

语义分割模型：DeepLab

张皓飞 (11921056)

2020 年 5 月 26 日

目录

1 引言	1
1.1 计算机视觉中的任务	1
1.2 语义分割任务	1
2 相关工作	1
2.1 语义分割任务	1
2.2 编-解码器架构	2
3 方法	2
3.1 DeepLab-V1	2
3.1.1 稠密图像分类器	2
3.1.2 全连接条件随机场	3
3.1.3 小结	4
3.2 DeepLab-V2	4
3.2.1 多孔卷积	4
3.3 DeepLab-V3	5
4 实验结果	5
4.1 数据集	5
4.2 其它设定	6
4.3 训练结果	6
5 总结	7
参考文献	9

1 引言

1.1 计算机视觉中的任务

计算机视觉想要解决的问题就是如何让机器真正理解一张输入图片，那么首先，机器必须要知道这张图片所描述的对象是什么。比较简单做法就是将所有图像描述的对象划分成一系列的类别，并对输入的图片根据划分好的类别进行分类，这样机器就可以获得输入图片的最基本语义信息。因此分类问题就是计算机视觉中是一个至关重要的任务。

传统的解决分类的任务为先根据输入的图片的大致分布人为构建一系列特征提取器，比较有名的特征提取器为 SIFT^[1]、SURF^[2] 和 HOG^[3] 等。这些特征提取器将输入的图像变换为具有固定长度的向量，之后再使用线性分类器如 Rosenblatt 感知机^[4]、贝叶斯分类器^[5] 和支持向量机 (SVM)^[6] 等或非线性的分类器如决策树^[7]、基于核方法的支持向量机^[6] 进行分类。

尽管这些传统的图像分类算法具有严格证明的数学定理和证明做支撑，但由于分类器以及研究得非常透彻，以至于大部分的算法构建时间都会用在如何从图像提取特征的问题，也称为“特征工程”。由于图像的分布对于不同的任务差异巨大，以至于这些特征工程算法很难进行泛化和推广，因此研究人员需要花费大量的时间进行模型的调参，从而得到令人满意的结果。

随着计算机软硬件和互联网的蓬勃发展，研究人员可以获得大量的数据，然而模型的效果并没有随着数据量的增加而显著提高，而于此同时，基于卷积神经网络的分类器却显著地由于传统的特征工程方法。在 2012 年的 ILSVRC^[8] 的比赛中 Alex Krizhevsky 通过训练卷积神经网络 AlexNet^[9] 并获得了冠军。

尽管卷积神经网络并没有像传统机器学习一样具有很强的可解释性，但由于其可表述更强的非线性和更大的参数空间，使得深度模型可以往往收敛到较传统方法更优的模型。

1.2 语义分割任务

有了对图片的分类，机器可以初步理解图像所表示的类别，但是通常图片里具有多个物体，每个物体可能都属于不同的类别，因此我们希望机器可以对一张图片的所有物体全部标记在图像中，从而可以获取到图片中更多的信息。从本质上讲，语义分割任务即对输入的图片的每个像素点进行分类，并将相同类别的像素点进行组合，从而得到物体对应的位置。

有了对图片的语义分割，我们可以获得更高层次的语义信息。在自动驾驶、人机交互和虚拟现实等诸多应用中都需要输入图片的更高层次的语义信息。因此语义分割任务也是计算机视觉里重要的任务之一。

本文中将介绍一种基于深度卷积网络的语义分割模型 DeepLab^[10-12] 系列，该算法在 Pascal VOC 数据集^[13] 和 MSCOCO 数据集^[14] 取得了当时较好的结果。

2 相关工作

2.1 语义分割任务

关于深度卷积网络 (DCNNs) 在像素级语义分割中有大致三种主要方法：

基于区域的语义分割 基于区域的方法通常基于目标检测架构如 R-CNN^[15] 等，对于分割的任务，R-CNN 首先利用选择性搜索提取大量的候选区域，并计算其对应的特征。之后再使用线性分类模型对每个区域进行分类，再将区域的预测转换为像素预测。这种方法通常无法端到端实现，计算量较大。

弱监督语义分割 通常对语义分割的数据集的标注是非常繁琐的，为此许多方法提出使用边框注释来作为监督信息训练模型，而非像素级的标注如 BoxSup^[16] 算法致力于通过使用带注释的边界框来实现语义分割。

全卷积网络语义分割 全卷积网络可以对输入的图像进行端到端的语义分割，从而使得语义分割模型可以快速且方便的训练。全卷积网络大致可看成为编码器-解码器架构，其中编码器部分从图像中提取特征，解码器部分将提取到的特征恢复为语义分割的图像，如图1所示。目前，FCNs^[17]、SegNet^[18] 以及 U-Net^[19] 系列模型都为基于全卷积的语义分割模型类别中。

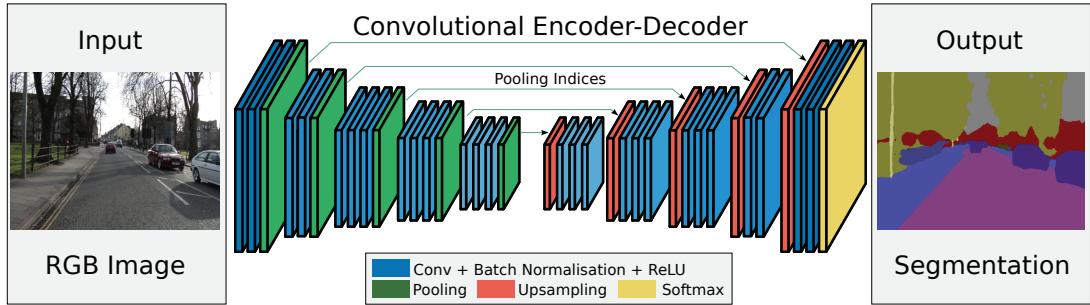


图 1: 编解码器示例^[18]

尽管 DeepLab 系列模型并不完全可以划分到上面三类，但由于其整体依旧可看作编-解码器类型，因此本文中将 DeepLab 划分到全卷积网络语义分割。为此我们需要简单介绍编解码器架构。

2.2 编-解码器架构

在编-解码的模型框架中，整体可以看作两部分：第一部分为编码器，负责将输入的图片转换为在编码子空间中对输入图像的表示（嵌入）；第二部分为解码器，负责将编码子空间中的表示恢复到维度更高的子空间中，如输入图像的子空间。在这种架构中，编码器可以将图像里复杂的目标提取出来，并剔除部分噪声。解码器部分则试图按照任务要求构将编码的信息部分恢复。在编-解码器的框架下，可以完成如语义分割，图像降噪，超分辨率等任务。

一般而言，编码器部分为深度卷积神经网络，且采用最大池化等操作进行维度的缩减，这样可以使得网络具有缩放不变性，平移不变性等。在解码器部分，通常采用转置卷积的操作以恢复维度。

3 方法

3.1 DeepLab-V1

在第一版的 DeepLab 模型中，使用 DCNNs 做为输入图像的特征提取器和稠密图像分类器，像素级的全连接条件随机场 (CRFs)^[20] 作为语义分割部分。其编码器部分可以看作是稠密图像分类器，解码器部分可以看作为全连接条件随机场。

3.1.1 稠密图像分类器

DeepLab 使用在 ImageNet 数据集预训练的 VGG-16 模型^[21] 作为骨干网络。然而在分类任务中强调目标的平移不变性，而在像素级语义分割中，则需要平移可变性。因此，DeepLab 方法将

VGG 模型中的后两次池化层去掉，并使将最后三次卷积操作变成空洞卷积如图2所示。从而将输入图像的尺寸降低到原来的 $\frac{1}{8}$ ，并作为图像的表示。

和池化操作或者步长不为 1 的卷积操作相比，空洞卷积操作可以在降低维度的同时具有更大的感受野，这样就可以在保留更多的空间信息的同时降低维度。尽管我们可以不设置池化操作，从而维持每层网络的输出不变，然而这样会导致计算量的增加，使得网络更加难以优化。

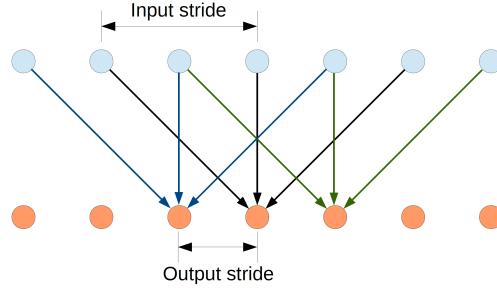


图 2: 一维空洞卷积示例，这里卷积核大小为 3，输入的步长为 2，输出的步长为 1

3.1.2 全连接条件随机场

在稠密图像分类器中，在一定程度上我们可以得到语义分割的结果（尺度为原来的 $\frac{1}{8}$ ），但由于深度卷积网络在准确率和定位上的矛盾，因此直接使用 DCNNs 的输出结果效果不佳。为了解决该问题，研究人员提出了两个解决的方向，一是采用多层输出的信息来对目标的边界进行更好的确定^[17, 22]；另外一类方法采用超像素的表示，将定位变成一种低级的分割方法^[23]。而在 DeepLab 算法中，给出了一种新的解决思路，即结合 DCNNs 的识别能力以及全连接条件随机场的细粒度定位能力。

传统上，条件随机场是用来对分割结果进行光滑处理，通常这些模型包含耦合相邻节点的能量项，倾向于将相同的标签分配给空间上最接近的像素。这些短程的条件随机场的主要功能为清楚奖励在手工设计的特征上的弱分类器的虚假预测。而从图3中我们可以看到，DCNNs 输出的结果已经非常光滑，因此我们的目的不是对其进行光滑处理，而是恢复细节信息。为此 DeepLab 引入全连接条件随机场的模型^[24] 来进行更细致的语义分割。

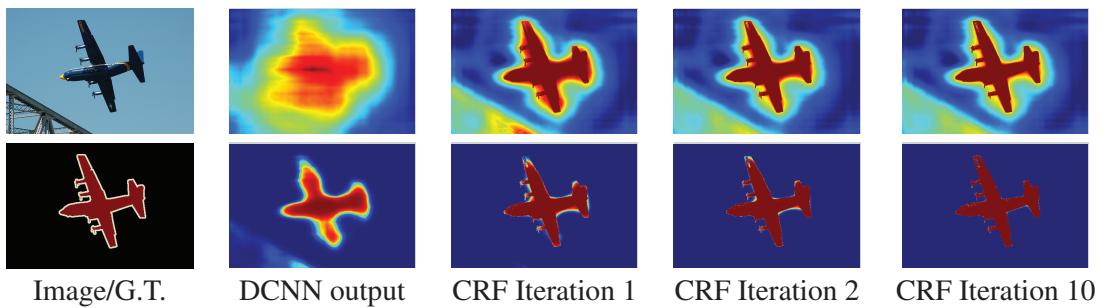


图 3: 使用 DCNN 和全连接条件随机场做分割的比较，其中最左面的为真实标注结果

定义能量函数

$$E(\mathbf{x}) = \sum_i \theta_i(x_i) + \sum_{i,j} \theta_{i,j}(x_i, x_j) \quad (1)$$

其中 \mathbf{x} 为对每个像素的标签。使用一元势函数 $\theta_i(x_i) = -\log P(x_i)$ ，其中函数 $P(x_i)$ 为通

过 DCNN 得到的对像素 i 的概率标签。二元势函数定义为

$$\theta_{i,j}(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^K w_m \cdot k^m(\mathbf{f}_i, \mathbf{f}_j)$$

其中，

$$\mu(x_i, x_j) = \begin{cases} 1 & x_i \neq x_j \\ 0 & x_i = x_j \end{cases}$$

每一个函数 $k_m(\cdot, \cdot)$ ，都为高斯核函数都依赖于第 i 和第 j 个像素点的特征。DeepLab 采用双边位置和颜色项并进行加权

$$w_1 \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\sigma_\beta^2}\right) + w_2 \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\gamma^2}\right) \quad (2)$$

其中第一项依赖于像素点 i 和 j 的距离和像素的值，而第二项仅依赖于距离， σ_α 、 σ_β 和 σ_γ 均为高斯核函数的超参数。

3.1.3 小结

DeepLab 第一个版本的整体流程如图4所示，首先输入的图片经过深度卷积网络得到粗粒度的分割图，经过双线性插值后恢复原图的尺寸，再通过全连接条件随机场获得最终的语义分割结果。

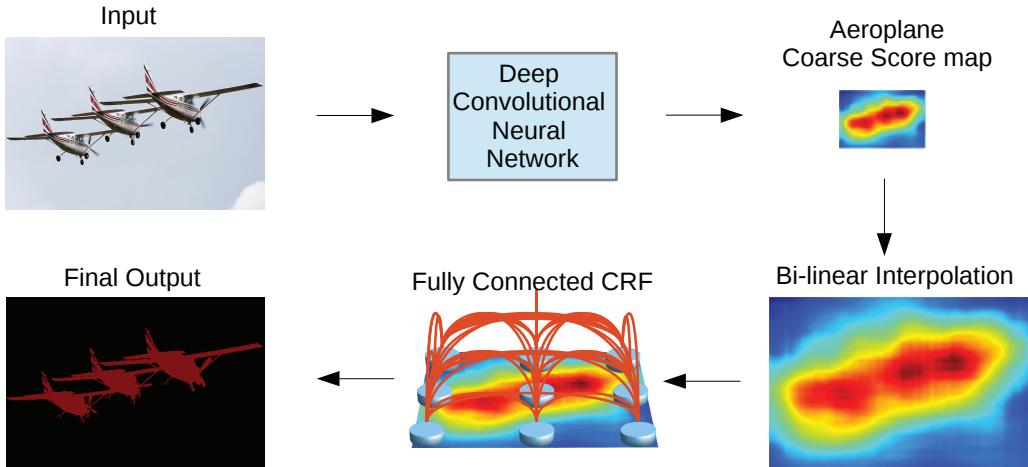


图 4: 模型整体流程

3.2 DeepLab-V2

DeepLab-V2 为对上述 DeepLab-V1 模型的改进，主要改进为将原模型中 DCNN 中的空洞卷积进行改进，采用多个上采样的卷积核组合构成多孔卷积（Atrous Convolution），此外在 DeepLab-V2 中骨架网络由原来的 VGG 改为 ResNet-101^[25]。

3.2.1 多孔卷积

在模型 DeepLab-V1 中，采用空洞卷积来避免对输入图片的空间信息的丢失。在本模型中，目标依然是进一步增加空间信息。为此，采用多个不同孔径大小的空洞卷积核组成多孔空间金字塔池化层（ASPP）如图5所示，其本质为多尺度的空洞卷积的叠加其具体实现如图6所示。

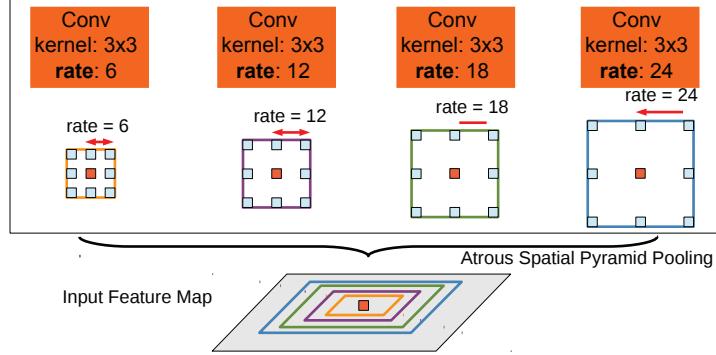


图 5: ASPP 示例

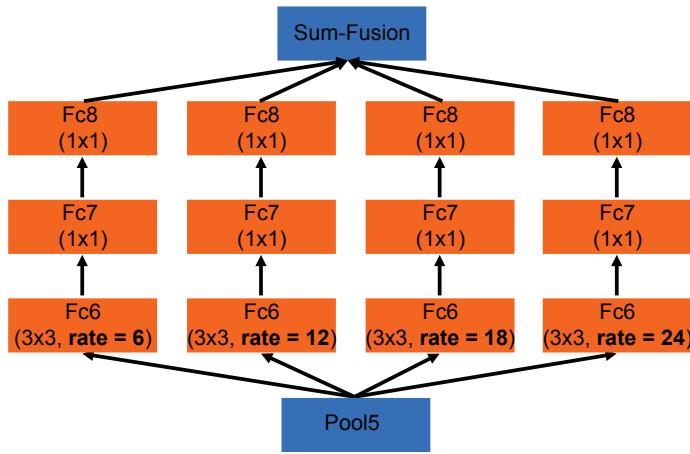


图 6: ASPP 实现方式

3.3 DeepLab-V3

在 DeepLab-V3 模型中，同样使用 ResNet 作为骨架网络如图7所示，网络前 3 个块直接使用了 ResNet 的前 3 个块，其中每个块里卷积层保持输入的维度，最后一层通过步长为 2 的卷积降低特征的维度。若直接使用 ResNet 的结构如7(a)，会导致最终输出的特征维度过低，从而语义分割的位置精度下降。

而采用空洞卷积来加深网络如7(b)，可以在维持特征维度的情况下获取到更大的感受野，从而提高精度。最终，DeepLab-V3 的改进如图8 所示。

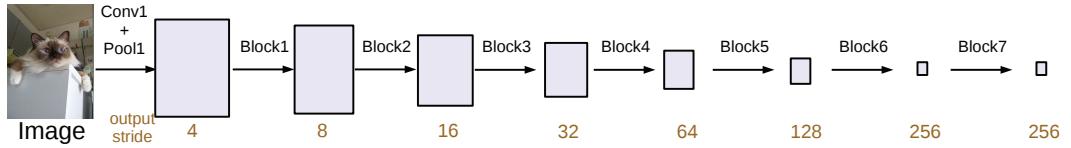
4 实验结果

本章中，我们将展示对模型 DeepLab-V3 复现的结果。

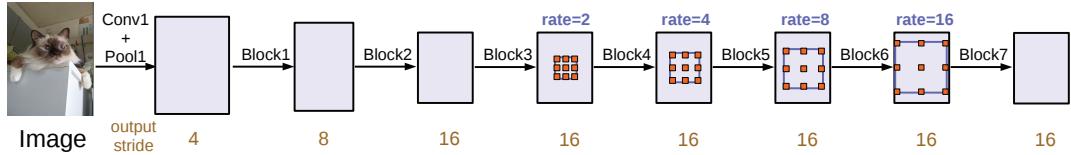
4.1 数据集

我们使用的数据集为 Pascal VOC 和 Cityscapes 数据集。

Pascal VOC 具体而言，我们使用 Pascal VOC 2012 分割数据集，分别包含 1464, 1449 和 1456 张训练、验证和测试的图片。包括背景。主要包含真实场景中的物体，共计 20 个类别。



(a) 不使用空洞卷积加深网络



(b) 使用空洞卷积加深网络

图 7: DCNNs 加深方式示意

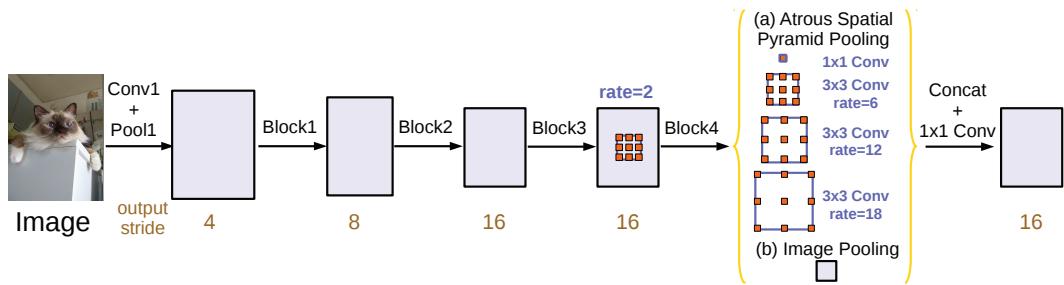


图 8: 改进 ASPP 实现方式

Cityscapes Cityscapes 拥有 5000 张在城市环境中驾驶场景的图像，由 19 个类别的语义分割标注。

4.2 其它设定

模型的骨架网络使用了 ResNet-50 和 ResNet-101。学习率按照表1设定。训练的循环次数按表2确定。

数据集	学习率
Pascal Voc	0.0001
Cityscapes	0.01

表 1: 学习率设定

4.3 训练结果

DeepLab-V3 的训练结果由表3给出。

在随机抽取的测试图片的分割结果如图9, 10所示，可以看出使用 ResNet-101 作为网络骨架训练出的结果要比 ResNet-50 分割效果更好。

数据集	循环次数
Pascal Voc	50
Cityscapes	120

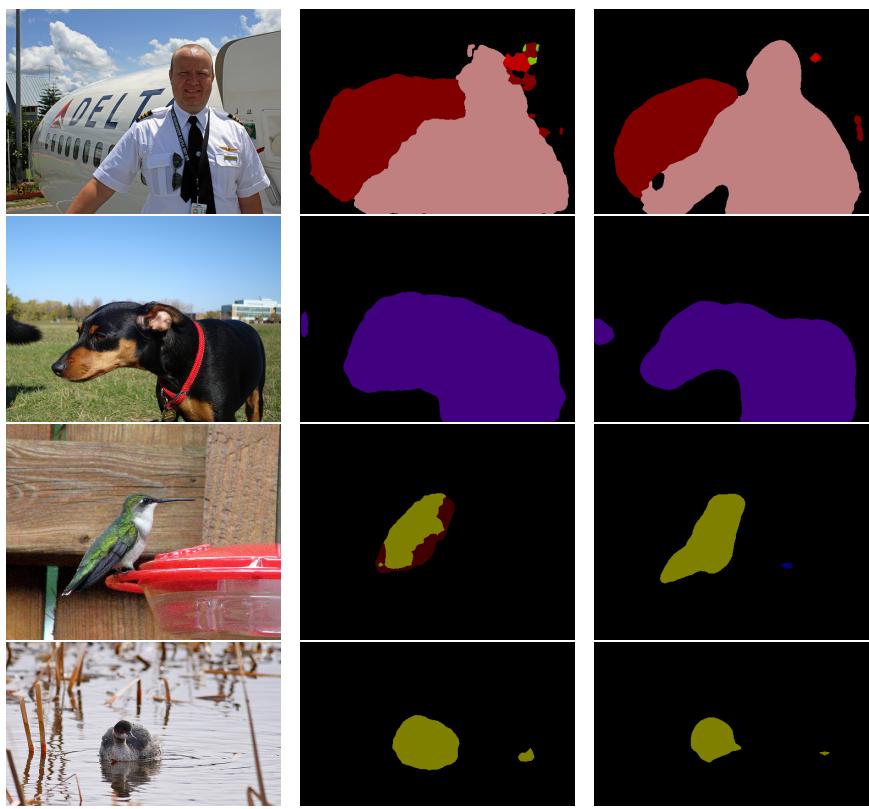
表 2: 训练循环次数

数据集	骨干模型	像素级准确率	mIoU
Pascal Voc	ResNet-50	84.618	56.120
Pascal Voc	ResNet-101	89.467	66.315
Cityscapes	ResNet-50	91.406	57.478
Cityscapes	ResNet-101	92.725	62.548

表 3: DeepLab-V3 模型在各数据集和骨干网络上的结果

5 总结

DeepLab 系列模型与常规的编-解码器框架不同点在于其采用了空洞卷积或组合的空洞卷积，并减少对池化或步长不为 1 的卷积的应用，使得能够在获得足够的感受野的情况下依然可以减少特征的尺寸，从而利用了深度卷积网络的分类准确率高的特性。此外，通过应用全连接条件随机场，将卷积网络输出的稠密分类图进行尺寸扩张和锐化边缘，使得输出的语义分割图像可以具有一方面像素分类的准确率高，另一方面对于物体的定位更加准确。

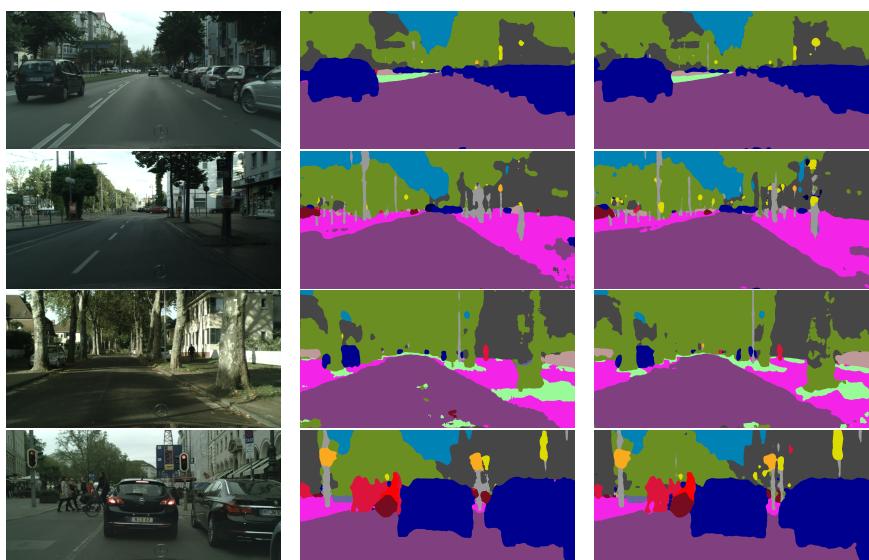


(a) 输入图像

(b) 骨干网络为 ResNet50

(c) 骨干网络为 ResNet101

图 9: DeepLab 不同的骨架网络在 Pascal Voc 数据集上用随机输入图片测试的结果



(a) 输入图像

(b) 骨干网络为 ResNet50

(c) 骨干网络为 ResNet101

图 10: DeepLab 不同的骨架网络在 Cityscapes 数据集上用随机输入图片测试的结果

参考文献

- [1] LOWE D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.
- [2] BAY H, TUYTELAARS T, GOOL L V. SURF: speeded up robust features[J]. European conference on computer vision, 2006, 3951: 404-417.
- [3] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05): vol. 1. [S.l. : s.n.], 2005: 886-893.
- [4] MULLIN A A, ROSENBLATT F. Principles Of Neurodynamics[M]. [S.l. : s.n.], 1962.
- [5] HAYKIN S S. Neural Networks and Learning Machines[M]. [S.l. : s.n.], 2010.
- [6] CRISTIANINI N, SHawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. [S.l. : s.n.], 2000.
- [7] TAN P N, STEINBACH M M, KUMAR V. Introduction to Data Mining[M]. [S.l. : s.n.], 2005.
- [8] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
- [9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of The ACM, 2017, 60(6): 84-90.
- [10] 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.
- [11] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[C]//International Conference on Learning Representations. [S.l. : s.n.], 2015.
- [12] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. ArXiv preprint arXiv:1706.05587, 2017.
- [13] EVERINGHAM M, GOOL L, WILLIAMS C K, et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.
- [14] LIN T Y, MAIRE M, BELONGIE S J, et al. Microsoft COCO: Common Objects in Context[C]//European Conference on Computer Vision. [S.l. : s.n.], 2014: 740-755.
- [15] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]//CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. [S.l. : s.n.], 2014: 580-587.
- [16] DAI J, HE K, SUN J. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation[C]//2015 IEEE International Conference on Computer Vision (ICCV). [S.l. : s.n.], 2015: 1635-1643.
- [17] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l. : s.n.], 2015: 3431-3440.
- [18] 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.
- [19] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. [S.l. : s.n.], 2015: 234-241.

- [20] COGSWELL M, LIN X, PURUSHWALKAM S, et al. Combining the Best of Graphical Models and ConvNets for Semantic Segmentation[J]. ArXiv preprint arXiv:1412.4313, 2014.
- [21] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]// ICLR 2015 : International Conference on Learning Representations 2015. [S.l. : s.n.], 2015.
- [22] EIGEN D, FERGUS R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture[C]//2015 IEEE International Conference on Computer Vision (ICCV). [S.l. : s.n.], 2015: 2650-2658.
- [23] MOSTAJABI M, YADOLLAHPOUR P, SHAKHNAROVICH G. Feedforward semantic segmentation with zoom-out features[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l. : s.n.], 2015: 3376-3385.
- [24] KRÄHENBÜHL P, KOLTUN V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials[C]//Advances in Neural Information Processing Systems 24. [S.l. : s.n.], 2011: 109-117.
- [25] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l. : s.n.], 2016: 770-778.
- [26] CHANG C C, LIN C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27.
- [27] CORDTS M, OMRAN M, RAMOS S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C]//Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l. : s.n.], 2016.