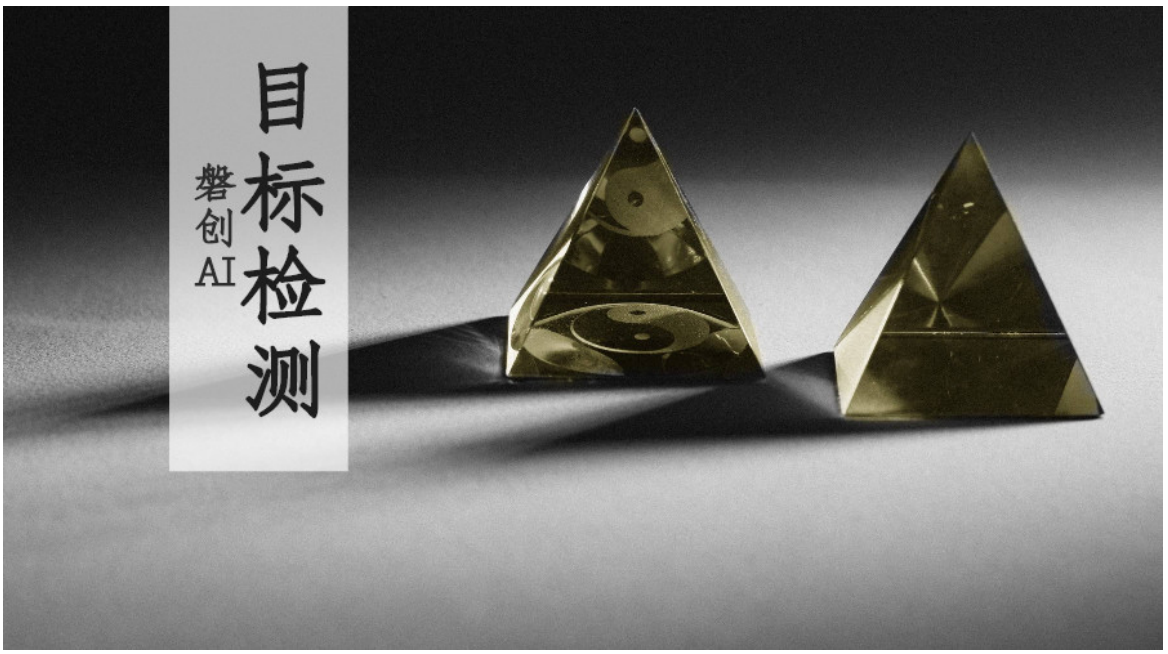
你好，这里有一份2019年目标检测指南

原创： Arno



编译 | Arno

来源 | Medium

目标检测(Object detection)是一种计算机视觉技术，旨在检测汽车、建筑物和人类等目标。这些目标通常可以通过图像或视频来识别。

目标检测在视频监控、自动驾驶汽车、人体跟踪等领域得到了广泛的应用。在本文中，我们将了解目标检测的基础知识，并回顾一些最常用的算法和一些全新的方法。

目标检测的原理

目标检测定位图像中目标的存在，并在该目标周围绘制一个边界框(bounding box)。这通常包括两个过程:预测目标的类型，然后在该目标周围绘制一个框。现在让我们来回顾一些用于目标检测的常见模型架构:

- R-CNN
- Fast R-CNN
- Faster R-CNN
- Mask R-CNN
- SSD (Single Shot MultiBox Defender)

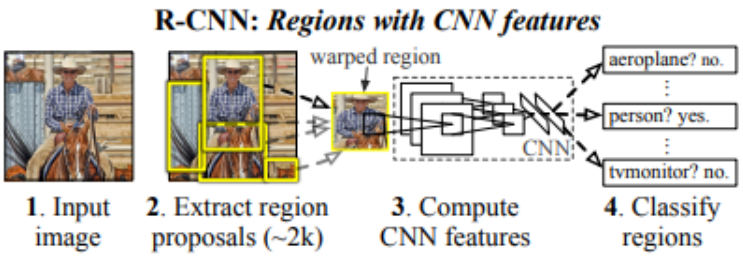
- YOLO (You Only Look Once)
- Objects as Points
- Data Augmentation Strategies for Object Detection

R-CNN 模型

该技术结合了两种主要方法:使用一个高容量的卷积神经网络将候选区域(region-proposals)自底向上的传播，用来定位和分割目标；如果有标签的训练数据比较少，可以使用训练好的参数作为辅助，进行微调(fine tuning)，能够得到非常好的识别效果提升。

> 论文链接: https://arxiv.org/abs/1311.2524?source=post_page----------

进行特定领域的微调，从而获得高性能的提升。由于将候选区域(region-proposals)与卷积神经网络相结合，论文的作者将该算法命名为R-CNN(Regions with CNN features)。



该模型在对每张图片提取了约2000个自底向上的候选区域。然后，它使用一个大型CNN计算每个区域的特征。然后，利用专门针对类别数据的线性支持向量机(SVMs)对每个区域进行分类。该模型在PASCAL VOC 2010上的平均精度达到53.7%。

该模型中的目标检测系统由三个模块组成。第一个负责生成类别无关的候选区域，这些区域定义了一个候选检测区域的集合。第二个模块是一个大型卷积神经网络，负责从每个区域提取固定长度的特征向量。第三个模块由一个指定类别的支持向量机组成。

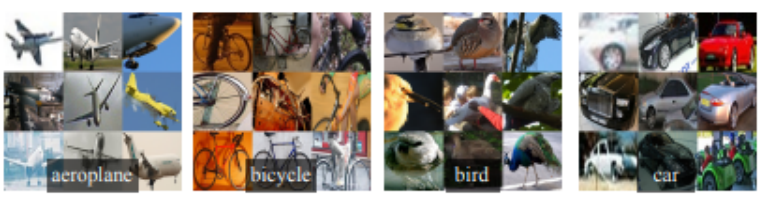


Figure 2: Warped training samples from VOC 2007 train.

该模型采用选择性搜索(selective search)方法来生成区域类别，根据颜色、纹理、形状和大小选择搜索对相似的区域进行分组。在特征提取方面，该模型使用CNN的一个Caffe实现版本对每个候选区域抽取一个4096维度的特征向量。将227×227 RGB图像通过5个卷积层和2个完全连接层进行前向传播，计算特征。论文中所解释的模型与之前在PASCAL VOC 2012的结果相比，取得了30%的相对改进。

而R-CNN的一些缺点是:

- **训练需要多阶段: 先用ConvNet进行微调，再用SVM进行分类，最后通过regression对 bounding box进行微调。**
- **训练空间喝时间成本大: 因为像VGG16这样的深度网络占用了大量的空间。**
- **目标检测慢: 因为其需要对每个目标候选进行前向计算。**

Fast R-CNN

下面的论文中提出了一种名为Fast Region-based Convolutional Network(Fast R-CNN)目标检测方法。

https://arxiv.org/abs/1504.08083?source=post_page----------

它是用Python和c++使用Caffe实现的。该模型在PASCAL VOC 2012上的平均精度为66%，而R-CNN的平均精度为62%。

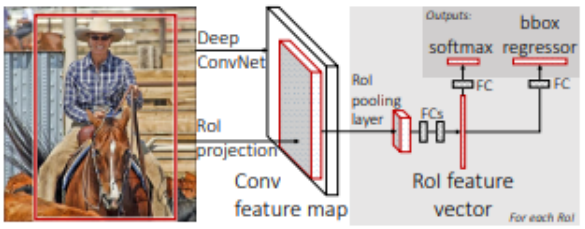


Figure 1. Fast R-CNN architecture. An input image and multiple regions of interest (RoIs) are input into a fully convolutional network. Each RoI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers (FCs). The network has two output vectors per RoI: softmax probabilities and per-class bounding-box regression offsets. The architecture is trained end-to-end with a multi-task loss.

与R-CNN相比，Fast R-CNN具有更高的平均精度、单阶段训练，训练更新所有网络层并且特征缓存不需要磁盘存储。

在其架构中， Fast R-CNN接收图像以及一组目标候选作为输入。然后通过卷积层和池化层对图像进行处理，生成卷积特征映射。然后，通过针对每个推荐区域，ROI池化层从每个特征映射中提取固定大小的特征向量。

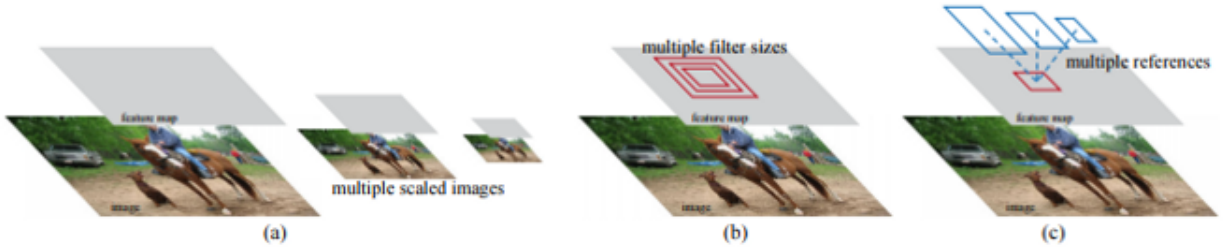
然后将特征向量提供给完全连接层。然后这些分支成两个输出层。其中一个为多个目标类生成softmax概率估计，而另一个为每个目标类生成4个实数值。这4个数字表示每个目标的边界框的位置。

Faster R-CNN

论文链接: https://arxiv.org/abs/1506.01497?source=post_page----------

论文提出了一种针对候选区域任务进行微调和针对目标检测进行微调的训练机制。

2



Faster R-CNN模型由两个模块组成:负责提出区域的深度卷积网络和使用这些区域的Fast R-CNN探测器。候选区域网络(Region Proposal Network)以图像为输入，生成矩形目标候选的输出。每个矩形都有一个objectness score。

4

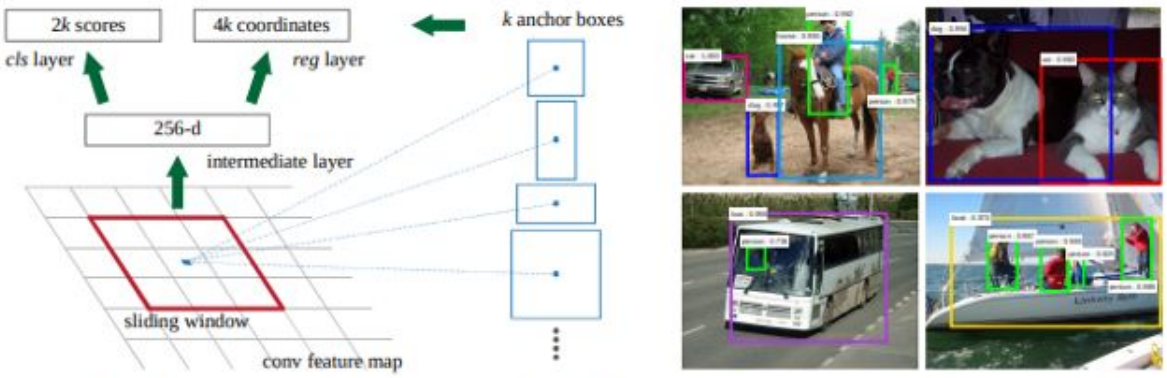


Figure 3: **Left**: Region Proposal Network (RPN). **Right**: Example detections using RPN proposals on PASCAL VOC 2007 test. Our method detects objects in a wide range of scales and aspect ratios.

Mask R-CNN

论文链接: https://arxiv.org/abs/1703.06870?source=post_page----------

论文提出的模型是上述Faster R-CNN架构的扩展。它还可以用于人体姿态估计。

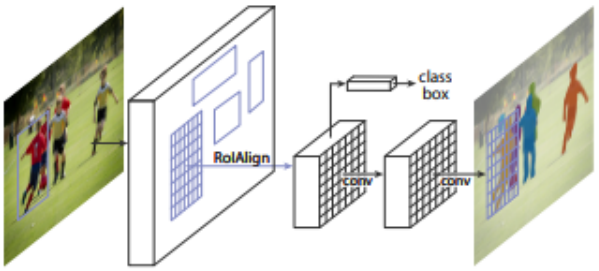


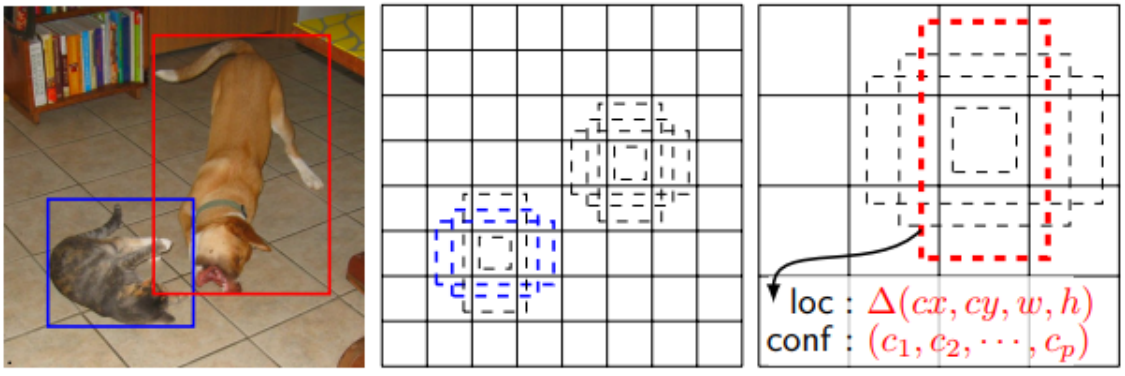
Figure 1. The **Mask R-CNN** framework for instance segmentation.

在该模型中，使用边界框和对每个像素点进行分类的语义分割对目标进行分类和定位。该模型通过在每个感兴趣区域(ROI)添加分割掩码(segmentation mask)的预测，扩展了Faster R-CNNR-CNN。Mask R-CNN产生两个输出:类标签和边界框。

SSD: Single Shot MultiBox Detector

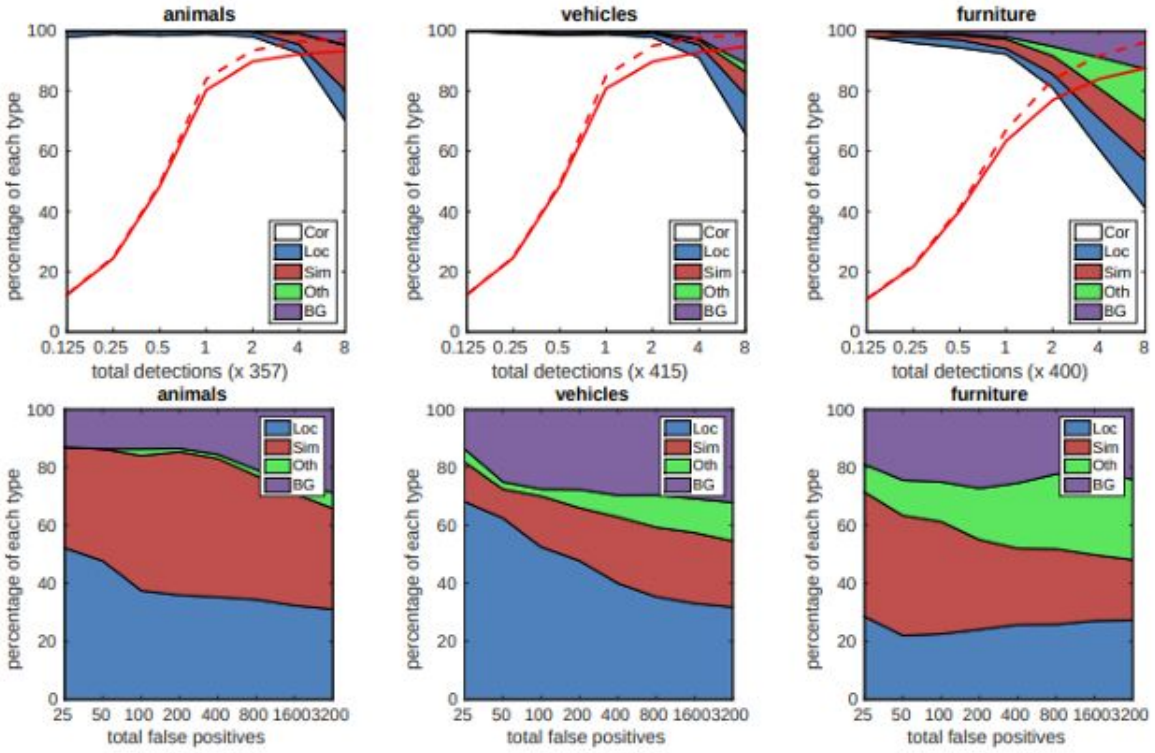
> 论文链接: https://arxiv.org/abs/1512.02325?source=post_page----------

论文提出了一种利用单个深度神经网络对图像中目标进行预测的模型。该网络使用应用于特征映射的小卷积滤波器为每个目标类别生成分数。



(a) Image with GT boxes (b) 8 × 8 feature map (c) 4 × 4 feature map

这种方法使用了一个前馈卷积神经网络，针对那些方框里的目标类别实例，产生一个固定大小的边界框的集合和分数。增加了卷积特征层，允许多比例特征映射检测。在这个模型中，每个特征映射单元(feature map cell)都链接到一组默认的边界框(default box)。下图显示了SSD512在动物、车辆和家具上的性能。



You Only Look Once (YOLO)

论文提出了一种基于神经网络的图像边界框和类概率预测方法。

> 论文链接: https://arxiv.org/abs/1506.02640?source=post_page-----------------------

YOLO模型每秒实时处理45帧。YOLO将图像检测看作是一个回归问题，使得它的管道非常简单。因为这个简单的管道，它非常快。

它可以实时处理流视频，延迟小于25秒。在训练过程中，YOLO可以看到整个图像，因此能够在目标检测中包含上下文。

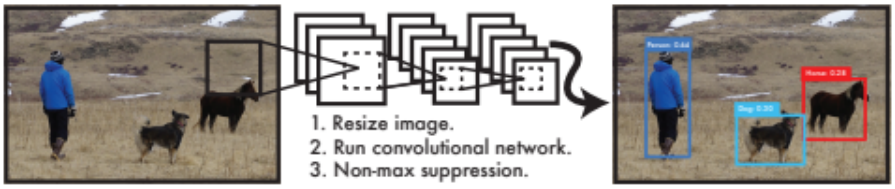
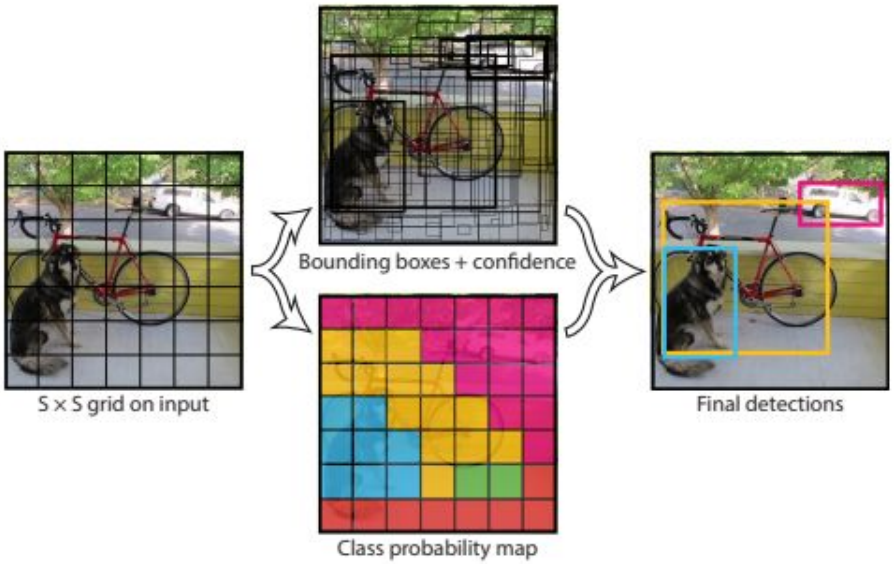


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

在YOLO中，每个边界框都由整个图像的特征来预测。每个边界框有5个预测:x，y，w，h，和置信度。(x，y)表示边界框的中心相对于网格单元格的边界。w和h是整个图像的预测宽度和高度。

该模型作为卷积神经网络实现，并在PASCAL VOC检测数据集上进行了评价。网络的卷积层负责提取特征，全连接层负责预测坐标和输出概率。



该模型的网络架构受到了用于图像分类的GoogLeNet模型的启发。该网络有24个卷积层和2个全连接层。该模型的主要挑战在于，它只能预测一个类，而且在鸟类等小目标上表现不佳。

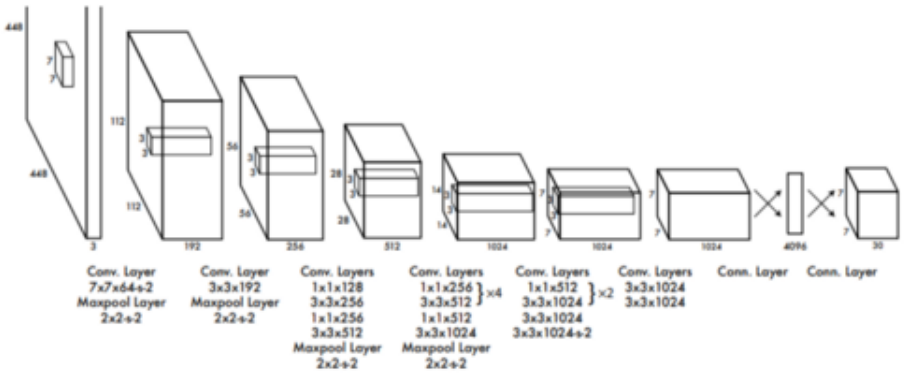


Figure 3: The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection.

该模型的平均精度达到52.7%，但有可能达到63.4%。

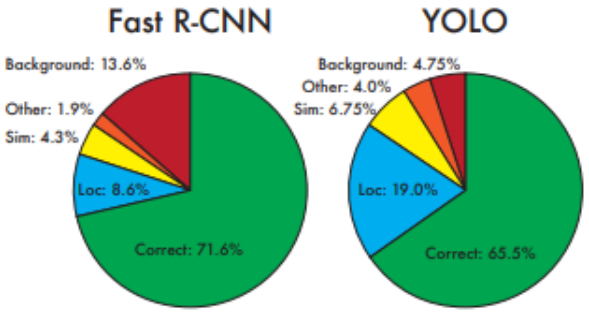


Figure 4: Error Analysis: Fast R-CNN vs. YOLO These charts show the percentage of localization and background errors in the top N detections for various categories (N = # objects in that category).

Objects as Points

论文提出将目标建模为单个点。它使用关键点估计来找到中心点，并回归到其他目标属性。

> 论文链接: https://arxiv.org/abs/1904.07850v2?source=post_page----------

这些属性包括3D位置、姿态和尺寸。它使用了CenterNet，这是一种基于中心点的方法，比其他边界框探测器更快、更准确。

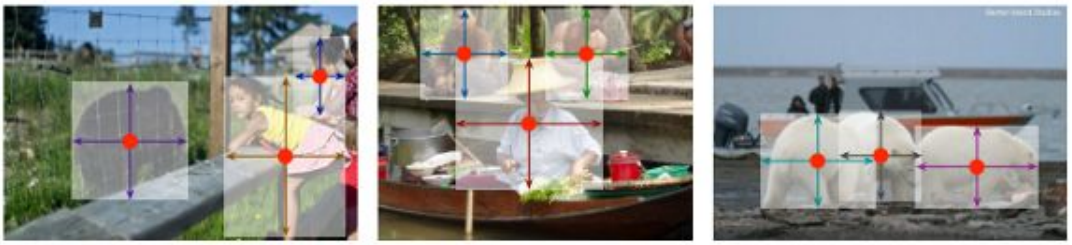


Figure 2: We model an object as the center point of its bounding box. The bounding box size and other object properties are inferred from the keypoint feature at the center. Best viewed in color.

目标大小和姿态等属性是由图像中心位置的特征回归得到的。该模型将图像输入卷积神经网络，生成热力图。这些热力图中的峰值表示图像中目标的中心。为了估计人体姿态，该模型检测关节点（2D joint）位置，并在中心点位置对其进行回归。

在COCO上，该模型以每秒1.4帧的速度实现了45.1%的平均精度。下图显示了与其他研究论文的结果进行比较的结果。

	Backbone	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MaskRCNN [21]	ResNeXt-101	**11**	39.8	62.3	43.4	22.1	43.2	51.2
Deform-v2 [63]	ResNet-101	-	46.0	67.9	50.8	27.8	49.1	59.5
SNIPER [48]	DPN-98	*2.5*	46.1	67.0	51.6	29.6	48.9	58.1
PANet [35]	ResNeXt-101	-	47.4	67.2	51.8	30.1	**51.7**	60.0
TridentNet [31]	ResNet-101-DCN	0.7	**48.4**	**69.7**	**53.5**	**31.8**	51.3	**60.3**
YOLOv3 [45]	DarkNet-53	20	33.0	57.9	34.4	18.3	25.4	41.9
RetinaNet [33]	ResNeXt-101-FPN	5.4	40.8	61.1	44.1	24.1	44.2	51.2
RefineDet [59]	ResNet-101	-	36.4 / 41.8	57.5 / 62.9	39.5 / 45.7	16.6 / 25.6	39.9 / 45.1	51.4 / 54.1
CornerNet [30]	Hourglass-104	4.1	40.5 / 42.1	56.5 / 57.8	43.1 / 45.3	19.4 / 20.8	42.7 / 44.8	**53.9** / 56.7
ExtremeNet [61]	Hourglass-104	3.1	40.2 / 43.7	55.5 / 60.5	43.2 / 47.0	20.4 / 24.1	43.2 / 46.9	53.1 / 57.6
FSAF [62]	ResNeXt-101	*2.7*	**42.9** / 44.6	**63.8 / 65.2**	**46.3** / 48.6	**26.6 / 29.7**	**46.2 / 47.1**	52.7 / 54.6
CenterNet-DLA	DLA-34	**28**	39.2 / 41.6	57.1 / 60.3	42.8 / 45.1	19.9 / 21.5	43.0 / 43.9	51.4 / 56.0
CenterNet-HG	Hourglass-104	7.8	42.1 / **45.1**	61.1 / 63.9	45.9 / **49.3**	24.1 / 26.6	45.5 / **47.1**	52.8 / **57.7**

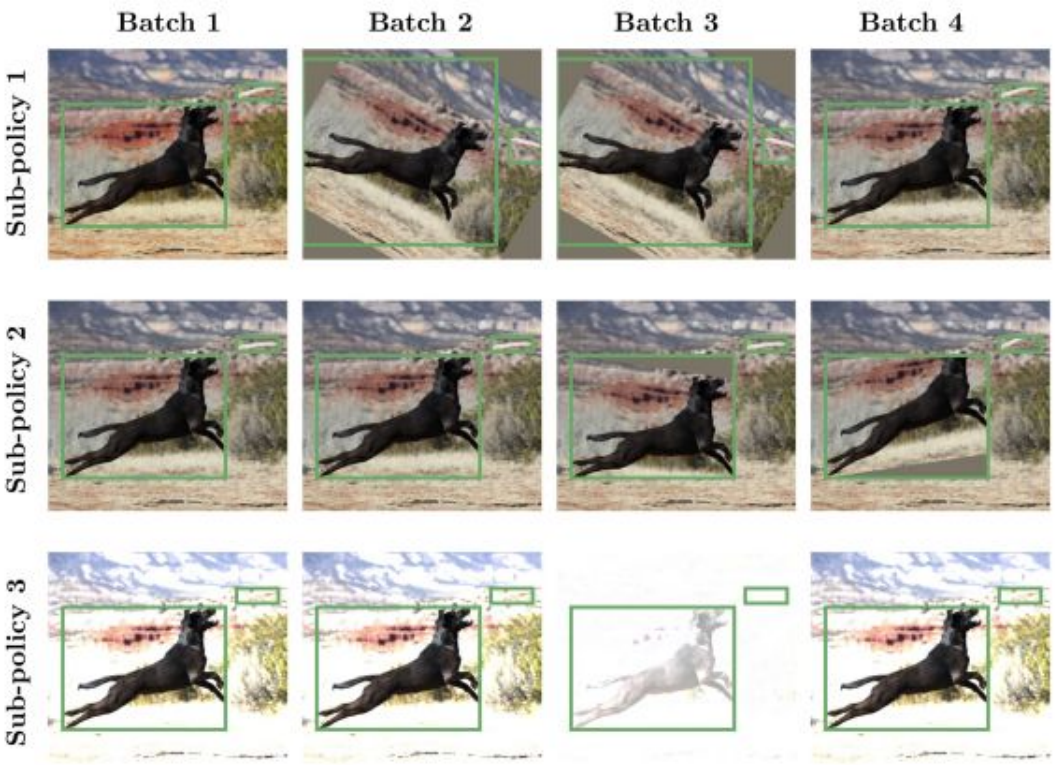
Table 2: State-of-the-art comparison on COCO test-dev. Top: two-stage detectors; bottom: one-stage detectors. We show single-scale / multi-scale testing for most one-stage detectors. Frame-per-second (FPS) were measured on the same machine whenever possible. Italic FPS highlight the cases, where the performance measure was copied from the original publication. A dash indicates methods for which neither code and models, nor public timings were available.

Learning Data Augmentation Strategies for Object Detection

数据增广包括通过旋转和调整大小等操作原始图像来创建新图像数据的过程。

论文链接: https://arxiv.org/abs/1906.11172v1?source=post_page----------

虽然这本身不是一个模型结构，但论文提出了可以应用于可以转移到其他目标检测数据集的目标检测数据集的变换的创建。转换通常在训练时应用。



该模型将增广策略定义为训练过程中随机选择的n个策略集合。该模型中应用的一些操作包括颜色变化、图像几何变化以及只变化bounding box annotations的像素内容。

在COCO数据集上的实验表明，优化数据增广策略可以使检测精度提高到+2.3以上的平均精度。这使得单个推理模型的平均精度达到50.7。

总结

现在，我们应该对在各种上下文中进行目标检测的一些最常见的技术(以及一些最新的技术)有所了解。

上面的论文/摘要也包含它们的代码实现的链接。希望能看到你在测试这些模型后得到的结果。

你也许还想看：

- NLPer入门指南 | 完美第一步
- 一文总结数据科学家常用的Python库（下）
- 一文看懂NLP神经网络发展历史中最重要的8个里程碑！

欢迎扫码关注：

点击下方 | 阅读原文 | 了解更多

阅读原文