人工神经网络原理文献报告

叶茂青

June 22, 2020

用中文写一篇文献报告。文献报告要求突出但不限于几个要点：文章内容简述，主要贡献，个人理解与体会。并且自选多个角度，用图表形式对这些文章进行对比分析

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

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. 使用了 SVM 进行分类

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

对于 R-CNN 中对图像进行裁剪后重新扩展到 227 × 227 作出了一点改进，提出了一种叫做 Spatial pyramid pooling network 的结构。

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 如上图，Spatial pyramid pooling network 通过将

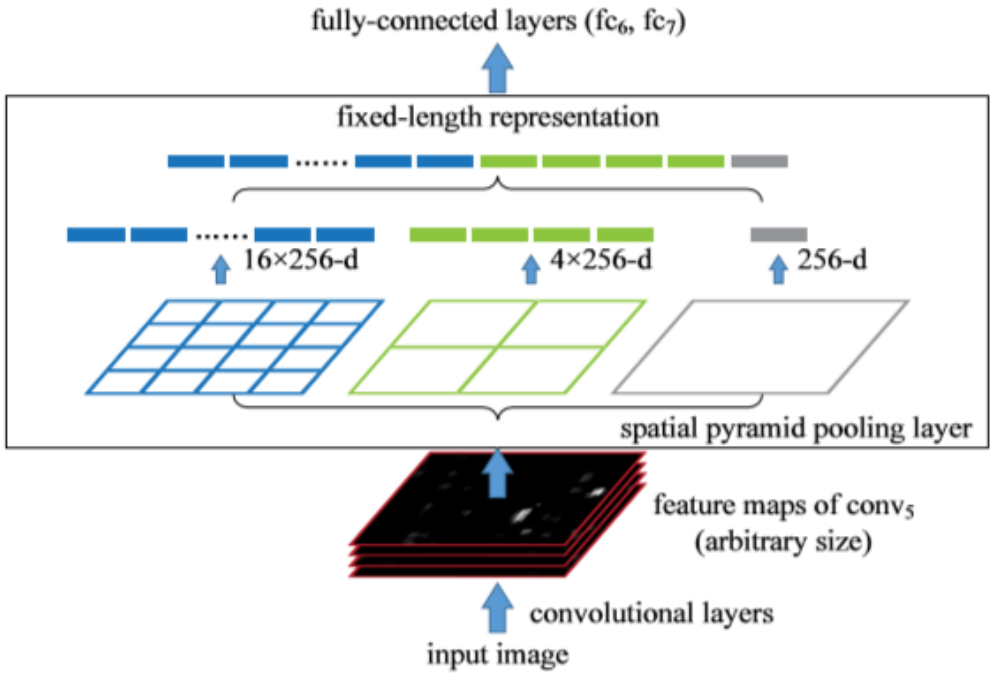

Figure 1: A network structure with a spatial pyramid pooling layer

最后一层卷积层的输出分割成 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

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，直接用 Softmax 对分类进行预测，并引入 multi-task loss。

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。

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

YOLO 网络的基本想法：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：

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

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。

6 Ssd: Single shot multibox detector[6]

xxxxxxxx

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.