基于 Mask R-CNN 的各类主干网络应用差异分析

朱圣果 李丹
四川大学锦城学院 四川 成都 611731

【摘 要】随着基于卷积神经网络的 R-CNN 在 VOC 2012 数据集上较此前最佳结果在平均精度（mAP）上获得了 30% 以上的提升，达到 53.3% mAP[1] 后，深度学习算法自此正式步入目标检测领域，凭借较强的通用性和特征迁移能力、高精确度、低优化及维护成本等优势迅速取代传统目标检测算法，逐步发展出 Fast R-CNN[2]、Faster R-CNN[3]、Mask R-CNN[4] 等一系列衍生，此间 SPP（Spatial Pyramid Pooling）、RoI Pooling（Region of Interest Pooling）、multi-task loss、RPN（Region Proposal Network）、FPN（Feature Pyramid Network）、Roi Align[4] 等概念的提出为根据实际应用场景调整主体算法跟主干网络结构提供了基础。

【关键词】Mask R-CNN；实例分割；backbone；ResNet；FBNet

1 引言

运算速度快存储空间大的计算机能够轻松处理许多复杂事务，自诞生之初便承担着帮助或替代人类进行重复性工作的任务。在诸如登记人员的流动情况、计算层层嵌套的数学公式等枯燥易错且耗时耗力但方法清晰逻辑明了的任务中利用计算机进行自动化处理无疑当属最优先决策。然而在理解自然语言，识别图像语音等方面却远不及人类那般手到擒来，而这正是 AI 将解决的问题 [7]。

一直以来，棋都被视为顶级人类智力与 AI 的试金石。五子棋、跳棋等游戏中很早便有 AI 的身影，通常仅供休闲娱乐也非专业人士的对手。1996 年 2 月 10 日的一场国际象棋人机大战中，美国 IBM 公司研发的深蓝在与顶尖棋手加里·卡斯帕罗夫的对弈中首次获得胜局后以 2：4 比分落败。然而次年 5 月 11 日，深蓝 [8] 再战加里·卡斯帕罗夫以 2 胜 1 负 3 平取胜后，人类棋手再未在国际象棋领域的顶尖人机对弈中取胜，可 AI 与人类的比试并未停歇。

深度学习是 AI 下属分支之一，从对人脑思考方式的模仿中建立起人工神经网络，以此解析图像、声音、文本等数据。2006 年《Science》一篇用神经网络对数据降维的文章面世并掀起深度学习研究浪潮 [9] 十年后，Google 的 AlphaGo 一举将自国际象棋领域沦陷后坚守近廿年、被称作人类最后智慧堡垒的围棋阵地攻破，展现出了深度学习的极大潜力。

2 目标检测

目标检测 (object detection) 结合了分类与定位，也是图像理解和 CV 的基石。如何提高特征表达能力、抗形变能力、分类器准确度、速度是传统目标检测算法关注要点。通常采用尺度不变特征变换 (Scale-invariant feature transform，SIFT)[10]、方向梯度直方图 (Histogram of Oriented Gradient，HOG)[11]、可变形部件模型 (Deformable Parts Model，DPM)[12] 等方法予以实现，但因依赖先验知识而缺乏自适应性与泛化能力。

3 R-CNN 系列算法

3.1R-CNN

Ross Girshick 等人提出的 R-CNN 中使用多尺度滑动窗口搭配 AlexNet 提取图像特征的思想源自对 OverFeat 算法的沿用与改进。首次在目标检测中应用深度学习并取得开创性成果的同时，也显露出运算冗余等弊端，为今后的研究提供了诸多方向。

3.2 Fast R-CNN

MSRA 在 SPP-Net 中通过 SPP 解决了冗余计算后，R-CNN 设计者之一 Ross Girshick 优化了空间开销，在提出 RoI Pooling 层的概念搭配多任务损失函数（multi-task loss）并增加一层 bounding box regressor 全连接层后实现了一体化的图像特征提取检测，亦即 Fast R-CNN[2]。

公式 1 multi-task loss

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*),\quad p_i = \begin{cases}1, & \text{正 anchor}\\ 0, & \text{负 anchor}\end{cases}$$

上式中：i 代表 mini-batch anchor 的索引；p_i 是目标预测的概率；t_i 是表示 bbox4 个参数化坐标的向量；N_{cls} 是 mini-batch 的大小；N_{reg} 是 anchor 数量；λ 是归一化平衡权重 [13]。

3.3 Faster R-CNN

不同于着重优化 R-CNN 后部构造的 Fast R-CNN，Faster R-CNN 目标在前部。通过 RPN(Region Proposal Networks) 提取候选区以取代 SS(Selective Search) 方法，组成了能端到端完成检测任务的 FCN(Fully Convolutional Network)。

3.4 Mask R-CNN

实例分割近似目标检测加语义分割，需在像素尺度识别目标轮廓。Faster R-CNN 与 Fast R-CNN 同是采用 RoI Pooling 池化来获取候选区域，其对边缘像素进行四舍五入产生的像素偏差无疑会对 bbox 的定位造成干扰。与目标检测不同，实例分割中这种干扰是不可忽视的。

对此， Facebook AI 提出的 Mask R-CNN 用 RoI Align 取代 RoI Pooling，通过双线性插值方法处理小数像素，实现了特征图与原始图像的精确对应。并在 Faster R-CNN 的基础上增加了与分类和回归网络并行的 FCN 层生成 binary mask 来判断各像素是否属于目标。

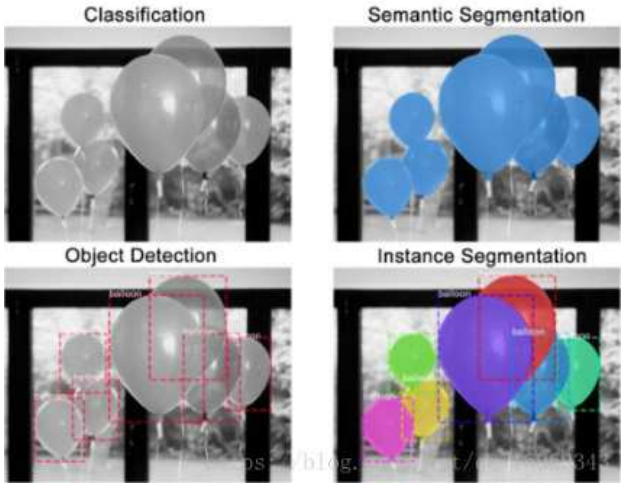


图 1 图像分类、语义分割、目标检测、实例分割

4 主干网络 Backbone

特征提取是深度学习目标检测算法的重要步骤，负责特征提取的则是 CNN 中的 backbone，算法的耗时程度基本取决于 backbone 的复杂度。提到目标检测 one-stage 方法中 SSD(Single Shot MultiBox Detector) 算法和 YOLO(You Only Look Once) 算法时，通常指 backbone 为 VGG16 的 SSD 和 darkNet 的 YOLO。因场景的限制或需求不同也有别的搭配，如 MobileNet-SSD[1]，Resnet-SSD，RefineDet-SSD[2]、MobileNet-YOLO[3] 等。本文对 ResNet、FBNet 跟 Mask R-CNN 的组合进行了实验分析。

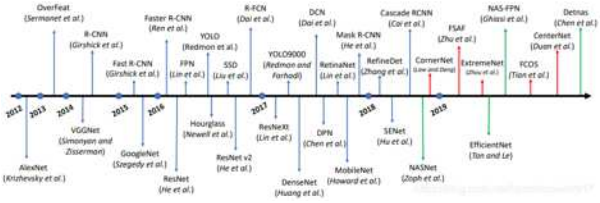


图 2 神经网络和 backbone 发展历程

Resnet-FPN

2015 年何凯明提出的残差网络 ResNet[14] 是图像识别的里程碑，首次使识别准确率超越人类。残差块的设计在预防梯度消失 / 爆炸的同时还丰富了各层网络包含的图像信息。2016 年特征金字塔网络 (Feature Pyramid Network，FPN) 将自顶向下的上采样结果与自底向上的 feature map 融合，解决了底层特征位置信息精准而语义信息不足跟顶层特征语义信息丰富而位置信息模糊的矛盾。这种通用架构结合 ResNet 后也应用于 Mask R-CNN。

FBNet

FBNet[15] 随着神经架构搜索 (Neural Architecture Search，NAS) 的发展而出现。此前各 CNN 的网络结构往往采用现成的设计，而静态网络结构有局限性，针对场景单独设计又很困难，NAS 便应运而生 [16]。FBNet 中网络结构的动态搜索策略是将搜索空间初始化为不同层组成的大图，利用可微的随机优化方法结合 Gumbel 采样技巧进行的。

5 实验设计与实现

纹理在图像中普遍存在又难以描述，作为重要的视觉线索，对纹理分类与分割的研究经久不衰 [17]，为试验拥有像素级实例分割能力的 Mask R-CNN 能否进行图像纹理特征提取，本实验将在 Kaggle 的 Dogs vs Cats 无标签图片集基础上手动针对毛皮进行标注后用作实验数据集。

实验由三部分构成：(1) 数据集构建，挑选无标签图片集并标注目标信息，进行随机旋转等预处理降低噪音干扰。(2) 模型训练，将训练集喂给各主干网络，按统一参数优化器 (solver) 训练后保存各参数模型及日志文件。(3) 用测试集进行测试并统计数据；用各模型单独预测无标签图片并保存。对比两种结果结合日志文件中训练过程参数变化进行分析总结。

5.1 数据集构建

为试验 Mask R-CNN 的图像纹理特征提取能力以及确保数据集中图片与标注信息准确有效，实验选用含 375000 张图片的 Kaggle 的 Dogs vs Cats 无标签图片集。剔除损坏文件后分别随机选取 cat/dog 各 500 张用于实验，

1 MobileNet-SSD https://github.com/chuanqi305/MobileNet-SSD.
2 RefineDet https://github.com/sfzhang15/RefineDet.
3 MobileNet-YOLO https://github.com/eric612/MobileNet-YOLO.

使用 labelme 软件针对各图片中具有毛皮纹理特征的猫狗目标进行手动标记后得到 1000 份 json 标注文件。

从中选取 cat/dog 各 50 张作测试集，其余作训练集后，将 900 份训练集与 100 份测试集图片的相应 json 文件分别转换成两个 coco2014 标注格式的训练 json 和测试 json 以备训练。

实验数据集总计 1000 张含标注图片，其中猫狗比例 1：1，训练测试比例 9：1。

5.2 模型训练

模型训练选用 Ubuntu 18.04 操作系统，Pytorch 1.2 环境，maskrcnn_benchmark 框架，Kaggle 的 Dogs vs Cats 无标签图片集子集， coco2014 格式自制图片标注信息，单 GPU 训练。训练所用模型配置文件采用 maskrcnn_benchmark 提供的 ① e2e-mask-rcnn-fbnet、② e2e-mask-rcnn-fbnet-600、③ e2e-mask-rcnn-fbnet-xirb16d-dsmask、④ e2e-mask-rcnn-fbnet-xirb16d-dsmask-600、⑤ e2e-mask-rcnn-R-50-C4-1x、⑥ e2e-mask-rcnn-R-50-FPN-1x、⑦ e2e-mask-rcnn-R-50-FPN-1x-periodically-testing、⑧ e2e-mask-rcnn-R-101-FPN-1x、⑨ e2e-mask-rcnn-X-101-32x8d-FPN-1x。

基于变量控制与算力限制等考量，实验中各主干网络使用相同参数的优化器进行训练。其中最大迭代次数 MAX_ITER=10000，学习率 BASE_LR=0.0025，每批图片数 IMS_PER_BATCH=1，权重衰减项 WEIGHT_DECAY=0.0001，预热学习率 WARMUP_FACTOR=1/3，预热轮数 WARMUP_ITERS=500。

6 实验结果与分析

经完整一轮实验发现 FBNet 系列网络对 bbox-(bounding box) 及 segm(mask) 的 AP 基本为 0 后，意识到不同于 ResNet 拥有现成的网络结构和初始权重，FBNet 额外需要动态探索适宜的网络结构，加之初始学习率相对较小，因此 1 万轮训练远不足以获得有价值的预测模型。所以紧接着 checkpoint 对其追加进行了 19 万次训练，记作⑩ FBNet(+)。

表 1 Bounding Box 平均精度

Backbone	AP	AP50	AP75	APS	APM	APL
① FBNet	0	0	0	—	0	0
② FBNet-600	0	0	0	—	0	0
③ FBNet- xirb16d-dsmask	0	0	0	—	0	0
④ FBNet- xirb16d-dsmask-600	0.0001	0.0004	0	—	0	0.0001
⑤ R-50-C4	0.5531	0.9382	0.6008	—	0	0.5587
⑥ R-50-FPN	0.5766	0.9399	0.6868	—	0	0.5834
⑦ R-50-FPN-periodically-testing	0.5206	0.9308	0.5262	—	0	0.5265
⑧ R-101-FPN	0.5804	0.9542	0.6970	—	0.0750	0.5853
⑨ X-101-32x8d-FPN	0.5670	0.9567	0.6343	—	0	0.5723
⑩ FBNet(+)	0.1145	0.1925	0.1199	—	0	0.1160

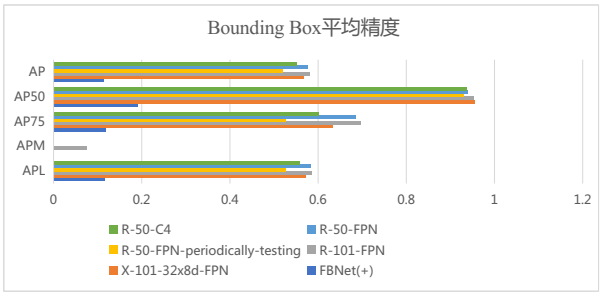
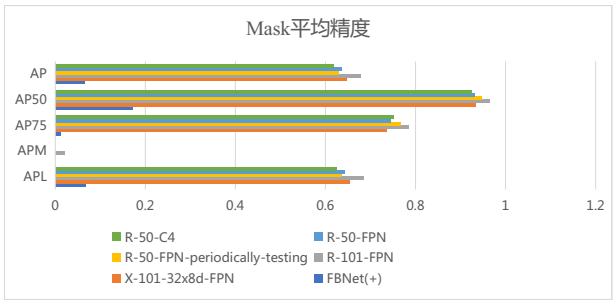


表 2 Mask 平均精度

	AP	AP50	AP75	APS	APM	APL
① FBNet	0	0	0	—	0	0
② FBNet-600	0	0	0	—	0	0
③ FBNet- xirb16d-dsmask	0	0	0	—	0	0
④ FBNet- xirb16d-dsmask-600	0.0002	0.0004	0	—	0	0.0002
⑤ R-50-C4	0.6191	0.9256	0.7528	—	0	0.6255
⑥ R-50-FPN	0.6368	0.9308	0.7455	—	0	0.6440
⑦ R-50-FPN-periodically-testing	0.6291	0.9481	0.7678	—	0	0.6364
⑧ R-101-FPN	0.6781	0.9655	0.7847	—	0.0222	0.6846
⑨ X-101-32x8d-FPN	0.6471	0.9344	0.7353	—	0	0.6541
⑩ FBNet(+)	0.0666	0.1722	0.0130	—	0	0.0677




```
Average Precision (AP):
  AP                % AP at IoU=.50:.05:.95 (primary challenge metric)
  AP^IoU=.50        % AP at IoU=.50 (PASCAL VOC metric)
  AP^IoU=.75        % AP at IoU=.75 (strict metric)
AP Across Scales:
  AP^small          % AP for small objects: area < 32^2
  AP^medium         % AP for medium objects: 32^2 < area < 96^2
  AP^large          % AP for large objects: area > 96^2
```

图 3 模型性能评价指标

分析测试数据发现模型⑧预测效果最佳。另因 Dogs vs Cats 是以猫狗为主体的图片集，所以关于标注信息中目标大小中没有 small、少有 medium，致使 APS 不存在且 APM 基本为 0。分析配置文件得知⑤使用 “R-50-C4” 卷积体，⑥⑦使用 "R-50-FPN" 卷积体，⑧⑨使用 "R-101-FPN" 卷积体。

两两比较得知：(1) ⑥⑦中⑦每 2500 步会中断一次训练进行测试，致使总体精度略低。(2) ⑧⑨中⑧初始权重存于 R-101.pkl；STRIDE_IN_1X1=True(获取 feature map 使用步长为 2 的 1x1 滤波器，适用于 original MSRA ResNet)；组正则数 NUM_GROUPS=1(适用于 ResNet 及 ResNeXt); 各组通道数 WIDTH_PER_GROUP=64，⑨初始

权重存于X-101-32x8d.pkl；STRIDE_IN_1X1=False(适用于C2 and Torch models)；NUM_GROUPS=32；WIDTH_PER_GROUP=8。且⑨默认最大迭代数与学习率衰减阈均为⑧两倍，由于本实验统一了优化器作参数，⑧训练了理想次数的九分之一，⑨训练了理想次数的十八分之一，所以⑧优于⑨。(3) ⑥⑧的卷积体与初始权重不同，⑥存于R-50.pkl。因参数相近、层数更深所以⑧优于⑥。(4)⑤⑥除卷积体不同外⑤输出层的输出通道数BACKBONE_OUT_CHANNELS=256*4；未使用FPN获取feature map，⑥ BACKBONE_OUT_CHANNELS=256；使用FPN获取feature map。事实证明，FPN综合高低各层语义信息和位置信息的处理对于算法精度的提升是确实存在的。

7 预测结果可视化分析

以下同种图像均按第一排：⑤、⑥、⑦，第二排：⑧、⑨、⑩排列。

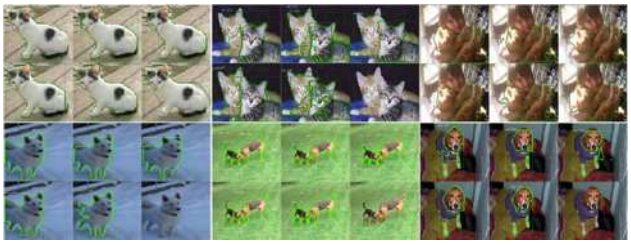


图4 各模型预测效果对比

观察分析各模型预测效果后发现：⑤训练程度不足，在部分易混淆的像素区域无法进行很好的区分。⑥⑦势均力敌，均能得出较为准确的预测结果。⑧预测效果最佳，对于猫狗身体的边缘覆盖十分细致，能够很好地避开遮挡物，但对于复数猫狗的检测存在部分过拟合现象。⑨的预测中存在较多过拟合。虽⑩经20万次训练后终于能得出大于70%IoU的预测，但依然有许多图片预测失败。

8 结束语

作为探索性参照组的FBNet因其特殊特性，结合实验结果发现此类不具备具体网络结构的backbone不适用于此类小规模任务，在经过其他网络二十倍的训练次数后，AP也仅在0.1附近徘徊，表明了NAS是个极其复杂的过程。虽然对于许多图片无法进行预测，但每一个预测出的bbox及mask对于相应图片而言，预测效果都比较准确，在极低的平均精度下有这等效果充分展现NAS类算法具有的极大潜力，在更复杂的任务中应该更能发挥其价值。对比ResNet系列主干网络的实验发现引入了SPP、FPN这类开创性思路的网络在精度上都有明显提升，但同时网络层数也不是越多越好，在小规模的数据集上一味增加网络深度容易导致精度不升反降。实验证明拥有像素级实例分割能力的Mask R-CNN确实可以提取图像纹理特征。使用绕开了各类障碍物并针对被毛发这类相同纹理包围的猫狗的身体进行标注的数据集所训练出的Mask R-CNN模型在预测有手臂、衣物等外物遮挡的图像时，能够很好地绕开障碍标注目标、并将表层看来四分五裂的身体视作同一目标准确预测的结果充分表明了Mask R-CNN能够实现从高层语义特征中提取图像纹理特征的功能。

【参考文献】

[1]R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524, 2013.

[2]Girshick R . Fast R-CNN[J]. Computer Science, 2015.

[3]Ren S , He K , Girshick R , et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(6).

[4]Kaiming H , Georgia G , Piotr D , et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018:1-1.

[5]He K , Zhang X , Ren S , et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16.

[6]Lin T Y , Doll á r, Piotr, Girshick R , et al. Feature Pyramid Networks for Object Detection[J]. 2016.

[7] 朱福喜 , 汤怡群 , 傅建明 . 人工智能原理 [M]. 武汉 : 武汉大学出版社 , 2002.

[8]Campblls M , Jr A J H , Hsu F H . Deep Blue[J]. Artificial intelligence, 2002, 134(1/2).

[9]Hinton G , Salakhutdinov R . Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786):p. 504-507.

[10]Lowe D G . Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[11]DALAL,N. Histgrams of oriented gradients for human detection[J]. proc of cvpr, 2005.

[12]Felzenszwalb P F , Mcallester D A , Ramanan D . A Discriminatively Trained, Multiscale, Deformable Part Model[C]// 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. IEEE, 2008.

[13] 周晓彦 , 王珂 , 李凌燕 . 基于深度学习的目标检测算法综述 [J]. 电子测量技术 , 2017(11):89-93.

[14]He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. 2015.

[15]Wu B , Dai X , Zhang P , et al. FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search[J]. 2018.

[16]Zoph B , Le Q V . Neural Architecture Search with Reinforcement Learning[J]. 2016.

[17] 刘丽 , 匡纲要 . 图像纹理特征提取方法综述 [J]. 中国图象图形学报 , 2009(04):63-76.