
计算机视觉五大技术：图像分类、对象检测、目标跟踪、语义分割和实例分割

作者：王柯凝译 来源：AI科技大本营

本文原地址：<https://www.iikx.com/news/progress/85.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

详解计算机视觉五大技术：图像分类、对象检测、目标跟踪、语义分割和实例分割



【AI科技大本营导读】目前，计算机视觉是深度学习领域最热门的研究领域之一。计算机视觉实际上是一个跨领域的交叉学科，包括计算机科学(图形、算法、理论、系统、体系结构)，数学(信息检索、机器学习)，工程学(机器人、语音、自然语言处理、图像处理)，物理学(光学)，生物学(神经科学)和心理学(认知科学)等等。许多科学家认为，计算机视觉为人工智能的发展开拓了道路。

那么什么是计算机视觉呢? 这里给出了几个比较严谨的定义：

“对图像中的客观对象构建明确而有意义的描述”(Ballard&Brown, 1982)

“ 从一个或多个数字图像中计算三维世界的特性 ” (Trucco&Verri, 1998)

“ 基于感知图像做出对客观对象和场景有用的决策 ” (Sackman&Shapiro, 2001)

■ 为什么要学习计算机视觉?

一个显而易见的答案就是，这个研究领域已经衍生出了一大批快速成长的、有实际作用的应用，例如：

人脸识别：Snapchat 和 Facebook 使用人脸检测算法来识别人脸。图像检索：Google Images 使用基于内容的查询来搜索相关图片，算法分析查询图像中的内容并根据最佳匹配内容返回结果。游戏和控制：使用立体视觉较为成功的游戏应用产品是：微软 Kinect。监测：用于监测可疑行为的监视摄像头遍布于各大公共场所中。生物识别技术：指纹、虹膜和人脸匹配仍然是生物识别领域的一些常用方法。智能汽车：计算机视觉仍然是检测交通标志、灯光和其他视觉特征的主要信息来源。

视觉识别是计算机视觉的关键组成部分，如图像分类、定位和检测。神经网络和深度学习的最新进展极大地推动了这些最先进的视觉识别系统的发展。在本文中，我将分享 5 种主要的计算机视觉技术，并介绍几种基于计算机视觉技术的深度学习模型与应用。

■ 1、图像分类



给定一组各自被标记为单一类别的图像，我们对一组新的测试图像的类别进行预测，并测量预测的准确性结果，这就是图像分类问题。图像分类问题需要面临以下几个挑战 🐼 🐼 🐼：

视点变化，尺度变化，类内变化，图像变形，图像遮挡，照明条件和背景杂斑

我们怎样来编写一个图像分类算法呢?



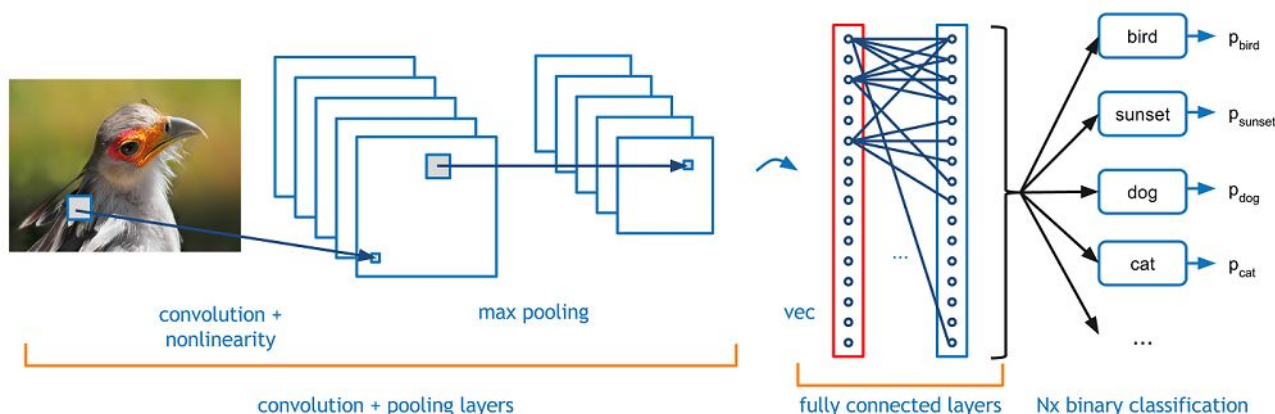
计算机视觉研究人员提出了一种基于数据驱动的方法。

该算法并不是直接在代码中指定每个感兴趣的图像类别，而是为计算机每个图像类别都提供许多示例，然后设计一个学习算法，查看这些示例并学习每个类别的视觉外观。也就是说，首先积累一个带有标记图像的训练集，然后将其输入到计算机中，由计算机来处理这些数据。

因此，可以按照下面的步骤来分解：

输入是由 N 个图像组成的训练集，共有 K 个类别，每个图像都被标记为其中一个类别。然后，使用该训练集训练一个分类器，来学习每个类别的外部特征。最后，预测一组新图像的类标签，评估分类器的性能，我们用分类器预测的类别标签与其真实的类别标签进行比较。

目前较为流行的图像分类架构是卷积神经网络(CNN)——将图像送入网络，然后网络对图像数据进行分类。卷积神经网络从输入“扫描仪”开始，该输入“扫描仪”也不会一次性解析所有的训练数据。比如输入一个大小为 100×100 的图像，你也不需要有一个有 10,000 个节点的网络层。相反，你只需要创建一个大小为 10×10 的扫描输入层，扫描图像的前 10×10 个像素。然后，扫描仪向右移动一个像素，再扫描下一个 10×10 的像素，这就是滑动窗口。

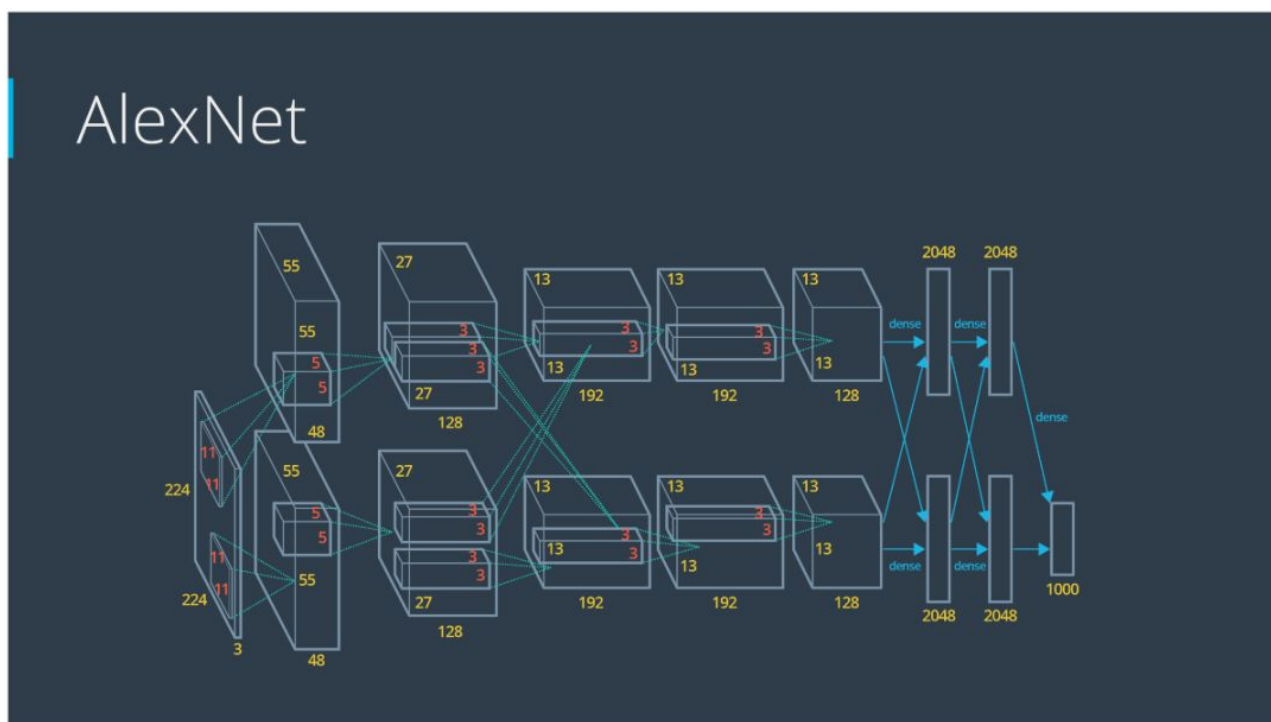


输入数据被送入卷积层，而不是普通层。每个节点只需要处理离自己最近的邻近节点，卷积层也随着扫描的深入而趋于收缩。除了卷积层之外，通常还会有池化层。池化是过滤细节的一种方法，常见的池化技术是最大池化，它用大小为 2×2 的矩阵传递拥有最多特定属性的像素。

现在，大部分图像分类技术都是在 ImageNet 数据集上训练的，ImageNet 数据集中包含了约 120 万张高分辨率训练图像。测试图像没有初始注释(即没有分割或标签)，并且算法必须产生标签来指定图像中存在哪些对象。

现存的很多计算机视觉算法，都是被来自牛津、INRIA 和 XRCE 等顶级的计算机视觉团队在 ImageNet 数据集上实现的。通常来说，计算机视觉系统使用复杂的多级管道，并且，早期阶段的算法都是通过优化几个参数来手动微调的。

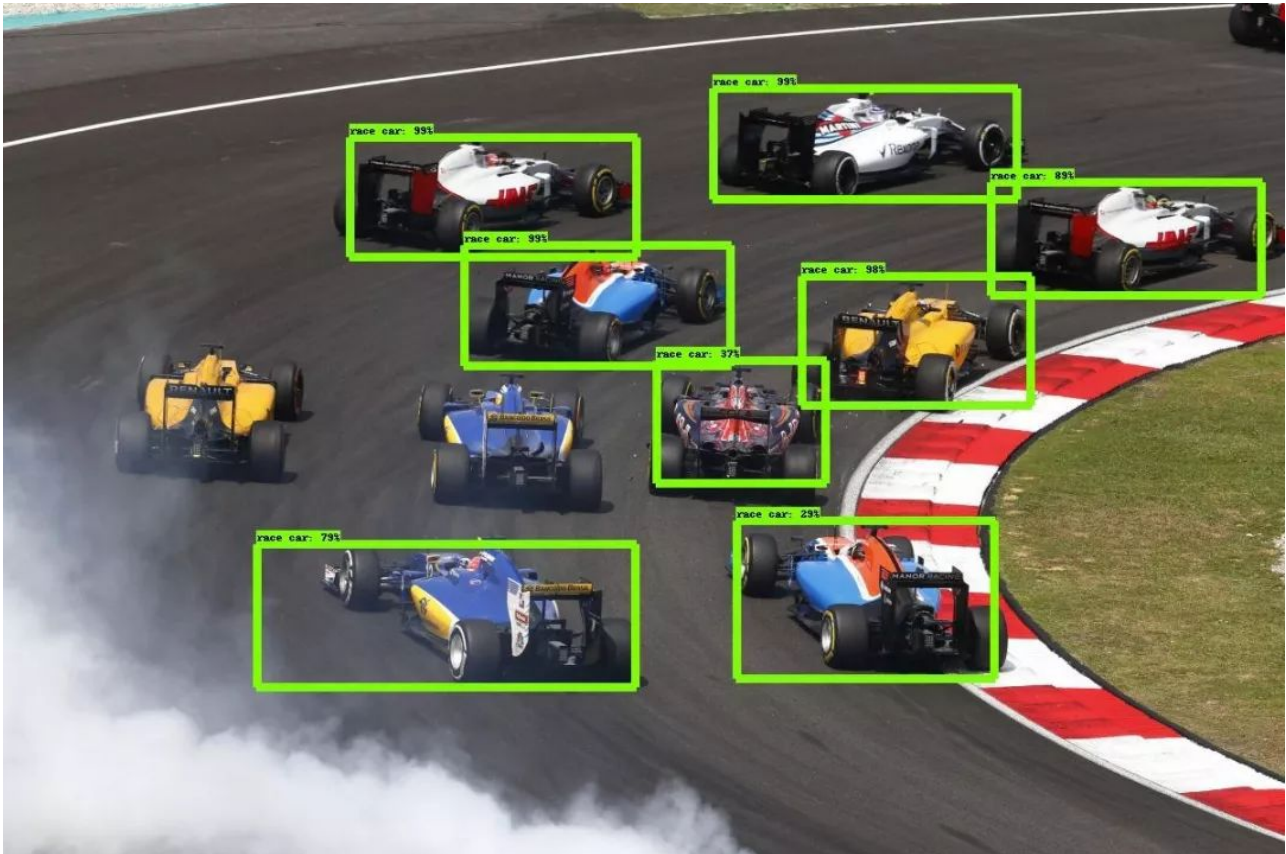
第一届 ImageNet 竞赛的获奖者是 Alex Krizhevsky(NIPS 2012)，他在 Yann LeCun 开创的神经网络类型基础上，设计了一个深度卷积神经网络。该网络架构除了一些最大池化层外，还包含 7 个隐藏层，前几层是卷积层，最后两层是全连接层。在每个隐藏层内，激活函数为线性的，要比逻辑单元的训练速度更快、性能更好。除此之外，当附近的单元有更强的活动时，它还使用竞争性标准化来压制隐藏活动，这有助于强度的变化。



就硬件要求而言，Alex 在 2 个 Nvidia GTX 580 GPU (速度超过 1000 个快速的小内核)上实现了非常高效的卷积网络。GPU 非常适合矩阵间的乘法且有非常高的内存带宽。这使他能在一周内完成训练，并在测试时快速的从 10 个块中组合出结果。如果我们能够以足够快的速度传输状态，就可以将网络分布在多个内核上。

随着内核越来越便宜，数据集越来越大，大型神经网络的速度要比老式计算机视觉系统更快。在这之后，已经有很多种使用卷积神经网络作为核心，并取得优秀成果模型，如 ZFNet(2013)，GoogLeNet(2014)，VGGNet(2014)，RESNET(2015)，DenseNet(2016)等。

■ 2、对象检测



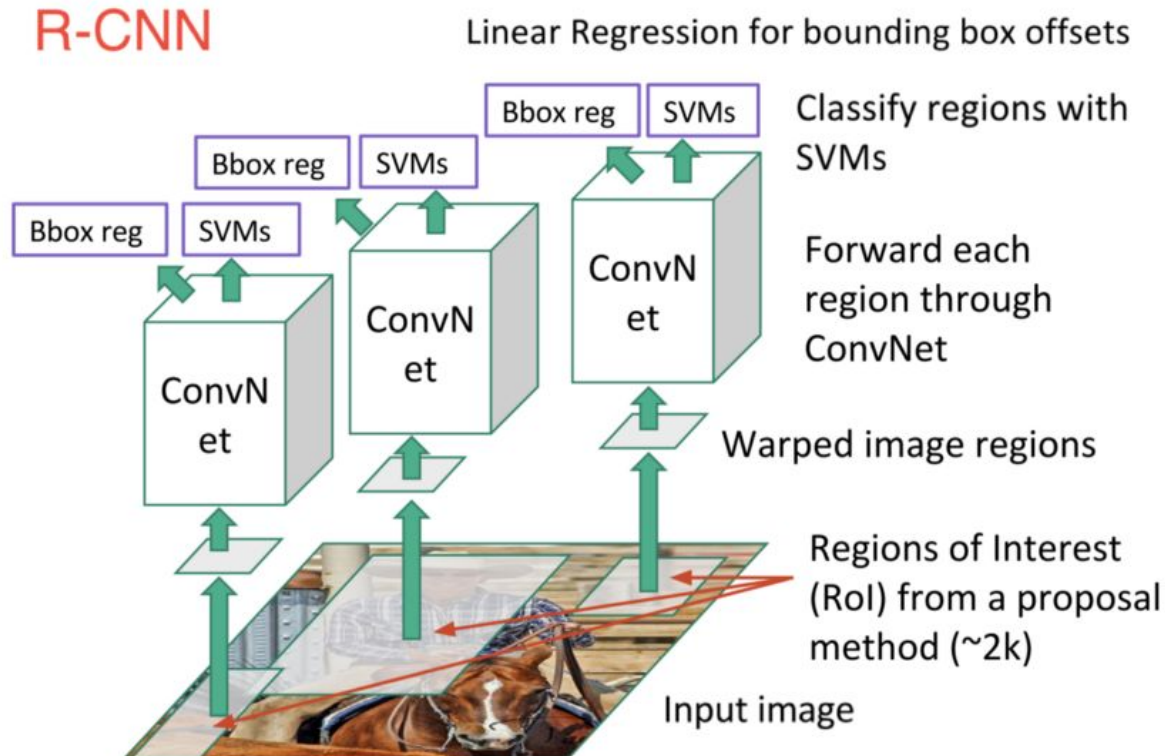
识别图像中的对象这一任务，通常会涉及到为各个对象输出边界框和标签。这不同于分类/定位任务——对很多对象进行分类和定位，而不仅仅是对个主体对象进行分类和定位。在对象检测中，你只有 2 个对象分类类别，即对象边界框和非对象边界框。例如，在汽车检测中，你必须使用边界框检测所给定图像中的所有汽车。

如果使用图像分类和定位图像这样的滑动窗口技术，我们则需要将卷积神经网络应用于图像上的很多不同物体上。由于卷积神经网络会将图像中的每个物体识别为对象或背景，因此我们需要在大量的位置和规模上使用卷积神经网络，但是这需要很大的计算量！

为了解决这一问题，神经网络研究人员建议使用区域(region)这一概念，这样我们会找到可能包含对象的“斑点”图像区域，这样运行速度就会大大提高。第一种模型是基于区域的卷积神经网络(R-CNN)，其算法原理如下：

在 R-CNN 中，首先使用选择性搜索算法扫描输入图像，寻找其中的可能对象，从而生成大约 2,000 个区域建议;然后，在这些区域建议上运行一个卷积神经网络;最后，将每个卷积神经网络的输出传给支持向量机(SVM)，使用一个线性回归收紧对象的边界框。

R-CNN

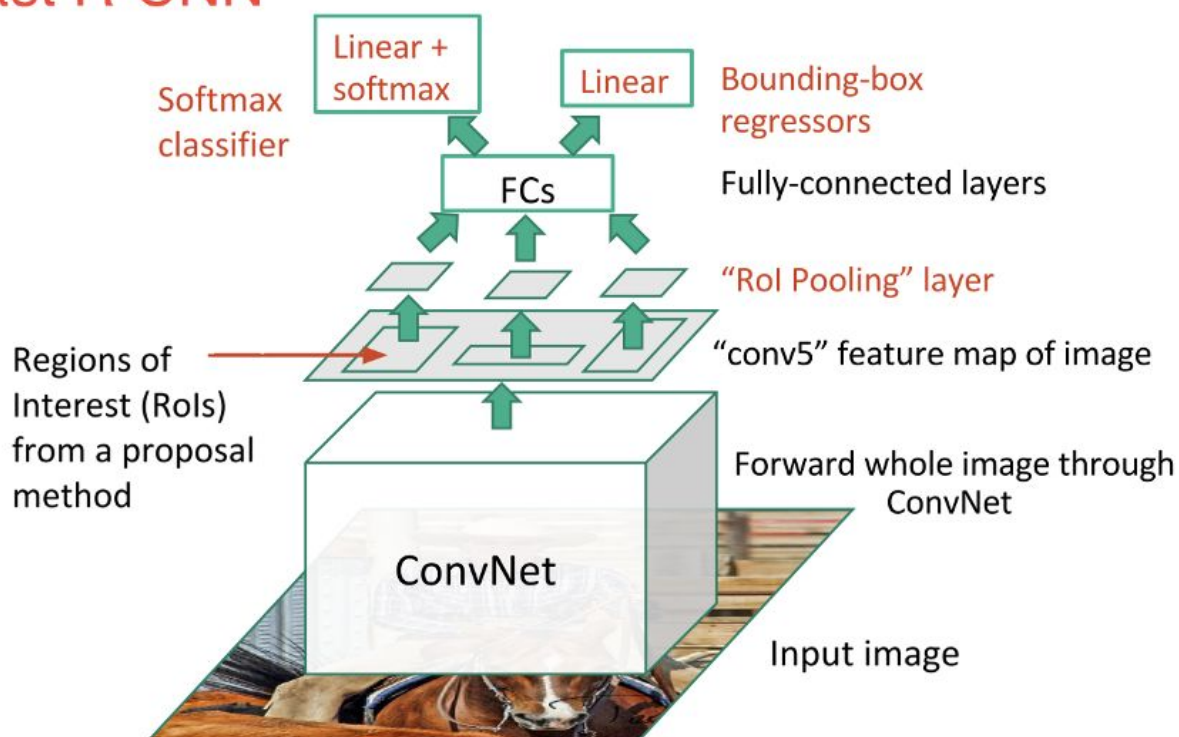


实质上，我们将对象检测转换为一个图像分类问题。但是也存在这些问题：训练速度慢，需要大量的磁盘空间，推理速度也很慢。

R-CNN 的第一个升级版本是 Fast R-CNN，通过使用了 2 次增强，大大提了检测速度：

在建议区域之前进行特征提取，因此在整幅图像上只能运行一次卷积神经网络;用一个 softmax 层代替支持向量机，对用于预测的神经网络进行扩展，而不是创建一个新的模型。

Fast R-CNN



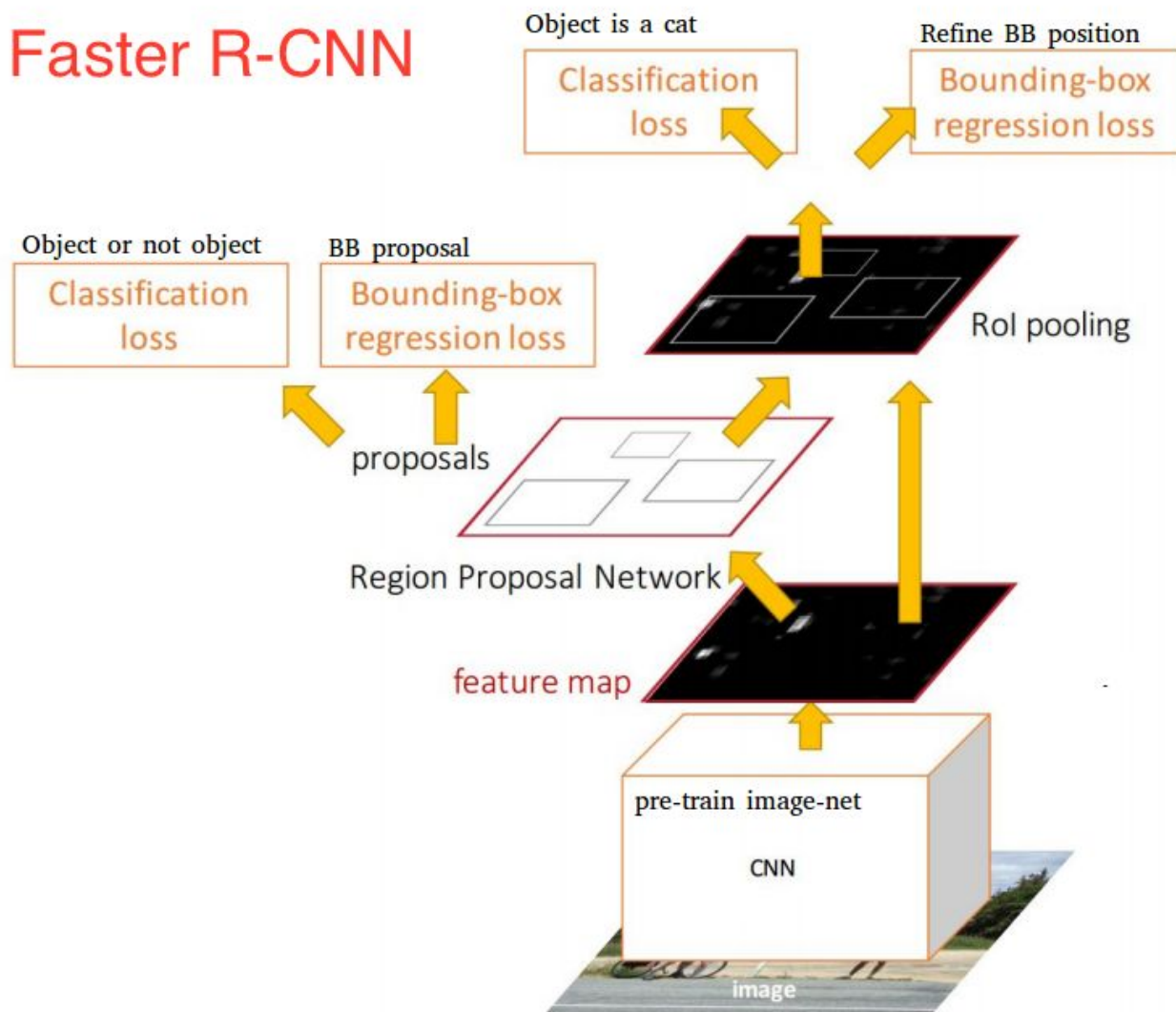
Fast R-CNN 的运行速度要比 R-CNN 快的多，因为在一幅图像上它只能训练一个 CNN。但是，择性搜索算法生成区域提议仍然要花费大量时间。

Faster R-CNN 是基于深度学习对象检测的一个典型案例。

该算法用一个快速神经网络代替了运算速度很慢的选择性搜索算法：通过插入区域提议网络(RPN)，来预测来自特征的建议。RPN 决定查看“哪里”，这样可以减少整个推理过程的计算量。

RPN 快速且高效地扫描每一个位置，来评估在给定的区域内是否需要作进一步处理，其实现方式如下：通过输出 k 个边界框建议，每个边界框建议都有 2 个值——代表每个位置包含目标对象和不包含目标对象的概率。

Faster R-CNN



一旦我们有了区域建议，就直接将它们送入 Fast R-CNN。并且，我们还添加了一个池化层、一些全连接层、一个 softmax 分类层以及一个边界框回归器。

总之，Faster R-CNN 的速度和准确度更高。值得注意的是，虽然以后的模型在提高检测速度方面做了很多工作，但很少有模型能够大幅度的超越 Faster R-CNN。换句话说，Faster R-CNN 可能不是最简单或最快速的目标检测方法，但仍然是性能最好的方法之一。

近年来，主要的目标检测算法已经转向更快、更高效的检测系统。这种趋势在 You Only Look Once(YOLO), Single Shot MultiBox Detector(SSD)和基于区域的全卷积网络(R-FCN)算法中尤为明显，这三种算法转向在整个图像上共享计算。因此，这三种算法和上述的3种造价较高的R-CNN 技术有所不同。

■ 3、目标跟踪



目标跟踪，是指在特定场景跟踪某一个或多个特定感兴趣对象的过程。传统的应用就是视频和真实世界的交互，在检测到初始对象之后进行观察。现在，目标跟踪在无人驾驶领域也很重要，例如 Uber 和特斯拉等公司的无人驾驶。

根据观察模型，目标跟踪算法可分成 2 类：生成算法和判别算法。

生成算法使用生成模型来描述外观特征，并将重建误差最小化来搜索目标，如主成分分析算法(PCA);判别算法用来区分物体和背景，其性能更稳健，并逐渐成为跟踪对象的主要手段(判别算法也称为 Tracking-by-Detection，深度学习也属于这一范畴)。

为了通过检测实现跟踪，我们检测所有帧的候选对象，并使用深度学习从候选对象中识别想要的对象。有两种可以使用的基本网络模型：堆叠自动编码器(SAE)和卷积神经网络(CNN)。

目前，最流行的使用 SAE 进行目标跟踪的网络是 Deep Learning Tracker(DLT)，它使用了离线预训练和在线微调。其过程如下：

离线无监督预训练使用大规模自然图像数据集获得通用的目标对象表示，对堆叠去噪自动编码器进行预训练。堆叠去噪自动编码器在输入图像中添加噪声并重构原始图像，可以获得更强大的特征表述能力。

将预训练网络的编码部分与分类器合并得到分类网络，然后使用从初始帧中获得的正负样本对网络进行微调，来区分当前的对象和背景。DLT 使用粒子滤波作为意向模型(motion model)，生成当前帧的候选块。

分类网络输出这些块的概率值，即分类的置信度，然后选择置信度最高的块作为对象。

在模型更新中，DLT 使用有限阈值。



鉴于 CNN

在图像分类和目标检测方面的优势，它已成为计算机视觉和视觉跟踪的主流深度模型。一般来说，大规模的卷积神经网络既可以作为分类器和跟踪器来训练。具有代表性的基于卷积神经网络的跟踪算法有全卷积网络跟踪器(FCNT)和多域卷积神经网络(MD Net)。

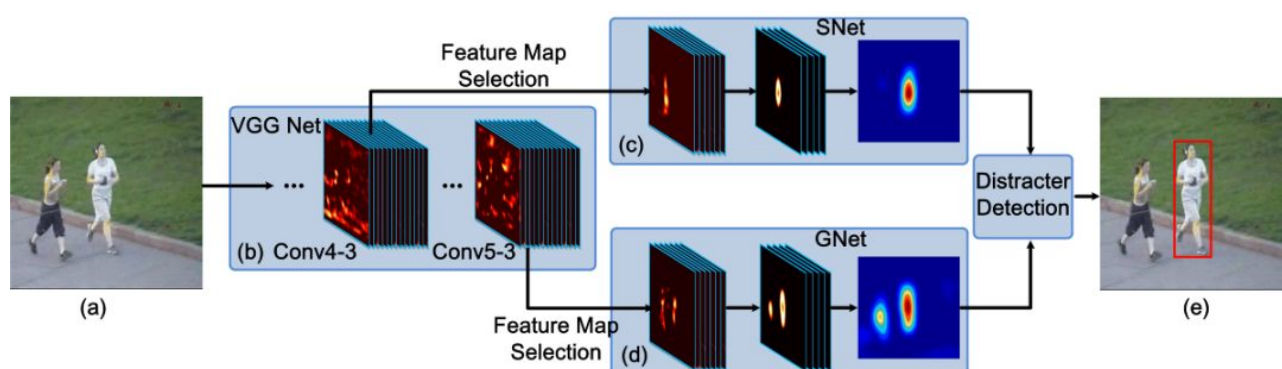
FCNT 充分分析并利用了 VGG 模型中的特征映射，这是一种预先训练好的 ImageNet 数据集，并有如下效果：

卷积神经网络特征映射可用于定位和跟踪。对于从背景中区分特定对象这一任务来说，很多卷积

神经网络特征映射是噪音或不相关的。较高层捕获对象类别的语义概念，而较低层编码更多的具有区性的特征，来捕获类别内的变形。

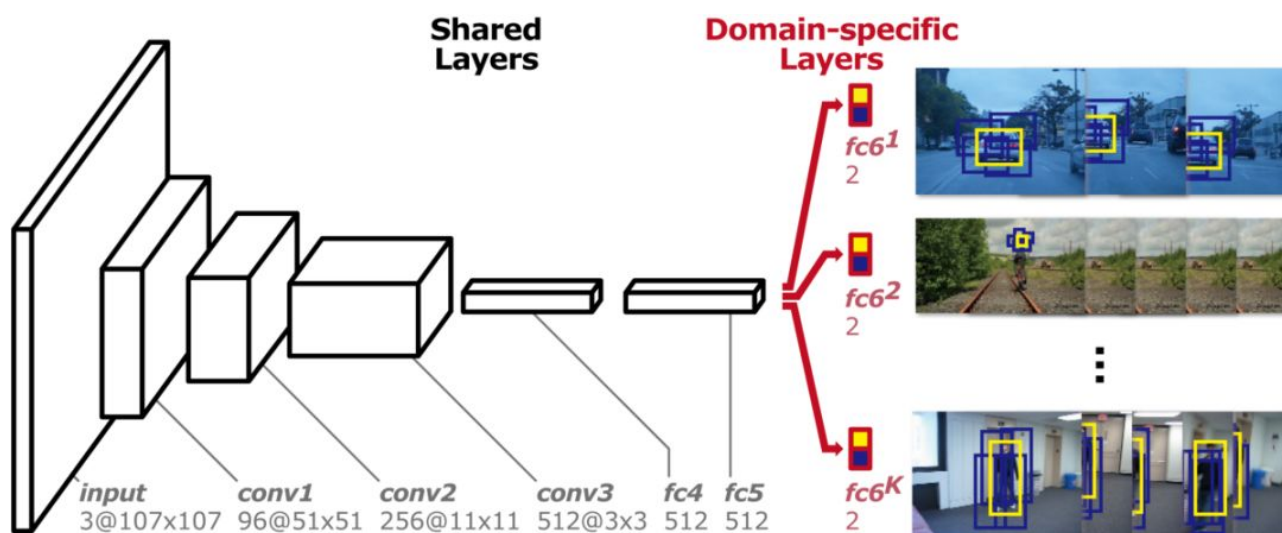
因此，FCNT 设计了特征选择网络，在 VGG 网络的卷积 4-3 和卷积 5-3 层上选择最相关的特征映射。然后为避免噪音的过拟合，FCNT 还为这两个层的特征映射单独设计了两个额外的通道(即 SNet 和 GNet)：GNet 捕获对象的类别信息; SNet 将该对象从具有相似外观的背景中区分出来。

这两个网络的运作流程如下：都使用第一帧中给定的边界框进行初始化，以获取对象的映射。而对于新的帧，对其进行剪切并传输最后一帧中的感兴趣区域，该感兴趣区域是以目标对象为中心。最后，通过 SNet 和 GNet，分类器得到两个预测热映射，而跟踪器根据是否存在干扰信息，来决定使用哪张热映射生成的跟踪结果。FCNT 的图如下所示。



与 FCNT 的思路不同，MD Net 使用视频的所有序列来跟踪对象的移动。上述网络使用不相关的图像数据来减少跟踪数据的训练需求，并且这种想法与跟踪有一些偏差。该视频中的一个类的对象可以是另一个视频中的背景，因此，MD Net 提出了“多域”这一概念，它能够在每个域中独立的区分对象和背景，而一个域表示一组包含相同类型对象的视频。

如下图所示，MD Net 可分为两个部分，即 K 个特定目标分支层和共享层：每个分支包含一个具有 softmax 损失的二进制分类层，用于区分每个域中的对象和背景;共享层与所有域共享，以保证通用表示。



近年来，深度学习研究人员尝试使用了不同的方法来适应视觉跟踪任务的特征，并且已经探索了很多方法：

应用到诸如循环神经网络(RNN)和深度信念网络(DBN)等其他网络模型;设计网络结构来适应视频处理和端到端学习，优化流程、结构和参数;或者将深度学习与传统的计算机视觉或其他领域的方法(如语言处理和语音识别)相结合。

■ 4、语义分割

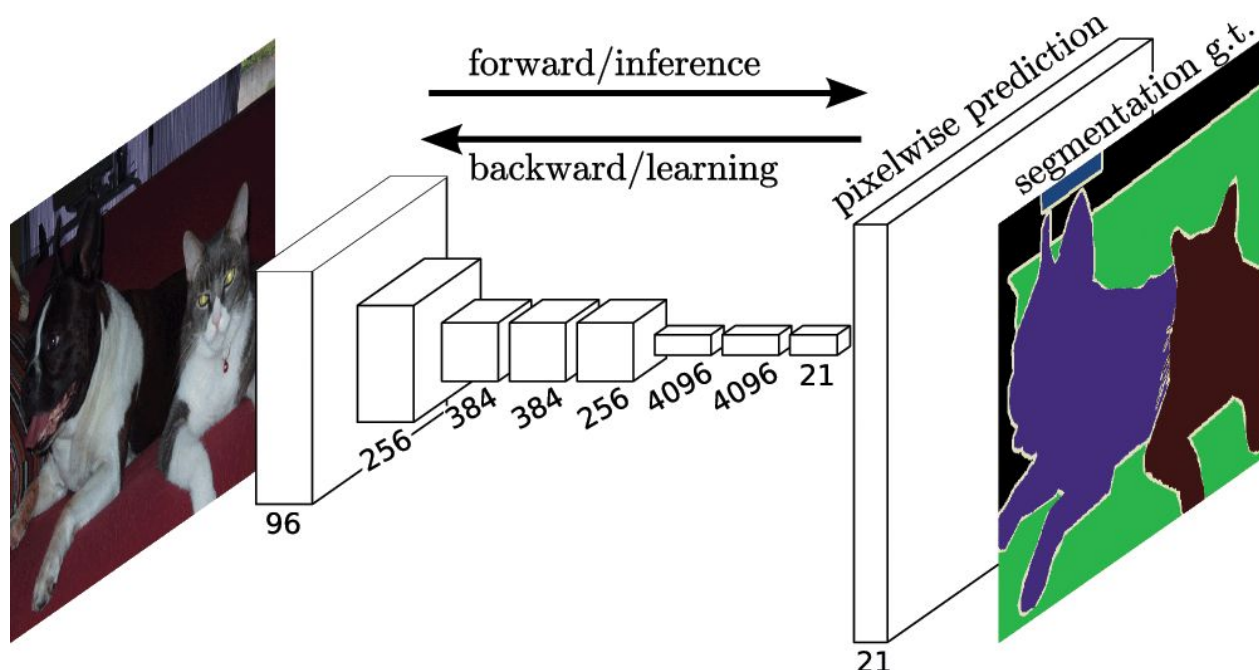


计算机视觉的核心是分割，它将整个图像分成一个个像素组，然后对其进行标记和分类。特别地，语义分割试图在语义上理解图像中每个像素的角色(比如，识别它是汽车、摩托车还是其他的类别)。如上图所示，除了识别人、道路、汽车、树木等之外，我们还必须确定每个物体的边界。因此，与分类不同，我们需要用模型对密集的像素进行预测。

与其他计算机视觉任务一样，卷积神经网络在分割任务上取得了巨大成功。最流行的原始方法之一是通过滑动窗口进行块分类，利用每个像素周围的图像块，对每个像素分别进行分类。但是其计算效率非常低，因为我们不能在重叠块之间重用共享特征。

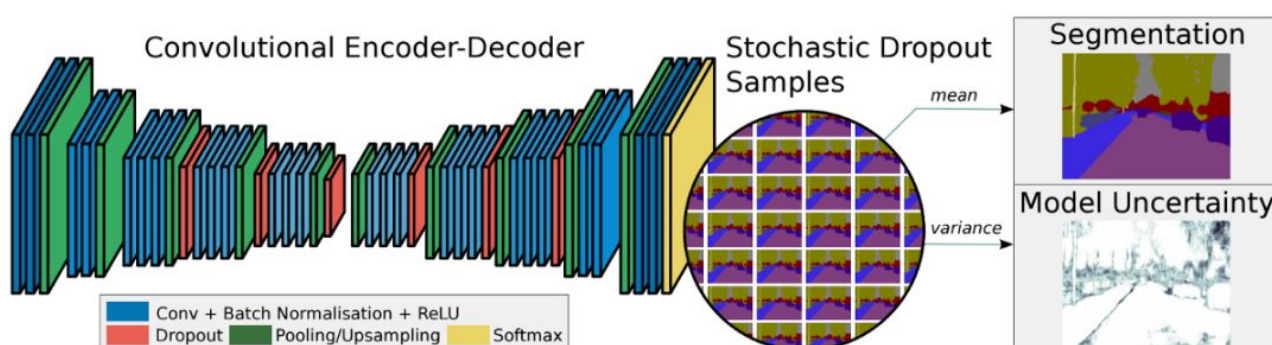
解决方案就是加州大学伯克利分校提出的全卷积网络(FCN)，它提出了端到端的卷积神经网络体系结构，在没有任何全连接层的情况下进行密集预测。

这种方法允许针对任何尺寸的图像生成分割映射，并且比块分类算法快得多，几乎后续所有的语义分割算法都采用了这种范式。



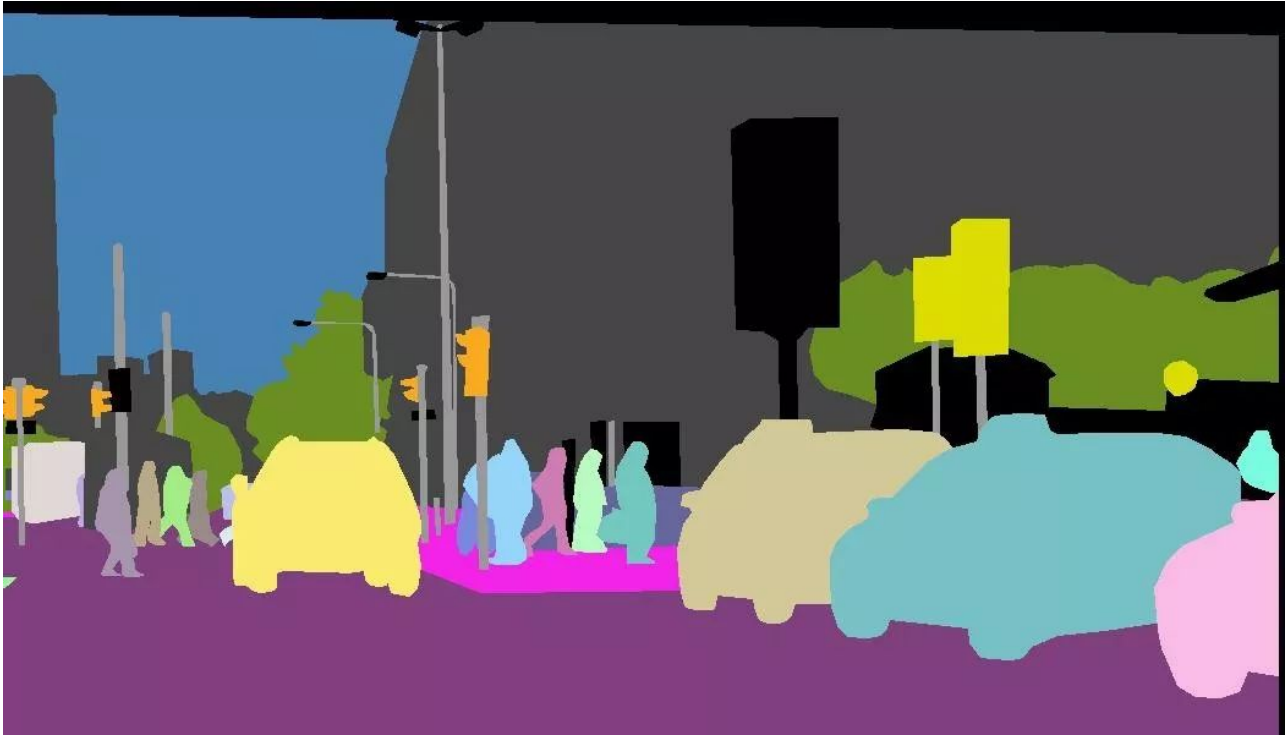
但是，这也仍然存在一个问题：在原始图像分辨率上进行卷积运算非常昂贵。为了解决这个问题，FCN 在网络内部使用了下采样和上采样：下采样层被称为条纹卷积(striped convolution);而上采样层被称为反卷积(transposed convolution)。

尽管采用了上采样和下采样层，但由于池化期间的信息丢失，FCN 会生成比较粗糙的分割映射。SegNet 是一种比 FCN (使用最大池化和编码解码框架)更高效的内存架构。在 SegNet 解码技术中，从更高分辨率的特征映射中引入了 shortcut/skip connections，以改善上采样和下采样后的粗糙分割映射。



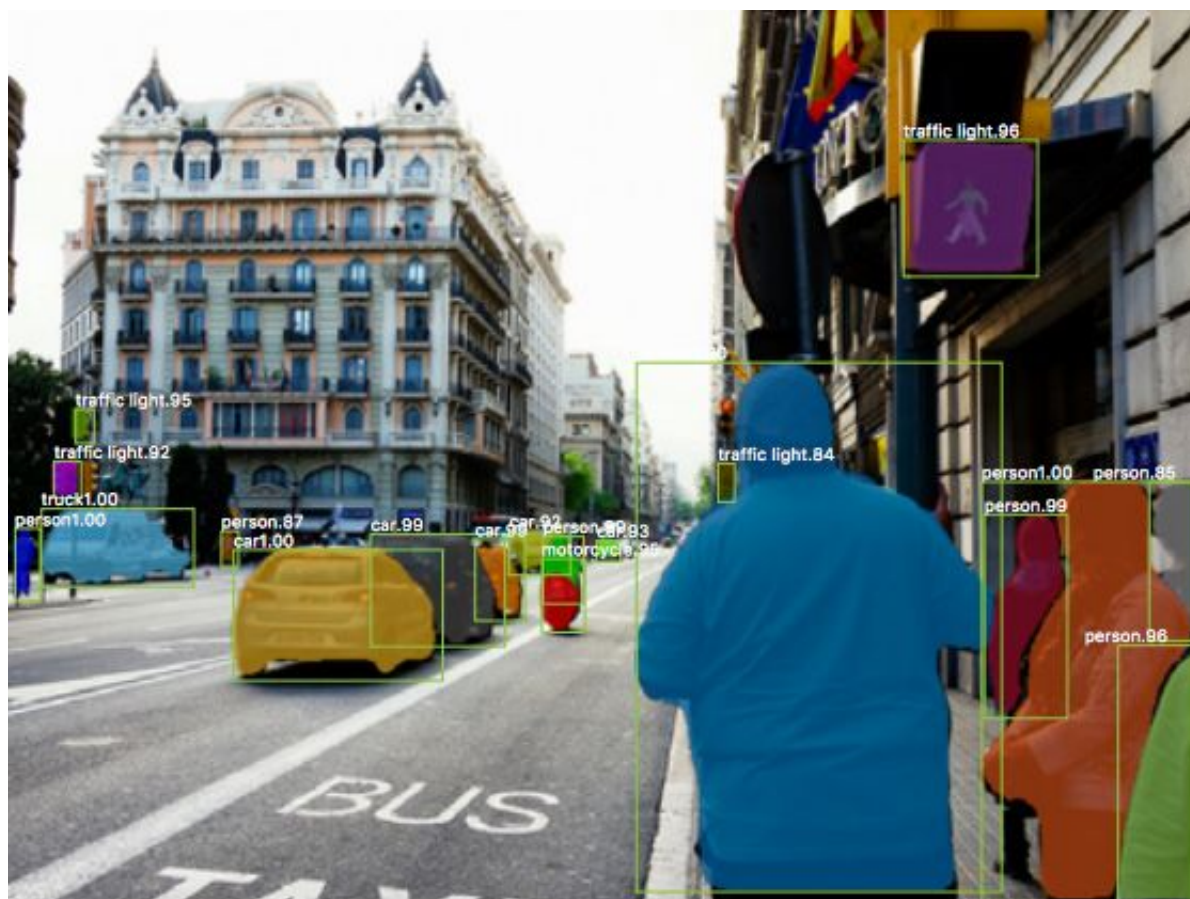
目前的语义分割研究都依赖于完全卷积网络，如空洞卷积 (Dilated Convolutions), DeepLab 和 RefineNet。

■ 5、实例分割



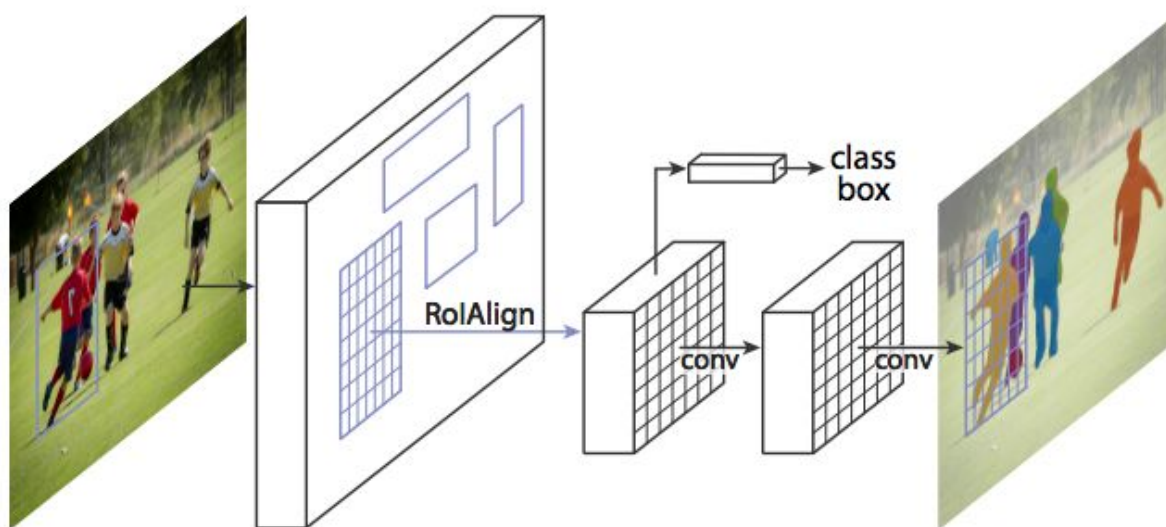
除了语义分割之外，实例分割将不同类型的实例进行分类，比如用 5 种不同颜色来标记 5 辆汽车。分类任务通常来说就是识别出包含单个对象的图像是什么，但在分割实例时，我们需要执行更复杂的任务。我们会看到多个重叠物体和不同背景的复杂景象，我们不仅需要将这些不同的对象进行分类，而且还要确定对象的边界、差异和彼此之间的关系！

到目前为止，我们已经看到了如何以多种有趣的方式使用卷积神经网络的特征，通过边界框有效定位图像中的不同对象。我们可以将这种技术进行扩展吗？也就是说，对每个对象的精确像素进行定位，而不仅仅是用边界框进行定位？Facebook AI 则使用了 Mask R-CNN 架构对实例分割问题进行了探索。



就像 Fast R-CNN 和 Faster R-CNN 一样，Mask R-CNN 的底层是鉴于 Faster R-CNN 在物体检测方面效果很好，我们是否可以将其扩展到像素级分割？

Mask R-CNN 通过向 Faster R-CNN 添加一个分支来进行像素级分割，该分支输出一个二进制掩码，该掩码表示给定像素是否为目标对象的一部分：该分支是基于卷积神经网络特征映射的全卷积网络。将给定的卷积神经网络特征映射作为输入，输出为一个矩阵，其中像素属于该对象的所有位置用 1 表示，其他位置则用 0 表示，这就是二进制掩码。



另外，当在原始 Faster R-CNN 架构上运行且没有做任何修改时，感兴趣池化区域(RoIPool)选择的特征映射区域或原始图像的区域稍微错开。由于图像分割具有像素级特性，这与边界框不同，自然会导致结果不准确。Mask R-CNN 通过调整 RoIPool 来解决这个问题，使用感兴趣区域对齐(Roialign)方法使其变的更精确。本质上，Roialign 使用双线性插值来避免舍入误差，这会导致检测和分割不准确。

一旦生成这些掩码，Mask R-CNN 将 RoIAlign 与来自 Faster R-CNN 的分类和边界框相结合，以便进行精确的分割：



■ 结语

上述这 5 种主要的计算机视觉技术可以协助计算机从单个或一系列图像中提取、分析和理解有用的信息。你还可以通过我的 GitHub 存储库(<https://github.com/khanhnamle1994/computer-vision>)获取所有的演讲幻灯片以及指南。

作者 | James Le

原文链接：<https://heartbeat.fritz.ai/the-5-computer-vision-techniques-that-will-change-how-you-see-the-world-1ee19334354b>

译者 | 王柯凝

出品 | AI科技大本营(公众号ID：rgznai100)

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发