1. 级联R-CNN深入研究高质量物体检测

Cascade R-CNN Delving into High Quality Object Detection

**——Zhaowei Cai，Nuno Vasconcelos（CVPR 2018）**

**一、科学问题**

**1.1 本文所涉及科学问题**

物体检测主要分为两个问题，物体分类和目标定位。本文涉及的问题是如何使用高IOU阈值的同时保证高精度的检测结果。

**1.2 同行专家如何解决**

一般常用阈值u=0.5，然而，常用的阈值u对阳性提出了相当宽松的要求。由此产生的探测器经常产生噪声边界框。

R-FCN提出了有效的区域完全卷积而没有精度损失，以避免Faster-RCNN的重区域CNN计算;而MS-CNN和FPN在多个输出层检测提议，以减轻RPN接收字段与实际对象大小之间的规模不匹配，从而实现高召回提议检测。

YOLO通过转发输入图像一次输出非常稀疏的检测结果。当使用高效的骨干网络实现时，它可以实现具有良好性能的实时对象检测。 SSD以类似于RPN的方式检测对象，但使用不同分辨率的多个特征图来覆盖各种尺度的对象。这些架构的主要限制是它们的精度通常低于两级探测器的精度

**1.3 本文所解决的问题**

问题在于，提议检测器中的假设分布通常严重偏向于低质量。通常，强制较大的IoU阈值会导致正数训练样本数量呈指数级增长。这对于神经网络来说尤其成问题，神经网络已知是非常示例性密集的，并且使得“高u”训练策略非常容易过度配置。另一个困难是探测器的质量与推断的测试假设之间的不匹配。如图1所示，高质量检测器仅对高质量假设最佳。当要求他们处理其他质量水平的假设时，检测可能不是最理想的。

提出了一种多级物体检测结构Cascade R-CNN来解决随着IOU阈值增加而出现的检测性能降低问题。它由一系列以增加的IoU阈值训练的探测器组成，依次更具选择性地抵抗近似误报。探测器逐步进行训练，利用观测器的输出是良好的分布，用于训练下一个更高质量的探测器。逐步改进的假设的重新采样保证了所有检测器都具有相同大小的一组正实例，从而减少了过度设置问题。

**1.4 本文解决方案效果**

定位：在图7（a）中比较级联回归和迭代BBox的定位性能。单个回归量的使用降低了高IoU假设的定位。当迭代地应用回归量时，这种效应累积，如在迭代BBox中，并且性能实际上下降。注意3次迭代后迭代BBox的性能非常差。相反，级联回归器在后期阶段具有更好的性能，在几乎所有IoU级别上都优于迭代BBox。

积分损耗：共用一个回归量的积分损耗检测器中所有分类器的检测性能如图7（b）所示。 u = 0.6的分类器在所有IoU级别中是最好的，而u = 0.7的分类器是最差的。所有分类师的合奏都没有明显的收益。表1显示，迭代BBox和积分损耗检测器均略微改善基线检测器。级联R-CNN对所有评估指标都具有最佳性能。对于较低的IoU阈值，增益是温和的，但对于较高的IoU阈值则是显着的。

检测性能：论文方法的实现优于原始检测器。 尽管如此，Cascade R-CNN仍然在这些基线上持续改进了2到4个点，与其强度无关。 这些增益在val和test-dev上也是一致的。 这些结果表明Cascade R-CNN在检测器架构中广泛适用。

参数和时序：级联RCNN参数的数量随级联级数的增加而增加。 该增加在基线探测器头的参数数量上是线性的。 另外，因为与RPN相比，检测头的计算成本通常较小，所以在训练和测试时，级联R-CNN的计算开销很小。

**二、研究内容**

**2.1 理论与方法介绍**

（论文主要研究内容的提出，主要技术路线、理论与方法介绍）

理论：RPN提案严重倾向于低质量。这不可避免地导致对高质量分类器的无效学习，Cascade R-CNN通过依赖级联回归作为重采样机制来解决该问题。这是由于在图1（c）中所有曲线都在对角灰线上方，即为某一u训练的边界框回归量倾向于产生更高IoU的边界框。

技术路线：级联检测。

方法：从一组示例（xi，bi）开始，级联回归连续地重新采样较高IoU的示例分布（x'i，b'i）。以这种方式，即使当检测器质量（IoU阈值）增加时，也可以将连续级的正例组保持在大致恒定的大小。这在图4中示出，其中在每个重采样步骤之后，分布更倾向于高质量示例。随之而来的是两个后果：首先，没有过度配置，因为各个层面都有丰富的例子。其次，深层阶段的探测器针对更高的IoU阈值进行了优化。

**2.2 验证分析与实验效果**

（论文中的实验分析和实验效果）

实验分析：图5（b）显示了将基础真值边界框添加到提议集中时获得的结果。虽然所有探测器都有所改进，但u = 0.7的探测器具有最大增益，在几乎所有IoU级别都能实现最佳性能。这些结果表明了两个结论。首先，对于精确检测，u = 0.5不是一个好的选择，对于低质量的提议更为稳健。其次，高精度检测需要与检测器质量相匹配的假设。接下来，原始探测器提议被更高质量的级联R-CNN提议所取代（u = 0.6且u = 0.7分别使用了第二和第三阶段提案）。图5（a）还表明，当测试建议与探测器质量更接近时，两个探测器的性能得到显着改善。

图6显示，当使用更精确的假设时，每个探测器都得到了改进，而更高质量的探测器具有更大的增益。例如，u = 0.7的检测器对于第一阶段的低质量提议表现不佳，但对于更深级联阶段可用的更精确假设更好。另外，即使在使用相同的提议时，图6的联合训练的检测器也优于图5（a）的经过单独训练的检测器。这表明探测器在Cascade R-CNN框架内得到了更好的训练。

5.3。与迭代BBox和积分损失的比较

在本节中，我们将Cascade R-CNN与迭代BBox和积分损耗检测器进行比较。通过迭代地应用FPN +基线三次实现迭代BBox。积分损耗检测器具有与Cascade R-CNN相同数量的分类磁头，U = {0.5,0.6,0.7}。

定位：在图7（a）中比较级联回归和迭代BBox的定位性能。单个回归量的使用降低了高IoU假设的定位。当迭代地应用回归量时，这种效应累积，如在迭代BBox中，并且性能实际上下降。注意3次迭代后迭代BBox的性能非常差。相反，级联回归器在后期阶段具有更好的性能，在几乎所有IoU级别上都优于迭代BBox。

积分损耗：共用一个回归量的积分损耗检测器中所有分类器的检测性能如图7（b）所示。 u = 0.6的分类器在所有IoU级别中是最好的，而u = 0.7的分类器是最差的。所有分类师的合奏都没有明显的收益。表1显示，迭代BBox和积分损耗检测器均略微改善基线检测器。级联R-CNN对所有评估指标都具有最佳性能。对于较低的IoU阈值，增益是温和的，但对于较高的IoU阈值则是显着的。

还进行了消融实验。

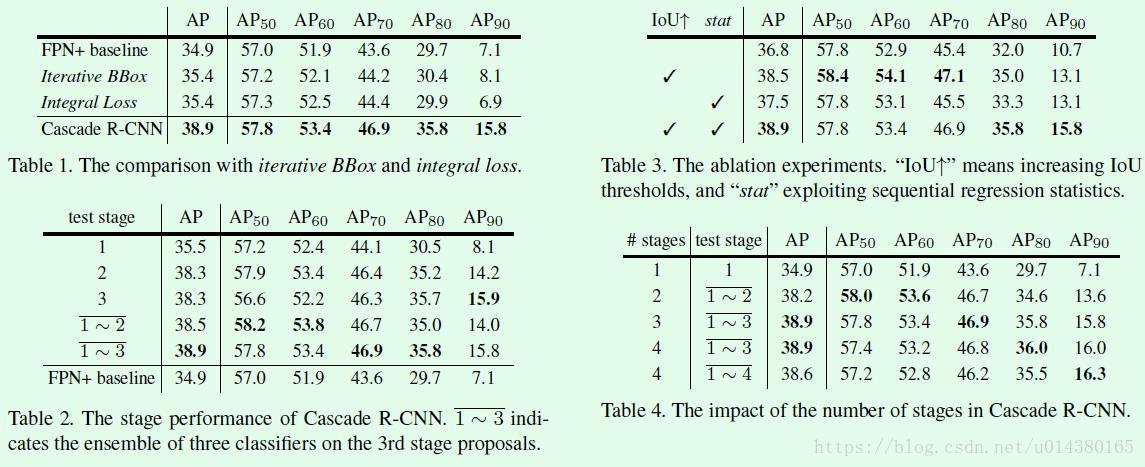
阶段性比较：表2总结了阶段性能。由于多阶段多任务学习的好处，第一阶段已经优于基线探测器。第2阶段大幅提高了性能，第3阶段相当于第2阶段。这与积分损耗检测器不同，其中较高的IOU分类器相对较弱。虽然前（后）阶段在低（高）IoU指标上更好，但所有分类的整体是最好的整体。

IoU阈值：对于所有磁头，使用相同的IoU阈值u = 0.5训练初级级联R-CNN。在这种情况下，阶段仅在他们收到的假设上有所不同。每个阶段都用相应的假设进行训练，即考虑图2的分布。表3的第一行表明级联在基线检测器上有所改进。这表明优化相应样本分布的阶段的重要性。第二行显示，通过增加阶段阈值u，可以使检测器对于紧密误报更具选择性，并专门用于更精确的假设，从而产生额外的增益。这支持了4.2节的结论。

回归统计：利用图2中逐步更新的回归统计数据，有助于有效的多任务学习分类和回归。通过比较表3中的有/无模型，可以看出它的好处。学习对这些统计不敏感。

阶段数：阶段数的影响总结在表4中。添加第二个检测阶段显着改善了基线检测器。三个检测阶段仍然会产生非平凡的改进，但增加第四阶段（u = 0.75）会导致轻微的性能下降。但请注意，虽然整体AP性能下降，但四级级联对于高IoU级别具有最佳性能。三级级联实现了最佳权衡。

实验效果：



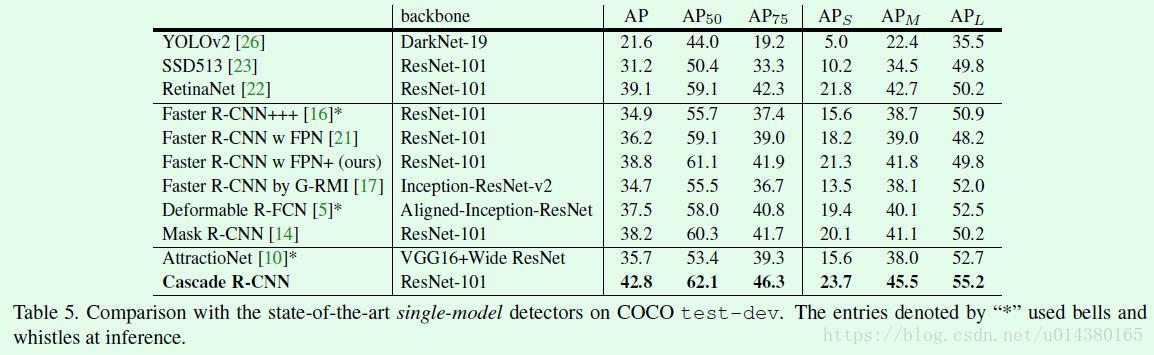
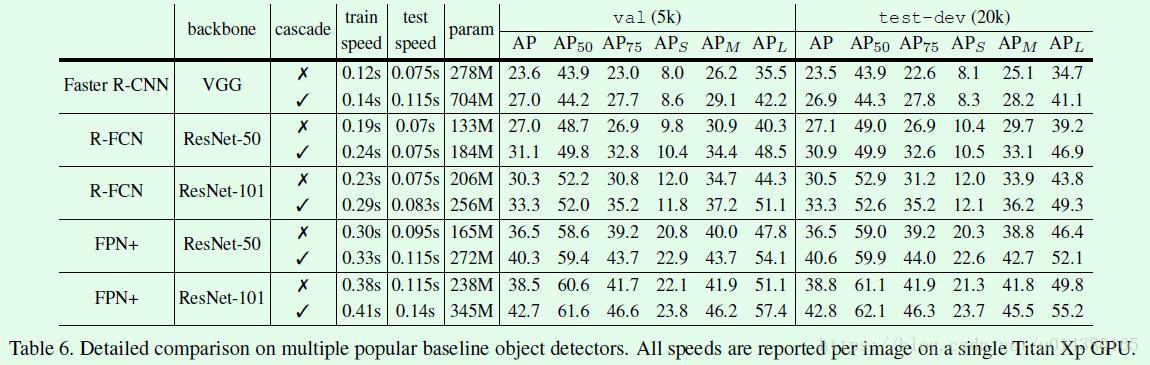


Table6主要通过在现有的two stage算法上添加cascade思想后的对比结果，另外还对比了训练、测试时间、参数量等信息。



**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

提出了新型的结构，但在算法方面没有大的改变。

**3.2 后续研究重点**

探索这种新的结构在各个方面的实际效果，针对这种结构探索适合的新的更优化的算法。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：基于级联式Faster RCNN的三维目标最优抓取方法研究

**（2）作者介绍**：陈丹 林清泉

**（3）摘要**: 机器人在三维目标识别和最优抓取方面的难点在于复杂的背景环境以及目标物体形状不规则,且要求机器人像人一样在识别不同三维目标的同时要确定该目标的最佳抓取部位的位姿。提出一种基于级联式模型的深度学习方法来识别目标物体及其最优抓取位姿。第1级提出了改进的Faster RCNN模型,该模型能识别成像小的目标物体,并能准确对其进行定位;第2级的Faster RCNN模型在前一级确定的目标物体上寻找该目标物体的最优抓取位姿,实现机器人的最优抓取。实验表明该方法能快速且准确地找到目标物体并确定其最优抓取位姿。

**4.2 相关论文二**

**（1）题目**：**CRRCNN: cascade rotational RCNN for dense arbitrary-oriented object detection**

**（2）作者介绍**：Jinduo Lei , Yali Li , Shengjin Wang

**（3）摘要**：In this work, we present a novel network named CRRCNN (Cascade Rotational Region-based CNN) to detect dense objects with oriented bounding boxes. The CRRCNN consists of a Faster RCNN and a Cascade RCNN with Rotational RoIAlign. The Faster RCNN consists of RPN (Region Proposal Network) and RCNN (Region-based CNN). RPN generates horizontal bounding boxes. Rotational region proposals are generated through quadrilateral vertices regression of RCNN, and therefore Faster RCNN is regarded as a Rotational Region Proposal Network (RRPN). To generate accurate rotational bounding boxes, a Cascade RCNN with Rotational RoIAlign is proposed following the Faster RCNN, which will be demonstrated to be crucial for accurate arbitrary-oriented object detection, especially for dense objects. Feature Pyramid Network is also employed to obtain rich context information. The two networks mentioned above are unified and learned end-to-end by jointly optimizing. Experiments on the challenging DOTA dataset demonstrate the effectiveness of our approach.

**4.3 相关论文三**

**（1）题目**：Hot Anchors: A Heuristic Anchors Sampling Method in RCNN-Based Object Detection.

**（2）作者介绍**：Zhang Jinpeng; Zhang Jinming; Yu Shan;

**（3）摘要**：In the image object detection task, a huge number of candidate boxes are generated to match with a relatively very small amount of ground-truth boxes, and through this method the learning samples can be created. But in fact the vast majority of the candidate boxes do not contain valid object instances and should be recognized and rejected during the training and evaluation of the network. This leads to extra high computation burden and a serious imbalance problem between object and none-object samples, thereby impeding the algorithm’s performance. Here we propose a new heuristic sampling method to generate candidate boxes for two-stage detection algorithms. It is generally applicable to the current two-stage detection algorithms to improve their detection performance. Experiments on COCO dataset showed that, relative to the baseline model, this new method could significantly increase the detection accuracy and efficiency.

1. 用于对象检测的关系网络

Relation Networks for Object Detection

**——Han Hu Jiayuan Gu Zheng Zhang Jifeng Dai Yichen Wei（CVPR 2018）**

**一、科学问题**

**1.1 本文所涉及科学问题**

目标检测中对象物体之间的关系

**1.2 同行专家如何解决**

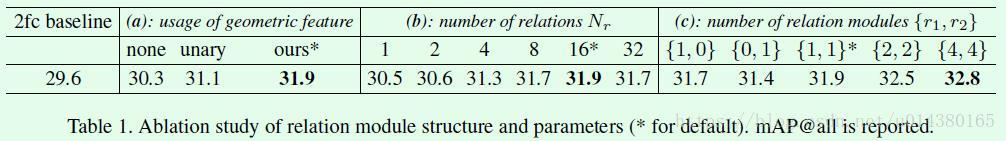
所有最先进的物体检测系统仍然依赖于单独识别物体中的物体，而不会在学习期间利用它们的关系。

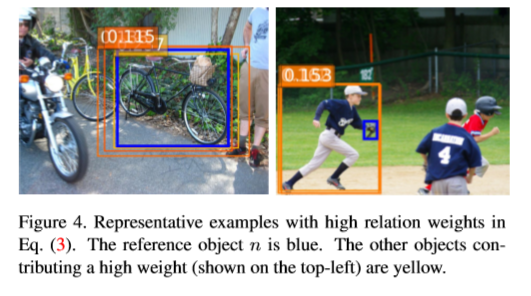
**1.3 本文所解决的问题**

通过引入object relation module来刻画object之间的关系，借助这样的attention来提升object detection的效果。

**1.4 本文解决方案效果**

全面的消融实验表明，关系模块已经在对象之间学习了信息，这些信息在对个体对象进行学习时缺失。 然而，它并不清楚关系模块的重要性，尤其是当被关注的多个人时。 为了理解，我们研究了表1（c）中{r1，r2} = {1,0}头部的（唯一）关系模块。 图4显示了具有高关系权重的示例性示例。 左边的例子表明，在同一个地面实况（自行车）上重叠的几个物体有助于定心物体。 正确的例子表明这个人为手套做出了贡献。 虽然这些示例很直观，但我们对关系模块如何工作的理解是初步的，并留作未来的工作。





**二、研究内容**

**2.1 理论与方法介绍**

这篇文章的出发点在于目前大部分的目标检测（object detection）算法都是独立地检测图像中的object，但显然如果模型能学到object之间的关系显然对于检测效果提升会有帮助，因此希望在检测过程中可以通过利用图像中object之间的相互关系或者叫图像内容（context）来优化检测效果，这种关系既包括相对位置关系也包括图像特征关系。

借鉴了最近几年火热的attention机制的启发，提出一个模块（module）：object relation module来描述object之间的关系，从而以attention的形式附加到原来的特征上最后进行回归和分类，同时将这种attention机制引入nms操作中，不仅实现了真正意义上的end-to-end训练，而且对于原本的检测网络也有提升。

在这篇文章中，object relation module主要放在两个全连接层后面和NMS模块，作者将目前目标检测算法分为4步：1、特征提取主网络；2、得到ROI及特征（RPN网络就包含在其中）；3、基于ROI做边框回归和object分类；4、NMS处理，去除重复框。从作者的分步情况和源码可以清晰地看出，这篇文章主要是基于Faster RCNN系列算法（Faster RCNN，FPN等）引入object relation module。

**2.2 验证分析与实验效果**

Table1主要做了3个验证：1、验证引入位置特征（geometric feature）的有效性（a）；2、验证关系特征数量的影响（b）；3、验证relation module在两个全连接层中的数量的影响（c）。

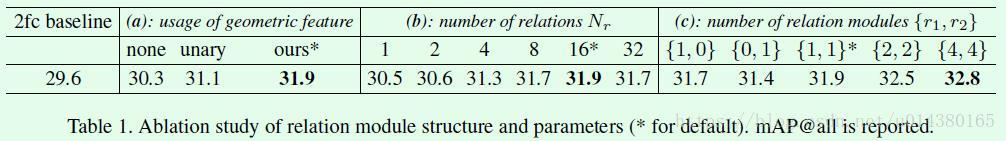


Table2主要是验证RM效果提升是否是因为参数量增加带来的。因此这里做了一些增加fc层、加宽fc层和增加residual结构的操作，最后表明效果的提升并非由于参数量提升带来。

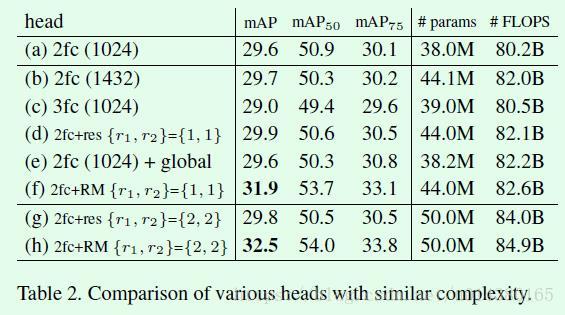
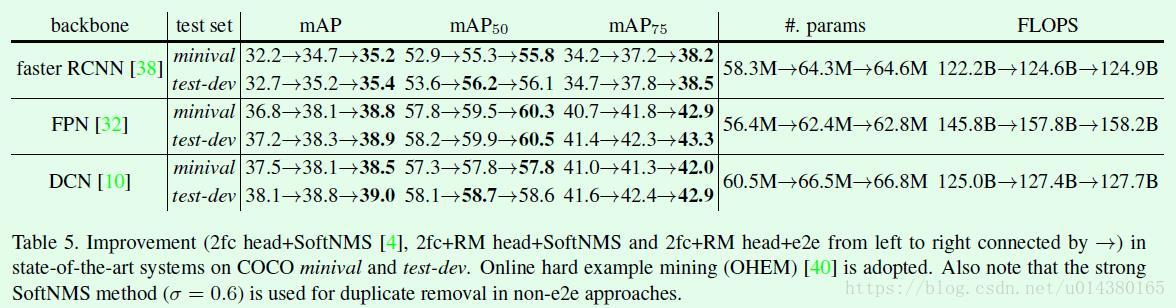


Table5是关于在不同算法上引入RM的效果



**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

论文方法的实现对最终检测效果的提升仍有很大的发展空间。

**3.2 后续研究重点**

改进算法和结构使利用图像中object之间的相互关系来优化检测的效果更出色。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：Context based object categorization: A critical survey

**（2）作者介绍**：Carolina Galleguillos; Serge Belongie;

**（3）摘要**: The goal of object categorization is to locate and identify instances of an object category within an image. Recognizing an object in an image is difficult when images include occlusion, poor quality, noise or background clutter, and this task becomes even more challenging when many objects are present in the same scene. Several models for object categorization use appearance and context information from objects to improve recognition accuracy. Appearance information, based on visual cues, can successfully identify object classes up to a certain extent. Context information, based on the interaction among objects in the scene or global scene statistics, can help successfully disambiguate appearance inputs in recognition tasks. In this work we address the problem of incorporating different types of contextual information for robust object categorization in computer vision. We review different ways of using contextual information in the field of object categorization, considering the most common levels of extraction of context and the different levels of contextual interactions. We also examine common machine learning models that integrate context information into object recognition frameworks and discuss scalability, optimizations and possible future approaches.

**4.2 相关论文二**

**（1）题目**：Rich feature hierarchies for accurate object detection and semantic segmentation

**（2）作者介绍**：Girshick, R. , Donahue, J. , Darrell, T. , Malik, J.

**（3）摘要**：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 http://www.cs.berkeley.edu/~rbg/rcnn.

**4.3 相关论文三**

**（1）题目**：Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

**（2）作者介绍**：He Kaiming; Zhang Xiangyu; Ren Shaoqing; Sun Jian

**（3）摘要**：Existing deep convolutional neural networks (CNNs) require a fixed-size (e.g., 224 × 224) input image. This requirement is "artificial" and may reduce the recognition accuracy for the images or sub-images of an arbitrary size/scale. In this work, we equip the networks with another pooling strategy, "spatial pyramid pooling", to eliminate the above requirement. The new network structure, called SPP-net, can generate a fixed-length representation regardless of image size/scale. Pyramid pooling is also robust to object deformations. With these advantages, SPP-net should in general improve all CNN-based image classification methods. On the ImageNet 2012 dataset, we demonstrate that SPP-net boosts the accuracy of a variety of CNN architectures despite their different designs. On the Pascal VOC 2007 and Caltech101 datasets, SPP-net achieves state-of-the-art classification results using a single full-image representation and no fine-tuning. The power of SPP-net is also significant in object detection. Using SPP-net, we compute the feature maps from the entire image only once, and then pool features in arbitrary regions (sub-images) to generate fixed-length representations for training the detectors. This method avoids repeatedly computing the convolutional features. In processing test images, our method is 24-102 × faster than the R-CNN method, while achieving better or comparable accuracy on Pascal VOC 2007. In ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, our methods rank #2 in object detection and #3 in image classification among all 38 teams. This manuscript also introduces the improvement made for this competition.