**富特征层级用于准确物体检测和语义分割**

**摘要**

物体检测性能，在权威的PASCAL VOC数据集上评估，在近几年已达到了稳定。性能最好的方法都是通常将多个低层图像特征和高层语义信息相结合的复杂集成系统。本文中，我们提出了一个简单可扩展的检测算法可以在VOC 2012最好的结果上提高相对30%的平均精度——达到了53.3%。我们的方法组合了两个关键见解：（1）一是可以应用高能力的卷积神经网络（CNNs）于自底向上的区域提案，以此来定位和分割物体。（2）当标注的训练数据稀少时，有监督的辅助任务预训练，接着进行特定领域调优会产生一个重大的性能提升。因为我们结合了区域提案和CNNs，我们称该方法为R-CNN：带有CNN特征的区域。我们也比较了R-CNN与OverFeat，一个最近提出的基于相似CNN结构的滑窗检测器。我们发现R-CNN在200类的ILSVRC2013检测数据集上大幅超越OverFeat。完整系统的源代码可在如下网址获取<http://www.cs.berkeley.edu/~rbj/rcnn>。

1. **简介**

特征很重要。多种视觉识别任务在过去十年的发展中一直是基于SIFT和HOG的使用。但如果我们看权威视觉识别任务，PASCAL VOC物体检测，普遍认为2010-2012年的发展很慢，建立集成系统和使用成功方法的微调变体所获收益很少。

SIFT和HOG都是blockwise orientation histograms，这种表示我们可以粗略的与V1（灵长类动物视觉路径的第一皮层区域）的复杂细胞相联系。但我们也知道识别发生在下游的几个阶段，这就表明可能存在分层结构，多阶段处理来计算视觉识别中更具信息的特征。

Fukushima的“neocognitron”，一个受生物学启发的层级、平移不变的模式识别模型，就是这样过程的一个早期尝试。但是neocognitron缺少一个有监督的训练算法。建立在Rumelhart等[33]的工作，LeCun等表明通过反向传播的随机梯度下降对训练神经网络十分有效，这是一种对neocognitron拓展的模型。

在1990s CNNs被广泛使用，但接着随着支持向量机的兴起又不再流行。在2012年，Krizhevsky等通过展示在ILSVRC上十分高的图像分类准确度又重燃了人们对CNNs的兴趣。他们的成功源于在120万标注图像上训练的大型CNN，以及对LeCun CNN的一些转变（如，max(x,0)整流非线性以及“dropout”正则化）。

ImageNet结果的重要性在ILSVRC 2012研讨会期间大有争议。核心问题可以归纳如下：CNN在ImageNet上的分类结果可以在多大程度上泛化到PASCAL VOC 挑战上的物体检测结果？

我们通过桥接图像分类和物体检测的间隙来回答该问题。本文是第一篇展示了在PASCAL VOC上，CNN与基于简单类似HOG特征的系统相比可以得到令人瞩目的物体检测结果。为了达到该结果，我们专注于两个问题：使用深度网络来定位物体以及使用一小部分标注的检测数据训练一个高能力模型。

与图像分类不同的是，检测需要在图像中定位物体。一个方法将定位视为回归问题。但是，Szegedy等的工作，与我们的工作一直，都表明该策略在实践中不太成功（他们报告了VOC 2007上30.5%的mAP而我们的方法达到了58.5%的mAP）。一个可供选择的方法就是构建一个滑窗检测器。这样使用CNNs已经至少20年了，尤其是在一些特定的类别，如人脸和行人。为了维持高的空间分辨率，这些CNNs通常只有两个卷积层和池化层。我们也想采用一个滑窗的方法。但是，我们网络中具有五个卷积层的高层单元在输入图像中有非常大的感知野（195x195像素）以及步长（32x32像素），这使得在滑窗范例中精确定位成为公认的技术挑战。

然而，我们通过在“识别使用区域”的范例（该方法在物体检测和语义分割中都很成功）中操作来解决CNN的定位问题。；在测试时，我们的方法在输入图像中产生了大概2000个类别独立的区域提案，使用CNN在每个提案中提取一个固定长度的特征向量，接着使用特定类别的线性SVMs来对每个区域进行分类。我们使用一个简单的技术（仿射图像扭曲）来从每个区域提案中计算一个固定大小的CNN输入，而不管区域的形状。图1展示了我们方法的一个概况并突出了一些我们的结果。因我我们的系统结合了区域提案和CNNs，我们成该方法为R-CNN：带有CNN特征的区域。

在本文最新的版本中，我们通过在200类ILSVRC2013检测数据集上运行R-CNN提供了一个R-CNN与最近提出的OverFeat检测系统的详细比较。OverFeat使用一个滑窗CNN来检测并且目前为止实在ILSVRC2013检测上表现最好的方法。我们表明了R-CNN比OverFeat表现的好很多，31.4%的mAP比24.3%mAP.

检测中面临的第二个问题是标注的数据是稀少的而且目前可用的数量对训练大型CNN来说是不足的。解决该方法的传统方法是使用无监督预训练，紧接着进行有监督调参。本文第二点原则的贡献是表明在大型辅助数据集（ILSVRC）上进行有监督预训练，紧接着进行特定领域的调参对于数据稀少时学习高能力的CNNs是十分有效的范例。在我们的实验中，为检测调参提高了mAP八个百分点。调参后，我们的系统在VOC 2010上达到了54%的mAP相比于高度调优，基于HOG的可变型部件模型（DPM）。我们也向读者指出同期Donahue等的工作表明了Krizhevsky的CNN（没有进行调参）可以用作黑箱特征提取器，在一些识别任务中表现很好，如场景分类，细粒度的子分类以及领域自适应。

我们的系统也很高效。唯一的特定类计算是相当小的矩阵向量相乘以及贪心非极大值抑制。这一计算性质是由所有类共享的特征产生的并且比之前使用的区域特征低两个数量级。

了解我们方法的失败模式对提升它是至关重要的，所以我们报告了Hoiem等人检测分析工具的结果。作为该分析的直接结果，我们证明了一个简单的边界框回归方法极大的减少了误定位（主要的错误模式）。

在开发详细技术之前，我们注意到因为R-CNN在区域上操作，它很自然被拓展到语义分割任务。经过小部分修改，我们也以VOC 2011测试集上47.9%的平均分割精度在PASCAL VOC语义任务上达到了具有竞争力的结果。

1. **使用R-CNN进行物体检测**

我们的物体检测体统包含三个模块。第一个模块产生类别独立的区域提案。这些提案定义了检测器可用的后选检测集。第二个模块是一个从每个区域提取一固定长度特征向量的卷积神经网络。第三个模块是特定类别线性SVM的集合。本节中，我们将展示每个模块的设计决策，描述他们测试的用法，详细说明每个参数如何学习并展示在PASCAL VOC 2010-12和ILSVRC2013上的检测结果。

* 1. **模块设计**

**区域提案。**最近大量的文章提供了生成类别独立的区域提案的方法。例如：objectness,selective search,category-independent object proposal,constrain-ed parametric min-cuts(CPMC),multi-scale combinatorial grouping,以及Ciresan等人通过在regulaly-spaced square crops（区域提案的一种特殊情况）上应用CNN来检测细胞有丝分裂。虽然R-CNN对特定区域提案方法是不可知的，但我们使用select-ive search使得与先前检测工作的可控对比可行。

**特征提取。**我们使用Caffe对Krizhevsky等的CNN的实现从每个区域提案中提取一个4096维的特征向量。特征是通过前向传播一个减去均值的227x227的RGB图像通过五个巻积层和两个全连接层。更多网络结构细节读者可参阅[24.25]。

为了计算某个区域提案的特征，我们首先必须要把该区域中的图像数据转换为与CNN兼容的形式（该结构需要固定227x227像素大小的输入）。在我们任意形状区域的所有可能形变中，我们选择最简单的。不管后选区域的尺寸或宽高比，我们将他周围的紧边界框中的所有像素扭曲到所需的大小。在扭曲之前，我们扩大了紧边界框，使得在扭曲尺寸上刚好在原始边界框附近有p个像素的扭曲图像上下文信息（我们使用p=16）。图2展示了扭曲的训练区域的随机采样。扭曲的可选方案在附录A中讨论。

* 1. **测试时检测**

在测试时，我们在测试图像上运行selective search来提取大约2000个区域提案（在所有实验中，我们都使用了selective search的“快速模式”）。我们扭曲每个提案并将其前向传播过CNN以此来计算特征。接着，对于每个类，我们使用该类训练的SVM来为每个提取的特征向量打分。给定所有图像中打分的区域，我们使用贪心非极大值抑制（对每个类独立使用）以抵制那些与具有高过学习阈值得分的选择区域有重叠的区域。

**运行时分析。**两个属性使检测高效。首先，CNN参数被所有类共享。第二，当与其他一般方法（如具有视觉词袋编码的空间金字塔）比较时，CNN所计算的特征向量是低维的。例如，UVA检测系统所使用的特征比我们的高两个数量级（360k比4k维）。

这种参数共享的结果就是计算区域提案和特征的时间（GPU上13s每幅图或CPU上53s每幅图）被分摊到所有类别上。唯一的特定类别计算知识特征和SVM权重间的点乘或非极大值抑制。实践中，所有的图像点乘都被批处理成一个单一矩阵相乘。特征矩阵通常是2000x4096而SVM矩阵是4096xN，N就是类别数。

分析表明R-CNN可以拓展到几千个物体类别二不需要依赖于类似的技术，如哈希。即使有100k个类，产生结果的矩阵乘法在现代多核CPU上也只花10秒。这种高效不仅仅是使用区域提案和共享特征的结果。UVA系统，由于它的高维特征，要比我们的方法慢两个数量级，并且需要134GB内存来存储100k个线性预测器，相比于该方法，我们只需要1.5GB内存来存储低维特征。

将R-CNN与最近Dean等人在可拓展检测上使用DPMs和哈希的工作相比较是十分有趣的。他们报告了在VOC 2007上大约16%的mAP,在引入10k个干扰类别时，运行速度为5分钟一张。使用我们的方法，10k个检测器在CPU上运行大约1分钟，由于没有近似，mAP将会保持在59%。

* 1. **训练**

**有监督预训练。**我们在大型辅助数据集上（ILSVRC 2012 分类）只使用图像级标注（边界框标注在该数据上不可用）来对CNN有区分的进行预训练。我们使用了开源的Caffe CNN库来进行预训练。简言之，我们的CNN基本与Krizhevsky等的性能一致，获得了前1的错误率，在ILSVRC 2012分类验证集上高出2.2个百分点。这种不符是由于训练工程的化简。

**特定领域调参。**为了使我们的CNN可以适应新任务（检测）以及新的领域（扭曲的提案窗口），我们继续只使用扭曲的区域提案来进行CNN参数的随即梯度下降。除了用随机初始化的（N+1）-路分类层（N是物体类别数量，加1是背景）来替换CNN的ImageNet特定1000-路分类层以外，CNN的结构是没变的。对于VOC，N=20而对于ILSVRC2013，N=200。我们将所有与ground-truth IoU>=0.5的区域提案视为对应类的正样本而剩下的为负样本。我们的SGD初始学习率为0.001（初始预训练率的十分之一），这使调参可以进行的同时不会破坏初始值。在每次SGD迭代中，我们一致的采样32个正样本窗口（所有类）以及96个背景窗口来构造大小为128的小批量。我们偏向正窗口采样因为他们与背景相比特别少。

**五题类别分类器。**考虑训练一个两分类器来检测车辆。很明显一个图像区域紧紧的包围一个车就应该为正样本。相似的，背景区域，与车没有任何关系，应为负样本。不太明确的是如何标注一个与车部分重叠的区域。我们使用一个IoU重叠阈值来解决该问题，低于该阈值的定义为负样本。重叠阈值为0.3，是通过在验证集上使用网格搜索｛0，0.1，…，0.5｝确定的。我们发现仔细的选择该阈值是重要的。设置其为0.5，如在[39]中一样，会降低mAP五个百分点。相似的，设置其为0则会降低mAP四个百分点。正样本被简单的定义为每个类的ground-truth边界框。

当特征被提取并且训练标注被应用后，我们就会为每个类调优一个线性SVM。由于训练数据太大而不能装入内存，我们采用了难负挖掘方法。难负挖掘收敛很快并且在实践中mAP会在单次通过所有图像后停止增长。

在附录B中我们讨论了为什么在调优和SVM训练中正负样本的定义不同。我们也讨论了涉及训练检测SVMs的权衡而不是简单的使用调优过后的CNN最后softmax层的输出。

* 1. **PASCAL VOC 2010-12上的结果**

根据PASCAL VOC上的最佳实践，我们在VOC 2007数据集上验证了所有设计决策和超参数。在VOC 2010-12数据集上的最终结果，我们在VOC 2012训练集上调优了CNN并在VOC 2012 trainval上调优了检测SVMs。我们对两个主要算法变体每个只向评估服务器提交了一次测试结果（有和没有边界框回归）。

表1展示了VOC 2010上完整的结果。我们比较了我们的方法与四个强基准方法，包括SegDPM,一个将语义分割系统和DPM检测器结合并使用附加inter检测器上下文和图像分类器重打分的方法。和Uijlings等人的UVA系统的比较是最恰当的，因为我们的系统使用了相同的区域提案方法。为了分类区域，他们的方法建立了一个四层的空间金字塔并使用密集采样的SIFT，拓展的OpponentSIFT以及RGB-SIFT描述符对其进行填充，每个向量都使用4000词的codebooks进行量化。分类是使用一个直方图相交核SVM。相比于他们的多特征，非线性核SVM方法，我们在mAP上达到了很大的提升，从35.1%到53.7%,并且速度更快。我们的方法在VOC 2011/12测试集上也达到了相似的表现（53.3%mAP）。

* 1. **ILSVRC2013检测上的结果**

我们在200类ILSVRC2013检测数据集上运行了R-CNN并使用了我们在PASCAL VOC上相同的系统超参数。我们按照同样的方法将测试结果提交至ILSVRC2013评估服务器上两次，一次有一次没有边界框回归。

表3比较了R-NN与ILSVRC 2013比赛条目与比赛后的OverFeat结果。R-CNN在mAP上达到了31.4%，比第二好的OverFeat的24.3%有大幅提升。为了给出AP对所有类贡献的直观感受，我们给出了箱型图以及文章最后的每个类AP的表格。大多数具有竞争力的提交结果（OverFeat,NEC-MU,UvA-Euvision,Toronto A以及UIUC-IFP）使用卷积神经网络，表明了如何将CNNs应用到物体检测具有很大的细微差距，也会导致不同的结果。

在第四节中，我们给出了ILSVRC2013检测数据集的概述并提供了我们运行R-CNN的选择细节。

1. **误差的可视化，消除和模式**
   1. **可视化学习的特征**

第一层过滤器可以被直接可视化并很容易理解。他们捕捉定向边缘和相对颜色。理解随后的几层更具挑战性。Zeiler和Fergus在[42]中提出了一个直观生动的反卷积法。我们提出了一个简单（补足的）的非参数方法可以直接展示出网络学到了什么。

思想就是在网络中挑出一个特别单元（特征）并将其作为独立的物体检测器使用。这样，我们在一大批held-out区域提案上计算单元的激活函数，将提案的激活从高到底排序，执行非极大值抑制然后显示得分最高的区域。我们的方法通过准确的展示那些输入正在被单元使用（fires on?）来使得选择的单元“为自己说话”。我们避免平均以观察不同的视觉模式并深入了解单元计算的不变性。

我们可视化了pool5层的单元，该层是网络第五层以及最后巻积层的最大池化输出。Pool5特征图是6x6x256=9216维的。忽略边界影响，每个pool5单元在227x227像素输入中都有一个195x195像素的接受区域。中心的pool5单元拥有几乎全局的视角，而边界附近的单元只有小的，一部分的支持。

图4每一行显示了我们在VOC 2007 trainval上调参的CNN中pool5层的前16个激活。6/256个功能独特单元被可视化了。这些单元被选出来以展示网络所学到内容的代表性样本。在第二行中，我们可以看到一个在狗脸和点阵上触发的单元。相对于第三行的单元是一个红色斑点检测器。还有人脸和抽象模式如文本和带有窗户的三角形结构的检测器。网络似乎会学习一个结合了小部分类别调整特征的代表和一个分布式包括形状，纹理，颜色以及材料特性的代表。之后的全连接层fc6有能力来对大量的这些富特征的组成部分建模。

* 1. **Ablation studies**

**每一层的性能，不使用调参。**为了了解哪些层对于检测性能来说是至关重要的，我们分析了CNN最后三层每一层在VOC 2007数据集上的结果。Pool5层在3.1节中已经简单的描述。最后两层可以总结如下。

Fc6层与pool5全连接。为了计算特征，它将一个4096x9216的权重矩阵与pool5特征图（变形为9216维向量）相乘然后与偏置向量相加。这个中间向量是分量半波整流的（x<-max(0,x)）。

Fc7是网络的最后一层。它是通过将fc6计算的特征与一个4096x4096维的权重矩阵相乘，并加上一个偏执向量以及应用半波整流实现的。

我们先看没有在PASCAL上调参的CNN结果，也就是所有CNN参数都只是在ILSVRC 2012上预训练的。一层一层的分析性能（表2 1-3行）显示了在泛化上，fc7产生的特征没有fc6产生的特征好。这意味着29%，或这说大约1680万的CNN参数可以被移除而不会降低mAP.更出人意料的是移除掉fc7和fc6也产生相当不错的结果尽管计算pool5特征只是用了CNN 6%的参数。CNN的表示能力大部分来自于它的巻积层，而不是大型密集的连接层。这一发现表明对于任意只存的图像只使用CNN的巻积层计算一个密集特征图，以HOG的角度来说，可能很实用。这种表示使得使用滑窗检测器，包括DPM，在pool5特征之上的实验可行。

**每一层的性能，使用调参。**我们现在来看在VOC 2007 trainval上调参之后我们CNN的结果。提升是显著的（表2 4-6行）：调参提升mAP八个百分点到54.2%。调参的在fc6和fc7上的提升比pool5上的提升大很多，这表明pool5从ImageNet中学到的特征是普遍的而大部分的提升是从其上学习领域特定的非线性分类器中获得的。

**与当前特征学习方法的比较。**相对较少的特征学习方法在PASCAL VOC检测上使用过。我们观察了最近两个基于可变型部件的方法。为了参考，我们也囊括了标准基于HOG的DPM结果。

第一个DPM特征学习方法，DPM ST，增加了具有“sketch token”概率直方图的HOG特征。直观上来说，sketch token就是通过图像patch中心的轮廓紧密分布。Sketch token概论是通过随机森林在每个像素上计算的，该随机森林是训练用来将35x35像素的patches分类到150个sketch tokens或背景之一。

第二种方法，DPM HSC，将HOG替换为稀疏编码的直方图（HSC）。为了计算HSC，使用一个具有100个7x7像素（灰度图）原子的学习字典来在每个像素上计算稀疏编码激活。结果激活被整流为三种方式（全波，两个半波），空间池化，单元l2归一化然后幂变换（x<-sign(x)|x|a）。

所有R-CNN变体都比三个DPM基准方法表现突出（表2 8-10行），包括两个使用特征学习的方法。与最新版本只使用HOG特征的DPM相比，我们的mAP高出20百分点：54.2%vs33.7%——61%的相对提升。HOG和sketch tokens的组合产生了比只使用HOG高出2.5个mAP点的结果，而HSC比HOG提升了4个mAP点（当与他们私有的DPM基准进行内部比较时——都使用比开源版本表现差的非公开实现）。这些方法分别在mAP上达到了29.1%和34.3%。

* 1. **网络结构**

本文中大部分结果使用了Krizhevsky等人的网络结构。但是我们发现结构的选择对R-CNN检测性能有着很大影响。表3中我们展示了使用最近由Simonyan 和Zisserman提出的16层深度网络在VOC 2007测试集上的结果。该网络是最近ILSVRC 2014分类挑战中表现最好的之一。网络具有均匀的结构，包括13层3x3的卷积核，以及散布五个最大池化层和最上面三个全连接层。我们将该网络称为“O-Net”代表OxfordNet以及称基准方法为“T-Net”代表TorontoNet。

为了在R-CNN中使用O-Net，我们从Caffe Model Zoo上下载了公开可用的VGG\_ILSVRC\_16\_layers模型的预训练网络权重。然后我们使用了与在T-Net上相同的协议来对网络调参。唯一的区别是使用更小的minibatches（24样本）以适应GPU显存。表3中的结果展示了使用O-Net的R-CNN大体上比使用T-Net的R-CNN表现要好，将mAP从58.5%提升到66.0%。但是在计算时间上有大量的缺点，O-Net的前传比T-Net要长大约7倍。

* 1. **检测误差分析**

我们使用了Hoiem等人的检测分析工具以显示我们方法的误差模式，理解调参是如何改变它们的并观察与DPM相比我们的误差类型。完整的分析同居总结超出了本文的范围，我们鼓励读者查阅来了解详情（例如“normalized AP”）。因为分析在相关图示中最好被吸收，我们在图5和图6的说明中给出了讨论。

* 1. **边界框回归**

基于误差分析，我们实现了一个简单方法来减少定位误差。受到DPM中所使用的边界框回归所启发，我们训练了一个回归模型，给定从selective search区域提案中的pool5特征预测新的检测窗口。详细信息在附录C中给出。表1，表2和图5的结果展示了这个简单方法修正了大量误定位检测，提升了mAP三到四个百分点。

* 1. **定性结果**

ILSVRC2013的定性检测结果在文末图8和图9给出。每个图像都是从val2 set中随机采样的并且展示了所有检测器的检测结果中准确率高于0.5的。注意这些并不是被设计的（curated）并给出了所使用检测器的现实印象。更多定性的结果在图10和图11给出，但是这些都是被设计的（curated）。我们选择每张图片是因为其包含了有趣的，惊喜的或好玩的结果。这里也是展示了所有准确度高于0.5的检测结果。

1. **ILSVRC2013检测数据集**

在第二节中我们给出了ILSVRC2013检测数据集上的结果。该数据集与PASCAL VOC相比没那么均匀，需要使用它的策略。因为这些决策并不琐碎，我们将在本节中介绍。

* 1. **数据集概览**

ILSVRC2013检测数据集被分为三部分：训练（395918），验证（20121）以及测试（40152）。验证和测试分割使用相同的图像分布。这些图像与PASCAL VOC图像是场景相似的并在复杂度上相似（物体数量，混乱程度，姿势多变性等）。Val和test都尽量标注了，意味着每张图中200类里的所有实例都用边界框标注了。相对的，训练集是服从ILSVRC2013分类图像分布。这些图像具有更高的可变复杂度，并对于单个中心物体的图像偏移。与val和test不同的是，训练图像（因为其数量巨大）并没有全部标注。在任何给定的训练图像，200类的实例可能也可能没有被标注。出来这些图像集外，每个类还有一个附加的负图像集。手动检查负图像来验证他们不包含与他们相关联类的实例。本工作中并没有使用负图像集。关于ILSVRC如何搜集和标注数据的详细信息可以在[11,36]中找到。

这些分割的性质为训练R-CNN提供了很多选择。训练图像不能用来难负挖掘，因为标注并不详尽。负样本应该从哪里来？而且，训练图像和val及test有着不同的统计信息。是否应该使用训练图像，如果需要，在多大程度上需要？虽然我们没有完整的评估大量的选择，我们基于前面的经验给出了一个似乎最明显的线路。

我们的一般策略是主要依赖于验证集并使用一些训练图像作为辅助正样本源。为了使用val来进行训练和验证，我们将其大致平均分成了“val1”和“val2”集。因为一些类在val中有很少的样本（最少的只有31，一半都少于110），产生一个近似类别平衡的划分是很重要的。为了做这项工作，产生了大量的候选分割，最后选择了一个具有最小类别不平衡极大值的。每个候选分割都是通过使用val图像类别的数量作为特征对其进行聚类，接着使用局部搜索来提高分割平衡性产生的。本处所使用的分割具有11%的最大相对不平衡以及4%的中等相对不平衡。Val1/val2分割以及产生它们的代码将会被公开使得其他研究者可以在本报告所使用的val分割上比较他们的方法。

* 1. **区域提案**

我们采用了与PASCAL检测上所使用的相同的区域提案方法。Selective search在val1,val2和test上的每张图像都以“fast mode”运行（但不是在train中的图像上）。需要一个小修改来解决selective search不是尺度不变的问题，因此产生的区域数取决于图像分辨率。ILSVRC图像尺寸从非常小到一些几百万像素不等，所以我们我们在运行selective search前将每张图像修改到固定宽度（500像素）。在val上，selective search在每张图上产生了平均2403个区域提案，并且对于所有ground-truth边界框具有91.6%的召回率（0.5IoU阈值）。该召回率比PASCAL上低很多（98%），表明区域提案阶段还有很大的提升空间。

* 1. **训练数据**

对于训练数据，我们形成了一个图像和边界框集包括所有val1中的selective search和ground-truth边界框以及train中每类最多N个ground-truth边界框（如果一个类具有少于N个ground-truth边界框，我们就全部选取）。我们将该数据集称为images and boxes val1+trainN.在ablation study中，我们展示了val2上N的mAP.

训练数据在R-CNN中需要经过三个步骤：（1）CNN 调参（2）检测器SVM训练（3）边界框回归器训练。CNN调参在val1+trainN上使用PASCAL上所用完全一致的设置来运行50kSGD迭代.使用Caffe在单个NVIDIA Tesla K20上调参需要花费13个小时。对于SVM训练，所有val1+trainN中的ground-truth边界框都用作他们相对类别的正样本。难负挖掘在val1中随机选取的一个5000张图像的子集上运行。一个初始实验表明从所有val1中进行难负挖掘与从5000图像的自己中挖掘，会导致mAP上0.5个百分点的下降，而减少一半的SVM训练时间。没有负样本从train中选取因为其标注不详尽。附加的验证负样本并没有被使用。边界框回归器是在val1上训练的。

* 1. **验证和评估**

在提交结果到评估服务器前，我们使用了上述的训练数据在val2集上验证了数据使用策略，调参的影响以及边界框回归。所有的系统超参数（如，SVM C超参数，区域扭曲的内边距，NMS阈值，边界框回归超参数）与PASCAL上使用的一致。毋庸置疑的是一些超参数的选择不是ILSVRC上最好的，但是本项工作的目标是在ILSVRC上不使用大量数据集调整来产生一个R-CNN预备结果。在选择过val2上的最好决策之后，我们提交了两个结果文件到ILSVRC2013评估服务器。第一个提交是没有边界框回归的而第二个是有的。对于这些提交，我们拓展了SVM和边界框回归器训练集来分别使用val+train1k和val。我们使用了在val1+train1k上调参的CNN以避免重新运行调参和特征计算。

* 1. **消除研究**

表4展示了对于训练数据量的不同，调参以及边界框回归的影响的消除研究。第一个观察到的就是在val2上的mAP与test上的mAP十分接近。这在val2上的mAP是对于测试集性能很好的指示器上给予了我们信心。第一个20.9%的结果是使用在ILSVRC2012分类数据集上预训练CNN的R-CNN达到的结果（没有调参）并且允许访问val1中少量的训练数据（回想起来，val1中一般的类具有15到55的样本）。将训练集拓展到val1+trainN将性能提升到24.1%，本质上N=500和N=1000是没有区别的。只使用val1中的样本来调参CNN产生一个适度的提升到26.5%，但是因为正训练样本的量少所以容易严重过拟合。将调参集拓展到val1+train1k，每类有1000个正样本来自train集，提升mAP到29.7%。边界框回归将结果提高到了31.0%，比在PASCAL上获得的增益相对要少。

* 1. **与OverFeat的关系**

R-CNN和OverFeat之间有很有趣的关系：OverFeat可以看作是R-CNN的一种特例。如果将selective search区域提案替换为多尺度正则矩形区域金字塔并将每类边界框回归器转换为一个单一边界框回归器，则两个系统就会非常相似（除去一些训练上潜在的显著差异：CNN检测调参，使用SVM等）。不值一提的是OverFeat在速度上比R-CNN快很多：大约快9倍，基于[34]中所引证的每幅图两秒。该速度是由于OverFeat的滑窗（也就是区域提案）没有在图像级别上进行扭曲因此重叠窗口何以共享计算。共享是通过在任意尺寸输入上以卷积的方式运行整个网络来实现的。提升R-CNN速度在很多方面是可能的并有待未来的工作去做。

1. **语义分割**

区域分类是语义分割的标准技术，使得我们很容易将R-CNN应用于PASCAL VOC 分割挑战。为了便于与目前领先的语义分割系统（称为O2P，代表“second-order pooling”）直接进行比较，我们在其开源框架中进行工作。O2P使用CPMC在每张图上产生150个区域提案接着使用SVR对每个类预测每个区域的质量。他们方法的高性能是因为CPMC区域的质量以及多种特征类型的强大的二阶池化（丰富了SIFT和LBP的变体）。我们也注意到Farabet等人证明了使用CNN作为多尺度每像素分类器在一些密集场景标注数据集（不包括PASCAL）上得到了不错的结果。

我们根据[2,4]将拓展了PASCAL分割训练机使其包含了Hariharan等人可用的附加标注。设计决策和超参数都在VOC 2011验证集上进行交叉验证。最终的测试结果只评估了一次。

**使用CNN特征分割。**我们评估了在CPMC区域上计算特征的三种策略，所有的这些开始先对区域附近的矩形窗口扭曲到227x227。第一个策略（full）忽略区域的形状并直接在扭曲的窗口上计算CNN特征，和我们进行检测时一样。但是，这些特征忽略了菲矩形的区域。两个区域可能据有非常相似的边界框而有很少的重叠。因此第二个策略（fg）只在区域的前景蒙板上计算CNN特征。我们将背景替换为平均输入使得背景区域在去除平均后为0。第三个策略（full+fg）简单的连接了full和fg的策略；我们的实验验证了他们的互补性

**VOC 2011上的结果。**表5展示了我们在VOC 2011验证集上的结果与O2P相比的总结（完整的每类结果见附录E）。在每个特征计算策略中，fc6层总是比fc7表现要好，接下来的讨论涉及的是fc6特征。Fg策略比full表现稍微好一点，表明了掩蔽的区域形状提供了一个更强的信号，与我们的直觉相符。但是，full+fg达到了一个平均准确度47.9%，与我们的结果有4.2%的差距（也比O2P好一点），表明了full特征提供的上下文信息具有很多的信息即使给出了fg特征。特别的，在我们full+fg特征上训练20个VR单核花费了一个小时，而在O2P特征上训练超过10小时。

表6中我们给出了VOC 2011测试集上的结果，以及我们表现最好的方法,fc6（full+fg），和两个强基准方法。我们的方法在21类中有11类上达到了最高的分割准确度，并达到了整体最高的分割准确度47.9%，所有类平均下来（但在任何合理的误差范围内与O2P结果相关联）。通过调参可能达到更好的结果。

1. **结论**

最近几年，物体检测性能停滞了。表现最好的结果是结合了多种低级图像特征与高级物体检测器和场景分类器中的上下文信息的复杂集成。本文提出了一个简单可拓展的物体检测算法并在PASCAL VOC 2012先前最好的结果上有相对30%的提升。

我们通过两个见解来达到该性能。第一个将高能力的卷积神经网络应用于自底向上的区域提案以定位和分割物体。第二个是当标注数据稀少时训练大型CNNs的范例。我们展示了使用大量数据（图像分类）在辅助任务上进行有监督网络预训练接着以数据稀少的（检测）目标任务对网络进行调参是十分高效的。我们推测“有监督预训练/特定领域调参”范例将对于许多数据稀少问题十分高效。

最后我们得出结论，使用计算机视觉中的分类工具和深度学习（自底向上的区域提案和卷积神经网络）的结合来实现这些结果是十分重要的。这两者是自然的，不可避免的伙伴而不是科学研究中相对的方面。