像特征，生成语义特征图；ASPP 模块则将空洞卷积与 SPP 进行结

**基金项目**：2020 年“沈阳工业大学大学生创新创业训练计划”(国家级)项目资助(202010142001X)。

**作者简介**：徐志凡(2000-)，男，本科生，主要研究方向：机器学习与计算机视觉；杜洪波(1977-)，男，硕士，副教授，主要研究方向：数据挖掘理论与应用；韩承霖(2000-)，男，本科生，主要研究方向：机器学习与计算机视觉；李恒岳(2000-)，男，本科生，主要研究方向：机器学习与计算机视觉；祁 新(1996-)，女，硕士研究生，主要研究方向：智能优化算法与可靠性理论；林凯迪(1995-)，女，硕士研究生，主要研究方向：数据挖掘理论与应用；黎 诗(1998-)，女，本科生，主要研究方向：机器学习与计算机视觉。

**通讯作者**：杜洪波 Email：sut_duhongbo@ sut.edu.cn

**收稿日期**：2021-04-30

合,对生成的特征图进行不同扩张率的空洞卷积采样,将得到的特征图 concat 融合后进行 1x1 的卷积,最后得到具有高级语义信息的特征图。解码器从基础网络 ResNet101 的某一个 block 中提取一张带有低级语义信息的特征图,将其与编码器所得的高级语义特征图进行 concat 融合,最后进行上采样得到与输入图像同样大小的语义分割图。该模型结构如图 1 所示。

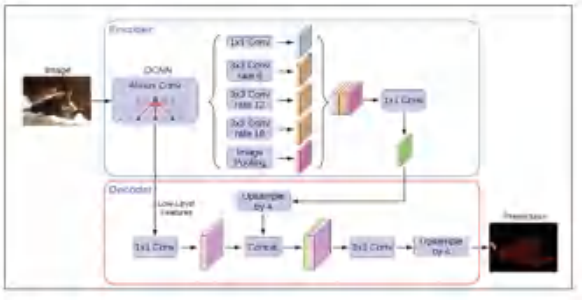

图 1 DeepLab V3+模型结构图

Fig. 1 DeepLab V3+ model structure diagram

### 1.2 改进的 DeepLab V3+模型

对于语义分割来说,在计算量减少的同时分割精细度越高越好。虽然 DeepLab V3+算法可以达到较高的分割精细度,但其在解码器部分对于特征图的多尺度连接并不充分,仅有高级语义特征图与低级语义特征图的连接会使模型的学习能力不足。为了提高模型的学习能力,得到更为精细的语义分割图,且在不增加计算量的前提下,可利用编码器结构中的 ASPP 模块,增加中级语义特征图。虽然在 ASPP 模块中对基础网络中得到的特征图进行了多尺度信息的提取,但是不同尺度的特征图包含的信息是不同的,且不同尺度的特征图中的信息差异较大,统一进行融合后很难学习。为此,本文模型引入中级语义特征图。中级语义特征图含有丰富的语义信息,使得解码器部分高级语义特征图与低级语义特征图的连接更为平滑,保留了更多的细节信息。经实验对比,改进后的模型分割精度有所提高。

在不增加计算量的前提下对编码器中的 ASPP 模块进行改进,一方面先将基础网络 ResNet101 所得的语义特征图并行处理,采用扩张率分别为 6、12、18 的 3x3 卷积提取特征,将多尺度信息做 concat 融合处理,并通过 1x1 卷积,调整中级语义特征图在语义分割预测图中所占的比重。另一方面将 ASPP 模块前两层输出的特征图同样以 concat 融合处理得到高级语义特征图,并和中级语义特征图一起输入到解码器部分,为图像语义分割做准备。改进后的整体模型的结构如图 2 所示。

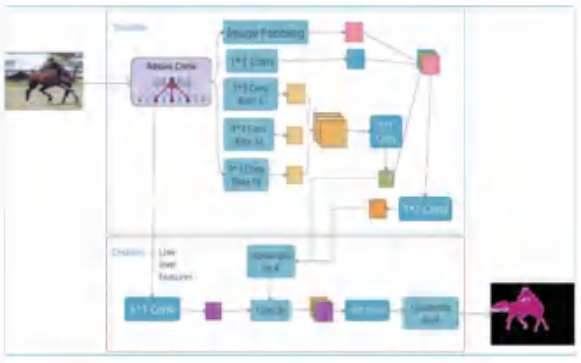

图 2 改进的 DeepLab V3+模型图

Fig. 2 Improved DeepLab V3+ model diagram

## 2 实验及结果分析

### 2.1 实验环境

实验仿真环境为 python3. 6、Anaconda3、TensorFlow1.15、Keras2.2.4。硬件环境为深度学习 GPU 运算塔式服务器主机,采用可支持两个 INTEL XEON SP 的可扩展处理器(10 核/20 线程 2.2G),内存为双 16G(24 个 DIMM 插槽),GPU 使用 1 块 GeForce RTX3070。

### 2.2 数据集

实验采用测试图像语义分割任务模型性能的 2 个主流图像数据集:COCO 2017 数据集[16]、PASCAL VOC 2012 增强版数据集[17]。其中 COCO 2017 数据集进行预训练,PASCAL VOC 2012 增强版数据集用于对模型进行测试和评价。

### 2.3 实验分析

在训练模型开始之前,将训练图像统一裁剪成 513x513 像素,出于高效读取数据的考虑,将图像转化为 Tfrecord 文件。为增强分割图片的显示效果,对真实结果和预测结果采用 RGB 彩色图显示。训练参数见表 1。

表 1 模型参数配置

Tab. 1 Training parameters

| 参数 | 值 |
|---|---|
| 初始学习率 | 0.007 |
| 最终学习率 | 0.000 001 |
| 动量 | 0.9 |
| 批尺寸 | 4 |
| 权重衰减 | 0.000 5 |
| 最大迭代次数 | 200 000 |
| 批正常衰减 | 0.999 7 |

学习率采用多项式自动衰减，当迭代次数超过 200 000 次，学习率为 0.000 001。对损失函数采用动量梯度下降法优化，在 PASCAL VOC 2012 增强版数据集上共计迭代150 307次。总损失函数为交叉熵损失，如式(1)所示：

$$L = -\sum_{c=1}^{M} \boldsymbol{y}_c \log(p_c). \tag{1}$$

其中：$M$ 代表类别数；$\boldsymbol{y}_c$ 是一个 one-hot 向量，元素只有 0 和 1 两种取值(若该类别和样本类别相同则取 1，否则取 0)；$p_c$ 表示预测样本属于 $c$ 的概率。

总损失如图 3 所示。由图中可见，总损失在大约 14 万次左右开始收敛。

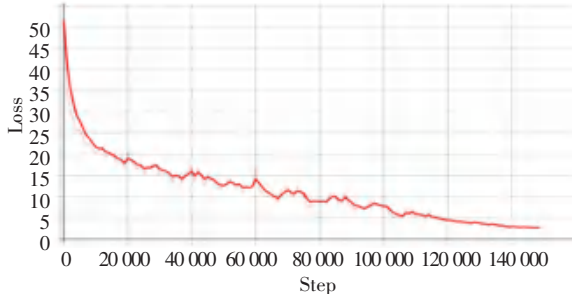


**图 3　总损失图**

**Fig. 3　Total loss graph**

如图 4 所示，改进后的模型不仅在单个目标的图像中(图 4 中第一行)有着良好的分割效果，在拥有多个目标的图像中(图 4 中第二行)也有不错的分割精细度。

(a) 原图　(b) 真实结果　(c) 预测结果

(a) The original figure　(b) Real results　(c) Prediction results

**图 4　改进后模型在验证集上效果**

**Fig. 4　Improved performance on the validation set model**

## 2.4　实验对比

通常在语义分割领域有 4 种经典评价指标：像素准确度(*PA*)、均像素准确度(*MPA*)、平均交并比(*MIoU*)以及频权交并比(*FWIoU*)。本实验选用 *MPA* 与 *MIoU* 作为衡量标准。

(1) *MPA*：计算分割正确的像素数量占像素总数的比例，再取平均：

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}. \tag{2}$$

(2) *MIoU*：计算分割图像与原始图像真值的重合度，再取平均：

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{3}$$

其中，$p_{ij}$ 表示真实值为 $i$，被预测为 $j$ 的数量；$p_{ii}$ 是真正的数量；$p_{ij}$ 表示预测为真但实际为假的数量；$p_{ji}$ 表示预测为假但实际为真的数量。

3 种模型对比测试结果见表 2。由此可见，改进后的 DeepLab V3+模型不仅相对于 PSPNet 模型在均像素精度(*MPA*)上提高了 16.08 %，平均交并比(*MIoU*)提高了 6.25 %，而且相对于原 DeepLab V3+模型在均像素精度(*MPA*)上提高了 0.54 %，平均交并比(*MIoU*)提高了 0.76%，验证了改进的模型有着更好的分割效果。

**表 2　3 种模型对比测试结果**

**Tab. 2　Comparison test results of three models**

| 模型 | *MPA*/ % | *MIoU*/ % |
|---|---|---|
| PSPNET | 79.04 | 73.36 |
| DeepLab V3+ | 94.58 | 78.85 |
| 改进的 DeepLab V3+ | 95.12 | 79.61 |

为进一步体现模型的分割性能，采用模型输出的语义分割图像来对比说明。在图 5 中：(a)为原图，(b)为 DeepLab V3+分割的效果，(c)为改进的 DeepLab V3+分割的效果。图中黄色圈所标注的是改进前后二者之间的差别，验证改进的 DeepLab V3+分割效果更优。例如：从图 5 中第一行可以看出，改进的 DeepLab V3+模型更为精细的分割出了飞机尾翼，而 DeepLab V3+模型并没有达到；由 5 中第二行可见，改进的 DeepLab V3+模型对鸟类的头，羽毛与尾部的边界分割相比于 DeepLab V3+更为精准。这表明改进的 DeepLab V3+模型在增加了中级语义特征图后模型的学习能力更强，对边界的分割的精细度更精准。

改进后的模型不仅分割效果更优而且在单张图片处理速度(*MS*)与模型大小(*MB*)上也更优。在单张图片的运行时间上，改进后的模型速度提高约 6.41%，且模型容量减少了 11.2%，详见表 3。

(a) 原图 (b) DeepLabV3+模型 (c) 改进的模型

(a) The original figure (b) DeepLab V3+ model (c) This model

**图 5 基于两种模型对比分割结果**

**Fig. 5 Compare segmentation results based on two models**

**表 3 两种模型对比测试结果**

**Tab. 3 Comparison test results of two models**

| 模型 | 单张图片处理时间 | 模型大小 |
|---|---|---|
| DeepLab V3+ | 191.7 | 487.2 |
| 改进的 DeepLab V3+ | 179.4 | 420.6 |

## 3 结束语

本文针对 DeepLab V3+模型在解码器部分对于特征图的多尺度连接不充分的问题,提出了一种基于 DeepLab V3+模型的改进算法,该算法对 DeepLab V3+网络模型进行了优化,在解码器部分增加了中级特征层,在 COCO 2017 数据集和 PASCAL VOC 2012 增强版数据集上进行验证,结果表明改进模型的 MIoU 有所提高。但是还存在计算量过大,对于移动端的实时分割还远远达不到要求等问题。因此,减少计算量,轻量化模型结构等将成为下一步的研究方向。

## 参考文献


[1] 徐金馨. VR 技术在医疗中的应用[J]. 通讯世界,2018(10):222-223.

[2] 瓦力斯·阿布力孜,罗岱. 基于虚拟现实的三维动画图像多屏显示系统设计[J]. 现代电子技术,2019,42(19):41-45.

[3] 游齐靖,万程. 基于深度学习的医学图像分割方法[J]. 中国临床新医学,2020,13(2):115-118.

[4] 亢寒,张荣国,陈宽. 基于深度学习的医学图像分割技术[J]. 人工智能,2018(4):30-37.

[5] 徐秋平. 基于人机交互式图割的目标快速提取[J]. 计算机工程与科学,2020,42(2):299-306.

[6] 陈超. 前景与背景分离的图像风格迁移系统设计与实现[J]. 信息通信,2019(4):60-62.

[7] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Date with Neural Networks[J]. Science, 2006, 313(5786): 504-507.

[8] 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.

[9] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder - Decoder Architecture for Image Segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(12): 2481-2495.

[10] Lin G,Milan A,Shen C, et al. RefineNet: Multi-path Refinement Networks for High-Resoluti -on Semantic Segmentation[C]. // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1925-1934.

[11] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network [C].//IEEE International Conference on Computer Vision and Pattern Recognition, 2017: 2881-2890.

[12] Chen L C, Papandreou G, Kokkinos I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs [J]. Computer Science, 2014(4): 357-361.

[13] 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, 2018, 40(4): 834-848.

[14] Chen L C, Papandreou G, Schroff F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. arXiv: 1706.05587, 2017.

[15] 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), 2018: 801-818.

[16] Lin T Y,Maire M, Belongie S, et al. Microsoft coco: common objects in context [C]. //European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[17] Chen X,Mottaghi R, Liu X, et al. Detect what you can: detecting and representing objects using holistic models and body parts[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1971-1978.