文章编号：2095-2163(2020)08-0001-04 中图分类号：TP391.4 文献标志码：A

# 基于学习的弱监督和半监督图像语义分割算法

汪 磊，左旺孟
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

**摘 要：**图像语义分割致力于将图像中的内容识别出来，即识别出图像中每个位置像素的类别。基于全卷积神经网络的语义分割方法取得了良好的进展，然而这种方法需要大量的且极其耗时的像素级别的标注，为了解决这个问题，基于弱监督和半监督的研究逐渐受到关注。在目前的弱监督和半监督算法中，大部分使用基于手工设计的算法来生成图像区域建议，没有充分利用图像的边界框标注信息。针对这个问题，本文提出了基于学习的弱监督和半监督图像语义分割算法。在全卷积分割网络基础上，利用边界框标注信息，学习出一个通用的图像二元分割模型，再生成图像区域建议，可以更好的利用图像的全局信息和边界框的位置信息。在基准数据集 Pascal VOC 上的实验结果证明，本文的算法性能已经超过目前的优秀的算法。

**关键词：**弱监督；卷积神经网络；图像语义分割

# Learning-based weakly- and semi-supervised for image semantic segmentation

WANG Lei, ZUO Wangmeng
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

**【Abstract】** Image semantic segmentation is dedicated to recognizing the content of the image, and will recognize the category of the pixel at each position in the image. The semantic segmentation method based on Fully Convolutional Networks (FCN) has made great progress. However, this method requires a large number and extremely time-consuming pixel-level labels. In order to solve this problem, research based on weak supervision and semi-supervision has gradually attracted attention. How to make full use of bounding box labels is a very difficult problem. In the current weakly supervised and semi-supervised algorithms, most of them use handcrafted algorithms to generate image region proposals, which is relatively blunt and rough, and does not make full use of the image´s bounding box annotations. In response to the above problems, this paper proposes an image semantic segmentation algorithm based on weakly and semi-supervised learning. Based on the fully convolutional segmentation network, the bounding box is used to annotate information to learn a general image binary segmentation model, and then generate image region suggestions. This learning-based algorithm generates image region suggestions, which can make better use of the global information of the image and the position information of the bounding box. The experimental results on the benchmark dataset Pascal VOC mIoU 67.6 prove that the performance of the algorithm in this paper has surpassed the current best state-of-the-art algorithm.

**【Key words】** weakly-supervised; CNN; semantic segmentation

## 0 引 言

深度学习技术需要大量的带标签的标注数据来训练模型。在图像分类任务中，需要对每张图像进行类别标签标注；在目标检测任务中，需要对每张图像中的物体进行边界框(bounding box)标注；在图像语义分割任务中，需要对图像中的每个像素打上类别标签。大量的数据标注带来了严重的人力、物力损失，特别是在图像语义分割任务中，需要对每个像素进行标注，耗费大量的时间，一幅大尺寸图像，作出精细化的标注需要几天甚至几个月。利用统计学统计，标注像素级别的分割数据需要的时间是标注图像中物体的边界框需要时间的 15 倍，是标注图像类别标签需要时间的 60 倍。因此，本文提出弱监督学习(Weakly-supervised Learning)的图像语义分割算法，只需要对图像类别标签或边界框的数据标注，标注耗时远小于像素级别的标注。

目前，基于弱监督学习的图像语义分割逐渐受到关注。弱监督学习的语义分割算法主要分为使用图像类别标签和使用边界框标签两种算法。其中使用图像类别标签的弱监督语义分割算法主要基于学习类激活图(Class Activation Map，CAM)，图像中不同类别的物体在 CAM 中有不同程度的响应结果，以此得到图像的区域建议标注(Region Proposals)；使用边界框标签的弱监督语义分割算法主要基于边界框作为真实标注(Ground Truth)，结合一些传统的阈值分割算法，以此得到区域建议标注。在得到图像的区域建议标注后，基于全卷积神经网络(Fully Convolutional Networks，FCN)设计出模型。然而，这些

作者简介：汪 磊(1995)，男，硕士研究生，主要研究方向：计算机视觉、图像语义分割；左旺孟(1977)，男，博士，教授，博士生导师，主要研究方向：图像处理、模式识别、计算机视觉等。

收稿日期：2020-06-19

算法未充分利用网上已有的公开数据,有一定缺陷。

本文算法使用边界框标签来进行弱监督学习的图像语义分割模型研究。利用网上公开的大量的像素级别的标注数据,学习出一个基本的二元分割模型,在基准数据集上生成图像的区域建议标注。借助这种元学习(Learning to Learn)的思想,本文学习的二元分割模型生成的区域建议标注更拟合 GT。在使用网上的像素级别标注数据中,本文并没有使用基准数据集上相应类别的数据,因此还是属于弱监督学习框架。实验结果表明,本文的学习框架有非常不错的表现。

## 1 相关工作

全监督相关研究 在图像语义分割任务中,大多数算法都是基于 Long 等人提出的 FCN 框架,FCN 简洁有效的框架为后续研究提供了一个基准。与之相似地,Liang-Chieh Chen 等人提出了 DeeLab-V1 算法,使用条件随机场(Conditional Random Fields, CRF)作为后处理,提升了在 benchmark 的性能。在医学细胞分割任务上,Ronneberger 等人提出了 U-Net,改进了特征提取的方式,基于编码器-解码器(Encoder - Decoder)架构和跨层链接(skip - connection)方式。Liang-Chieh Chen 在 DeepLab-V2 中,使用了多尺度(multi-scale)输入,提出空洞空间金字塔池化(Atrous Spatial Pyramid Pooling, ASPP),进一步提升性能,随后又在 DeepLab-V3 中重新设计了空洞卷积(Dilated Convolution)。在 Zhao 等人提出的金字塔场景解析网络(Pyramid Scene Parsing Network,PSPNet)中,使用金字塔池化更好的融合了全局信息(global priors)。在 FCN 框架中,基于残差模块(Residual Module)的不断改进,调整网络结构为监督学习的图像语义分割带来了巨大的提升。

基于边界框标签的弱监督相关研究 基于边界框标注的弱监督学习方法中,由于边界框具有较强的物体位置信息和类别信息,因此涌现的方法性能相较于前面所说的只基于类别标签的方法好很多。在目前的主流算法中,核心都是借助于边界框的信息和一些传统的非深度学习的分割算法生成物体的 Region Proposals。Dai Jifeng 等人提出的 BoxSup[1],使用多尺度连接(Multiscale Combinatorial Grouping, MCG)结合边界框生成 Region Proposals,对于 MCG 生成的 2k 个候选 Region Proposals,选取与边界框最相近的迭代训练模型,相较于 GrabCut[2] 生成的提升较大。George Papandreou 等人提出的 WSSL[3] 算法,使用边界框结合 CRF 生成 Region Proposals,由于 CRF 利用了图像底层的空间位置和颜色等信息,因此生成的 Region Proposals 在空间连续性上保持得更好,而且 WSSL 在生成 Region proposals,利用了边界框中心部分是概率意义上更大可能为分割结果这一先验。

在 WSSL 的基础上,Anna Khoreva 等人提出的 SDI[4] 框架中,尝试了 GrabCut 和 MCG 两种方式来生成 Region Proposals,并将其与边界框里不相近的部分作为忽略区域,而且 SDI 改进了 GrabCut 和 MCG 生成的方式,利用整体最近边缘检测(Holistically-nested Edge Detection,HED)边界,进一步优化成 GrabCut+和 MCG+算法,HED 算法充分利用图像的 RGB 颜色信息,更有利于生成 Region Proposals。Chunfeng Song 等人提出的 BCM-FR[5] 框架中,为了更好地利用边界框的全局掩码(global mask)信息,提出了边界框驱动的不同类掩码(Box-driven Class-wise Mask),提供了每一类边界框的线索(hints),而且考虑了不同类物体在边界框面积占比不同这一先验知识,提出了填充比(Filling Rate)损失,达到了目前的最好结果。

在上面所介绍的研究中,基于边界框监督(box supervision)的算法,如 Box - Sup、WSSL、SDI 和 BCM-FR 均有一个大问题,它们利用生成 Region Proposals 当作 FCN 框架中的真实标签(Ground Truth),假设 Region Proposals 与 Ground Truth 不一致,训练出的 FCN 无法解决这个差距(gap)。虽然在 BCM-FR 发现了这个问题并作出相应的改进,但是依然是不够的。为了解决这个问题,本文提出了基于学习(Learning-based)的弱监督和半监督学习图像语义分割和实例分割框架。

## 2 算法详述

本文基于元学习方法提出了弱监督学习的分割模型,使用额外的数据集 MS COCO 学习出二元分割模型,用这个二元分割模型生成 benchmark 数据集 Pascal VOC 的 Region Proposals,最后训练出基于 FCN 框架的语义分割模型。本文弱监督图像语义分割模型总体上分为 3 个部分:

(1)基于 FCN 的全监督语义分割模型。本文基于 DeepLab-V1 的 LargeFOV 作为 backbone,其为 VGG16 网络的改进版。在 MS COCO 数据集上,去除了 benchmark 数据集 Pascal VOC 数据集上含有 20 个类别的物体,得到剩余 60 类物体的图像。在这 60 类物体的图像上,使用二元分割模型生成

pixel－level 的 Region Proposals，使用这些 Region Proposals 训练出基于 FCN 的语义分割模型。

（2）二元分割模型。二元分割模型的输入为 MS COCO 数据集上的 60 类物体的图像，每一类物体对应的边界框掩码（bounding box mask）和 FCN 网络输出 feature maps 的相应类别的归一化后的热度图（score map），经过二元 FCN 的分割模型，预测出一个二分类的前景背景的二元分割结果。二元分割模型需要使用 pixel-level 的语义标注，由于本文是弱监督学习的图像语义分割研究，因此没有使用 benchmark 数据集 Pascal VOC 的 20 类物体的标注，使用的是 MS COCO 数据集上去除这 20 类的剩余 60 类物体的语义标注。

（3）生成图像区域建议（Region Proposals）。训练好二元分割模型后，既需要在 MS COCO 数据集上生成 60 类物体图像的 pixel－level 的 Region Proposals，也需要生成在 benchmark 数据集 Pascal VOC 上的 pixel-level 的 Region Proposals。本文设计了具体的生成图像区域建议的算法。

弱监督学习模型框架如图 1 所示。本文弱监督模型的整体流程为：对于输入的图像 $I$，经过基于 FCN 的分割网络，得到 softmax 归一化后的 score maps，取出第 $m$ 类对应的 score map $s_m$ 和第 $m$ 类物体的 bounding box mask $B_m$，再和图像 I 连接起来（concat 方式），输入到基于 FCN 的二元分割网络中，预测出第 $m$ 类物体的前景图 $F_m$，和 Ground Truth $Y^{gt}$ 计算二元分割损失 Seg2 Loss，反向传播计算梯度，更新二元分割网络 SegNet2；更新二元分割网络后，输入 $S_m$、$B_m$ 和 $I$，生成其图像区域建议 Region Proposals $Y^{gt}$，再与 score map $S$ 计算分割损失 Seg1 Loss，反向传播计算梯度，更新分割网络 SegNet1。如此迭代，训练分割网络 SegNet2、训练 SegNet1、训练 SegNet2、训练 SegNet1……。

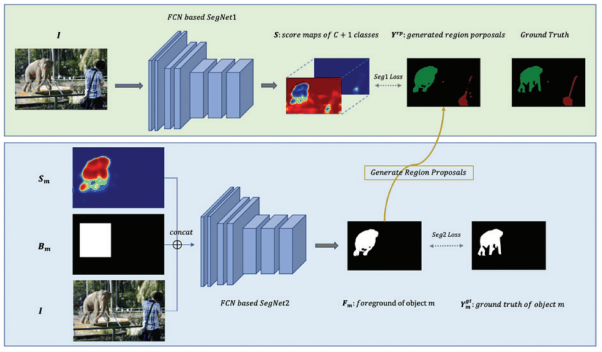


**图 1　弱监督学习模型框架图**

**Fig. 1　The framework of weakly supervised semantic segmentation**

全监督模型的损失函数为 pixel-level 的交叉熵损失函数，Seg1 Loss 为式（1）：

$$l_{seg1} = -\frac{1}{W_d * H_d}\sum_{i=1}^{W_d}\sum_{j=1}^{W_d} Y_{i,j}^{rp} log\ (S_{i,j}(I;\theta)). \quad (1)$$

二元分割模型的损失函数也是 pixel-level 的交叉熵损失函数，Seg2 Loss 为式（2）：

$$l_{seg2} = -\frac{1}{W_d * H_d}\sum_{i=1}^{W_d}\sum_{j=1}^{W_d} Y_{i,j}^{gt} log\ (F_{mi,j}(S_m, B_m, I;\theta)). \quad (2)$$

## 3　实　验

### 3.1　实验参数

对于训练全监督分割网络 SegNet1，在实际训练过程中，本文基于 DeeLab－V1 的 LargeFOV 的 VGG16 网络。SegNet1 的输入大小为 321×321×3，$W_d = H_d = 321$，下采样 8 倍，$W = H = 41$。在 MS COCO 上，$C = 60$，而在 Pascal VOC 上 $C = 20$。实验训练参数 batch 取 30，学习率取 0.001，优化器取带动量（Momentum）的随机梯度下降（Stochastic Gradient Descent，SGD），共训练 2 万次迭代。SegNet1 模型在 MS COCO 和 Pascal VOC 数据集上训练和推理。对于二元分割模型 SegNet2，采用了 HourglassNet 作为 backbone，输入为 $S_m$、$B_m$ 和 $I$ 连接，大小为 512×512×5，$W = H = 512$，下采样 4 倍，$W_d = H_d = 128$。实验训练参数 batch 取 32，学习率取 0.000 1，优化器为自适应的动量估计（Adaptive Moment Estimation，Adam）SGD，每一轮共训练 6 万次迭代。SegNet2 模型只在 MS COCO 构造的二元分割数据集上训练和推理，而在 Pascal VOC 数据集上只推理。在第一轮 Round1 时，SegNet2 输入大小为 512×512×4。本文使用了 PyTorch 深度学习框架来训练和推理模型，显卡数量为 1 张，型号为英伟达 RTX 2080 Ti。

### 3.2　实验结果

本文对比了四篇弱监督学习的语义分割论文提出的方法 BoxSup、WSSL、SDI 和 BCM-FR，在评价指标上，本文基于元学习的弱监督模型达到了 68.3%，为目前最高。在弱监督学习图像语义分割模型中，在 Pascal VOC 2012 数据集上，本文只使用了边界框标注信息，在二元分割模型上生成图像 Region Proposals，训练的模型在 VOC 验证集上，mIoU 达到 68.3%。在全监督模型下，使用 pixel－level 的分割标注训练模型，在 VOC 验证集上 mIoU 为 69.6%，对比结果，见表1。本文的弱监督模型结果已经非常

（下转第 9 页）