Mask R-CNN

Abstract: We present a conceptually simple, flexible, and general framework for object instance segmentation. Our approach efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The method, called Mask R-CNN, extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN, running at 5 fps. Moreover, Mask R-CNN is easy to generalize to other tasks, e.g., allowing us to estimate human poses in the same framework. We show top results in all three tracks of the COCO suite of challenges, including instance segmentation, boundingbox object detection, and person keypoint detection. Without bells and whistles, Mask R-CNN outperforms all existing, single-model entries on every task, including the COCO 2016 challenge winners. We hope our simple and effective approach will serve as a solid baseline and help ease future research in instance-level recognition.

摘　要：提出了一个概念上简单，灵活和通用的目标分割框架。方法有效地检测图像中的目标，同时为每个实例生成高质量的分割掩码。称为 Mask R-CNN 的方法通过添加一个与现有目标检测框回归并行的，用于预测目标掩码的分支来扩展 Faster R-CNN。Mask R-CNN 训练简单，相对于 Faster R-CNN，只需增加一个较小的开销，运行速度可达 5 FPS。此外，Mask R-CNN 很容易推广到其他任务，例如，允许在同一个框架中估计人的姿势。在 COCO 挑战的所有三个项目中取得了最佳成绩，包括目标分割，目标检测和人体关键点检测。在没有使用额外技巧的情况下，Mask R-CNN 优于所有现有的单一模型，包括 COCO 2016 挑战优胜者。这种简单且有效的方法将成为一个促进未来目标级识别领域研究的坚实基础。

一、精读论文

1.1 网络基本框架

下图 Mask R-CNN 的框架。众说周知，Mask R-CNN 是对 faster r-cnn 的扩展，与 bbox 识别并行的增加一个预测每一个 ROI 的分割 mask 的分支。mask 分支是应用到每一个 ROI 上的一个小的 FCN（Fully Convolutional Network），以 pix2pix 的方式预测分割 mask。目标掩码与已有的 class 和 box 输出的不同在于它需要对目标的空间布局有一个更精细的提取。因为原有的 Faster R-CNN 架构缺少精细的像素对齐，故此 Mask R-CNN 进行了一定程度的弥补。Mask RCNN 比 Faster RCNN 速度慢一些，达到了 5fps。

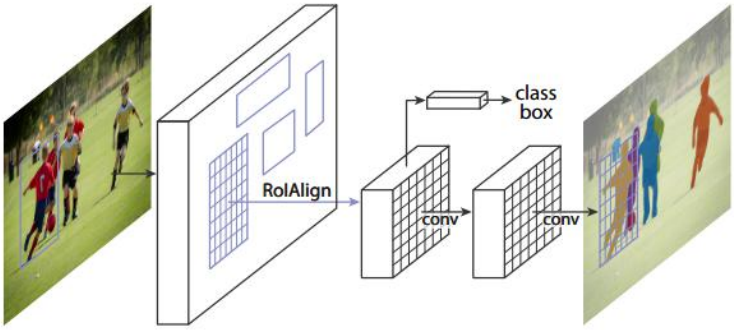


Figure 1. The **Mask R-CNN** framework for instance segmentation.

图 1 Mask R-CNN 的基本结构

如何正确的设计 mask 分支是结果好坏的关键。主要的要点有以下几点：

1. 最重要的一点是 Faster R-CNN 没有设计网络输入与输出的 pixel to pixel 的对齐机制。特别明显的是 ROIpooling 对特征提取执行非常粗糙的空间量化（空间量化指什么还没搞懂）。为了改进未对齐的缺点，本文提出了 quantization-free layer 叫作 RoIAlign，它准确的保存空间位置。尽管是很小的变化，但是作用很明显。提高相对 mask 准确率 10%~50%。

2. 非常必要的对 mask 和 class prediction 去耦合。本文对每个类别独立的预测一个二值 mask，不依赖分类分支的预测结果。

1.2 Mask R-CNN 的特点

- 在边框识别的基础上添加分支网络，用于语义 Mask 识别
- 训练简单，在 Faster RCNN 上仅增加一个小的 Overhead，可以跑到 5FP

- 可以方便的扩展到其他任务，比如人的姿态估计等
- 不借助 Trick，在每个任务上，效果优于目前所有的 single-model entries(包括 COCO 2016 的 Winners)

1.3 Mask R-CNN 技术要点

- 技术要点 1 - 强化的基础网络：通过 ResNeXt-101+FPN 用作特征提取网络，达到 state-of-the-art 的效果。
- 技术要点 2 - ROIAlign：采用 ROIAlign 替代 RoiPooling（改进池化操作）。引入了一个插值过程，先通过双线性插值到 14*14，再 pooling 到 7*7，很大程度上解决了仅通过 Pooling 直接采样带来的 Misalignment 对齐问题。虽然 Misalignment 在分类问题上影响并不大，但在 Pixel 级别的 Mask 上会存在较大误差。后面我们把结果对比贴出来，能够看到 ROIAlign 带来较大的改进，可以看到，Stride 越大改进越明显。
- 技术要点 3 - Loss Function：每个 ROIAlign 对应 K * m^2 维度的输出。K 对应类别个数，即输出 K 个 mask，m 对应 池化分辨率（7*7）。Loss 函数定义： Lmask(Cls_k) = Sigmoid (Cls_k)，平均二值交叉熵（average binary cross-entropy）Loss，通过逐像素的 Sigmoid 计算得到。通过对每个 Class 对应一个 Mask 可以有效避免类间竞争。

二、算法细节

Mask R-CNN 主要分为两个阶段：(1)生成候选框区域。该流程与 Faster R-CNN 相同，都是使用的 RPN（Region Proposal Network）。（2）在候选框区域上使用 RoIPool 来提取特征并进行分类和边界框回归，同时为每个 RoI 生成了一个二元掩码。这与当前大部分系统不一样，当前这些系统的类别分类依赖于 mask 的预测。我们还是沿袭了 Fast R-CNN 的精神，它将矩形框分类和坐标回归并行的进行，这么做很大的简化了 R-CNN 的流程。

2.1 LOSS fuction

多任务损失函数对于每一个 ROI，L=L_cls+L_box+L_mask.其中 L_cls 和 L_box 与 Faster R-CNN 一样。mask 分支对每一个 ROI 有 Km^2 维输出。表示分辨率为 m*m 的 K 个二值 mask。K 是类别数，每一类一个。对每个像素实行一个 sigmoid，定义 L_mask 是平均二值 cross-entropy loss。对于一个 ROI 的 ground truth 是第 k 类，L_mask 只定义在第 k 个 mask 上（其他 mask 输出对于损失没有贡献）。

图 2 human pose estimation

2.2 Mask Representation

mask 覆盖输入目标的空间位置，所以不能像类标和 bbox 一样通过全连接层坍塌到很短的向量。提取空间结构很自然的想到利用卷积的 pixel to pixel 对应的特性。具体的对每一个 ROI 预测一个 mm 大小的 mask 用 FCN。这能保证 mask 分支的每一层都明确的保持 mm 目标的空间布局，不会坍塌成缺少空间维度的向量。与前人工作使用全连接层预测 mask 相比，本文的 FCN 需要更少的参数，得到更好的效果 。pixel to pixel 的任务需要 ROI 特征与原始输入图像有很好对齐来保持每个像素的空间对应。这就是提出 RoIAlign 层的动机。

2.3 RoIAlign

ROIpool 是对 ROI 提取小的特征映射（e.g. 7*7）标准的操作符。量化导致了 ROI 和特征层的不对齐。这对分类任务没什么影响，但是对 pixel to pixel 的任务就有很大的负面影响。为了解决这个问题，本文提出了 RoIAlign 层，移除 ROIPool 粗糙的量化,正确的对齐特征和输入。提出的改变非常简单：避免任何 ROI 边界或者 bins 的量化，即用 x/16 代替[x/16]。用双向性插值法输入特征在每个 ROI bin 的四个采样点的精确值。

RoIPool 的目的是为了从 RPN 网络确定的 ROI 中导出较小的特征图(a small feature map)，ROI 的大小各不相同，但是 RoIPool 后都变成了 7x7 大小。RPN 网络会提出若干 RoI 的坐标以[x,y,w,h]表示，然后输入 RoI Pooling，输出 7x7 大小的特征图供分类和定位使用。问题就出在 RoI Pooling 的输出大小是 7x7 上，如果 RON 网络输出的 RoI 大小是 8*8 的，那么无法保证输入像素和输出像素是一一对应，首先他们包含的信息量不同（有的是 1 对 1，有的是 1 对 2），其次他们的坐标无法和输入对应起来（1 对 2 的那个 RoI 输出像素该对应哪个输入像素的坐标？）。这对分类没什么影响，但是对分割却影响很大。RoIAlign 的输出坐标使用插值算法得到，不再量化；每个 grid 中的值也不再使用 max，同样使用差值算法。

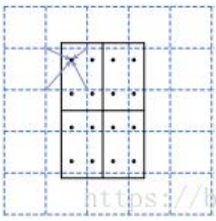


Figure 3. **RoIAlign:** The dashed grid represents a feature map, the solid lines an RoI (with 2×2 bins in this example), and the dots the 4 sampling points in each bin. RoIAlign computes the value of each sampling point by bilinear interpolation from the nearby grid points on the feature map. No quantization is performed on any coordinates involved in the RoI, its bins, or the sampling points.

图 3 RoIAlign

2.4 Network Architecture

将整个网络分成两部分:

1. 卷积主干结构用来提取整幅图像的特征。
2. 网络头用来对 ROI 进行 bbox 识别和 mask 预测。

分别考察 50 层和 101 层 Resnet 和 ResNeXt 网络作为卷积主干结构。还探索另一种有效的主干结构，叫作 FPN（ Feature Pyramid Network）。在 Mask R-CNN 中，在 Faster R-CNN 的 CNN 特征的顶部添加了一个简单的完全卷积网络（FCN），以生成 mask（分割输出）。请注意它是如何与 Faster R-CNN 的分类和边界框回归网络并行的。Mask R-CNN 通过简单地向 Faster R-CNN 添加一个分支来输出二进制 mask，以说明给定像素是否是目标的一部分。

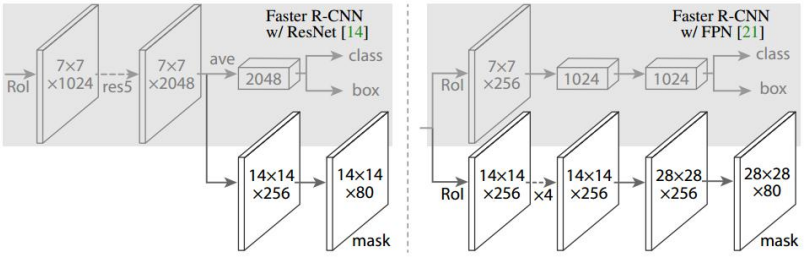


图 3 网络头结构(head 框架介绍图)

三、论文复现

Network Architecture：为了表述清晰，有两种分类方法。使用不同的backbone：resnet-50，resnet-101，resnext-50，resnext-101；或者使用不同的 head Architecture：Faster RCNN 使用 resnet50 时，从 CONV4 导出特征供 RPN 使用，这种叫做 ResNet-50-C4。作者使用除了使用上述这些结构外，还使用了一种更加高效的 backbone—FPN。

3.1 Training

使用 Fast/Faster 相同的超参数，同样适用于 Mask RCNN。

- 当 IoU 与 Ground Truth 的 IoU 大于 0.5 时才会被认为有效的 RoI，只把有效 RoI 计算进去。
- 采用 image-centric training，图像短边 resize 到 800，每个 GPU 的 mini-batch 设置为 2，每个图像生成 N 个 RoI，对于 backbone 的 N=64，对于 FPN 作为 backbone 的，N=512。作者服务器中使用了 8 块 GPU，所以总的 minibatch 是 16，迭代了 160k 次，初始 lr=0.02，在迭代到 120k 次时，将 lr 设定到 lr=0.002，另外学习率的 weight_decay=0.0001，momentum = 0.9。如果是 resnext，初始 lr=0.01,每个 GPU 的 mini-batch 是 1。
- RPN 的 anchors 有 5 种 scale，3 种 ratios。为了方便剥离、如果没有特别指出，则 RPN 网络是单独训练的且不与 Mask RCNN 共享权重。但是在本论文中，RPN 和 Mask R-CNN 使用一个 backbone，所以他们的权重是共享的。Ablation Experiments 为了方便研究整个网络中哪个部分其的作用到底有多大，需要把各部分剥离开。

3.2 Inference

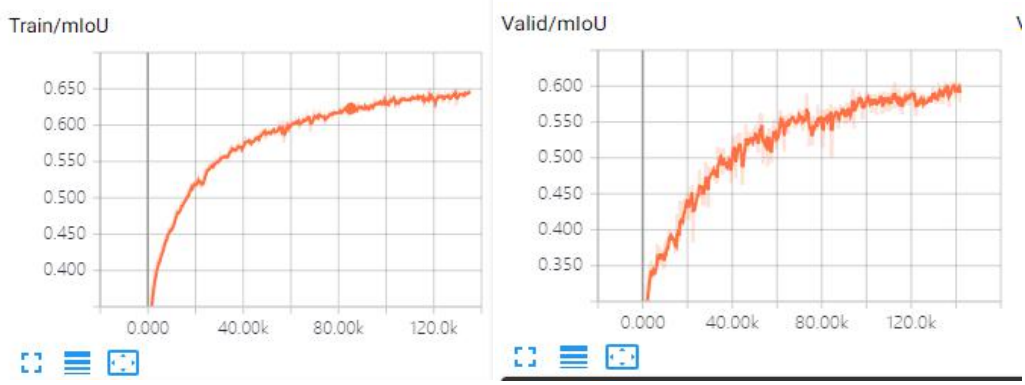
- 在测试时，使用 C4 backbone 情况下 proposal number=300，使用 FPN 时 proposal number=1000。然后在这些 proposal 上运行 bbox 预测，接着进行非极大值抑制。mask 分支只应用在得分最高的 100 个 proposal 上。顺序和 train 是不同的，但这样做可以提高速度和精度。mask 分支对于每个 roi

可以预测 k 个类别，但是我们只要背景和前景两种，所以只用 k-th mask，k 是根据分类分支得到的类型。然后把 k-th mask resize 成 roi 大小，同时使用阈值分割(threshold=0.5)二值化。

3.3 测试细节

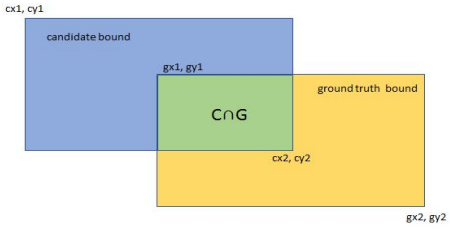
对于一张要测试的图片，将其按一定的分辨率(1024 * 1024)分成多张图片，随机切割并打乱顺序。随机切割的同时对标签也施加同样的操作，保证每一张子图与相应的子标签对应。

将数据集按一定比列分成训练集和验证集，将训练集送入网络进行训练，在 loss 函数收敛时，完成训练，保存模型参数同时进行测试模型的性能。



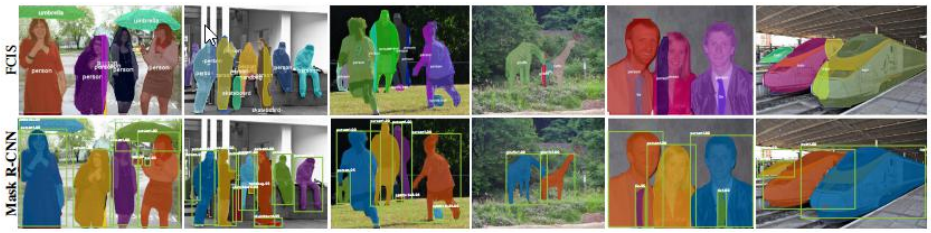
绘制 mIOU 曲线，模型收敛时在训练集上的 mIOU 为 0.66，在测试集上的 mIOU 为 0.6 左右。

IOU 交并比是目标检测中使用的一个概念，是产生的候选框（candidate bound）与原标记框（ground truth bound）的交叠率，即它们的交集与并集的比值。最理想情况是完全重叠，即比值为 1。



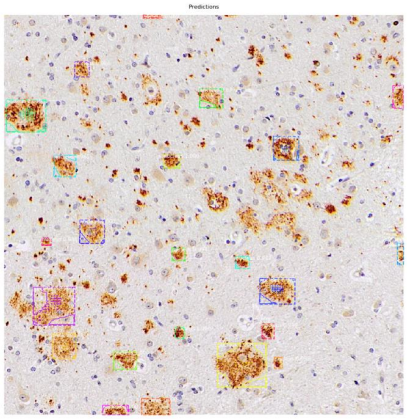
四、对比实验

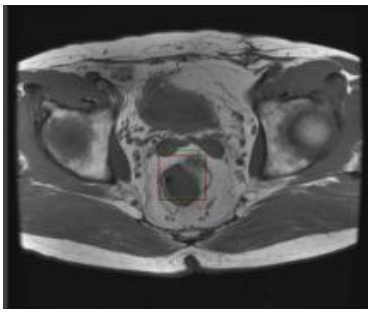
对比 FCIS 与 Mask RCNN， FCIS 的分割结果中都会出现一条竖着的线(systematic artifacts)，这线主要出现在物体重的部分，作者认为这是 FCIS 架构的问题，无法解决。但是在 Mask RCNN 中没有出现。



五、新数据集实验

将实现的网络 Mask RCNN 用于肿瘤检测与细胞检测。效果如下图：





六、总结

在目标检测领域，目前的主要有两种做法，一种是基于 RCNN 的带有 reign proposal 的方法；一种是先生成一些默认的检测 boxes，再根据网络的学习与 loss 函数将这些默认的 boxes 与 ground truth 不断靠近，以得到一个较好的识别效果，这样的方法诸如 YOLO、SSD 等，其优势是速度较快但在精度上往往有劣势。本项目主要根据论文实现 Mask RCNN，属于第一种做法。通过实现不同的网络 layers，复现论文的结果，并进行新的实验，将该方法应用到医学图像处理领域，如肿瘤位置检测、细胞检测等，并将结果展示。