(二)计算机视觉四大基本任务(分类、定位、检测、分割)

知 zhuanlan.zhihu.com/p/31727402

机器学习 计算机视觉 | 个人主页 http://t.cn/EL6QPsx

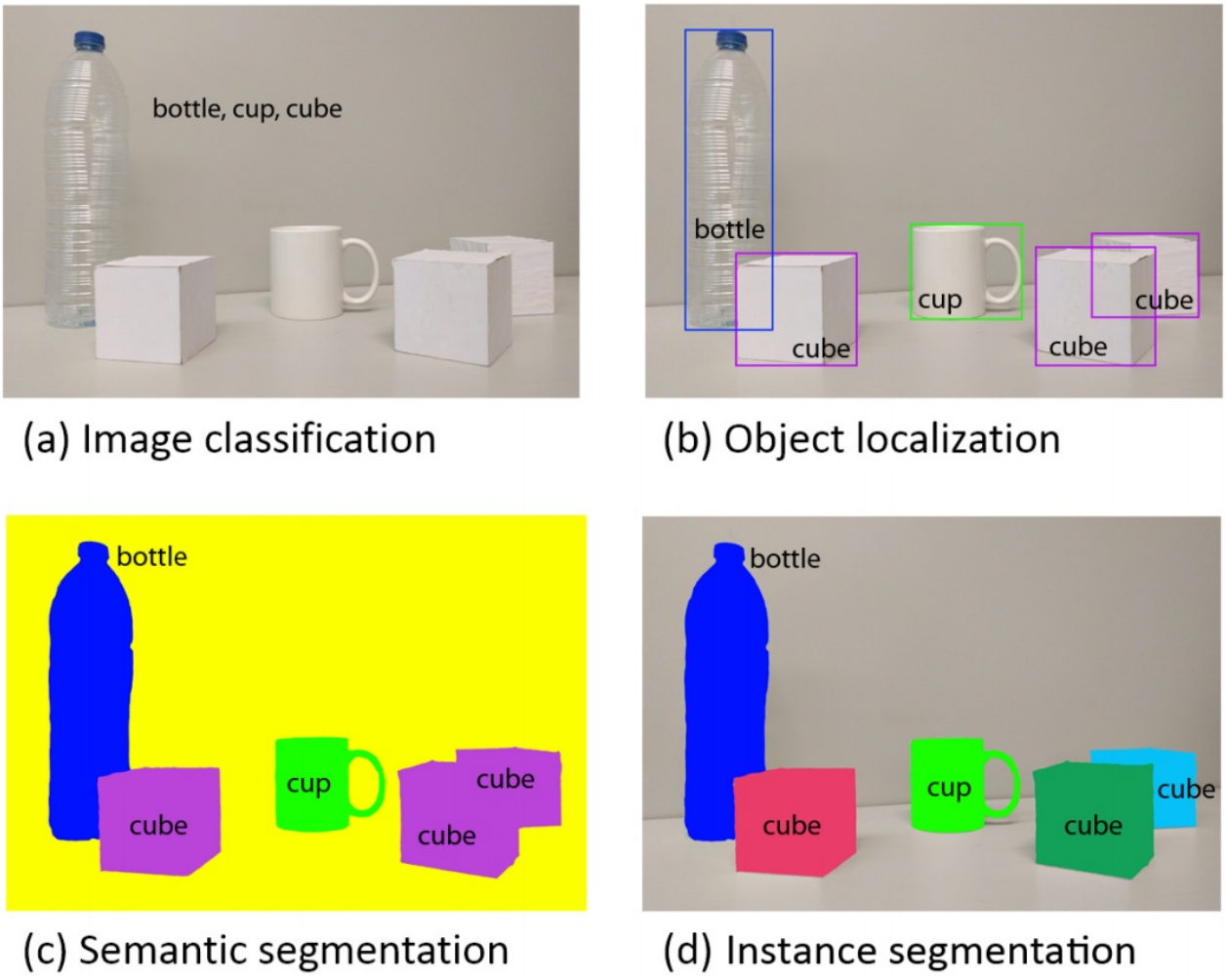
JustDoIT
等

引言

深度学习目前已成为发展最快、最令人兴奋的机器学习领域之一，许多卓有建树的论文已经发表，而且已有很多高质量的开源深度学习框架可供使用。然而，论文通常非常简明扼要并假设读者已对深度学习有相当的理解，这使得初学者经常卡在一些概念的理解上，读论文似懂非懂，十分吃力。另一方面，即使有了简单易用的深度学习框架，如果对深度学习常见概念和基本思路不了解，面对现实任务时不知道如何设计、诊断、及调试网络，最终仍会束手无策。

本系列文章旨在直观系统地梳理深度学习各领域常见概念与基本思想，使读者对深度学习的重要概念与思想有一直观理解，做到"知其然，又知其所以然"，从而降低后续理解论文及实际应用的难度。本系列文章力图用简练的语言加以描述，避免数学公式和繁杂细节。本文是该系列文章中的第二篇，旨在介绍深度学习在计算机视觉领域四大基本任务中的应用，包括分类(图a)、定位、检测(图b)、语义分割(图c)、和实例分割(图d)。后续文章将关注深度学习在计算机视觉领域的其他任务的应用，以及自然语言处理和语音识别。

(本文作者为我本人，部分内容首发于新智元)



计算机视觉(computer vision)简介

计算机视觉旨在识别和理解图像/视频中的内容。其诞生于1966年MIT AI Group的"the summer vision project"。当时，人工智能其他分支的研究已经有一些初步成果。由于人类可以很轻易地进行视觉认知，MIT的教授们希望通过一个暑期项目解决计算机视觉问题。当然，计算机视觉没有被一个暑期内解决，但计算机视觉经过50余年发展已成为一个十分活跃的研究领域。如今，互联网上超过70%的数据是图像/视频，全世界的监控摄像头数目已超过人口数，每天有超过八亿小时的监控视频数据生成。如此大的数据量亟待自动化的视觉理解与分析技术。

计算机视觉的难点在于语义鸿沟。这个现象不仅出现在计算机视觉领域，Moravec悖论发现，高级的推理只需要非常少的计算资源，而低级的对外界的感知却需要极大的计算资源。要让计算机如成人般地下棋是相对容易的，但是要让电脑有如一岁小孩般的感知和行动能力却是相当困难甚至是不可能的。

语义鸿沟(semantic gap) 人类可以轻松地从图像中识别出目标，而计算机看到的图像只是一组0到255之间的整数。

计算机视觉任务的其他困难 拍摄视角变化、目标占据图像的比例变化、光照变化、背景融合、目标形变、遮挡等。

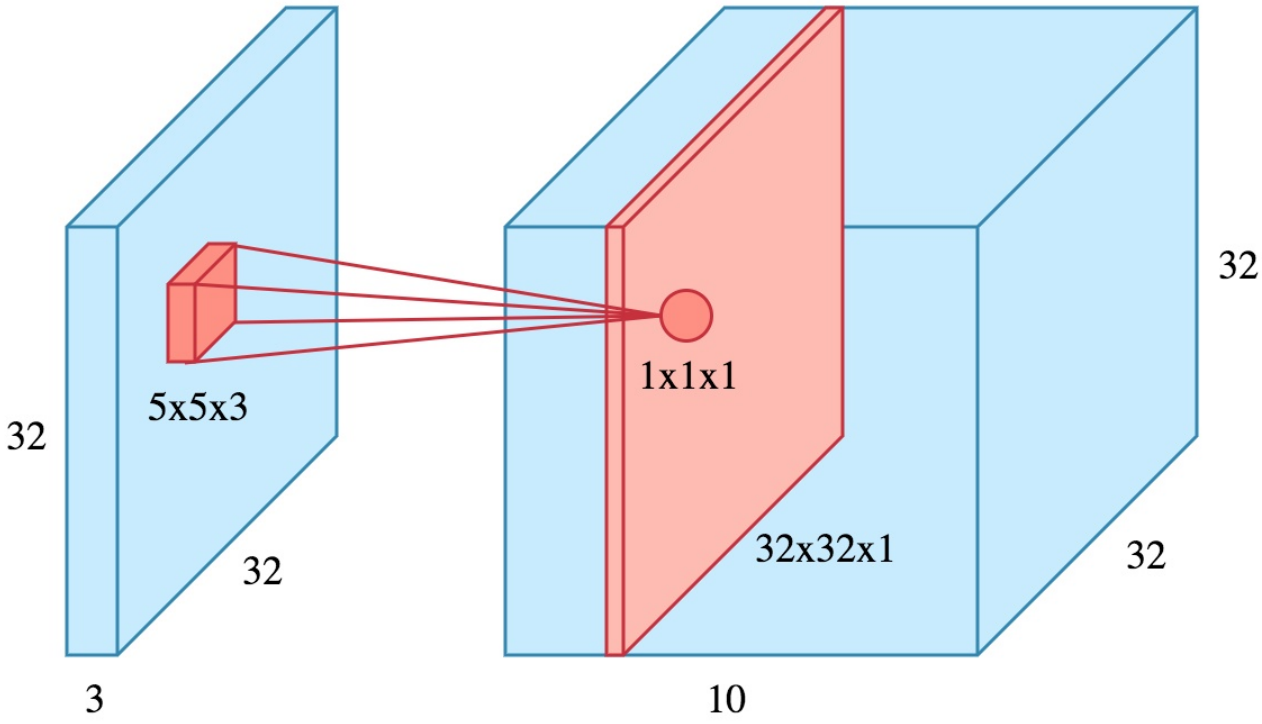
计算机视觉的顶级会议和期刊 顶级会议有CVPR、ICCV、和ECCV，此外ICLR也有不少计算机视觉论文。顶级期刊有IJCV和TPAMI。由于计算机视觉领域发展十分迅速，不论身处学术界或产业界，通过阅读顶级会议和期刊论文了解计算机视觉的最近研究成果都十分必要。

卷积神经网络(convolutional neural networks, CNN)

经典的多层感知机由一系列全连接层组成，卷积神经网络中除全连接层外，还有卷积层和汇合(pooling)层。

(1) 卷积层

为什么要用卷积层 输入图像通常很维数很高，例如，1,000×1,000大小的彩色图像对应于三百万维特征。因此，继续沿用多层感知机中的全连接层会导致庞大的参数量。大参数量需要繁重的计算，而更重要的是，大参数量会有更高的过拟合风险。卷积是局部连接、共享参数版的全连接层。这两个特性使参数量大大降低。卷积层中的权值通常被成为滤波器(filter)或卷积核(convolution kernel)。



局部连接 在全连接层中，每个输出通过权值(weight)和所有输入相连。而在视觉识别中，关键性的图像特征、边缘、角点等只占据了整张图像的一小部分，图像中相距很远的两个像素之间有相互影响的可能性很小。因此，在卷积层中，每个输出神经元在通道方向保持全连接，而在空间方向上只和一小部分输入神经元相连。

共享参数 如果一组权值可以在图像中某个区域提取出有效的表示，那么它们也能在图像的另外区域中提取出有效的表示。也就是说，如果一个模式(pattern)出现在图像中的某个区域，那么它们也可以出现在图像中的其他任何区域。因此，卷积层不同空间位置的神经元共享权值，用

于发现图像中不同空间位置的模式。共享参数是深度学习一个重要的思想，其在减少网络参数的同时仍然能保持很高的网络容量(capacity)。卷积层在空间方向共享参数，而循环神经网络(recurrent neural networks)在时间方向共享参数。

卷积层的作用 通过卷积，我们可以捕获图像的局部信息。通过多层卷积层堆叠，各层提取到特征逐渐由边缘、纹理、方向等低层级特征过度到文字、车轮、人脸等高层级特征。

卷积层中的卷积和数学教材中的卷积是什么关系 基本没有关系。卷积层中的卷积实质是输入和权值的互相关(cross-correlation)函数，而不是数学教材中的卷积。

描述卷积的四个量 一个卷积层的配置由如下四个量确定。1. **滤波器个数**。使用一个滤波器对输入进行卷积会得到一个二维的特征图(feature map)。我们可以用时使用多个滤波器对输入进行卷积，以得到多个特征图。2. **感受野(receptive field)** *F*，即滤波器空间局部连接大小。3. **零填补(zero-padding)** *P*。随着卷积的进行，图像大小将缩小，图像边缘的信息将逐渐丢失。因此，在卷积前，我们在图像上下左右填补一些0，使得我们可以控制输出特征图的大小。4. **步长(stride)** *S*。滤波器在输入每移动*S*个位置计算一个输出神经元。

卷积输入输出的大小关系 假设输入高和宽为*H*和*W*，输出高和宽为*H*'和*W*', 则*H*'=(*H*-*F*+2*P*)/*S*+1, *W*'=(*W*-*F*+2*P*)/*S*+1. 当*S*=1时，通过设定*P*=(*F*-1)/2, 可以保证输入输出空间大小相同。例如，3*3的卷积需要填补一个像素使得输入输出空间大小不变。

应该使用多大的滤波器 尽量使用小的滤波器，如3×3卷积。通过堆叠多层3×3卷积，可以取得与大滤波器相同的感受野，例如三层3×3卷积等效于一层7×7卷积的感受野。但使用小滤波器有以下两点好处。1. **更少的参数量**。假设通道数为*D*，三层3×3卷积的参数量为3×(*D*×*D*×3×3)=27*D*^2, 而一层7×7卷积的参数量为*D*×*D*×7×7=49*D*^2。2. **更多非线性。**由于每层卷积层后都有非线性激活函数，三层3×3卷积一共经过三次非线性激活函数，而一层7×7卷积只经过一次。

1×1卷积 旨在对每个空间位置的*D*维向量做一个相同的线性变换。通常用于增加非线性，或降维，这相当于在通道数方向上进行了压缩。1×1卷积是减少网络计算量和参数的重要方式。

全连接层的卷积层等效 由于全连接层和卷积层都是做点乘，这两种操作可以相互等效。全连接层的卷积层等效只需要设定好卷积层的四个量：滤波器个数等于原全连接层输出神经元个数、感受野等于输入的空间大小、没有零填补、步长为1。

为什么要将全连接层等效为卷积层 全连接层只能处理固定大小的输入，而卷积层可以处理任意大小输入。假设训练图像大小是224×224，而当测试图像大小是256×256。如果不进行全连接层的卷积层等效，我们需要从测试图像中裁剪出多个224×224区域分别前馈网络。而进行卷积层等效后，我们只需要将256×256输入前馈网络一次，即可达到多次前馈224×224区域的效果。

卷积结果的两种视角 卷积结果是一个*D*×*H*×*W*的三维张量。其可以被认为是有*D*个通道，每个通道是一个二维的特征图，从输入中捕获了某种特定的特征。也可以被认为是有*H*×*W*个空间位置，每个空间位置是一个*D*维的描述向量，描述了对应感受野的图像局部区域的语义特征。

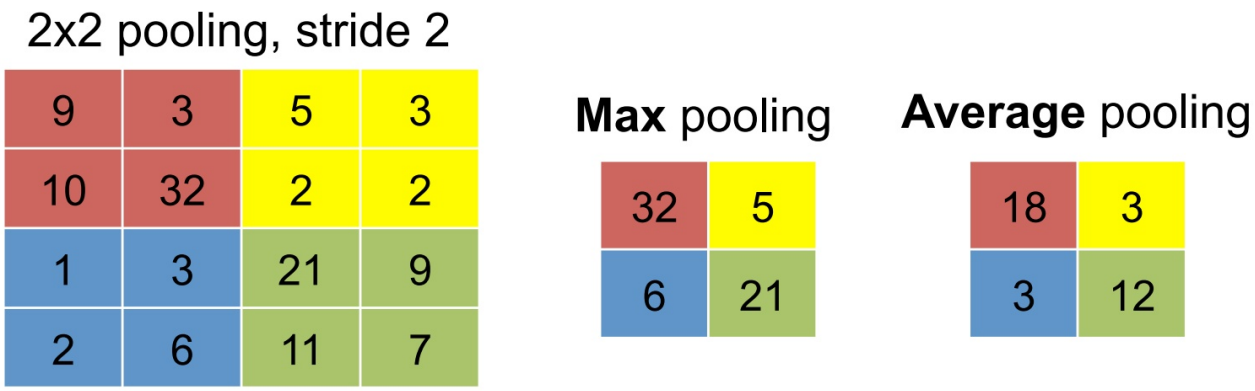
卷积结果的分布式表示 卷积结果的各通道之间不是独立的。卷积结果的各通道的神经元和语义概念之间是一个“多对多”的映射。即，每个语义概念由多个通道神经元一起表示，而每个神经

元又同时参与到多个语义概念中去。并且，神经元响应是稀疏的，即大部分的神经元输出为0。

卷积操作的实现 有如下几种基本思路。1. **快速傅里叶变换(FFT)**。通过变换到频域，卷积运算将变为普通矩阵乘法。实际中，当滤波器尺寸大时效果好，而对于通常使用的1×1和3×3卷积，加速不明显。2. **im2col (image to column)**。im2col将与每个输出神经元相连的局部输入区域展成一个列向量，并将所有得到的向量拼接成一个矩阵。这样卷积运算可以用矩阵乘法实现。im2col的优点是可以利用矩阵乘法的高效实现，而弊端是会占用很大存储，因为输入元素会在生成的矩阵中多次出现。此外，Strassen矩阵乘法和Winograd也常被使用。现有的计算库如MKL和cuDNN，会根据滤波器大小选择合适的算法。

(2) 汇合层

汇合层 根据特征图上的局部统计信息进行下采样，在保留有用信息的同时减少特征图的大小。和卷积层不同的是，汇合层不包含需要学习的参数。最大汇合(max-pooling)在一个局部区域选最大值作为输出，而平均汇合(average pooling)计算一个局部区域的均值作为输出。局部区域汇合中最大汇合使用更多，而全局平均汇合(global average pooling)是更常用的全局汇合方法。



汇合层的作用 汇合层主要有以下三点作用。**1. 增加特征平移不变性。**汇合可以提高网络对微小位移的容忍能力。**2. 减小特征图大小。**汇合层对空间局部区域进行下采样，使下一层需要的参数量和计算量减少，并降低过拟合风险。**3. 最大汇合可以带来非线性。**这是目前最大汇合更常用的原因之一。近年来，有人使用步长为2的卷积层代替汇合层。而在生成式模型中，有研究发现，不使用汇合层会使网络更容易训练。

图像分类(image classification)

给定一张输入图像，图像分类任务旨在判断该图像所属类别。

(1) 图像分类常用数据集

以下是几种常用分类数据集，难度依次递增。 rodrigob.github.io/are_ 列举了各算法在各数据集上的性能排名。

MNIST 60k训练图像、10k测试图像、10个类别、图像大小1×28×28、内容是0-9手写数字。

CIFAR-10 50k训练图像、10k测试图像、10个类别、图像大小3×32×32。

CIFAR-100 50k训练图像、10k测试图像、100个类别、图像大小3×32×32。

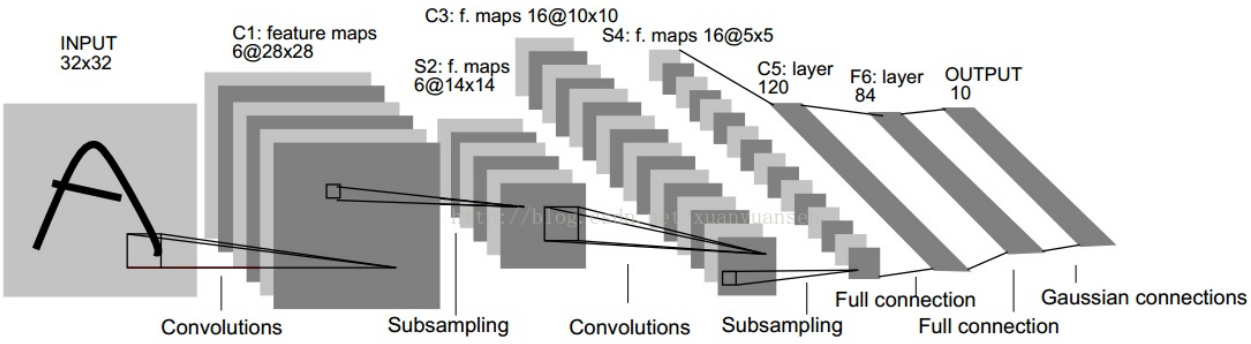
ImageNet 1.2M训练图像、50k验证图像、1k个类别。2017年及之前，每年会举行基于ImageNet数据集的ILSVRC竞赛，这相当于计算机视觉界奥林匹克。

(2) 图像分类经典网络结构

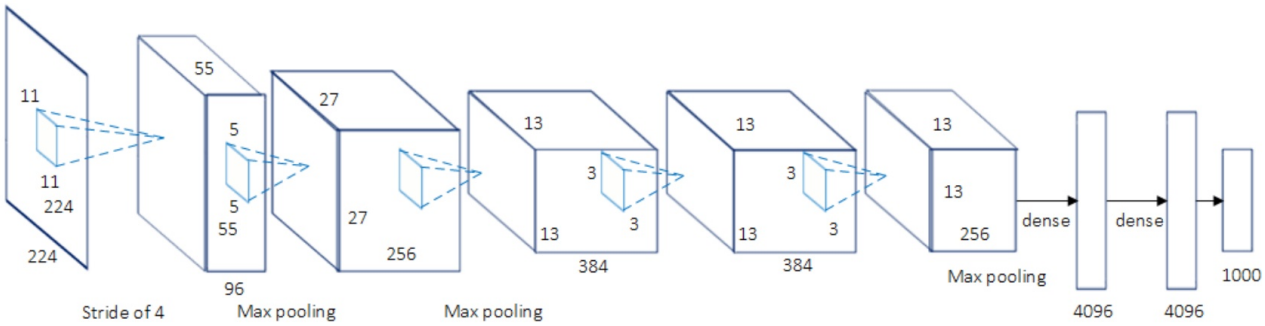
基本架构 我们用conv代表卷积层、bn代表批量归一层、pool代表汇合层。最常见的网络结构顺序是conv -> bn -> relu -> pool，其中卷积层用于提取特征、汇合层用于减少空间大小。随着网络深度的进行，图像的空间大小将越来越小，而通道数会越来越大。

针对你的任务，如何设计网络？ 当面对你的实际任务时，如果你的目标是解决该任务而不是发明新算法，那么不要试图自己设计全新的网络结构，也不要试图从零复现现有的网络结构。找已经公开的实现和预训练模型进行微调。去掉最后一个全连接层和对应softmax，加上对应你任务的全连接层和softmax，再固定住前面的层，只训练你加的部分。如果你的训练数据比较多，那么可以多微调几层，甚至微调所有层。

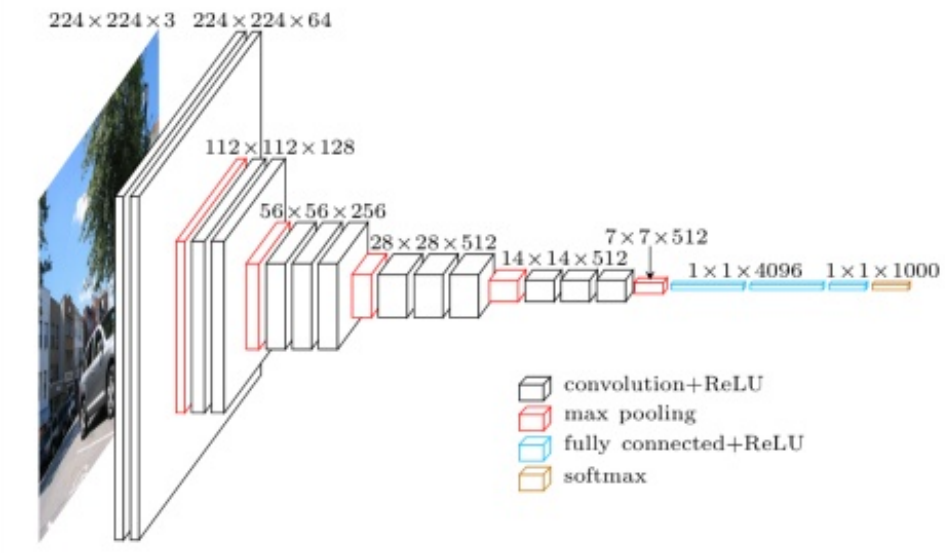
LeNet-5 60k参数。网络基本架构为：conv1 (6) -> pool1 -> conv2 (16) -> pool2 -> fc3 (120) -> fc4 (84) -> fc5 (10) -> softmax。括号中的数字代表通道数，网络名称中有5表示它有5层conv/fc层。当时，LeNet-5被成功用于ATM以对支票中的手写数字进行识别。LeNet取名源自其作者姓LeCun。



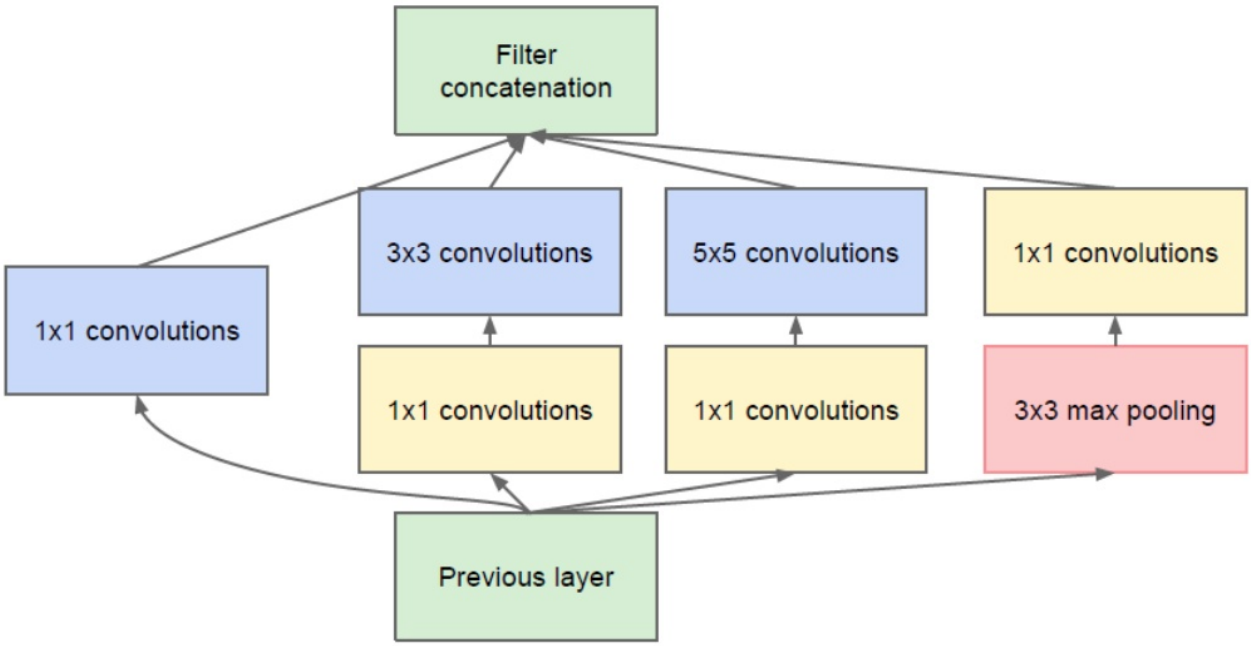
AlexNet 60M参数，ILSVRC 2012的冠军网络。网络基本架构为：conv1 (96) -> pool1 -> conv2 (256) -> pool2 -> conv3 (384) -> conv4 (384) -> conv5 (256) -> pool5 -> fc6 (4096) -> fc7 (4096) -> fc8 (1000) -> softmax。AlexNet有着和LeNet-5相似网络结构，但更深、有更多参数。conv1使用11×11的滤波器、步长为4使空间大小迅速减小(227×227 -> 55×55)。AlexNet的关键点是：(1). 使用了**ReLU**激活函数，使之有更好的梯度特性、训练更快。(2). 使用了**随机失活(dropout)**。(3). 大量使用**数据扩充**技术。AlexNet的意义在于它以高出第二名10%的性能取得了当年ILSVRC竞赛的冠军，这使人们意识到卷积神经网络的优势。此外，AlexNet也使人们意识到可以利用GPU加速卷积神经网络训练。AlexNet取名源自其作者名Alex。



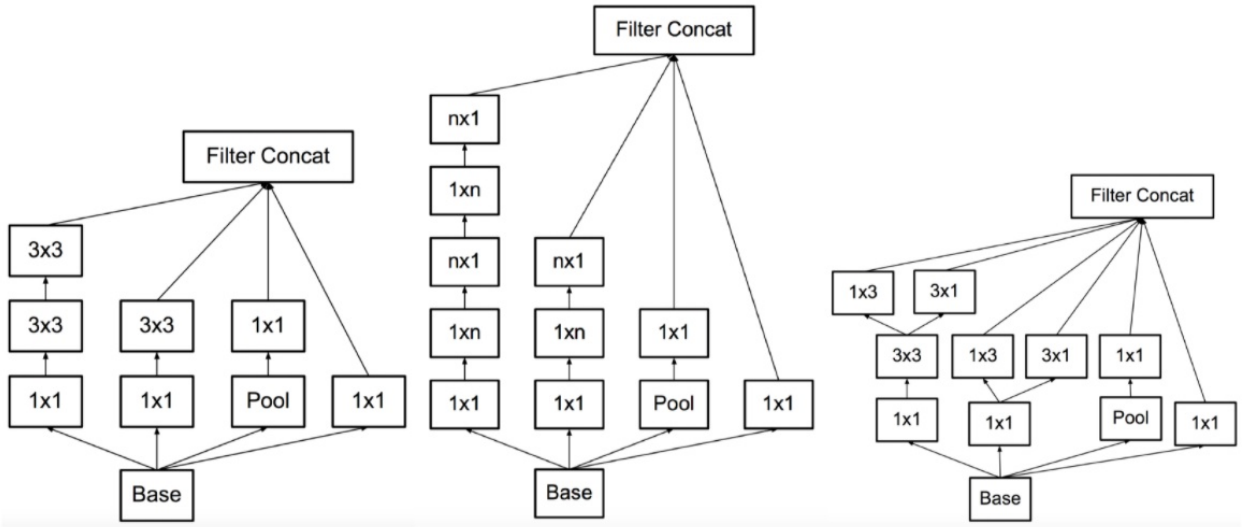
VGG-16/VGG-19 138M参数，ILSVRC 2014的亚军网络。VGG-16的基本架构为：conv1^2 (64) -> pool1 -> conv2^2 (128) -> pool2 -> conv3^3 (256) -> pool3 -> conv4^3 (512) -> pool4 -> conv5^3 (512) -> pool5 -> fc6 (4096) -> fc7 (4096) -> fc8 (1000) -> softmax。 ^3代表重复3次。VGG网络的关键点是：(1). **结构简单**，只有3×3卷积和2×2汇合两种配置，并且**重复堆叠**相同的模块组合。卷积层不改变空间大小，每经过一次汇合层，空间大小减半。(2). **参数量大**，而且大部分的参数集中在全连接层中。网络名称中有16表示它有16层conv/fc层。(3). 合适的网络**初始化**和使用批量归一(batch normalization)层对训练深层网络很重要。在原论文中无法直接训练深层VGG网络，因此先训练浅层网络，并使用浅层网络对深层网络进行初始化。在BN出现之后，伴随其他技术，后续提出的深层网络可以直接得以训练。VGG-19结构类似于VGG-16，有略好于VGG-16的性能，但VGG-19需要消耗更大的资源，因此实际中VGG-16使用得更多。由于VGG-16网络结构十分简单，并且很适合迁移学习，因此至今VGG-16仍在广泛使用。VGG-16和VGG-19取名源自作者所处研究组名(Visual Geometry Group)。



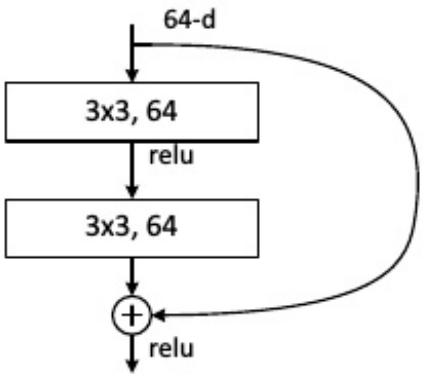
GoogLeNet 5M参数，ILSVRC 2014的冠军网络。GoogLeNet试图回答在设计网络时究竟应该选多大尺寸的卷积、或者应该选汇合层。其提出了Inception模块，同时用1×1、3×3、5×5卷积和3×3汇合，并保留所有结果。网络基本架构为：conv1 (64) -> pool1 -> conv2^2 (64, 192) -> pool2 -> inc3 (256, 480) -> pool3 -> inc4^5 (512, 512, 512, 528, 832) -> pool4 -> inc5^2 (832, 1024) -> pool5 -> fc (1000)。GoogLeNet的关键点是：(1). **多分支**分别处理，并级联结果。(2). 为了降低计算量，用了**1×1卷积**降维。GoogLeNet使用了全局平均汇合替代全连接层，使网络参数大幅减少。GoogLeNet取名源自作者所处单位(Google)，其中L大写是为了向LeNet致敬，而Inception的名字来源于盗梦空间中的"we need to go deeper"梗。

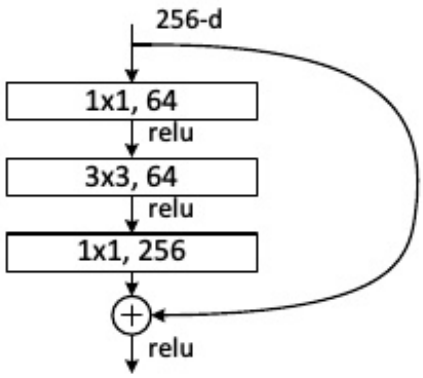


Inception v3/v4 在GoogLeNet的基础上进一步降低参数。其和GoogLeNet有相似的Inception模块，但将7×7和5×5卷积分解成若干等效3×3卷积，并在网络中后部分把3×3卷积分解为1×3和3×1卷积。这使得在相似的网络参数下网络可以部署到42层。此外，Inception v3使用了批量归一层。Inception v3是GoogLeNet计算量的2.5倍，而错误率较后者下降了3%。Inception v4在Inception模块基础上结合了residual模块(见下文)，进一步降低了0.4%的错误率。



ResNet ILSVRC 2015的冠军网络。ResNet旨在解决网络加深后训练难度增大的现象。其提出了residual模块，包含两个3×3卷积和一个短路连接(左图)。短路连接可以有效缓解反向传播时由于深度过深导致的梯度消失现象，这使得网络加深之后性能不会变差。短路连接是深度学习又一重要思想，除计算机视觉外，短路连接也被用到了机器翻译、语音识别/合成领域。此外，具有短路连接的ResNet可以看作是许多不同深度而共享参数的网络的集成，网络数目随层数指数增加。ResNet的关键点是：(1). 使用**短路连接**，使训练深层网络更容易，并且**重复堆叠**相同的模块组合。(2). ResNet大量使用了**批量归一层**。(3). 对于很深的网络(超过50层)，ResNet使用了更高效的**瓶颈(bottleneck)**结构(右图)。ResNet在ImageNet上取得了超过人的准确率。

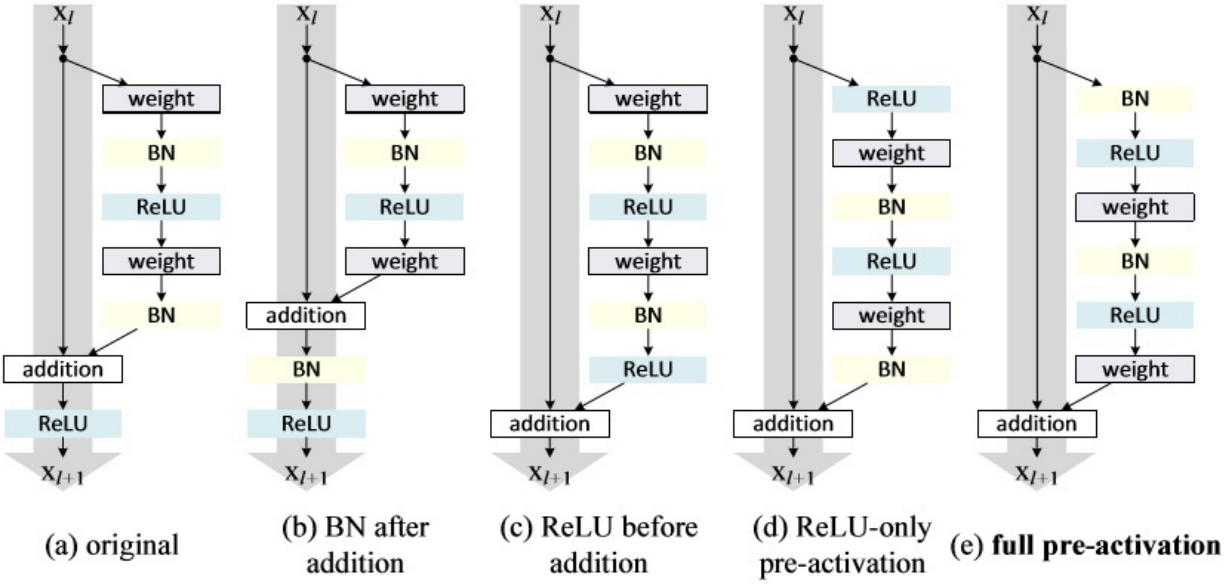




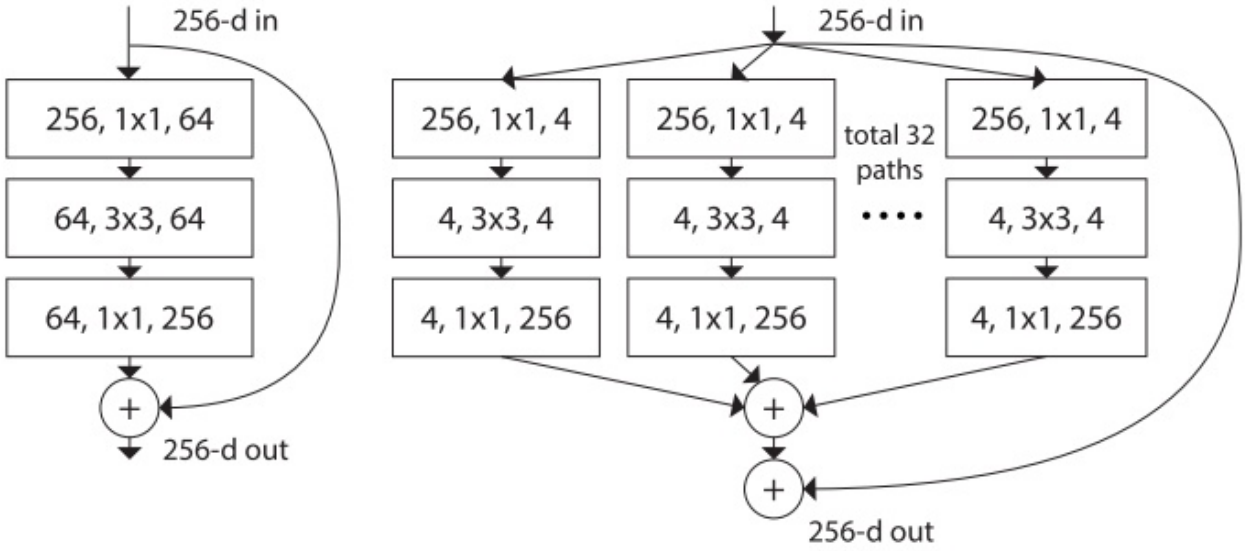
下图对比了上述几种网络结构。

Metrics	LeNet 5	AlexNet	Overfeat fast	VGG 16	GoogLeNet v1	ResNet 50
Top-5 error†	n/a	16.4	14.2	7.4	6.7	5.3
Top-5 error (single crop)†	n/a	19.8	17.0	8.8	10.7	7.0
Input Size	28×28	227×227	231×231	224×224	224×224	224×224
# of CONV Layers	2	5	5	13	57	53
Depth in # of CONV Layers	2	5	5	13	21	49
Filter Sizes	5	3,5,11	3,5,11	3	1,3,5,7	1,3,7
# of Channels	1, 20	3-256	3-1024	3-512	3-832	3-2048
# of Filters	20, 50	96-384	96-1024	64-512	16-384	64-2048
Stride	1	1,4	1,4	1	1,2	1,2
Weights	2.6k	2.3M	16M	14.7M	6.0M	23.5M
MACs	283k	666M	2.67G	15.3G	1.43G	3.86G
# of FC Layers	2	3	3	3	1	1
Filter Sizes	1,4	1,6	1,6,12	1,7	1	1
# of Channels	50, 500	256-4096	1024-4096	512-4096	1024	2048
# of Filters	10, 500	1000-4096	1000-4096	1000-4096	1000	1000
Weights	58k	58.6M	130M	124M	1M	2M
MACs	58k	58.6M	130M	124M	1M	2M
Total Weights	60k	61M	146M	138M	7M	25.5M
Total MACs	341k	724M	2.8G	15.5G	1.43G	3.9G

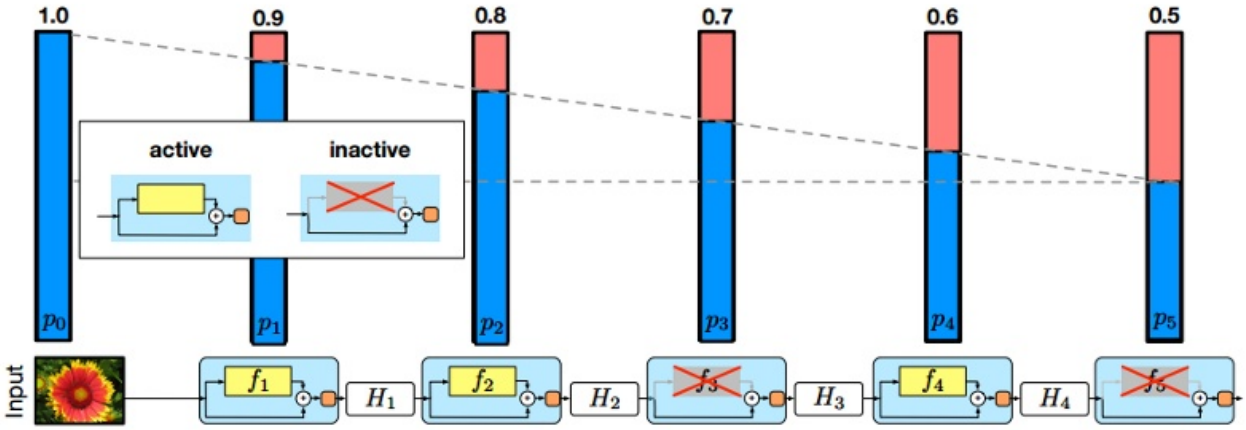
preResNet ResNet的改进。preResNet整了residual模块中各层的顺序。相比经典residual模块(a)，(b)将BN共享会更加影响信息的短路传播，使网络更难训练、性能也更差；(c)直接将ReLU移到BN后会使该分支的输出始终非负，使网络表示能力下降；(d)将ReLU提前解决了(e)的非负问题，但ReLU无法享受BN的效果；(e)将ReLU和BN都提前解决了(d)的问题。preResNet的短路连接(e)能更加直接的传递信息，进而取得了比ResNet更好的性能。



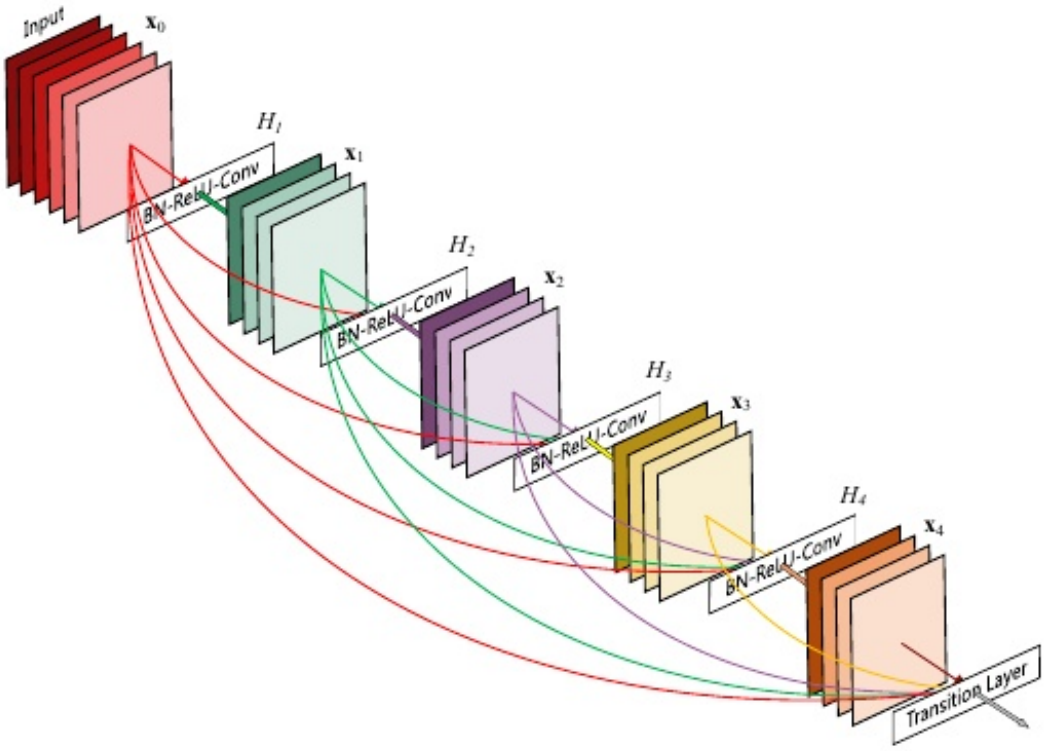
ResNeXt ResNet的另一改进。传统的方法通常是靠加深或加宽网络来提升性能，但计算开销也会随之增加。ResNeXt旨在不改变模型复杂度的情况下提升性能。受精简而高效的Inception模块启发，ResNeXt将ResNet中非短路那一分支变为多个分支。和Inception不同的是，每个分支的结构都相同。ResNeXt的关键点是：(1). 沿用ResNet的**短路连接**，并且重复堆叠相同的模块组合。(2). **多分支**分别处理。(3). 使用**1×1卷积**降低计算量。其综合了ResNet和Inception的优点。此外，ResNeXt巧妙地利用分组卷积进行实现。ResNeXt发现，增加分支数是比加深或加宽更有效地提升网络性能的方式。ResNeXt的命名旨在说明这是下一代(next)的ResNet。



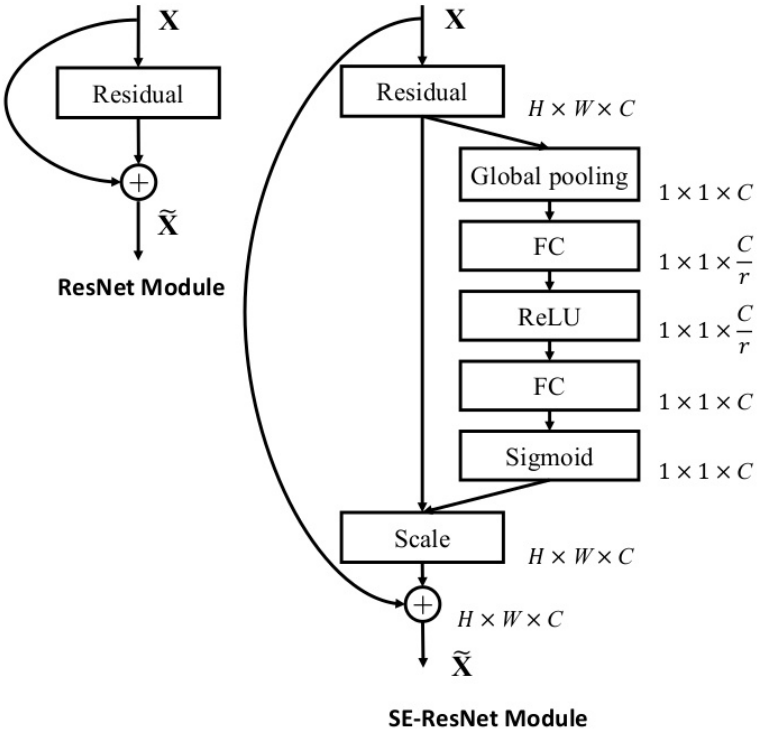
随机深度 ResNet的改进。旨在缓解梯度消失和加速训练。类似于随机失活(dropout)，其以一定概率随机将residual模块失活。失活的模块直接由短路分支输出，而不经过有参数的分支。在测试时，前馈经过全部模块。随机深度说明residual模块是有信息冗余的。



DenseNet 其目的也是避免梯度消失。和residual模块不同，dense模块中任意两层之间均有短路连接。也就是说，每一层的输入通过级联(concatenation)包含了之前所有层的结果，即包含由低到高所有层次的特征。和之前方法不同的是，DenseNet中卷积层的滤波器数很少。DenseNet只用ResNet一半的参数即可达到ResNet的性能。实现方面，作者在大会报告指出，直接将输出级联会占用很大GPU存储。后来，通过共享存储，可以在相同的GPU存储资源下训练更深的DenseNet。但由于有些中间结果需要重复计算，该实现会增加训练时间。



SENet ILSVRC 2017的冠军网络。SENet通过额外的分支(gap-fc-fc-sigm)来得到每个通道的[0, 1]权重，自适应地校正原各通道激活值响应。以提升有用的通道响应并抑制对当前任务用处不大的通道响应。



目标定位(object localization)

在图像分类的基础上，我们还想知道图像中的目标具体在图像的什么位置，通常是以包围盒的(bounding box)形式。

基本思路 多任务学习，网络带有两个输出分支。一个分支用于做图像分类，即全连接+softmax判断目标类别，和单纯图像分类区别在于这里还另外需要一个“背景”类。另一个分支用于判断目标位置，即完成回归任务输出四个数字标记包围盒位置(例如中心点横纵坐标和包围盒长宽)，该分支输出结果只有在分类分支判断不为“背景”时才使用。

人体位姿定位/人脸定位 目标定位的思路也可以用于人体位姿定位或人脸定位。这两者都需要我们对一系列的人体关节或人脸关键点进行回归。

弱监督定位 由于目标定位是相对比较简单的任务，近期的研究热点是在只有标记信息的条件下进行目标定位。其基本思路是从卷积结果中找到一些较高响应的显著性区域，认为这个区域对应图像中的目标。

目标检测(object detection)

在目标定位中，通常只有一个或固定数目的目标，而目标检测更一般化，其图像中出现的目标种类和数目都不定。因此，目标检测是比目标定位更具挑战性的任务。

(1) 目标检测常用数据集

PASCAL VOC 包含20个类别。通常是用VOC07和VOC12的trainval并集作为训练，用VOC07的测试集作为测试。

MS COCO COCO比VOC更困难。COCO包含80k训练图像、40k验证图像、和20k没有公开标记的测试图像(test-dev)，80个类别，平均每张图7.2个目标。通常是用80k训练和35k验证图像的并集作为训练，其余5k图像作为验证，20k测试图像用于线上测试。

mAP (mean average precision) 目标检测中的常用评价指标，计算方法如下。当预测的包围盒和真实包围盒的交并比大于某一阈值(通常为0.5)，则认为该预测正确。对每个类别，我们画出它的查准率-查全率(precision-recall)曲线，平均准确率是曲线下的面积。之后再对所有类别的平均准确率求平均，即可得到mAP，其取值为[0, 100%]。

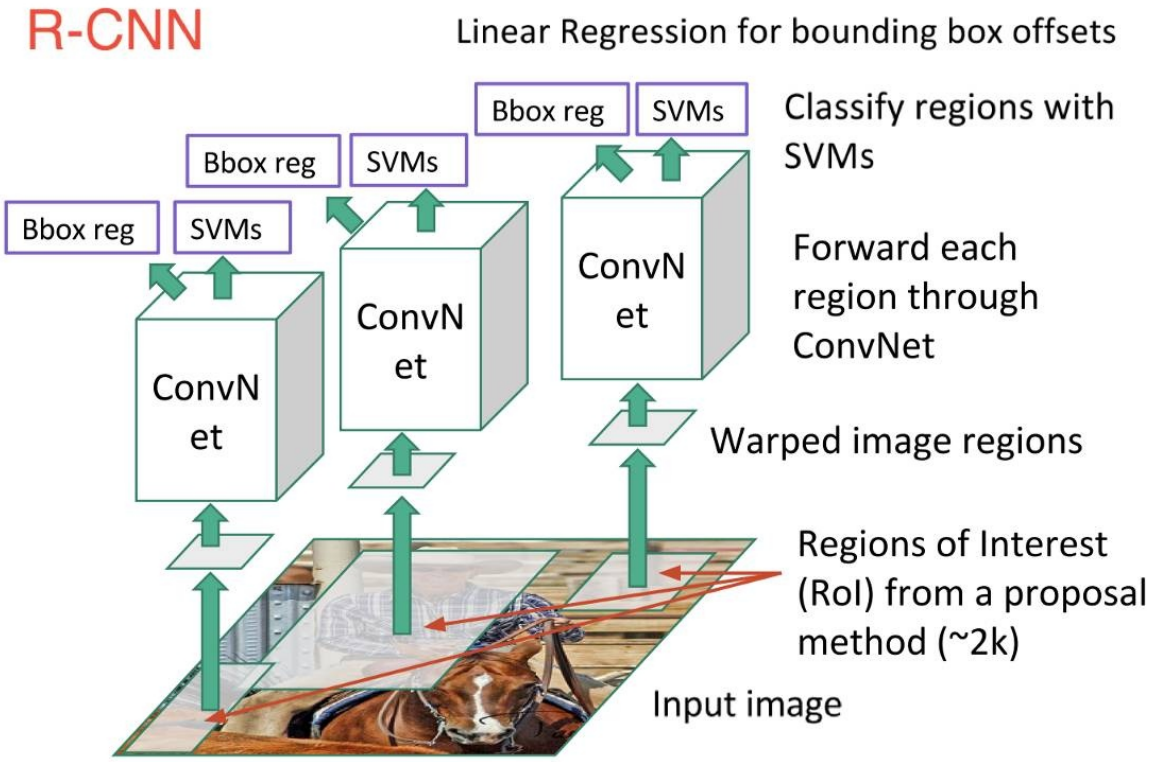
交并比(intersection over union, IoU) 算法预测的包围盒和真实包围盒交集的面积除以这两个包围盒并集的面积，取值为[0, 1]。交并比度量了算法预测的包围盒和真实包围盒的接近程度，交并比越大，两个包围盒的重叠程度越高。

(2) 基于候选区域的目标检测算法

基本思路 使用不同大小的窗口在图像上滑动，在每个区域，对窗口内的区域进行目标定位。即，将每个窗口内的区域前馈网络，其分类分支用于判断该区域的类别，回归分支用于输出包围盒。基于滑动窗的目标检测动机是，尽管原图中可能包含多个目标，但滑动窗对应的图像局部区域内通常只会有一个目标(或没有)。因此，我们可以沿用目标定位的思路对窗口内区域逐个进行处理。但是，由于该方法要把图像所有区域都滑动一遍，而且滑动窗大小不一，这会带来很大的计算开销。

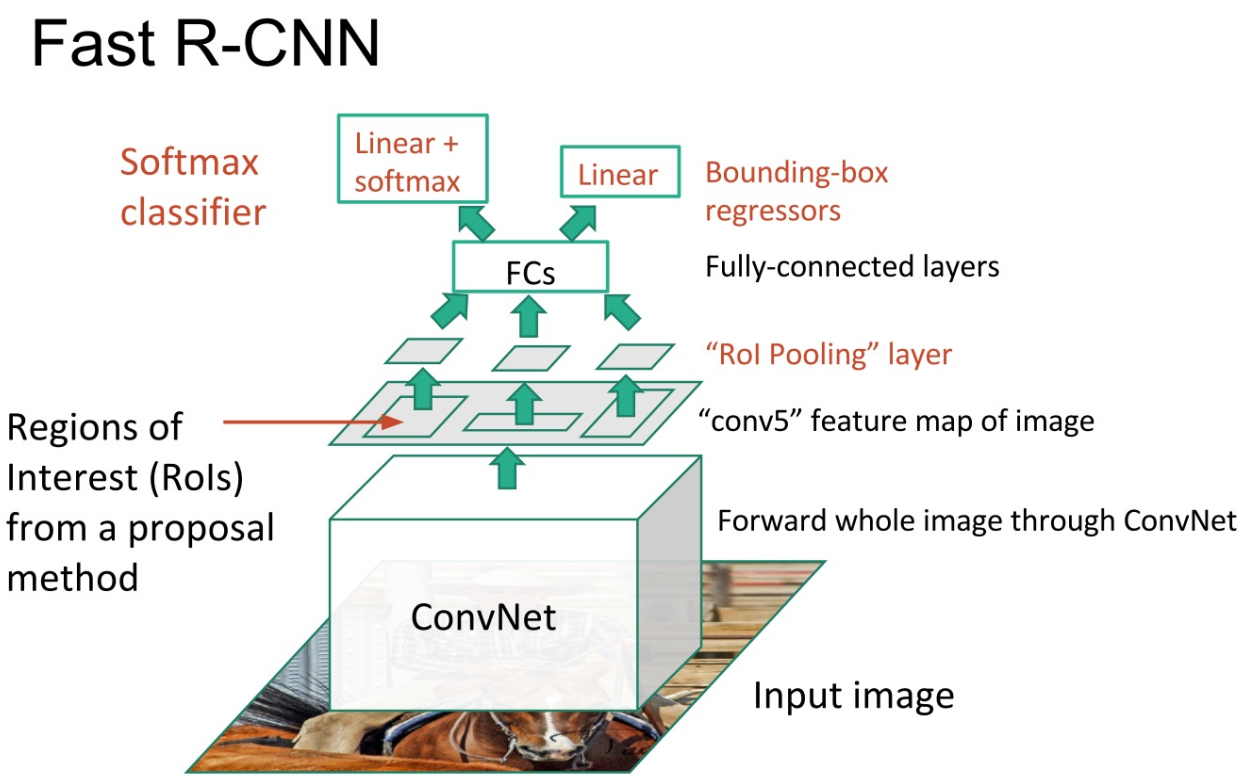
R-CNN 先利用一些非深度学习的类别无关的无监督方法，在图像中找到一些可能包含目标的候选区域。之后，对每个候选区域前馈网络，进行目标定位，即两分支(分类+回归)输出。其中，我们仍然需要回归分支的原因是，候选区域只是对包含目标区域的一个粗略的估计，我们需要有监督地利用回归分支得到更精确的包围盒预测结果。R-CNN的重要性在于当时目标检测已接近瓶颈期，而R-CNN利于在ImageNet预训练模型微调的方法一举将VOC上mAP由35.1%提升至53.7%，确定了深度学习下目标检测的基本思路。一个有趣之处是R-CNN论文开篇第一句只有两个词"Features matter." 这点明了深度学习方法的核心。

候选区域(region proposal) 候选区域生成算法通常基于图像的颜色、纹理、面积、位置等合并相似的像素，最终可以得到一系列的候选矩阵区域。这些算法，如selective search或EdgeBoxes，通常只需要几秒的CPU时间，而且，一个典型的候选区域数目是2k，相比于用滑动窗把图像所有区域都滑动一遍，基于候选区域的方法十分高效。另一方面，这些候选区域生成算法的查准率(precision)一般，但查全率(recall)通常比较高，这使得我们不容易遗漏图像中的目标。



Fast R-CNN R-CNN的弊端是需要多次前馈网络，这使得R-CNN的运行效率不高，预测一张图像需要47秒。Fast R-CNN同样基于候选区域进行目标检测，但受SPPNet启发，在Fast R-CNN中，不同候选区域的卷积特征提取部分是共享的。也就是说，我们先将整副图像前馈网络，并提取conv5卷积特征。之后，基于在原始图像上运行候选区域生成算法的结果在卷积特征上进行采样，这一步称为兴趣区域汇合。最后，对每个候选区域，进行目标定位，即两分支(分类+回归)输出。

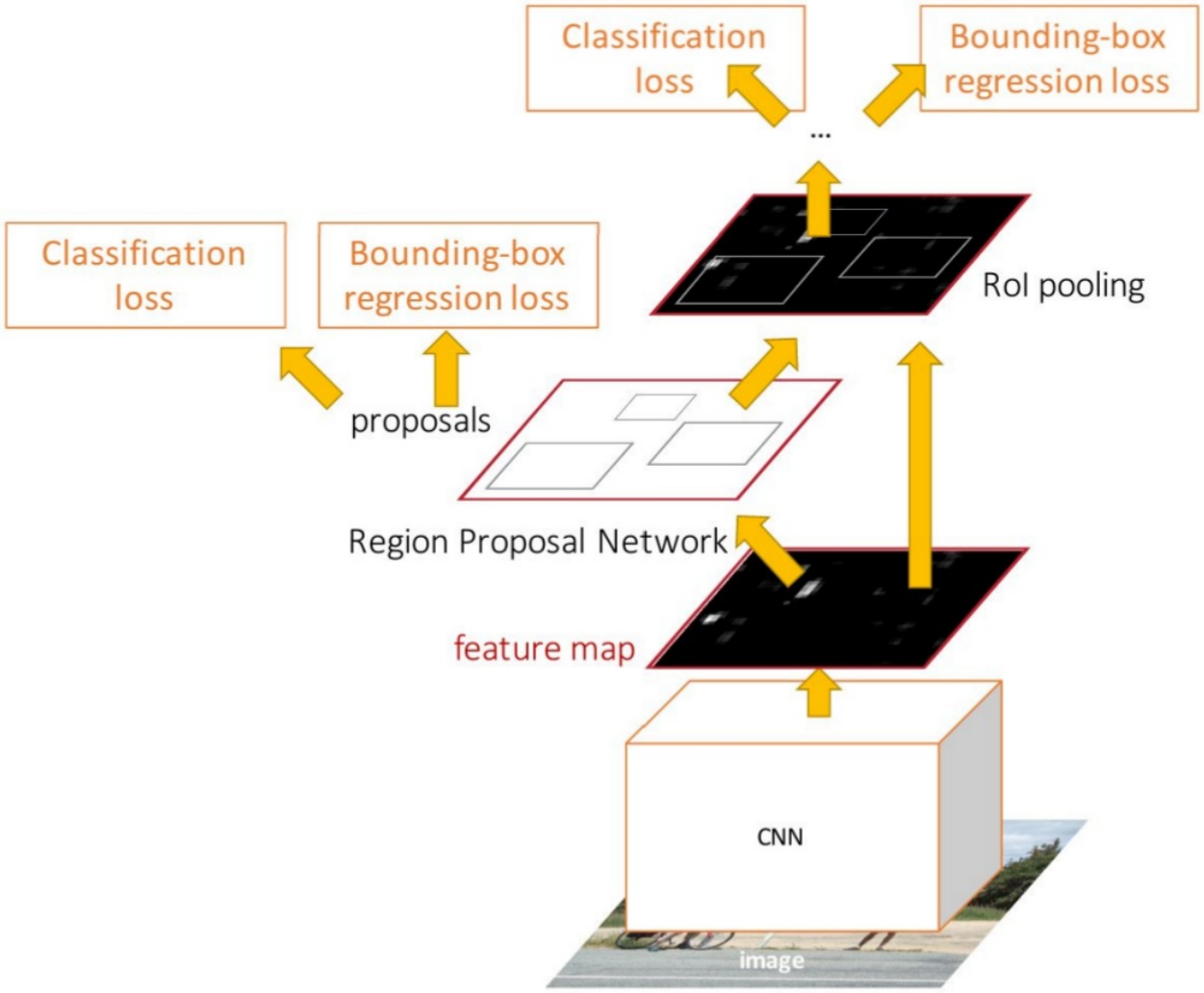
兴趣区域汇合(region of interest pooling, RoI pooling) 兴趣区域汇合旨在由任意大小的候选区域对应的局部卷积特征提取得到固定大小的特征，这是因为下一步的两分支网络由于有全连接层，需要其输入大小固定。其做法是，先将候选区域投影到卷积特征上，再把对应的卷积特征区域空间上划分成固定数目的网格(数目根据下一步网络希望的输入大小确定，例如VGGNet需要7×7的网格)，最后在每个小的网格区域内进行最大汇合，以得到固定大小的汇合结果。和经典最大汇合一致，每个通道的兴趣区域汇合是独立的。



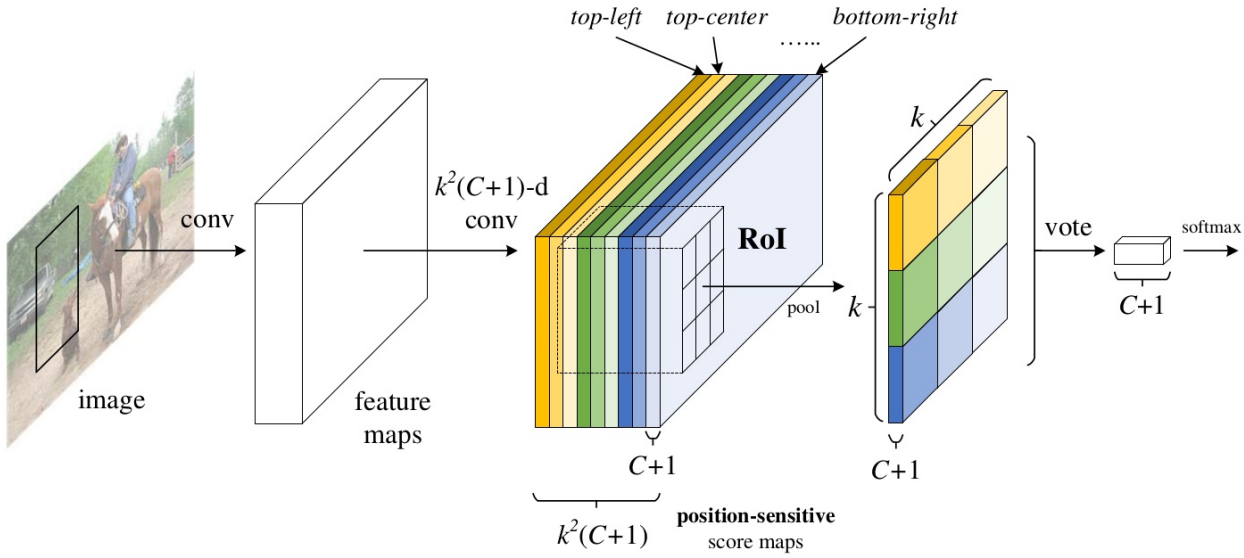
Faster R-CNN Fast R-CNN测试时每张图像前馈网络只需0.2秒，但瓶颈在于提取候选区域需要2秒。Faster R-CNN不再使用现有的无监督候选区域生成算法，而利用候选区域网络从conv5特征中产生候选区域，并且将候选区域网络集成到整个网络中端到端训练。Faster R-CNN的测试时间是0.2秒，接近实时。后来有研究发现，通过使用更少的候选区域，可以在性能损失不大的条件下进一步提速。

候选区域网络(region proposal networks, RPN) 在卷积特征上的通过两层卷积(3×3和1×1卷积)，输出两个分支。其中，一个分支用于判断每个锚盒是否包含了目标，另一个分支对每个锚盒输出候选区域的4个坐标。候选区域网络实际上延续了基于滑动窗进行目标定位的思路，不同之处在于候选区域网络在卷积特征而不是在原图上进行滑动。由于卷积特征的空间大小很小而感受野很大，即使使用3×3的滑动窗，也能对应于很大的原图区域。Faster R-CNN实际使用了3组大小(128×128、256×256、512×512)、3组长宽比(1:1、1:2、2:1)，共计9个锚盒，这里锚盒的大小已经超过conv5特征感受野的大小。对一张1000×600的图像，可以得到20k个锚盒。

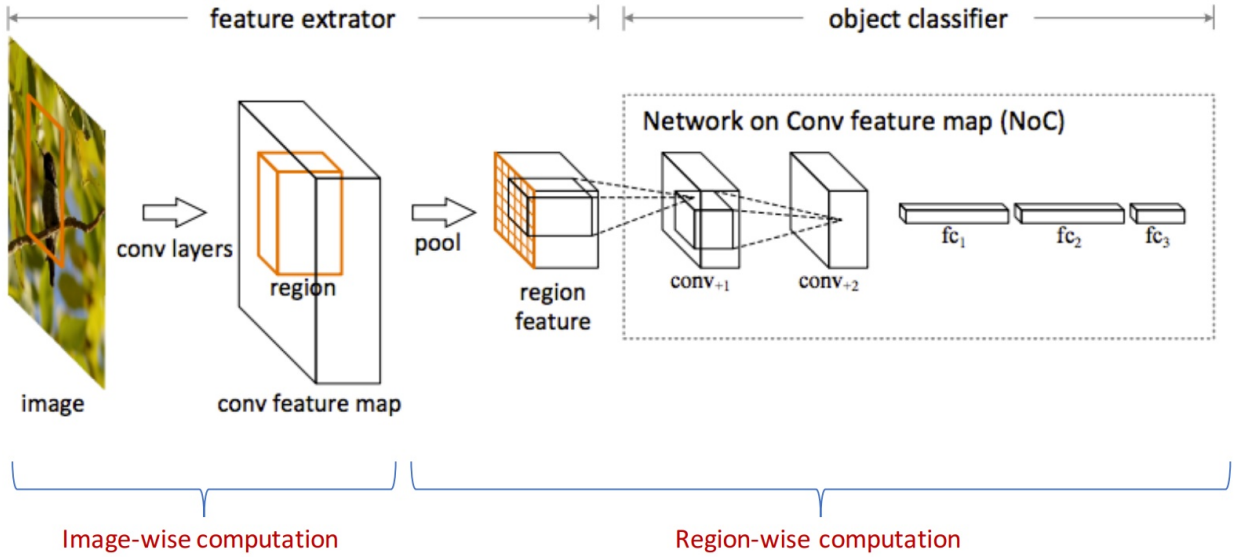
为什么要使用锚盒(anchor box) 锚盒是预先定义形状和大小的包围盒。使用锚盒的原因包括：(1). 图像中的候选区域大小和长宽比不同，直接回归比对锚盒坐标修正训练起来更困难。(2). conv5特征感受野很大，很可能该感受野内包含了不止一个目标，使用多个锚盒可以同时对感受野内出现的多个目标进行预测。(3). 使用锚盒也可以认为这是向神经网络引入先验知识的一种方式。我们可以根据数据中包围盒通常出现的形状和大小设定一组锚盒。锚盒之间是独立的，不同的锚盒对应不同的目标，比如高瘦的锚盒对应于人，而矮胖的锚盒对应于车辆。



R-FCN Faster R-CNN在RoI pooling之后，需要对每个候选区域单独进行两分支预测。R-FCN旨在使几乎所有的计算共享，以进一步加快速度。由于图像分类任务不关心目标具体在图像的位置，网络具有平移不变性。但目标检测中由于要回归出目标的位置，所以网络输出应当受目标平移的影响。为了缓和这两者的矛盾，R-FCN显式地给予深度卷积特征各通道以位置关系。在RoI汇合时，先将候选区域划分成3×3的网格，之后将不同网格对应于候选卷积特征的不同通道，最后每个网格分别进行平均汇合。R-FCN同样采用了两分支(分类+回归)输出。



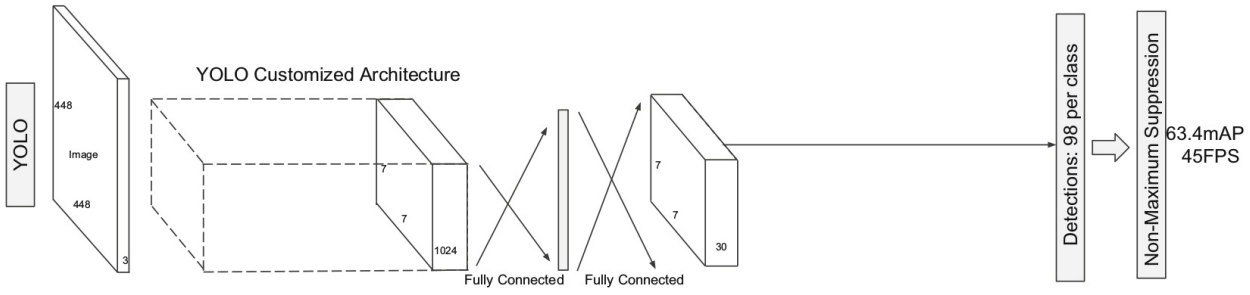
小结 基于候选区域的目标检测算法通常需要两步：第一步是从图像中提取深度特征，第二步是对每个候选区域进行定位(包括分类和回归)。其中，第一步是图像级别计算，一张图像只需要前馈该部分网络一次，而第二步是区域级别计算，每个候选区域都分别需要前馈该部分网络一次。因此，第二步占用了整体主要的计算开销。R-CNN, Fast R-CNN, Faster R-CNN, R-FCN这些算法的演进思路是逐渐提高网络中图像级别计算的比例，同时降低区域级别计算的比例。R-CNN中几乎所有的计算都是区域级别计算，而R-FCN中几乎所有的计算都是图像级别计算。



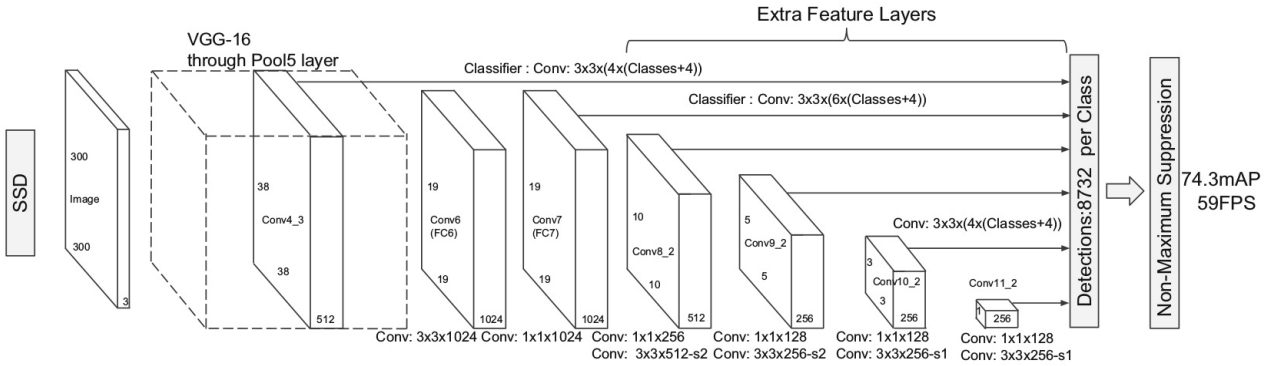
(3) 基于直接回归的目标检测算法

基本思路 基于候选区域的方法由于有两步操作，虽然检测性能比较好，但速度上离实时仍有一些差距。基于直接回归的方法不需要候选区域，直接输出分类/回归结果。这类方法由于图像只需前馈网络一次，速度通常更快，可以达到实时。

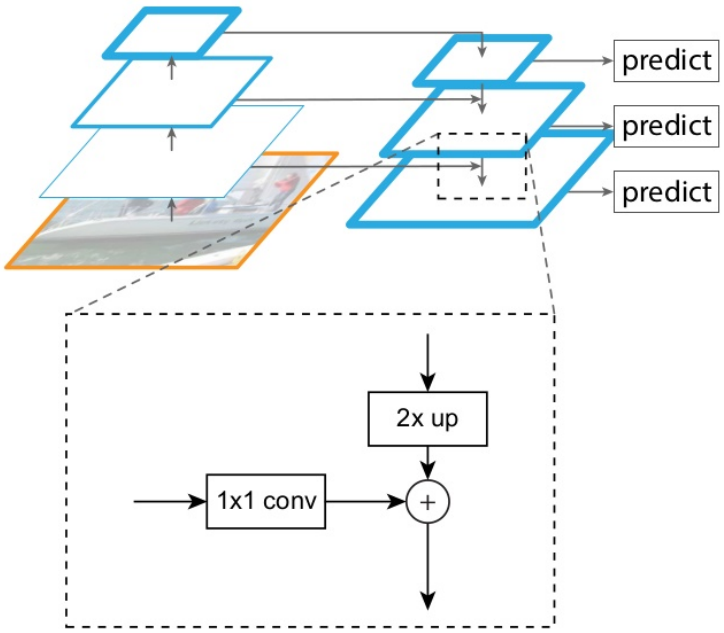
YOLO 将图像划分成7×7的网格，其中图像中的真实目标被其划分到目标中心所在的网格及其最接近的锚盒。对每个网格区域，网络需要预测：每个锚盒包含目标的概率(不包含目标时应为0，否则为锚盒和真实包围盒的IoU)、每个锚盒的4个坐标、该网格的类别概率分布。每个锚盒的类别概率分布等于每个锚盒包含目标的概率乘以该网格的类别概率分布。相比基于候选区域的方法，YOLO需要预测包含目标的概率的原因是，图像中大部分的区域不包含目标，而训练时只有目标存在时才对坐标和类别概率分布进行更新。YOLO的优点在于：(1). 基于候选区域的方法的感受野是图像中的局部区域，而YOLO可以利用整张图像的信息。(2). 有更好的泛化能力。YOLO的局限在于：(1). 不能很好处理网格中目标数超过预设固定值，或网格中有多个目标同时属于一个锚盒的情况。(2). 对小目标的检测能力不够好。(3). 对不常见长宽比的包围盒的检测能力不强。(4). 计算损失时没有考虑包围盒大小。大的包围盒中的小偏移和小的包围盒中的小偏移应有不同的影响。



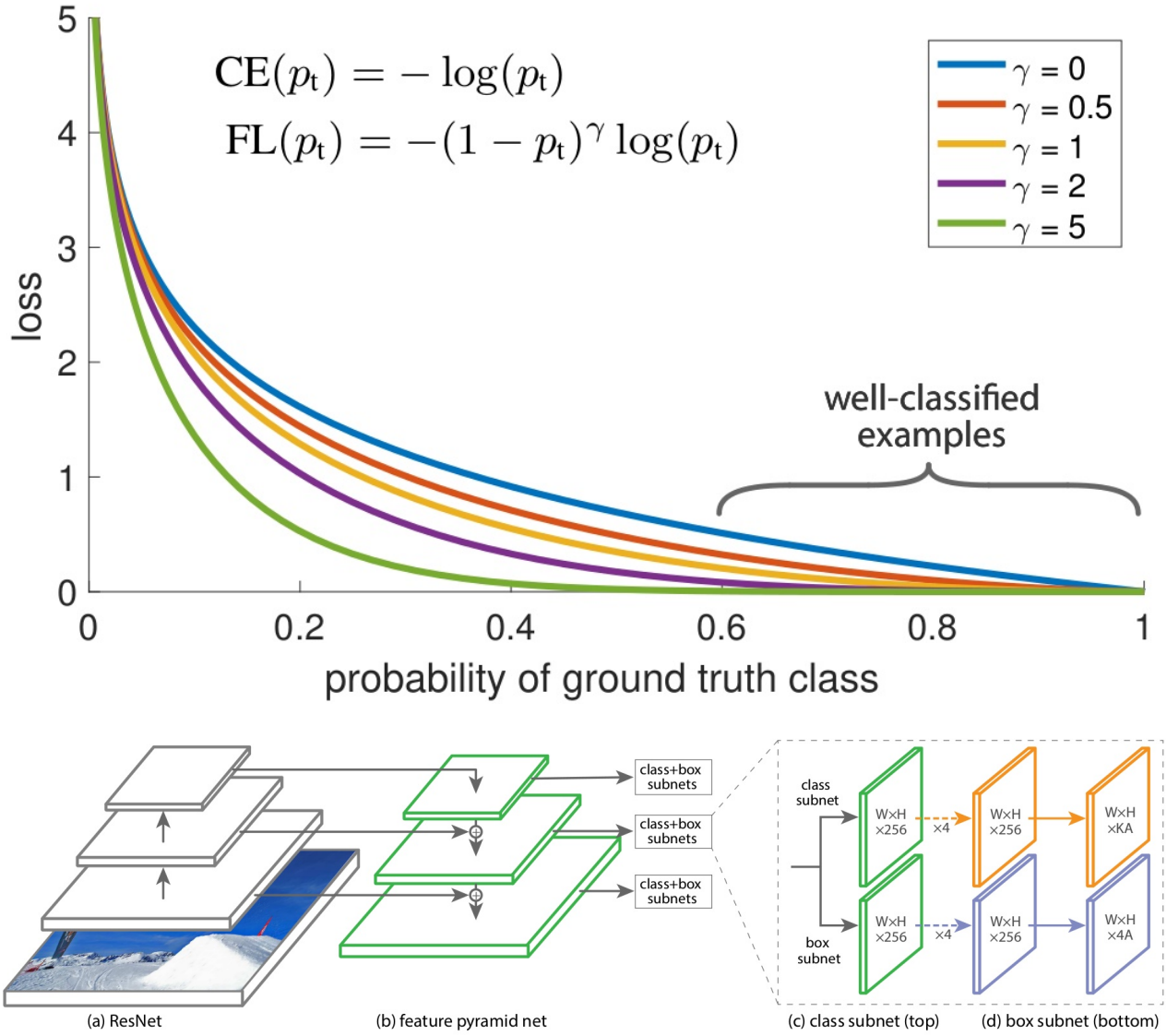
SSD 相比YOLO，SSD在卷积特征后加了若干卷积层以减小特征空间大小，并通过综合多层卷积层的检测结果以检测不同大小的目标。此外，类似于Faster R-CNN的RPN，SSD使用3×3卷积取代了YOLO中的全连接层，以对不同大小和长宽比的锚盒来进行分类/回归。SSD取得了比YOLO更快，接近Faster R-CNN的检测性能。后来有研究发现，相比其他方法，SSD受基础模型性能的影响相对较小。



FPN 之前的方法都是取高层卷积特征。但由于高层特征会损失一些细节信息，FPN融合多层特征，以综合高层、低分辨率、强语义信息和低层、高分辨率、弱语义信息来增强网络对小目标的处理能力。此外，和通常用多层融合的结果做预测的方法不同，FPN在不同层独立进行预测。FPN既可以与基于候选区域的方法结合，也可以与基于直接回归的方法结合。FPN在和Faster R-CNN结合后，在基本不增加原有模型计算量的情况下，大幅提高对小目标的检测性能。



RetinaNet RetinaNet认为，基于直接回归的方法性能通常不如基于候选区域方法的原因是，前者会面临极端的类别不平衡现象。基于候选区域的方法可以通过候选区域过滤掉大部分的背景区域，但基于直接回归的方法需要直接面对类别不平衡。因此，RetinaNet通过改进经典的交叉熵损失以降低对已经分的很好的样例的损失值，提出了焦点(focal)损失函数，以使模型训练时更加关注到困难的样例上。RetinaNet取得了接近基于直接回归方法的速度，和超过基于候选区域的方法的性能。



(4) 目标检测常用技巧

非最大抑制(non-max suppression, NMS) 目标检测可能会出现的一个问题是，模型会对同一目标做出多次预测，得到多个包围盒。NMS旨在保留最接近真实包围盒的那一个预测结果，而抑制其他的预测结果。NMS的做法是，首先，对每个类别，NMS先统计每个预测结果输出的属于该类别概率，并将预测结果按该概率由高至低排序。其次，NMS认为对应概率很小的预测结果并没有找到目标，所以将其抑制。然后，NMS在剩余的预测结果中，找到对应概率最大的预测结果，将其输出，并抑制和该包围盒有很大重叠(如IoU大于0.3)的其他包围盒。重复上一步，直到所有的预测结果均被处理。

在线困难样例挖掘(online hard example mining, OHEM) 目标检测的另一个问题是类别不平衡，图像中大部分的区域是不包含目标的，而只有小部分区域包含目标。此外，不同目标的检测难度也有很大差异，绝大部分的目标很容易被检测到，而有一小部分目标却十分困难。OHEM和Boosting的思路类似，其根据损失值将所有候选区域进行排序，并选择损失值最高的一部分候选区域进行优化，使网络更关注于图像中更困难的目标。此外，为了避免选到相互重叠很大的候选区域，OHEM对候选区域根据损失值进行NMS。

在对数空间回归 回归相比分类优化难度大了很多。ℓ_2 损失对异常值比较敏感，由于有平方，异常值会有大的损失值，同时会有很大的梯度，使训练时很容易发生梯度爆炸。而 ℓ_1 损失的

梯度不连续。在对数空间中，由于数值的动态范围小了很多，回归训练起来也会容易很多。此外，也有人用平滑的 ℓ_1 损失进行优化。预先将回归目标规范化也会有助于训练。

语义分割(semantic segmentation)

语义分割是目标检测更进阶的任务，目标检测只需要框出每个目标的包围盒，语义分割需要进一步判断图像中哪些像素属于哪个目标。

(1) 语义分割常用数据集

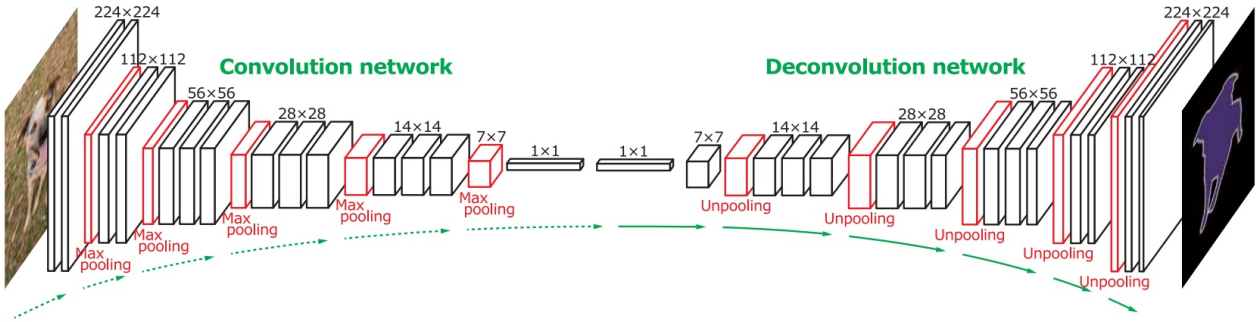
PASCAL VOC 2012 1.5k训练图像，1.5k验证图像，20个类别(包含背景)。

MS COCO COCO比VOC更困难。有83k训练图像，41k验证图像，80k测试图像，80个类别。

(2) 语义分割基本思路

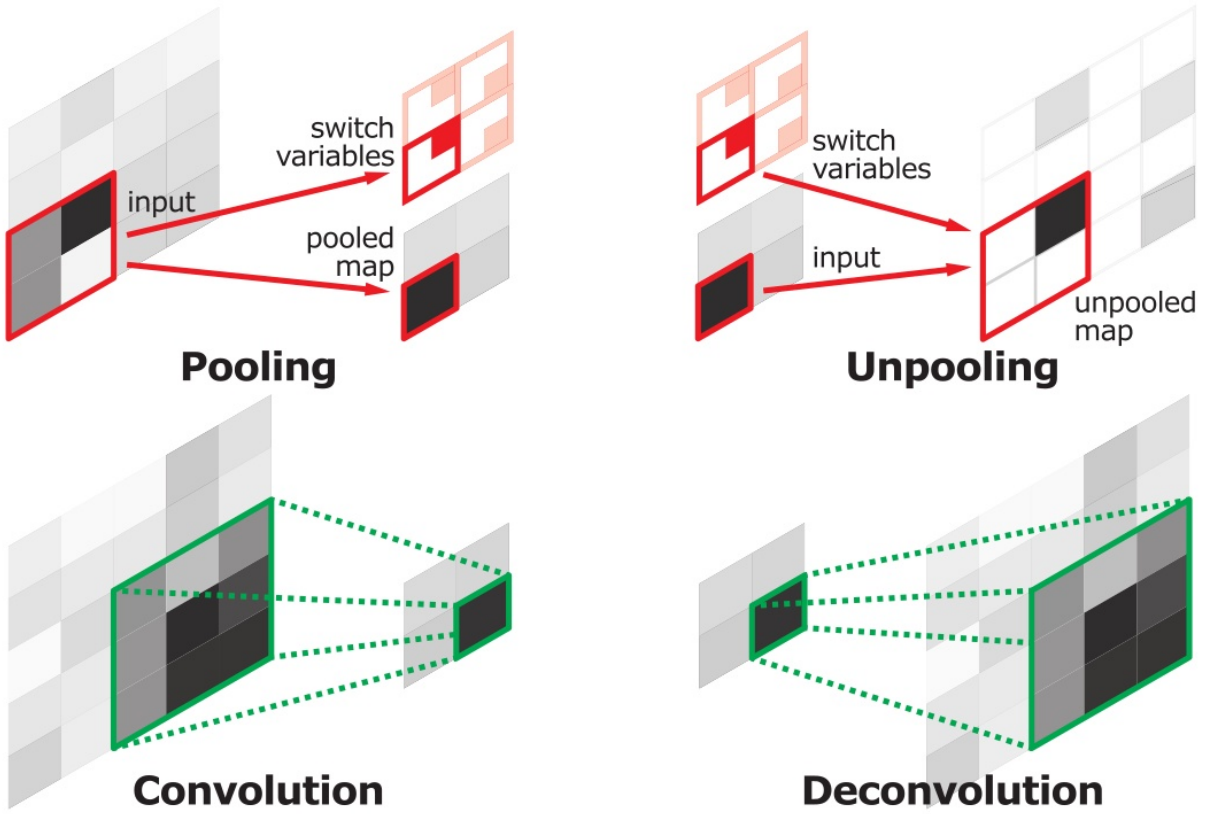
基本思路 逐像素进行图像分类。我们将整张图像输入网络，使输出的空间大小和输入一致，通道数等于类别数，分别代表了各空间位置属于各类别的概率，即可以逐像素地进行分类。

全卷积网络+反卷积网络 为使得输出具有三维结构，全卷积网络中没有全连接层，只有卷积层和汇合层。但是随着卷积和汇合的进行，图像通道数越来越大，而空间大小越来越小。要想使输出和输入有相同的空间大小，全卷积网络需要使用反卷积和反汇合来增大空间大小。



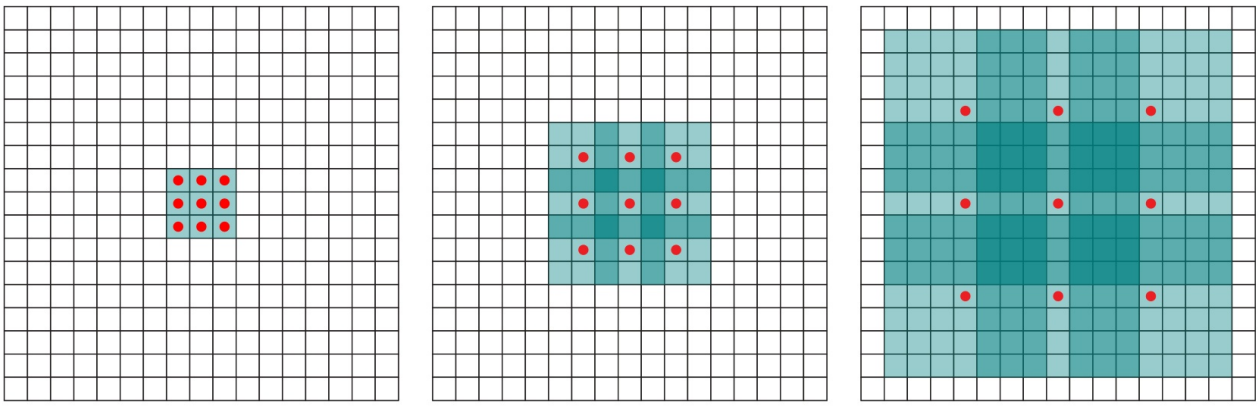
反卷积(deconvolution)/转置卷积(transpose convolution) 标准卷积的滤波器在输入图像中进行滑动，每次和输入图像局部区域点乘得到一个输出，而反卷积的滤波器在输出图像中进行滑动，每个由一个输入神经元乘以滤波器得到一个输出局部区域。反卷积的前向过程和卷积的反向过程完成的是相同的数学运算。和标准卷积的滤波器一样，反卷积的滤波器也是从数据中学到的。

反最大汇合(max-unpooling) 通常全卷积网络是对称的结构，在最大汇合时需要记下最大值所处局部区域位置，在对应反最大汇合时将对应位置输出置为输入，其余位置补零。反最大汇合可以弥补最大汇合时丢失的空间信息。反最大汇合的前向过程和最大汇合的反向过程完成的是相同的数学运算。



(3) 语义分割常用技巧

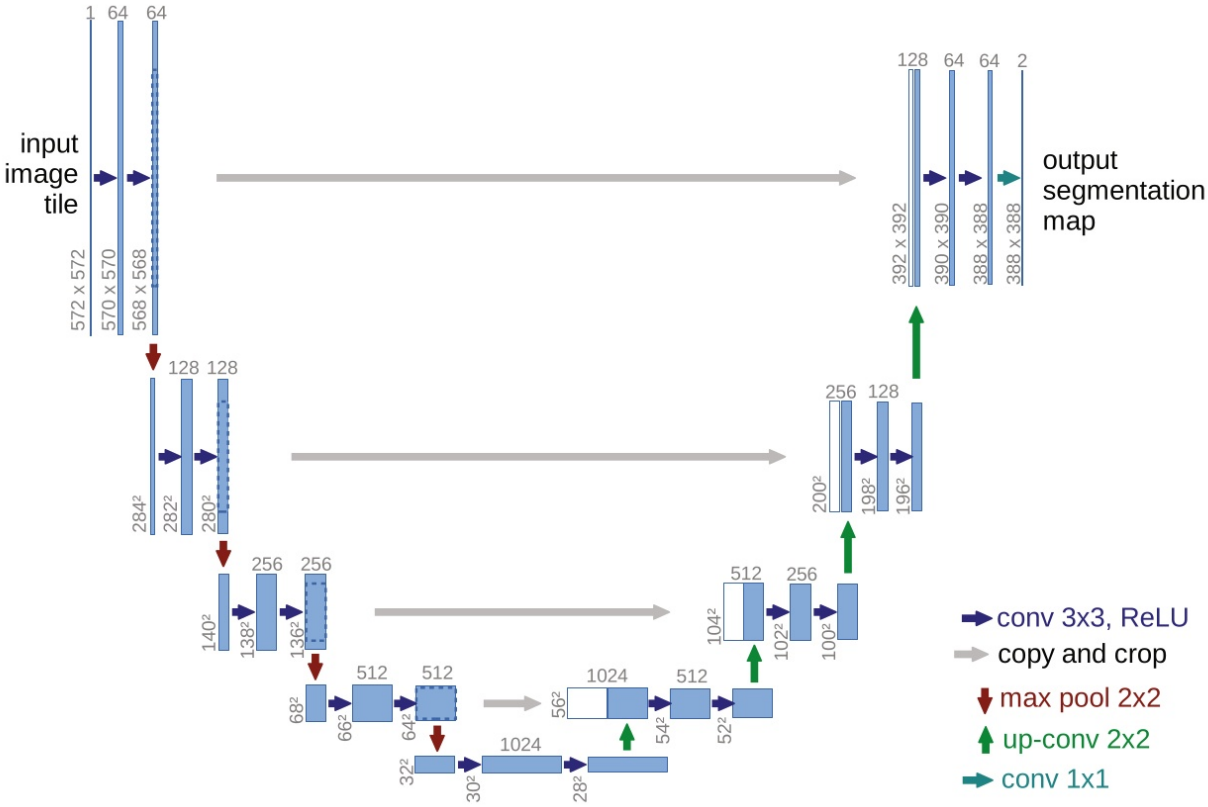
扩张卷积(dilated convolution) 经常用于分割任务以增大有效感受野的一个技巧。标准卷积操作中每个输出神经元对应的输入局部区域是连续的，而扩张卷积对应的输入局部区域在空间位置上不连续。扩张卷积向标准卷积运算中引入了一个新的超参数扩张量(dilation)，用于描述输入局部区域在空间位置上的间距。当扩张量为1时，扩张卷积退化为标准卷积。扩张卷积可以在参数量不变的情况下有效提高感受野。例如，当有多层3×3标准卷积堆叠时，第l 层卷积(l从1开始)的输出神经元的感受野为2l +1。与之相比，当有多层3×3扩张卷积堆叠，其中第l 层卷积的扩张量为2^{l-1}时，第l 层卷积的输出神经元的感受野为2^{l +1}-1。感受野越大，神经元能利用的相关信息越多。和经典计算机视觉手工特征相比，大的感受野是深度学习方法能取得优异性能的重要原因之一。



条件随机场(conditional random field, CRF) 条件随机场是一种概率图模型，常被用于微修全卷积网络的输出结果，使细节信息更好。其动机是距离相近的像素、或像素值相近的像素更可能属于相同的类别。此外，有研究工作用循环神经网络(recurrent neural networks)近似条

件随机场。条件随机场的另一弊端是会考虑两两像素之间的关系，这使其运行效率不高。

利用低层信息 综合利用低层结果可以弥补随着网络加深丢失的细节和边缘信息，利用方式可以是加和(如FCN)或沿通道方向拼接(如U-net)，后者效果通常会更好一些。

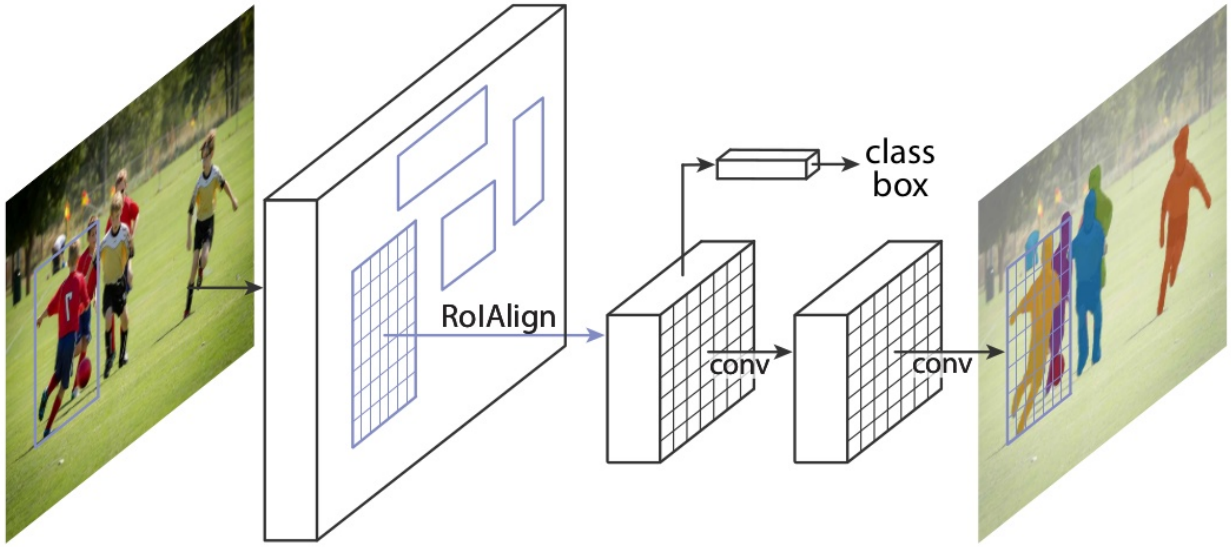


实例分割(instance segmentation)

语义分割不区分属于相同类别的不同实例。例如，当图像中有多只猫时，语义分割会将两只猫整体的所有像素预测为“猫”这个类别。与此不同的是，实例分割需要区分出哪些像素属于第一只猫、哪些像素属于第二只猫。

基本思路 目标检测+语义分割。先用目标检测方法将图像中的不同实例框出，再用语义分割方法在不同包围盒内进行逐像素标记。

Mask R-CNN 用FPN进行目标检测，并通过添加额外分支进行语义分割(额外分割分支和原检测分支不共享参数)，即Mask R-CNN有三个输出分支(分类、坐标回归、和分割)。此外，Mask R-CNN的其他改进有：(1). 改进了RoI汇合，通过双线性差值使候选区域和卷积特征的对齐不因量化而损失信息。(2). 在分割时，Mask R-CNN将判断类别和输出模板(mask)这两个任务解耦合，用sigmoid配合对率(logistic)损失函数对每个类别的模板单独处理，取得了比经典分割方法用softmax让所有类别一起竞争更好的效果。


class
box
RoIAlign
conv
conv