《Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision》复现

欧振华

摘要

在本文中，作者通过研究标记数据和额外的无标记数据来研究半监督语义分割问题。他们提出了一种新的一致性正则化方法，称为交叉伪监督 (CPS)。对于相同的输入图像，他们的方法将一致性强加于两个初始化不同的分割网络上。由一个扰动分割网络输出的伪 one-hot 标签图被用来监督另一个具有标准交叉熵损失的分割网络，反之亦然。CPS 一致性有两个作用: 鼓励两个扰动网络对同一输入图像的预测高度相似，并通过使用带有伪标签的未标记数据扩展训练数据。实验结果表明，该方法在 Cityscape 和 PASCAL VOC 2012 上都达到了最先进的半监督分割性能。

关键词：半监督语义分割；一致性正则化

1 引言

图像语义分割是计算机视觉中的一项基本识别任务。语义分割训练数据需要像素级的人工标记，与其他视觉任务如图像分类和目标检测相比，成本要高得多。这使得半监督分割成为通过使用标记数据以及附加的未标记数据来学习分割模型的一个重要问题。

一致性正则化在半监督语义分割中得到了广泛的研究。它在多种扰动下加强了预测的一致性，例如通过增加输入图像进行输入扰动[11]、特征扰乱[27]和网络扰乱[18]。自我训练也被用于半监督分割[6]。它在未标记的图像上加入了从标记图像上训练的分割模型获得的伪分割图，以扩大训练数据，并重新训练分割模型。

作者提出了一种新颖而简单的网络扰动一致性正则化方法，称为交叉伪监督。所提出的方法是将已标记和未标记的图像输入到两个结构相同但初始化不同的分割网络。这两个网络在标记数据上的输出分别由相应的 ground-truth 分割图进行监督。作者的重点在于交叉伪监督，它可以强制执行两个分割网络之间的一致性。每个输入图像的分割网络估计一个分割结果，称为伪分割图。伪分割图被用作监督其他分割网络的额外信号。

交叉伪监管方案的好处有两个方面。一方面，与之前的一致性正则化方法一样，该方法鼓励不同初始化网络对同一输入图像的预测是一致的，并且预测决策边界位于低密度区域。另一方面，在后期的优化阶段，伪分割变得更加稳定，比只在标记数据上进行正常监督训练的结果更准确。伪标记数据表现为对训练数据的扩展，从而提高了分割网络的训练质量。

在 Cityscape 和 PASCAL VOC 2012 两个基准上的不同设置下的实验结果表明，所提出的交叉伪监督方法优于现有的半监督分割一致性方案。作者的方法在两个基准上都实现了最先进的半监督分割性能。

2 相关工作

2.1 语义分割

现代深度学习的语义分割方法大多基于全卷积网络 (FCN)[23]。后续研究主要从分辨率、上下文和边缘三个方面对模型进行了研究。分辨率扩大方面的工作包括调解分类网络中造成的空间损失，例如使用编码器-解码器方案[5]或扩展卷积[4]，以及保持高分辨率，如 HRNet[3]。

利用上下文的工作包括空间上下文，例如，PSPNet[14]和 ASPP[4]，对象上下文[2]，和自我注意力的应用[9]。提高边缘区域的分割质量包括 Gated-SCNN[13]，PointRend[20]，和 SegFix[24]。本文主要研究了如何利用无标签数据，主要使用 DeepLabv3+ 进行实验，并在 HRNet 上报告了实验结果。

2.2 半监督语义分割

语义分割的手动像素注释是非常耗费时间和成本的。探索可用的未标记的图像来帮助学习分割模型是很有价值的。

一致性正则化被广泛研究用于半监督性分割。它在各种扰动下加强了预测/中间特征的一致性。输入扰动方法[19]对输入图像进行随机增强，并对增强后图像的预测之间施加一致性约束，使决策函数位于低密度区域。

特征扰动提出了一种使用多个解码器的特征扰动方案,实现了解码器[27]输出之间的一致性。GCT[17]方法通过使用两个结构相同但初始化不同的分割网络进一步进行网络扰动，并加强扰动网络预测之间的一致性。作者的方法与 GCT 不同，通过使用伪分割图来强制执行一致性，还有一个额外的好处，如扩大训练数据。

除了对一个图像强制执行各种扰动之间的一致性外，基于 GAN 的方法[25]还强制执行标记数据的 ground-truth 分割图和未标记数据的预测分割图的统计特征之间的一致性。统计特征是从一个判别器网络中提取的，该网络被学习用来区分 ground-truth 的分割和预测的分割。

自训练，又称自我学习、自我标记或决策导向学习，最初是为了在分类中使用未标记的数据而开发的[15,10,1,22,3]。最近它被应用于半监督分割[6,9,13,25,24,14]。它结合了以前在标记数据上训练的分割模型获得的未标记数据上的伪分割图，用于重新训练分割模型。这个过程可以重复几次。关于如何决定伪分割图，介绍了多种方案。例如，基于 gan 的方法[13,25]使用学习到的鉴别器来区分预测和地面真理分割，在未标记的图像上选择高置信度的分割预测作为伪分割。

PseudoSeg，与作者的工作同时进行，也探索了半监督性分割的伪分割。与作者的方法至少有两个不同之处。PseudoSeg 遵循 FixMatch 方案，通过使用弱增强图像的伪分割来监督基于单个分割网络的强增强图像的分割。作者的方法采用两个相同的、独立初始化的分割网络，输入相同的图像，并使用每个网络的伪分割图来监督其他网络。另一方面，作者的方法对两个分割网络都进行了反向传播，而 PseudoSeg 只对强增强的图像进行了反向传播。

3 本文方法

3.1 本文方法概述

给定一组由 N 个标记图像组成的 D^1 和一组由 M 个未标记图像组成的 D^u，半监督的语义分割任务旨在通过探索标记和未标记的图像来学习分割网络。交叉伪监管, 提出的方法由两个并行分割网络

组成:

$$P_1 = f(X;\theta_1) \tag{1}$$

$$P_2 = f(X;\theta_2) \tag{2}$$

这两个网络具有相同的结构，它们的权重，即 θ_1 和 θ_2，被初始化得不同。输入 X 是有相同的增量，P_1（P_2）是分割置信度图，是 softmax 归一化后的网络输出。

3.2 损失函数定义

训练目标包含两个损失：监督损失 $\mathcal{L}_s$ 和交叉伪监督损失 $\mathcal{L}_{cps}$。监督损失 $\mathcal{L}_s$ 是使用标准的像素级交叉熵损失对两个平行分割网络的标记图像进行制定的。

$$\mathcal{L}_s = \frac{1}{|\mathcal{D}^l|}\sum_{\mathbf{X}\in\mathcal{D}^l}\frac{1}{W\times H}\sum_{i=0}^{W\times H}\left(\ell_{ce}\left(\mathbf{p}_{1i},\mathbf{y}_{1i}^*\right) + \ell_{ce}\left(\mathbf{p}_{2i},\mathbf{y}_{2i}^*\right)\right) \tag{3}$$

其中 $\mathcal{L}_{ce}$ 是交叉熵损失函数，$\mathbf{y}_{1i}^*(\mathbf{y}_{2i}^*)$ 是 ground truth。W 和 H 代表输入图像的宽度和高度。交叉伪监督损失是双向的。一个是从 $f(\theta_1)$ 到 $f(\theta_2)$。我们用一个网络 $f(\theta_1)$ 输出的像素级 one-hot 标签图 Y_1 来监督另一个网络 $f(\theta_2)$ 的像素级信心图 P_2，另一个是从 $f(\theta_2)$ 到 $f(\theta_1)$。无标签数据的交叉伪监督损失写为:

$$\mathcal{L}_{cps}^u = \frac{1}{|\mathcal{D}^u|}\sum_{\mathbf{X}\in\mathcal{D}^u}\frac{1}{W\times H}\sum_{i=0}^{W\times H}\left(\ell_{ce}\left(\mathbf{p}_{1i},\mathbf{y}_{2i}\right) + \ell_{ce}\left(\mathbf{p}_{2i},\mathbf{y}_{1i}\right)\right) \tag{4}$$

我们还以同样的方式定义了标记数据的交叉伪监督损失 $\mathcal{L}_{cps}^l$。整个交叉伪监督损失是有标签和无标签数据的损失的组合:$\mathcal{L}_{cps} = \mathcal{L}_{cps}^l + \mathcal{L}_{cps}^u$。

整个训练目标被写成：

$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_{cps} \tag{5}$$

其中 λ 是权衡权重。与 CutMix 增量相结合。CutMix 增强方案被应用于半监督性分割的平均教师框架[11]。我们还在我们的方法中应用了 CutMix 增强功能。我们将 CutMixed 图像输入到两个网络 $f(\theta_1)$ 和 $f(\theta_2)$。我们使用类似于[11]的方式从两个网络中生成伪分割图：将两张源图像（用于生成 CutMix 图像）输入到每个分割网络，并将这两张伪分割图混合作为另一个分割网络的监督。

4 复现细节

4.1 实验环境搭建

本次实验运行在 Linux 上，所需要的相关软件和库有：Python3.6、CUDA 10.1、Pytorch 1.0.0、torchvision 0.9.1 等。数据集来源于 PASCAL VOC 2012[8]和 Cityscapes[7]。PASCAL VOC 2012[8]是一个标准的以物体为中心的语义分割数据集，它由超过 13000 张图像组成，有 20 个物体类别和 1 个背景类别。标准训练集、验证集和测试集分别由 1464，1449 和 1456 张图像组成。我们遵循以前的工作，使用增强集[12]（10582 张图像）作为我们的完整训练集。Cityscapes[7]主要是为城市场景理解而设计的。该官方分集有 2975 张图像用于训练，500 张用于验证，1525 张用于测试。每张图片的分辨率为 2048×1024，

并对 19 个语义类别的像素级标签进行了精细标注。

4.2 实施细节

本次实验基于 PyTorch 框架来实现。我们用在 ImageNet 上预训练的相同权重和两个分割头的权重随机地初始化两个分割网络中的两个骨干。我们采用带动量的迷你批次 SGD 来训练我们的模型与 Sync-BN[16]。动量被固定为 0.9，权重衰减被设定为 0.0005。我们采用聚能学习率策略，初始学习率乘以 $(1-\frac{iter}{max\_iter})^{0.9}$。对于在完整训练集上训练的监督基线，如果没有指定，我们使用随机水平翻转和多尺度作为数据增强。我们对 PASCAL VOC 2012 训练了 60 个 epochs，基础学习率设置为 0.01，对 Cityscapes 训练了 240 个 epochs，基础学习率设置为 0.04。在 Cityscapes 上使用了 OHEM 损失。

4.3 创新点

对于 $f(\theta_1)$ 和 $f(\theta_2)$ 这两个网络，作者使用相同的结构，但是不同的初始化。作者用 PyTorch 框架中的 kaiming_normal 进行两次随机初始化，而没有对初始化的分布做特定的约束。在论文作者没有给出相关的分割模型超参数让我完美的复现的情况下，CPS 性能效果较差，我设计了特定的初始化，CPS 性能提升了 1.46%，但是还是比原论文的效果低一点。

5 实验结果分析

(1) 分割预测的定量结果

我们在 PASCAL VOC 数据集上可视化了一些分割的预测结果。（c）列是仅使用 labeled data 进行训练的结果，（d）（e）列是我们的预测，（b）列是 ground truth 标签。可以看出，由于标注数据很少，（c）的结果不能准确识别物体的语义和边界，而改进的 CPS 可以更好地处理这些问题。

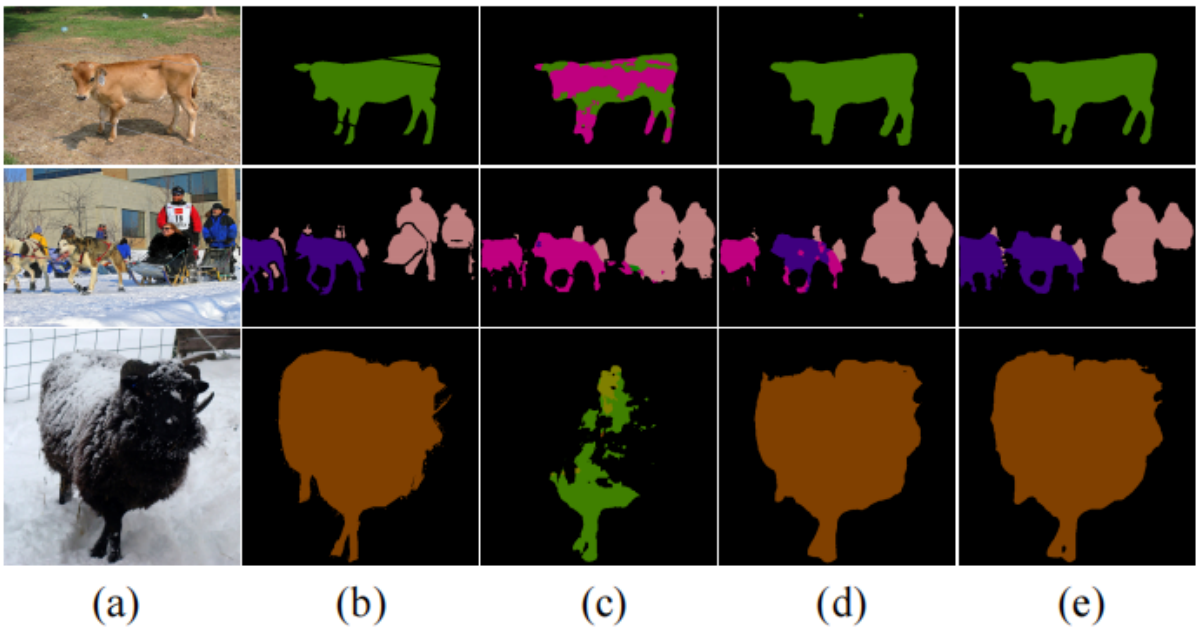

图 1: PASCAL VOC 数据集上的定量结果示例

6 总结与展望

通过课堂上的学习到的知识和参考大量网上的资料，复现该论文并对其进行了一点改进。复现论文过程使我对语义分割技术有了更加深刻的理解，尤其是对论文的模型架构进行改进的时候。从提出改进办法、搭建网络、训练网络、调整参数的实现过程中，不断的学习, 不断地提高自己的技术水平。

当然，在实现的过程中也会遇到一些困难，例如：环境搭建失败，原文复现效果较差，性能提升困难等问题。但在对科研的热爱下，所有的问题都一个一个被解决。每次解决了一个问题，心中的自豪感和满足感就会激励着我不断前进，这大概就是科研的魅力吧。

交叉伪监督方法通过使用一个网络获得的单次伪分割图来监督另一个网络，使两个具有相同结构和不同初始化的网络之间保持一致。另一方面，带有伪分割图的未标记数据在后面的训练阶段更加准确，可以作为扩大训练数据来提高性能。对于 $f(\theta_1)$ 和 $f(\theta_2)$ 这两个网络，作者使用相同的结构，但是不同的初始化。作者用 PyTorch 框架中的 kaiming_normal 进行两次随机初始化，而没有对初始化的分布做特定的约束。在论文作者没有给出相关的分割模型超参数让我完美的复现的情况下，CPS 性能效果较差，我设计了特定的初始化，CPS 性能提升了 1.46%，但是还是比原论文的效果低一点。在未来我计划进一步探索性能不如原论文的原因，尽量能超过原论文的性能。

参考文献

[1] LEE S, LEE M, LEE J, et al. Railroad is not a Train: Saliency as Pseudo-pixel Supervision for Weakly Supervised Semantic Segmentation[Z]. 2021.

[2] FAN J, ZHANG Z, TAN T, et al. CIAN: Cross-Image Affinity Net for Weakly Supervised Semantic Segmentation[C]//National Conference on Artificial Intelligence. 2020.

[3] LIU Y, WU Y H, WEN P, et al. Leveraging Instance-, Image- and Dataset-Level Information for Weakly Supervised Instance Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1415-1428. DOI: 10.1109/TPAMI.2020.3023152.

[4] SUN G, WANG W, DAI J, et al. Mining Cross-Image Semantics for Weakly Supervised Semantic Segmentation[C]//ECCV. 2020.

[5] QI X, LIU Z, SHI J, et al. Augmented Feedback in Semantic Segmentation Under Image Level Supervision[C]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision –ECCV 2016. Cham: Springer International Publishing, 2016: 90-105.

[6] DONG Z, HANWANG Z, JINHUI T, et al. Causal Intervention for Weakly Supervised Semantic Segmentation[C]//NeurIPS. 2020.

[7] KUMAR SINGH K, JAE LEE Y. Hide-and-seek: Forcing a network to be meticulous for weaklysupervised object and action localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3524-3533.

[8] LI K, WU Z, PENG K C, et al. Tell me where to look: Guided attention inference network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9215-9223.

[9] WEI Y, FENG J, LIANG X, et al. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1568-1576.

[10] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4981-4990.

[11] HUANG Z, WANG X, WANG J, et al. Weakly-supervised semantic segmentation network with deep seeded region growing[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7014-7023.

[12] WEI Y, XIAO H, SHI H, et al. Revisiting dilated convolution: A simple approach for weakly-and semisupervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7268-7277.

[13] LEE S, LEE J, LEE J, et al. Robust tumor localization with pyramid grad-cam[J]. arXiv preprint arXiv:1805.11393, 2018.

[14] HOU Q, CHENG M M, HU X, et al. Deeply supervised salient object detection with short connections [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3203-3212.

[15] WANG L, LU H, WANG Y, et al. Learning to detect salient objects with image-level supervision[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 136-145.

[16] YANG C, ZHANG L, LU H, et al. Saliency detection via graph-based manifold ranking[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3166-3173.

[17] JIANG Z, DAVIS L S. Submodular salient region detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2043-2050.

[18] LIU N, HAN J. Dhsnet: Deep hierarchical saliency network for salient object detection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 678-686.

[19] WEI Y, LIANG X, CHEN Y, et al. Stc: A simple to complex framework for weakly-supervised semantic segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11): 2314-2320

[20] WANG X, YOU S, LI X, et al. Weakly-supervised semantic segmentation by iteratively mining common object features[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1354-1362.

[21] YAO Q, GONG X. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation[J]. IEEE Access, 2020, 8: 14413-14423.

[22] FAN R, HOU Q, CHENG M M, et al. Associating inter-image salient instances for weakly supervised semantic segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 367-383.

[23] ZENG Y, ZHUGE Y, LU H, et al. Joint learning of saliency detection and weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 7223-7233.

[24] WANG L, LU H, WANG Y, et al. Learning to detect salient objects with image-level supervision[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 136-145.

[25] ZHAO T, WU X. Pyramid feature attention network for saliency detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3085-3094.

[26] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4981-4990.

[27] EVERINGHAM M, ESLAMI S, VAN GOOL L, et al. The pascal visual object classes challenge: A retrospective[J]. International journal of computer vision, 2015, 111(1): 98-136.

[28] HARIHARAN B, ARBELÁEZ P, BOURDEV L, et al. Semantic contours from inverse detectors[C]// 2011 international conference on computer vision. 2011: 991-998