## 摘要

在对象检测中，需要定义交并比（IoU）的阈值。以低IoU阈值训练的物体检测器，例如0.5，这通常会产生噪声检测。但是，随着IoU阈值的增加，检测性能趋于下降。造成这种情况的主要原因有两个：1）训练期间的过拟合，在训练期间由于正样本呈指数消失而过度拟合；2）探测器最佳的IoU的获得时间与输入假设的IoU之间的推理时间不匹配。为了解决这些问题，我们提出了一种多级目标检测架构Cascade R-CNN。它由一系列检测器组成，这些检测器由不断提高的IoU阈值训练而成，从而对接近的误报具有更高的选择性。我们对检测器的训练是逐步进行的，利用了这样的观察：一个检测器的输出的一个良好的分布，能够训练出下一个更高质量检测器。对逐步改进的假设的重新采样保证了所有检测器都有一组等效大小的正向例子，从而减少了过拟合问题。（时间）推理时采用相同的级联过程，从而使假设与每个阶段的检测器质量之间更紧密地匹配。对Cascade R-CNN的简单实现，在COCO数据集上，可以看到它的性能能够超越所有单模型对象检测器。实验还表明，Cascade R-CNN可广泛应用于各种检测器架构，从而获得稳定的增益，而与基线检测器强度无关。该代码将在https://github.com/zhaoweicai/cascade-rcnn提供。

## 介绍

对象检测是一个复杂的问题，需要解决两个主要任务。首先，检测器必须解决识别问题，以区分前景对象和背景，并为其分配适当的对象类别标签。其次，检测器必须解决定位问题，以便为不同的对象分配准确的边界框。两者都特别困难，因为检测器面临许多“接近”的假阳性，对应于“接近但不正确”的边界框。检测器必须在抑制这些接近的假阳性的同时找到真阳性。许多最近提出的对象检测器都基于“二阶段R-CNN框架”[12、11、27、21]，其中检测被构造为结合了分类和边界框回归的多任务学习问题。与对象识别不同，交并比（IoU）的阈值被用来定义正极/负极。然而，一般使用的阈值u（通常为u = 0.5）对正极建立了相当宽松的要求。生成的检测器经常会产生噪声边界框，如图1（a）所示。大多数人会认为接近的假阳性的假设经常通过IoU≥0.5的测试。尽管根据u = 0.5准则获得的示例丰富多样，但会难以训练能够有效拒绝近似误报的检测器。

在这项工作中，我们将假设的质量作为IoU的真值，将检测器的质量定义为用于训练它的IoU阈值。目标是调查迄今为止学习质量较高的目标检测器的研究不足，该检测器的输出几乎没有误报，如图1（b）所示。基本思想是，单个检测器只能针对单个质量级别进行优化。这在成本敏感型学习文献[7，24]中是已知的，其中，接收机工作特性（ROC）的不同点的优化需要不同的损耗函数。主要区别是，我们考虑针对给定IoU阈值的优化，而不是误报率。

图1（c）和（d）展示了这种想法。图1（c）和（d）分别展示了三个不同IoU阈值u = 0.5、0.6、0.7训练得到的检测器的定位和检测性能。如COCO [20]所示，根据输入建议的IoU评估定位性能（定位性能是建议输入IoU值的函数），并根据IoU阈值评估检测性能（检测性能是IoU门限值的函数）。请注意，在图1（c）中，每个边界框回归器在IoU接近检测器训练阈值的示例中表现最佳。这也适用于检测性能，甚至过拟合。图1（d）显示，对于低IoU实例，u = 0.5的检测器优于u = 0.6的检测器，在较高的IoU级别下其性能不如后者。通常，单个IoU级别优化的检测器在其他级别不一定是最佳的。这些观察结果表明，更高质量的检测要求检测器与假设之间的质量更紧密匹配。通常，只有在提出高质量建议时，检测器才能具有高质量。

然而，为了生产高质量的检测器，仅在训练期间提高u是不够的。实际上，从图1（d）中u = 0.7的检测器可以看出，这会降低检测性能。问题在于，在低质量下，提议的检测器发出的假设分布严重失衡。通常，强制的较大的IoU阈值会导致正训练样本数量成倍减少。对于神经网络而言，这尤其成问题，因为神经网络被认为是非常注重示例性的，并且使得“高u值”的训练策略非常容易过度拟合。另一个困难是，在时间推断时，检测器质量和测试假设质量之间的不匹配。如图1所示，对于高值的假设，高质量的检测器仅是最佳的。当要求他们根据其他水平的值的假设进行工作（训练）时，检测可能不是最佳的。

在本文中，我们提出了新型检测器架构Cascade R-CNN，它能够解决这些问题。它是R-CNN的多级扩展，其中级联更深的检测器级依次对接近的假阳性更具选择性。依次训练R-CNN阶段的级联，使用一个阶段的输出来训练下一个阶段。因为我们观察到，即回归器的输出IoU几乎总是好于输入IoU。可以在图1（c）中进行观察，其中所有图都在灰线上方。这表明按照某个IoU阈值训练得到的检测器的输出是良好的分布的，可以来训练下一个更高的IoU阈值的检测器。这类似于对象检测文献[31，8]中通常用于组装数据集的引导方法。主要区别在于，Cascade R-CNN的重采样过程并非旨在挖掘硬负极。相反，通过调整边界框，每个阶段的目的是找到一组好的假阳性集，以训练下一阶段。当以这种方式操作时，IoU值越来越高的的一系列检测器可以克服过度拟合的问题，因此可以得到有效的训练。推断时，将应用相同的级联过程。逐步改进的假设在每个阶段都与不断提高的检测器质量更好地匹配。如图1（c）和（d）所示，这可以实现更高的检测精度。

Cascade R-CNN的实现和端到端的训练非常简单。我们的结果表明，在具有挑战性的COCO检测任务[20]上，尤其是在更高质量的评估指标下，没有任何花哨的实现方式就大大超越了所有先前的最新单模型检测器。此外，可以使用任何基于R-CNN框架的两级对象检测器来构建Cascade R-CNN。我们已经观察到了一致的增益（2-4点），计算量略有增加。该增益与基线目标检测器的强度无关。因此，我们认为，这种简单有效的检测体系结构可能会引起许多对象检测研究的兴趣。

## 相关工作

由于R-CNN [12]体系结构的成功，通过组合“提议检测器”和“区域分类器”两级解决目标检测问题近来成为主流。为了减少R-CNN中的冗余CNN计算，SPP-Net [15]和Fast-RCNN [11]引入了区域特征提取的思想，显着加快了整个检测器的速度。后来，Faster-RCNN [27]通过引入区域提议网络（RPN）进一步提高了速度。该体系结构已成为领先的对象检测框架。最近的一些工作已将其扩展到能解决各种细节问题。例如，R-FCN [4]提出了有效的区域级全卷积，而没有精度损失，从而避免了Faster-RCNN繁重的区域级CNN计算。而MS-CNN [1]和FPN [21]在多个输出层检测提议，以减轻RPN接收字段与实际对象大小之间的比例失配，以进行高召回提议检测。

而一级目标检测架构也已变得很流行，这主要是由于它们的计算效率。这些架构接近经典的滑动窗口策略[31，8]。 YOLO [26]通过转发一次输入图像，输出非常稀疏的检测结果。当通过高效的骨干网络实现时，它可以以合理的性能实现实时对象检测。SSD [23]以类似于RPN [27]的方式检测物体，但是使用具有不同分辨率的多个特征图来覆盖各种比例的物体。这些架构的主要局限在于其精度通常低于两级检测器的精度。最近，RetinaNet [22]被提出来解决密集物体检测中的极端前景-背景类不平衡问题，其效果比最新的两级物体检测器更好。

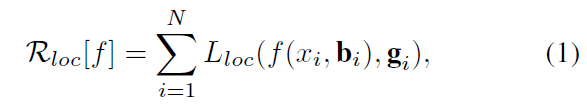
还有对多阶段目标检测的一些探索。多区域检测器[9]引入了迭代边界框回归，其中多次应用了R-CNN，以生成更好的边界框。CRAFT [33]和AttractioNet [10]使用了多级程序来生成准确的建议，并将其转发给Fast-RCNN。 [19，25]在对象检测网络中嵌入了[31]的经典级联体系结构。 [3]交替进行检测和分割任务，例如分割。

## 目标检测

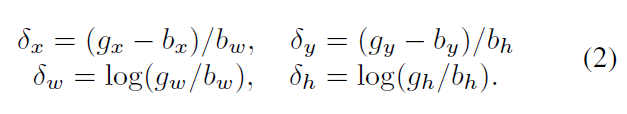
在本文中，我们扩展了Faster-RCNN的两阶段体系结构[27，21]，如图3（a）所示。 第一阶段是提议子（“ H0”），应用于整个图像，以产生初步的检测假设，称为目标提议。 在第二阶段，这些假设然后由表示为检测头的感兴趣区域检测子网（“ H1”）处理。每个假设都由一个最终分类得分（“ C”）和边界框（“ B”）。 我们专注于对多级检测子网进行建模，并采用但不限于RPN [27]进行提议检测。

### 边界框回归

边界框b =（bx，by，bw，bh）包含图像块x的四个坐标。边界框回归的任务是使用回归函数f（x，b）将候选边界框b回归到目标边界框g。这是从训练样本{gi，bi}中学到的，以便最小化边界框风险，



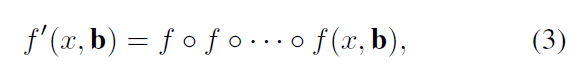
其中Lloc是R-CNN中的L2损失函数[12]，但在Fast-RCNN [11]中更新为平滑的L1损失函数。为了奖励尺度和位置的回归不变量，Lloc对距离矢量▲=（x，y，w，h）的定义为：

由于边界框回归通常对b进行较小的调整，因此（2）的数值可以很小。因此，（1）的风险通常比分类风险小得多。为了提高多任务学习的有效性，▲一般通过其均值和方差对其进行归一化，即

这在文献中被广泛使用[27，1，4，21，14]。由于包围盒回归通常会对b进行较小的调整，因此（2）的数值可能会很小。因此，（1）的风险通常比分类风险小得多。为了提高多任务学习的有效性，通常通过其均值和方差对其进行归一化，即x替换为'x =（x-μx）/ x。

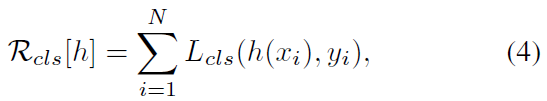
这在文献中被广泛使用[27，1，4，21，14]。

一些工作[9，10，16]认为f的单个回归步骤不足以进行精确定位。取而代之的是，迭代地应用f，作为完善边界框b的后处理步骤。



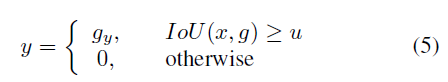
这称为迭代边界框回归，表示为迭代BBox。可以使用图3（b）的推理体系结构来实现，其中所有头都相同。但是，这个想法忽略了两个问题。首先，如图1所示，对于较高IoU的假设，在u = 0.5处训练的回归因子f次优。实际上，它会使大于0.85的IoU边界框降级。其次，如图2所示，边界框的分布在每次迭代后都会发生显著变化。虽然回归变量对于初始分布是最佳的，但是在那之后它可能是次优的。由于这些问题，迭代的BBox需要大量的人工工程，包括提案累积，框选择等形式[9，10，16]，并且增益有些不可靠。通常使用超过2次f没任何好处。

### 分类

分类器是一个函数h（x），它将图像块x分配给M + 1个类之一，其中类0包含背景和余下要检测的对象。 h（x）是所有类的后验分布的M + 1维估计，即hk（x）= p（y = k | x），其中y是类别标签。 给定训练集（xi，yi），可以通过最小化分类风险来学习其中Lcls是经典的交叉熵损失。

### 检测质量

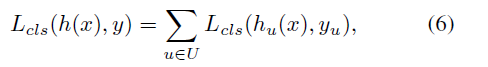
由于边界框通常包含一个对象和一定数量的背景，因此很难确定检测结果是正还是负。这通常通过IoU度量解决。如果IoU高于阈值u，则将该部分视为该类的示例。因此，假设x的类别标签是u的函数，



其中gy是真值框g的类标签。此IoU阈值u定义了检测器的质量。

对象检测具有挑战性，因为无论阈值如何，检测设置都具有很高的对抗性。当u高时，阳值包含较少的背景信息，但是很难收集足够的正向训练样例。当u低时，可以使用更丰富，更多样化的阳性训练集，但是训练好的检测器几乎没有动力拒绝接近的假阳性。通常，很难要求单个分类器在所有IoU级别上均一地表现良好。推断时，由于提议检测器产生的大多数假设，例如RPN [27]或选择性搜索[30]，它们的质量较低，对于较低质量的假设，检测器必须更具判别力。这些相互矛盾的要求之间的标准折中方案是使u = 0.5。但是，这是一个相对较低的阈值，导致大多数人认为是相似假阳性的低质量的检测，如图1（a）所示。

简单的解决方案是使用图3（c）的架构开发一组分类器，并针对不同质量级别的损失进行优化：



其中U是一组IoU阈值。这与[34]的积分损失密切相关，其中U = {0.5，0.55，···，0.75}，其设计，适应CoCO挑战的评价标准。根据定义，分类器需要在推理时进行组合。该解决方案无法解决以下问题：（6）的不同损失作用于不同数量的正数。如图4的第一个图所示，正样本集随u迅速减少。这尤其成问题，因为高质量的分类器容易过拟合。另外，这些高质量的分类器被要求在推理时处理压倒性低质量的提案，但它们并不是最优的。由于所有这些，（6）的集合无法在大多数质量级别上实现更高的精度，并且该架构与图3（a）相比，增益很小。

## Cascade R-CNN

在本节中，我们将介绍图3(D)中提出的Cascade R-CNN对象检测体系结构。

### 级联边界框回归

如图1（c）所示，很难要求单个回归器在所有质量级别上都能完美地统一执行。受级联姿态回归[6]和面部对齐[2，32]的启发，可以将困难的回归任务分解为一系列更简单的步骤。在Cascade R-CNN中，它以图3（d）的体系结构构造为级联回归问题。这依赖于级联的专业回归函数：



其中T是级联级的总数。请注意，级联中的每个回归变量ft根据针对到达相应阶段的样本分布{bt}而不是{b1}的初始分布进行了优化。这种级联逐步改善了假设。

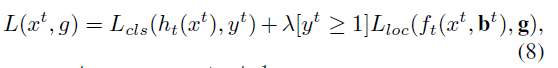
它在几个方面与图3（b）的迭代BBox体系结构不同。首先，虽然迭代BBox是用于改善边界框的后处理过程，但级联回归是一种重采样过程，可以更改要由不同阶段处理的假设的分布。其次，由于它既用于训练又用于推理，因此训练和推理分布之间没有差异。第三，针对不同阶段的重采样分布，优化了多个专用回归变量{fT，fT-1，...，f1}。这与（3）的单个f相反，后者仅对于初始分布是最佳的。这些差异可以实现比迭代BBox更精确的定位，而无需进行进一步的人工操作。

如第3.1节所述，（2）中的需要通过其均值和方差进行归一化，以进行有效的多任务学习。在每个回归阶段之后，这些统计信息将顺序发展，如图2所示。在训练中，将使用相应的统计信息在每个阶段进行归一化。

### 级联检测

如图4左侧所示，初始假设的分布（例如RPN提案）在很大程度上倾向于低质量。这不可避免地导致对高质量分类器的无效学习。Cascade R-CNN通过将级联回归作为重采样机制来解决此问题。这是由于在图1（c）中所有曲线都位于对角线灰色线上方，即针对某个u进行训练的边界回归器往往会产生更高IoU的边界框。因此，从一组示例（xi，bi）开始，级联回归依次对较高IoU的示例分布（x'i，b'i）进行重采样。以这种方式，即使当检测器质量（IoU阈值）增加时，也可以将连续阶段的正例的集合保持在大致恒定的大小。如图4所示，在每个重采样步骤之后，分布向高质量示例的倾斜程度更大。随之而来的是两种后果。首先，没有过度拟合，因为各个级别的示例很多。其次，针对较高的IOU阈值，对较深阶段的检测器进行了优化。请注意，如图2所示，通过增加IoU阈值可以顺序除去一些离群值，从而可以更好地训练专用检测器序列。

在每个阶段t，R-CNN包括分类器ht和针对IoU阈值ut优化的回归器ft，其中ut> ut-1。通过最小化损耗来保证这一点：



其中，g是xt的真值，λ= 1是折衷系数，[·]指标函数，yt是xt由（5）给出。与（6）的积分损失不同，这保证了一系列训练有素的检测器质量不断提高。在推论中，通过应用相同的级联过程，假设的质量得到了依次改善，并且只需要较高质量的检测器即可对较高质量的假设进行操作。如图1（c）和（d）所示，这可以实现高质量的物体检测。

## 实验结果

Cascade R-CNN在MS-Coco2017[20]上进行了评估，其中包含用于培训的∼118 k图像、用于验证的5k图像和用于在没有提供注释的情况下进行测试的∼20k(test-dev)图像。CoCo风格的平均精度(AP)跨越IOU阈值的平均值为0.5至0.95，间隔为0.05。这些评价指标衡量各种质量的检测性能。所有模型均在COCO训练集上进行训练，并在Val集上进行评估。最后的结果也报告了test-dev集。

### 实现细节

对于简单性而言，所有的回归者都是类不可知的。级联R-CNN的所有级联检测阶段都具有相同的体系结构，是基线检测网络的核心。总之，Cascade R-CNN有四个阶段，一个RPN和三个检测U={0.5，0.6，0.7}，除非另有说明。第一检测阶段的取样如下[11，27]。在以下阶段，重采样是通过简单地使用前一阶段的回归输出来实现的，如第4.2节所示。除标准水平图像翻转外，不使用数据增强。推理是在单一的图像尺度上进行的，没有进一步的铃铛和哨声。所有基线探测器都用CAFFE[18]重新实现，在相同的代码基上进行公平比较。

### 基线网络

为了测试Cascade R-CNN的多功能性，对三个流行的基线检测器进行了实验：以VGG-Net为骨干的Faster RCNN[29]，以ResNet[16]为骨干的R-FCN[4]和FPN[21]。这些基线具有广泛的检测性能。除非提到，否则将使用它们的默认设置。用端到端训练代替多步训练

**Faster RCNN**: 网络头有两个完全连接的层。为了减少参数，我们使用[13]修剪不太重要的连接。每个完全连接的层保留了2048个单元，删除了dropout层。训练以0.002的学习速度开始，60k和90K迭代时减少10倍，在100 k迭代时停止，在2个同步GPU上，每个GPU每次迭代保存4幅图像。每幅图像使用128份ROI。

**R-FCN**: R-FCN为ResNet添加了一个卷积、一个边界框回归和一个分类层.所有级联R-CNN的头目都有这个结构。没有使用在线Hard negative mining[28]。训练开始时的学习率为0.003，在160 k和240 k迭代时下降了10倍，在280 k迭代时停止,在4个同步GPU上，每次迭代保持一幅图像。每幅图像使用256份ROIS。

**FPN**： 由于还没有公开的FPN源代码，我们的实现细节可能会有所不同。RoIAlign[14]用于更强的基线。这被表示为FPN+，并被用于所有的消融研究。与往常一样，ResNet-50用于消融研究，ResNet-101用于最终检测.在8个同步GPU上，训练使用了120 k迭代的学习速率为0.005，在接下来的60k迭代中使用了0.0005，每个GPU每次迭代保存一幅图像。每幅图像使用256份ROIS。

### 质量失配

图5(A)显示了三个单独训练的探测器增加IOU阈值U={0.5、0.6、0.7}的AP曲线。在低IOU水平下，u=0.5的检测器优于u=0.6的检测器，但在较高的水平上性能较差。然而，u=0.7的检测器性能不如其他两个。为了理解为什么会出现这种情况，我们在推论中改变了建议的质量。图5(B)显示了将真值框添加到提案集中时所获得的结果。虽然所有检测器都有所改进，但u=0.7的检测器收益最大，几乎在所有IOU级别上都取得了最佳性能。这些结果提出了两个结论。首先，u=0.5对于精确检测来说不是一个很好的选择，只是对低质量的方案更有鲁棒性。其次，高精度检测需要假设与检测器质量相匹配的假设，其次，原来的检测器方案被质量较高的Cascade R-CNN方案所取代(u=0.6和u=0.7，分别采用第二和第三阶段方案)。图5(A)还表明，当测试方案与检测器质量更接近时，两种检测器的性能都有了显著提高。

在所有级联阶段测试所有级联R-CNN探测器产生了类似的观测结果。图6显示，当使用更精确的假设时，每个探测器都得到了改进，而高质量的探测器具有更大的增益。例如，u=0.7的检测器对于第一阶段的低质量建议表现不佳，但对于在更深的级联阶段有更精确的假设则要好得多。此外，图6的联合训练检测器的性能优于单独训练的图5(A)检测器，即使使用了同样的建议。这表明探测器在Cascade R-CNN框架内得到了更好的训练。

### 与迭代BBOX和积分损耗的比较

在这一部分中，我们比较了Cascade R-CNN与迭代BBOX和积分损耗检测器。迭代BBOX是通过迭代三次应用FPN+基线来实现的。积分损耗检测器具有与Cascade R-CNN相同的分类头数，U={0.5，0.6，0.7}。

定位图7(a)比较了级联回归和迭代BBox的定位性能。当迭代中应用回归方程时，这种效应累积，如迭代Bbox中的，性能实际上下降。注意迭代BBox在3次迭代后的性能非常差。相反，级联回归器在后期具有更好的性能，在几乎所有IoU级别上都优于迭代BBox。

积分损失：图7(B)显示了共享一个回归器的积分损失检测器中所有分类器的检测性能。在所有IOU水平上，u=0.6的分类器最好，u=0.7的分类器最差。所有分类器的集合都没有明显的增益。

表1显示，迭代BBOX和积分损耗检测器都略微改进了基线检测器。Cascade R-CNN对所有的评估指标都有最好的性能。低IOU阈值的收益较小，而较高的IOU阈值则显著。

### 消融实验

我们还进行了消融实验。

**阶段比较**：表2总结了阶段性能。由于多阶段多任务学习的优点，第一阶段的性能已经优于基线检测器。第二阶段大大提高性能，第三阶段相当于第二阶段。这与积分损耗检测器不同，后者较高的IOU分类器相对较弱。虽然前(后)阶段在低(高)IOU度量时更好，但所有分类器的集合总体上是最好的。

**IoU 阈值**：使用相同的IOU阈值u=0.5对所有头部进行了初步的级联R-CNN训练，在这种情况下，阶段只是在他们收到的假设不同。每个阶段都使用相应的假设进行训练，即计算图2的分布。表3的第一行显示，级联改进了基线检测器。这表明了为相应的样本分布优化阶段的重要性。第二行表明，通过增加级阈值u，可以使检测器对防止相似假阳性更具选择性，并专门用于更精确的假设，从而带来额外的增益。这论证了第4.2节的结论。

**回归统计**：利用图2中逐步更新的回归统计数据，有助于对分类和回归进行有效的多任务学习。它的好处是通过比较表3中有或没有它的模型来实现的。学习对这些统计数字不敏感。

**阶段数量**：表4总结了阶段数的影响。增加第二检测级显着地改进了基线检测器。三个检测阶段仍然产生非平凡的改善，但第四个阶段(u=0.75)的增加导致了轻微的性能下降。然而，请注意，虽然AP的整体性能下降，但对于高IOU级别，四阶段级联具有最佳的性能。三级级联实现了最佳的权衡。

### 与最先进的检测器比较

基于FPN+和ResNet-101主干网的Cascade R-CNN与表5中最先进的单模型物体检测器进行了比较。这些设置如第5.1.1节所述，但总共运行了280 k的训练迭代，在160 k和240 k的迭代中，学习率下降了。Rois的数量也增加到512。表5中的第一组检测器为一级（one-stage）检测器，第二组为two-stage检测器，最后一组为多级检测器(3-stages+RPN的Cascade R-CNN)。所有比较的先进的探测器都是用u=0.5训练的。需要指出的是，我们的FPN+实现比最初的FPN[21]更好，提供了一个非常强大的基准。此外，从FPN+到Cascade R-CNN的扩展提高了∼4点的性能。在所有的评估指标下，Cascade R-cnn的性能也大大超过了所有单一型号的探测器。这包括2015年和2016年CoCo挑战赛获奖者的单模参赛作品(Faster R-CNN[16]和G-RMI[17])，以及最近Deformable R-FCN[5]、RetinaNet[22]和MASK R-CNN[14]。CoCO上最好的多级检测器，AttractioNet[10]，使用迭代BBOX来生成提案。虽然AttractioNet中使用了许多增强功能，但普通的Cascade R-CNN的性能仍然比它高7.1分。注意，与MASKR-CNN不同的是，在Cascade R-CNN中没有利用分段信息。最后，普通的单模级联R-cnn也超过了2015年和2016年获得CoCo挑战的重工设计集成探测器(AP 37.4和41.6)1。

### 泛化能力

表6比较了三种基线检测器的三级Cascade R-CNN。所有设置均如上，对5.5节的FPN+作了更改。

**检测性能**：同样，我们的实现比原来的检测器[27，4，21]更好。尽管如此，Cascade R-CNN在这些基线上持续改进2∼4点，独立于他们的力量。这些收益在Val和test-dev上也是一致的。这些结果表明Cascade R-CNN在检测器体系结构中具有广泛的应用前景。

**参数和时间**：Cascade R-CNN参数的数目随着级联级的数目而增加。基线探测器头参数的增加是线性的。此外，由于检测头的计算成本通常比RPN小，因此Cascade R-CNN的计算开销在训练和测试中都是很小的。