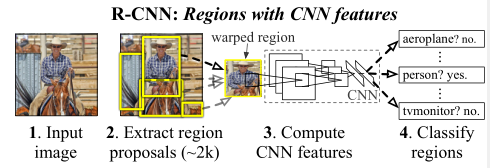
**[Rich feature hierarchies for accurate object detection and semantic segmentation]**

**CVPR2014 Ross Girshick**

**主要思想：**Regions with CNN features。RCNN模型主要分为四个部分：1)区域建议算法(region proposal)采用**selective search**；2)特征提取采用CNN(AlexNet)；3)分类器是线性SVM；4)对边界框采用修正回归模型。算法流程是：1)selective search推荐1000~2000个候选框；2)对所有区域进行归一化和缩放，为了让CNN能够接受；3)采用**AlexNet提取特征**，其中conv5和fc7层的特征会被后续用到；4)fc7的特征分别经过**20个线性SVM分类器**（20类），每个分类器都对当前区域打分确定区域类别，所有区域都操作；5)对上一步获得的所有带标签的区域使用带阈值的**非极大值抑制**，剩下的框就是最后要框出来的；6)conv5层的特征送入**Bounding Box 回归模型**，修正边界框的位置；7)根据前面得到的类别和修正边界框，输出结果。模型的训练量很大，20个（类别数取决于数据集）的SVM分类器、20个Bounding Box 回归模型、和一个CNN模型都需要单独训练，其中CNN模型的训练采用在大型数据集上预训练并在小型数据集上微调的方式。

**贡献/缺点：**RCNN有两个大的贡献：A）第一次实现采用大型**CNN提取特征**，构建了bottom-up的区域推荐模型用于定位和分割，大幅提高了HOG,DPM等算法的结果；B)实现并证明，当训练数据不足时，先在其他的大型数据集上进行有监督的**预训练**，再在小型数据集上进行特定的**微调**，构建高性能的CNN。正因如此，RCNN成为了卷积神经网络在目标检测任务中的开山之作，之前的CNN都是分类模型。RCNN的主要缺点是训练复杂且耗时长，检测时间也超长。好在后续的Fast-RCNN, Faster RCNN有不断的改进。RCNN**本质上是基于候选框的分类模型**，但超多的候选框使它实现了定位。

**论文细节：**



**Region Proposal。**采用selective search产生1000~2000个与类别无关的区域，这个阶段是独立的，不需要进行训练，采用ss是为了实现和其他算法进行受控比较。

**Feature Extraction。**特征提取模型采用的是AlexNet，有5个conv+2个fc，提取到的特征是**fc7层**输出的4096维特征。AlexNet的输入上固定的227x227，而ss产生的区域大小和比例是可变的，经过对比，作者在AlexNet之前对所有区域进行warping（padding16+**各向异性缩放**，就是padding16后直接放射变化到227，不考虑变形，对比各向同性缩放）。

**Test-time Detection。**对2000个region都提取特征之后，也就得到了2000x4096的特征向量。每个region，使用基于每个类别（N=20类）训练出来的SVM打分，即每张图像有2000个region乘 N个分数，对这2000N基于每一类单独使用NMS拒绝掉和高分区域IOU大于阈值的框，得到检测结果。

**Bounding Box Regression。conv5层**的特征输入到下文训练的Bounding box回归模型，对当前region的边界框进行修正，得到修正后的边界框，结合上一步类别的NMS结果，作为最终输出。

**训练：（训练是分为三部分单独训练，比较繁琐）**

**训练CNN。**由于数据不够，先在ImageNet上**pre-train**的AlexNet模型，再在R-CNN中进行**fine-tune**，fine-tune的过程是将AlexNet的Softmax改为任务需要的类别数（N=21,包括背景类），然后还是当做一个分类模型来训练，训练样本的构建使用ss生成的子图，当这些图与实际样本的框（Ground-truth）的IoU大于等于0.5时，认为是某一个类的正样本，这样的类一共有20个；IoU小于0.5时，认为是负样本。然后就可以AlexNet做pre-train了，pre-train之后AlexNet的1000类输出的Softmax层就被扔掉了，只剩下其余层的参数，这套参数保持不变，用来做特征提取，接下来训练20个SVM。

**训练SVM：** SVM的输入特征是AlexNet fc7的4096位特征向量输出，然后SVM做二分类，一共有20个SVM模型。对于其中某一个分类器来说，它的正样本是所有Ground-truth区域经过AlexNet后输出的特征，负样本是与Ground-truth区域重合IoU小于0.3的区域经过AlexNet后输出的特征，特征和标签确定了，就可以训练SVM了。（hard negative mining method，选择具有代表性的样本继续训练提高SVM性能，同时不占据太多硬件资源）

**训练Bounding box回归模型。**Bounding box回归模型也是20个，它的输入是AlexNet conv5的特征，20是类的个数，对一个Bounding box来说，它有4套参数，因为一个Bounding box回归模型分别对对边界框四个属性值做回归，得到**修正后的边界框**。IoU>0.6的ss区域才能经过卷积作为输入特征。

精髓：<https://blog.csdn.net/chaipp0607/article/details/78112861>

翻译：https://blog.csdn.net/v1\_vivian/article/details/78599229