149 | 计算机视觉高级话题（一）：图像物体识别和分割

2018-09-14 洪亮劼

讲述：初明明
时长 05:50　大小 2.68M

从今天开始，我们讨论几个相对比较高级的计算机视觉话题。这些话题都不是简单的分类或者回归任务，而是需要在一些现有的模型上进行改进。

我们聊的第一个话题就是图像中的**物体识别**（Object Recognition）和**分割**（Segmentation）。我们前面介绍过物体识别和分割。通俗地讲，就是给定一个输入的图像，我们希望模型可以分析这个图像里究竟有哪些物体，并能够定位这些物体在整个图像中的位置，对于图像中的每一个像素，能够分析其属于哪一个物体。

这一类型任务的目的是更加仔细地理解图像中的物体，包括图片分类、对图像里面的物体位置进行分析，以及在像素级别进行分割，这无疑是一个充满挑战的任务。

R-CNN

深度模型，特别是卷积神经网络（CNN）在 AlexNet 中的成功应用，很大程度上开启了神经网络在图像分类问题上的应用。这之后，不少学者就开始考虑把这样的思想利用到物体识别上。第一个比较成功的早期工作来自加州大学伯克利分校 [1]，这就是我们接下来要介绍的**R-CNN 模型**。

首先，R-CNN 的输入是一个图片，输出是一个“选定框”（Bounding Box）和对应的标签。R-CNN 采用了一种直观的方法来生成选定框：尽可能多地生成选定框，然后来看究竟哪一个选定框对应了一个物体。

具体来说，针对图像，R-CNN 先用不同大小的选定框来扫描，并且尝试把临近的具有相似色块、类型、密度的像素都划归到一起去。然后，再利用一个 AlexNet 的变形来对这些待定（Proposal）的选定框进行特征提取（Feature Extraction）。在模型的最后一层，R-CNN 加入了一个支持向量机（Support Vector Machine）来判断待选定框是否是某个物体。判断好了选定框以后，R-CNN 再运行一个线性回归来对选定框的坐标进行微调。

R-CNN 虽然证明了在物体识别这样的任务中，CNN 的确可以超越传统的模型，但整个模型由多个模块组成，相对比较繁琐。

Fast R-CNN

意识到了 R-CNN 的问题以后，一些学者开始考虑如何在这个模型上进行改进。第一个重大改进来自于 R-CNN 原文中的第一作者罗斯·吉尔什克（Ross Girshick）。吉尔什克这个时候已经来到了微软研究院，他把自己改进的模型叫作**Fast R-CNN**[2]。

Fast R-CNN 的一个重要特点就是观察到我们刚才介绍 R-CNN 中的第二步骤，也就是每一个待定的选定框都需要进行特征提取。这里的特征提取其实就是一个神经网络，往往非常消耗资源。而且很多待定的选定框有很多重叠的部分，可以想象就会有很多神经网络的计算是重复多余的。

那么，有没有什么办法我们可以针对一个图片仅仅运行一次神经网络，但是又可以针对不同的待选定框共享呢？这其实就是 Fast R-CNN 的核心思想。Fast R-CNN 的另外一个特点就是尝试用一个神经网络架构去替代 R-CNN 中间的四个模块。这样两个改进的结果是怎样的呢？Fast R-CNN 和 R-CNN 相比在效果上差不多，但是训练时间快了 9 倍以上。

Faster R-CNN 和 Mask R-CNN

在 Fast R-CNN 的技术上，一群当时在微软研究院的学者们把对 R-CNN 的加速往前推进了一步，这就是模型**Faster R-CNN**[3]。Faster R-CNN 是在如何提出待定的选定框上做了进一步的改进，使得这部分不依赖一个单独的步骤，而依赖我们已经训练的 CNN 网络。这在速度上比 Fast R-CNN 又快了不少。

在 Faster R-CNN 的基础上，**Mask R-CNN**不仅能够做到对图像中的物体进行判别，而且还能够做到像素级的抽取 [4]。前面我们在讲 2017 年 ICCV 最佳研究论文的时候，介绍过这部分内容。这里我带你做一个简单的回顾。

Faster R-CNN 分为两个阶段。第一个阶段是“区域提交网络”（Region Proposal Network），目的是从图像中提出可能存在候选矩形框。第二个阶段，从这些候选框中使用“RoIPool”这个技术来提取特征从而进行标签分类和矩形框位置定位这两个任务。这两个阶段的一些特征可以共享。

区域提交网络的大体流程是什么样的？大体来说，最原始的输入图像经过经典的卷积层变换之后形成了一个图像特征层。在这个新的图像特征层上，模型使用了一个移动的小窗口来对区域进行建模。

这个移动小窗口有这么三个任务需要考虑。首先移动小窗口所覆盖的特征经过一个变换达到一个中间层，然后经过这个中间层，直接串联到两个任务，也就是物体的分类和位置的定位。其次，移动的小窗口用于提出一个候选区域，也就是矩形框。而这个矩形框也参与刚才所说的定位信息的预测。当区域提交网络“框”出了物体的大致区域和类别之后，模型再使用一个“物体检测”的网络来对物体进行最终的检测。

Mask R-CNN 的第一部分完全使用 Faster R-CNN 所提出的区域提交网络，模型对第二部分进行了更改。那 Mask R-CNN 的第二部分都输出什么呢？不仅仅输出区域的类别和框的相对位置，同时还输出具体的像素分割。和很多类似工作的区别是，像素分割、类别判断、位置预测是三个独立的任务，并没有互相的依赖，这是作者们认为 Mask R-CNN 能够成功的一个重要的关键。

小结

今天我为你讲了计算机视觉高级话题之一的物体识别和分割技术。我们总结了从最早的 R-CNN 到加速的 Fast R-CNN 和更快的 Faster R-CNN，以及最后能够进行像素分割的 Mask R-CNN。

最后，给你留一个思考题，从这一系列模型的发展中，你能总结出一些心得体会吗？

欢迎你给我留言，和我一起讨论。

参考文献

1. Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. **Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation**. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580-587, 2014.
2. Ross Girshick. **Fast R-CNN**. The IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448, 2015.
3. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. **Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks**. Conference on Neural Information Processing Systems (NIPS), 2015.
4. K. He, G. Gkioxari, P. Dollar and R. Girshick. **Mask R-CNN**. In IEEE Transactions on Pattern Analysis and Machine Intelligence.

上一篇　148 | 计算机视觉领域的深度学习模型（三）：ResNet

下一篇　150 | 计算机视觉高级话题（二）：视觉问答

精选留言

写留言

由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。