Object Detection with Deep Learning: A Survey*

Xingwang Xiong †
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Haiding Distinct, Beijing
xiongxingwang@ict.ac.cn

Abstract

目标检测，作为计算机视觉研究中最基础的并且最富有挑战性的任务之一，一直以来备受研究人员的关注。深度神经网络可以从数据中学习特征，具有强大的表征能力。随着深度神经网络的发展，特别是深度卷积网络，目标检测领域迎来了突破性的进展。本文将主要介绍近几年主流的基于深度学习的目标检测算法及其主要的算法原理。

1 Introduction

目标检测，解决的问题是从图像中检测并标记出实例，既包括了边框级别的预测输出，也包括了像素级别的预测输出。目标检测也是许多其他计算机视觉的基础任务之一，如实例分割，目标跟踪，看图说话等。

目标检测的发展，得益于算法/模型研究的突破、计算能力的发展、和大规模的、高质量的数据集的开放。

目前目标检测的深度学习模型主要包括两类：Two-stage 的检测算法和 One-stage 的检测算法。前者往往具有更好的检测精度，而后者通常在速度上更有优势。

Two-stage 的检测算法基于 Regional Proposal，针对每个样本生成若干的候选框，然后用卷积神经网络（CNN）对候选区域进行特征提取、分类。这类算法的代表性工作有：RCNN [1]、SPPNet [2]、Fast RCNN [3]、Faster RCNN [4]、R-FCN、Mask RCNN [5] 等。

*本文为三维视觉与机器学习课程论文

†S.No: 2018E8013261007; My web homepage: https://xingw-xiong.ac.cn

One-stage 的检测算法直接一步对目标的位置进行回归，代表性工作有：SSD [6]，YOLO [7] 等。

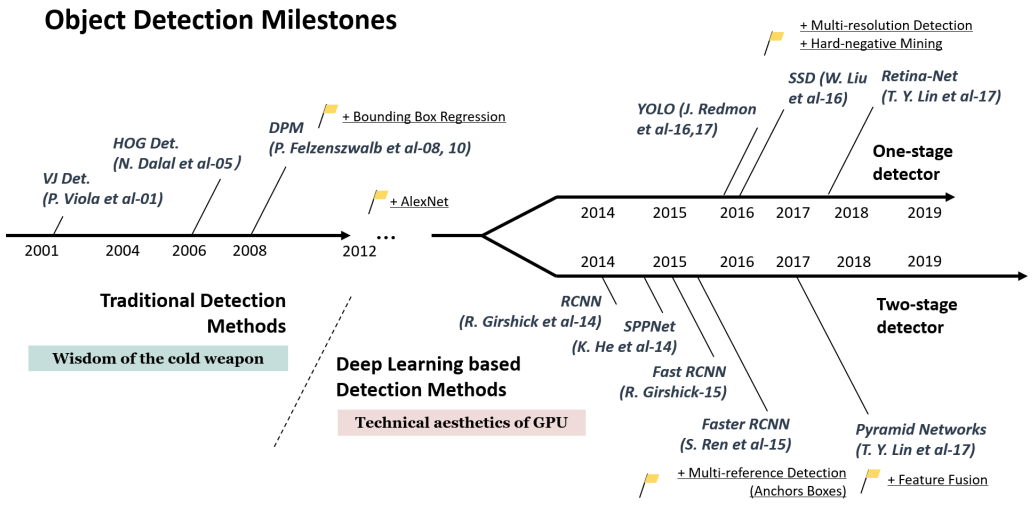

图 1: 目标检测算法发展路线，引用自 [8]

2 Two-stage 目标检测算法

Two-stage 的检测算法基于 Regional Proposal，针对每个样本生成若干的候选框，然后用卷积神经网络（CNN）对候选区域进行特征提取、分类，然后再做边框的回归。

2.1 R-CNN

R-CNN (Regions with CNN features) [1] 在 2014 年，首次提出将 CNN 架构加入到目标检测任务中，在这之前的目标检测算法，如 OverFeat，都是基于滑动窗口、HoG 检测器等方法。

R-CNN 算法的大致思路就是通过选择性搜索（Selective Search）生成一些候选框，然后把这些候选框放缩成一个固定大小的图像（227*227 的图像），并通过在 ImageNet 上预训练之后的 CNN 模型对其提取特征。最后通过线性 SVM 分类器对提取之后的特征分类。R-CNN 在 Pascal VOC 2010 数据集上的 mean Average Precision (mAP) 提升到了 53.7%。

相比于之前的目标检测算法，RCNN 虽然带来了很大的精度提高，但是运算速度是很慢的，因为存在着大量的特征冗余计算。RCNN 会在单张图片上产生超过 2K 个候选框，其中就有很多相互重叠的候选框，并且要对每个候选框单独提取特征。大量的特征的冗余计算导致 RCNN 算法在单 GPU 上每 14s 才能识别一张图像。

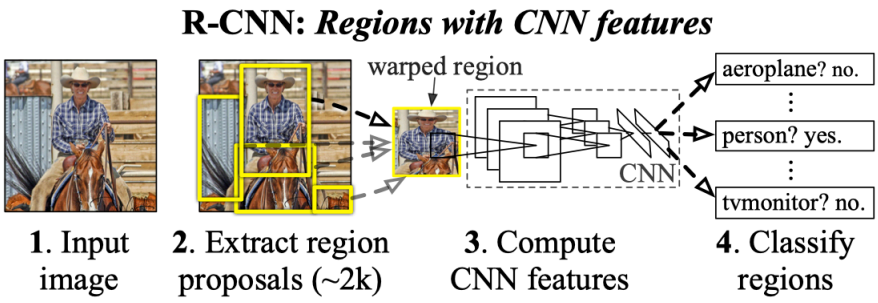

图 2: R-CNN 架构，引用自 [1]

2.2 SPPNet

R-CNN 需要对候选框进行放缩到统一大小的图像（224*224），这样不可避免给图像带来或多或少的失真。为解决这一问题，2014 年，何凯明提出了 **S**patial **P**yramid **P**ooling **Net**work (SPPNet) [2]，通过 CNN 生成一个固定长度的向量表达，而不需要对图像进行放缩。

如图3所示，SPPNet 将卷积网络提取到的任意大小的特征图划分别分成 16、4、1 块，然后对每一块做 Max Pooling，然后将池化之后的特征拼接得到一个固定大小的向量输出。

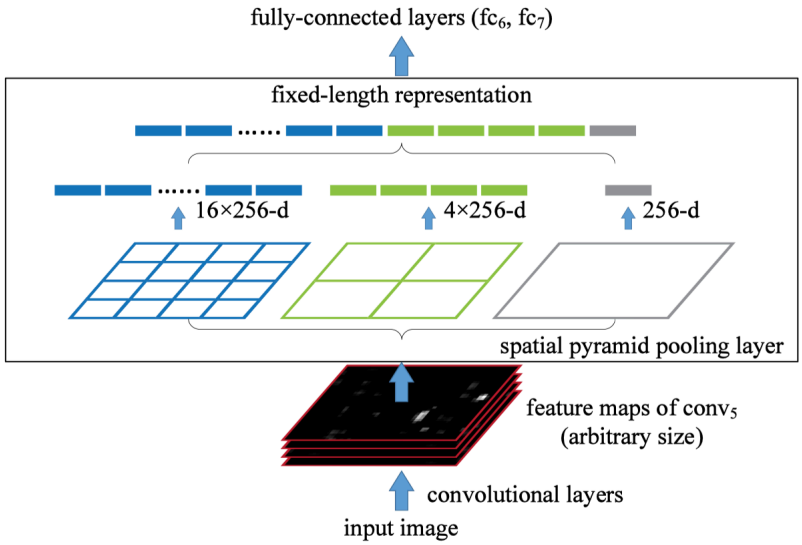

图 3: SPPNet 架构，引用自 [2]

SPPNet 在 Pascal VOC 2007 数据集上的检测速度比 RCNN 快了 24-102 倍，并且具有更好的准确率。

2.3 Fast R-CNN

不过和 RCNN 类似的是，SPPNet 既需要训练 CNN 提取特征，并需要训练 SVM 分类这些特征。因此，这需要分两步训练，而且需要耗费巨大的存储空间来存储 CNN 提取之后的特征。针对这些问题，Fast R-CNN 提出了一种一步训练的方式，同时训练检测器和边框回归器。

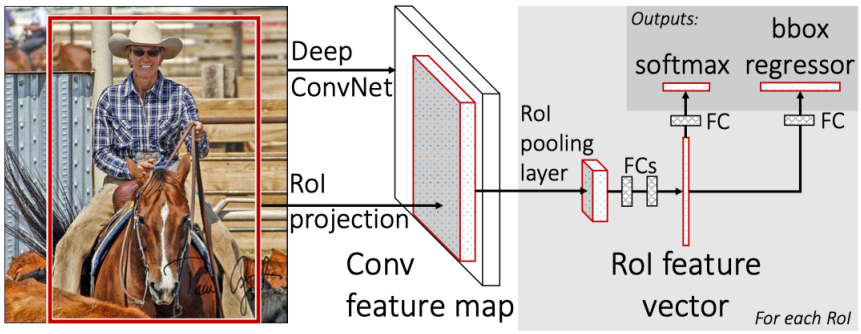

图 4: Fast R-CNN 架构，引用自 [3]

与 R-CNN 相比，Fast R-CNN 训练 VGG16 网络的速度要快 9 倍，测试速度快 213 倍；与 SPPNet 相比，Fast R-CNN 训练 VGG16 网络的速度要快 3 倍，测试速度要快 10 倍。并且，Fast R-CNN 比 R-CNN、SPPNet 具有更好的检测精度。

2.4 Faster R-CNN

Fast R-CNN 通过 Selective Search 来生成候选区域，该算法在 CPU 上运行，且比较耗时，Selective Search 也已经是 Fast R-CNN 算法的性能瓶颈。Faster R-CNN [4] 提出了 Region Proposal Network (RPN)，与监测网络共享卷积特征，实现了几乎零成本的候选区域生成。

RPN 是一个全卷积网络（Fully Convolutional Network, a.k.a., FCN），同时预测目标边界及其置信度。通过端到端的训练，RPN 生成高质量的侯选区，送到 Fast R-CNN 网络中做检测。简而言之，Faster R-CNN 用 RPN 代替了 Selective Search，因此也可以把 Faster R-CNN 看成 RPN + Fast R-CNN 检测。

Faster R-CNN 的在 GPU 的目标检测整个过程（生成候选框 + 检测分类、回归）的速度达到了 5fps，每张图片也仅生成了 300 个候选框。

2.5 Feature Pyramid Networks

Feature **P**yramid **N**etworks (FPN) [9] 由 Facebook 在 2016 年提出，主要解决的是物体检测中的多尺度问题。对于卷积神经网络而言，不同深度对应着不同层次的语义特征，浅层网络分辨率高，学的更多是细节特征，深层网络分辨率低，学的更多是语义特征，具有更好的位置信息。

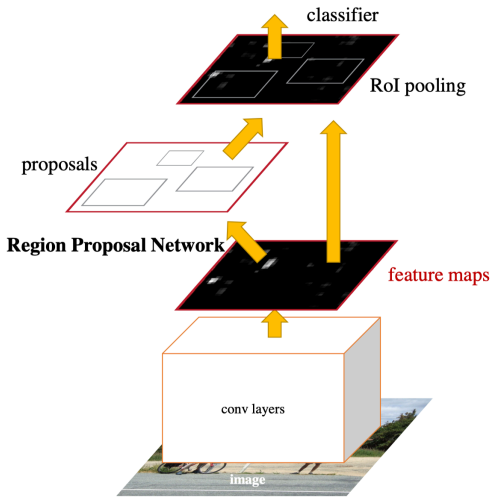

图 5: Faster R-CNN 架构，引用自 [4]

为了解决多尺度目标检测，图6a 的做法是对图像进行多尺度裁剪，缩放，并对每个尺度下的图像单独计算特征，这种做法是相当慢的，并且不实用；如图6b 所示，而一些检测系统为了更快的检测，则仅采用高层特征；图6c 所示是分别对各层特征单独预测；图6d 的方法就是 FPN，FPN 与 (b), (c) 的方法有同样的速度，但具有更好的检测精度。FPN 的思想是：首先对图像做一次前向计算提取出高层特征，然后对高层特征进行上采样，并与底层特征相加之后再进行分类预测。

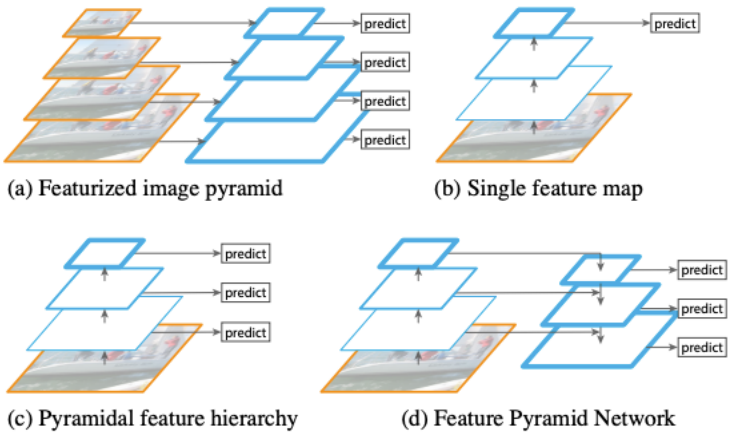

图 6: 各个多尺度目标检测方案，引用自 [9]

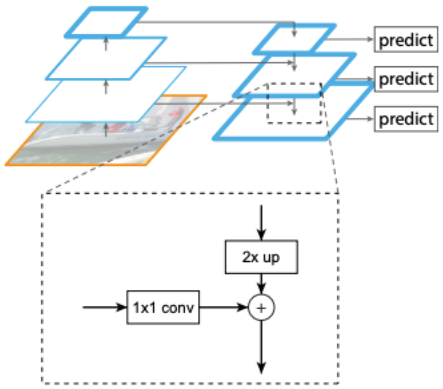

图 7: FPN 的内部连接，引用自 [9]

3 One-stage 目标检测算法

尽管 Two-stage 目标检测算法经过多次改进，Faster R-CNN 在识别精度上已经达到了很好的水平，但是在速度上不能满足实时的要求。One-stage 目标检测算法虽然在识别精度上不如 Two-stage 目标检测算法，但是在实时性要求比较高的场景下，One-stage 表现出它的不可替代性。One-stage 中的代表性算法有 YOLO [7], SSD [6]。

3.1 YOLO

2015 年，Joseph Redmon 等人提出 YOLO 算法 (**Y**ou **O**nly **L**ook **O**nce)，该算法是基于图像的全局信息进行预测，网络结构比较简单，通过将输入图像调整到一个 448*448 的固定尺寸大小，并将图像划分为 7*7 网格区域，通过 CNN 提取特征训练，直接预测每个网格内的边框坐标和每个类别置信度。

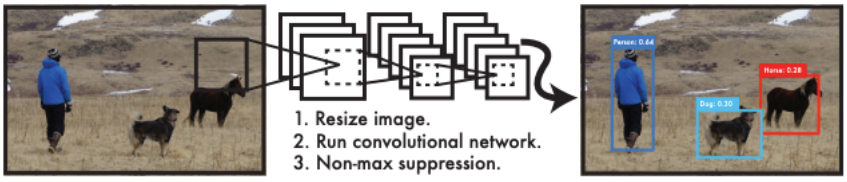

图 8: YOLO 检测系统，引用自 [7]

YOLO 检测速度快，在 GPU 上能够达到 45fps 的检测速度。但是这类算法存在的问题有：定位不准、召回率不如基于 Regional Proposal 的算法，并且当距离很近或者尺寸很小的物体检测效果不好，泛化能力相对较弱。

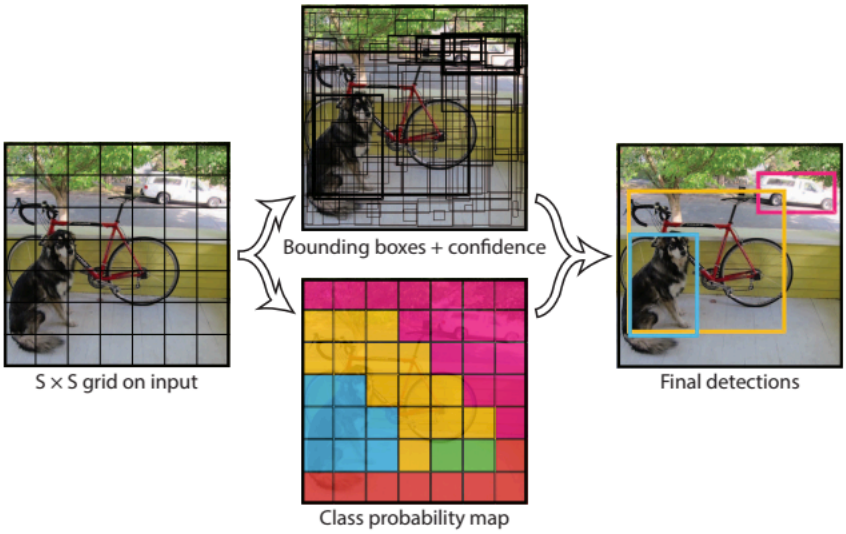

图 9: YOLO 模型，引用自 [7]

3.2 SSD

针对 YOLO 算法定位精度不准的问题，2016 年 Wei Liu 等人提出 SSD [6] 算法，它将 YOLO 的回归思想和 Faster R-CNN 的 anchor box 机制结合起来。通过在不同卷积层的特征图上预测物体区域，输出离散化的多尺度、多比例的 default boxes 坐标，同时利用小卷积核预测一系列候选框的边框坐标补偿和每个类别的置信度。

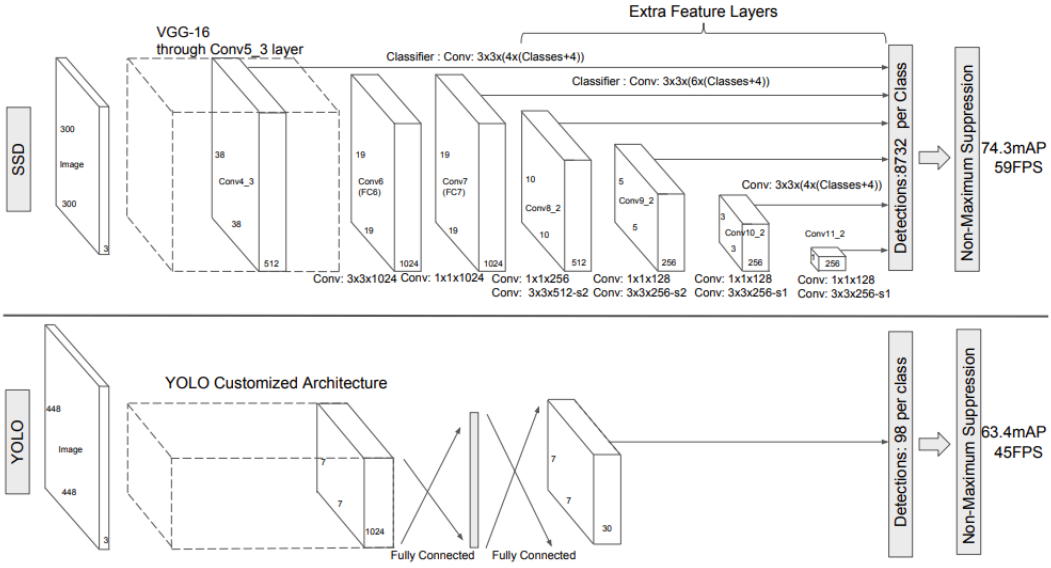

图 10: SSD 架构与 YOLO 架构比较，引用自 [6]

在整幅图像上各个位置用多尺度区域的局部特征图边框回归，保持 YOLO 算法快速特性的同时，也保证了边框定位效果和 Faster R-CNN 类似。但因 SSD 利用了多层次特征分类，导致其对于小目标依旧检测困难。最后一个卷积层的感受野范围很大，使得小目标特征不明显。

4 Conclusion

One-stage、Two-stage 两种算法各有优势，我们需要针对不同任务场景，选择不同的算法。

Low-level Vs. high-level features 对于卷积神经网络而言，不同深度对应着不同层次的语义特征，浅层网络分辨率高，学的更多是细节特征，深层网络分辨率低，学的更多是语义特征，具有更好的位置信息。

Stage Cascade in 2-stage object detection Regional Proposal 关注召回率 (Average Recall, abbr. AR)，尽可能多选出目标，所以允许我们生成大量的候选框；Detection 关注正检率 (Average Precision, abbr. AP)，尽可能让目标分类正确。

Skip architecture 跳连结构在很多 CNN 中都被用到，如 ResNet、FPN、UNet、DenseNet 等。将多层特征通过 add/concat 方式融合，在保留了深层语义特征的同时，还融合了浅层细节特征。

References

[1] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR*, abs/1406.4729, 2014.

[3] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[8] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *CoRR*, abs/1905.05055, 2019.

[9] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.