

在ILSVRC

2012研讨会上进行了讨论的核心问题可以总结如下：ImageNet上的CNN分类结果在多大程度上可以推广到PASCAL

VOC挑战中的目标检测结果？我们通过弥合图像分类与目标检测之间的差距来回答这个问题。本文是第一个展示CNN相比基于简单HOG特征的系统，可以在PASCAL VOC上显著提高目标检测性能的研究。为了实现这一结果，我们专注于两个问题：使用深度网络定位对象以及仅利用少量标注的目标检测数据训练高容量模型。

与图像分类不同，检测需要在图像中定位（可能是多个）对象。一种方法是将定位视为回归问题。然而，Szegedy等人的工作[38]（与我们的工作同时进行）表明，这种方法在实践中可能表现不佳（他们报告在VOC 2007上的mAP为30.5%，而我们的方法达到了58.5%）。另一种方法是构建滑动窗口检测器。CNN至少在过去的二十年里以这种方式用于受限的对象类别，例如人脸[32, 40]和行人[35]。为了保持高空间分辨率，这些CNN通常只有两层卷积和池化层。我们也考虑采用滑动窗口的方法。然而，我们的网络（具有五层卷积层）中高层单元在输入图像中有非常大的感受野（ 195×195 像素）和步幅（ 32×32 像素），这使得在滑动窗口范式内进行精确定位成为一个开放的技术难题。

相反，我们通过在“基于区域识别”的范式[21]内操作来解决CNN定位问题，该范式在目标检测[39]和语义分割[5]方面都取得了成功。在测试时，我们的方法为输入图像生成大约2000个独立于类别的区域提议，使用CNN从每个提议中提取固定长度的特征向量，然后使用特定于类别的线性SVM对每个区域进行分类。我们使用简单的技术（仿射图像变形）从每个区域提议中计算固定大小的CNN输入，无论区域的形状如何。图1概述了我们的方法并突出了部分结果。由于我们的系统结合了区域提议和CNN，我们将该方法命名为R-CNN：带有CNN特征的区域。

在这篇更新版本的论文中，我们在200类ILSVRC2013检测数据集上运行R-CNN，并与最近提出的OverFeat[34]检测系统进行直接比较。OverFeat使用滑动窗口CNN进行检测，在此之前一直是ILSVRC2013检测的最佳方法。我们展示了R-CNN显著优于OverFeat，mAP分别为31.4%和24.3%。

检测面临的第二个挑战是标注数据不足。

数据稀缺，目前可用的数据量不足以训练大型卷积神经网络（CNN）。解决这个问题的传统方法是使用无监督预训练，随后进行有监督微调（例如，[35]）。本文的第二个主要贡献是证明，在数据稀缺的情况下，先在大规模辅助数据集（ILSVRC）上进行有监督预训练，然后在小规模特定领域数据集（PASCAL）上进行领域特定微调，是一种有效的学习高容量CNN的范式。在我们的实验中，检测任务的微调使平均精度（mAP）提高了8个百分点。经过微调后，我们的系统在VOC 2010上的mAP达到了54%，而高度优化的基于HOG的可变形部件模型（DPM）[17]

[20]仅为33%。我们还提请读者参阅Donahue等人的同期工作[12]，他们展示了Krizhevsky的CNN可以在不进行微调的情况下作为黑盒特征提取器，在包括场景分类、细粒度子分类和领域适应在内的多个识别任务中表现出色。

我们的系统也非常高效。唯一的类别特定计算是一个合理大小的矩阵向量乘法和贪婪非极大值抑制。这种计算特性源于所有类别共享的特征，并且这些特征的维度比之前使用的区域特征低两个数量级（参见[39]）。理解我们方法的失败模式对于改进它至关重要，因此我们报告了Hoiem等人[23]的检测分析工具的结果。作为这一分析的直接结果，我们展示了简单的边界框回归方法显著减少了定位错误，这是主要的错误模式。

在展开技术细节之前，我们注意到由于R-CNN操作于区域，因此自然可以将其扩展到语义分割任务。经过一些小的修改，我们在PASCAL VOC分割任务上也取得了具有竞争力的结果，在VOC 2011测试集上的平均分割准确率为47.9%。

我们的目标检测系统由三个模块组成。

第一个模块生成类别无关的区域提议。

这些提议定义了可供检测器使用的候选检测集合。第二个模块是一个大型卷积神经网络，从每个区域提取固定长度的特征向量。第三个模块是一组类特定的线性支持向量机（SVM）。在本节中，我们介绍了每个模块的设计决策，描述了它们在测试时间的使用方法，详细说明了如何学习其参数，并展示了在PASCAL VOC 2010-12和ILSVRC2013上的检测结果。



例子包括：objectness[1]、选择性搜索[39]、类别无关的对象提议[14]、约束参数最小割（CPMC）[5]、多尺度组合分组[3]，以及Ciresan等人[6]，他们通过将卷积神经网络（CNN）应用于规则间隔的正方形裁剪来检测有丝分裂细胞，这些裁剪是区域提议的一个特例。虽然R-CNN对具体的区域提议方法是无感知的，但我们使用选择性搜索来进行与先前检测工作（例如[39, 41]）的受控比较。

目标类别分类器。考虑训练一个二元分类器来检测汽车。

显然，紧密包围一辆汽车的图像区域应该是一个正例。类似地，显然一个与汽车无关的背景区域应该是一个负例。不太清楚的是如何标注部分重叠汽车的区域。我们通过交并比（IoU）重叠阈值解决了这个问题，在该阈值之下，区域被定义为负例。重叠阈值0.3是通过对{0, 0.1, ..., 0.5}进行网格搜索选择的验证集上的最佳值。我们发现仔细选择这个阈值很重要。将其设置为0.5（如[39]中所述）会使得平均精度均值（mAP）下降5个百分点。同样，将其设置为0会使得mAP下降4个百分点。正例简单地定义为每个类别的真实边界框。

一旦提取了特征并应用了训练标签，我们针对每种类别优化一个线性支持向量机（SVM）。由于训练数据太大而无法全部放入内存，我们采用了标准的难负样本挖掘方法[17,

37]。难负样本挖掘收敛得很快，并且实际上在遍历所有图像后mAP停止增加。

在附录B中，我们讨论了为什么在微调和SVM训练中正负样本的定义不同。我们还讨论了训练检测SVM而不是简单使用微调CNN的最终softmax层输出所涉及的权衡。

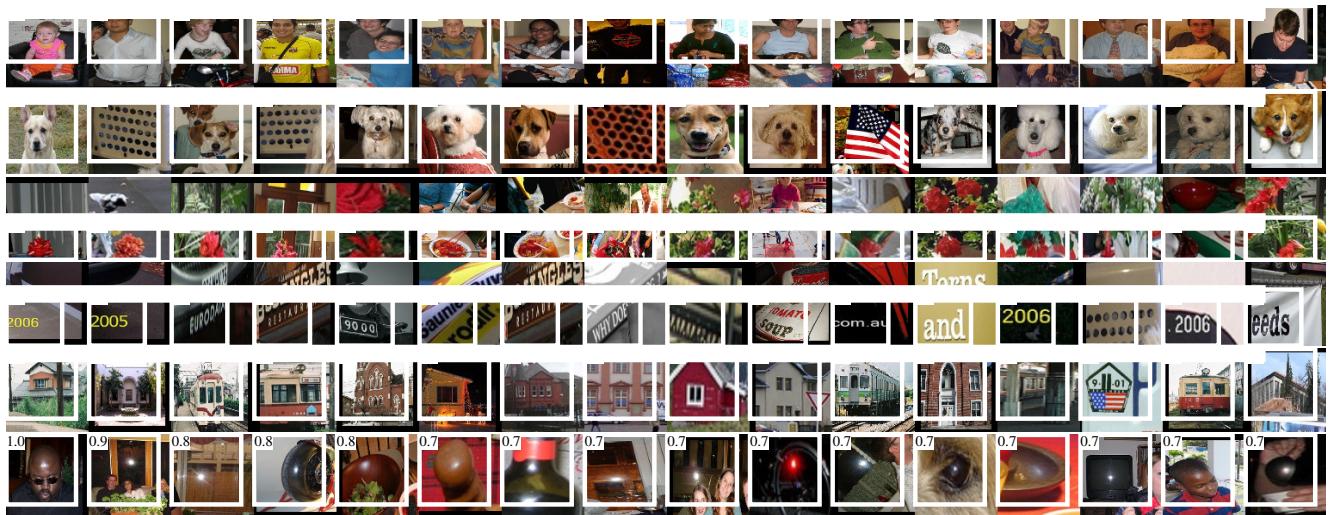
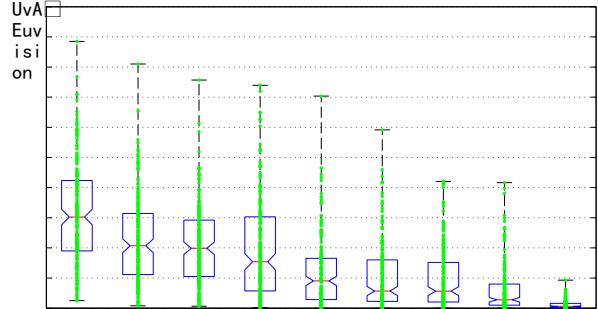
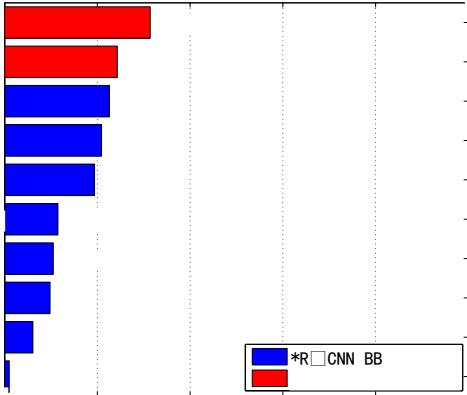


Figure 4: Top regions for six pool₅ units. Receptive fields and activation values are drawn in white. Some units are aligned to concepts, such as people (row 1) or text (4). Other units capture texture and material properties, such as dot arrays (2) and specular reflections (6).

我们可视化了来自第五层和最终卷积层的最大池化输出层 pool5 的单元。

pool5 特征图是 $6 \times 6 \times 256 = 9216$ 维的。忽略边界效应，每个 pool5 单元在原始 227×227 像素输入中的感受野为 195×195 像素。中央的 pool5

单元具有几乎全局视野，而接近边缘的单元则有较小且被截断的支持区域。

图 4 中的每一行显示了我们在 VOC 2007 训练验证集上微调的 CNN 的一个 pool5 单元的前 16 个激活值。可视化了 256 个功能上独特的单元中的 6 个（附录 D）。

包含更多）。这些单元被选出来展示网络所学习内容的一个代表性样本。在第二行中，我们可以看到一个对狗脸和点阵触发的单元。第三行对应的单元是一个红色斑块检测器。还有用于检测人脸、文字等更抽象模式以及带窗户的三角结构的检测器。该网络似乎学习了一种结合少量类别调谐特征与形状、纹理、颜色和材料属性的分布式表示的表示形式。后续的全连接层 fc6 具有能力建模这些丰富特征的大组合集。

我们应用了 Hoiem 等人提出的优秀检测分析工具 [23]，以揭示我们方法的错误模式，理解微调如何改变这些模式，并比较我们的错误类型与 DPM 的差异。该分析工具的完整总结超出了本文的范围，建议读者查阅 [23] 了解一些更详细的细节（例如“标准化 AP”）。由于该分析最好结合相关的图表来理解，因此我们在图 5 和图 6 的图例中进行讨论。

本文中的大多数结果使用了来自 Krizhevsky 等人的网络架构 [25]。然而，我们发现架构的选择对 R-CNN 检测性能有很大影响。在表 3 中，我们展示了使用 Simonyan 和 Zisserman 最近提出的 16 层深度网络在 VOC 2007 测试集上的结果 [43]。该网络是近期 ILSVRC 2014 分类挑战中的顶级表现者之一。该网络具有同质结构，由 13 层 3×3 卷积核组成，并间插有五层最大池化层，顶部配有三层全连接层。我们将此网络称为 “O-Net” (OxfordNet)，并将基准网络称为 “T-Net” (TorontoNet)。

¹<https://github.com/BVLC/caffe/wiki/Model-Zoo>

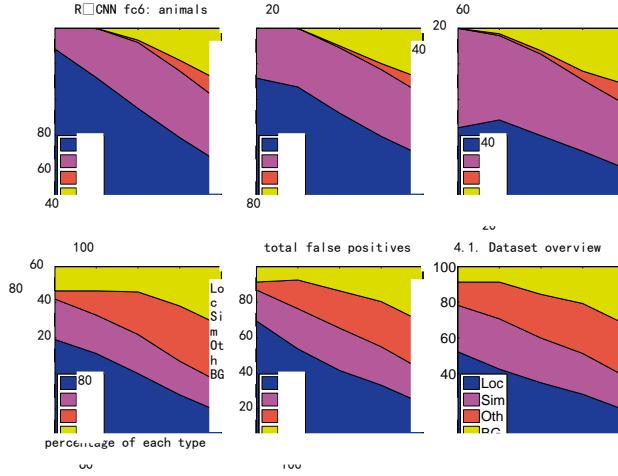
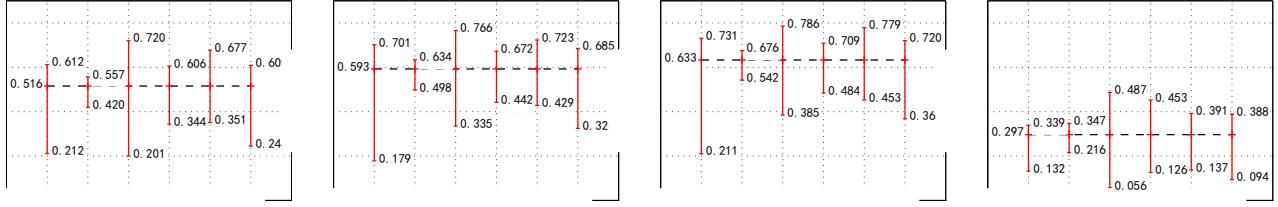


Figure 5: Distribution of top-ranked false positive (FP) types. Each plot shows the evolving distribution of FP types as more FPs are considered in order of decreasing score. Each FP is categorized into 1 of 4 types: Loc—poor localization (a detection with an IoU overlap with the correct class between 0.1 and 0.5, or a duplicate); Sim—confusion with a similar category; Oth—confusion with a dissimilar object category; BG—a FP that fired on background. Compared with DPM (see [23]), significantly more of our errors result from poor localization, rather than confusion with background or other object classes, indicating that the CNN features are much more discriminative than HOG. Loose localization likely results from our use of bottom-up region proposals and the positional invariance learned from pre-training the CNN for whole-image classification. Column three shows how our simple bounding-box regression method fixes many localization errors.

PASCAL VOC, requiring choices about how to use it. Since these decisions are non-trivial, we cover them in this section.

4.1. Dataset overview

The ILSVRC2013 detection dataset is split into three sets: train (395,918), val (20,121), and test (40,152), where the number of images in each set is in parentheses. The

val and test splits are drawn from the same image distribution. These images are scene-like and similar in complexity (number of objects, amount of clutter, pose variability, etc.) to PASCAL VOC images. The val and test splits are exhaustively annotated, meaning that in each image all instances from all 200 classes are labeled with bounding boxes. The train set, in contrast, is drawn from the ILSVRC2013 *classification* image distribution. These images have more variable complexity with a skew towards images of a single centered object. Unlike val and test, the train images (due to their large number) are not exhaustively annotated. In any given train image, instances from the 200 classes may or may not be labeled. In addition to these image sets, each class has an extra set of negative images. Negative images are manually checked to validate that they do not contain any instances of their associated class. The negative image sets were not used in this work. More information on how ILSVRC was collected and annotated can be found in [11, 36].

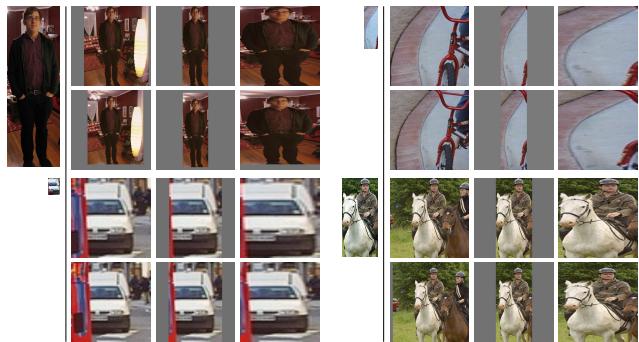
The nature of these splits presents a number of choices for training R-CNN. The train images cannot be used for hard negative mining, because annotations are not exhaustive. Where should negative examples come from? Also, the train images have different statistics than val and test. Should the train images be used at all, and if so, to what extent? While we have not thoroughly evaluated a large number of choices, we present what seemed like the most obvious path based on previous experience.

Our general strategy is to rely heavily on the val set and use some of the train images as an auxiliary source of positive examples. To use val for both training and validation, we split it into roughly equally sized “val₁” and “val₂” sets. Since some classes have very few examples in val (the smallest has only 31 and half have fewer than 110), it is important to produce an approximately class-balanced partition. To do this, a large number of candidate splits were generated and the one with the smallest maximum relative

²Relative imbalance is measured as $|a - b|/(a + b)$ where a and b are class counts in each half of the split.

使用多尺度金字塔的正方形区域提议，并将每个类别的边界框回归器改为单个边界框回归器，那么这些系统将会非常相似（除了它们的训练方式可能存在一些显著差异：CNN检测微调、使用SVM等）。值得注意的是，OverFeat在速度上相对于R-CNN具有显著优势：它大约快了9倍，根据引用[34]中的数据，每张图片处理时间为2秒。这种速度上的提升来源于OverFeat的滑动窗口（即区域提议）在图像级别上没有被扭曲，因此可以在重叠窗口之间轻松共享计算。通过以卷积的方式在整个网络中运行任意大小的输入来实现共享。加快R-CNN的速度可以通过多种方式实现，这仍然是未来的工作。

区域分类是一种标准的技术，用于语义分割，使我们能够轻松地将R-CNN应用于PASCAL VOC分割挑战。为了与当前领先的语义分割系统（称为O2P，即“二级池化”）[4]进行直接比较，我们在他们的开源框架内工作。O2P使用CPMC为每张图像生成150个区域建议，然后使用支持向量回归（SVR）预测每个类别的每个区域的质量。他们方法的高性能归因于CPMC区域的质量以及多种特征类型的强大的二级池化（SIFT和LBP的丰富变体）。我们还注意到，Farabet等人[16]最近在几个密集场景标注数据集（不包括PASCAL）上使用CNN作为多尺度逐像素分类器取得了良好结果。我们遵循[2, 4]，并将PASCAL分割训练集扩展为包含Hariharan等人[22]提供的额外标注。设计决策和超参数在VOC 2011验证集上进行了交叉验证。最终测试结果只评估了一次。



致谢。本研究部分得到了DARPA Mind's Eye和MSEE项目、NSF奖项IIS-0905647、IIS-1134072和IIS-1212798以及MURI N000014-10-1-0933的支持，并且得到了丰田公司的支持。本研究中使用的GPU由NVIDIA公司慷慨捐赠。

我们使用一个简单的边界框回归阶段来提高定位性能。在使用类别特定的检测SVM对每个选择性搜索建议进行评分后，我们利用类别特定的边界框回归器预测一个新的检测边界框。

这与可变形部件模型[17]中使用的边界框回归在精神上相似。两种方法的主要区别在于，这里我们从CNN计算出的特征进行回归，而不是从推断出的DPM部件位置计算出的几何特征。

我们的训练算法输入是一组N个训练对 $\{(P_i, G_i)\}_{i=1, \dots, N}$ ，其中 $P_i = (P_{ix}, P_{iy}, P_{iw},$

$P_{ih})$ 指定了提议 P_i 的边界框中心的像素坐标及其宽度和高度（以像素为单位）。接下来，除非需要，我们将省略上标 i 。每个真实边界框 G 也是以同样的方式指定： $G = (G_x, G_y, G_w, G_h)$ 。我们的目标是学习一种将提议框 P 映射到真实边界框 G 的变换。

我们用四个函数 $dx(P)$, $dy(P)$, $dw(P)$ 和 $dh(P)$ 来参数化这种变换。前两个函数指定 P 的边界框中心的尺度不变平移，而后两个函数则指定 P 的边界框宽度和高度的日志空间平移。学习了这些函数之后，我们可以通过对输入提议 P 应用该变换将其转换为预测的真实边界框 \hat{G} 。

图12展示了20个pool5单元的附加可视化结果。对于每个单元，我们显示了从VOC 2007测试集中大约1000万个区域中最大化激活该单元的24个区域提议。我们通过其在 $6 \times 6 \times 256$ 维pool5特征图中的(y, x, channel)位置来标记每个单元。在每个通道内，CNN对输入区域计算完全相同的函数，而(y, x)位置的变化仅影响感受野。

- [13] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg 和 C. Schmid. Web 规模图像搜索中 GIST 描述符的评估. 在 ACM 国际图像与视频检索会议论文集, 2009. 13
- [14] I. Endres 和 D. Hoiem. 类别无关的对象提议. 在 ECCV, 2010. 3
- [15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn 和 A. Zisserman. PASCAL 视觉对象类 (VOC) 挑战. IJCV, 2010. 1, 4
- [16] C. Farabet, C. Couprie, L. Najman 和 Y. LeCun. 学习层次特征用于场景标注. TPAMI, 2013. 10
- [17] P. Felzenszwalb, R. Girshick, D. McAllester 和 D. Ramanan. 基于部件的区分训练模型进行目标检测. TPAMI, 2010. 2, 4, 7, 12
- [18] S. Fidler, R. Mottaghi, A. Yuille 和 R. Urtasun. 自底向上分割用于自顶向下检测. 在 CVPR, 2013. 4, 5
- [19] K. Fukushima. Neocognitron:
一种不受位置移动影响的模式识别自组织神经网络模型.
生物控制论, 36(4):193 - 202, 1980. 1
- [20] R. Girshick, P. Felzenszwalb 和 D. McAllester. 区分训练可变形部件模型, 版本 5.
<http://www.cs.berkeley.edu/~rbg/latent-v5/>. 2, 5, 6, 7
- [21] C. Gu, J. J. Lim, P. Arbeláez 和 J. Malik. 使用区域进行识别. 在 CVPR, 2009. 2
- [22] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji 和 J. Malik. 从逆检测器生成语义轮廓. 在 ICCV, 2011. 10
- [23] D. Hoiem, Y. Chodpathumwan 和 Q. Dai. 目标检测器误差诊断. 在 ECCV. 2012. 2, 7, 8
- [24] Y. Jia. Caffe:
一个开源卷积架构用于快速特征嵌入.
<http://caffe.berkeleyvision.org/>, 2013. 3
- [25] A. Krizhevsky, I. Sutskever 和 G. Hinton. 利用深度卷积神经网络进行 ImageNet 分类. 在 NIPS, 2012. 1, 3, 4, 7
- [26] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard 和 L. Jackel. 反向传播应用于手写邮政编码识别. 神经计算, 1989. 1
- [27] Y. LeCun, L. Bottou, Y. Bengio 和 P. Haffner. 梯度下降法在文档识别中的应用. IEEE 会议录, 1998. 1
- [28] J. J. Lim, C. L. Zitnick 和 P. Dollár. 笔记符号: 用于轮廓和目标检测的学习中级表示. 在 CVPR, 2013. 6, 7

