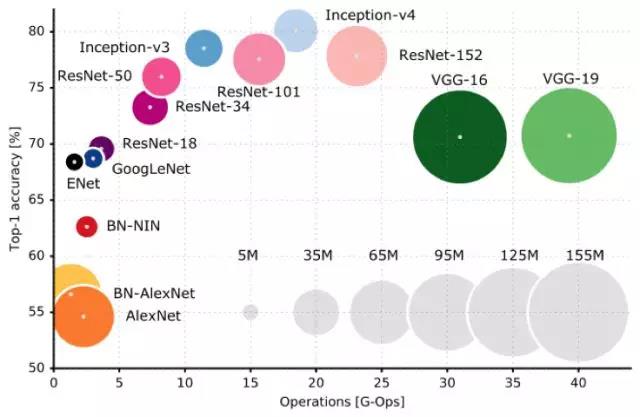
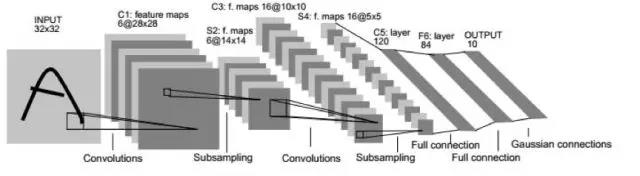
卷积神经网络

自2012年AlexNet网络出现之后，卷积网络得到了急速地发展。卷积神经网络在深度神经网络中应用广泛，在各个领域中都有应用。接下来将介绍一些主流的深度神经卷积网络。



LeNet

LeNet神经网络由Yan LeCun提出。LeNet主要用来进行手写体识别，并在美国的银行中投入了使用。LeNet的实现确立了CNN的结构，现在神经网络中的许多内容在LeNet的网络结构中都能看到，例如卷积层，Pooling层，ReLU层。LeNet的网络结构如下图所示：

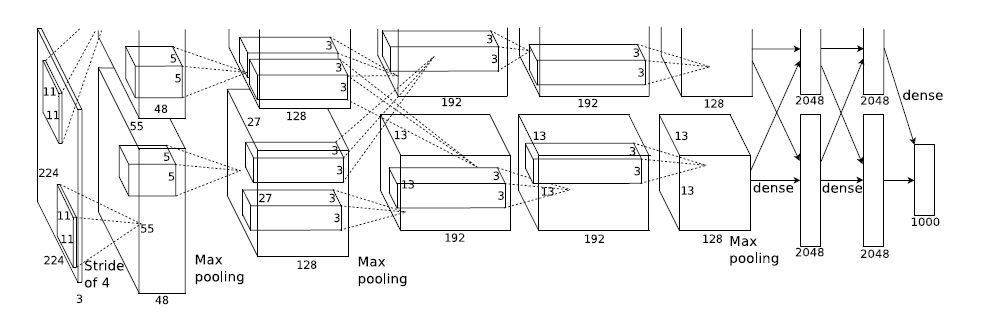


该网络的输入为32×32的图像，整个网络有2个卷积层，2个池化层，两个全连接层和一个输出层。输出层有10个神经元，代表输出的10个数字，即0到9。网络的激活函数使用的是tanh函数，损失函数为均方误差函数，训练方法使用的是随机梯度下降和反向传播方法。

虽然LeNet早在20世纪90年代就已经提出来了，但是由于当时缺乏大规模的训练数据、计算机计算能力较低，因此LeNet神经网络在处理复杂的问题是效果并不理想。

AlexNet

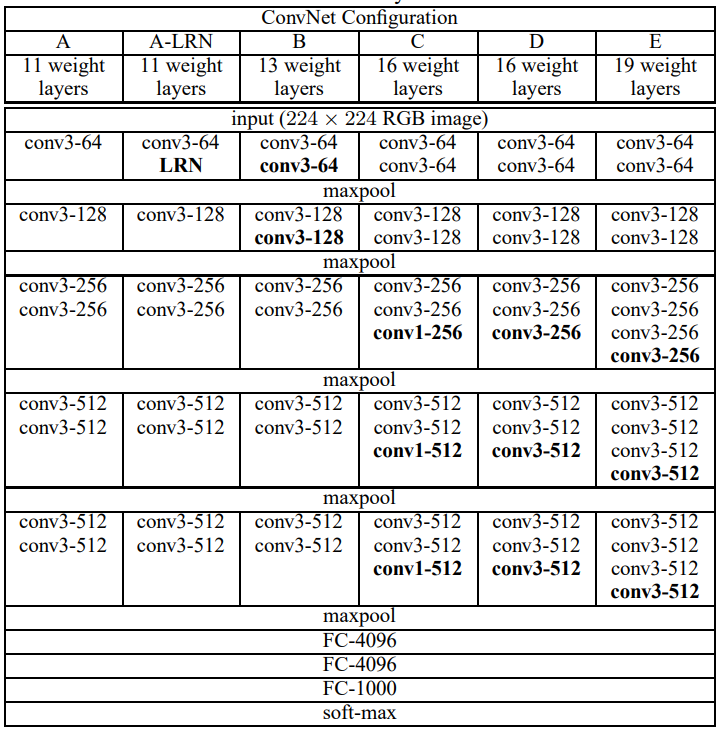
AlexNet是在2012年被提出的，该网络模型赢得了2012年ILSVRC（ImageNet Large Scale Visual Recognition Compitition）的冠军。



该模型结构和LeNet结构非常相似，与LeNet相比，AlexNet包含5个卷积层，两个全连接层和一个输出层，输出层中有1000个单元，代表ImageNet中的1000类。和LeNet传统卷积神经网络相比，AlexNet网络有一些改动，对原始数据做了数据增强，添加Dropout层用于防止过拟合，使用新的激活函数Relu，使用局部响应归一化策略，同时为了加快网络的训练时间，在多个GPU上进行训练。

VGG

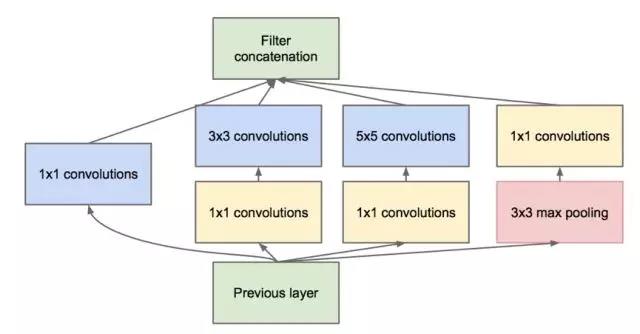
VGG网络由牛津大学视觉组2014年提出，该网络获得了ILSVRC 2014比赛分类任务的第二名和检测任务的第一名。VGGNet网络结构非常简单，整个网络都是使用同样大小的卷积核尺寸和池化尺寸。VGGNet探索了卷积神经网络的深度与其性能之间的关系，通过反复堆叠3×3的卷积层和2×2的最大池化层来构建网络。



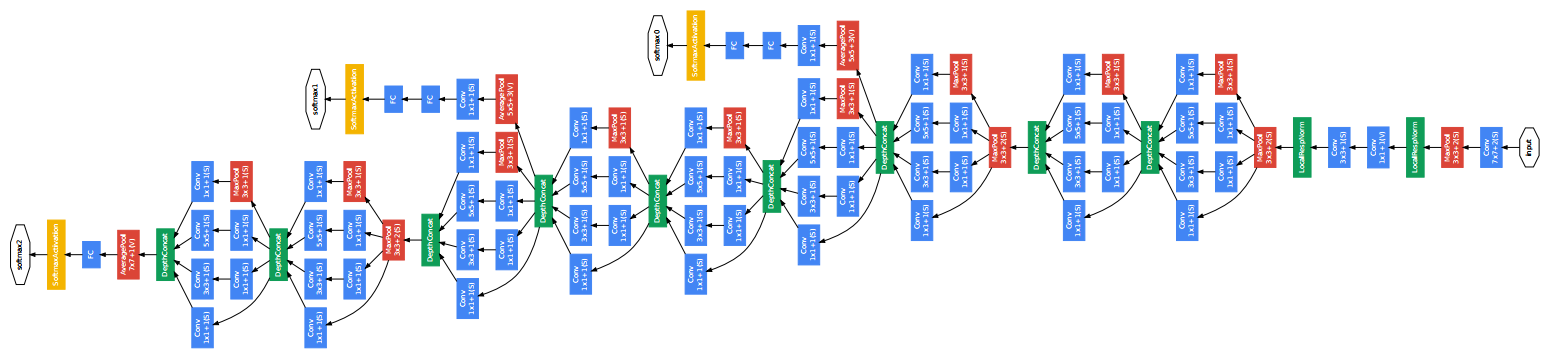
VGG网络和AlexNet相比，去掉了LRN层，采用多个较小的卷积来代替较大尺寸的卷积层。

GoogLeNet

相对于AlexNet，VGG的想法是从深度方向对网络结构进行改进，而GoogLeNet则是从宽度方向对网络结构进行改进。GoogLeNet也是在2014年被提出，在ILSVRC2014中取得分类任务第一名，其主要创新是Inception机制，即对图像进行多尺度处理。这种机制带来的好处是大幅度减小模型的参数数量。Inception的模块结构图如下所示：

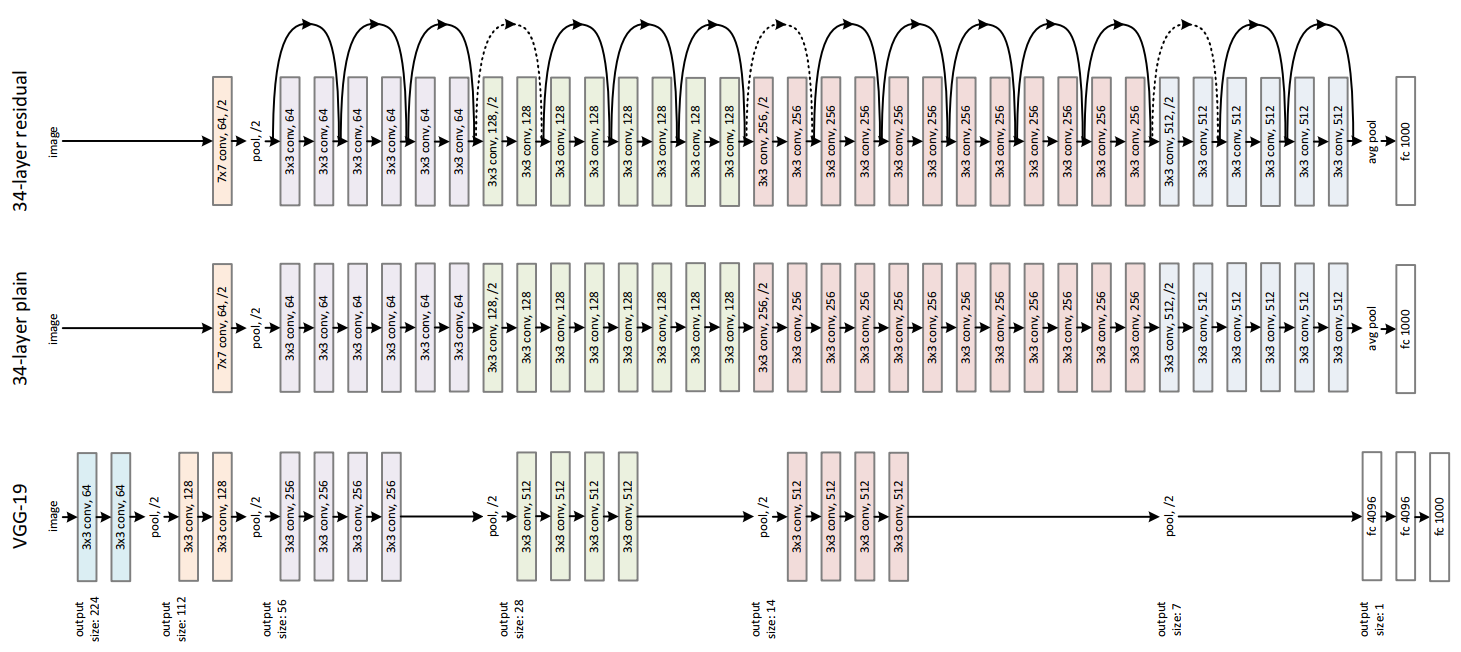


该模块由3组并行的卷积层和一个最大池化层组成，它们对输入图像并行进行处理，然后将输出进行拼接。其中1×1的卷积层主要用于降维，从而减少运算量。GoogLeNet网络的深度达到了22层，其结构如下图所示：



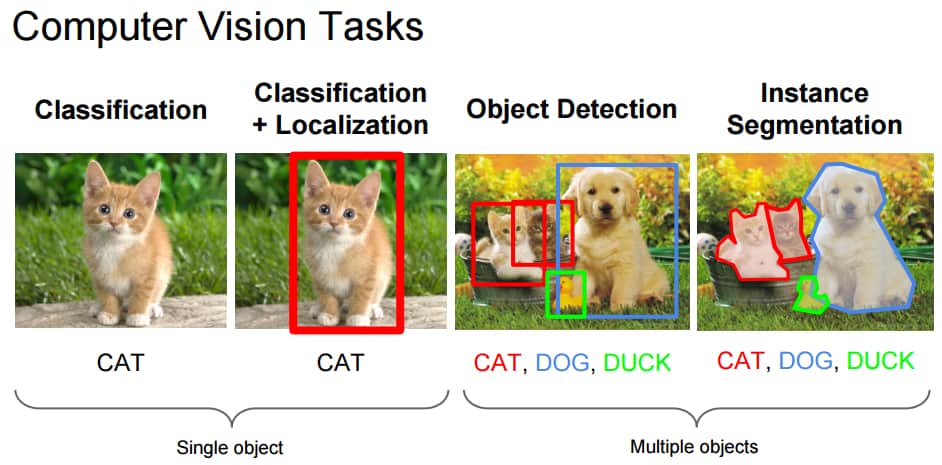
ResNet

在神经网络结构中，增加网络的层数会提高网络的性能，但增加到一定的程度之后，随着层次的增加，神经网络的训练误差和测试误差会增大，这个问题被称为退化。因此，非常深的网络是很难训练的。残差网络用跨层连接拟合残差项的手段来解决了极深度条件下深度卷积网络性能退化的问题。在ImageNet数据集上，作者使用152层的残差网络来进行实验，与VGG网络相比，残差网络深度差不多是VGG网络深度的8倍，但是复杂度却比VGG网络低。该网络模型获得了ILSVRC 2015分类任务、检测任务以及COCO检测和分割的第一名。残差网络的结构图如下所示:



目标检测发展

计算机视觉是指研究计算机视觉的能力的学科，是的机器能够像人一样观察并分析周围的环境。计算机视觉的研究目标通常是静态图像或者视频，计算机视觉被定义为对单个图像或者序列图像进行自动提取、分析和理解。

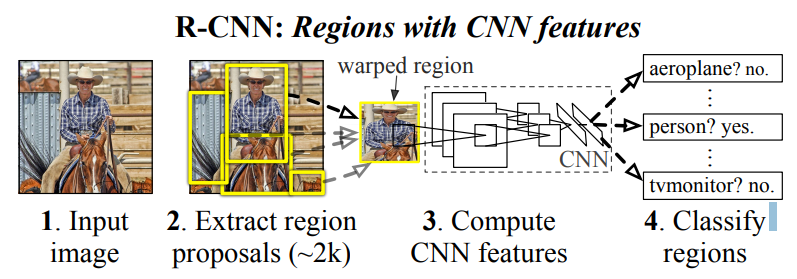


计算机视觉的任务主要包括目标分类定位、目标检测、目标跟踪以及目标分割，如下图所示。图像的分类定位任务是指给一张图像指定一个标签，例如：“猫”，然后在图像中定位出所指定的标签的目标，定位的表达形式通常是用目标边框的坐标。目标检测是指输出图像中每一个目标的边框坐标和类别标签，目标检测和目标分类定位的主要区别就是目标检测应用于多个目标上。目标跟踪是指在视频中对感兴趣的指定目标进行检测和追踪。目标分割是指将图像分为可以被标定和分类的像素群。

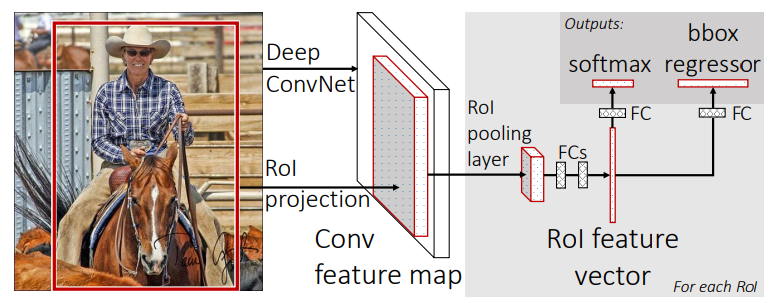
本论文主要研究的是目标检测，因此接下来将对近些年基于深度学习的目标检测的发展状况进行描述。基于深度学习的目标检测方法主要分为两大类别，一类是two-stage的目标检测方法，该方法第一阶段先从图像中提取目标候选框特征，然后再对候选框特征进行分类回归实现目标检测，该类方法主要包括RCNN，Fast RCNN，Faster RCNN，RFCN等；另一类是one-stage的目标检测方法，该类检测方法不会提取目标候选框而直接对图像特征图进行分类回归实现目标检测，该检测方法主要包括YOLO，SSD以及它们的改进方法。接下来将对每一类的方法进行详细描述。

Two-stage目标检测方法

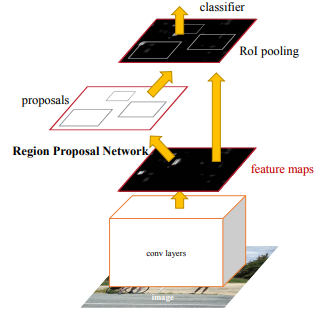
基于深度学习的二阶的目标检测方法可以从RCNN（Region-based Convolutional Neural Network）说起。RCNN可以分为四个步骤：首先输入一张图像，然后使用选择搜索在图象上提取2000个左右的候选框，并将这些框调整到固定的输入尺寸，接着将固定尺寸大小的目标候选框送入卷基层网络中提取特征，最后将这些特征送入到一个线性SVM分类器中进行分类，同时也送入到一个回归器中得到目标框坐标。



RCNN的主要缺点就是慢，在原图像上提取2000个候选框，然后再对这些候选框进行卷积获取特征，而这些候选框具有大量的重叠部分，因此在提取特征是就有很多重复的计算，使得运算量巨大，运算速度慢。因此之后提出了一种对RCNN的改进方法，叫做Fast RCNN，该方法只对图像做一次卷积，然后再特征图上提取候选框，这样能够减少计算量。Fast RCNN首先在整个图像上做卷积提取特征图，然后在特征图上利用选择性搜索方法选取2000个感兴趣的区域，被称作ROIs（Region Of Interest）,之后再将提取的ROIs送入到全连接层中，最后将全连接层的输出分别送入一个softmax激活函数的全连接层进行分类和一个回归器进行回归以获得目标框的坐标。Fast RCNN使用一个softmax层来输出类别的概率，相比于RCNN训练多个SVM来对目标进行分类，现在只需要训练一个网络就可以，这样训练将会更简单一些，同时运算时间也会减少。



Fast RCNN相对于RCNN来说运算速度有所提高，但是运算也是相对慢的，主要是选择性搜索限制的，因此，研究者在Fast RCNN的基础上又提出了改进，提出了Faster RCNN方法，去除了选择性搜索模块，添加了RPN网络，使得整个网络可以端到端进行训练，因此运算时间将大大减少，同时训练也变得更加容易。Faster RCNN具体的算法将在之后进行详细说明。



Faster RCNN在检测速度和效果上可以说已经很好了，但是由于该网络中全连接层神经元数量比较大，导致全连接层参数较多，同时使得运算量大，运算时间长，因此，有研究者针对这个问题提出了新的目标检测方法，称作R-FCN。该方法的主要思想是使得更多的层能够共享参数以便减少运算时间。该方法使用ResNet101作为骨架，为了使得能够极大地共享参数，该方法去掉了全局平均层和分类用的全连接层，并添加了一个1×1的卷基层来降维，使得整个网络仅包含卷基层。但是这其中还面临着一个问题，也就是图像分类的平移不变性和图像目标定位的平移变换性相互矛盾，为了解决这个这个问题，该方法使用了位置敏感分数特征图。

