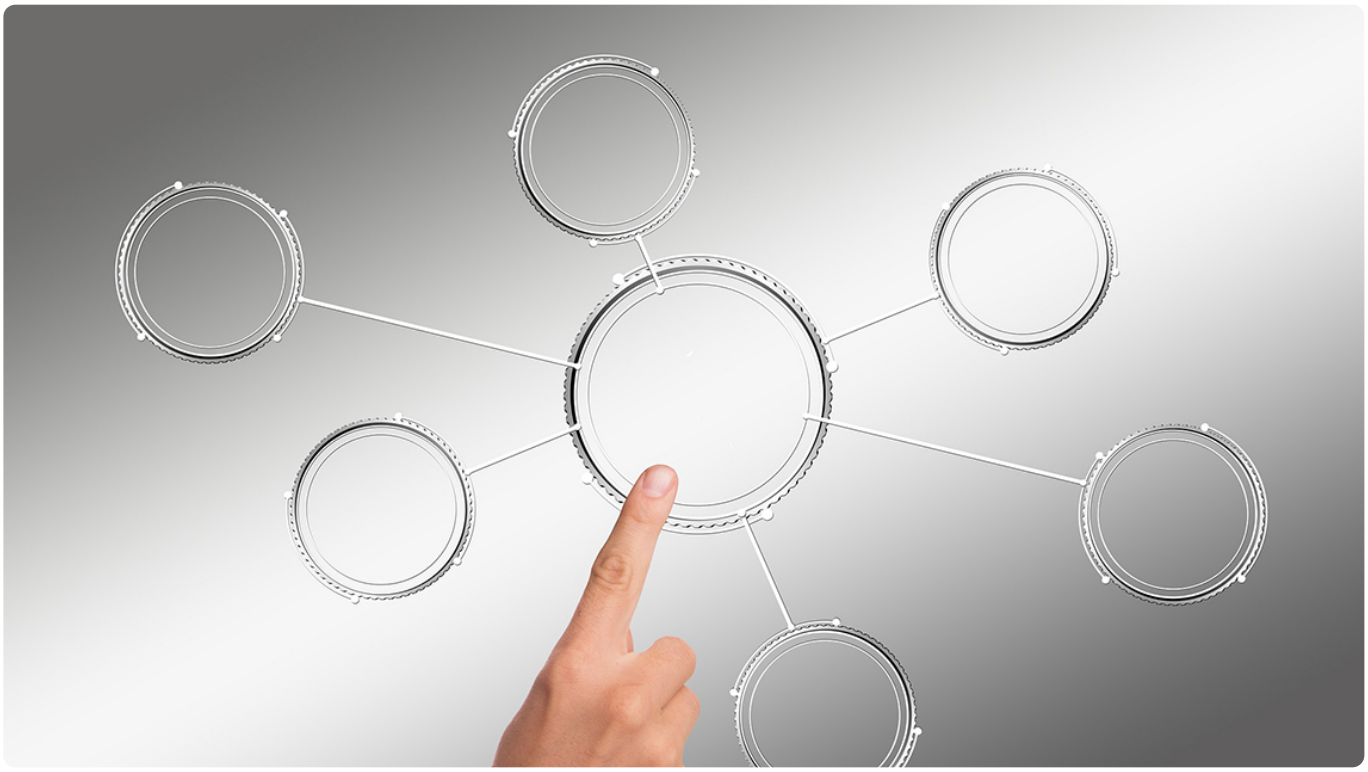
048 | 精读2017年ICCV最佳研究论文

2018-01-22 洪亮劼

AI技术内参

进入课程 >



讲述：初明明
时长 09:14　大小 4.23M

ICCV（International Conference on Computer Vision，国际计算机视觉大会），是每两年举办一次的计算机视觉顶级会议。从 1987 年开始举办，已经有 30 年的历史。2017 年的 ICCV 大会于 10 月 22 日至 29 日在意大利的水城威尼斯举行。

在每届 ICCV 大会上，都会从众多学术论文中挑选出两篇最有新意和价值的论文作为最佳研究论文和最佳学生论文。ICCV 的最佳论文奖又叫作“马尔奖项”（Marr Prize），是为了纪念英国的心理学家和神经科学家大卫·马尔（David Marr）而设计的奖项。马尔将心理学、人工智能和神经生理学的研究成果结合起来，提出了全新的关于视觉处理的理论，他被认为是计算神经科学的创始人。

今天，我就来带你认真剖析一下 ICCV 2017 年的最佳研究论文“Mask R-CNN”。这篇论文是一个集大成的工作，介绍了一个新的方法可以用于同时解决图像的“**物体识**

别”（Object Detection）、“**语义分割**”（Semantic Segmentation）和“**数据点分割**”（Instance Segmentation）的工作。

什么意思呢？通俗地讲，那就是给定一个输入的图像，利用这篇论文提出的模型可以分析这个图像里究竟有哪些物体，比如是一只猫，还是一条狗；同时能够定位这些物体在整个图像中的位置；并且还能针对图像中的每一个像素，知道其属于哪一个物体，也就是我们经常所说的，把物体从图像中“抠”出来。

作者群信息介绍

这篇论文的作者全部来自 Facebook 的人工智能研究院（Facebook AI Research）。

第一作者就是近几年在计算机视觉领域升起的学术之星何恺明博士（Kaiming He）。他于 2016 年加入 Facebook 人工智能研究院，之前在微软亚洲研究院进行计算机视觉的研究工作；他还是 CVPR 2016 年和 CVPR 2009 年的最佳论文得主。目前，何恺明在计算机视觉领域有三项重大贡献。

第一，他与其他合作者发明的 ResNet 从 2016 年以来成为了计算机视觉深度学习架构中的重要力量，被应用到了计算机视觉以外的一些领域，比如机器翻译和 AlphaGo 等，相关论文引用数超过 5 千次。

第二，他与其他合作者开发的 Faster R-CNN 技术，发表于 NIPS 2015 上，是图像物体识别和语义分析的重要技术手段，也是今天我们要讨论的这篇论文的基础，论文引用数超过 2 千次。

第三，他与其他合作者在 ICCV 2015 年发表论文《深入研究整流器：在 ImageNet 分类上超越人类水平》（Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification），研究了一种改进的 ReLU（Rectified Linear Unit，线性整流函数，又称修正线性单元）结构从而达到了更好的效果，论文引用数近 2 千次。

第二作者乔治亚·吉克里奥夏里（Georgia Gkioxari）目前是 Facebook 人工智能研究院的博士后研究员。乔治亚可以说是师出名门，在 Facebook 工作之前才从加州大学伯克利毕业，师从计算机视觉泰斗吉腾德拉·马利克（Jitendra Malik）。乔治亚之前还分别在谷歌大脑和谷歌研究院实习过。在过去几年中，乔治亚在计算机视觉界已经发表了多篇高质量论文。

第三作者皮奥特·多拉（Piotr Dollár）是 Facebook 人工智能研究院的一名经理。2007 年从加州大学圣地亚哥分校获得博士学位，2014 年加入 Facebook，这之前在微软研究院工作。皮奥特长期从事计算机视觉的研究工作。

最后一个作者罗斯·吉尔什克（Ross Girshick）是 Facebook 人工智能研究院的一名科学家。他于 2012 年毕业于芝加哥大学，获得计算机博士。罗斯之前也在微软研究院工作，也曾在计算机视觉泰斗吉腾德拉的实验室里担任博士后的研究工作。

论文的主要贡献

我们首先来看一下这篇文章的主要贡献。还是要先去理解，这篇文章主要解决的是一个什么场景下的问题。

刚才我们已经简单地谈到了，这篇文章要解决的问题，就是对输入图像的物体识别、语义分割，以及数据点分割，是这三个任务的一个集成。在之前的一个工作中，“Faster R-CNN”[1] 已经解决了前两个任务。那么，这篇论文其实就是 Faster R-CNN 在逻辑上的一个扩展。然而，这个扩展也并不是那么显而易见的。**为了解决数据点分割的任务，Mask R-CNN 提出了深度学习网络结构上的一个创新，这是本篇论文的一个重要贡献。**

本文提出的模型不仅在数据点分割的标准数据集 COCO 上表现强劲，击败所有之前提出的模型以外，还能够很容易地扩展到其他的任务中，比如“人体形态估计”（Human Pose Estimation），从而**奠定了 Mask R-CNN 作为一个普适性框架的地位。**

论文的核心方法

要想理解 Mask R-CNN 的核心思想，我们就必须先简要理解 Faster R-CNN 的一些基本原理。刚才说到了，Mask R-CNN 就是在其之上的一种改进和延伸。

Faster R-CNN 对于每一个输入图像中的每一个候选物体，都会有两个输出，一个是候选物体的**标签**（比如，猫、狗、马等），还有一个就是一个**矩形框**（Bounding Box），用于表达这个物体在图像中的位置。第一个**标签输出是一个分类问题**（Classification），而第二个**位置预测则是一个回归问题**（Regression）。

Faster R-CNN 分为两个阶段（Stage）。第一个阶段叫作“**区域提交网络**”（Region Proposal Network），目的是从图像中提出可能存在的候选矩形框。第二个阶段，从这些

候选框中使用一个叫“**RoIPool**”的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性可以共享。

区域提交网络的大体流程是这样的。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（Sliding Window）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，有时候也叫 ROI，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”（Object Detection）的网络来对物体进行最终的检测。在这里，物体检测实际是使用了 Fast R-CNN[2] 的架构。所以，也就是为什么 Faster R-CNN 的名字里用“Faster”来做区分。Faster R-CNN 的贡献，在于区域提交网络和 Fast R-CNN 的部分，也就是物体检测的部分达到了共享参数，或者叫共享网络架构，这样也就起到了加速的作用。

Mask R-CNN 在第一部分完全使用 Faster R-CNN 所提出的区域提交网络，在此基础上，对第二部分进行了更改。也就是说，不仅仅在第二部分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。然而，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。对比之前的一些工作，像素分割成了类别判断的依赖，从而导致这几个任务之间互相干扰。

Mask R-CNN 在进行像素分割的时候，因为要在原始的图像上进行分割，因此需要在整个流程中保留原始图像的位置关系。这个需求是类别判断和位置预测所不具备的。而在 Faster R-CNN 中，因为不需要这个需求，因此类别判断和位置预测所依赖的信息是一个压缩过后的中间层。那么很明显，Mask R-CNN 依靠这个压缩层就不够了。在这篇文章中，作者们**提出了一个叫 RoIAlign 的技术来保证中间提取的特征能够反映在最原始的像素中。**如果对这部分内容感兴趣，建议你去细读文章。

方法的实验效果

作者们使用 Mask R-CNN 在目前流行的图像物体检测任务数据集 COCO 2015 和 COCO 2016 上做了检测，相对于之前的这两个竞赛的冠军，实验结果表明 Mask R-CNN 的精度

都大幅度增加。在一个“平均精度”（Average Precision）的度量上，Mask R-CNN 比 COCO 2015 的最佳结果好了近 13%，而比 COCO 2016 的最佳结果好了 4%，可以说效果非常明显。在实验结果中，作者们非常细致地测试了整个 Mask R-CNN 中每一个部件的效果。其中，把三个任务分开、以及 RoIAlign 方法都有非常显著的作用，证明了这些模型组件是优秀结果的必要步骤。

小结

今天我为你讲了 ICCV 2017 年的最佳研究论文，这篇文章介绍了目前在图像物体识别中的最新算法 Mask R-CNN 的大概内容。

一起来回顾下要点：第一，我们简要介绍了这篇文章的作者群信息。第二，我们详细介绍了这篇文章要解决的问题以及贡献 。第三，我们简要地介绍了文章提出方法的核心内容 。

最后，给你留一个思考题，你觉得为什么 Mask R-CNN，包括之前的一些工作，要把物体检测的工作分为两步，第一步先分析一个大的矩形框，第二步进行物体检测，这两步都是必要的吗？

欢迎你给我留言，和我一起讨论。

参考文献

1. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 6 (June 2017), 1137-1149, 2017.
2. Ross Girshick. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15). IEEE Computer Society, Washington, DC, USA, 1440-1448, 2015.

上一篇　047 | 职场话题：聊聊数据科学家的职场规划

下一篇　049 | 精读2017年ICCV最佳学生论文

精选留言 (1)

写留言

林彦
2018-01-24

第一步先分析一个大的矩形框，第二步进行物体检测。因为要打标签，第二步肯定是需要的。因为最终目标是物体检测，为了检测中图片中是否有物体，是什么物体。先把完整包含物体可能性最大的区域框出来，然后做里面的物体分类。第一步也是必须的。可以看成是最初CNN图片分类的升级，也更接近我们人类对于复杂图片中的物体识别方法。

展开