人工神经网络原理文献报告作业

陈政培 17363011 智能科学与技术

综合横向对比分析

从传统方法到R-CNN	从R-CNN到SPP	从R-CNN&SPP到Fast R-CNN	从Fast R-CNN到Faster R-CNN	从Faster R-CNN到YOLO	从YOLO到SSD
R-CNN的三大步骤：得到候选区域，用cnn提取特征，训练分类器（后两步放在一个网络中，用softmax做分类器也可以）	R-CNN必须限制输入图像大小（全连接层要求）	Fast R-CNN通过ROI pooling（一层的SPP）,multi-task等改进大大提高速度	Faster R-CNN对于Fast R-CNN的改进在于把region proposal的步骤换成一个CNN网络，也就是RPN	YOLO的pipeline	SSD的pipeline和关键技术
深度模型应用于目标检测的开创性工作之一，改变了物体检测的总思路	SPP的两大优势：1. 可变输入大小 2. 各patch块之间卷积计算是共享的	比R-CNN，SPP-net更高的检测精度，单步训练	Faster R-CNN的base model: VGG16（base model的中间卷积层输出即为要输入到RPN的那个feature map）	YOLO的网络结构、模型、损失函数	SSD的网络结构、多尺度特征图
	SPP的缺陷：multi-stage，训练和测试都比较慢		Faster R-CNN的锚点anchor box、损失函数、四步训练	模型构造简单，可以直接在全图像上训练，推动实时目标检测技术	SSD的default box、SSD的default box与faster r-cnn的anchor box的对比、SSD的default box和尺度选择、SSD的default box与faster r-cnn的anchor box的对比

two-stage方法，如R-CNN系列算法
 -- 第一步选取候选框
 -- 第二步对这些候选框分类或者回归

one-stage方法，如Yolo和SSD
 -- 其主要思路是均匀地在图片的不同位置进行密集抽样
 -- 抽样时可以采用不同尺度和长宽比，然后利用CNN提取特征后直接进行分类与回归
 -- 整个过程只需要一步，所以其优势是速度快

综合来看，从R-CNN → SPP → Fast R-CNN → Faster R-CNN → YOLO → SSD整体在准确率和速度上都在提高。但对小目标的识别YOLO和SSD仍比较差，还达不到faster rcnn的水准，但实时处理优势明显。

区域卷积神经网络 R-CNN系列

R-CNN

背景

长达十几年，目标检测算法以SVM等分类器为主，且mAP出现瓶颈很难产生新的技术突破。

文献

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In **Proceedings of the IEEE conference on computer vision and pattern recognition** (pp. 580-587)

简述

文中提供了一种简单且可扩展的目标检测算法，其mAP相较于以前的方法53.3%提升了超过30%。其核心在于应用大容量卷积神经网络进行自下而上的区域训练，当训练数据稀少时对大型cnn的监督训练。

其中涉及到非极大值抑制（NMS）的概念。NMS用于在目标检测中用于提取分数最高的窗口的。滑动窗口经提取特征，经分类器分类识别后，每个窗口都会得到一个分数。但是滑动窗口会导致很多窗口与其他窗口存在包含或者大部分交叉的情况。这时就需要用到NMS来选取那些邻域里分数最高，并且抑制那些分数低的窗口。非极大值抑制就是在解决当前的梯度值在梯度方向上是一个局部最大值的问题。

先用了ILSVRC2012图片分类训练数据库，先进行网络的图片分类训练。这个数据库有大量的标注数据，共包含了1000种类别物体，因此预训练阶段cnn模型的输出是1000个神经元。

测试阶段，这个方法对每一个输入的图片产生近2000个不分种类的“region proposals，使用CNN+SVM的方法，提取特征向量并对特定种类的线性SVM进行分类。图片分类与物体检测不同，物体检测需要定位出物体的位置，这种就相当于回归问题，求解一个包含物体的方框。而图片分类其实是逻辑回归。

候选框搜索阶段，文中考虑过使用滑动窗口，但由于滑动窗口会带来更深的网络，更大的输入图片和滑动步长，使得使用滑动窗口来定位的方法计算量异常恐怖。最终采用selective search搜索出2000个搜索框。但CNN对输入图片的大小是有限制的，所以每个搜索框还需要采用各向异性缩放，padding=16的精度缩放到固定大小。

CNN特征提取阶段，由于标签训练数据少，文中直接采用了Alexnet并直接采用了其参数作为初始值，引入了迁移学习的概念。然后进行fine-tuning训练。文中还证明了一个理论，如果不进行fine-tuning，而直接使用Alexnet模型，不针对特定的任务。然后把提取的特征用于分类，精度不会明显变化，如果进行fine-tuning了，那么提取到的特征最会训练的svm分类器的精度就会显著提升。

SVM训练、测试阶段，文中通过IOU阈值来定义正负样本，搜索框是否包含整个物体，而文中得到的IOU为0.3，效果最好。最后将特征放入SVM分类器就可以完成整个任务。

主要贡献

1. 将常年来目标检测老方法的平均精度从53.3%提高了30%。
2. 深度模型应用于目标检测的开创性工作之一，改变了物体检测的总思路，后续许多文献关于深度学习的慕白检测算法都有继承其中思想。
3. 物体检测和图片分类的区别：图片分类不需要定位，而物体检测需要定位出物体的位置，也就是相当于把物体的框出来，物体检测是要把所有图片中的物体都识别定位出来。
4. 通过实验数据证明了fine-tuning过程对于训练效果的显著提升。
5. 得到了一个对R-CNN合适的IOU阈值，以区分正负样本。
6. 为少量训练数据得到高质量模型提供了方法，合理的运用了迁移学习的思想。

个人理解与体会

R-CNN的三大步骤：得到候选区域，用cnn提取特征，训练分类器（cnn和分类器在一个网络中，分类器可以为SVM、softmax等）

SPPnet

背景

由于之前的大部分CNN模型的输入图像都是固定大小的（大小，长宽比），而不同大小的输入图像需要通过crop或者warp来生成一个固定大小的图像输入到网络中。这样子就存在问题，1.尺度的选择具有主观性，对于不同的目标，其最适合的尺寸大小可能不一样，2.对于不同的尺寸大小的图像和长宽比的图像，强制变换到固定的大小会损失信息；3.crop的图像可能不包含完整的图像，warp的图像可能导致几何形变。所以说固定输入到网络的图像的大小可能会影响到他们的识别特别是检测的准确率。

文献

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. **IEEE transactions on pattern analysis and machine intelligence**, **37**(9), 1904-1916

简述

文中，提出了利用空间金字塔池化（spatial pyramid pooling，SPP）来实现对图像大小和不同长宽比的处理，产生新的网络SPP-Net，可以不论图像的大小产生相同大小长度的表示特征；这样的网络用在分类和检测上面都刷新的记录；并且速度比较快，快30-170倍，因为之前的检测方法都是采用：1.滑动窗口（慢） 2.对可能的几个目标（显著性目标窗口，可能有几千个）的每一个都进行识别然后再选出最大值作为检测到的目标；利用这种网络，我们只需要计算完整图像的特征图（feature maps）一次，然后池化子窗口的特征，这样就产生了固定长度的表示，它可以用来训练检测器；

将SPP层接到最后一个卷积层后面，SPP层池化特征并且产生固定大小的输出，它的输出然后再送到第一个全连接层。在卷积层和全连接层之前，导入了一个新的层，它可以接受不同大小的输入但是产生相同大小的输出；这样就可以避免在网络的输入口处就要求它们大小相同，也就实现了可以接受任意输入尺度。

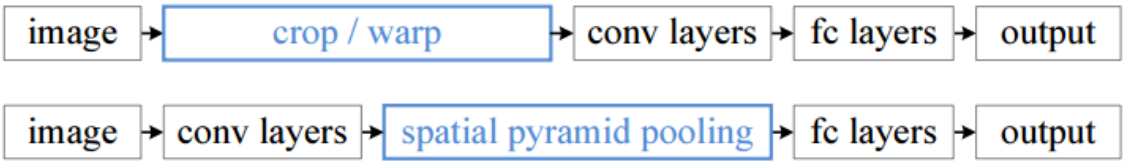


Figure 1. Top: cropping or warping to fit a fixed size. Middle: a conventional deep convolutional network structure. Bottom: our spatial pyramid pooling network structure.

空间金字塔匹配（spatial pyramid matching or SPM）是BoW的一个扩展，它把一张图片划分为从不同的分辨率级别然后聚合这些不同分辨率的图像，SSP有一些很好的特征：1.它可以不论输入数据的大小而产生相同大小的输出，而卷积就不行 2.SPP使用多级别的空间块，也就是说它可以保留了很大一部分的分辨率无关性；3.SPP可以池化从不同尺度图像提取的特征。

在利用SPP层替换最后一个卷积层后面的池化层中，池化层（Poolinglayer）在滑动窗口的角度下，也可以看作为卷积层，卷积层的输出称之为featuremap，它表示了响应的强度和位置信息。在每一个空间块中，池化每一个滤波器的响应，所以SPP层的输出为256M维度，其中256是滤波器的个数，这样不同输入图像大小的输出就可以相同了。

主要贡献

1. 他可以解决输入图片大小不一造成的缺陷。
2. 由于把一个feature map从不同的角度进行特征提取，再聚合的特点，显示了算法的robust的特性。
3. 同时也在object recongtion增加了精度。

个人理解与体会

对比于R-CNN，R-CNN更耗时，因为它是通过对图像的几千个区域，通过显著性提取特征表示，而SPP只需要运行卷积层一次，对整幅图像无论大小进行卷积，然后利用SPP层来提取特征，它提取的特征长度是相同的，所以说它减少了卷积的次数，所以比R-CNN快了几十倍到一百多倍的速度。

Fast R-CNN

背景

R-CNN虽然通过预训练的卷积神经网络有效抽取了图像特征，但它的主要缺点是速度慢。想象一下，我们可能从一张图像中选出上千个提议区域，对该图像做目标检测将导致上千次的卷积神经网络的前向计算。这个巨大的计算量令R-CNN难以在实际应用中被广泛采用。

文献

Girshick, R. (2015). Fast r-cnn. In **Proceedings of the IEEE international conference on computer vision** (pp. 1440-1448)

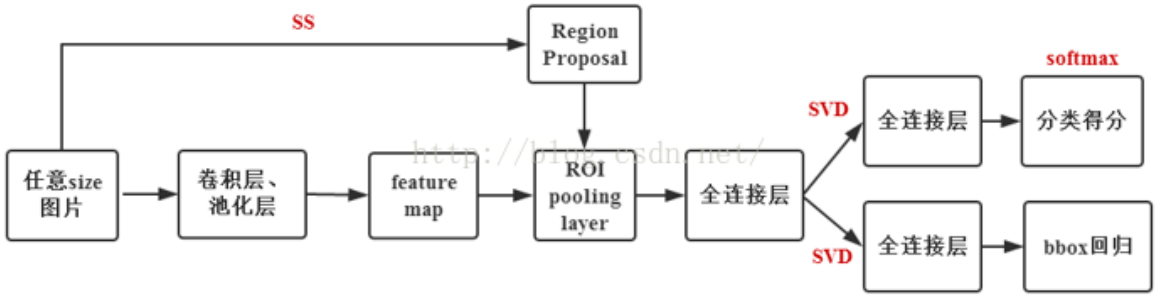
简述

提出了一种基于区域的快速卷积网络方法用于目标检测。Fast R-CNN在前人工作的基础上，利用深卷积网络对目标方案进行有效分类。与以往的工作相比，Fast R-CNN在提高训练和测试速度的同时，也提高了检测精度。快速R-CNN训练通过VGG16网络比R-CNN快，并在PASCAL VOC 2012上实现更高的mAP。与SPPnet相比，通过VGG16训练的Fast R-CNN更快，更准确。

总结了R-CNN的缺陷：1.训练是多阶段多步骤的。2.由于要训练SVM和回归器，所以空间和时间上开销都非常的大。3.目标特征提取速度慢。而SPPnet的缺陷继承了大多R-CNN的缺陷，但由于SPP的微调算法，RCNN提出的fine-tuning方法无法直接用在SPP-net，限制了非常深网络的精度。

Fast R-CNN主要计算步骤：

1. 与R-CNN相比，Fast R-CNN用来提取特征的卷积神经网络的输入是整个图像，而不是各个提议区域。而且，这个网络通常会参与训练，即更新模型参数。
2. 选择性搜索生成n个搜索框。这些搜索框在卷积神经网络的输出上分别标出兴趣区域。这些兴趣区域需要抽取出形状相同的特征以便于连结全连接层后输出。Fast R-CNN引入兴趣区域池化层（region of interest pooling，RoI池化），将卷积神经网络的输出和提议区域作为输入，输出连结后的各个提议区域抽取的特征。
3. 通过全连接层将输出形状变换
4. 预测类别时，将全连接层的输出的形状再变换并使用softmax回归。预测边界框时，将全连接层的输出的形状变换。也就是为每个提议区域预测类别和边界框。



在RoI池化层中，我们通过设置池化窗口、填充和步幅来控制输出形状。而RoI层对每个区域的输出形状是可以直接指定的，RoI层可从形状各异的兴趣区域中均抽取出形状相同的特征。

实验对比了三个预训练网络：CaffeNet，VGG_CNN_M_1024，VGG-16，每个网络有五个最大池层和五到十三个转换层。Fast R-CNN用反向传播训练所有网络权重。说明了SPPnet无法更新空间金字塔池化层之前的层的权重。提出了一种更有效的训练方法，利用训练期间的特征共享。在Fast RCNN训练中，随机梯度下降（SGD）小批量计算被分层采样，并使用一个精简的训练过程，一次微调中联合优化softmax分类器和bbox回归。经历多任务损失、小批量取样、RoI pooling层的反向传播、SGD超参数完成微调工作。

除了模型提出以外，文章还进行了很多对比实验验证了多任务损失的作用，softmax和SVM的性能对比，提出了加快训练的方法mini-batch和加快检测时间的方法通过SVD来加速全连接层的计算。

从尺度不变性上，通过对比实验给出了指导性建议值，图像单一尺寸与多个尺寸效果相似。验证了fine-tuning不同卷积层的更新对最终结果的影响，没有必要对所有卷积层进行fine-tuning。并得出selective search滑动窗口提取数量2000已经足够，更多的proposal反而会导致mAP的下降。

主要贡献

1. 比R-CNN，SPP-net更高的检测精度
2. 训练是单步的(single-stage), 并且使用了多任务学习的损失函数，把多个任务的损失函数写到了一起，实现单级的训练过程
3. 训练的时候可以更新所有层的所有网络参数
4. 在训练SVM的时候不需要额外的硬盘存储特征
5. 给出了诸多Fast R-CNN评估结论

个人理解与体会

R-CNN的主要性能瓶颈在于需要对每个提议区域独立抽取特征。由于这些区域通常有大量重叠，独立的特征抽取会导致大量的重复计算。Fast R-CNN对R-CNN的一个主要改进在于只对整个图像做卷积神经网络的前向计算。

Faster R-CNN

背景

上述的目标检测网络依靠region proposal算法来假设目标的位置，SPPnet和Fast R-CNN进行的优化已经减少了这些检测网络的运行时间，但region proposal的计算仍然是一个瓶颈。

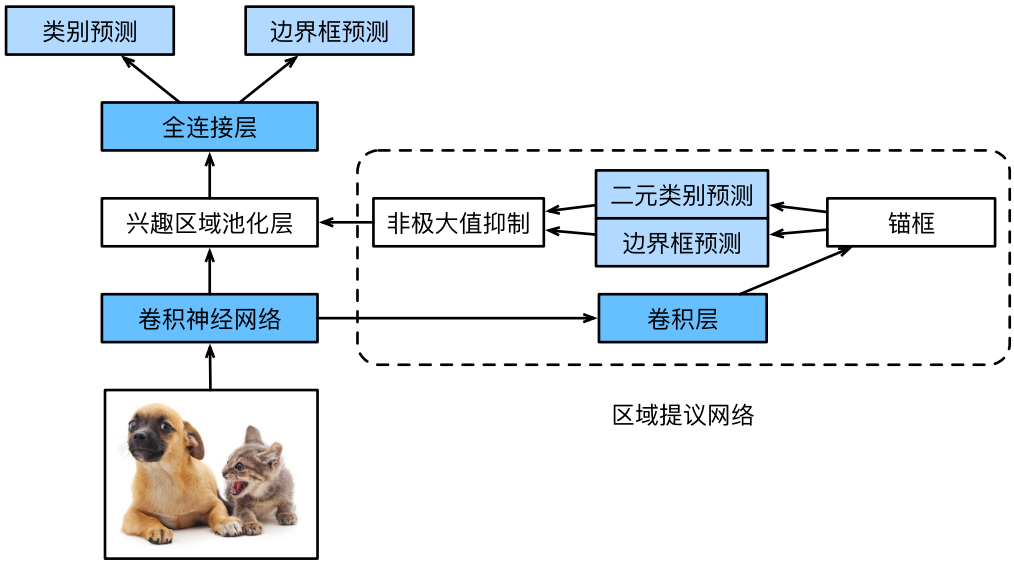
文献

Ren, S., He, K., Girshick, R., & Sunx, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In **Advances in neural information processing systems** (pp. 91-99).

简述

这篇论文提出了一种RegionProposal Network(RPN),它能够和检测网络共享整张图像的卷积特征，从而使计算量大大减小。RPN是一种全卷积的网络，能够同时预测目标的边界以及对objectness得分。通过共享卷及特征进一步将RPN和Fast R-CNN合并成一个网络，使用attention机制，RPN组件能够告诉网络看向哪里。对于VGG-16模型，检测系统在GPU上的帧率为5帧（包含所有步骤），同时仅用每张图300个proposals取得了PASCAL VOC2007,2012以及MS COCO数据集的最好检测精度。

文中首先分析了region proposal使用CPU导致的耗时长，而且其计算过程错过了将网络结构特征共享计算的机会，即使使用GPU仍然开销大。从而详细介绍了RPN，如何完美的与目标检测网络共享卷积层。从而减少提议区域的生成数量，并保证目标检测的精度。



上图描述了Faster R-CNN模型。与Fast R-CNN相比，只有生成提议区域的方法从选择性搜索变成了区域提议网络，而其他部分均保持不变。

文中的Faster RCNN，主要由两个模块组成：第一层是深度全卷积网络来提取区域，第二层为Fast R-CNN检测器。整个系统是一个对象检测的独立、统一的网络。其中提出了锚点Anchor的概念结合平移不变性，建立了一个anchor的金字塔，更加高效，仅仅依赖于单一尺度的图片和feature map，并且使用单一大小的滑窗。

为训练RPN定义了损失函数，其中，i为一个anchor在一个mini-batch中的下标，pi是anchor i为一个object的预测可能性。如果这个anchor是positive的，则ground-truth标签pi为1，否则为0。ti表示预测bounding box的4个参数化坐标，ti是这个positive anchor对应的ground-truth box。分类的损失（classification loss）Lcls是一个二值分类器（是object或者不是）的softmax loss。回归损失（regression loss），其中R是Fast R-CNN中定义的robust ross function (smooth L1)。piLreg表示回归损失只有在positive anchor（pi=1)的时候才会被激活。cls与reg层的输出分别包含{pi}和{ti}。

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*).$$

区域提议网络作为Faster R-CNN的一部分，是和整个模型一起训练得到的。也就是说，Faster R-CNN的目标函数既包括目标检测中的类别和边界框预测，又包括区域提议网络中锚框的二元类别和边界框预测。最终，区域提议网络能够学习到如何生成高质量的提议区域，从而在减少提议区域数量的情况下也能保证目标检测的精度。通过交替优化来学习共享特征，训练算法包含4步：

1. 训练RPN，使用提前训练好的ImageNet模型来初始化，然后对region proposal task进行微调。
2. 得到的proposal和Fast R-CNN来训练一个单独的detector network。这个detector network也是使用提前训练好的ImageNet模型来初始化。
3. 训练好的detector network来初始化RPN，然后训练。这里训练的时候固定共享卷积网络，只微调RPN部分的网络层。这个时候两个网络共享卷积层。
4. 保持共享卷积层固定，只微调Fast R-CNN部分的网络层。

两个网络共享了相同的网络层，形成了一个统一的网络。

主要贡献

1. Faster R-CNN将Fast R-CNN中的选择性搜索替换成区域提议网络，从而减少提议区域的生成数量，并保证目标检测的精度。
2. 提出了使用 RPN来生成 region proposals ，然后使用共享权值减少了网络参数， 使得区域提案的步骤几乎是无损耗的。
3. 通过与其后的检测网络共享卷积特征，使一个一致的，基于深度学习的目标检测系统以近乎实时的帧率运行。

个人理解与体会

Faster R-CNN 可以看做是对 Fast R-CNN 的进一步加速，最主要解决的如何快速获得 proposal，一般的做法都是利用Selective search走一遍待检测图，得到proposal。基于区域的深度卷积网络虽然使用了 GPU 进行加速，但是区域提案方法确却都是在CPU上实现的，这就大大地拖慢了整个系统的速度。然后文中提出，卷积后的特征图同样也是可以用来生成区域提案的。通过增加两个卷积层来实现RPN, 一个用来将每个feature map的位置编码成一个向量，另一个则是对每一个位置输出一个objectness score。RPNs 是一种全卷积网络，为了与Fast R-CNN统一，提出了固定proposal，为生成proposal和目标检测这两个任务交替微调网络 。

目标检测之YOLO系列

YOLO

背景

先前的目标检测工作需要利用分类器来执行检测。而YOLO将目标检测作为一个回归问题来处理空间分离的边界框和相关的类概率。

文献

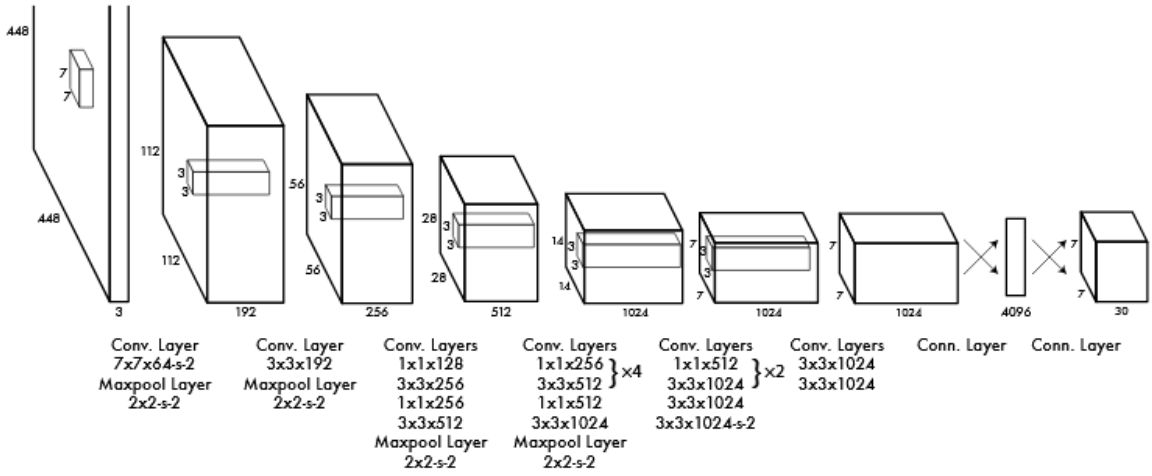
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In **Proceedings of the IEEE conference on computer vision and pattern recognition** (pp. 779-788)

简述

文中提出了一种新的目标检测方法YOLO。单个神经网络在一次评估中直接从完整图像预测包围盒和类概率。由于整个检测管道是一个单一的网络，因此可以直接对检测性能进行端到端的优化。YOLO的统一架构速度非常快。我们的基本YOLO模型以每秒45帧的速度实时处理图像。另一个更小版本的网络Fast YOLO每秒处理155帧。与最先进的检测系统相比，YOLO定位误差更大，但在背景下预测误报的可能性较小。

在准确性上，YOLO算法仍然落后于最先进的检测系统。 虽然它可以快速识别图像中的对象，但它很难精确定位某些对象，特别是小对象。统一检测中将目标检测统一到一个神经网络。网络使用整个图像中的特征来预测每个边界框。 它也是同时预测图像的所有类的所有边界框。 网络学习到的完整图像和图中所有的对象。YOLO设计可实现端到端训练和实时的速度，同时保持较高的平均精度。

YOLO的网络模型受到GoogleNet分类模型的启发，网络有24个卷积层，后面是2个全连接层。然后使用1x1降维层，后面是3x3卷积层。在ImageNet分类任务上以一半的分辨率(224x224的输入图像）预训练卷积层，然后将分辨率加倍来进行检测。同时文中还提供了一种快速版本的YOLO，可以更快速的目标检测，使用了更少的滤波器。除了网络规模外，所有训练和参数都是相同的。



训练中的损失函数，YOLO的损失函数会同样的对待小边界框与大边界框的误差。大边界框的小误差通常是良性的，但小边界框的小误差对IOU的影响要大得多。主要错误来源是不正确的定位。

但YOLO仍然有缺陷，YOLO的每一个网格只预测两个边界框，一种类别。这导致模型对相邻目标预测准确率下降。因此，YOLO对成队列的目标识别准确率较低。由于模型学习从数据中预测边界框，因此它很难泛化到新的、不常见角度的目标。

文中将YOLO与其他几种目标检测算法框架进行比较，突出了关键的相似性和差异性。

- 可变形部件模型。可变形零件模型（DPM）使用滑动窗口方法进行目标检测。DPM使用不相交的流程来提取静态特征，对区域进行分类，预测高评分区域的边界框等。我们的系统用单个卷积神经网络替换所有这些不同的部分。网络同时进行特征提取，边界框预测，非极大值抑制和上下文推理。代替静态特征，网络内嵌地训练特征并为检测任务优化它们。我们的统一架构导致了比DPM更快，更准确的模型。
- R-CNN。R-CNN及其变种使用区域提名而不是滑动窗口来查找图像中的目标。选择性搜索产生潜在的边界框，卷积网络提取特征，SVM对边界框进行评分，线性模型调整边界框，非极大值抑制消除重复检测。这个复杂流程的每个阶段都必须独立地进行精确调整，所得到的系统非常慢，测试时每张图像需要超过40秒。

YOLO与R-CNN有一些相似之处。每个网格单元提出潜在的边界框并使用卷积特征对这些框进行评分。但是，我们的系统对网格单元提出进行了空间限制，这有助于缓解对同一目标的多次检测。我们的系统还提出了更少的边界框，每张图像只有98个，而选择性搜索则只有2000个左右。最后，我们的系统将这些单独的组件组合成一个单一的，共同优化的模型。

- 其它快速检测器。Fast和Faster的R-CNN通过共享计算和使用神经网络替代选择性搜索来提出区域加速R-CNN框架。虽然它们提供了比R-CNN更快的速度和更高的准确度，但两者仍然不能达到实时性能。

主要贡献

1. YOLO模型构造简单，可以直接在全图像上训练。与基于分类器的方法不同，YOLO是基于与检测性能直接对应的损失函数来训练的，整个模型是联合训练的。
2. 快速YOLO是当时速度最快的通用目标检测器，推动了实时目标检测技术的发展。
3. YOLO还可以很好地推广到新的领域，使其成为依赖于快速、健壮的对象检测的应用程序的理想选择。

个人理解与体会

YOLO不是试图优化大型检测流程的单个组件，而是完全抛弃流程，被设计为快速检测。像人脸或行人等单类别的检测器可以高度优化，因为他们必须处理更少的变化。YOLO是一种通用的检测器，可以学习同时检测多个目标。

YOLO是一种快速，精确的目标检测器，非常适合计算机视觉应用。将YOLO连接到网络摄像头，并验证它是否能保持实时性能，包括从摄像头获取图像并显示检测结果的时间。由此产生的系统是交互式和参与式的。虽然YOLO单独处理图像，但当连接到网络摄像头时，其功能类似于跟踪系统，可在目标移动和外观变化时检测目标。

目标检测之SSD系列

SSD

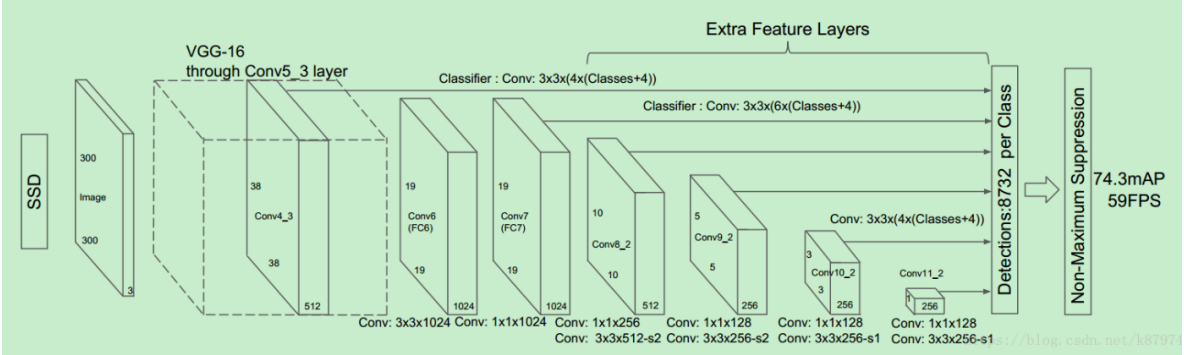
背景

在faster rcnn中，anchors只作用在最后的特征图上，这对检测小物体及位置来说是有不足的，所以SSD想在多个特征图上用anchors回归检测物体。高层特征图的语义信息丰富对分类有益，而低层特征图语义信息少但位置信息多，利于定位。

文献

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In **European conference on computer vision** (pp. 21-37). Springer, Cham

简述



SSD基本网络结构如上。SSD的其他一些trick，负样本挖掘，在生成一系列的 predictions 之后，会产生很多个符合 ground truth box 的 predictions boxes，但同时，不符合 ground truth boxes 也很多，而且这个 negative boxes，远多于 positive boxes。这会造成 negative boxes、positive boxes 之间的不均衡。训练时难以收敛。

文中还作了以下的观察：更多的default boxes会带来更精确的检测，但耗时增加。由于detector以多种分辨率运行于特征上，因此在多个图层上使用MultiBox也会导致更好的检测。80%的时间花在基础VGG-16网络上：这意味着，使用更快，同样精确的网络，SSD的性能可能会更好。SSD将具有相似类别的对象（例如动物）混淆。SSD在较小的对象上产生较差的性能，因为它们可能不会出现在所有功能地图中。增加输入图像分辨率缓解了这个问题，但并未完全解决这个问题。

主要贡献

1. YOLO主要利用conv5_3上的信息进行预测，而SSD利用了多个特征图。进一步提升了识别速度和精度。
2. 预设好anchoris后的如何卷积训练
3. 多特征图上预测，划分特征图

个人理解与体会

YOLO也是单阶段，但SSD比YOLO快还准。SSD虽然在多个特征图上进行分类回归，但是对小目标的识别仍比较差，还达不到faster rcnn的水准。这主要是因为小尺寸的目标多用较低层级的anchor来训练(因为小尺寸目标在较低层级IOU较大)，较低层级的特征非线性程度不够，无法训练到足够的精确度。

参考文献

1. 区域卷积神经网络（R-CNN）系列 http://zh.gluon.ai/chapter_computer-vision/rcnn.html#%E5%8F%82%E8%80%83%E6%96%87%E7%8C%AE
2. RCNN学习笔记(2):Rich feature hierarchies for accurate object detection and semantic segmentation https://blog.csdn.net/u011534057/article/details/51218250
3. 检测之旧文新读(三)-Fast R-CNN https://www.jianshu.com/p/5cdf4058c910
4. Fast RCNN论文总结 https://blog.csdn.net/qq_30159015/article/details/80088444
5. 【论文翻译】Faster R-CNN https://blog.csdn.net/xiaqunfeng123/article/details/78716106
6. 【论文笔记】Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks https://blog.csdn.net/tmylzq187/article/details/51441553
7. 第三十五节，目标检测之YOLO算法详解 https://blog.csdn.net/javastart/article/details/82860718
8. 深度学习笔记（一）空间金字塔池化阅读笔记Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition https://blog.csdn.net/liyaohhh/article/details/50614380?locationNum=11
9. SSD：Single Shot MultiBox Detector解读 https://blog.csdn.net/hancoder/article/details/89387236
10. 目标检测方法系列——R-CNN, SPP, Fast R-CNN, Faster R-CNN, YOLO, SSD https://blog.csdn.net/majinlei121/article/details/53870433
11. CV感悟：YOLO与R-CNN的比较 https://blog.csdn.net/yangzixuan_0608/article/details/103662994