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

**1引言**

特征在目标检测中非常重要。近10年来的进步主要得益于SIFT和HOG的应用，但近两年的发展逐渐放缓。

作者认为从人类对目标进行识别的过程来看，“识别”这一动作发生在信息处理的较深层次，这意味着为了实现准确的目标检测和语义分割，需要使用深度网络模型来逐层地提取丰富的特征和语义信息。此外，*deep neural networks for object detection*论文表明将目标识别问题作为回归问题来解决的结果并不理想，作者认为应该使用滑动窗口检测器来进行识别。那么很自然地就能想到卷积神经网络。作者使用了5层卷积神经网络，且使用195x195的超大感受野以及32x32的步长。这也是论文中所提出的*Rich feature hierarchies*（丰富的特征层次结构）的核心思想。

此外，由于目标检测任务的标记数据远远少于图像识别任务，并不足以支撑深度卷积神经网络模型的训练，作者通过先使用无监督的预训练，然后使用有监督地微调来训练深度卷积神经网络，以解决标记数据少的问题，这是典型的迁移学习的案例，使得在标注数据集较少的情况下也能训练高性能的深度卷积神经网络。

**2 R-CNN的目标检测**

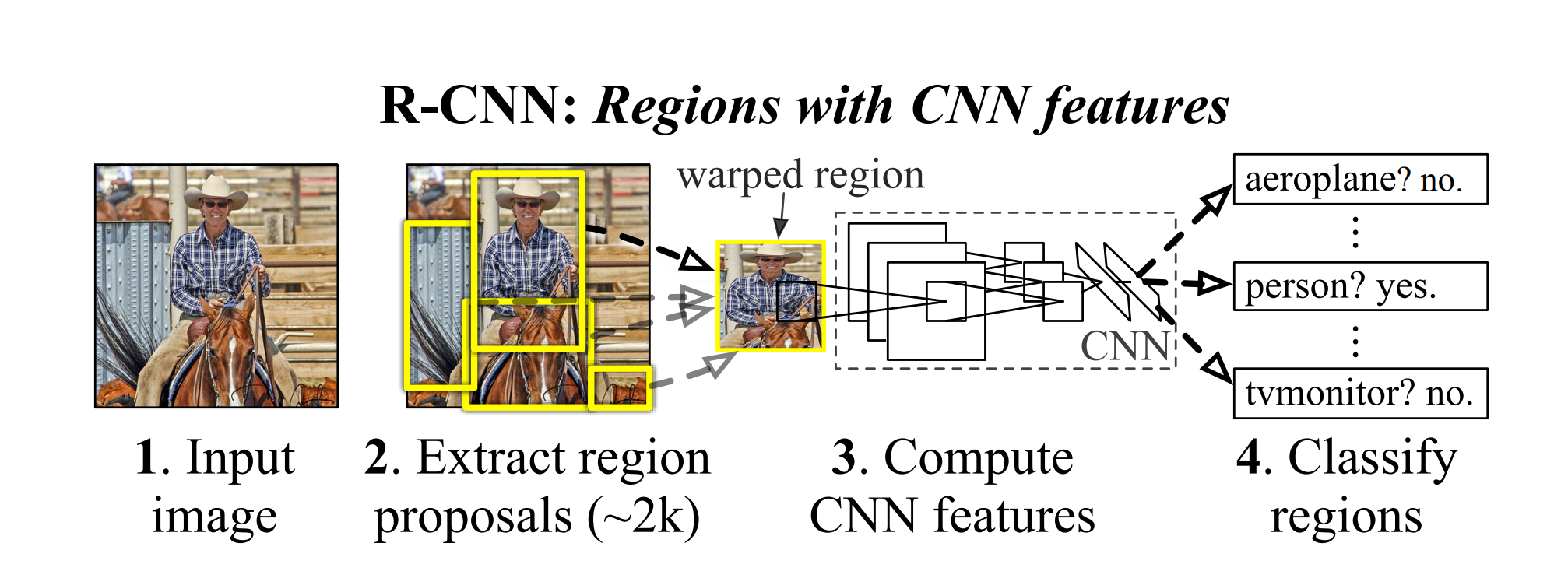


图2-1 RCNN结构

RCNN主要有以下模块组成：

1、区域建议生成（与目标无关）；

2、深度卷积神经网络；

3、20个SVM进行二分类，对卷积神经网络计算得到的边界框进行打分并筛选；

4、边界框回归（Bounding box regression），用于对卷积神经网络计算的边界框进行微调。

下面对上述模块中的细节进行介绍。

**2.1 区域建议算法**

首先需要说明的是，R-CNN并不需要指定区域建议算法，任何区域建议算法都可以使用。本文中，为了和其他的目标检测算法进行比较，选取了selective search算法[1]作为区域建议算法。Selective search算法的主要思想是通过分割和合并的方法，将图像划分为一组具有相似纹理、颜色和结构的区域，并通过计算不同区域之间的相似程度，将这些区域逐步合并成更大的候选框。

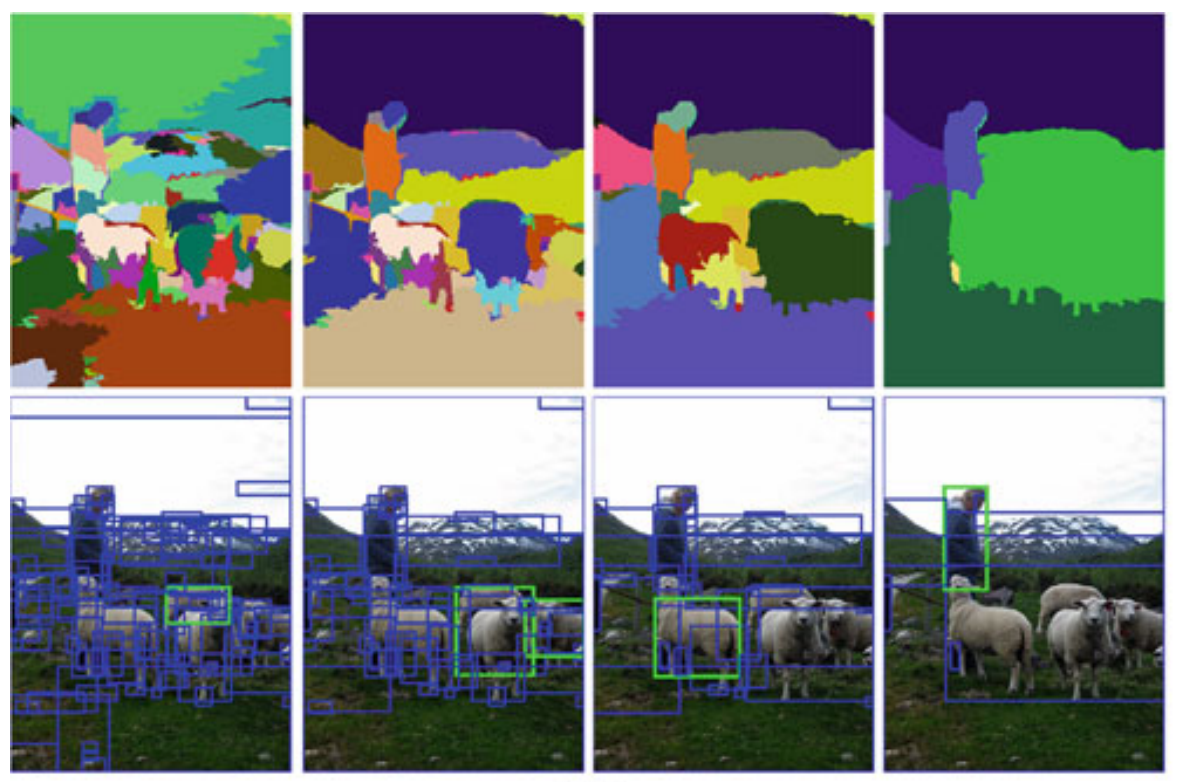


图2-2 selective search框选过程示例

**2.2深度卷积神经网络**

在RCNN中，深度卷积网络主要用于目标边界框的计算。本文主要采用AlexNet[2]的网络架构，对数据集进行训练。然而，相对于图像识别领域的开放数据集规模，目标检测领域的标注数据集不足以支撑AlexNet这样的深度卷积神经网络的训练。于是作者采用迁移学习的思想，先用ImageNet2012数据集对卷积神经网络进行预训练，然后对网络进行fine-tuning（微调）。

首先使用2.1中介绍的selective search对VOC数据集中的图像提取了2000个候选框（tight bounding box），为了与卷积神经网络的输入尺寸（227\*227）匹配，作者首先对候选框进行了16个像素值上下文填充，然后将填充后的候选框放缩为227\*227尺寸的图像，作为fine-tuning部分中卷积神经网络的输入。作者设置mini-batch为128，其中32张为positive windows（含全部目标种类），96张为background windows。这里的positive是使用通过计算卷积神经网络得到的边界框和真实边界框之间的IoU，0.5以上的设为positive得到的。

此外，作者将原来的1000维（ImageNet数据集的1000类目标）特征向量改为21维（VOC数据集中的20类目标以及背景）。

**2.3 目标分类**

本文中训练了20个SVM支持向量机分别进行20类目标的分类。通过2.1节和2.2节中的方法得到适用于目标检测的深度神经网络模型，将VOC数据集的test集进行2.1节的相同操作后作为模型的输入，将pool5得到的特征图作为SVM的输入进行打分。结合上下文我们可以发现，一定会有很多冗余的候选框出现，即多个候选框对应某一个目标。为了消除冗余，提高后续计算的效率，作者使用非极大值抑制来消除冗余。

非极大值抑制过程如下:

1、对SVM打分后的候选框进行按分数的降序排序（显然是对于同一目标而言）。

2、计算分数最高的候选框与其余候选框的IoU，若大于0.3（作者实验后得到），说明存在冗余，则丢弃；反之则保留。

3、重复过程2，直到没有候选框被丢弃。

**2.4边界框回归（Bounding box Regression）**

将2.3中消除冗余后的候选框与真实候选框进行回归，拟合出一个回归模型用于对深度卷积神经网络计算得到的边界框进行微调，已达到更好的效果。

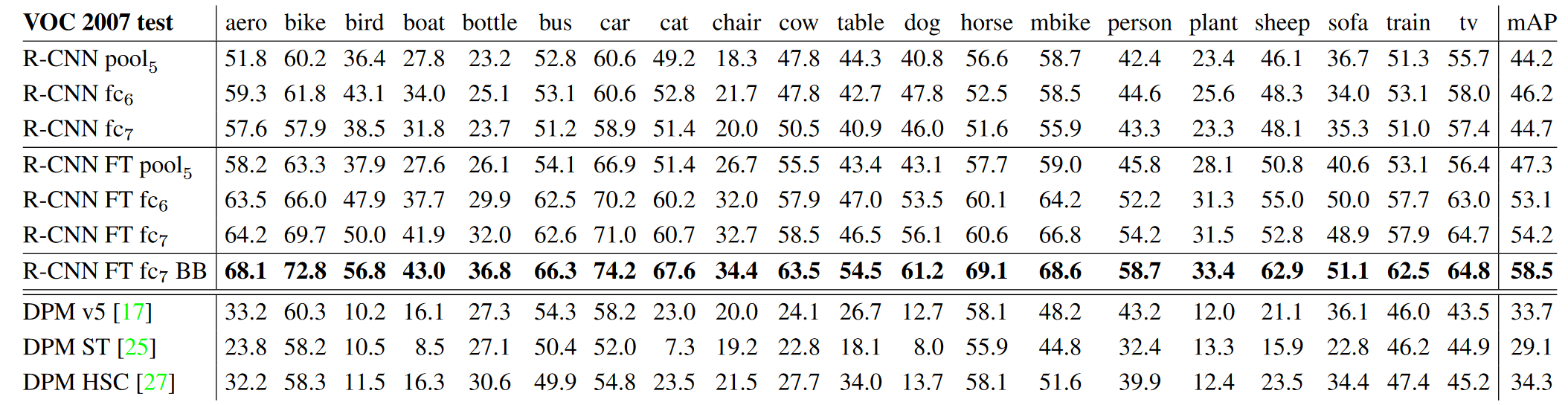
**3、可视化，消融以及误差分析**

**3.1可视化**

作者使用ZFNet[3]对响应最强烈的特征进行可视化，并标注相应候选框以及SVM计算得到的分数，结果如图XX所示，作者表示，部分单元捕捉人像和文本，部分单元捕捉纹理和材质（点阵和反光）。

**3.2 消融实验**

首先，作者为了理解哪一层对性能影响较大，分析了在VOC 2007数据集上训练的有微调和无微调的CNN的最后三层，即Pool5，FC6和FC7。结果如表XX所示。结果表明CNN的的表达能力主要来自卷积层，而不是全连结层。此外，通过比较有微调和无微调的结果可以发现，有微调的CNN网络的mAP提高了8%，达到了54.2%，效果显著。



**4 讨论**

作者提出的简单且可扩展的物体检测方法，将在VOC数据集的训练性能提高了30%。这主要得益于两方面的工作：1、将高性能的深度剪辑神经网络应用于自底向上的区域建议算法以定位以及分割目标。2、在标签数据集较少的情况下，使用预训练加微调的方式训练模型。

缺陷：

1、训练是多阶段的，而不是端到端的；

2、训练过程非常耗费空间和时间；

3、目标检测非常慢