人工神经网络原理文献报告

17363092 叶茂青

June 23, 2020

1 Rich feature hierarchies for accurate object detection and semantic segmentation[1]

提出了 R-CNN 这一网络架构，在传统的 CNN 架构上加入 Region Proposal 的结构，从而完成图像目标检测的任务。R-CNN 完成图像目标检测的步骤如下：

1. 使用 Region proposals 在原图像中取出大约 2000 个区域
2. 使用一个 CNN 网络将提取出来的区域变为一个固定长度的特征
3. 使用 SVM 进行分类

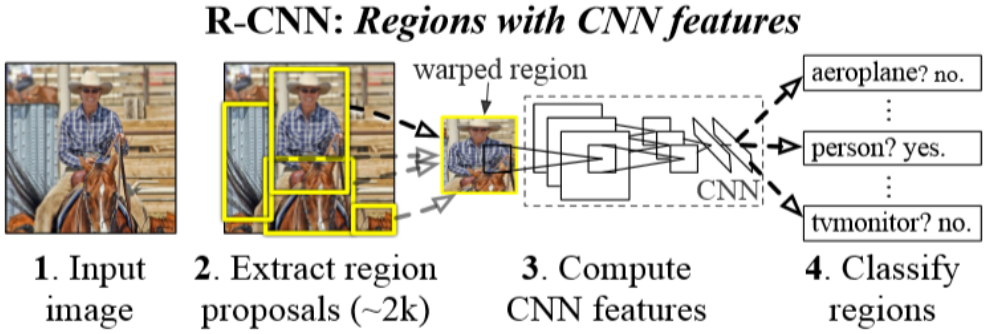

Figure 1: R-CNN architecture

第一步 Region proposals 的方法采用 Selective search，第二步使用预训练的 Alexnet 对输入为 227×227 的 RGB 图片抽取 4096 维的特征向量，最后采用 SVM 进行分类，并通过回归的方法修正 Bounding box 的误差。

R-CNN 的一些缺点：

1. Region proposals 提取的区域会有重叠，送入 CNN 网络时会有很多重复计算

2. CNN 网络需要固定尺寸的图片

3. 网络分为了多个过程，训练过程复杂

2 Spatial pyramid pooling in deep convolutional networks for visual recognition[2]

对于前面所提到的 R-CNN 的缺点 1、2 作出了一点改进，提出了一种叫做 Spatial pyramid pooling network 的结构。

对于第一个缺点，通过映射关系，从 featrue map 中找到对应 region 产生的 feature，从而避免重复的计算，只需要将整张图片经过一次 CNN 网络即可。

对于第二个缺点，R-CNN 中需要对裁剪的图像重新进行处理的原因是 Alexnet 最后为全连接层，需要保证输入的维数固定，Spatial pyramid pooling network 通过将最后一层卷积层的输出，经过池化操作变为固定长度的特征向量，从而让网络可以接受任意大小的图片。

如 Figure 2 中所示，在整幅图像的卷积结果上找到对应 region 产生的 feature，Spatial pyramid pooling network 将 feature 分割成 16 份、4 份、1 份，再使用 Max Pooling 对每一份内的特征进行池化，最后拼接起来形成 $21 \times dimension$ 的特征向量作为全连接层的输入。

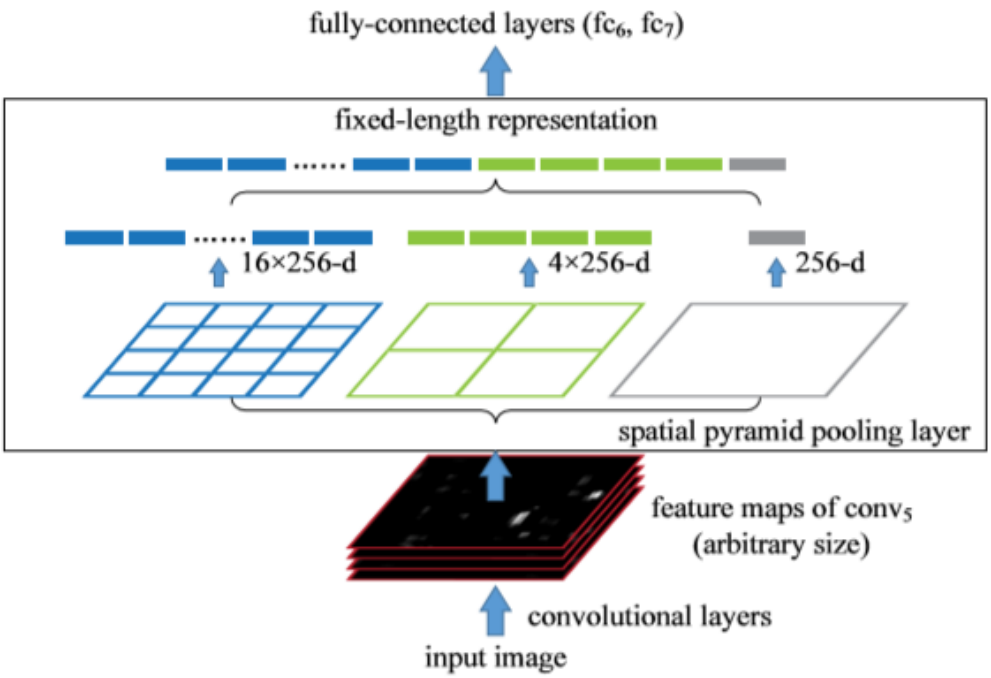

Figure 2: A network structure with a spatial pyramid pooling layer

3 Fast r-cnn[3]

R-CNN 需要把 Region proposals 选出的 2000 个区域都送入 CNN 进行计算，极大的降低了目标检测的效率，Fast R-CNN 受 Spatial pyramid pooling network 的启发，在特征图上进行 Region proposals。同时抛弃了 R-CNN 使用 SVM 进行分类的做法，引入 multi-task loss，直接用 Softmax 对分类进行预测。

Fast R-CNN 的想法与 Spatial pyramid pooling network 中的类似，相比于 Spatial pyramid pooling network 中将 feature 做 4×4、2×2、1×1 的分割，Fast R-CNN 提出的 RoI pooling layer 直接将 feature 划分为 7×7 的网格并做 Max Pooling。

其他的 trick 包括使用 smoothL1 loss 计算 bounding box 误差，使用 SVD 分解加快推测速度，采用更大的 CNN 网络。

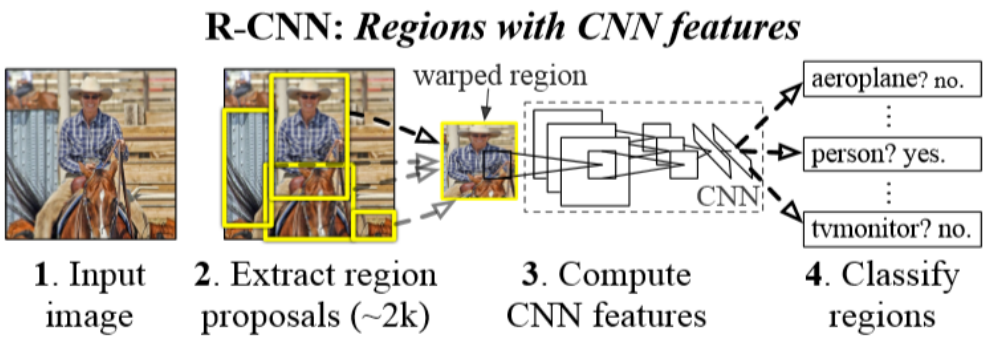

Figure 3: Fast R-CNN architecture

4 Faster r-cnn: Towards real-time object detection with region proposal networks[4]

Fast R-CNN 并没有完全做到 end-to-end 的训练，虽然只需要过一次 CNN 网络，但 Region proposals 部分依旧需要很长的时间，Faster R-CNN 将 Region proposals 部分也融入到网络中，极大的提高了运算效率。相比之前的 Selective search 方法，Faster R-CNN 引入了 Region Proposal Networks 来完成 Region Proposal 的任务。

Region Proposal Networks 取最后一层卷积层的输出作为输入，用 sliding window（论文设定为 3×3）滑过 feature map，对于每一个中心点建立 k 个 region，再交由后面的 cls layer 分辨是否存在物体，reg layer 修正 region 的范围，由于实际中不包含物体的 region 较多，为了保证正负样本的均衡，在每副图像上随机采样 256 个 anchors 计算 loss，如果正样本少于 128 个，则对负样本进行补充。

5 You only look once: Unified, real-time object detection[5]

YOLO 网络的基本想法：

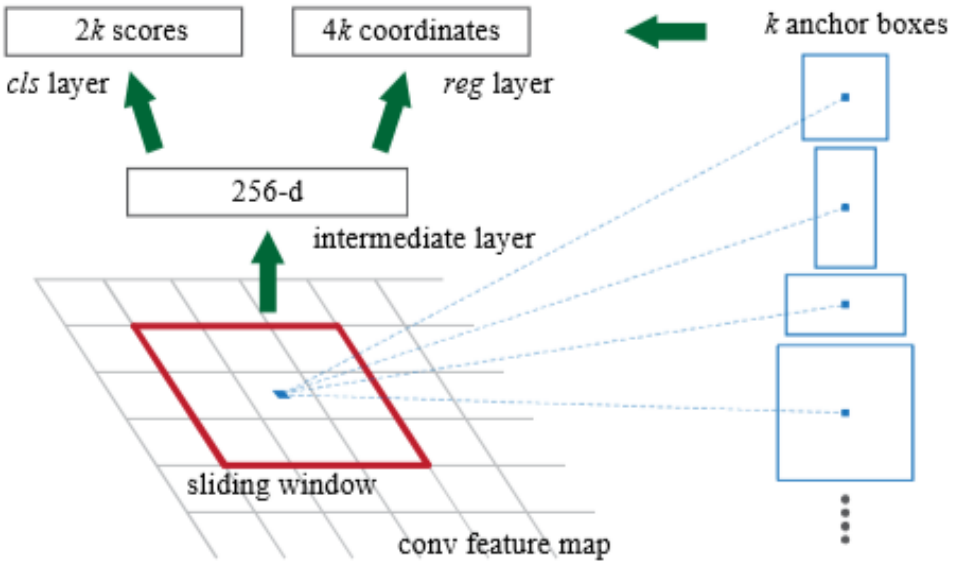

Figure 4: Region Proposal Network

系统将输入图片分为 $S*S$ 的网格单元。如果物体的中心落入某个格子，那么这个格子将会用来检测这个物体。每个网格单元会预测 B 个 bounding box 以及这些框的置信值。每个 bounding box 会有 5 个预测值：x,y,w,h 和置信值 confidence

$$confidence = Pr(Object) * IOU_{pred}^{truth} \tag{1}$$

如果 object 落在一个网格单元里，$Pr(Object)$ 取 1，否则取 0。检测评价函数 intersection-over-union：

$$IOU = \frac{DetectionResult \cap GroundTruth}{DetectionResult \cup GroundTruth} \tag{2}$$

每个网格单元也预测 C 个条件类概率，$Pr(Class_i|Object)$，在一个网格单元包含一个物体的前提下，它属于某个类的概率。我们只为每个网格单元预测一组类概率，而不考虑框 B 的数量。在测试的时候，通过如下公式来给出对某一个 box 来说某一类的 confidence score：

$$Pr(Class_i|Object) * Pr(Object) * IOU_{pred}^{truth} = Pr(Class_i) * IOU_{pred}^{truth} \quad (3)$$

然后设定阀值，过滤掉得分低的，对保留的 boxes 进行 NMS 处理，得到最终的检测结果。

网络存在的问题：

1. 由于输出层为全连接层，只能检测与训练图像相同分辨率的图像。

2. 虽然每个格子可以预测 B 个 bounding box，但是最终只选择只选择 IOU 最高的 bounding box 作为物体检测输出，即每个格子最多只预测出一个物体。当物体占画面比例较小，如图像中包含畜群或鸟群时，每个格子包含多个物体，但却只能检测出其中一个。

3. 对于 bounding box 的误差采用平方误差，没有衡量 bounding box 的大小。

6 Ssd: Single shot multibox detector[6]

SSD 改进了 YOLO 中一个网格只能预测一个物体的缺点，同时改进网络结构，引入多尺度训练，在保证模型速度的同时提高了模型的准确度。如 Figure 5 所示，相比 YOLO，SSD 加入了更多的卷积层，并使用多个不同层次的卷积层做预测，同时用卷积层代替了中间的全连接层，提升了网络的速度。对于不同层的特征图，SSD 设置的 box 会有不同，靠前的特征图用于检测小物体，靠后的特征图用于检测大物体，一定程度上解决了多尺度目标检测的问题。

对于正负样本差异的问题，SSD 通过 confidence loss 选择样本，使得正负样本比例最低为 1:3。其他的 trick 包括数据增强、引入更大 COCO 数据集，增大输入图片大小等。

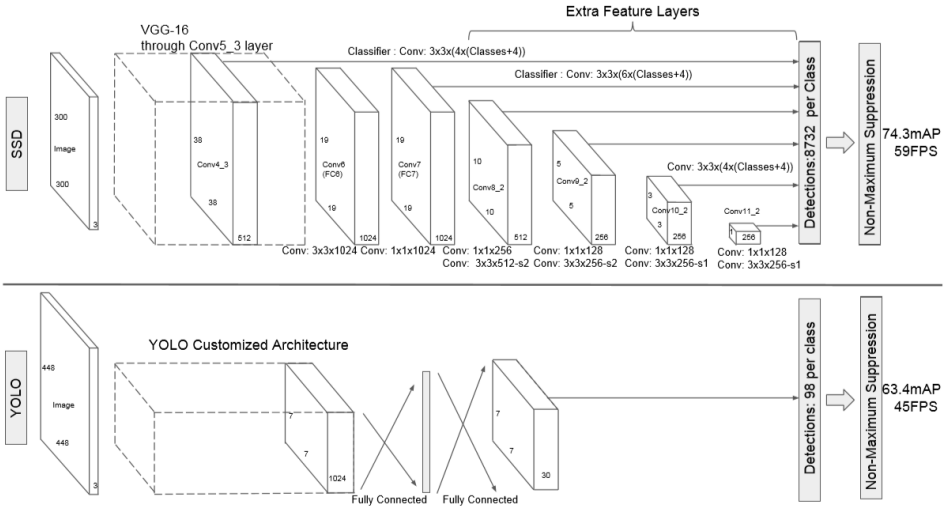

Figure 5: A comparison between two single shot detection models: SSD and YOLO

7 图表对比

Architecture	Performance(VOC 2007)	Speed
R-CNN	59.2mAP	∼15s/image
SPP Net	60.9mAP	∼0.4s/image
Fast R-CNN	70.0mAP	∼0.3s/image
Faster R-CNN	73.2mAP	7FPS
YOLO	63.4mAP	45FPS
SSD	74.3mAP	58FPS

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[3] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.