doi:10.3772/j.issn.1002-0470.2017.09－10.005

图像语义分割深度学习模型综述①

张新明②* 祝晓斌③* 蔡 强* 刘新亮* 邵 玮** 王 磊***

(*北京工商大学计算机与信息工程学院 北京 100048)

(**北京中电高科技电视发展有限公司 北京 100036)

(***广播科学研究院信息研究所 北京 100866)

摘 要 阐述了图像语义分割实质上是像素级别的密集分类的概念,及其在计算机视觉领域中的核心地位和应用意义。全面综述了图像语义分割算法的常用分类方法及最新成果,详尽比较了图像语义分割深度学习模型在 PASCAL VOC 2012 数据集上的像素准确率、平均像素正确率、平均交叠率、频率加权交叠率四个方面的实际表现性能,同时给出了模型的平均耗时、底层框架、实现语言、代码可读性、部署难度信息。最后对语义分割领域的发展进行了总结和展望,指出了模型面临的训练数据集不足、参数优化困难和结构单一的问题。

关键词 深度学习, 语义分割, PASCAL VOC, 卷积神经网络

0 引 言

图像语义分割(image semantic segmentation)是像素级别的密集分类问题,其目标是对图像中的每个像素进行语义信息标注。语义分割广泛应用于多个领域,例如,在自动驾驶领域,语义分割可为汽车智能控制程序提供行人、车辆位置和其他重要路况信息,在医学图像识别领域,语义分割模型提供的人体脏器区域信息可用于肝癌检测。因此,语义分割的研究具有广阔的应用前景,正成为当今的一大研究热点。

语义分割在整个计算机视觉任务中有着举足轻重的地位,算法的优劣直接影响后续的图像处理任务。然而,图像语义分割问题是计算机视觉领域的非常具有挑战性的难题之一,与物体检测任务类似,其主要难点来自于以下三个方面:物体层次、类别层次和背景层次[1]。这意味着语义分割任务需要在复杂多变的背景中正确标记出语义信息,并区分具有高度相似外观的不同类别物体。不仅如此,物体间还常常伴随遮挡、割裂现象,进一步增加了语义分割任务的难度。大型深度学习模型参数优化过程会消耗大量资源,为避免维度灾难,减小内存需求,神经网络常采取池化(pooling)、主成分分析(principal components analysis,PCA)等方式降低维度、减小计算量,但导致含有重要语义信息的小尺寸区域被漏检。另一方面,模型泛化能力不高是目前语义分割领域普遍存在的问题,深度学习模型很难找到一个通用的高性能算法,大部分模型只对一个或两个数据集进行优化。在模型迁移时,算法不能达到预期效果,使得模型不能大范围地广泛使用。

即便如此,深度学习技术的应用正逐渐克服语义分割带来的挑战性难题,并加快了商业化应用的发展,激烈的商业竞争对实验性模型到商业化应用的时间提出了更高的要求。然而,在本文作者知识范围内,尚未有文献给出模型在工程应用维度的性

① 国家自然科学基金(61402023),北京市自然科学基金(4162019)和食品安全知识图谱及大数据平台研制(Z161100001616004)资助项目。

② 男,1992 年生,硕士生;研究方向:媒体大数据;E-mail: zhangxm@st.btbu.edu.cn

③ 通信作者,E-mail: brucezhucas@gmail.com

(收稿日期:2017-05-18)

能信息,如单张图片耗时、模型部署难度、代码易读性等。这些信息对学术研究和应用有着重要的参考意义。本文根据 PASCAL VOCChallenge 排行榜,实际测试了采用深度学习技术的模型在图像语义分割领域的主流数据集 PASCALVOC 2012[2] 上的性能。在评测其主流参数的基础上,额外分析了模型的平均耗时、底层框架、实现语言、代码可读性、部署难度信息,为学术研究及应用提供重要参考价值。

本文对图像语义分割深度学习模型进行了综述。

1 相关工作

图像语义分割方法分为传统图像语义分割方法和基于神经网络的图像语义分割方法。传统图像语义分割方法不使用神经网络及深度学习技术,具体可分为以大量的专业知识为基础的显式特征方法、基于概率图模型的方法、不含语义信息的无监督学习方法。值得注意的是,图像语义分割方法的界定一直以来都十分模糊[3]。基于显式特征及无监督学习的传统方法自身无法建立起像素与语义的映射关系,后期需要人工参与分割过程才能真正完成语义分割过程,所以严格意义上属于非语义分割(non-semantic segmentation)方法。深度学习技术在计算机视觉领域的应用,使包括图像语义分割在内的多个领域取得了长足进步。相比于传统分割方法,基于神经网络的语义分割方法隐式地建立了像素到语义的映射,不需要后期人工的参与依然能完成整个分割过程。基于这一点,本文将应用神经网络及深度学习技术的图像语义分割方法归为第二类语义分割方法,本小节将对传统语义分割方法和基于深度学习技术的语义分割方法进行简单梳理总结。

1.1 传统图像语义分割方法

传统语义分割方法十分依赖特征的选择,特征选择的质量影响算法性能。无监督学习方法虽然不需要显示地指定特征,但学习规则的制定依然需要专业知识,分割过程不含语义信息。这些传统方法运行快速但分割性能普遍不高。

(1) 基于概率图模型的分割方法

随机决策森林(random decision forests)是一种多决策树组合分类器,最早由文献[4]提出。分类的结果有多个决策树共同完成。马尔科夫随机场(Markov random fields,MRFs)是计算机视觉中应用普遍的模型,属于无向图的概率图模型,其主要思想是为每个特征和像素分配一个随机向量,通过计算每个像素属于每个类的概率确定该像素的分类。条件随机场(conditional random fields,CRFs)在 MRFs 的基础上,对随机向量加入观察值(observations),从本质上讲,CRFs 是给定了观察值集合的 MRFs。文献[5]使用 CRFs 在 PASCAL VOC 2010[6] 进行了语义分割任务。

(2) 无监督学习方法

文献[7]使用聚类算法对像素进行无监督分类,语义分割领域最常用的是 k-means 与 mean-shift 算法,如 k-means 用在医学图片的分割上[8]。基于图形的图像分割(graph based image segmentation)算法通常将像素解释为顶点,边缘权重与差异性相关,如颜色差异。文献[9]使用图形的图像分割方法在 Pascal VOC 2010[6] 挑战上获得第二名的成绩。随机行走(random walks)方法是基于图形的图像分割方法的一种,属于交互式分割方法。随机行走方法通过一定的方法(可以是人工放置或其他方法以实现全自动过程)放置种子点代表不同的类别,其余像素的类别为最高概率到达种子点的类。主动轮廓模型(active contour model)是基于图像物体边缘的算法并尽量找到最平滑的边界,该算法可以直接进行图像分割,也可以对图像分割的结果进行微调。

1.2 基于深度学习的图像语义分割方法

深度学习模型需要进行训练才能达到预期效果,图 1 展示了深度学习模型训练与测试的处理流程。实线表示训练过程,原始图像与标签不断地送

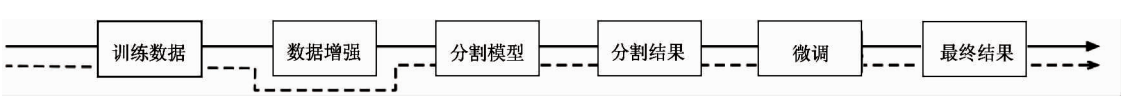

图 1 深度学习模型处理流程

入模型以进行参数优化,在图像输入模型之前,通常会使用数据增强技术对图像进行处理,如反转、遮挡、变形等[10]。模型输出预测标签后,可选择性使用概率图模型算法微调以得到更好的结果。虚线表示测试或应用过程,与训练过程不同的地方在于,测试过程不使用数据增强技术。一般地,由于概率图模型运算复杂度较高,耗时较长,通常作为一种可选方案。

针对数据集标注的粒度不同发展出了全监督学习方法与弱监督学习方法。在全监督学习方法中,使用的数据集为像素级别的标注图像,而弱监督学习方法中使用的数据集为图像级别或物体级别的标注。由于含有完善的标注信息,全监督学习领域的研究发展较快,效果较好。

(1) 全监督学习

2015 年,用于语义分割的全卷积网络(fully convolutional networks,FCN)由文献[11]提出。FCN 改编自一个预训练(pre-trained)的分类网络,并将全连接层转置为卷积层,学习像素到像素的映射。在 PASVAL VOC2012 上取得了平均交叠率(mean intersection over union, Mean IU)62.2% 的表现,领先之前的同时检则和分割算法[12](simultaneous detection and segmentation,SDC) 51.6%,并且运算时间大幅降低。图像语义分割领域进入了快速发展期, DeconvNet[13]、DeepLab[14]、CRFasRNN[15]、DialatConvNet[16]、ResNet-38[17]、PSPNet[18]等模型相继被提出,模型特点描述在第 2 节给出。PSPNet 在 PASVAL VOC2012 上取得了平均交叠率 85.4% 的性能表现,曾跃居 PASCAL VOC 2012 排行榜第一名。

(2) 弱监督学习

深度学习技术的优势在于其高度复合的函数几乎可以对任意两个空间进行映射,人工标注的数据集加上强大的计算能力,使得这种映射可以部分地被找到。但全监督学习方式需要大量精细标注的图像数据集,人工成本与时间成本较高,而弱监督学习方法的优势在于其使用粗标注图像完成训练分割过程,同时难度更高。文献[19]提出了基于物体边框(bounding boxes)的训练方法,并结合 FCN 进行网络训练。该方法在 PASCAL VOC2012 的验证集上得到了平均交叠率 62.0% 的性能表现,相比于基准线 63.8% 取得了相对不错的效果。类似地,文献[20]同样使用物体边框作为标注信息的数据集进行训练并结合 DeepLab-CRF[21]方法,最终在 PASCAL VOC2012 验证集上获得 70% 的性能表现。图像级别标注的数据集比边框级别标注的数据集更容易获得,但也带来了巨大的挑战。近年来,基于图像标签(Label)的语义分割方法倍受关注。文献[22]通过从 Web 存贮库中检索与目标标签相关的视频并生成分割标签,克服训练模型过度关注图像的畸变部分(discriminative parts)并完成训练过程,最终模型在 PASCAL VOC2012 验证集上平均交叠率的表现为 58.1%,测试集上的表现为 58.7%。文献[23]同样在图像标注级别的语义分割领域取得了出色性能。

2 深度学习模型

本小节描述了本文参数评测涉及到的深度神经网络模型,相对独立地对基于深度学习技术的模型进行分析。

(1) 语义分割 FCN

语义分割全卷积网络(FCN)的核心观点是建立“全卷积”网络,输入任意尺寸,经过有效的推理和学习产生相应尺寸的输出,学习像素到像素的映射,见图 2。FCN 是语义分割深度学习模型的开山之作,它确定了一种用于图像语义分割通用模型框架:改编基本分类网络(如 AlexNet[24], VGG-Net[25], GoogLeNet[26]提取粗特征,再进行分类预测生成分割图,最后再对生成的分割图像进行优化输出。SegNet[27], DeconvNet[13], DeepLab[14] 等网络结构,其整体模型框架均与 FCN 保持一致。FCN 提出了像素正确率(pixel accuracy),平均像素正确率(mean pixel accuracy),平均交叠率(mean IU),频率加权交叠率(frequency-weighted IU)4 项参数计算方法,用于衡量模型的性能表现。优化输出的方案主要有:条件随机场(CRF)、全连接条件随机场(DenseCRF[28])、马尔科夫随机场(MRF)、高斯条件随机场(G-CRF[29])。

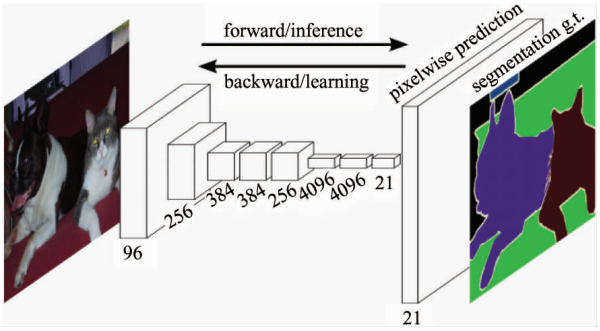

图 2　FCN 整体结构[11]

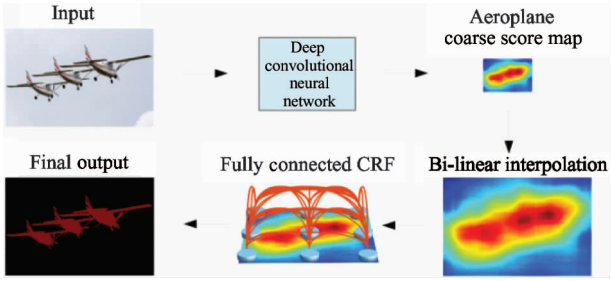

图 3　Deeplab 结构图

(2) DeconvNet

文献[13]首次在语义分割领域提出反卷积网络(deconvolution network, DeconvNet)。DeconvNet 是一个多层网络,由反卷积层(deconvolution layer)、上采样层(unpooling layer)、整流线性单元层(rectified linear unit)构成。给定一幅图像,DeconvNet 首先产生足够多的候选分割图,最后将这些候选分割图合并成一个语义分割分类图像。为了得到更优秀的表现,DeconvNet 还可以结合 FCN-8s 的预测结果。此外,与基于 FCN 的方法相比,DeconvNet 克服了物体尺度带来的高识别错误问题。

(3) DeepLab

DeepLab 是一个相对成熟的结构,文献[14]用深度学习处理语义图像分割的任务,不同的是卷积过程使用了 Atrous 卷积[30],Atrous 卷积可以明确地控制在深卷积神经网络中计算特征响应的分辨率,还使网络能够有效地扩大滤波器的视野,从而在不增加参数数量或计算量的情况下整合更大的上下文。DeepLab(见图 3)基于一个深度卷积神经网络,如 VGG-16 或者 ResNet-101[31]。双线性插值阶段将特征图放大到原始图像分辨率,最后应用全连接的 CRF 优化分割结果,得到更好的边缘[14]。DeepLab 将深度卷积神经网络(deep convolutional neural networks, DCNNs)层的响应与完全连接的条件随机场(CRF)相结合来克服定位精度问题。DeepLab 解决了 3 个对 DCNNs 具有挑战性的问题:越来越小的特征分辨率、不同尺度带来的特征重合问题、空间特征丢失导致定位正确率下降问题。

(4) CRFasRNN

利用深度学习技术的图像识别技术来处理像素级标签任务的核心问题是深度学习技术限制视觉对象的能力有限。为了解决这个问题,文献[15]引入了一种新的卷积神经网络形式,结合了卷积神经网络(CNN)和条件随机场(CRF)的概率图形建模的优势。系统将 CRF 建模与 CNN 完全集成,可以通过常规的反向传播算法对整个深度网络进行端到端的训练,避免了后处理方法的对象划分。端到端(end-to-end)神经网络可以得到从原始输入到最终的输出,缩减人工预处理和后续处理,使得模型可以根据数据自我优化参数,提高模型的整体契合度。

(5) DialatConvNet

语义分割之类的密集预测问题在结构上与图像分类不同。文献[16]提出一种专门用于密集预测的新卷积网络模块,模块使用扩张卷积(dilated convolution)来系统地聚合多尺度的上下文信息而不损失分辨率。扩张卷积使得特征图像素的感受野随着网络层数的增加成指数扩展。

(6) ResNet-38

越来越深的神经网络的趋势已经被普遍的观察所推动,即增加的深度增加了网络的性能。但简单地增加深度可能不是提高绩效的最佳方法。对深层次的网络的调查也表明,它们实际上可能不是单一的深度网络,而是作为许多相对较浅的网络的一个特征。ResNet-38[17]是一种新的、低成本的残留网络架构:ResNet-38 Multi-scale,在 Image-Net 分类数据集上显著超过更深层次的模型,如 ResNet-200[32]。ResNet-38 网络的架构超越了其比较器,包括非常深的 ResNets,而且更有效地利用内存。

（7）PSPNet

场景解析对于无限制的开放式和多样化的场景来说非常具有挑战性。文献[18]，通过空间池模块以及空间场景解析网络，利用不同区域的上下文聚合全局上下文信息的能力。PSPNet 为像素级预测任务提供了优越的框架。CNN 阶段采用的是经过修改的 ResNet-101[31]，并整合扩张卷积，上采样阶段使用双线性插值。为减小维度和维持全局特征的权重，空间池模块（pyramid pooling module）中使用 1x1 的卷积核。改进后的 PSPNet 曾跃居 PASCAL VOC 2012 排行榜第一名。

3　实　验

3.1　测试数据

本文中所有模型均在 PASCAL VOC2012 的验证集上进行测试。PASCAL VOC2012 全称 The PASCAL Visual Object Classes（VOC）Challenge。PASCAL VOC 挑战是视觉对象类别识别和检测的基准，为视觉和机器学习社区提供了标准的图像和注释数据集，以及标准评估程序。从 2005 年到现在的每年组织、挑战及其相关数据集已被接受为对象检测的基准。

VOC2012 共 21 个类别（见图 4）、11530 幅图片，其中分割图 6929 幅，用于图像的分割。

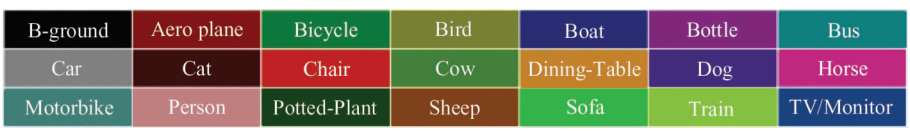

图 4　VOC2012 分类及色板

3.2　软硬件环境

采用深度神经网络的模型在训练和测试时往往对机器性能要求较高，尤其是在训练时会进行大量计算并消耗大量内存和显存，但受限于价格与实验环境，平台也在追求一个平衡。深度学习框架如 Caffe、Tensorflow、Theano、Torch、Keares 对各个平台的支持性和安装难度也不一样。本文根据经验，按照主流配置搭建基础学习实验环境，其基础配置见表 1，重要软件配置见表 2。截至实验结束，所有软件均为最新版本，降低部署难度充分发挥硬件性能，以便能给出更加客观的结果。参考互联网文章，即可搭建起基本深度学习平台，之后针对各个模型的情况，适当安装相应软件包即可快速运行代码。

表 1　基础系统平台配置

	系统	CPU	内存	硬盘	显卡
明细	Ubuntu16.04.2 至强 2603V		ECC-8GB	500GB	TaitanX

表 2　重要软件配置

	GPU-Driver	CUDA	OPENCV	Matlab	Python
明细	378	8.0.44	3.2	2016b	2.7.12

3.3　实验结果

像素准确率指的是分类正确的像素正确率（pixel accuracy），当测试集出现分类失衡时，像素正确率并不能客观地反应模型性能。论文[11]提出了平均像素正确率，平均交叠率，频率加权交叠率三种评价标准，其中平均交叠率（Mean IU，有的论文中也写作 Mean IOU）是最重要的性能评价指标，它比平均像素准确率更能反应模型的准确程度。使 n_{ij} 表示分类为 i 的像素分类为 j 的个数，n_{cl} 表示不同种类的个数，$t_i = \sum_j n_{ij}$ 表示类别 i 的像素数目，各个参数的计算方式如下：

- 像素正确率：$\sum_i n_{ii} / \sum_i t_i$
- 平均像素正确率：$(1/n_{cl}) \sum_i n_{ii}/t_i$

- 平均交叠率:$(1/n_{cl})\sum_i n_{ii}/(t_i + \sum_j n_{ji} - n_{ii})$
- 频率加权交叠率:

$$(\sum_k t_k)^{-1}\sum_i t_i n_{ii}/(t_i + \sum_i n_{ji} - n_{ii})$$

像素正确率、平均像素正确率、平均交叠率、频率加权交叠率是模型性能的主要评价指标。各个模型的性能表现见表3。除此之外,模型的执行效率关系到模型的实用程度及使用场景,表3展示了各个模型的完成一次分割平均耗时以及模型的底层框架、实现语言;代码可读性最高5分,分数越高代表越容易阅读修改;部署难度最高5分,分数越高表示部署难度越大。

表3 衡量模型性能的重要参数

	像素准确率	平均像素正确率	平均交叠率	频率加权交叠率	平均耗时(s)	底层框架	实现语言	代码可读性	部署难度
FCN	0.90	0.81	0.71	0.90	0.14	CF	PTH	5	2
PSPNet	0.95	0.89	0.89	0.92	1.25	CF	MTL	4	3
DialatConvNet	0.91	0.83	0.74	0.86	0.39	CF	PTH	4	3
CRFasRNN	0.94	0.86	0.79	0.90	4.63	CF	PTH	4	3
ResNet	0.94	0.89	0.80	0.90	0.39	MX	PTH	4	3
DeepLab	0.95	0.86	0.78	0.90	0.23	TF	PTH	3	1

实验结果表明,平均交叠率能更好地反映模型性能。全卷积网络(FCN)较高像素正确率及较低的平均交叠率表明图像背景在整个数据集中占有比率较高,在前景错误划分的情况下仍能保持高正确率,进一步说明平均交叠率更能反映模型的优劣。大部分模型可以在1s左右完成语义分割任务,经过改进并迁移到移动端的难度相对较低。CRFasRNN由于使用了概率图模型耗时较长,计算量大,不适合实时任务。最早出现的Caffe是主流深度学习框架,Python由于其易读性及易用性成为首要采用的实现语言。在代码可读性及部署难度方面,各个模型相差不大,模型总体处理逻辑基本一致,代码风格略有差异。总体来讲,在实际应用中,还需要针对具体任务,选用不同模型,并做针对性修改才能符合任务要求。

4 结 论

本文综述了图像语义分割研究背景,整理了语义分割的常用方法及分类,简要分析了基于深度学习模型的语义分割方法并实际测试了模型的实际性能。深度学习模型的应用为图像语义分割领域注入了新的活力,并将得到进一步发展。尽管深度学习模型已经取得了一定的进步,但仍面临着许多挑战,未来需要解决以下3个问题。(1)训练数据集匮乏。语义分割需要的标注需要精确到像素级别,数据集的标注十分昂贵,未来可能会通过弱监督训练得到缓解,但仍不能从根本上解决问题。(2)参数优化困难。模型参数需要经过大量运算获得,更优秀的优化算法将成为重要研究方向。(3)模型结构单一。深度学习模型结构解释性差,通过理论指导创造更有效的模型变得十分困难,结构创新发展缓慢,迁移学习成本增高。

参考文献

[1] 黄凯奇, 任伟强, 谭铁牛. 图像物体分类与检测算法综述[J]. 计算机学报, 2014, 37(6): 1225-1240

[2] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes challenge 2012 (voc2012) results[EB/OL]. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html: University of Leeds, 2011

[3] Zhang X Y. Simultaneous optimization for robust correlation estimation in partially observed social network[J]. *Neurocomputing*, 2016, 205: 455-462

[4] Zhang C, Xue Z, Zhu X, et al. Boosted random contextual semantic space based representation for visual recognition[J]. *Information Sciences*, 2016, 369: 160-170

[5] Tian D P. Semantic image annotation based on GMM and random walk model[J]. *High Technology Letters*, 2017, 23(2): 221-228

[6] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. *International Journal of Computer Vision*, 2010, 88(2): 303-338

[7] Wang M, Wang Y B, Li Y. SCMR: a semantic-based coherence micro-cluster recognition algorithm for hybrid web data stream[J]. *High Technology Letters*, 2016 (2): 224-232

[8] Chen C W, Luo J, Parker K J. Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications[J]. *IEEE Transactions on Image Processing*, 1998, 7(12): 1673-1683

[9] Carreira J, Sminchisescu C. Constrained parametric min-cuts for automatic object segmentation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2010, 34(7):3241-3248

[10] Zhang X Y, Wang S, Yun X. Bidirectional active learning: a two-way exploration into unlabeled and labeled data set[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(12): 3034-3044

[11] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2014, 39(4):640

[12] Hariharan B, Arbeláez P, Girshick R, et al. Simultaneous Detection and Segmentation[C]. In: Proceeding of the 13th European Conference on Computer Vision, Zürich, Switzerland, 2014. 297-312

[13] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015. 1520-1528

[14] 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 & Machine Intelligence*, 2017, PP(99):1-1

[15] Arnab A, Jayasumana S, Zheng S, et al. Higher order conditional random fields in deep neural networks[C]. In: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Netherlands, 2016. 524-540

[16] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. *arXiv preprint arXiv*, 2015: 1511.07122.

[17] Wu Z, Shen C, Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition[J]. *arXiv preprint arXiv*, 2016: 1611.10080.

[18] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[J]. *arXiv preprint arXiv*, 2016: 1612.01105.

[19] Dai J, He K, Sun J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation[C]. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015. 1635-1643

[20] Papandreou G, Chen L C, Murphy K, et al. Weakly-and semi-supervised learning of a DCNN for semantic image segmentation[J]. *arXiv preprint arXiv*, 2015: 1502.02734.

[21] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. *arXiv preprint arXiv*, 2016:1606.00915.

[22] Hong S, Yeo D, Kwak S, et al. Weakly supervised semantic segmentation using web-crawled videos[J]. *arXiv preprint arXiv*, 2017:1701.00352.

[23] Saleh F, Akbarian M S A, Salzmann M, et al. Built-in foreground/background prior for weakly-supervised semantic segmentation[C]. In: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Netherlands, 2016. 413-432

[24] Souly N, Spampinato C, Shah M. Semi and weakly supervised semantic segmentation using generative adversarial network[J]. *arXiv preprint arXiv*, 2017:1703.09695

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. *arXiv preprint arXiv*, 2014:1409.1556

[26] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. In: Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition, Boston, USA, 2015. 1-9

[27] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for scene segmentation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, PP(99):2481-2495

[28] Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials[C]. In: Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 2011. 109-117

[29] Vemulapalli R, Tuzel O, Liu M Y, et al. Gaussian conditional random field network for semantic segmentation [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016. 3224-3223

[30] Zhang X Y. Simultaneous optimization for robust correlation estimation in partially observed social network[J]. *Neurocomputing*, 2016, 205: 455-462

[31] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016. 770-778

[32] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]. In: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Netherlands, 2016. 630-645

Survey of the deep learning models for image semantic segmentation

Zhang Xinming*, Zhu Xiaobin*, Cai Qiang*, Liu Xinliang*, Shao Wei**, Wang Lei***

(*School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048)

(**CCTV High-Tech Television Development Limited Company, Beijing 100036)

(***Institute of Information Technology, Academy of Broadcasting Science, Beijing 100866)

Abstract

The concept that image semantic segmentation is essentially the dense classification in pixel level, as well as its core position and practical significance are interpreted. Then, the commonly used classifications and latest achievements in image semantic segmentation are comprehensively reviewed, and for some deep learning models for image semantic segmentation, the pixel accuracy, average pixel accuracy, mean intersection over union and frequency-weighted intersection over union on the PASCAL VOC 2012 dataset are compared in detail. Meanwhile, the models' other performance indexes including the average time-consuming, program framework, language used, code readability, difficulty of deployment are shown. Finally, the developments of image semantic segmentation are summarized and discussed, and the challeges facing the models such as lack of training data sets, difficulty of parameter optimization and single structure are pointed out.

Key words: deep learning, semantic segmentation, PASCAL VOC, convolutional neural networks