基于全景图像分割的研究综述

方炫苏
21835076
数学科学学院
793137669@qq.com

摘要

文章来自MIT，Google和Berkeley联合出品，研究的内容是全景图像的快速高效分割和解析任务。提出了一种bottoom-up，single-shot全景图像分析方法。全景图像分析包括“stuff”形式（类别）的语义分割和“thing”形式的实例分割（区分不同的个体）。目前，经典的全景图像分析方法由语义分割任务和实例分割任务的独立模块组成，需要进行多个inference操作。相比之下，本文提出了一种相对简单的全卷积方式的图像场景分析。语义分割和实例分割任务以single-shot的方式处理，以获得具有更快处理速度的流线型模型。对于定量分析，本文使用基于实例的全景质量-PQ度量和区域推荐覆盖率分析-PC度量，可以更好地捕获“stuff”类别和更大目标实例的图像质量。基于Mapillary Vistas数据集实验，本文的单一模型基于GPU实现 31.95%（val）和 31.6%（test）PQ和 55.26%（val）PC。以 3fps或接近实时速度 22.6fps运行，但准确率会下降。

关键词：bottoom-up；single-shot；全景图像分割

Abstract

This paper proposes a bottoom-up, single-shot panoramic image analysis method. Panoramic image analysis includes semantic segmentation of the "stuff" form (category) and instance segmentation of the "thing" form (distinguishing different individuals). At present, the classic method of panoramic image analysis is composed of independent modules of semantic segmentation tasks and instance segmentation tasks, and it needs to perform multiple inference operations. In contrast, this paper proposes a scene analysis of the image in a relatively simple full-convolution manner. The semantic-segmentation and instance-segmentation tasks are processed in a single-shot manner to obtain a streamlined model with faster processing speed. For quantitative analysis, this paper uses the instance-based panoramic quality-PQ metric and the regional recommendation coverage analysis-PC metric, which can better capture the image quality of the "stuff" category and larger target instances. Based on the Mapillary Vistas dataset experiment, the single model of this paper is based on GPU to achieve 31.95% (val) and 31.6% (test) PQ and 55.26% (val) PC. Running at 3fps or near real-time speed of 22.6fps, but the accuracy will drop.

Keywords: bottoom-up; single-shot; panoramic image segmentation

1. 概述

本文致力于解决有效的进行全景分析问题，图像解析是计算机视觉任务中的一个长期未解决的问题，同时，也是现实中许多应用的组件之一，比如自动驾驶。图像解析的难点在于统一了语义分割及实例分割两个具有挑战性的任务。语义分割重点是将图像中的区域划分为具有语义信息的几个区域，其语义类别可以是可统计（“thing”）的类，也可以是不可统计("stuff")的类。与之相反的是，实例分割只是处理与“thing”类别相关，但需要区分不同的实例部分。将主题与图像分析进行结合可以将包含“stuff”类与“thing”类的整个图片进行分割，同时，不同分离不同的"thing"实例。

有许多用于解决图像分析中问题的相关工作，但大都未考虑效率问题。而对于将模型部署到实际生活中，其效率是十分重要的。图像解析由于需要经过复杂的网络进行处理，因此会增加计算量，而且，随着输入分辨率的增加，计算量还会继续增加。

本文提出一些用于高效图像解析的神经网络设计策略，显著降低高分辨率输入的内存占用情况。这些创新包括深度可分离卷积的扩展应用，使用带两层预测头的共享解码输出，增大内核大小而不是使用更深的网络，使用空间到深度和深度到空间的变换而不是上采样操作，采用困难样本挖掘策略，详细的消融研究显示了实践中这些策略的影响。

基于以上设计策略，提出了一种一次性高效，自底向上的图像解析网络，DeeperLab。在Mapillary Vistas数据集上，所提出的基本模型Xception-71 达到 31.95%的验证PQ、31.6%的测试PQ以及 55.26%的验证PC，GPU上每秒可以处理 3 帧图像；加宽版本的MobileNetV2 基础模型能够在CPU上达到接近实时的性能（22.61fps），

准确率稍有下降。故提出一种称为Parsing Covering的指标替代用于评估基于区域远景的图像解析结果。

本文设计了一种图像解析器，综合考虑了准确率及效率二者之间的关系。提出了一种single-shot,bottom-up的图像解析器-DeeperLab。如下图，DeeperLab基于single-pass的全卷积网络来产生语义及实例分割的预测mask。最后通过一个快速的算法将预测结果进行融合得到解析的结果。DeeperLab在执行时，检测物体的实例个数对其影响甚小，因此，该模型可以适用于更加复杂的场景。

对于定量评估，作者认为最近提出的基于实例的Panoptic Quality（PQ）评估,过分注重于小物体，其对"thing"类别的关注度要超过“stuff”类别。为了弥补此影响，本文提出了假定的基于区域的Parsing Covering（PC）度量，与适用于类别不匹配评估Covering评估相匹配，针对图像解析任务，本文使用了PQ及PC两个标准进行评估。推荐本文的理由如下：

(1) 本文针对图像解析器设计了几种神经网络策略，尤其是对于较高分辨率输入的情形，降低其内存的占用。创新性的做法包括：广泛的使用深度可分离卷积，共享由两层网络组成的预测端。扩大卷积核的大小而不是增加网络的深度，应用space-to-depth及depth-to-space的方法而不是上采样，同时，执行hard data mining，本文也详细介绍了融合的相关研究进而证明本模型在实际应用中的有效性。

(2)本文提出了一种高效的signel-shot,bottom-up的图像解析器-DeeperLab。

(3)本文提出了一种交替的评价标准- Parsing Covering，基于区域的角度来评判图像解析结果。

2. 相关工作

Image parsing:Imae parsing的作用是将图像分解为连续的视觉模式，像纹理及检测目标等，其涵盖了分割，检测，识别等任务。首次使用基于贝叶斯框架进行Image parsing，后来基于AND-OR图，Exemplars及条件随机场等方法进行全场景理解任务。早期这些任务的评估标准是独立的，比如，检测有检测的评估标准，分割有分割的评估标准。随着基于实例的全景质量（PQ）评估引入多个benchmarks中，全景分割越来越受到关注。

2.1. 语义分割

大多数state-of-art的分割模型在基于FCN的基础上进行一些创新性改进得到的。比如，上下文信息对像素级的标记十分重要，因此，有些工作使用图像金字塔对不同尺寸的输入图像进行编码操作。PSPNet提出了基于不同网格尺寸的图像金字塔池化结构，DeepLab提出了使用不同rate的并行的空洞卷积结构（ASPP）从而可以有效的利用上下文信息。另一个有效的方法是使用encoder-decoder结构。在encoder阶段得到图像的上下文信息，而在解码阶段对边界进行恢复。DeeperLab利用FCN,ASPP，encoder-decoder等结构来最大化image parsing的准确率。

2.2. 实例分割

当前实例分割的方法可以归类为top-down及bottom-up的方法。top-down的方法通过增强state-of-the-art检测器得到的框获得instance masks。其中，FCIS使用位置敏感性score maps。Mask R-CNN基于FPN的基础上进行搭建，在Faster R-CNN上增加了另一个分割分支，取得较好的效果。另一方面，bottom-up的方法采用两阶段的处理过程，由分割模型得到的像素级预测按照实例预测的方式进行聚合。PersonLab预测人体的关键点及进行人体实例分割，而DeNet及CornerNet通过预测边界框的角点来检测实例。

2.3. 评价标准

语义分割的结果可以通过基于区域或者轮廓的指标来进行评估。基于区域的评估标准定量评价标记正确的像素所占比例，包括：overall pixel accuracy,mean class accuracy,mean IOU。而基于轮廓的度量则关注分割边界的标记精度。比如，在分割边界较窄的三角地带评估像素级的准确率及IOU。对于类别不可知的分割可以使用covering标准来度量。实例分割可以看作是mask检测，是边界框检测的增强。因此，此类任务通常使用APr进行度量，像计算mask的IOU而不是边界框的IOU。在0.5到0.95不同重叠率阈值下计算AP的平均值进而评估分割结果。基于区域覆盖指标来评估实例分割结果，该方法适用于无法计算预测值重叠率的情形。图像解析结果可以通过Panoptic Quality（PQ）指标进行评估，同时将具有相同“stuff”类别的图像区域作为单个实例。而PQ度量存在的一个问题是，无论目标物的尺寸为多大，都视为相同的，因此，PQ度量可能会过度强调小的物体，像“thing”类别的而不是“stuff”类。

3. 模型

本文受DeepLab与PersonLab启发，提出了一种single-shot,bottom-up的有效的神经网络模型用于图像解析。网络结构如下图。

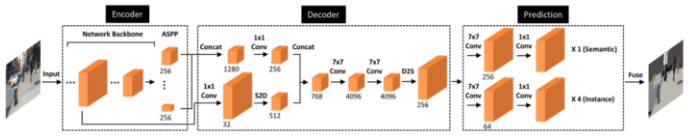

图 1 网络结构图

网络结构采用encoder-decoder的形式，为了提高效率，语义分割与实例分割共享decoder的输出,将二者的输出结果进行融合作为最终图像解析的结果。

对于图像解析任务来说，输入需要较高的分辨率（本文基于 Mapillary Vistas数据集，分辨率大小为1441x1441）会造成大量的内存占用及冗余。本文详细介绍了如何克服上述问题，进而在准确率及内存占用上得到一个最优的平衡处理。

3.1. 编码

本文基于高效的深度可分离卷积实验了两个网络结构：1）标准的Xception-71 用于获得较高的准确率，2）更宽的MobileNetV2 用于更快的推理。虽然标准的MobilenetV2 在输入大小为 224x224 的ImageNet图像分类任务中表现较好。但是对于较高输入分辨率的图像解析任务，其有限的感受野（491x491）无法捕捉大范围的上下文信息。正如Xception-71 那样，叠加更多 3x3 的卷积是增大感受野的一种方式，然而增加的额外的网络层会造成大量的内存占用。考虑到计算资源有限，将MobileNetV2 中的所有 3x3 的卷积替换为 5x5 的卷积。这种方法在不增加内存占用的条件下有效的增加了感受野的大小（981x981），计算量会稍有增加。本文称其为更宽的MobileNetV2。

本文增加了网络中的ASPP结构，encoder输出的feature map stride为 16，其空间分辨率为输入上每个分辨率以 16 的倍数进行降采样得到的分辨率。

3.2. 解码

decoder的目的是恢复目标物边界的细节信息，参考DeeplabV3+，本文采用将encoder输出的激活层的feature map(stride=16)与backbone的较低层次的feature map（stride=4）进行融合。ASPP的输出与低层feature map的通道数首先经过 1x1 的卷积进行降维处理来减少通道数。DeeplabV3+考虑到不同分辨率的将降维后的ASPP输出进行基于双线性插值进行上采样，然而上采样的操作会大大增加内存的消耗。本文对低层次的feature map采用了space-to-depth的操作，如下图，使占用的内存不会发生改变。

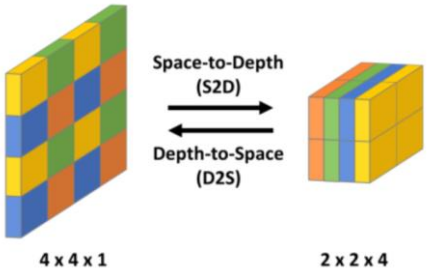

图 2 S2D和D2S算法图

与encoder相似，decoder使用大小为 7x7 的深度卷积来增加感受野的范围。通道数为 4096，然后，通过depth-to-space操作来实现上采样操作，得到一个通道数为 256，stride为 4 的feature map，作为image parsing处理的输入。

3.3. 图像解析

图像解析部分的顶部包含五大部分，每一个都单独的拼接到共享的decoder输出，同时包含两个卷积大小分别为 7x7 及 1x1 的两个卷积层。一端用于语义分割（第一个 7x7 的卷积核的通道数为 256），其余 4 个用于类别不可分的实例分割（第一个 7x7 的卷积核的通道数为64）

3.4. 语义分割

基于引导性交叉熵损失对分割进行训练，即将每个像素按照其对应的交叉熵损失进行排序，只对其前K个位置的像素进行反向传播（hard example mining），本文设置K的大小为 0.158xN，N为图像中所有像素的个数。此外，根据实例的大小，对像素的损失进行了加权重操作，从而更加关注小样本。

3.5. 实例分割

本文采用基于关键点的形式对目标实例进行表示，本文考虑边界框的四个顶点及中心作为P为 5 的目标物关键点。参照PersonLab网络，定义四个head用于实例分割。

与PersonLab相类似，定义了四个用于实例分割的四个预测heads：a keypoint heatmap , long-range, short-range及 middle range offset maps。这些heads用于预测每个像素点与对应实例关键点之间的联系。四个heads如下图。

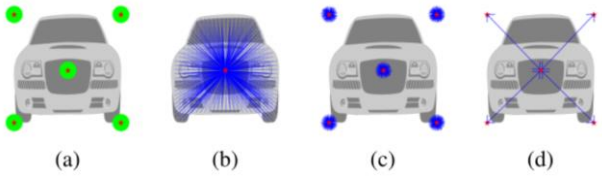

图 3 heads图

The keypoint heatmap：用于预测像素是否位于以对应关键点为中心半径为R的disks中。如果在其中，则目标激活值为 1，否则为 0.无论实例多大，统一设置R为 25。对于每个关键点，预测得到的kepoint heatmap通道数为P，基于标准的sigmoid交叉熵损失对误差进行惩罚。

The long-range offset map：用于预测一个像素相对于所有关键点的偏移，编码每个像素的长距离信息。得到的长偏移map通道数为 2P。每两个通道用于预测每个关键

点水平及垂直方向上的偏移，使用L1 损失，只激活属于目标实例的像素。

The short-range offset map：与 long-range offset map相似，只是半径变为 25，损失函数同上。

The middle-range offset map：用于预测有向关键点练习图（DKRG）关键点对之间的偏移。该map用于组合来自形同目标实例的关键点。本文采用star-graph，质心点与四角的点基于双线性连接。预测得到的map通道数为2E，其中E为DKRG中有向边的数量（E=8).损失函数同上。

3.6. 预测融合

本文首先说明如何将前面得到的四个预测maps融合为一个类别不可分的实例分割map。基于预测出的语义及实例分割maps,对于图像中的每个像素将语义及实例label都进行融合。

Instance Prediction：与PersonLab相类似，根据实例相关的四个预测maps生成实例分割map。

Recursive offset refinement:本文观察到距离关键点越近则预测的准确率就越高，因此，像PersonLab一样递归的增强偏移maps.

Keypoint localization:对于每个关键点，在short-range偏移map上使用霍夫投票，同时使用对应关键点heatmap的激活值作为投票权重用于从而生成short-range score map。同样，在long-range偏移map上使用霍夫投票，权重都为 1，生成long-range的score map。两个score maps按照权重相加进行融合。通过在融合后的score maps中寻找局部最大值从而定位关键点。最后，使用ExpectedOKS对所有关键点重新打分。

Instance detection：为了实例检测，基于贪婪算法对关键点进行聚合。首先，所有关键点被推入一个优先队列中，同时，一次只弹出一个来。如果被弹出的关键点已经在检测的实例中存在，则丢弃继续执行。否则，根据middle-range offsets 来确认保留四个关键点的位置，从而形成一个新的检测实例。新检测实例的置信分数为各个关键点分数的平均值。检测完所有实例后使用NMS去除重叠较大的实例。

Assignment of pixels to instances:最后，对检测到的实例通过使用long-range offset map进行label操作。将每个像素分配到检测到的实例中，该实例中的关键点与该像素预测到的关键点二者之间的L2 距离最短。

Semantic and Instance Prediction Fusion：本文选择简单的融合方法，考虑“stuff”类别(像天空)及“thing”类别（人）两种情况从预测的语义分割开始。对于预测的像素为“stuff”类的则标记一个独一无二的实例Label.对于其他像素，实例标签由实例分割结果确定，而语义标签，则由语义分割中投票数较多的情况决定。

4. 实验结果

正文中报告的是本文提出的Xception-71、Wider MobileNetV2 以及Light Wider MobileNetV2 三种网络在Mapillary Vistas数据集上进行的消融实验，并没有与其他算法进行对比的实验结果。所有模型参数都是端到端训练没有采用分段预训练每个组件的过程，除了主干网络使用ImageNet-1K预训练之外。

表 1 卷积核大小消融实验

Kernel Size	ASPP	PQ (%)	PC (%)
3 × 3		16.17	34.80
5 × 5		17.92	39.34
7 × 7		18.27	40.33
3 × 3	✓	19.21	41.07
5 × 5	✓	19.85	42.98
7 × 7	✓	20.14	43.40

表 2 解码器与预测头的设计消融实验

BU	S2D	DH	LK	PQ (%)	PC (%)
				19.85	42.98
✓				20.78	43.83
✓		✓		21.59	44.95
✓		✓	✓	22.31	44.62
	✓			21.12	44.86
	✓	✓		22.45	46.30
	✓	✓	✓	23.48	46.33

表 3 难例挖掘消融实验

HPM	SI	PQ (%)	PC (%)
		24.07	48.23
✓		24.64	48.34
	✓	24.27	48.42
✓	✓	24.99	49.23

表 4 关键点关联图建模方法

Star	Rectangle	PQ (%)	PC (%)
✓		24.07	48.23
	✓	23.54	46.44

表 5 主干网络深度消融实验

ASPP	BU	S2D	HPM	SI	PQ (%)	PC (%)
✓	✓				27.79	50.80
✓		✓			28.34	50.88
✓		✓	✓		30.04	53.35
		✓	✓	✓	30.46	54.55

如上表所示，文章分别从主干网络卷积核大小、解码器和预测头的设计、难例挖掘、关键点关联图建模方法、主干网络深度等五个方面进行了消融实验，数据还是非常详细且有说服力的。

最后，下面的表显示了文章所提出的DeeperLab在不同分辨率的验证集和测试集上的有效性和效率。

表 6 验证集上的实验结果

Method	Input Size	PQ (%)	PC (%)	fps (CPU)	fps (GPU)	Merge (ms)
Light Wider MNV2	721 × 721	17.59	43.43	0.77	22.61	45
Light Wider MNV2	1441 × 1441	22.36	47.52	0.18	9.37	145
Wider MNV2	1441 × 1441	25.20	49.80	0.09	6.19	145
Xception-71	1441 × 1441	31.95	55.26	0.06	3.09	145

表 7 测试集上的实验结果

Method	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
Light Wider MNV2 †	17.3	66.2	22.7	9.1	62.6	12.8	28.2	71.0	35.8
Light Wider MNV2	22.6	69.6	29.3	15.4	67.2	21.1	32.1	72.9	40.2
Wider MNV2	25.3	70.6	32.3	17.6	65.5	23.4	35.5	77.4	44.0
Xception-71	31.6	75.5	40.1	25.0	73.4	33.1	40.3	78.3	49.3

参考文献

[1] Yang T J , Collins M D , Zhu Y , et al. DeeperLab: Single-Shot Image Parser[J]. 2019.

[2] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár,“Panoptic segmentation,” arXiv:1801.00868, 2018. 1, 2, 3,5, 6

[3] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder,“The mapillary vistas dataset for semantic understanding of street scenes.,” in ICCV, 2017. 1, 2, 6

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” PAMI,2011. 2, 3, 5, 6

[5] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in CVPR, 2017. 2, 3, 6, 8

[6] H. Qi, Z. Zhang, B. Xiao, H. Hu, B. Cheng, Y. Wei, and J. Dai, “Deformable convolutional networks – coco detection and segmentation challenge 2017 entry,” ICCV COCO Challenge Workshop, 2017. 2, 3, 6, 8

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in ECCV, 2018.2, 3, 6, 8

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv:1704.04861, 2017. 2, 3

[9] S.-C. Zhu and D. Mumford, “A stochastic grammar of images,”Foundations and Trends in Computer Graphics and Vision, 2007. 2

[10] L. Zhu, Y. Chen, Y. Lin, C. Lin, and A. Yuille, “Recursive segmentation and recognition templates for image parsing,” TPAMI. 2

[11] T. Malisiewicz and A. A. Efros, “Recognition by association via learning per-exemplar distances,” in CVPR, 2008.2

[12] J. Tighe and S. Lazebnik, “Finding things: Image parsing with regions and per-exemplar detectors,” in CVPR, 2013.2

[13] P. Isola and C. Liu, “Scene collaging: Analysis and synthesis of natural images with semantic layers,” in ICCV, 2013.2

[14] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, “Objects in context,” in ICCV, 2007. 2

[15] S. Gould, T. Gao, and D. Koller, “Region-based segmentation and object detection,” in NIPS, 2009. 2

[16] G. Heitz, S. Gould, A. Saxena, and D. Koller, “Cascaded classification models: Combining models for holistic scene understanding,” in NIPS, 2009. 2