

引用格式:熊炜,童磊,李利荣,等. 基于可分离空洞卷积与联合归一化的语义分割算法研究[J]. 微电子学与计算机,2020,37(10):18-23. [XIONG W, TONG L, LI L R, et al. Semantic segmentation algorithm based on separable dilated convolution and joint normalization method[J]. Microelectronics & Computer, 2020, 37(10):18-23.]

基于可分离空洞卷积与联合归一化的语义分割算法研究

熊 炜^{1,2}, 童 磊¹, 李利荣¹, 李 敏^{1,2}

(1 湖北工业大学 电气与电子工程学院, 湖北 武汉 430068;

2 美国南卡罗来纳大学 计算机科学与工程系, 南卡 哥伦比亚 29201)

摘 要: 图像语义分割是图像理解的重要一环, 已被广泛应用于自动驾驶等场景中. 针对信息丢失和语义分割速度慢的问题, 本文提出一种基于可分离空洞卷积和联合归一化的语义分割算法. 首先结合可分离卷积和空洞卷积提取 ResNet101 的后三层输出, 然后在语义分割中应用实例归一化方法, 与应用批量归一化对比, 验证了实例归一化的有效性, 最后提出了两种联合归一化方法, 验证了这两种归一化方法对语义分割效果的提升. 本文方法在 Pascal VOC 2012 数据集进行了实验, 结果表明, 本文方法加速了网络的训练、验证和预测, 交并集之比最高到达了 80.62%.

关键词: 图像语义分割; 可分离空洞卷积; 实例归一化; 批量归一化

中图分类号: TP391.41

文献标识码: A

文章编号: 1000-7180(2020)10-0018-06

Semantic segmentation algorithm based on separable dilated convolution and joint normalization method

XIONG Wei^{1,2}, TONG Lei¹, LI Li-rong¹, LI Ming^{1,2}

(1 School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China;

2 Department of Computer Science and Engineering, University of South Carolina, Columbia 29201, USA)

Abstract: Image semantic segmentation is an important part of image understanding, which is applied to automatic driving. In this paper, we use the Pascal VOC 2012 data set and ResNet101 as the basic network. We propose semantic segmentation algorithm based on separable dilated convolution and improved normalization method to solve the problem of information loss and slow speed. Firstly, we combine separable convolution and dilated convolution to extract the last three layers' output of ResNet101. Compared with standard dilated convolution, separable dilated convolution accelerates the training, validation and prediction of the network. Then, in the semantic segmentation, the instance normalization method is applied and compared with the application batch normalization to verify the effectiveness of batch normalization. Finally, two normalization methods combining batch normalization and instance normalization are proposed to improve the effect of semantic segmentation. This method is tested in Pascal VOC 2012 data set. The results show that our method accelerates the training, validation and prediction of the network. The highest mean intersection over union of this method in Pascal VOC 2012 data set is 80.62%.

Key words: image semantic segmentation; separable dilated convolution; instance normalization; batch normalization

收稿日期: 2019-12-24; 修回日期: 2020-02-08

基金项目: 国家自然科学基金项目(61571182, 61601177); 国家留学基金项目(201808420418); 湖北省自然科学基金项目(2019CFB530)

1 引言

图像分割是一个热门研究方向,包括实例分割和语义分割,虽然都是像素级别的分割,但是实例分割将图像中单个物体分为一类,而语义分割是将图像中相同类别物体都分为一类^[1],广泛应用于自动驾驶场景中。

得益于深度卷积神经网络的发展,图像的语义分割也在不断加速发展。首先,完全卷积网络(FCN)^[2]的提出令端到端的语义分割得以实现,然后 SegNet 的提出极大地提高了语义分割的效果,确立了语义分割的编解码结构^[3]。当前的研究基本都使用已提出的深度卷积模型作为基本骨架,以 ResNet 使用最多。在其基础上,Zhao 等人^[4]提出金字塔模型,结合图像空间尺度信息进行语义分割。Yu 等人^[5]从改进标准卷积出发,提出了空洞卷积(Dilated convolution),旨在有相同的卷积核,扩大感受野的区域,使得卷积过程提取更广泛的图像信息。Chen 等人^[6]结合金字塔模型和空洞卷积,提出空间金字塔池化模型(Atrous Spatial Pyramid Pooling, ASPP),而 Li 等人^[7]结合金字塔模型和注意力机制提出了注意力金字塔模型(PAN),在不少模型中,也应用了池化层对图像进行降维处理。

目前的研究中使用 ResNet101 为骨架网络的较多,将 ResNet101 的最后一层的输出接入相应模型作为编码结构,输出高层次特征,将 ResNet101 的低一层输出作为低层次特征,解码结构连接高层次特征和低层次特征,以此完成图像的语义分割。这些结构均未完全利用 ResNet101 的输出,造成了信息丢失,同时采用的池化操作也在同时增加信息丢失。本文结合空洞卷积与可分离卷积,设计了可分离空洞卷积,利用深度可分离空洞卷积同时提取 ResNet101 最后三层输出,与低层次特征结合作为编码结构,解码结构应用全局池化结合全局信息,从而完成语义分割,并在整个过程中,使用双线性插值恢复图像尺寸。本文使用 SoftMaxCrossEntropy 作为损失函数,ReLU 函数为激活函数,应用 Dropout,在归一化操作上,首先使用 BN,然后使用实例归一化(Instance Normalization, IN)^[8],验证 IN 的效果,最后结合 BN 和 IN 改进归一化方法,提出了 BN、IN 并联归一化(ParallelBatch and Instance Normalization, PBIN)方法和 BN、IN 并联归一化(CascadedBatch and Instance Normalization, CBIN)方法,提高语义分割效果。

2 本文方法介绍

2.1 总体框架

本文的总体框架如图 1 所示,图中省去卷积之后的归一化、DropOut 和 ReLu 激活函数等部分操作。首先,在编码结构上,选择 ResNet101 作为骨架网络进行特征提取,ResNet101 的后三层特征输出如图 1 中 f_1 、 f_2 和 f_3 所示,结合可分离卷积和空洞卷积对 f_1 、 f_2 和 f_3 进行特征提取,通过插值处理成相同尺寸,进行联合;然后,在解码结构上,将 f_1 作为低层次信息,进行全局池化,再插值处理使得与编码输出的尺寸相同,与解码输出 f_4 进行联合,通过 3×3 卷积进一步处理联合后的特征,再经过 1×1 卷积将特征的深度降为 21,与 Pascal VOC 2012 数据集图像的类别数相同;最后,使用 SoftMaxCrossEntropy 函数作为损失函数,训练语义分割模型,应用双线性插值和图像变换,输出语义分割结果。

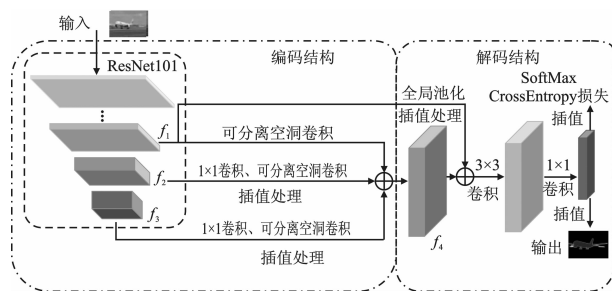


图 1 本文总体框架

2.2 可分离空洞卷积

本文 ResNet101 后的解码结构采用可分离空洞卷积,空洞卷积如图 2 所示,卷积核为 3×3 ,空洞卷积率(rate)为 1 时,卷积为标准的 3×3 卷积,而 rate 为 2 时,在 3×3 卷积核之间插入一个为 0 的值,卷积的感受野与标准的 5×5 卷积相同,以此类推,在语义分割任务中,卷积核相同时,感受野增大极大的提高了语义分割效果。

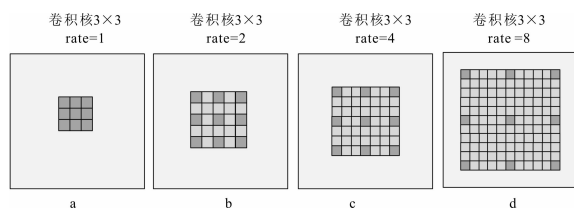


图 2 空洞卷积

标准卷积如图 3 所示,特征图像的每个通道特征都与卷积核进行卷积运算,然后融合为一个通道

的特征图像,完成卷积过程后通道数改变为卷积核的个数。

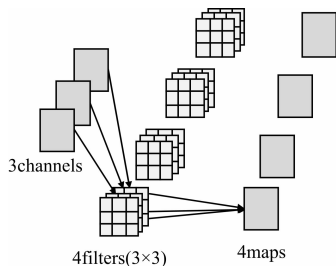


图3 标准卷积

可分离卷积如图4所示,特征图像的每个通道特征单独与每个卷积核进行卷积运算,完成卷积后不改变特征图像的通道数,加入 1×1 卷积改变通道数,到达与标准卷积同步的输出。

设可分离卷积的卷积核大小为3,而 Res-Net101 后的三层特征图像的通道数分别为 512、1 024 和 2 048,则可分离卷积可减少 8~9 倍的计算量,因此本文模型在训练、验证和预测的时间上比采用标准卷积要少。

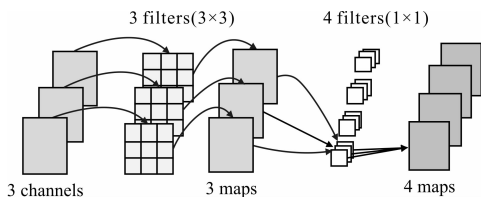


图4 可分离卷积

2.3 归一化方法

目前语义分割网络使用的归一化方法均为 BN, BN 是在一个批量图像的每个通道上进行归一化, batchsize 设置太小意义不大,而在风格迁移任务中广泛使用的归一化方法为 IN, IN 是在每个图像上进行归一化,不受 batchsize 和通道影响,两种方法各异. 本文结合 BN 和 IN,在 ResNet101 之后的编码部分使用了如图5所示的四种归一化方法, BN 的定义如式(1):

$$y_{i_BN}^C(x_i^C) = \alpha \times \frac{x_i^C - \frac{1}{m} \sum_{i=1}^m x_i^C}{\sqrt{\frac{1}{m} \sum_{i=1}^m (x_i^C - \frac{1}{m} \sum_{i=1}^m x_i^C)^2 + \epsilon}} + \beta \quad (1)$$

式中, x_i^C 表示输入特征图像; $y_{i_BN}^C$ 表示批量归一化结果; m 为 batchsize 值; C 表示通道数; ϵ 为常量; α 和 β 为可学习参数。

IN 的定义如式(2):

$$y_{i_IN}^C(x_i^C) = \frac{x_i^C}{\sum_{i=1}^W \sum_{j=1}^H x_i^C} \quad (2)$$

式中, H 和 W 表示特征图像的高度和宽度(尺寸); $y_{i_IN}^C$ 表示实例归一化结果。

图5(c)所示 PBIN 方法的定义如式(3)所示:

$$y_{i_PBIN}^C = \begin{cases} y_{i_BN}^{C_1} & C_1 \in [1, C_1) \\ y_{i_IN}^{C_2} & C_2 \in (C_2, C] \end{cases} \quad (3)$$

BN 和 IN 均分特征通道进行归一化,其中 $C_1 = \frac{C}{2}$ 和 $C_2 = \frac{C}{2} + 1$, $y_{i_PBIN}^C$ 表示 PBIN 结果。

图5(d)所示 CBIN 方法的定义如式(4):

$$y_{i_CBIN}^C = y_{i_IN}^C(y_{i_BN}^C) \quad (4)$$

式中, $y_{i_BN}^C$ 和 $y_{i_IN}^C$ 分别如式(1)和式(2)所示, $y_{i_CBIN}^C$ 表示 CBIN 结果。

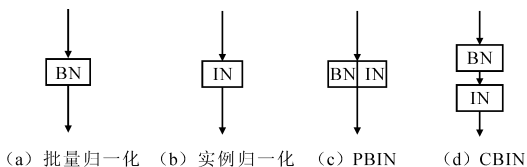


图5 归一化

2.4 损失函数

语义分割任务使用的损失函数基本都是 Soft-MaxCrossEntropy 函数,其中, CrossEntropy 函数如式(5):

$$\text{SoftMax CrossEntropy} = - \sum_i y'_i \log(y_i) \quad (5)$$

y_i 表示预测图像的概率值,如式(6)

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^k e^{x_j}} = \frac{\exp(x_i)}{\sum_{j=1}^k \exp(x_j)} \quad (6)$$

式中, x_i 和 x_j 表示预测图像第 i 和第 j 类的值,本文使用 Pascal VOC 2012 数据集,因此 $i, j, k \in \{1, 2, 3, \dots, 21\}$ 。

y'_i 表示标签图像的概率值,如式(7)

$$y'_i = \frac{e^{x'_i}}{\sum_{j=1}^k e^{x'_j}} = \frac{\exp(x'_i)}{\sum_{j=1}^k \exp(x'_j)} \quad (7)$$

式中, x'_i 和 x'_j 表示标签图像第 i 和第 j 类的值。

3 实验结果与分析

本文使用 Pascal VOC 2012 数据集,包含了 10 582 张训练图像,1 446 张验证图像和 1 456 张测试图像, Pascal VOC 2012 数据集图像大小 300~500 不等,标签图像总共有 21 个不同类别物体,每种类

别物体在标签图像上表示为不同像素颜色,Pascal VOC 2012 公布了其中的训练集和验证集的原始图像和标签,测试集只公布了原始图像,因此本文网络训练是在训练集上进行的,指标评价是在验证集上进行的,语义分割预测结果则使用测试集.

本文使用 Python 作为编程语言,Pytorch 作为深度学习框架,建立语义分割模型,GPU 型号为 8G GeForce RTX 2070,性能评估指标使用平均交并集之比(mIoU),其值越高则表示语义分割效果越好.实验中,ResNet101 骨架网络使用的预训练参数是在 ImageNet 上进行预训练的,将输入图像大小调整为 400×400 ,然后随机裁剪为 384×384 (预处理)输入网络中,设置学习率为 0.001,学习率衰减为 0.9,权重衰减为 0.000 1,迭代周期为 240,batch-size 为 8.

3.1 可分离空洞卷积实验对比

首先,在 Pascal VOC 2012 数据集上对本文网络模型进行训练,对比标准空洞卷积和可分离空洞卷积的实验结果如表 1 所示.

表 1 标准空洞卷积和可分离空洞卷积实验对比

方法	mIoU (%)	train (hour)	Val (fps)	Pred (fps)
标准空洞卷积	78.77	91.94	32.45	28.54
可分离空洞卷积	78.48	83.19	35.17	33.09

标准空洞卷积和可分离空洞卷积两种方法相比,mIoU 值相差不大,可分离空洞卷积方法训练时间减少 8.75 个小时,使用可分离空洞卷积方法验证时的帧率提高近 3 fps,预测时,使用可分离空洞卷积方法的帧率提高近 4 fps,表明,可分离空洞卷积方法的语义分割效果与应用标准空洞卷积方法的效果基本相同,但是可分离空洞卷积减少了模型计算量,在训练、验证和预测时,速度更快,表现出可分离空洞卷积的优势.

3.2 归一化方法实验对比

本本文首先在 ResNet101 之后的编码部分分别使用 BN、IN、PBIN 和 CBIN 四种归一化方法,以此对比归一化方法的效果,其对比结果如表 2 所示.

首先,IN 方法比 BN 方法的效果要好,但是效果提升并不大,有接近 0.005 的效果提升.通过对比,我们发现 IN 方法在椅子(chair)、植物(plant)、沙发(sofa)等具有小细节的类别物体上效果显著,桌子(table)和羊(sheep)等类别物体效果明显不如 BN 方法,其他类别物体上,两种方法的效果差距并

不大.因此,语义分割任务当中使用 IN 方法是可行的,BN 和 IN 在不同类别物体的表现上各有优势,就语义分割整体效果而言,IN 方法较好与 BN 方法.

表 2 归一化方法实验对比

方法	BN	IN	PBIN	CBIN
background	93.94	93.77	84.08	94.26
aeroplane	91.08	91.66	93.82	93.11
bike	63.47	62.74	45.76	46.95
bird	91.35	89.97	93.57	92.67
boat	74.24	74.10	75.26	72.72
bottle	77.92	76.13	85.06	84.14
bus	90.34	90.68	93.57	93.71
car	86.84	88.50	92.03	90.21
cat	91.34	92.15	95.16	91.77
chair	35.25	38.53	44.16	44.54
cow	89.90	91.79	93.15	91.80
table	64.82	61.90	54.04	50.44
dog	88.05	87.53	90.98	89.90
horse	87.47	89.41	90.75	90.55
motorbike	83.99	85.34	88.07	87.49
person	85.80	85.84	88.42	88.39
plant	58.69	60.80	69.41	72.86
sheep	87.23	84.93	90.08	84.85
sofa	45.46	51.53	51.67	55.26
train	85.34	85.03	89.70	90.47
television	75.53	75.19	74.29	76.31
mIoU(%)	78.48	78.93	80.62	80.11

其次,PBIN 方法比 CBIN 方法效果好,效果提升了 0.005.通过对比,我们发现在自行车(bike)和桌子(table)这两类物体上,两种方法均比 BN 和 IN 方法在效果上有大量下降,而在椅子(chair)和植物(plant)两类物体上有大量的效果提升,分析有以下原因:首先,这四类物体都有大量的细节部位,本身存在语义分割的难度;其次,这四类物体的细节类型上也存在差异;最后,BN 和 IN 两种方法之间存在相互影响,因此,在这四类物体上效果存在很大差异,不过不影响这两种方法的有效性. PBIN 方法中,BN 和 IN 的处理上是分离的,BN 和 IN 处理结果是在不同的通道上,相互之间在计算没有影响,只在输入到下一层的时候影响特征的提取,而 CBIN 方法是先使用 BN 后使用 IN,BN 的计算结果直接对之后的 IN 产生了一些影响,BN 使得一些数据分

布出错,导致之后 IN 处理效果不好,最终 PBIN 方法比 CBIN 方法处理效果好. 因此,本文提出的 PBIN 方法和 CBIN 方法能够有效提升语义分割的效果,而 PBIN 方法比 CBIN 方法好.

最后,作者将本文提出的语义分割方法同近年其他语义分割的方法进行评价指标的对比,mIoU 值均为验证集上的结果,如表 3 所示.

本文提出的模型加入 PBIN 后,mIoU 到达了 80.62%,较 DFN、PAN 和 DUpsamling 三种方法有 1%的性能提升,较 DeeplabV3+方法有近 2%的性能提升,较 MsNet-4 方法也有 5%的性能提升,而使用 CBIN 方法后,mIoU 也到达了 80.11%,比其他方法都高. 因此,可以证明本文提出的语义分割方法是有效的.

表 3 语义分割方法对比

方法	MIoU(%)
MsNet-4 ^[9]	75.80
DFN ^[10]	79.54
DeeplabV3+ ^[6]	78.85
PAN ^[7]	79.38
DUpsamling ^[11]	79.67
PBIN(ours)	80.62
CBIN(ours)	80.11

3.3 可视化结果

使用本文语义分割的模型,分别使用四种归一化方法,在 Pascal VOC 2012 测试集上生成语义分割的可视化结果,如图 6 所示,BN、IN、PBIN 和 CBIN 对应四种归一化方法生成的结果. 忽略不同方法训练模型的误差,可以看到,PBIN 和 CBIN 这两张方法比 BN 和 IN 这两种方法好,在大部分物体类别的细节方面也处理的更好,而在 bike 这类物体上也是处理的较差,与在验证集上的评价指标符合.

4 结束语

本本文提出的语义分割方法,通过可分离空洞卷积融合 ResNet101 网络的三层输出,能够减少网络的计算量,在训练、验证和预测时速度更快,在 Pascal VOC 2012 训练集上的训练时间减少了 8.75 小时,验证和预测时的帧率提高了 3~4 fps,验证了 IN 方法在语义分割中的效果,同时结合 BN 与 IN 两种归一化方法,有效的提高了语义分割的效果,本文使用的 PBIN 归一化方法 mIoU 达到了 80.62%,较 DFN、PAN 和 DUpsamling 三种方法有 1%的效

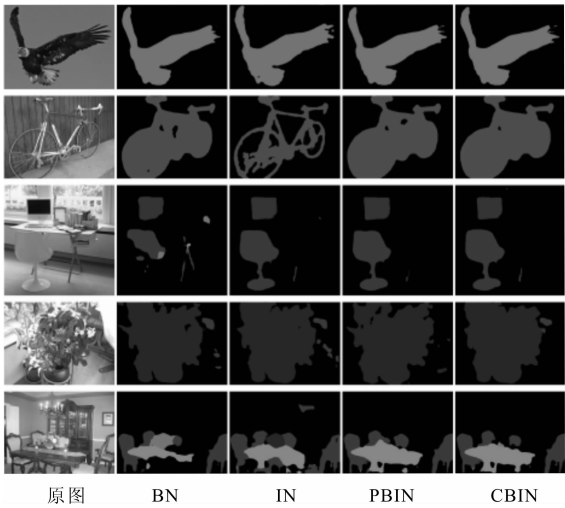


图 6 Pascal VOC 2012 预测集语义分割结果

果提升,较其他方法也有 2%~5%的效果提升. 本文使用 ResNet101 作为深度神经网络提取特征,有进一步改进空间来提升语义分割训练、验证和预测的速度,同时网络在不规则物体上的语义分割效果也有待改进.

参考文献:

[1] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, NV, USA, 2016: 770-778.

[2] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015: 3431-3440.

[3] CHAURASIA A, CULURCIO E. LinkNet: Exploiting encoder representations for efficient semantic segmentation [C]// Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 2017: 1-4.

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

[5] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions [EB/OL]. arXiv preprint arXiv: 1511. 07122, 2015. <https://arxiv.org/abs/1511.07122>.

[6] CHEN L C, ZHU Y K, Papandreou G, et al. En-

coder-decoder with atrous separable convolution for semantic image segmentation [C]// Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich Germany, 2018: 801-818.

[7] LI H C, XIONG P F, AN J, et al. Pyramid attention network for semantic segmentation [EB/OL]. arXiv preprint arXiv:1805.10180, 2018. <https://arxiv.org/abs/1805.10180>.

[8] ULYANOV D, VEDALDI A, LEMPITSKY V. Instance Normalization: The missing ingredient for fast stylization [EB/OL]. arXiv preprint arXiv: 1607.08022, 2016. <https://arxiv.org/abs/1607.08022>.

[9] CHEN Z. Research on semantic segmentation based on convolutional neural network [D]. Beijin;Beijing Jiaotong University, 2018.

[10] YU C Q, WANG J B, PENG C, et al. Learning a discriminative feature network for semantic segmentation [C]// Proceedings of the 2018 IEEE/CVF Conference

on Computer Vision and Pattern Recognition(CVPR), Salt Lake City, UT, USA, 2018: 1857-1866.

[11] TIAN Z, HE T, SHEN C H, et al. Decoders matter for semantic segmentation: data-dependent decoding enables flexible feature aggregation [EB/OL]. arXiv preprint arXiv:1903.02120, 2019. <https://arxiv.org/abs/1903.02120>.

作者简介:

熊 炜 男,(1976-),博士,副教授. 研究方向为数字图像处理和计算机视觉.

童 磊 男,(1993-),硕士研究生. 研究方向为计算机视觉与图像语义分割.

李利荣(通讯作者) 女,(1974-),博士,副教授. 研究方向为模式识别与智能系统. E-mail:Rongli@hbut.edu.cn

李 敏 女,(1976-),高级工程师. 研究方向为计算机视觉与模式识别.