Deep Metric Learning for Open World Semantic Segmentation

杜俩鹏

摘要

经典的闭集语义分割网络假设测试中的所有类都已经在训练期间参与，检测异常像素的能力有限，这可能对自动驾驶等安全关键型应用造成灾难性后果。增量学习这些带有少量注释的未知对象是扩大深度学习模型知识库的理想方法。本文所复现的论文提出了一个开放世界语义分割系统，包括两个模块:(1) 开放集语义分割模块，用于检测分布内对象和 OOD 对象。(2) 增量的少数镜头学习模块，逐渐将这些 OOD 对象合并到现有的知识库中。这个开放世界语义分割系统像人一样，能够识别 OOD 对象，并在相应的监督下逐步学习。采用带有对比聚类的深度度量学习网络 (DMLNet) 来实现开放集语义分割，无需使用额外的数据或生成模型，与其他开放集语义分割方法相比，DML-Net 在三个具有挑战性的开放集语义分割数据集上实现了最先进的性能。本次课程的论文工作通过实现具有对比聚类的深度度量学习网络来实现开放集语义分割，在此基础上，用增量少样本学习方法，通过对未知对象的注释逐步改进 DMLNet，在实际应用中更加稳健实用。

关键词：语义分割；开集识别；深度学习

1 引言

深度卷积网络在语义分割任务中取得了巨大的成功，受益于高质量的数据集。然而，封闭集语义分割的前提是测试中的所有类都已经在训练过程中涉及到，这在开放世界中是不成立的。目前，在语义分割中，我们的目标是训练分类器来为图像中的所有像素分配类别标签，其中有标签的训练图像和未标签的测试图像共享相同的标签集，但是，在开放的世界中，未标记的测试图像可能包含未知类别，并且与标记的图像具有不同的分布。如果封闭集系统错误地将分布内标签分配给 OOD 对象，可能会在安全关键型应用程序 (如自动驾驶) 中造成灾难性的后果。与此同时，静态感知系统无法根据所见内容更新知识库，因此，它仅限于特定的场景，需要在一段时间后重新训练。为了解决这些问题，本文所复现的论文提出了一种开放集的动态感知系统，称为开放世界语义分割系统[1]。它包含两个模块:(1) 开放集语义分割模块，用于检测 OOD 对象，并为分布内的对象分配正确的标签。(2) 增量的少数镜头学习模块，将未知对象逐步纳入现有知识库。在开放集语义分割系统中，训练好的分类器需要能够识别未知类像素，也要能很好地对已知类像素进行分类。

对于开集语义分割，最基本的部分是在一幅图像的所有像素中识别出 OOD 像素，称为异常分割。异常分割的典型方法是将图像级开集分类方法应用于像素级开集分类。我们提出使用 DMLNet 来解决开放世界的语义分割问题，因为 DMLNet 的分类原理基于对比聚类，对异常目标的识别非常有效，同时，DMLNet 与原型相结合，其增量学习可以通过添加新的原型来实现，非常适用于少镜头任务，这是一种自然而有用的方法，在现实应用中更加健壮和实用。本次课程的论文工作通过实现具有对比聚类的深度度量学习网络（DMLNet）来实现开放集语义分割，在此基础上，用增量少样本学习方法，通过对未知对象的注释逐步改进。

2 相关工作

此章节对基于深度度量学习的面向开放世界语义分割任务的定义以及前人工作做进一步的介绍。

2.1 语义分割

语义分割是计算机视觉领域的关键问题之一。语义分割对图像的每一个像素点进行密集的预测，对每一个像素点给予标签来实现了图像像素级的分类。图像语义分割方法主要分为传统的图像语义分割和基于深度学习的图像语义分割方法，在过去的几年里，深度学习在图像分类方面的突破迅速转移到了语义分割任务上。这个任务涉及分割和分类，首先是基于深度卷积神经网络的语义分割系统通常采用级联的自下而上图像分割，然后是基于深度卷积神经网络的区域分类。卷积神经网络用于图像分割，方法是根据其类别对区域的中心像素进行分类，在所有像素位置上通过使用快捷方式来解决简单堆叠卷积的限制浅层和深层之间的反向梯度。但为了将单独的像素映射给标签，我们需要将标准卷积神经网络编码器扩展为编码器-解码器架构。在这个架构中，编码器使用卷积层和池化层将特征图尺寸缩小，使其成为更低维的表征。解码器接收到这一表征，用通过转置卷积执行上采样而恢复空间维度，这样每一个转置卷积都能扩展特征图尺寸。最终，解码器生成一个表示原始图像标签的数组。

2.2 开放世界分类和检测

开放世界分类和检测，最近几年在计算机视觉领域备受关注，其目标是在已知类数据集下训练模型，使其既能对已知类别进行识别，也能对训练集没有的未知类别进行识别，对未知类别具有鲁棒性[2]。最近 Joseph 等提出了一种基于对比聚类、未知感知提议网络和基于能量的未知识别准则的开放世界目标检测系统。本文的开放世界语义分割系统与他们的相似，除了两个重要的差异:(1) 在他们的开放集检测模块中，他们依赖于区域提议网络 (RPN) 是类不可知的事实，因此没有标记的潜在 OOD 对象也可以被检测到。这样，OOD 样本的信息对训练是有效的。然而，我们专注于语义分割，其中在训练中使用的每个像素都被分配一个分布内标签，因此没有 OOD 样本可以添加到训练中。(2) 在增量学习模块中，他们使用了新类的所有标记数据，而我们关注的是自然难度更大的少数镜头条件。目前研究较少的是增量少镜头学习，主要包括分类增量少镜头学习、目标检测增量少镜头学习和语义分割增量少镜头学习[3]。

2.3 深度度量学习网络

深度度量学习网络已被广泛应用于视频理解和人物再识别等领域。深度度量学习网络将这些问题转化为使用欧几里得距离、马氏距离或马图西塔距离来计算度量空间中的嵌入式特征相似度。卷积原型网络和深度度量学习网络通常一起用于解决特定问题，例如检测图像级 OOD 样本和用于语义分割的少镜头学习。我们还按照这种组合构建了第一个用于开放世界语义分割的深度度量学习网络。

3 本文方法

3.1 本文方法概述

本文提出了一种开放集的动态感知系统，称为开放世界语义分割系统。该系统包括两个模块：(1) 开放集语义分割模块，用于检测分布内对象和面向对象对象。(2) 增量的少数镜头学习模块，将这些面向对象设计对象逐步纳入其现有知识库。在开放集语义分割系统中，训练好的分类器需要能够识别未知类像素，也要能很好地对已知类像素进行分类。所提出的开放世界语义分割系统的整个流程如图

1所示：

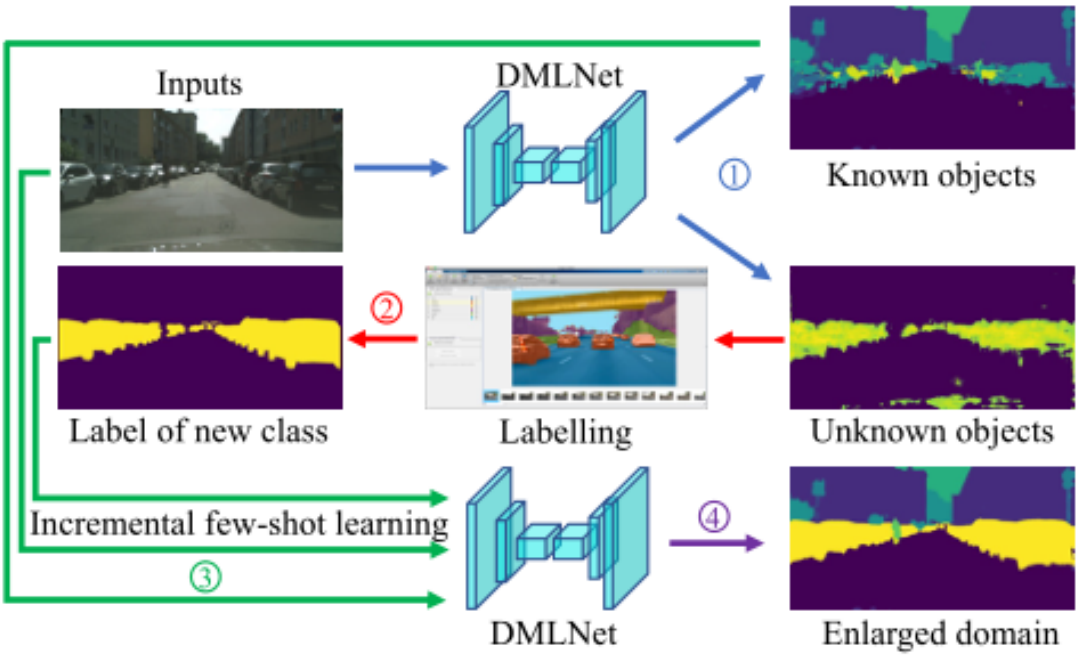

图 1: 开放世界语义分割系统

该系统流程为检测异常分割，创建潜在新对象的集群 (创建伪标签)，进行新类增量学习。具体来说，第一步是对输入的图像识别已知和未知物体 (蓝色箭头)；第二步是为未知对象添加注释 (红色箭头)；第三步是应用 incremental few-shot learning 来训练增加网络的分类范围 (绿色箭头)；第四步是用经过 incremental few-shot learning 训练后的 DMLNet 在 larger domain 中输出结果，分割更多类别 (紫色箭头)。

本次课程的论文复现工作实现具有对比聚类的深度度量学习网络来实现开放集语义分割，在此基础上，用增量少样本学习方法，通过对未知对象的注释逐步改进开放语义分割系统。将开放语义分割系统修改应用到 Cityscapes[4]数据集上进行训练并对结果分析。

3.2 开集语义分割模块

开放集语义分割模块由封闭集语义分割子模块和异常分割子模块组成。开放集语义分割模块的流程如图 2所示。蓝色虚线框内包含封闭集语义分割子模块，红色虚线框内包含异常分割子模块。开集分割是这两个子模块生成的结果的组合。在开放集分割图中预测了分布内类和 OOD 类。

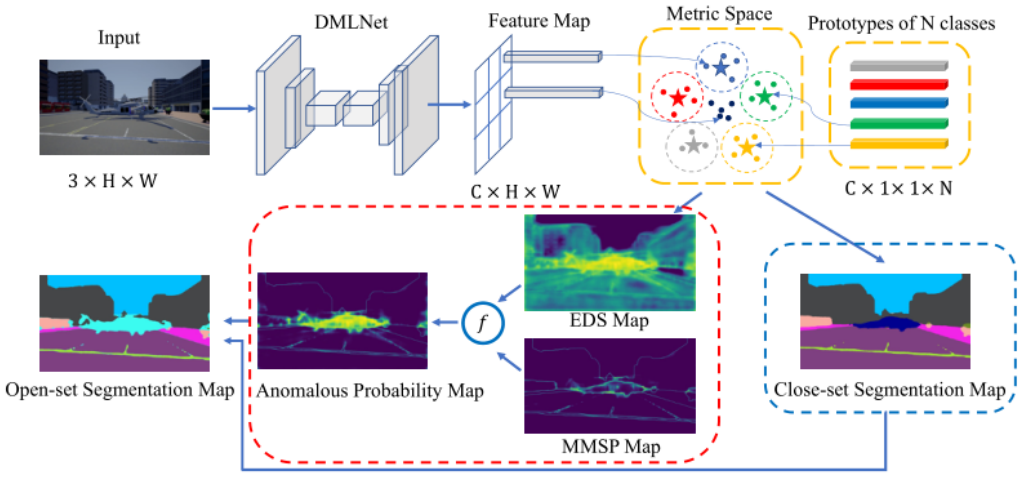

图 2: 开集语义分割模块

3.3 深度度量学习网络

经典的基于 CNN 的语义分割网络可以被分解为两部分: 特征提取器 f(X;θ_f)，得到每个像素的嵌入向量和分类器 g(f(X;θ_f);θ_g) 用于生成决策边界，其中 X、θ_f、θ_g 分别表示输入图像、特征提取器参数、分类器参数。

这种可学习的分类器不适合 OOD 检测，因为它将所有的特征空间分配给已知的类，而没有为 OOD 类留下任何空间。相比之下，在 DMLNet 中分类器被欧几里得距离表示取代，所有原型 $M_{in} = \{m_t \in R^{1\times N} | t \in \{1, 2, ...N\}\}$，$m_t$ 指的是 $C_{in,t}$ 类的原型。

特征提取器 f(X;θ_f) 学习将输入 X 映射到度量空间中与原型长度相同的特征向量。对于闭集分割任务，一个像素 $X_{i,j}$ 属于类别 $C_{in,t}$ 的概率被公式化为：

$$p_t(X_{i,j}) = \frac{exp(-||f(X;\theta_f)_{i,j} - m_t||^2)}{\sum_{t'=1}^{N} exp(-||f(X;\theta_f)_{i,j} - m_{t'}||^2)} \tag{1}$$

在这种基于欧氏距离概率的基础上，判别性交叉熵（DCE）的损失函数 $L_{DCE}(X_{i,j}, Y_{i,j}; \theta_f, M_{in})$ 被定义为：

$$L_{DCE} = \sum_{i,j} -log(\frac{exp(-||f(X;\theta_f)_{i,j} - m_{Y_{i,j}}||^2)}{\sum_{k=1}^{N} exp(-||f(X;\theta_f)_{i,j} - m_k||^2)}) \tag{2}$$

其中 Y 是输入图像 X 的标签。L_{DCE} 的分子和分母分别指的是吸引力和排斥力。我们制定了另一个损失函数，称为方差损失（VL）函数 $L_{VL}(X_{i,j}, Y_{i,j}; \theta_f, M_{in})$，其定义如下:

$$L_{VL} = \sum_{i,j} ||f(X;\theta_f)_{i,j} - m_{Y_{i,j}}||^2 \tag{3}$$

L_{VL} 只有吸引力的作用，没有排斥力的作用。对于 DCE 和 VL，混合损失定义为 $\square L = L_{DCE} + \lambda_{VL} L_{VL}$，其中 $_{VL}$ 是权重参数。

4 复现细节

4.1 与已有开源代码对比

本次论文复现过程中代码参考复现论文 DMLNet(https://github.com/Jun-CEN/Open-World-Semantic-Segmentation) 中的代码。

DMLNet 为每个类设置固定的中心嵌入，即特征空间中的独热编码 one-hot vector。虽然可以有效地在不同类之间制造距离，但是忽略了类之间的相对相似性。例如，在 Cityscapes 数据集中，该方法未能揭示人和骑手之间的差异小于他们与天空之间的差异。我们将 one-hot 设置替换为 Circleloss[5]作为度量学习的目标，这不仅保持了良好的类间距离，而且使组内分布更加集中。部分代码如图 3 所示：

```
features = x.permute(0, 2, 3, 1).contiguous()  # batch * h * w * c
features_out = x.clone()
shape = features.size()
features = features.view(shape[0], shape[1] * shape[2], shape[3])  # batch * hw * c
num_classes = self.centers.size()[0]
features_shape = features.size()
features = features.unsqueeze(2).expand(features_shape[0], features_shape[1], num_classes,
                                        features_shape[2])  # batch * hw * num_classes * c
dists = features - self.centers.cuda()  # batch * hw * num_classes * c
dist2mean = -torch.sum(dists ** 2, 3)  # batch * hw * num_classes
x = dist2mean.permute(0, 2, 1).contiguous().view(output_size[0], num_classes, output_size[2],
                                                 output_size[3])
```

图 3: Metric Space 改进

我们将开放语义分割系统修改应用到 Cityscapes 数据集上，用伪标签法对数据集进行操作，产生部分未知类标签。部分代码如图 4 所示：

```
if cls.unknown_target != None:
    cont = 0
    for h_c in cls.unknown_target:

        target[target == h_c - cont] = 255
        for c in range(h_c - cont + 1, 19):
            target[target == c] = c - 1
            target_true[target_true == c] = c - 1
        cont = cont + 1
    target[target == 255] = -1
```

图 4: 伪标签法

4.2 实验环境搭建

本次复现环境使用 NVIDIA A100 单张 GPU，python3.8，torch1.7.1 进行训练以及测试。

4.3 创新点

DMLNet 使用特征空间中的独热编码 one-hot vector，为每个类设置固定的中心嵌入。虽然可以有效地在不同类之间制造距离，但是忽略了类之间的相对相似性。例如，在 Cityscapes 数据集中，该方法未能揭示人和骑手之间的差异小于他们与天空之间的差异。

我们将 one-hot 设置替换为 Circleloss 作为度量学习的目标，这不仅保持了良好的类间距离，而且使组内分布更加集中。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

对于 StreetHazards 数据集，我们遵循了与原复现论文相同的训练程序，在 StreetHazards 训练集上训练 PSPNet[6]和 ResNet50[7]提取特征，Embedding 表示使用了 Metric Space。混合损耗的 λVL 为 0.01，所有原型的非零元素 T 都是 3。实验结果如表 1 所示：

表 1: StreetHazards 的开集分割结果

Method	AUROC	AUPR	FPR95	mIoU	Acc
resnet50+PSPNet	64.45	4.18	77.66	49.33	89.4
resnet50+PSPNet+embedding	93.75	14.68	17.12	49.76	89.66

我们将开放语义分割系统修改应用到 Cityscapes 数据集上，遵循了上述相同的训练程序，所得结果如表 2 所示：

表 2: Cityscapes 的开集分割结果

Method	AUROC	AUPR	FPR95	mIoU	Acc
resnet50+PSPNet+embedding	87.95	3.2	18.8	45.71	88.54

6 总结与展望

本次论文复现工作 DMLNet 包含两个模块: 一个开集分割模块和一个增量式少镜头学习模块。开集分割模块基于深度度量学习网络，并使用欧几里德距离来实现，在数据集 Cityscapes 上基本达到和论文中一样的精度。开放世界语义分割系统的两个模块仍然有很大的改进空间，未来将进一步研究以提高性能。

综上，本次复现任务受益颇多，不仅提升了代码能力，以动手实践的方式更加理解论文，更深的认识如何训练一个模型，更加深刻地认识到一篇顶刊论文工作量的庞大与严谨性。

参考文献

[1] CEN J, YUN P, CAI J, et al. Deep metric learning for open world semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 15333-15342.

[2] DONG H, CHEN Z, YUAN M, et al. Region-aware metric learning for open world semantic segmentation via meta-channel aggregation[J]. arXiv preprint arXiv:2205.08083, 2022.

[3] UHLEMEYER S, ROTTMANN M, GOTTSCHALK H. Towards unsupervised open world semantic segmentation[C]//Uncertainty in Artificial Intelligence. 2022: 1981-1991.

[4] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3213-3223.

[5] SUN Y, CHENG C, ZHANG Y, et al. Circle loss: A unified perspective of pair similarity optimization [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6398-6407.

[6] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[7] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.