基于 Mask R-CNN 上不同犬种的分类分割性能对比分析

罗又天 李丹
四川大学锦城学院 四川 成都 611731

【摘 要】不同图片数据的分类在现实生活中是一项很常见任务，早在卷积神经网络就得以实现，并且随着R-CNN(Region CNN) 出现，简单的分类任务开始向目标检测任务上发展，在之后便是各种分割任务。而 Mask R-CNN[3] 在 Faster R-CNN[4] 之后，在目标检测的同时实现了实例分割。该算法模型需要在图片面板中辨别不同的目标物体，并学习大量的特性来表示每个目标物体的细节。而狗狗数据图片中，狗的种类特征之间 (比如毛色，纹理，体型) 存在相异性和相似性。所以在本文中，笔者选取并自己制作了两个不同狗类数据集，分别代表特征具有较大相异性和相似性的目标。用这两个狗类数据集来进行 Mask R-CNN 目标检测性能与实例分割性能的对比分析。

【关键词】Mask R-CNN；分类；分割；对比

引言

目标检测任务的开端就是分类任务。普通的卷积神经网络 CNN 在特征提取上具备较大的优势，所以最初用以图片分类。RCNN 引入了区域的概念，将区域与 CNN 相结合，成功得以应用到目标检测问题上。而实例分割需要正确的找出一幅图中的所有不同类别的目标，并将这些目标所在一一分割出来。Mask R-CNN 基于 faster R-CNN，引入特征金字塔网络 [2] 来提取候选区域 (Region Proposal) 以及构建一条平行于预测 (类别判断和边框偏移) 分支并行的 Mask 分支，它在图像中检测目标所在并通过 Mask 分支预测一个 class-aware Mask。该方法能够更精确的提取出目标，在实例分割中取得了很大的成就。

为了分析 Mask R-CNN 在不同数据集上的分类与分割性能上的表现。本文中，笔者选取了两个不同品种狗的数据集，数据部分来源于斯坦福高校的 Dogs Dataset[1]，笔者在这些类别中分别选取了两不相似品种 (体型，毛色均存在较大差异) 和两相似品种 (体型、毛色相似性较高) 的数据并进行整合，并分别扩充数据集至 200 张左右，该两数据集分别代表特征具有较大差异的图片集和特征差异较小的图片集。基于早先提出的 Mask R-CNN，笔者使用该两数据集进行训练，用所得出的结果对 Mask R-CNN 的在二者上的分类分割性能做出比较分析。

进行实验比较后发现，Mask R-CNN 在处理相似特征目标的时候，在分类分割性能上具有较强的鲁棒性。

1 相关工作

1.1 算法环境搭建

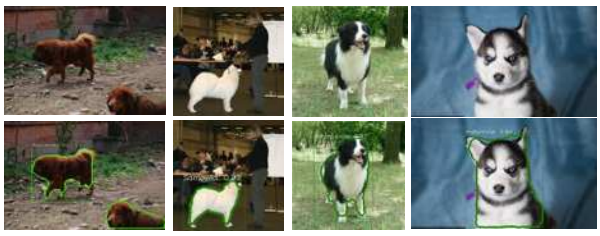

图 1 狗狗分类分割结果

硬件环境：介于 Mask R-CNN 有着相当的计算量，在原文中，对 coco 数据集的训练使用了 8 个 Nvidia Tesla M40 GPU，对硬件设施的要求较高，所以笔者将 Mask R-CNN 部署在一个 GeForceRTX2080 显卡的服务器上，以满足算法基本算力的要求。

软件环境：基于 Ubuntu 系统采用 pytorch 框架实现算法代码，并安装对应 cuda 的 pytorch 版本以及相应的依赖。此外，因制作的数据集采用 coco 数据集的格式，还需安装相应的 coco 处理模块。

1.2 数据集的制作

在数据集的制作上，部分图片数据选取斯坦福大学的 Dogs Dataset，分别扩充数据集至 200 张左右进行图片数据量的增加。使用 labelme 标注软件，人工对图片进行标注，每个数据集分两类进行标注，种类特征相异性大的数据集中目标分别标注为 Tibetan Mastiff (藏獒) 类和 Samoyed(萨摩耶) 类，后文中简称 TM&S，种类特征相似的数据集中目标分别标注为 Border collie(边境牧羊

犬) 类和 malamute(哈士奇) 类，后文简称 BC&M，每个张图片标注后生成一一对用的 json 文件，将 json 文件整合合并。训练集和测试集的比例为 5：1。最终，图片以 coco 数据集的形式保存。

1.3 数据的对比分析

为了比较 Mask R-CNN 在两数据集上的分类与分割性能，在本文中，分别比较 Bounding Box（目标所在区域，同一般目标检测）的平均精确率和 segment（实例分割）的平均精确率。本文的所涉及的种类数为 3（待分类的两个类别以及背景），平均准确率是指所有待分类种类的准确率均值。在训练过程中，每经过一定的迭代次数，记录一次二者的平均精确率，观察二者平均精确率随时间的变化。

比较并不局限于同一主干网络的选择下，还可以选取不同主干网络以及调整各参数后对结果进行比较分析。

2 Mask R-CNN

Mask R-CNN 基于 Faster R-CNN 的结构之上，沿用其思想。该算法平行于 Faster R-CNN 的预测（类判别和边框偏移）模块，增加了一个 Mask 分支，该 Mask 分支由几个卷积层构成，可以预测目标所在的精确位置 (目标掩码 Mask)。所以，基于这个新增的分支结构，Mask R-CNN 能在进行传统目标检测的同时进行实例分割。

2.1 Faster R-CNN

由于 Mask R-CNN 沿用了 Faster R-CNN 的思维，采用了两阶段的构造，所以 Faster R-CNN 中的一些构造需要被探讨。

RPN(region proposal network)：模型的第一阶段就是用 RPN 提取 ROI。RPN 是在 faster RCNN 中提出的，目标是提升候选框的提取速度。RPN 的思想就是在 feature map 上通过一定规则选择出若干候选区 anchors(不同尺度和宽度)，再将这些 anchors 分别传给分类分支和 anchor box 回归分支，最后再根据 anchor 的分类概率和设定的 IoU 阈值筛选出符合标准的 anchor 作为 RoI。

预测分支：该分支解耦成了 2 个小分支，一个分支用于为分类，另一个分支用于检测框的回归修正。这两个小分支都是通过全连接层来完成对相应功能的实现。

2.2 ROI ALIGEN

早先的 RoIPool 方法一般选择的是 max pooling 的方法，并且在候选区域的选择以及划分区域时，为了便于操作，进行了两次取整，这种方法在下采样中是粗糙的，会对候选框的位置产生一些细小的偏差，对于一般的分类任务，检测框的判定往往对这些细小的偏差具有很强的鲁棒性，但是分割任务中，对目标区域的划分往往是像素级别的，这个细小的偏差可能对于分类没有影响，但是对于 Mask 分支的分割任务是十分负面的。在 Mask RCNN 中提出了 RoIAlign 来替换 RoIPool 操作，RoIAlign 在小数级的像素点上下采样，该方法在降维的同时不会对区域的判定结果产生偏差，满足了分割的要求。

2.3 Mask 分支

Mask 分支是 Mask R-CNN 的不同之处，在延用 Faster R-CNN 的结构的基础上，在 RoIAlign 操作后平行于原预测分支新构建一条 Mask 分支，能够进行实例分割的任务。该分支中采用全卷积网络 FCN[5] 取代了全连接层，对于输入该分支 RoI，全连接层一般是用于整合在前面结构中所提取的目标特征，所以全连接层能根据所有的目标特征进行目标整体判别，因此常用于进行分类。而卷积的作用在于提取目标的特征，越深层，其就越具有代表性，所以用全卷积的形式取代全连接层，这就是卷积化，所以全卷积能处理分割任务。在全卷积的最后，为了得到目标掩码所在，还需要进行反卷积到原始图片大小的操作。

2.4 FPN

Mask 分支中还采用了特征金字塔网络的结构 [2]。在原文所进行的实验中发现，使用 FPN 的 ResNet 主干架构比一般的 ResNet 主干架构在精度和速度上都有很大的提升。

在整个网络中，相异层次的 feature map 具有不同强弱的语义信息，浅层 feature map 分辨率较高，但是语义信息弱，深层 feature map 则具有相反的特性。识别不同大小的目标是实例分割中的一个基本任务，但深层粗分辨率的 feature maps 不能提供更多的细节，还需要浅层高分辨率的细节部分。因此，FPN 的作用就是融合多层上的语义信息，用于不同大小目标的检测。

3 实验

在本节中，将基于 Mask R-CNN，对两特征相似性差距较大的数据集进行训练，比较 Mask R-CNN 在两数据集上分类和分割任务的性能。

3.1 网络选择与评估标准

笔者在一台 GPU 为 GeForceRTX2080 的服务器上对 Mask R-CNN 进行了算法部署。在网络的选取上，两个数据集均采用深度为 50 的 ResNet 网络作为主干进行训练，该网络中，引入了 FPN 用作特征提取，命名为 ResNet-50-FPN。此外，笔者还尝试了使用 ResNet-50-C4 网络作为主干，表示在 ResNet 网络中，在第四阶段的最后一个卷积层上进行特征提取。对于实验结果，

采用 coco 标准衡量指标进行评估，包括 AP、AP50、AP75，它们分别表示综合平均精确率，IoU 阈值为 0.5 的平均精确率，IoU 阈值为 0.75 的平均精确率。

3.2 实验过程

Tibetan Mastiff 和 Samoyed 数据

该数据集对应目标特征差异较大的数据。对于 ResNet-50-FPN 网络，训练中批处理的大小设置为 2 个图像 / 每 GPU，测试中批处理的大小同样设置为 2 个图像 / 每 GPU。在进行图样标注之后，每张图样的高为 600 个像素 , 宽为 800 个像素，所以在数据增强上，笔者将训练图片输入的最小边长设为 [800,1200], 最大边长设为 1333，这样短边会在 [800,1200] 进行随机缩放，而长边会按照一定比例进行缩放，但最大不会超过 1333。测试集的设置也是如此。迭代次数为 20k 次，每经过 5k 次迭代输出一次准确率，初始学习率为 0.00025。

结果表明，使用 ResNet-50-FPN 网络训练 20k 耗时 1.25h，bounding box 在 AP 上表现为 68.34，AP50 为 99.72，AP75 为 79.45。segm 在 AP 上表现为 79.76，AP50 为 97.80，AP75 为 93.14。如表 1 所示。

对于 ResNet-50-C4 网络使用相同参数训练 20k 耗时 1.80h，但效果很不理想，重新调整学习率至 0.001，训练 20k 次耗时 1.5h，bounding box 在 AP 上表现为 55.00，AP50 为 89.02，AP75 为 66.32。segm 在 AP 上表现为 63.30，AP50 为 89.80，AP75 为 76.25。如表 2 所示。

表 1 TM&S 数据集上使用 ResNet-50-FPN 网络训练结果

ResNet-50-FPN	AP	AP50	AP75
bounding box	68.34	99.72	79.45
segm	79.76	97.80	93.14

表 2 TM&S 数据集上使用 ResNet-50-C4 网络训练结果

ResNet-50-C4	AP	AP50	AP75
bounding box	55.00	89.02	66.32
segm	63.30	89.80	76.25

Border collie 和 malamute 数据

该数据集对应特征差异不大的数据。因为该数据集的图片大小占比和前者差不多，为了起到对比的作用，在参数的选择上采用相同的参数进行训练。

结果表明使用 ResNet-50-FPN 网络训练 20k 耗时 1.36h，bounding box 在 AP 上表现为 62.11，AP50 为 95.38，AP75 为 72.58。segm 在 AP 上表现为 71.96，AP50 为 95.18，AP75 为 86.94。如表 3 所示。

表 3 BC&M 数据集上使用 ResNet-50-FPN 网络训练结果

ResNet-50-FPN	AP	AP50	AP75
bounding box	62.11	95.38	72.58
segm	71.96	95.18	86.94

3.3 增加迭代次数

增加迭代次数常常可以在性能上得到显著提升。将两数据集的迭代次数增加至 30k 次，并在迭代至 20k 时，学习率衰减到 10%(0.000025)。通过对迭代次数的提升，在 corgi 和 Samoyed 数据上，bounding box 的 AP 表现为 73.24，AP50 为 100.00，AP75 为 92.72。segm 在 AP 上表现为 86.43，AP50 为 100.00，AP75 为 98.02。如表 4 所示。

在 Border collie 和 malamute 数据上，bounding box 的 AP 表现为 70.00，AP50 为 99.88，AP75 为 83.27。segm 在 AP 上表现为 82.57，AP50 为 99.88，AP75 为 97.97。如表 5 所示。

表 4 TM&S 数据集上训练迭代次数提高至 30k 次

TM&S	AP	AP50	AP75
bounding box	73.24	100.00	92.72
segm	86.43	100.00	98.02

表 5 BC&M 数据集上训练迭代次数提高至 30k 次

BC&M	AP	AP50	AP75
bounding box	70.00	99.88	83.27
segm	82.57	99.88	97.97

3.4 数据对比分析

通过对训练测试结果的 AP 值对比发现，TM&S 数据上使用 FPN 的 ResNet-50-FPN 网络对比一般的 ResNet-50-C4 网络，训练 20k 次，在 bounding box 上，AP 相差 13.34 个百分点，在 segm 上 AP 相差 16.46 个百分点，时间上相差 0.25h，所以可见无论是在速度还是精度上使用 FPN 都有显著的优势，所以未对 BC&M 数据集使用 ResNet-50-C4 网络进行训练。在使用 ResNet-50-FPN 网络所得结果中，两数据集所得的 coco 标准指标均远大于原文在 coco 数据集上训练测试所得的指标，这可能是数据总量偏小以及分类数量较少的原因。

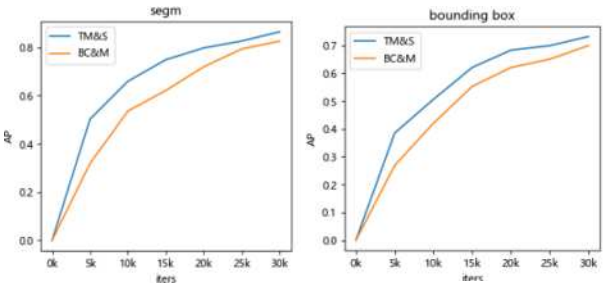

图 2 两数据集 segm 和 bbox 的平均精确率变化曲线

对于 30k 次迭代，两数据集 bbox 和 segm 的平均精确率的变化曲线如图 2 所示。对比两数据集的训练数据结果，TM&S 数据集相较于 BC&M 数据集，训练次数达到 20k 时，在 bounding box 上，AP 相差 6.23 个百分点，

在 segm 上 AP 相差 7.80 个百分点，而当训练次数达到 30k 时，在 bounding box 上，AP 相差 13.21 个百分点，在 segm 上 AP 相差 3.86 个百分点。由分类和分割的平均准确率 AP 的差异可见，Mask R-CNN 对于较大特征差异目标的分类分割性能上的提升较快，而对于较小特征差异目标的分类分割性能上的提升较慢。该结果也在情理之中，对于特征相差较大的目标，Mask R-CNN 能够很快地收敛，所以在较少训练次数的情况下就能有较高的准确率，而当特征差异较小的时候，收敛速度就会有一定的下降，在达到相同准确率的情况下就需要更多的迭代次数。

所以总体上随着迭代次数的增加，精确率的差异在逐渐减小。由此可见 Mask R-CNN 在处理相似特征目标的时候，同样在分类分割性能上具有较强的鲁棒性。

总结

Mask R-CNN 在不同数据集上，即使是目标特征属性相似时，也具有较强的鲁棒性。Mask R-CNN 因其独特的思路与结构，在实例分割上取得巨大的成功，就 Mask 分支而言，他所做到的实例分割性能就已经超过了当时的所有的模型在实例分割上的性能。Mask R-CNN 在当时 (2017 年) 的实例分割任务上取得了领先地位，原文中也表明代码开源，希望 Mask R-CNN 能成为之后更先进算法在相关任务上的一个框架。而事实表明，Mask R-CNN 的提出也确实为后续实例分割，乃至实例级人体分割 [6] 等任务上做出了巨大的贡献。

【参考文献】

[1]Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao and Li Fei-Fei. Novel dataset for Fine-Grained Image Categorization. First Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[2]T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In CVPR, 2017.

[3]K. He, G. Gkioxari, P. Doll´ar, and R. Girshick. Mask r-cnn. In ICCV, 2017.

[4]S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[5]J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. InCVPR,2015.

[6] Yang L , Song Q , Wang Z , et al. Parsing R-CNN for Instance-Level Human Analysis[J]. 2018.