Rich feature hierarchies for accurate object detection and semantic segmentation.

论文：https://arxiv.org/pdf/1311.2524.pdf
源码：https://github.com/rbgirshick/rcnn

开篇膜拜RBG大神。

作为RBG一系列论文（RCNN, Fast-RCNN [1], Faster-RCNN [2], YOLO [3] ）在物体检测方向的开山之作，RCNN在时间和精度上虽然不如后面的几篇论文，但后几篇论文乃至其他作者的一些论文例如SPPNet [4]都是继承自这篇论文的思路，其意义也是不言自明的。

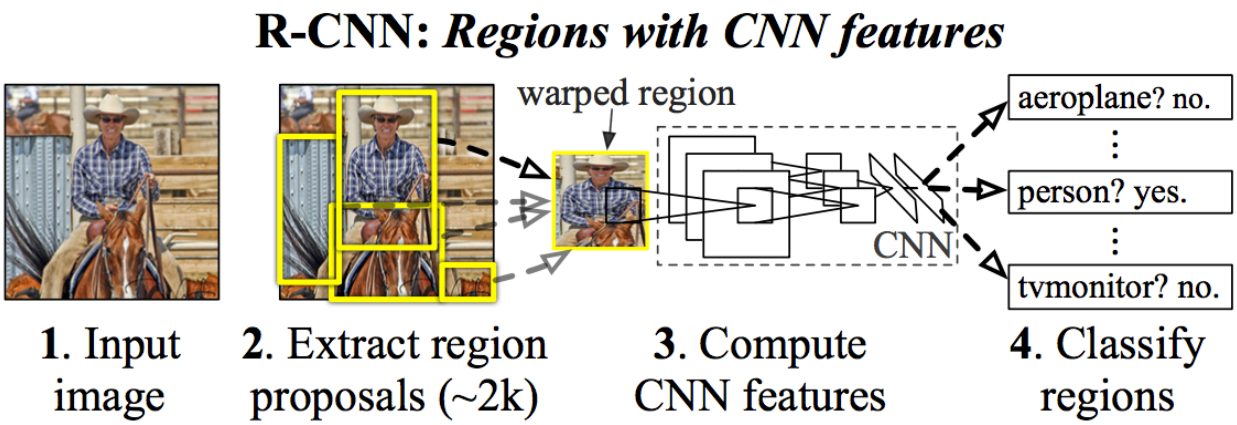
1. 主要贡献

按照论文中所说，作者认为自己的贡献主要有两点：

1. 使用CNN进行精确的目标定位与分割。 (Apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects.)
2. 当训练数据数量不足的时候，可以使用预训练加微调的方式。(When labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost.)

2. RCNN训练

RCNN的训练流程相对比较粗糙。如图1所示，主要分成四步：



1. 使用Selective Search [5]选择候选区域；
2. 使用CNN提取每个候选区域的特征；
3. 使用SVM对每个候选区域进行分类；
4. 使用回归器对位置进行精调。

2.1. Selective Search

Selective Search 的流程：

1. 使用 [6]的方法，将图像分成若干个小区域
2. 计算相似度，合并相似度较高的区域，直到小区域全部合并完毕
3. 输出所有存在过的区域，即候选区域

如图2代码：

Algorithm 1: Hierarchical Grouping Algorithm

Input: (colour) image
Output: Set of object location hypotheses L

Obtain initial regions $R = \{r_1, \cdots, r_n\}$ using [13]
Initialise similarity set $S = \emptyset$
foreach *Neighbouring region pair* (r_i, r_j) **do**
  Calculate similarity $s(r_i, r_j)$
  $S = S \cup s(r_i, r_j)$

while $S \neq \emptyset$ **do**
  Get highest similarity $s(r_i, r_j) = \max(S)$
  Merge corresponding regions $r_t = r_i \cup r_j$
  Remove similarities regarding r_i : $S = S \setminus s(r_i, r_*)$
  Remove similarities regarding r_j : $S = S \setminus s(r_*, r_j)$
  Calculate similarity set S_t between r_t and its neighbours
  $S = S \cup S_t$
  $R = R \cup r_t$

Extract object location boxes L from all regions in R

图2：Selective Search 伪代码

区域的合并规则是：

1. 优先合并颜色相近的
2. 优先合并纹理相近的
3. 优先合并合并后总面积小的
4. 合并后，总面积在其BBOX中所占比例大的优先合并

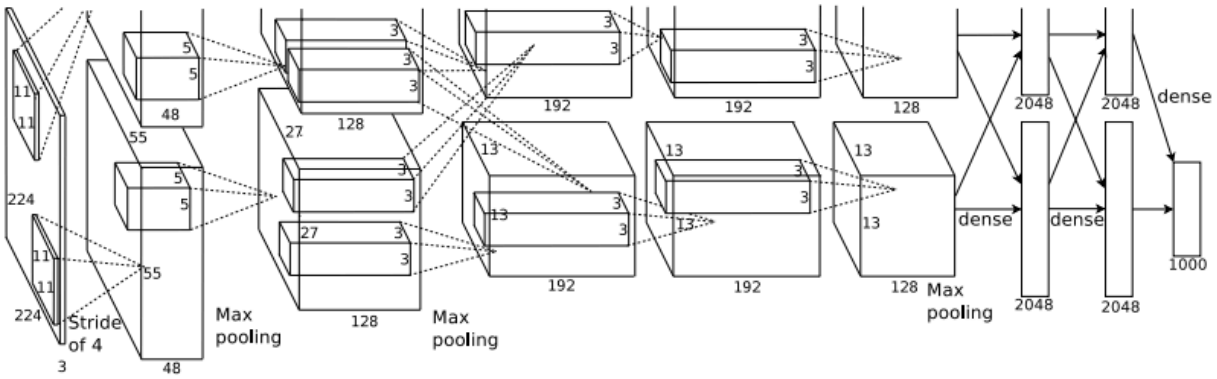
2.2. CNN

2.2.1.候选区域预处理

为了能够进行特征提取，论文中提出要把候选区域归一化为227×227像素点。论文中采用的直接wrap的方法。作者也在附录A中讨论了例如加入padding等各种方法，这些不是RCNN的重点，不进行深入探讨。

2.2.2.网络结构

论文采用了Hinton在2012年NIPS上使用的经典的Alex-Net网络结构 （图3） 。同样的，作者也对其他网络例如ZF，T-Net进行了尝试。Alex-Net的输入时一张图像，输出的是这个图像的4096维的特征向量



2.2.3 训练

作者指出，由于存在训练数据样本不够的问题，作者采用了在大数据集上（ILSVRC）进行预训练并在特用数据集(PASCAL_VOC)上进行微调的训练方式。

（1）监督预训练 （Supervised pre-training）

这个过程又叫做预训练，主要是为了初始化网络的权重。使用的是ILSVRC2012数据集进行分类网络的训练。此阶段的分类精度并不是最重要的

（2）领域微调 （Domain-specific fine-tuning）

数据：论文中使用的数据是2.1中提取出来的候选区域。训练一个N+1（N个不同物体和1个背景类）的分类器。如果候选区域与某类的Ground Truth的重合部分（IOU）的比例大于0.5，则被视为该类的正样本，否则的话则视为背景。

参数：使用的是SGD，学习率是预训练的1/10, mini-batch中32个正样本，96个负样本。

2.3. SVM

SVM是一个特定类别的线性分类器，此处的输入时CNN提取的4096维的特征向量，用来判断该候选区域是否属于该类。

数据: 与CNN分类器不同的是，SVM中的候选区域IoU大于0.3时，便被设为正样本。为什么选择0.3，以及为什么选择SVM而非Softmax作为分类器在论文的附录B中给予了阐述。

训练：由于训练数据量太大，为了节约内存，SVM使用的是[7] (DPM) 中所描述的standard hard negative mining method。

2.4. 位置精修

为了精确确定候选区域的位置，作者使用了DPM [7] 中提到的bounding-box regression对位置进行校正 (附录C)。

精修过程是一个四个输出（x, y, w, h）的线性脊回归器，输入是conv5层的4096维特征。bounding-box regression提升了约3-4个点的mAP。

3. RCNN测试

1. 使用Selective Search在图像中提取2000个候选区域；
2. 候选区域归一化到227×227；
3. 输入训练好的CNN中，得到该候选区域的4096维特征向量；
4. 使用SVM对特征向量进行打分；
5. 使用非极大值抑制（NMS）去除相交的多余的框；

3.1.非极大值抑制

非极大值抑制（NMS）先计算出每一个bounding box的面积，然后根据score进行排序，把score最大的bounding box作为选定的框，计算其余bounding box与当前最大score与box的IoU，去除IoU大于设定的阈值的bounding box。然后重复上面的过程，直至候选bounding box为空，然后再将score小于一定阈值的选定框删除得到一类的结果。

Reference：

[1]. R. B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. 4, 5, 6

[2]. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: To- wards real-time object detection with region proposal net- works. *arXiv preprint arXiv:1506.01497*, 2015. 2, 3, 4, 5, 6

[3]. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015. 4, 5

[4]. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: ECCV. (2014)

[5]. J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 2013. 1, 2, 3, 4, 5, 9

[6]. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient GraphBased Image Segmentation. IJCV, 59:167–181, 2004. 1, 3, 4, 5, 7

[7]. P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ra- manan. Object detection with discriminatively trained part based models. *TPAMI*, 2010. 2, 4, 7, 12

[8]. http://blog.csdn.net/u011534057/article/details/51218218

[9]. http://blog.csdn.net/u011534057/article/details/51240387