Rich feature hierarchies for accurate object detection and semantic segmentation
用于精确物体定位和语义分割的丰富特征层次结构

论文引用：

Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

ABSTRACT

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012 -- achieving a mAP of 53.3%. Our approach combines two key insights: (1) one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. Since we combine region proposals with CNNs, we call our method R-CNN: Regions with CNN features. We also present experiments that provide insight into what the network learns, revealing a rich hierarchy of image features. Source code for the complete system is available at https://www-cs-berkeley-edu.vpn.seu.edu.cn/~rbg/rcnn.

摘要

过去几年，在权威数据集 PASCAL 上，物体检测的效果已经达到一个稳定水平。效果最好的方法是融合了多种低维图像特征和高维上下文环境的复杂融合系统。在这篇论文里，我们提出了一种简单并且可扩展的检测算法，可以将 mAP 在 VOC2012 最好结果的基础上提高 30%以上——达到了 53.3%。我们的方法结合了两个关键的因素：(1) 在候选区域上自下而上使用大型卷积神经网络(CNNs)，用以定位和分割物体; (2) 当带标签的训练数据不足时，先针对辅助任务进行有监督预训练，再进行特定任务的调优，就可以产生明显的性能提升。由于我们将 region proposal 与 CNN 结合起来，我们将方法称为 R-CNN：Regions with CNN features（具有 CNN 特征的区域）。我们还展开实验，提供对网络学习内容的深入了解，揭示丰富的图像特征层次结构。有关完整系统的源代码，请访问 https://www-cs-berkeley-edu.vpn.seu.edu.cn/~rbg/rcnn。

1. Introduction

1. 引文

Features matter. The last decade of progress on various visual recognition tasks has been based considerably on the use of SIFT [27] and HOG [7]. But if we look at performance on the canonical visual recognition task, PASCAL VOC object detection [13], it is generally acknowledged that progress has been slow during 2010–2012, with small gains obtained by building ensemble systems and employing minor variants of successful methods.

特征很重要。在过去十年，各类视觉识别任务基本都建立在对 SIFT[27]和 HOG[7]特征的使用。但如果我们关注一下 PASCAL VOC 对象检测[13]这个经典的视觉识别任务，就会发现，2010-2012 年进展缓慢，取得的微小进步都是通过组合不同模型和使用已有方法的变种才达到的。

SIFT and HOG are blockwise orientation histograms, a representation we could associate roughly with complex cells in V1, the first cortical area in the primate visual pathway. But we also know that recognition occurs several stages downstream, which suggests that there might be hierarchical, multi-stage processes for computing features that are even more informative for visual recognition.

SIFT 和 HOG 是逐块定向直方图(blockwise orientation histograms)，一种类似大脑初级皮层 V1 层复杂细胞的表示方法。但我们知道识别发生在多个下游阶段，(我们是先看到了一些特征，然后才意识到这是什么东西) 也就是说对于视觉识别来说，更有价值的信息，是层次化的，多阶段的特征。

Fukushima's "neocognitron" [17], a biologically-inspired hierarchical and shift-invariant model for pattern recognition, was an early attempt at just such a process. The neocognitron, however, lacked a supervised training algorithm. Building on Rumelhart et al. [30], LeCun et al. [24] showed that stochastic gradient descent via backpropagation was effective for training convolutional neural networks (CNNs), a class of models that extend the neocognitrona.

Fukushima 的"神经认知机"[17]是一种受生物学启发的分层和偏移不变的模式识别模型，是早期的对于这样一个过程的尝试。但是，神经认知机缺乏监督训练算法。卷积神经网络（CNN）是一类扩展神经认知机的模型，建立在 Rumelhart 等[30] 和 LeCun 等[24]提出的通过反向传播进行的随机梯度下降的基础之上。

CNNs saw heavy use in the 1990s (e.g., [25]), but then fell out of fashion with the rise of support vector machines. In 2012, Krizhevsky et al. [23] rekindled interest in CNNs by showing substantially higher image classification accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [9], [10]. Their success resulted from training a large CNN on 1.2 million labeled images, together with a few twists on LeCun's CNN (e.g., max(x, 0) rectifying non-linearities and "dropout" regularization).

CNN 在 20 世纪 90 年代有广泛的使用（例如[25]），但是随着支持向量机的兴起，CNN 已经逐渐淡出了公众视野。2012 年，Krizhevsky 等[23]通过在 ImageNet 大型视觉识别挑战(ILSVRC) [9], [10]上显示出更高的图像分类准确度，重新唤起了人们对 CNN 的兴趣。他们的成功是通过使用大型 CNN 训练 120 万张带标签图像，以及对 LeCun 的 CNN(例如，max(x, 0)非线性整流和"Dropout"正则化)的一些改进。

The significance of the ImageNet result was vigorously debated during the ILSVRC 2012 workshop. The central issue can be distilled to the following: To what extent do the CNN classification results on ImageNet generalize to object detection results on the PASCAL VOC Challenge?

ImageNet 结果的重要性在 ILSVRC 2012 研讨会期间大有争议。中心问题可以归结为：ImageNet 上的 CNN 分类结果在何种程度上能够应用到 PASCAL VOC 挑战的物体检测任务上。

We answer this question by bridging the gap between image classification and object detection. This paper is the first to show that a CNN can lead to dramatically higher object detection performance on PASCAL VOC as compared to systems based on simpler HOG-like features. To achieve this result, we focused on two problems: localizing objects with a deep network and training a high-capacity model with only a small quantity of annotated detection data.

我们通过弥合图像分类和对象检测之间的差距，回答了这个问题。本论文是第一个提出：与基于更简单的 HOG 类特征的系统相比，CNN 可以显著提高 PASCAL VOC 的目标检测性能。为了实现这一结果，我们主要关注两个问题：使用深度网络定位物体和在小规模的标注数据集上进行大型网络模型的训练。

Unlike image classification, detection requires localizing (likely many) objects within an image. One approach frames localization as a regression problem. However, work from Szegedy et al. [33], concurrent with our own, indicates that this strategy may not fare well in practice (they report a mAP of 30.5% on VOC 2007 compared to the 58.5% achieved by our method). An alternative is to build a sliding-window detector. CNNs have been used in this way for at least two decades, typically on constrained object categories, such as faces [29], [35] and pedestrians [31]. In order to maintain high spatial resolution, these CNNs typically only have two convolutional and pooling layers. We also considered adopting a sliding-window approach. However, units high

与图像分类不同的是检测需要定位一个图像内的许多物体。一个方法是将框定位看作是回归问题。但 Szegedy 等人的工作[33]说明这种策略在实践中可能不会很好（在 VOC2007 上他们的 mAP 是 30.5%，而我们的达到了 58.5%）。另一个可替代的方法是使用滑动窗口探测器，CNN 已经以这种方式被使用了至少二十年，通常用于受限物体如人脸[29], [35]和行人[31] 上。为了保持高空间分辨率，这些 CNN 通常只有两个卷积和池化层。我们本来也考虑过使用滑动窗口的方法，但是由于网络层次更深，输入图片有非常大的感受野（195×195）和 步长（32×32），这使得在滑动窗口内的精确定位成为开放的技术挑战。

up in our network, which has five convolutional layers, have very large receptive fields (195×195 pixels) and strides (32×32 pixels), in the input image, which makes precise localization within the sliding-window paradigm an open technical challenge.

Instead, we solve the CNN localization problem by operating within the "recognition using regions" paradigm [19], which has been successful for both object detection [34] and semantic segmentation [5]. At test time, our method generates around 2000 category-independent region proposals for the input image, extracts a fixed-length feature vector from each proposal using a CNN, and then classifies each region with category-specific linear SVMs. We use a simple technique (affine image warping) to compute a fixed-size CNN input from each region proposal, regardless of the region's shape. Figure 1 presents an overview of our method and highlights some of our results. Since our system combines region proposals with CNNs, we dub the method R-CNN: Regions with CNN features.

相反，我们通过在"识别使用区域"范式[19]中操作来解决 CNN 的定位问题，这已经成功实现了目标检测[34]和语义分割[5]。测试时，我们的方法为输入图像生成大约 2000 个类别无关的候选区域，使用 CNN 从每个区域中提取固定长度的特征向量，然后借助专门针对特定类别数据的线性 SVM 对每个区域进行分类。我们使用简单的技术（图像仿射变换）来计算每个候选区域的固定大小的 CNN 输入，而不管区域的形状。图 1 展示了我们方法的全貌并突出展示了一些实验结果。由于我们结合了 Region proposals 和 CNNs，所以起名 R-CNN：Regions with CNN features。

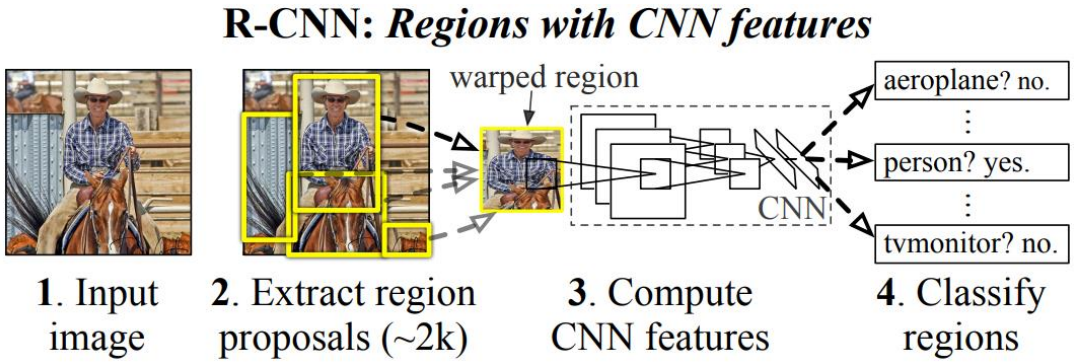

Figure 1. Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean average precision (mAP) of 53.7% on PASCAL VOC 2010. For comparison, [32] reports 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part models perform at 33.4%.

图 1. 物体检测系统概述。我们的系统（1）采用输入图像，（2）提取大约 2000 个自下而上候选区域，（3）使用大型卷积神经网络（CNN）计算每个区域的特征，然后（4）使用特定类的线性 SVM 对每个区域进行分类。R-CNN 在 PASCAL VOC 2010 上实现了 53.7％的平均精确度（mAP）。相比之下，[32] 使用相同的候选区域方法，但采用空间金字塔和视觉词袋方法，达到了 35.1％的 mAP。流行的可变形零件模型的性能为 33.4％。

A second challenge faced in detection is that labeled data is scarce and the amount currently available is insufficient for training a large CNN. The conventional solution to this problem is to use unsupervised pre-training, followed by supervised fine-tuning (e.g., [31]). The second principle contribution of this paper is to show that supervised pre-training on a large auxiliary dataset (ILSVRC), followed by domain-specific fine-tuning on a small dataset (PASCAL), is an effective paradigm for learning high-capacity CNNs

检测面临的第二个挑战是带标签的数据很少，目前可用的数量不足以训练大型 CNN。这个问题的常规解决方案是使用无监督的预训练，然后进行辅助微调（例如[31]）。本文的第二个主要贡献是在大型辅助数据集(ILSVRC)上进行监督预训练，然后对小数据集(PASCAL)进行域特定的微调，这是在数据稀缺时训练高容量 CNN 模型的有效范例。在我们的实验中，微调将检测的 mAP 性能提高了 8 个百分点。微调后，我们的系统在 VOC 2010 上实现了 54％的 mAP，而高度优化的基于 HOG 的可变

when data is scarce. In our experiments, fine-tuning for detection improves mAP performance by 8 percentage points. After fine-tuning, our system achieves a mAP of 54% on VOC 2010 compared to 33% for the highly-tuned, HOG-based deformable part model (DPM) [15], [18]. We also point readers to contemporaneous work by Donahue et al. [11], who show that Krizhevsky's CNN can be used (without fine-tuning) as a blackbox feature extractor, yielding excellent performance on several recognition tasks including scene classification, fine-grained sub-categorization, and domain adaptation.

部件模型(DPM)为 33％[15], [18]。Donahue 等人同时进行的工作 21 表明可以使用 Krizhevsky 的 CNN（无需微调）作为黑盒特征提取器，在多个识别任务（包括场景分类，细粒度子分类和域适配）中表现出色。

Our system is also quite efficient. The only class-specific computations are a reasonably small matrix-vector product and greedy non-maximum suppression. This computational property follows from features that are shared across all categories and that are also two orders of magnitude lower-dimensional than previously used region features (cf [34]).

我们的系统也很有效率。唯一的特定类计算是相当小的矩阵向量乘积和贪心非极大值抑制。这种计算属性来自于所有样本共享的特征，并且比以前使用的区域特征维度还低两个数量级（参见[34]）。

Understanding the failure modes of our approach is also critical for improving it, and so we report results from the detection analysis tool of Hoiem et al. [21]. As an immediate consequence of this analysis, we demonstrate that a simple bounding box regression method significantly reduces mislocalizations, which are the dominant error mode.

了解我们的方法的失败模式对于改进它也是至关重要的，因此我们给出了由 Hoiem 等人[21]提出的检测分析工具的结果。 作为这一分析的中间后果，我们证明了一种简单的边界回归方法显著地减少了定位误差，这是主要的误差模式。

Before developing technical details, we note that because R-CNN operates on regions it is natural to extend it to the task of semantic segmentation. With minor modifications, we also achieve competitive results on the PASCAL VOC segmentation task, with an average segmentation accuracy of 47.9% on the VOC 2011 test set.

在发掘技术细节之前，我们注意到，由于 R-CNN 在区域上运行，将其扩展到语义分割的任务是很自然的。经过少量的修改，我们也在 PASCAL VOC 分割任务中取得了有竞争力的成果，VOC 2011 测试集的平均分割精度为 47.9％。

2. Object Detection with R-CNN

2. 用 RCNN 进行目标检测

Our object detection system consists of three modules. The first generates category-independent region proposals. These proposals define the set of candidate detections available to our detector. The second module is a large convolutional neural network that extracts a fixed-length feature vector from each region. The third module is a set of class-specific linear SVMs. In this section, we present our design decisions for each module, describe their test-time usage, detail how their parameters are learned, and show results on PASCAL VOC 2010–12.

我们的目标检测系统由三个模块组成。第一个生成类别无关区域提案。这些提案定义了可用于我们的检测器的候选检测集。第二个模块是从每个区域提取固定长度特征向量的大型卷积神经网络。第三个模块是一组特定类别的线性 SVM。在本节中，我们介绍每个模块的设计思路，描述其测试时使用情况，详细介绍其参数的学习方式，并给出在 PASCAL VOC 2010-12 和 ILSVRC2013 上的检测结果。

2.1 Module Design

2.1 模型设计

Region Proposals：A variety of recent papers offer methods for generating category-independent region proposals.

候选区域：各种最近的论文提供了生成类别无关区域提案的方法。例子包括：对象性 23，选择性搜索 16，类别

Examples include: objectness [1], selective search [34], category-independent object proposals [12], constrained parametric min-cuts (CPMC) [5], multi-scale combinatorial grouping [3], and Cireşan et al. [6], who detect mitotic cells by applying a CNN to regularly-spaced square crops, which are a special case of region proposals. While R-CNN is agnostic to the particular region proposal method, we use selective search to enable a controlled comparison with prior detection work (e.g., [34], [36]).

无关对象提议 24，约束参数最小化(CPMC)17，多尺度组合分组 25 和 Cireşan 等提出的 26，通过将 CNN 应用于特定间隔的方块来检测有丝分裂细胞，这是区域提案的特殊情况。具体的区域提案方法对于 R-CNN 是透明的，但我们使用**选择性搜索**以便于与先前检测工作的对照比较（例如[34], [36]）。

Feature Extraction： We extract a 4096-dimensional feature vector from each region proposal using the Caffe [22] implementation of the CNN described by Krizhevsky et al. [23]. Features are computed by forward propagating a mean-subtracted 227×227 RGB image through five convolutional layers and two fully connected layers. We refer readers to [22], [23] for more network architecture details.

特征提取：我们使用 Krizhevsky 等人[23]提出的 CNN 的 Caffe[22]实现，从每个区域提案中提取 4096 维特征向量。将减去像素平均值的 227×227 分辨率的 RGB 图像通过五个卷积层和两个全连接层向前传播来计算特征。可以参考[22], [23]以获得更多的网络架构细节。

In order to compute features for a region proposal, we must first convert the image data in that region into a form that is compatible with the CNN (its architecture requires inputs of a fixed 227×227 pixel size). Of the many possible transformations of our arbitrary-shaped regions, we opt for the simplest. Regardless of the size or aspect ratio of the candidate region, we warp all pixels in a tight bounding box around it to the required size. Prior to warping, we dilate the tight bounding box so that at the warped size there are exactl y p pixels of warped image context around the original box (we use p=16). Figure 2 shows a random sampling of warped training regions. The supplementary material discusses alternatives to warping.

为了计算区域提案的特征，我们必须首先将该区域中的图像数据转换为与 CNN 兼容的格式（其架构需要固定 227×227 像素大小的输入）。在许多可能的针对任意形状区域的变换中，我们选择最简单的。不管候选区域的大小或横纵比如何，我们将整个区域不保持横纵比缩放到所需的大小。在缩放之前，我们扩展紧密的边界框，以便在变形的尺寸上，原始框周围有变形的图像上下文的精确像素（我们使用 p = 16）。图 2 显示了变形训练区域的随机抽样。补充材料讨论了形变的替代方案。

Figure 2. Warped training samples from voc 2007 train.

图 2. 变形 voc 2007 训练集中的训练样本

2.2 Test-time detection

At test time, we run selective search on the test image to extract around 2000 region proposals (we use selective search's “fast mode” in all experiments). We warp each proposal and forward propagate it through the CNN in order to read off features from the desired layer. Then, for each

2.2 检测测试

在测试时，我们对测试图像进行选择性搜索，以提取大约 2000 个区域提案（我们在所有实验中使用选择性搜索的“快速模式”）。然后缩放每个区域，并通过 CNN 向前传播，以计算特征。最后，对于每个类，我们使用针对该类训练的 SVM 来对每个提取的特征向量进行评

class, we score each extracted feature vector using the SVM trained for that class. Given all scored regions in an image, we apply a greedy non-maximum suppression (for each class independently) that rejects a region if it has an intersection-over-union (IoU) overlap with a higher scoring selected region larger than a learned threshold.

分。给定图像中的所有区域的得分，我们应用贪婪非极大值抑制（每个类别独立进行），在训练时学习一个阈值，如果其与得分较高的区域的重叠部分(IoU)高于这个阈值，则丢弃这个区域。

Run-Time Analysis: Two properties make detection efficient. First, all CNN parameters are shared across all categories. Second, the feature vectors computed by the CNN are low-dimensional when compared to other common approaches, such as spatial pyramids with bag-of-visual-word encodings. The features used in the UVA detection system [34], for example, are two orders of magnitude larger than ours (360k vs. 4k-dimensional).

性能分析：两种性质使检测效率高。首先，所有 CNN 参数都在所有类别中共享。其次，与其他常见方法比较，由 CNN 计算出的特征向量是低维度的，例如具有空间金字塔和视像单词的方法。UVA 检测系统[34]中使用的特征比我们（维度，360k 对比 4k）大两个数量级。

The result of such sharing is that the time spent computing region proposals and features (13s/image on a GPU or 53s/image on a CPU) is amortized over all classes. The only class-specific computations are dot products between features and SVM weights and non-maximum suppression. In practice, all dot products for an image are batched into a single matrix-matrix product. The feature matrix is typically 2000×4096 and the SVM weight matrix is 4096×N, where N is the number of classes.

这种共享的结果是计算区域提案和特征（GPU 上的 13 秒/图像或 CPU 上的 53 秒/图像）的时间在所有类别上进行摊销。唯一的类特定计算是特征与 SVM 权重和非极大值抑制之间的点积。在实践中，图像的所有点积运算都被整合为单个矩阵与矩阵的相乘。特征矩阵通常为 2000×4096，SVM 权重矩阵为 4096×N，其中 N 为类别数。

This analysis shows that R-CNN can scale to thousands of object classes without resorting to approximate techniques, such as hashing. Even if there were 100k classes, the resulting matrix multiplication takes only 10 seconds on a modern multi-core CPU. This efficiency is not merely the result of using region proposals and shared features. The UVA system, due to its high-dimensional features, would be two orders of magnitude slower while requiring 134GB of memory just to store 100k linear predictors, compared to just 1.5GB for our lower-dimensional features.

分析表明，R-CNN 可以扩展到数千个类，而不需要使用如散列这样的技术。即使有 10 万个类，在现代多核 CPU 上产生的矩阵乘法只需 10 秒。这种效率不仅仅是使用区域提案和共享特征的结果。由于其高维度特征，UVA 系统的速度将会降低两个数量级，并且需要 134GB 的内存来存储 10 万个线性预测器。而对于低维度特征而言，仅需要 1.5GB 内存。

It is also interesting to contrast R-CNN with the recent work from Dean et al. on scalable detection using DPMs and hashing [8]. They report a mAP of around 16% on VOC 2007 at a run-time of 5 minutes per image when introducing 10k distractor classes. With our approach, 10k detectors can run in about a minute on a CPU, and because no approximations are made mAP would remain at 59% (Section 3.2).

将 R-CNN 与 Dean 等人最近的工作对比也是有趣的。使用 DPM 和散列的可扩展检测[8]。在引入 1 万个干扰类的情况下，每个图像的运行时间为 5 分钟，其在 VOC 2007 上的的 mAP 约为 16%。通过我们的方法，1 万个检测器可以在 CPU 上运行大约一分钟，而且由于没有逼近，可以使 mAP 保持在 59%（见消融研究）。

2.3 Training

2.3 训练

Supervised Pre-Training: We discriminatively pre-trained the CNN on a large auxiliary dataset (ILSVRC 2012) with

监督预训练：我们仅通过使用图像级标记来区分性地对大型辅助数据集（ILSVRC2012 分类）进行 CNN 预训练

image-level annotations (i.e., no bounding box labels). Pretraining was performed using the open source Caffe CNN library [22]. In brief, our CNN nearly matches the performance of Krizhevsky et al. [23], obtaining a top-1 error rate 2.2 percentage points higher on the ILSVRC 2012 validation set. This discrepancy is due to simplifications in the training process.

（此数据没有检测框标记）。使用开源的 Caffe CNN 库进行预训练[22]。简而言之，我们的 CNN 几乎符合 Krizhevsky 等人的论文中[23]的表现，ILSVRC2012 分类验证集获得的 top-1 错误率高出 2.2 个百分点。这种差异是由于训练过程中的简化造成的。

Domain-Specific Fine-Tuning：To adapt our CNN to the new task (detection) and the new domain (warped VOC windows), we continue stochastic gradient descent (SGD) training of the CNN parameters using only warped region proposals from VOC. Aside from replacing the CNN's ImageNet-specific 1000-way classification layer with a randomly initialized 21-way classification layer (for the 20 VOC classes plus background), the CNN architecture is unchanged. We treat all region proposals with ≥0.5 IoU overlap with a ground-truth box as positives for that box's class and the rest as negatives. We start SGD at a learning rate of 0.001 (1/10th of the initial pre-training rate), which allows fine-tuning to make progress while not clobbering the initialization. In each SGD iteration, we uniformly sample 32 positive windows (over all classes) and 96 background windows to construct a mini-batch of size 128. We bias the sampling towards positive windows because they are extremely rare compared to background.

特定域的微调：为了使 CNN 适应新任务（检测）和新域（缩放的提案窗口），我们仅使用缩放后的区域提案继续进行 CNN 参数的随机梯度下降(SGD)训练。除了用随机初始化的(N + 1)路分类层（其中 N 是类别数，加 1 为背景）替换 CNN 的 ImageNet 特有的 1000 路分类层，CNN 架构不变。对于 VOC，N = 20，对于 ILSVRC2013，N = 200。我们将所有区域提案与检测框真值 IoU ≥0.5 的区域作为正样本，其余的作为负样本。我们以 0.001（初始学习率的 1/10）的学习率开始 SGD，这样可以在不破坏初始化的情况下进行微调。在每个 SGD 迭代中，我们统一采样 32 个正样本（在所有类别中）和 96 个负样本，以构建大小为 128 的小批量。将采样的正样本较少是因为它们与背景相比非常罕见。

Object Category Classifiers：Consider training a binary classifier to detect cars. It's clear that an image region tightly enclosing a car should be a positive example. Similarly, it's clear that a background region, which has nothing to do with cars, should be a negative example. Less clear is how to label a region that partially overlaps a car. We resolve this issue with an IoU overlap threshold, below which regions are defined as negatives. The overlap threshold, 0.3, was selected by a grid search over {0,0.1,…,0.5} on a validation set. We found that selecting this threshold carefully is important. Setting it to 0.5, as in [34], decreased mAP by 5 points. Similarly, setting it to 0 decreased mAP by 4 points. Positive examples are defined simply to be the ground-truth bounding boxes for each class.

目标类别分类器：考虑训练二分类器来检测汽车。很明显，紧紧围绕汽车的图像区域应该是一个正样本子，一个与汽车无关的背景区域应该是一个负样本。较不清楚的是如何标注部分重叠汽车的区域。我们用 IoU 重叠阈值来解决这个问题，在这个阈值以下的区域被定义为负样本。重叠阈值 0.3 是通过在验证集上尝试了 0,0.1,…,0.5 的不同阈值选择出来的。我们发现选择这个阈值是很重要的。将其设置为 0.5，如[34]，mAP 会降低 5 个点。同样，将其设置为 0 会将 mAP 降低 4 个点。正样本被简单地定义为每个类的检测框真值。

Once features are extracted and training labels are applied, we optimize one linear SVM per class. Since the training data is too large to fit in memory, we adopt the standard hard negative mining method [15], [32]. Hard negative mining converges quickly and in practice mAP stops increasing after only a single pass over all images.

一旦提取了特征并应用了训练标签，我们就可以优化每类线性 SVM。由于训练数据太大内存不够，我们采用标准的难分样本挖掘方法[15], [32]。难分样本挖掘可以快速收敛，实际上所有图像遍历一边，mAP 就停止增长了。

In supplementary material we discuss why the positive and negative examples are defined differently in fine-tuning versus SVM training. We also discuss why it's necessary to train detection classifiers rather than simply use outputs from the final layer (fc8) of the fine-tuned CNN.

在附录材料中，我们将讨论为什么在微调与 SVM 训练中，正样本和负样本的数量不同。我们还讨论了涉及训练检测 SVM 的权衡，而不是简单地使用微调 CNN 的最终 softmax 层的输出。

2.4 Results on Pascal Voc 2010–12

Following the PASCAL VOC best practices [13], we validated all design decisions and hyperparameters on the VOC 2007 dataset (Section 3.2). For final results on the VOC 2010–12 datasets, we fine-tuned the CNN on VOC 2012 train and optimized our detection SVMs on VOC 2012 trainval. We submitted test results to the evaluation server only once for each of the two major algorithm variants (with and without bounding box regression).

2.4 PASCAL VOC 2010-12 上的结果

根据 PASCAL VOC 最佳实践[13]，我们在 VOC 2007 数据集上验证了所有设计和超参数（见消融研究）。对于 VOC 2010-12 数据集的最终结果，我们对 VOC 2012 train 上对 CNN 进行了微调，并在 VOC 2012 trainval 上优化检测 SVM。我们将测试结果提交给评估服务器，对于两种主要算法变体（带有和不带有检测框回归）的每一种，都只提交一次。

Table 1 shows complete results on VOC 2010. We compare our method against four strong baselines, including SegDPM [16], which combines DPM detectors with the output of a semantic segmentation system [4] and uses additional inter-detector context and image-classifier rescoring. The most germane comparison is to the UVA system from Uijlings et al. [34], since our systems use the same region proposal algorithm. To classify regions, their method builds a four-level spatial pyramid and populates it with densely sampled SIFT, Extended OpponentSIFT, and RGB-SIFT descriptors, each vector quantized with 4000-word codebooks. Classification is performed with a histogram intersection kernel SVM. Compared to their multi-feature, non-linear kernel SVM approach, we achieve a large improvement in mAP, from 35.1% to 53.7% mAP, while also being much faster (Section 2.2). Our method achieves similar performance (53.3% mAP) on VOC 2011/12 test.

表 1 展示了在 VOC2010 的结果，我们将自己的方法同四种先进基准方法作对比，其中包括 SegDPM[16]，这种方法将 DPM 检测子与语义分割系统相结合并且使用附加的 inter-detector 的环境和图片检测器。更加恰当的比较是同 Uijling 的 UVA[34]系统比较，因为我们的方法同样基于候选框算法。对于候选区域的分类，他们通过构建一个四层的金字塔，并且将之与 SIFT 模板结合，SIFT 为扩展的 OpponentSIFT 和 RGB-SIFT 描述子，每一个向量被量化为 4000-word 的 codebook。分类任务由一个交叉核的 SVM 承担，对比这种方法的多特征方法，非线性内核的 SVM 方法，我们在 mAP 达到一个更大的提升，从 35.1%提升至 53.7%，而且速度更快。我们的方法在 VOC2011/2012 测试集上达到了相似的检测效果 mAP53.3%。

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 [17]†	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 [32]	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 [35]	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 [15]†	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**

Table 1: Detection average precision (%) on VOC 2010 test. R-CNN is most directly comparable to UVA and Regionlets since all methods use selective search region proposals. Bounding box regression (BB) is described in Section 3.4. At publication time, SegDPM was the top-performer on the PASCAL VOC leaderboard. †DPM and SegDPM use context rescoring not used by the other methods.

3. Visualization, ablation, and modes of error

3.1 Visualizing Learned Features

First-layer filters can be visualized directly and are easy to understand [23]. They capture oriented edges and opponent colors. Understanding the subsequent layers is more challenging. Zeiler and Fergus present a visually attractive deconvolutional approach in [37]. We propose a simple (and complementary) non-parametric method that directly shows what the network learned.

The idea is to single out a particular unit (feature) in the network and use it as if it were an object detector in its own right. That is, we compute the unit's activations on a large set of held-out region proposals (about 10 million), sort the proposals from highest to lowest activation, perform non-maximum suppression, and then display the top-scoring regions. Our method lets the selected unit "speak for itself" by showing exactly which inputs it fires on. We avoid averaging in order to see different visual modes and gain insight into the invariances computed by the unit.

We visualize units from layer pool5, which is the max-pooled output of the network's fifth and final convolutional layer. The pool5 feature map is 6×6×256=9216-dimensional. Ignoring boundary effects, each pool5 unit has a receptive field of 195×195 pixels in the original 227×227 pixel input. A central pool5 unit has a nearly global view, while one near the edge has a smaller, clipped support.

Each row in Figure 3 displays the top 16 activations for a pool5 unit from a CNN that we fine-tuned on VOC 2007 trainval. Six of the 256 functionally unique units are visualized (the supplementary material includes more). These units were selected to show a representative sample of what the network learns. In the second row, we see a unit that fires on dog faces and dot arrays. The unit corresponding to the third row is a red blob detector. There are also detectors for human faces and more abstract patterns such as text and triangular structures with windows. The network appears to learn a representation that combines a small number of class-tuned features together with a distributed representation of shape, texture, color, and material properties. The subsequent fully connected layer fc6 has the ability to model a large set of compositions of these rich features.

3. 可视化、消融和模型错误

3.1 可视化学习的特征

第一层卷积核可以直观可视化，易于理解[23]。它们捕获定向边缘和相对颜色。了解后续层次更具挑战性。Zeiler 和 Fergus 在[37]中提出了一种有视觉吸引力的反卷积方法。我们提出一个简单（和补充）非参数方法，直接显示网络学到的内容。

这个想法是单一输出网络中一个特定单元（特征），然后把它当作一个正确类别的物体检测器来使用。方法是这样的，先计算所有抽取出来的推荐区域（大约 1000 万），计算每个区域所导致的对应单元的激活值，然后按激活值对这些区域进行排序，然后进行最大值抑制，最后展示分值最高的若干个区域。这个方法让被选中的单元在遇到他想激活的输入时"自己说话"。我们避免平均化是为了看到不同的视觉模式和深入观察单元计算出来的不变性。

我们可视化了第五层的池化层 pool5，是卷积网络的最后一层，feature_map(卷积核和特征数的总称)的大小是 6 x 6 x 256 = 9216 维。忽略边界效应，每个 pool5 单元拥有 195×195 的感受野，输入是 227×227。pool5 中间的单元，几乎是一个全局视角，而边缘的单元有较小的带裁切的支持。

图 3 的每一行显示了对于一个 pool5 单元的最高 16 个激活区域情况，这个实例来自于 VOC 2007 上我们调优的 CNN，这里只展示了 256 个单元中的 6 个（附录 D 包含更多）。我们看看这些单元都学到了什么。第二行，有一个单元看到狗和斑点的时候就会激活，第三行对应红斑点，还有人脸，当然还有一些抽象的模式，比如文字和带窗户的三角结构。这个网络似乎学到了一些类别调优相关的特征，这些特征都是形状、纹理、颜色和材质特性的分布式表示。而后续的 fc6 层则对这些丰富的特征建立大量的组合来表达各种不同的事物。

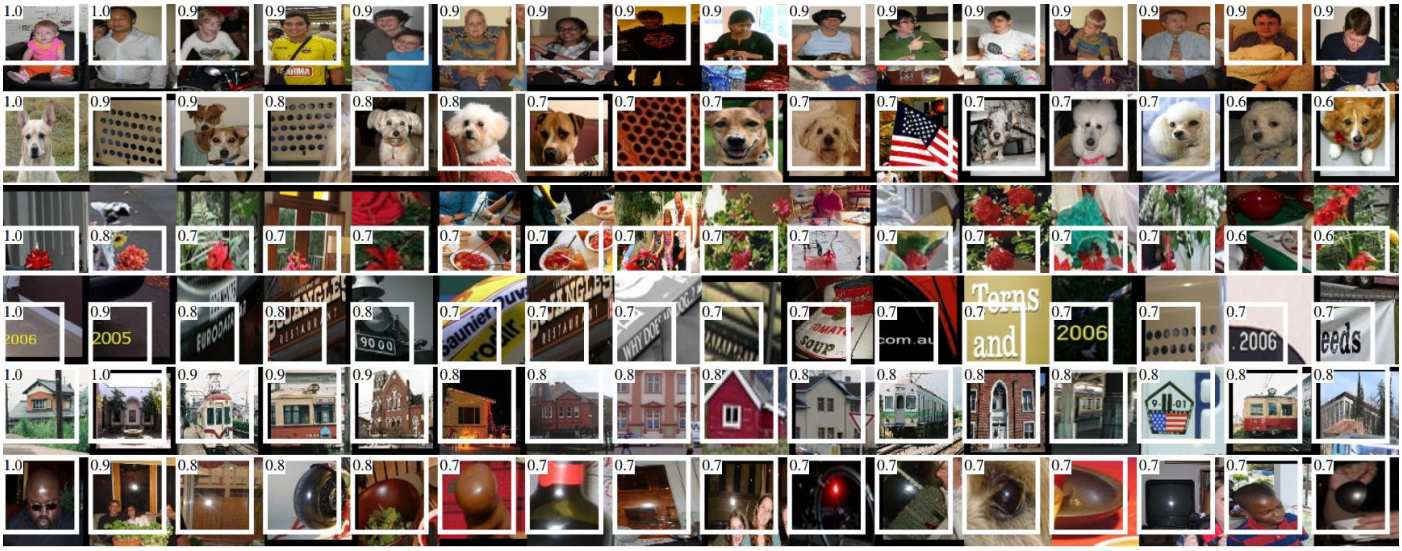

Figure 3. Top regions for six pool5 units. Receptive fields and activation values are drawn in white. Some units are aligned to concepts, such as people (row 1) or text (4). Other units capture texture and material properties, such as dot arrays (2) and specular reflections (6).

图 3. 6 个 pool5 单位的顶级区域。接收字段和激活值以白色绘制。一些单元与概念对齐，例如人（第 1 行）或文本（4）。其他单位捕获纹理和材质属性，例如点阵列（2）和镜面反射（6）。

3.2 Ablation Studies

Performance Layer-by-Layer, without Fine-Tuning. To understand which layers are critical for detection performance, we analyzed results on the VOC 2007 dataset for each of the CNN's last three layers. Layer pool5 was briefly described in Section 3.1. The final two layers are summarized below.

Layer fc6 is fully connected to pool5. To compute features, it multiplies a 4096×9216 weight matrix by the pool-feature map (reshaped as a 9216-dimensional vector) and then adds a vector of biases. This intermediate vector is component-wise half-wave rectified (x←max(0,x)).

Layer fc7 is the final layer of the network. It is implemented by multiplying the features computed by fc6 by a 4096×4096 weight matrix, and similarly adding a vector of biases and applying half-wave rectification.

We start by looking at results from the CNN without fine-tuning on PASCAL, i.e. all CNN parameters were pre-trained on ILSVRC 2012 only. Analyzing performance layer-by-layer (Table 2 rows 1–3) reveals that features from fc7 generalize worse than features from fc6. This means that 29%, or about 16.8 million, of the CNN's parameters can be removed without degrading mAP. More surprising is that removing both fc7 and fc6 produces quite good results even though pool5 features are computed using only 6% of the CNN's parameters. Much of the CNN's representational

3.2 消融研究

逐层分析性能，没有微调：为了理解哪一层对于检测的性能十分重要，我们分析了 CNN 最后三层的每一层在 VOC2007 上面的结果。Pool5 在 3.1 中做过简短的表述。最后两层下面来总结一下。

fc6 是一个与 pool5 连接的全连接层。为了计算特征，它和 pool5 的 feature map（reshape 成一个 9216 维度的向量）做了一个 4096×9216 的矩阵乘法，并添加了一个 bias 向量。中间的向量是逐个组件的半波整流（component-wise half-wave rectified）(x←max(0,x))

fc7 是网络的最后一层。跟 fc6 之间通过一个 4096×4096 的矩阵相乘。也是添加了 bias 向量和应用了 ReLU。

我们先来看看没有调优的 CNN 在 PASCAL 上的表现，没有调优是指所有的 CNN 参数就是在 ILSVRC2012 上训练后的状态。分析每一层的性能显示来自 fc7 的特征泛化能力不如 fc6 的特征。这意味 29%的 CNN 参数，也就是 1680 万的参数可以移除掉，而且不影响 mAP。更多的惊喜是即使同时移除 fc6 和 fc7，仅仅使用 pool5 的特征，只使用 CNN 参数的 6%也能有非常好的结果。可见 CNN 的主要表达力来自于卷积层，而不是全连接层。这一发现表明通过仅使用 CNN 的卷积层来计算任意大小图像的类似 HOG 意义上的密集特征图的潜在实

power comes from its convolutional layers, rather than from the much larger densely connected layers. This finding suggests potential utility in computing a dense feature map, in the sense of HOG, of an arbitrary-sized image by using only the convolutional layers of the CNN. This representation would enable experimentation with sliding-window detectors, including DPM, on top of pool5 features.

用性。这种表示方式可以在 pool5 特征之上实现包括 DPM 在内的滑动窗口检测器。

Performance Layer-by-Layer, with Fine-Tuning. We now look at results from our CNN after having fine-tuned its parameters on VOC 2007 trainval. The improvement is striking (Table 2 rows 4–6): fine-tuning increases mAP by 8.0 percentage points to 54.2%. The boost from fine-tuning is much larger for fc6 and fc7 than for pool5 which suggests that the pools features learned from ImageNet are general and that most of the improvement is gained from learning domain-specific non-linear classifiers on top of them.

逐层分析性能，微调：现在我们来看看在 PASCAL 上进行微调的 CNN 的结果。改善情况引人注目（表 2 第 4-6 行）：微调使 mAP 提高 8.0 个百分点至 54.2%。对于 fc6 和 fc7，微调的提升比对 pool5 大得多，这表明从 ImageNet 中学习的 pool5 特性是一般性的，并且大多数改进是从学习域特定的非线性分类器获得的。

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 [17]	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 [25]	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 [27]	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

Table 2: Detection average precision (%) on VOC 2007 test. Rows 1-3 show R-CNN performance without fine-tuning. Rows 4-6 show results for the CNN pre-trained on ILSVRC 2012 and then fine-tuned (FT) on VOC 2007 trainval. Row 7 includes a simple bounding box regression (BB) stage that reduces localization errors (Section 3.4). Rows 8-10 present DPM methods as a strong baseline. The first uses only HOG, while the next two use different feature learning approaches to augment or replace HOG.

Comparison to Recent Feature Learning Methods. Relatively few feature learning methods have been tried on PASCAL VOC detection. We look at two recent approaches that build on deformable part models. For reference, we also include results for the standard HOG-based DPM [18].

与近期特征学习方法的比较：近期在 PASCAL VOC 检测中已经开始尝试了一些特征学习方法。我们来看两种最新的基于 DPM 模型的方法。作为参考，我们还包括基于标准 HOG 的 DPM 的结果[18]。

The first DPM feature learning method, DPM ST [26], augments HOG features with histograms of "sketch token" probabilities. Intuitively, a sketch token is a tight distribution of contours passing through the center of an image patch. Sketch token probabilities are computed at each pixel by a random forest that was trained to classify 35×35 pixel patches into one of 150 sketch tokens or background.

第一个 DPM 特征学习方法，DPM ST[26]，使用"草图表征"概率直方图增强了 HOG 特征。直观地，草图表征是通过图像片中心的轮廓的紧密分布。草图表征概率在每个像素处被随机森林计算，该森林经过训练，将 35 x 35 像素的图像片分类为 150 个草图表征或背景之一。

The second method, DPM HSC [28], replaces HOG with histograms of sparse codes (HSC). To compute an HSC, sparse code activations are solved for at each pixel using a learned dictionary of 100 7×7 pixel (grayscale) atoms. The

第二种方法，DPM HSC[28]，使用稀疏码直方图(HSC)替代 HOG。为了计算 HSC，使用 100 个 7 x 7 像素（灰度）元素的学习词典，在每个像素处求解稀疏代码激活。所得到的激活以三种方式整流（全部和两个半波），空

resulting activations are rectified in three ways (full and both half-waves), spatially pooled, unit ℓ2 normalized, and then power transformed (x←sign(x)|x|α).

间合并，单位 L2 归一化，和功率变换(x←sign(x)|x|α)。

All R-CNN variants strongly outperform the three DPM baselines (Table 2 rows 8–10), including the two that use feature learning. Compared to the latest version of DPM, which uses only HOG features, our mAP is more than 20 percentage points higher: 54.2% vs. 33.7%-a 61% relative improvement. The combination of HOG and sketch tokens yields 2.5 mAP points over HOG alone, while HSC improves over HOG by 4 mAP points (when compared internally to their private DPM baselines—both use nonpublic implementations of DPM that underperform the open source version [18]). These methods achieve mAPs of 29.1 % and 34.3%, respectively.

所有 R-CNN 变体的都优于三个 DPM 基线（表 2 第 8-10 行），包括使用特征学习的两个。与仅使用 HOG 特征的最新版本的 DPM 相比，我们的 mAP 提高了 20 个百分点以上：54.2%对比 33.7%，相对改进 61%。HOG 和草图表征的组合与单独的 HOG 相比 mAP 提高 2.5 个点，而 HSC 在 HOG 上 mAP 提高了 4 个点（使用内部私有的 DPM 基线进行比较，两者都使用非公开实现的 DPM，低于开源版本 20）。这些方法的 mAP 分别达到 29.1%和 34.3%。

3.3 Detection Error Analysis

We applied the excellent detection analysis tool from Hoiem et al. [21] in order to reveal our method's error modes, understand how fine-tuning changes them, and to see how our error types compare with DPM. A full summary of the analysis tool is beyond the scope of this paper and we encourage readers to consult [21] to understand some finer details (such as "normalized AP"). Since the analysis is best absorbed in the context of the associated plots, we present the discussion within the captions of Figure 4 and Figure 5.

3.3 检测错误分析

为了揭示我们的方法的错误模式，我们应用了 Hoiem 等人的优秀检测分析工具[21]，以了解微调如何改变它们，并将我们的错误类型与 DPM 比较。分析工具的完整介绍超出了本文的范围，可以参考[21]了解更多的细节（如"标准化 AP"）。千言万语不如一张图，我们在下图（图 4 和图 5）中讨论。

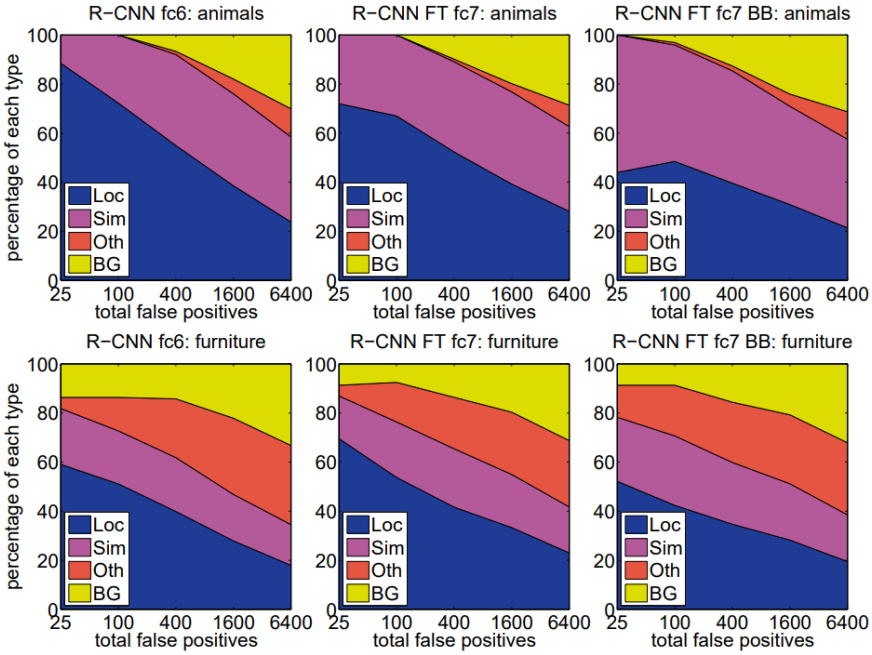

Figure 4. Distribution of top-ranked false positive (fp) types. Each plot shows the evolving distribution of fp types as more fps are considered in order of decreasing score. Each fp is categorized into 1 of 4 types: loc—poor localization (a detection with an iou overlap with the correct class between 0.1 and 0.5, or a duplicate); sim-

图 4. 最多的假阳性（FP）类型分布。每个图表显示 FP 类型的演变分布，按照 FP 数量降序排列。FP 分为 4 种类型：Loc（定位精度差，检测框与真值的 IoU 在 0.1 到 0.5 之间或重复的检测）。Sim（与相似类别混淆）。Oth（与不相似的类别混淆）。BG（检测框标在了背景上）。与 DPM（参见[21]）相比，我们的 Loc

confusion with a similar category; oth-confusion with a dissimilar object category; bg—a fp that fired on background. Compared with dpm (see [21]), 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.

显著增加，而不是 Oth 和 BG，表明 CNN 特征比 HOG 更具区分度。Loc 增加的原因可能是我们使用自下而上的区域提案可能产生松散的定位位置，以及 CNN 进行全图像分类的预训练模型所获得的位置不变性。第三列显示了我们的简单边界回归方法如何修复许多 Loc。

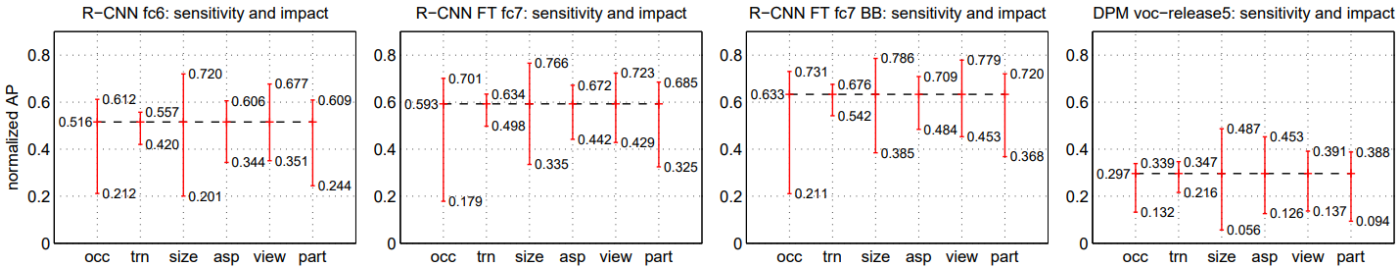

Figure 5. Sensitivity to object characteristics. Each plot shows the mean (over classes) normalized ap (see [21]) for the highest and lowest performing subsets within six different object characteristics (occlusion, truncation, bounding box area, aspect ratio, viewpoint, part visibility). We show plots for our method (r-cnn) with and without fine-tuning (ft) and bounding box regression (bb) as well as for dpm voc-release5. Overall, fine-tuning does not reduce sensitivity (the difference between max and min), but does substantially improve both the highest and lowest performing subsets for nearly all characteristics. This indicates that fine-tuning does more than simply improve the lowest performing subsets for aspect ratio and bounding box area, as one might conjecture based on how we warp network inputs. Instead, fine-tuning improves robustness for all characteristics including occlusion, truncation, viewpoint, and part visibility.

图 5. 对目标特点的敏感度。每个图显示六个不同目标特点（遮挡，截断，边界区域，纵横比，视角，局部可视性）内最高和最低性能的子集的平均值（跨类别）归一化 AP（见 22）。我们展示了我们的方法（R-CNN）有或没有微调（FT）和边界回归（BB）以及 DPM voc-release5 的图。总体而言，微调并不会降低敏感度（最大和最小值之间的差异），而且对于几乎所有的特点，都能极大地提高最高和最低性能的子集的性能。这表明微调不仅仅是简单地提高纵横比和边界区域的最低性能子集的性能（在分析之前，基于我们如何缩放网络输入而推测）。相反，微调可以改善所有特点的鲁棒性，包括遮挡，截断，视角和局部可视性。

3.4 Bounding Box Regression

Based on the error analysis, we implemented a simple method to reduce localization errors. Inspired by the bounding box regression employed in DPM [15], we train a linear regression model to predict a new detection window given the pool5 features for a selective search region proposal. Full details are given in the supplementary material. Results in Table 1, Table 2, and Figure 4 show that this simple approach fixes a large number of mislocalized detections, boosting mAP by 3 to 4 points.

3.4 检测框回归

基于错误分析，我们实现了一种简单的方法来减少定位错误。受 DPM 中使用的检测框回归的启发[15]，我们训练一个线性回归模型使用在区域提案上提取的 pool5 特征来预测一个新的检测框。完整的细节在附录中给出。表 1，表 2 和图 4 中的结果表明，这种简单的方法解决了大量的定位错误，将 mAP 提高了 3 到 4 个点。

4. Semantic Segmentation

Region classification is a standard technique for semantic segmentation, allowing us to easily apply R-CNN to the PASCAL VOC segmentation challenge. To facilitate a direct comparison with the current leading semantic segmentation system (called O2P for "second-order pooling") [4], we work within their open source framework. O2P uses CPMC to generate 150 region proposals per image and then predicts the quality of each region, for each class, using support vector regression (SVR). The high performance of their approach is due to the quality of the CPMC regions and the powerful second-order pooling of multiple feature types (enriched variants of SIFT and LBP). We also note that Farabet et al. [14] recently demonstrated good results on several dense scene labeling datasets (not including PAS-CAL) using a CNN as a multi-scale per-pixel classifier.

We follow [2], [4] and extend the PASCAL segmentation training set to include the extra annotations made available by Hariharan et al. [20]. Design decisions and hyperparameters were cross-validated on the VOC 2011 validation set. Final test results were evaluated only once.

CNN Features for Segmentation: We evaluate three strategies for computing features on CPMC regions, all of which begin by warping the rectangular window around the region to 227×227. The first strategy (full) ignores the region's shape and computes CNN features directly on the warped window, exactly as we did for detection. However, these features ignore the non-rectangular shape of the region. Two regions might have very similar bounding boxes while having very little overlap. Therefore, the second strategy (fg) computes CNN features only on a region's foreground mask. We replace the background with the mean input so that background regions are zero after mean subtraction. The third strategy (full+fg) simply concatenates the full and fg features; our experiments validate their complementarity.

Results on Voc 2011: Table 3 shows a summary of our results on the VOC 2011 validation set compared with O2P. (See supplementary material for complete per-category results.) Within each feature computation strategy, layer fc6 always outperforms fc7 and the following discussion refers to the fc6 features. The fg strategy slightly outperforms full, indicating that the masked region shape provides a stronger signal, matching our intuition. However, full+fg achieves an

4. 语义分割

区域分类是语义分割的基础，这使我们可以轻松地将 R-CNN 应用于 PASCAL VOC 分割挑战。为了便于与当前领先的语义分割系统（称为"二阶池化"的 O2P）[4]的直接比较，我们在其开源框架内修改。O2P 使用 CPMC 为每个图像生成 150 个区域提案，然后使用支持向量回归(SVR)来预测对于每个类别的每个区域的质量。他们的方法的高性能是由于 CPMC 区域的高质量和强大的多种特征类型（SIFT 和 LBP 的丰富变体）的二阶池化。我们还注意到，Farabet 等[14]最近使用 CNN 作为多尺度像素级分类器在几个密集场景标记数据集（不包括 PAS-CAL）上取得了良好的结果。

我们遵循[2], [4]并扩展 PASCAL 分割训练集，以包含 Hariharan 等提供的额外注释[20]。在 VOC 2011 验证集上，交叉验证我们的设计决策和超参数。最终测试结果仅提交一次。

用于分割的 CNN 特征：我们评估了在 CPMC 区域上计算特征的三个策略，所有这些策略都是将区域缩放为 227 x 227。第一个策略（full）忽略了该区域的形状，并直接在缩放后的区域上计算 CNN 特征，就像我们缩放区域提案那样。然而，这些特征忽略了区域的非矩形形状。两个区域可能具有非常相似的边界框，同时具有非常小的重叠。因此，第二个策略（fg）仅在区域的前景掩码上计算 CNN 特征。我们用图像均值替换背景，使得背景区域在减去图像均值后为零。第三个策略（full + fg）简单地连接 full 和 fg 特征。我们的实验验证了它们的互补性。

表 3 显示了与 O2P 相比较的 VOC 2011 验证集的结果（每个类别的计算结果见补充材料）。在每个特征计算策略中，fc6 总是优于 fc7，下面就针对 fc6 进行讨论。fg 策略略优于 full，表明掩蔽区域形状提供了更强的信号，匹配我们的直觉。然而，full+fg 的平均精度为 47.9%，比 fg 优 4.2%（也稍优于 O2P），这表明即使提供了 fg 特征，由 full 特征提供的上下文也是有很多信息。值得注意的是，训练 20 个 SVR，在我们的 full+fg 特征在单

average accuracy of 47.9%, our best result by a margin of 4.2% (also modestly outperforming O2P), indicating that the context provided by the full features is highly informative even given the fg features. Notably, training the 20 SVRs on our full+fg features takes an hour on a single core, compared to 10+ hours for training on O2P features.

核上需要 1 小时，而在 O2P 特征则需要 10 个小时。

In Table 4 we present results on the VOC 2011 test set, comparing our best-performing method, fc6(full+fg), against two strong baselines. Our method achieves the highest segmentation accuracy for 11 out of 21 categories, and the highest overall segmentation accuracy of 47.9%, averaged across categories (but likely ties with the O2P result under any reasonable margin of error). Still better performance could likely be achieved by fine-tuning.

在表 4 中，我们提供了 VOC 2011 测试集的结果，将我们的最佳表现方法 fc6(full + fg)与两个强大的基线进行了比较。我们的方法在 21 个类别中的 11 个中达到了最高的分割精度，最高的分割精度为 47.9%，跨类别平均（但可能与任何合理的误差范围内的 O2P 结果有关）。微调可能会实现更好的性能。

	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

Table 3: Segmentation mean accuracy (%) on VOC 2011 validation. Column 1 presents O_2P; 2-7 use our CNN pre-trained on ILSVRC 2012.

VOC 2011 test	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
R&P [2]	83.4	46.8	18.9	36.6	31.2	42.7	57.3	47.4	44.1	8.1	39.4	**36.1**	36.3	49.5	48.3	50.7	26.3	47.2	22.1	42.0	43.2	40.8
O_2P [4]	**85.4**	**69.7**	22.3	45.2	**44.4**	46.9	66.7	57.8	56.2	**13.5**	**46.1**	32.3	41.2	**59.1**	55.3	51.0	**36.2**	50.4	**27.8**	46.9	**44.6**	47.6
ours (*full+fg* R-CNN fc_6)	84.2	66.9	**23.7**	**58.3**	37.4	**55.4**	**73.3**	**58.7**	**56.5**	9.7	45.5	29.5	**49.3**	40.1	**57.8**	**53.9**	33.8	**60.7**	22.7	**47.1**	41.3	**47.9**

Table 4: Segmentation accuracy (%) on VOC 2011 test. We compare against two strong baselines: the "Regions and Parts" (R&P) method of [2] and the second-order pooling (O_2P) method of [4]. Without any fine-tuning, our CNN achieves top segmentation performance, outperforming R&P and roughly matching O_2P.

5. Conclusion

In recent years, object detection performance had stagnated. The best performing systems were complex ensembles combining multiple low-level image features with high-level context from object detectors and scene classifiers. This paper presents a simple and scalable object detection algorithm that gives a 30% relative improvement over the best previous results on PASCAL VOC 2012.

5. 结论

近年来，物体检测性能停滞不前。性能最好的系统是复杂的组合，将多个低级图像特征与来自物体检测器和场景分类器的高级语境相结合。本文提出了一种简单且可扩展的对象检测算法，相对于 PASCAL VOC 2012 上的前最佳结果，相对改进了 30%。

We achieved this performance through two insights. The first is to apply high-capacity convolutional neural networks to bottom-up region proposals in order to localize and segment objects. The second is a paradigm for training large CNNs when labeled training data is scarce. We show that it is highly effective to pre-train the network-withsupervision—for a auxiliary task with abundant data (image classification) and then to fine-tune the network for the target task where data is scarce (detection). We conjecture that the "supervised pre-training/domain-specific fine-tuning" paradigm will be highly effective for a variety of data-scarce vision problems.

我们通过两个关键的改进实现了这一效果。第一个是将大容量卷积神经网络应用于自下而上的区域提案，以便定位和分割对象。第二个是在有标记的训练数据很少的情况下训练大型 CNN 的方法。我们发现，通过使用大量的图像分类数据对辅助任务进行有监督的预训练，然后对数据稀缺的目标检测任务进行微调，是非常有效的。我们相信，"监督的预训练/领域特定的微调"的方法对于各种数据缺乏的视觉问题都将是非常有效的。

We conclude by noting that it is significant that we achieved these results by using a combination of classical tools from computer vision and deep learning (bottom-up region proposals and convolutional neural networks). Rather than opposing lines of scientific inquiry, the two are natural and inevitable partners.

我们通过使用计算机视觉中的经典工具与深度学习（自下而上的区域提案和卷积神经网络）的组合达到了很好的效果。而不是仅仅依靠纯粹的科学探究。

ACKNOWLEDGMENTS

This research was supported in part by DARPA Mind's Eye and MSEE programs, by NSF awards IIS-0905647, IIS-1134072, and IIS-1212798, MURI N000014-10-1-0933, and by support from Toyota. The GPUs used in this research were generously donated by the NVIDIA Corporation.

致谢

该研究部分由 DARPA Mind 的 Eye 与 MSEE 项目支持，NSF 授予了 IIS-0905647，IIS-1134072 和 IIS-1212798，以及丰田支持的 MURI N000014-10-1-0933。本研究中使用的 GPU 由 NVIDIA 公司慷慨捐赠。

4. References

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

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

[3] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multi-scale 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] I. Endres and D. Hoiem. Category independent object proposals. In ECCV, 2010.

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

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

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

[15] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In CVPR, 2013.

[16] 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.

[17] R. Girshick, P. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://www.cs.berkeley.edu/~rbg/latent-v5/.

[18] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In CVPR, 2009.

[19] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In ICCV, 2011.

[20] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In ECCV. 2012.

[21] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[22] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.
[23] 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.
[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998.
[25] J. J. Lim, C. L. Zitnick, and P. Dollar. Sketch tokens: A learned mid-level representation for contour and object detection. In CVPR,2013.
[26] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.
[27] X. Ren and D. Ramanan. Histograms of sparse codes for object detection. In CVPR, 2013.
[28] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. TPAMI, 1998.
[29] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In CVPR,2013.
[30] 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.
[31] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In NIPS, 2013.
[32] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. IJCV, 2013.
[33] 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.
[34] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. HOGgles: visualizing object detection features. ICCV, 2013.
[35] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. In ICCV, 2013.
[36] M. Zeiler, G. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In CVPR, 2011.