指导教师：杨涛　　提交时间：2015-3-29

The task of Digital Image Processing

数字图像处理

School of Computer Science

No：1

姓名：刘静

学号：2012302481

班号：10011204

在精确的目标检测和语义分割中存在的丰富的功能层次结构

Ross Girshick　Jeff Donahue　Trevor Darrell　Jitendra Malik
伯克利大学
{rbg,jdonahue,trevor,malik}@eecs.berkeley.edu

摘要

基于规范的 PASCALL VOC 数据集进行测量的目标检测的执行效果在过去几年里已停滞不前。拥有最好效果的方法是一个通常结合了多个低层图像特征和高层次的环境的复杂的集成系统。在这篇论文中，我们提出了一个简单而且可扩展的检测算法，该算法比之前在 2012 年 VOC 中实现了 53.3%的平均精度的最好效果又提高了 30%。我们的方法结合了两个关键的想法：（1）为了定位和分割目标，我们可以在自底向上的区域建议中应用高容量的卷积神经网络。（2）当被标记的训练数据很稀缺的时候，对于一个辅助的任务的监督预训练，随后是特定区域的微调，得到了显著的性能改善。因为我们结合了区域建议和卷积神经网络，所以我们将我们的方法称为 **R-CNN：Regions with CNN features。**完整系统的源码可从以下网站获得：
http://www.cs.berkeley.edu/~rbg/rcnn。

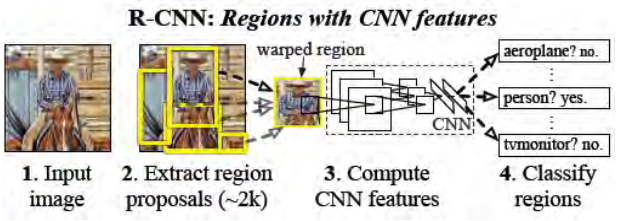

图 1：目标检测系统概要。我们的系统（1）需要输入一个图像,（2）提取大概 2000 个自底向上的区域建议，（3）利用一个卷积神经网络（CNN）为每个建议计算特征，然后（4）用特定类的线性 SVM 对每一个区域进行分类。R-CNN 在 2012 年 PASCAL VOC 上实现了 53.7%的平均精度。作为对比，一个利用空间金字塔和视觉词袋的方法使用相同的区域建议只得到了 35.1%的平均精度。很流行的变形零件模型达到了 33.4%。

1. 介绍

特征问题。过去十年在不同的视觉识别任务的进步大量地都是基于 SIFT 算法和 HOG 算法的应用。但是如果我们看一看规范的视觉识别任务的表现，PASCALL VOC 目标检测，我们就会知道在 2010 年到 2012 年之间，发展很慢，在建立集成系统并采用成功方法的微小不同的过程中有很小的收获。

SIFT 算法和 HOG 算法是块的方向直方图，是一个我们可以将位于 V1 区域的复杂细胞粗略的联系起来的展示方法，其中 V1 区域是灵长类动物视觉通路中第一个皮质区域。但是我们也知道识别发生在那些阶段的下游，这表明了对于计算功能可能存在分层的，多阶段的过程，对于视觉识别甚至会更具信息性。

Fukushima 的神经认知机就是在这样的过程中的一个早期的尝试，它是一个生物激励的层次结构和在模式识别中应用的平移不变的模型。然而，神经认知机缺少一个被监督的训练算法。在 Rumelhart 等人的基础上，LeCun 等人表示通过反向传播出现的随机梯度下降现象对训练卷积神经网络是很有效的，这是延伸神经认知机的一类模型。

卷积神经网络在 20 世纪 90 年代被大量的使用，但是却随着支持向量机的不断发展，日渐衰落了。在 2012 年，由于在大型视觉辨识挑战(ILSVRC)中，卷积神经网络基本上都显示出了很高的图像分类精度，Krizhevsky 等人对卷积神经网络又重新燃起了兴趣。他们的成功是因为通过 120 万张被编辑的图像训练了一个很大的卷积神经网络，还有对 LeCun 的据那几神经网络的一些改进（比如：max(x,0) 矫正非线性和 ）。

在 2012 年大型视觉辨识挑战研讨会期间，图像网络结论的重要意义被大力的争论。争论的焦点如下：在 ImageNet 数据库中应用卷积神经网络分类的结果推广到 PASCAL VOC 挑战赛中的目标识别中，能达到什么程度？

我们通过弥合图像分类和目标识别的差距回答了该问题。这篇论文是第一次通过和基于简单的 HOG 特征的系统进行比较表明卷积神经网络可以在 PASCAL VOC 挑战赛中实现非常好的目标检测效果。为了达到该效果，我们仅关注两个问题：用一个深度网络定位目标、通过较少数量

的注释过的检测数据训练一个高容量的模型。

不同于图像分类，检测要求在一个图像中定位目标。有的方法把定位看做是回归问题。然而，Szegedy 等人和我们的工作都表明这种策略在实践中可能不会很顺利（他们在 2007 年的 VOC 报告会上提出了 30.5%的平均精度，而我们的方法可实现 58.5%）。一种可能的解决方法就是建立一个滑窗检测器。卷积神经网络在过去至少 20 年内一直使用这种方式，尤其是在一些被约束的对象类别上，例如脸部，行人。为了满足高空间分辨率，这些卷积神经网络只有两个卷积和汇聚层。我们也考虑了使用滑窗的方法。然而，在我们拥有 5 个卷积层的网络中，网络单元很多。在输入图像中，这些单元都有很大的接受域（195*195 像素）和步态（32*32 像素），这使得位于滑动窗口模式下的精确定位成为一个开放的技术挑战。

相反，我们通过在识别区域模式下操作解决了卷积神经网络定位的问题，而且已经成功的应用到了目标检测和语义分割。在测试时，我们的方法对输入的图像生成了大概 2000 个与类别无关的区域建议，通过使用一个卷积神经网络，从每一个建议中提取出一个固定长度的特征向量，然后利用特定类别的线性支持向量积对每一个建议进行分类。我们用一个简单的技术（仿射图形变换）从每一个区域建议中计算出一个特定大小的卷积神经网络输入，而不管这个区域的形状是什么。图像 1 是我们方法的一个大体概述，也展示了我们结果中一些最出众的地方。由于我们的系统结合了区域建议和卷积神经网络两部分，我们将它称作 R-CNN：即区域和卷积神经网络特征。

在检测中面临的第二个挑战是稀缺的被标记的数据，而且现在可得到的数量对于训练一个大规模的卷积神经网络效率很低。常规的解决这个问题的方法是使用无监督的预训练，之后进行受监督的微调。这篇论文的第二个原则上的贡献是显示当数据稀缺时，基于大量辅助数据的监督预训练以及随后的基于一个小数据集的特定领域的微调的方法对于高容量的卷积神经网络是一种有效的模式。在我们的试验中，对检测的微调提高了 8 个百分点的平均精度的性能。进行微调之后，相比于进行高度调整后的基于 HOG 的 DPM 目标检测算法实现的 33%的平均精度，我们的系统在 2010 年 VOC 上实现了 54%的平均精度。我们也向读者指出了由 Donahue 等人完成的当时的工作，这些工作表明了 Krizhevsky 的卷积神经网络（无微调）可以被用作一个特征提取的黑箱，而且在一些包括场景分类，细粒度次范畴，域适应的识别任务中表现出良好的性能。

我们的系统效率也非常高。仅有的特定类的计算就是一个很小的矩阵向量积的计算和贪婪非最大抑制。计算的属性是从一些特征得来的，这些特征被所有的类别所共享，而且相比于之前使用的区域特征低两个数量级的维度。

理解了我们的方法失败的模式对于提高它也是至关重要的，所以我们公布了使用 Hoiem 等人的检测分析工具得到的结果。通过这次分析的一个直接的效果，我们证明了一个简单的边框回归方法能够大大地减少主要的错误模式，也就是错误定位。

在发展技术细节之前，我们注意到由于 R-CNN 算法是在区域上进行操作，将其扩展到语义分割的任务上市很自然的。进行了微小的改正后，我们在 PASCAL VOC 分割任务中也获得了很不错的效果。在 2011 年 VOC 测试数据中，我们达到了 47.9%的平均分割精度。

2. 利用 R-CNN 算法进行目标检测

我们的目标检测系统包含三个模块。第一个生成了无类别的区域建议。这些建议定义了一系列的对我们的检测器可用的候选检测。第二个模块是一个大规模的卷积神经网络，可从每一个区域中提取一个定长的特征向量。第三个模块是一系列的特定类的线性支持向量机。在本节中，我们将展示我们为每个模块设计的策略，描述它们在测试时的用处，详细地介绍它们的参数是如何被学习的，最后向大家展示在 2010 年 12 月份 PASCAL VOC 挑战赛的结果。

2.1.模块设计

区域建议。最近很多的论文提出了如何生成无类别的区域建议的方法。例如：对象性[1]，选择搜索[34]，无类别的目标建议[12]，CMPC 算法[5]，多尺度组合分组算法[3]，还有 Ciresan 等人[6]的研究。Ciresan 通过应用卷积神经网络到规则排列的方形的农作物上检测到了有丝分裂的细胞，这是区域建议的一个特殊的例子。因为 R-CNN 相对于特定的区域建议方法是不可知的，

我们使用有选择性的搜索策略以便和之前的检测工作进行受控的比较。(例如[34,36])。

特征提取。利用由 Krizhevsky 等人描述的卷积神经网络的 Caff 实现，我们从每一个区域建议中提取了一个 4096 维度的特征向量。特征的计算是通过 5 个卷积层和 2 个完全连接层向前传播一个均值减去 227*227 RGB 的图像。读者可以参考[22, 23]，以获得更多的网络体系结构的细节。

图 2：从 2007 年 VOC 训练中得到的扭曲的训练样本

为了对一个区域建议计算特征，我们首先必须把这个区域的图像数据转换成一中可以和卷积神经网络（其结构要求输入必须为固定的 227*227 像素大小）兼容的形式。在将所有的任意形状的区域进行可实现的转换后，我们在其中选择了最简单的。不论这些候选区域的大小、高宽比如何，我们扭曲了所有的像素到一个包围盒里以达到要求的大小。在扭曲之前，我们扩大了这个紧密的包围盒，以便在设定的大小时相比原始的盒子有精确的 p 个像素的上下浮动（我们使用的 p 为 16）。图 2 展示了扭曲的训练区域的一个任意的样本。这些补充的材料使得扭曲成为可能。

2.2.检测实验

在测试时，我们对测试的图像进行了选择搜索，提取出了大概 2000 个区域建议（我们在所有的实验中使用了选择搜索的快速模式）。我们对每一个建议进行了变形，并通过卷积神经网络将它向前传播以从所需的层中读出特征。然后，我们使用为每一层训练的支持向量机对每一个提取的特征进行打分。考虑到在一个图像所有已经打分的区域，我们利用贪婪非最大抑制算法（对每一层都是独立的），如果它和一个比学习阈值得分更高的被选择的区域有 IoU 的重叠,这个区域就会被拒绝。

运行分析。两种特性使得检测很有效率。第一，所有的卷积神经网络参数被所有的类别所共同使用。第二，由卷积神经网络计算出来的特征向量相比于其他常见的方法是低维度的。例如，带有视觉词袋编码的空间金字塔模型。在 UVA 检测系统中使用的特征比我们的大了两个数量级（36 万 vs 4 千的维度）。

分享使用的结果就是计算区域建议和特征（在一个图形处理器上的 13 秒的图像或者在一个中央处理器上的 53 秒的图像）花费的时间分散到了所有的级别的计算上了。仅有的特定类的计算是特征、支持向量机权重以及非最大抑制之间的点积。在实际应用中，对一个图像的所有的点积的计算，被分批分散到了一个单个的矩阵和矩阵之间的乘机。特征矩阵典型的是 2000*4096，而支持向量机权重矩阵是 4096*N，其中 N 是类别的个数。

这种分析表明了 R-CNN 算法可以延展到数以千计的目标类别而不用采取类似的技术，例如哈希散列的方法。即使我们 10 万个类别，在一个现代的多核 CPU 上计算出矩阵乘积的结果仅仅花费 10 秒。这种效率不仅是使用了区域建议和被分享的特征的结果。由于高维度的特征，当使用 134G 的内容仅用来存储 10 万的线性预测器，UVA 系统的速度将减慢两个数量级相比于使用 1.5G 的内存存储我们的低维度的特征。

将 R-CNN 算法和 Dean 等人最近利用 DPMs 和哈希散列进行大规模识别的工作进行对比也是很有趣的。他们公布了自己在 2007 年 VOC 挑战赛上对 1 万个牵引器类进行分类时 5 分钟一个图片的速度得到的大概 16%的平均精度。而使用我们的方法，在一个 CPU 上，1 万个检测器大概运行 1 分钟的时间。因为没有进行逼近，平均精度将保持在 59%（3.2 小节）。

2.3.训练

监督预训练。基于一个带有映象级注解（没有包围盒标记）的大规模的辅助数据集（2012 年的 ILSVRC）我们特别地对卷积神经网络进行了预训练。当使用开源 Caffe CNN 库的时候，预训练会被执行。简要的说，我们的卷积神经网络基本上能与 Krizhevsky 等人的研究相媲美，在 2012 年 ILSVRC 验证组获得了高出最好误码率 2.2 个百分点的好成绩。误差是由于在训练过程中的简化。

特定领域的微调。为了使我们的卷积神经网络适应新的任务（检测）和新的领域（扭曲的 VOC 窗口），我们仅仅利用从 VOC 获得的扭曲的区域建议继续进行 CNN 参数的随机梯度下降（SGD）训练。除了用随机初始化的 21 路分类

VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5 [18]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA [34]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
Regionlets [36]	65.0	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54.0	45.9	39.7
SegDPM [16]†	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32.0	30.5	56.4	57.2	65.9	27.0	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	**71.8**	**65.8**	**53.0**	**36.8**	**35.9**	**59.7**	**60.0**	**69.9**	**27.9**	**50.6**	**41.4**	**70.0**	**62.0**	**69.0**	**58.1**	**29.5**	**59.4**	**39.3**	**61.2**	**52.4**	**53.7**

表 1：在 2010 年测试的检测平均精度（%）。因为所有的方法都是使用了选择搜索区域建议，R-CNN 是最直接的和 UVA、Regionlets 进行比较。边框抑制（BB）在 3.4 节进行了描述。在公开时间，SegDPM 在 PASCAL VOC 排行榜上是执行效果最好的。†DPM 和 SegDPM 使用了上下文解码，在其他方法中没有涉及到。

层（20 个 VOC 层加上背景）替换特定图像网络 1000 路分类层，CNN 的架构是没有改变的。我们将所有的区域建议与一个实地的盒子作为正面，其余的作为反面。我们以 0.001 的学习速率（起始预训练率的十分之一）开始了随机梯度下降，这个过程允许微调以便当没有完成初始化时可以不断改进。在每一个随机梯度下降的迭代过程中，我们随机选取 32 个正面的窗口（覆盖所有的类别）和 96 个背景窗口，以此构造一个大小为 128 的小批量系统。我们更喜欢正面窗口的采样，因为它们相对于背景来说太少见了。

目标类别分类器。考虑训练一个二进制的分类器用来检测汽车。很明显一个图像区域里完全是一辆汽车是很难得的现象。同样的，一个与汽车无关的背景区域却很常见。不太清晰的是如何标记一个区域，部分地重叠一辆车。我们用 IoU 重叠阈值解决了这个问题，低于这个值得区域被定义为背景。重叠阈值 0.3 是在一个验证组上通过表格搜索{0,0.1,…,0.5}选择出的。我们发现仔细地搜索这个阈值是很重要的。将它置为 0.5，就像在[34]中，降低了 5 点的平均精度。相似的，将它置为 0 会降低 4 点的平均精度。较好的情况被简单的确定为实地包围盒。

一旦特征被提取出来、训练的标记被应用，对于每一类我们就会优化得到一个线性的支持向量机。因为训练数据太大难以存放在内存中，我们采用标准的硬负开发方法[15, 32]。硬负开发收敛迅速，在实际应用中，所有的图像中仅仅一个传递之后，平均精度就会停止增加。

在辅助材料中我们讨论了为什么阳性和阴性在微调与支持向量机训练中的定义不同。我们也讨论了为什么训练检测分类器比简单的使用从调整好的 CNN 的最终的层中得到的输出更有必要。

2.4. 2010-12 PASCAL VOC 中的结果

随着在 PASCAL VOC 获得最佳实践[13]，我们验证了基于 2007 年的 VOC 数据集所有的设计决策和参数（3.2 节）。对于 2010 年 12 月份 VOC 数据集的最终结果，在 VOC2012 年的训练中我们对卷积神经网络进行了微调，并优化了我们的检测支持向量机。对于两种主要的算法的不同（带有边框回归和不带有边框回归），我们向评测的服务器仅提交了一次测试结果。

表 1 展示了在 2010 年 VOC 全部的结果。我们将我们的方法同 4 条严格的基线进行比较。其中包括 SegDPM[16]，该算法结合了 DPM 检测器和一个语义分割系统的结果，并且使用了附加的检测器间的背景信息和图像分类器的记录信息。最恰当的比较是和来自 Uijlings 等人的 UVA 系统[34]的比较，因为我们的系统使用了相同的区域建议算法。为了对区域进行分类，他们的方法是建立了一个四层的空间金字塔，并用密集的采样 SIFT、扩展对立 SIFT 和 RGB-SIFT 描述符，每一个向量都用一个 4000 个单词的码本来量化。分类的进行是使用直方图交叉核 SVM。相比于他们的多特征、非线性的支持向量机方法，我们在平均精度上实现了一个很大的进步，从原来的 35.1%到现在的 53.7%，当然我们的算法也更加迅速（2.2 节）。我们的方法在 2011 年 12 月 VOC 测试中也达到了相似的效果（53.3%的平均精度）。

3. 可视化，消融，以及错误模式

3.1. 可视化学习特征

第一层的滤波器可以直接被看到，而且很容易理解。它们会捕获导向的边缘和对立的颜色。但是弄清楚随后要介绍的这一层是很有挑战性的。Zeiler 和 Fergus 展示了一种视觉上很吸引人的解卷积的方法在[37]中。我们提出了一个简单的(补充的)非参数的方法能够直接显示出网络在学习什么。

图 3：6 个 $pool_5$ 单元中的最好的区域。接受域和激活值用白色画出。一些单元和概念是相一致的。例如人（行 1）或者文字（行 4）.其他单元捕获纹理和材料特性，例如点阵（行 2）和镜面反射（行 6）。

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN $pool_5$	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc_6	59.3	61.8	43.1	34.0	25.1	53.1	60.6	52.8	21.7	47.8	42.7	47.8	52.5	58.5	44.6	25.6	48.3	34.0	53.1	58.0	46.2
R-CNN fc_7	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT $pool_5$	58.2	63.3	37.9	27.6	26.1	54.1	66.9	51.4	26.7	55.5	43.4	43.1	57.7	59.0	45.8	28.1	50.8	40.6	53.1	56.4	47.3
R-CNN FT fc_6	63.5	66.0	47.9	37.7	29.9	62.5	70.2	60.2	32.0	57.9	47.0	53.5	60.1	64.2	52.2	31.3	55.0	50.0	57.7	63.0	53.1
R-CNN FT fc_7	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN FT fc_7 BB	**68.1**	**72.8**	**56.8**	**43.0**	**36.8**	**66.3**	**74.2**	**67.6**	**34.4**	**63.5**	**54.5**	**61.2**	**69.1**	**68.6**	**58.7**	**33.4**	**62.9**	**51.1**	**62.5**	**64.8**	**58.5**
DPM v5 [18]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
DPM ST [26]	23.8	58.2	10.5	8.5	27.1	50.4	52.0	7.3	19.2	22.8	18.1	8.0	55.9	44.8	32.4	13.3	15.9	22.8	46.2	44.9	29.1
DPM HSC [28]	32.2	58.3	11.5	16.3	30.6	49.9	54.8	23.5	21.5	27.7	34.0	13.7	58.1	51.6	39.9	12.4	23.5	34.4	47.4	45.2	34.3

表 2：在 2007 年 VOC 测试中检测平均精度（%）。1-3 行展示了不带微调的 R-CNN 的效果。4-6 行展示了在 2012 年 ILSVRC 卷积神经网络预训练的结果，然后是在 2007 年 VOC 训练进行微调后的结果。第 7 行包括了一个简单的边框抑制（BB）的阶段，降低了定位误差（3.4 节）。8-10 行展示了 DPM 作为一个有力的基线。第一个仅仅使用了 HOG 算法，而下面两个使用了两个不同的特征学习方法用来补充或者替代 HOG。

这个想法是为了找出在这个网络中一个特别的单元（也就是特征），然后使用它好像在它自己看来它是一个目标检测器。这就是说，我们在我们持有的区域建议的一个很大的集合下（大约 1 千万）计算这个单元的激活量，将这些建议按照从最高到最低的激活量进行排列，执行非最大抑制算法，然后展示得分最高的区域。我们的方法让被选择的单元通过展示 dandified 是哪个输入触发它来“为自己说话”。我们避免平均，是为了能够看到不同的可视模式，深入了解由这个单元得出的不变性质。

我们从层池 5 可以看到单元，这个层池是网络的第五层和最后的卷积层输出的最大汇总。层池 5 特征图是 6*6*256=9216 维度的。不考虑边界影响，每一个池 5 单元在原始 227*227 像素输出上有一个 195*195 像素的可接受域。一个中心池 5 单元会有一个几乎全局的视野，然而一个接近边缘的就会有一个较小的受限制的支持。

在图 3 中的每一行展示了一个池 5 单元的前 16 的激活量，这个池 5 单元室来自我们在 2007VOC 训练中进行过微调的卷积神经网络。256 个功能唯一的单元中的 6 个是可视的（辅助材料包含的会更多）。这些单元室用来展示一个网络学习的典型的样例。在第二行，我们可以看到一个聚集了点面和点阵列的单元。这个单元相对于第三行是一个红色斑点检测器。这里当然也有用于人脸的检测器和更多的抽象模式，例如文字以及带有窗口的三角的结构网络出现时为了学习一种表示，这种表示将少数的类调谐特征和形状，质地，颜色以及材料特性的一种分布式的表示进行了结合。随后的完全连接的层 fc_6 能够对大规模的丰富的特征的组成进行建模。

3.2.消融研究

逐层的性能，不带有微调。为了了解是哪一层对

检测的性能是至关重要的，我们对每一个卷积神经网络最后三层在 2007 年 VOC 数据集的结果都进行了分析。层池 5 会在 3.1 节进行简要的描述。接下来我们将对最后的两层做概述。

层 fc_6 和池 5 是完全连接的。为了计算特征，池 5 特征图乘以了一个 4096*9216 维的权重矩阵（被改造为一个 9216 维的向量），然后加上一个偏差的向量。中间向量是组成分明的半波整流（x←max(0,x)）。

层 fc_7 是这个网络的最后一层。它是通过由 fc_6 计算出来的特征乘以一个 4096*4096 维的权重矩阵，同样的再加上一个偏差向量并应用半波整流得到的。

我们开始研究是通过观察在 PASCAL 上没有进行微调的卷积神经网络的结果，即所有的卷积神经网络的参数仅仅在 2012 年的 ILSVRC 是被预训练过的。逐层地分析性能（表 2 的第 1-3 行）揭示了来自 fc_7 的特征比来自 fc_6 的特征更难推广。这意味着 29%，或者大约 1680 万个卷积神经网络的参数可以被删除而不降低平均精度。更令人惊奇的是，删除 fc_7 和 fc_6 甚至会产生更好的效果，即使池 5 的特征的计算仅仅使用了 6%的卷积神经网络的参数。卷积神经网络的具有代表性的强大能力来自于它的的卷积层，而不是来自那些紧密连接的层。这个发现表明了在计算一个紧密的特征图的潜在的效用。从 HOG 的方面来说，就是一个任意大小的图像的计算仅仅通过使用卷积神经网络的卷积层就可得到。这种表示方法将使滑动窗口检测器的实验可以实现，包括 DPM，池 5 特征的最顶层。

逐层的性能，带有微调。在 2007 年的 VOC 训练中已经调好参数之后，我们现在观察来自我们的卷积神经网络的结果。进步是显著的（表 24-6 行）：微调提高了 8%的平均精度，实现了 54.2%。对于 fc_7 和 fc_6，微调带来的效果比池 5 大的多，这表明了从图像集中学习到的池 5 特征更一般，大部分的进步是通过学习在它们之上的特定领域的非线性的分类器得到的。

和最近的特征学习方法的比较。相对较少的特征学习方法在 PASCAL VOC 检测中已经被尝试过了。我们观察了两种最近出现的建立在可变形的部分模型的方法。作为参考，我们也包含了标准的基于 HOG 的 DPM 算法的结果。

第一个 DPM 特征学习方法，DPM ST [26]，增强的 HOG 特征和和“素描令牌”的概率直方图。更直观地来说，一个素描令牌就是一个穿过一个图像块中心的紧密分布的轮廓。素描令牌的概率在每一个像素上的计算是通过一个任意森林，它被训练用来将 35*35 像素块分类成一个 150 个素描令牌或背景。

第二个方法，DPM HSC，用稀疏编码直方图（HSC）替代了 HOG。为了计算一个 HSC，我们利用了 1007*7 像素（灰度）原子的学习字典在每一个像素级上解决了稀疏编码的激活。结果的激活使用了三种方式进行整流（全波和半波），然后空间混合，l_2 单元的归一化，最后是能量转换（x←sign(x) |x|α）。

所有的 R-CNN 变型的性能比三种 DPM 基线（表 2 的 8-10 行）更好，包括使用了特征学习的两种。相比于仅仅使用了 HOG 特征的 DPM 最新的版本的性能，我们的平均精度高出 20 个百分点：54.2% vs. 33.7%——相对进步了 61%。HOG 和素描令牌的结合比单独使用 HOG 多出 2.5 个平均精度点，然而 HSC 比 HOG 提高了 4 个平均精度点（当和他们私人的 DPM 基线进行内部比较时，都使用不公开的 DPM 的实现方法，而且这个 DPM 比开源的版本[18]的性能差）。这些方法达到的平均精度分别是 29.1%和 34.3%。

3.3.错误检测的分析

为了暴露我们所使用的方法的错误模式，理解微调是如何影响它们的，看到我们的错误类型和 DPM 相比是怎样的，我们应用了 Hoiem 等人[21]研究的效果很好的检测分析工具。这个分析工具的完整概述超过了这篇论文的范围，但是我们鼓励读者去查询以便理解一些细节（例如“归一化 AP”）。因为这个分析已经充分吸收在在相关图篇的上下文中，所以我们将在图 4 和图 5 的说明中进行讨论。

3.4.边框抑制

基于对错误的分析，我们实现了一个很简单的方法去降低定位错误的发生。由在 DPM 中使用的边框回归得到的灵感，我们训练了一个线性回归模型，用来预测一个新的检测窗口，这个窗口给出了用来进行选择搜索区域建议的池 5 特征。在辅助材料中给出了完整的细节。在表 1，表 2 和图 4 的结果显示了这种简单的方法适合大量的错误定位检测的情况，并且提高了 3 到 4 点的平均精度。

4.语义分割

区域分类对语义分割来说是一个标准的技术，允许我们很容易的应用 R-CNN 算法到 PASCAL VOC 分割挑战赛中。为了和最近出现的最重要的语义分割系统（称作 O_2P 二阶池）[4]进行一个直接的比较，我们在他们的开源框架下进行了实验。O_2P 使用 CPMC，每个图像能够生成 150 个区域建议，然后预测每一个区域的质量，在每一层，则使用支持向量回归（SVR）。他们的方法的高性能是由于 CPMC 区域的质量以及强大的二阶池的多特征类型（丰富了 SIFT 和 LBP 的变化特性）。我们也注意到 Farabet 等人最近证明了在一些紧密的场景标记的数据集上（不包括 PASCAL）使用卷积神经网络作为一个多尺度像素级的分类器能得到了不错的结果。

我们在[2, 4]研究的基础上，扩展 PASCAL 分割训练集使其包括由 Hariharan 等人[20]提供的额外的注释。设计决策和参数交替对 2011 年 VOC 验证集进行验证。最终的结果只进行一次评估。

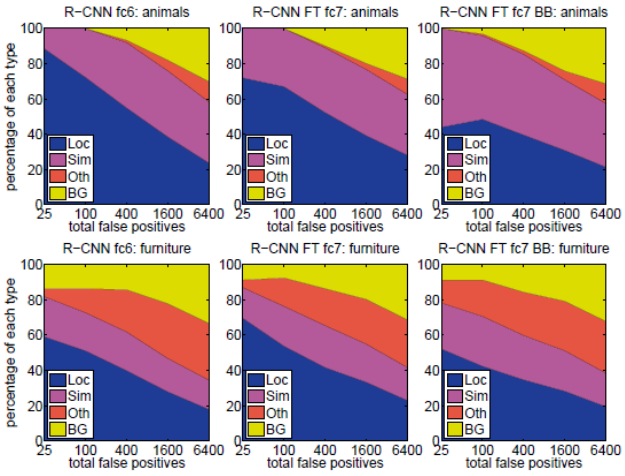

图 4：排名第一的假阳性（FP）类型的分布。每张图显示了 FP 类型的不断变化的分布，因为更多的 FP 被考虑进来由于降低的分数。每一个 FP 被分成 1-4 类：Loc——效果很差的定位（一个带有 IoU 值重叠的检测，重叠发生在 0.1 和 0.5 之间正确的类，或者一个副本）；Sim——和相似类的混淆；Oth——和不相似的目标类别的混淆；BG——由背景触发的假阳性。和 DPM 相比（见[21]），由差的定位导致的很多的错误，而不是和背景或者其他目标的混淆，表明了 CNN 特征比 HOG 更有判别性。就像由于使用了自底向上的区域建议和从为整个图像的分类的 CNN 预训练中学习得到的位置不变形，定位也就变得松散。栏目 3 展示了我们简单的边框回归的方法是如何弥补定位错误的。

分割中的 CNN 特征。我们评价了三种用来计算 CMPC 区域特征的策略，所有的策略都是通过将区域周围的正方形的窗口弯曲到 227*227 的大小开始的。第一个策略（full）忽略了区域的形状，直接在弯曲后的窗口上计算 CNN 特征，也就是我们为检测所做的事。然而，这些特征忽略了不是正方形形状的区域。两个区域可能会有非常相似的边界，尽管它们有很少的交叠。因此，第二个策略（fg）仅仅在一个进行了前景屏蔽的区域上计算 CNN 特征。我们用平均输入来代替背景，因此背景区域在进行平均减法后结果是 0。第三个策略（full+fg）就是简单了结合了 full 和 fg 特征，我们的实验验证了它们的互补性。

	full R-CNN		Fg R-CNN		full+fg R-CNN	
O_2P[4]	fc_6	fc_7	fc_6	fc_7	fc_6	fc_7
46.4	43.0	42.5	43.7	42.1	**47.9**	45.8

表 3：在 2011 年 VOC 验证集上的平均分割精度（%）。第一列展示了 O_2P；2-7 列使用了我们在 2012 年 ILSVRC 的 CNN 预训练。

2011 年 VOC 上的结果。表 3 展示了我们在 2011 年 VOC 验证集上比较 O_2P 的结果的一个概要。（完整的每个章节的结果请阅读补充材料。）使用了每一个特征计算的策略，fc_6 层总是比 fc_7 的效果更好，而且接下来的讨论提到了 fc_6 特征。fg 策略稍稍比 full 策略好些，表明了屏蔽的区域形状提供了一个很强的信号，和我们的预想一致。然而，full+fg 实现了 47.9%的平均精度，和我们最好的结果相差 4.2%（也比 O_2P 较好），表明了由 full 特征提供的内容更具信息性，即使得到了 fg 特征。值得注意的是，在单核的 CPU 上，用我们的 full+fg 特征训练 20 个支持向量机大概花费一个小时的时间，对于训练 O_2P 特征使用 10 多个小时。

在表 4 中我们展示了在 2011 年 VOC 测试集上的结果，比较了我们有最好效果的方法，fc_6（full+fg），和两条基线的差别。在 21 个类别中我们的方法有 11 个都实现了最高的分割精度，而且得到了总体的最高的分割精度 47.9%，总体的精度是计算所有类别的平均得到的（但是可能是任何合理的错误余量导致的和 O_2P 的关系）。通过微调仍有可能得到更好的效果。

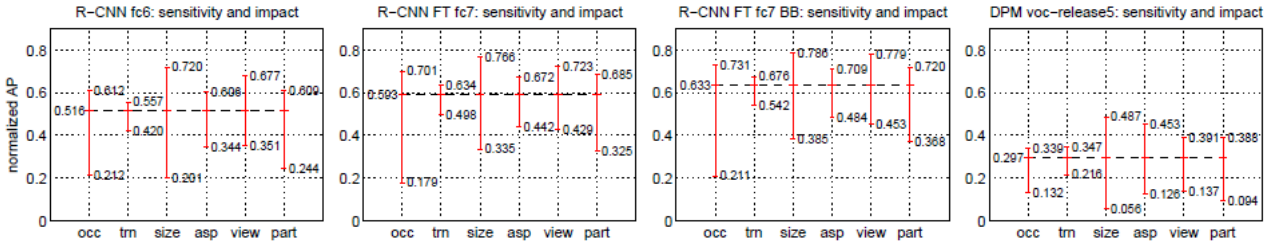

图 5：对目标特性的敏感性。每一个图表展示了在 6 个不同的目标特性（阻塞，截断，边框面积，长宽比，观点，部分能见度）中最高和最低的执行子集的归一化的 AP（见[21]）。我们展示了我们带有和不带有微调以及边框抑制的方法的图表。总的来看，微调没有降低敏感性（大和小之间的不同），但是没有显著提高几乎所有的特性的最高和最低执行子集。这表明了微调不仅仅提高了最低执行子集的长宽比和边框面积，作为一个基于我们如何弯曲网络的输入的可能的猜测。相反，微调提高了所有的特性包括阻塞，截断，观点和部分能见度的鲁棒性。

5.结论

最近几年，目标检测的发展一直停滞不前。拥有最好效果的系统是一个通常结合了多个低层图像特征和来自目标检测器和场景分类器的高层次的内容的复杂的集成系统。在这篇论文中，我们展示了一个简单而且可扩展的检测算法，该算法比之前在 2012 年 VOC 中最好效果的平均精度又提高了 30%。

我们主要通过两个想法实现了这种效果。第一个就是在自底向上的区域建议中应用高容量的卷积神经网络，以便定位和分割目标。第二个是为当标记训练数据稀缺时需要训练大量的 CNN 建立的一个范式。我们展示了预训练网络并带有监督是非常有效的——对于一个有丰富的数据（图像分类）的辅助任务，当数据稀缺时（检测）我们可以微调目标任务的神经网络。我们猜测“辅助预训练或者特定类的微调”的这种范式对很多的数据稀缺的视觉问题都会很有效。

通过结合视觉计算的分类工具和深度学习（自底向上的区域建议和卷积神经网络）实现的效果是有重要意义的。这两者不是科学探究的对立，而是自然与必然的合作伙伴。

致谢。这个研究受到了很多机构的支持，DARPA Mind’s Eye，MSEE programs，NSF awards IIS-0905647，IIS-1134072，and IIS-1212798，MURI N000014-10-1-0933，同时也得到了来自东京的支持。在这个研究中用到的 GPU 是由 NVIDIA 公司慷慨赞助。

参考

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. TPAMI, 2012.

[2] P. Arbel´aez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik.
Semantic segmentation using regions and parts. In CVPR, 2012.

[3] P. Arbel´aez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale
combinatorial grouping. In CVPR, 2014.

[4] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In ECCV, 2012.

[5] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation
using constrained parametric min-cuts. TPAMI, 2012.

[6] D. Cires¸an, A. Giusti, L. Gambardella, and J. Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In MICCAI, 2013.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.

[8] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In CVPR, 2013.

[9] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Competition 2012 (ILSVRC2012). http://www.image-net.org/challenges/LSVRC/2012/.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet:
A large-scale hierarchical image database. In CVPR, 2009.

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and
T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In ICML, 2014.

[12] I. Endres and D. Hoiem. Category independent object proposals. In ECCV, 2010.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman.
The PASCAL Visual Object Classes (VOC) Challenge.
IJCV, 2010.

[14] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical
features for scene labeling. TPAMI, 2013.

[15] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. TPAMI, 2010.

[16] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation
for top-down detection. In CVPR, 2013.
[17] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift
in position. Biological cybernetics, 36(4):193–202, 1980.
[18] R. Girshick, P. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://www.cs.
berkeley.edu/~rbg/latent-v5/.
[19] C. Gu, J. J. Lim, P. Arbel´aez, and J. Malik. Recognition using regions.
In CVPR, 2009.
[20] B. Hariharan, P. Arbel´aez, L. Bourdev, S. Maji, and J. Malik. Semantic
contours from inverse detectors. In ICCV, 2011.
[21] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In ECCV. 2012.
[22] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/,
2013.
[23] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification
with deep convolutional neural networks. In NIPS, 2012.
[24] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard,
and L. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Comp., 1989.
[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998.
[26] J. J. Lim, C. L. Zitnick, and P. Doll´ar. Sketch tokens: A learned mid-level representation for contour and object detection. In CVPR,
2013.
[27] D. Lowe. Distinctive image features from scale-invariant keypoints.
IJCV, 2004.
[28] X. Ren and D. Ramanan. Histograms of sparse codes for object detection.
In CVPR, 2013.
[29] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face
detection. TPAMI, 1998.
[30] D. E. Rumelhart, G. E. Hinton, and R. J.Williams. Learning internal
representations by error propagation. Parallel Distributed Processing, 1:318–362, 1986.
[31] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian
detection with unsupervised multi-stage feature learning. In CVPR,
2013.
[32] K. Sung and T. Poggio. Example-based learning for view-based human
face detection. Technical Report A.I. Memo No. 1521, Massachussets Institute of Technology, 1994.
[33] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In NIPS, 2013.
[34] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective
search for object recognition. IJCV, 2013.
[35] R. Vaillant, C. Monrocq, and Y. LeCun. Original approach for the localisation of objects in images. IEE Proc on Vision, Image, and Signal Processing, 1994.
[36] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object
detection. In ICCV, 2013.
[37] M. Zeiler, G. Taylor, and R. Fergus. Adaptive deconvolutional networks
for mid and high level feature learning. In CVPR, 2011.