華中科技大學

计 算 机 视 觉 课 程 报 告

题目： 基于前馈神经网络的分类任务设计

学　　号　1111

姓　　名　thezmmm

专　　业　数据科学与大数据技术

班　　级　000

指 导 教 师　杨卫

计算机科学与技术学院

摘　要

本文综述了目标检测领域的相关技术。目标检测是计算机视觉中的重要任务，旨在从图像或视频中准确地识别和定位多个目标。本文主要介绍了两类主流的目标检测框架：基于区域提议和基于回归/分类。

在基于区域提议的框架中，我们介绍了几种经典方法，包括 R-CNN、SPP-Net、Fast R-CNN、R-FCN 和 Mask-CNN。这些方法通过生成候选区域并使用深度学习模型对候选区域进行分类和边界框回归，实现目标的检测和定位。其中，Mask-CNN 还支持目标实例的精确分割。

在基于回归/分类的框架中，我们介绍了先前经验、YOLO 和 SSD 等方法。这些方法通过将目标检测任务转化为回归或分类问题来实现实时性能。它们利用先验知识、多层次的特征提取和多尺度的边界框预测，实现对图像中多个目标的高效检测。

关键词： 目标检测，图像识别

目　录

一　介绍

目标检测是一项重要的计算机视觉任务,涉及在数字图像中检测特定类别的视觉对象实例(例如人类、动物或汽车)。目标检测的目标是开发计算模型和技术,为计算机视觉应用提供最基本的知识之一:哪些对象在哪里?目标检测的两个最重要的指标是准确性(包括分类准确性和定位准确性)和速度。目标检测是许多其他计算机视觉任务的基础,例如实例分割,图像字幕生成和目标跟踪。近年来,深度学习技术的快速发展极大地推动了目标检测的进步,取得了显著突破,并将其推向了一个备受关注的研究热点。目标检测现在已广泛应用于许多实际应用中,例如自动驾驶、机器人视觉和视频监控。

为了获得对图像的完整理解,我们不仅应该集中精力对不同图像进行分类,还应该尝试准确估计每个图像中包含的对象的概念和位置。这个任务被称为目标检测,通常包括不同的子任务,如人脸检测,行人检测和骨架检测。作为计算机视觉的基本问题之一,目标检测能够为图像和视频的语义理解提供有价值的信息,并与许多应用相关,包括图像分类,人体行为分析,人脸识别和自动驾驶。与此同时,继承自神经网络和相关的学习系统,这些领域的进展将发展神经网络算法,并且也对可以被视为学习系统的目标检测技术产生重大影响。然而,由于视点、姿态、遮挡和光照条件的巨大变化,要想完美地完成目标检测任务以及额外的目标定位任务是困难的。因此,近年来这个领域引起了广泛关注。

由于不同的检测任务具有完全不同的目标和约束条件,它们的困难程度可能会有所不同。除了一些其他计算机视觉任务中的共同挑战(例如不同视角、光照和类内变化下的对象),目标检测中的挑战包括但不限于以下几个方面:对象旋转和尺度变化(例如小对象),准确的对象定位,密集和遮挡的对象检测,加速检测等。

本文旨在全面了解目标检测的相关技术,使得我们能对目标检测技术进行一个大致的掌握,同时比较不同技术的优缺点,从而在进行目标检测任务时能够选择更加合适的技术。

二　目标检测相关技术

通用目标检测旨在在图像中定位和分类现有的对象，并使用矩形边界框对它们进行标记以显示存在的置信度。通用目标检测方法的框架主要可以分为两类。一种遵循传统的目标检测流程，首先生成区域提议，然后将每个提议分类为不同的对象类别。另一种将目标检测视为回归或分类问题，采用统一的框架直接实现最终结果（类别和位置）。基于区域提议的方法主要包括 R-CNN、空间金字塔池化（SPP-net）、快速 R-CNN 、Faster R-CNN 、基于区域的全卷积网络（R-FCN）、特征金字塔网络（FPN）和掩膜 R-CNN ，其中一些方法彼此之间存在相关性（例如，SPP-net 通过添加 SPP 层修改了 R-CNN）。基于回归/分类的方法主要包括 MultiBox 、AttentionNet、G-CNN、YOLO、单次多框检测器（SSD）、YOLOv2、去卷积单次检测器（DSSD）和深度监督目标检测器（DSOD）。这两种流程之间的关联通过 Faster R-CNN 中引入的锚点进行桥接。以下是这些方法的详细信息。

2.1　基于区域提议的框架

基于区域提议的框架是一个两步的过程，在某种程度上类似于人脑的注意机制，首先对整个场景进行粗略扫描，然后将注意力集中在感兴趣的区域（RoIs）上。在先前的相关工作中，最具代表性的是 Overfeat。该模型将 CNN 插入到滑动窗口方法中，从最高层特征图的位置直接预测边界框，并获取潜在对象类别的置信度。

2.1.1　R-CNN

提高候选边界框的质量和采用深度架构来提取高层次特征非常重要。为了解决这些问题，Girshick 等人提出了 R-CNN ，在 PASCAL VOC 2012 数据集上取得了 53.3% 的平均精确度（mean average precision，mAP），相比之前最佳结果（基于 DPM 直方图的稀疏编码）有了 30% 以上的改善。图 3 显示了 R-CNN 的流程图，可以分为以下三个阶段：尽管 R-CNN 相对传统方法有所改进，并在将 CNN 引入实际目标检测方面具有重要意义，但仍存在一些缺点。

由于存在全连接（FC）层，CNN 需要一个固定大小（例如 227×227）的输入图像，在每个评估的区域中需要重新计算整个 CNN，导致测试期间耗时较长。R-CNN 的训练是一个多阶段的流程。首先，对对象提议上的卷积网络（ConvNet）进行微调。然后，用 SVM 替换通过微调学习的 softmax 分类器，以适应 ConvNet 的特征。最后，训练边界框回归器。训练的空间和时间成本高。从不同的区域提议中提取特征并存储在磁盘上。使用非常深的网络（如 VGG16）处理相对较小的训练集将花费很长时间。同时，这些特征所需的存储

内存也应引起关注。尽管选择性搜索可以生成相对高召回率的区域提议，但所得到的区域提议仍然存在冗余，并且这个过程耗时(提取 2000 个区域提议约需 2 秒)。

2.1.2 SPP-Net

全连接层需要接受固定尺寸的输入。这就是为什么 R-CNN 选择将每个区域提议进行扭曲或裁剪为相同尺寸的原因。然而，对象可能部分存在于裁剪的区域中，并且由于扭曲操作可能产生不需要的几何畸变。这些内容损失或畸变将降低识别准确性，特别是当对象的尺度变化时。为了解决这个问题，何凯明等人考虑了空间金字塔匹配(SPM)理论，并提出了一种名为 SPP-net 的新型 CNN 架构。SPM 使用多个从细到粗的尺度将图像分割成多个区域，并将量化的局部特征聚合成中层表示。

与 R-CNN 不同，SPP-net 重用了第五个卷积层(conv5)的特征图，将任意尺寸的区域提议投影到固定长度的特征向量上。这些特征图可重用的可行性是因为这些特征图不仅包含局部响应的强度，还与它们的空间位置有关。最终卷积层之后的层被称为 SPP 层。

SPP-net 不仅能够在相应尺度上正确估计不同区域提议，并且在测试阶段通过在 SPP 层之前共享计算成本，提高了检测效率，减少了计算开销的重复。

2.1.3 Fast R-CNN

尽管 SPP-net 在准确性和效率方面相比 R-CNN 取得了令人瞩目的改进，但它仍然有一些明显的缺点。SPP-net 几乎与 R-CNN 具有相同的多阶段流水线，包括特征提取、网络微调、SVM 训练和边界框回归器拟合。因此，仍然需要额外的存储空间开销。此外，SPP 层之前的卷积层无法使用中介绍的微调算法进行更新。因此，非常深的网络的准确性下降是可以预料的。为此，Girshick 引入了一种在分类和边界框回归上的多任务损失，并提出了一种名为 Fast R-CNN 的新型 CNN 架构。

与 SPP-net 类似，整个图像经过卷积层处理以生成特征图。然后，使用 RoI 池化层从每个区域提议中提取出固定长度的特征向量。RoI 池化层是 SPP 层的特殊情况，只有一个金字塔级别。然后，将每个特征向量输入到一系列的全连接层中，最后分为两个并行的输出层。一个输出层负责为所有 C + 1 个类别(C 个对象类别加上一个“背景”类别)生成 softmax 概率，另一个输出层使用四个实数编码了精细调整的边界框位置。在这些步骤中的所有参数(除了区域提议的生成)都通过一个端到端的多任务损失进行优化。

为了加速 Fast R-CNN 的流程，还需要另外两个技巧。一方面，如果训练样本 (即 RoIs) 来自不同的图像，通过 SPP 层的反向传播变得非常低效。Fast R-CNN 以分层方式对小批量样本进行采样，即首先随机采样 N 个图像，然后在每个图像中进行 R/N 个 RoIs 的采样，其中 R 表示 RoIs 的数量。关键是，正向和反向传播时，来自同一图像的 RoIs 共享计算和内存。另一方面，在正向传播中计算 FC 层需要很多时间。截断奇异值分解

(SVD)可用于压缩大型 FC 层并加速测试过程。

在 Fast R-CNN 中, 除了区域提议生成, 所有网络层的训练可以在单个阶段中进行, 使用多任务损失。这节省了额外的存储空间开销, 并通过更合理的训练方案提高了准确性和效率。

2.1.4 R-FCN

通过 RoI 池化层的划分, 一种常见的目标检测深度网络家族由两个子网络组成: 一个共享的全卷积子网络(与 RoIs 无关)和一个非共享的 RoI-wise 子网络。这种分解源自先驱的分类架构(例如 AlexNet 和 VGG16), 它们由一个卷积子网络和若干个全连接层组成, 中间由一个特定的空间池化层分隔。最近的最先进的图像分类网络, 如 ResNets 和 GoogLeNets, 都是全卷积的。为了适应这些架构, 自然而然地构建一个没有 RoI-wise 子网络的全卷积目标检测网络。然而, 这种简单的解决方案却表现不佳。这种不一致性是由于在目标检测中尊重平移不变性与在图像分类中增加平移不变性之间的两难选择。换句话说, 在图像分类中, 将图像中的对象移动位置应该是无差别的, 而在目标检测中, 边界框中对象的任何平移可能是有意义的。将 RoI 池化层手动插入卷积中可以破坏平移不变性, 但需要额外的非共享区域层。因此, Dai 等人提出了 R-FCN(见补充材料中的图 S2)。

与 Faster R-CNN 不同, 对于每个类别, R-FCN 的最后一个卷积层产生了一共 k2 个位置敏感的分数图, 其中 k×k 是一个固定网格, 然后在其后附加了一个位置敏感的 RoI 池化层, 以聚合这些分数图的响应。最后, 在每个 RoI 中, 平均 k2 个位置敏感得分以生成一个 C + 1 维向量, 并计算跨类别的 softmax 响应。另外, 还附加了一个 4k2 维的卷积层来获取与类别无关的边界框。

通过 R-FCN, 可以采用更强大的分类网络, 在完全卷积的架构中共享几乎所有的层, 同时在 PASCAL VOC 和 Microsoft COCO 数据集上实现了最先进的结果, 测试速度为每张图像 170 毫秒。

2.1.5 Mask-CNN

R-CNN: 实例分割是一项具有挑战性的任务, 需要检测图像中的所有对象并对每个实例进行分割(语义分割)。这两个任务通常被视为两个独立的过程。多任务方案会在重叠实例上产生虚假边缘并展示系统性错误。为了解决这个问题, 与 Faster R-CNN 中用于分类和边界框回归的现有分支并行, Mask R-CNN 添加了一个分支以像素对像素的方式预测分割掩码(图 8)。

与其他两个分支不可避免地通过全连接层折叠为短输出向量不同, 分割掩码分支编码了一个 m×m 的掩码以保持明确的对象空间布局。这种完全卷积表示需要更少的参数,

但比中的表示更准确。形式上,除了用于分类和边界框回归的公式(1)中的两个损失之外,还定义了一个额外的损失函数,用于分割掩码分支以实现多任务损失。这个损失只与真实类别相关,并依赖于分类分支来预测类别。

由于 RoI 池化是 Faster R-CNN 中的核心操作,用于特征提取时执行粗糙的空间量化,导致 RoI 和特征之间存在不对齐。这对分类影响较小,因为分类对小平移具有鲁棒性。然而,它对像素对像素的掩码预测产生了较大的负面影响。为了解决这个问题,Mask R-CNN 采用了一种简单且无量化的层,即 RoIAlign,以准确地保留像素级的空间对应关系。RoIAlign 通过用双线性插值替换 RoI 池化的严格量化,在每个 RoI bin 中的四个定期采样位置计算输入特征的精确值。尽管它很简单,但这个看似微小的改变极大地提高了掩码的准确性,特别是在严格的定位度量下。

在 Faster R-CNN 框架下,掩码分支只增加了少量的计算负担,并且它与其他任务的合作为目标检测提供了互补信息。因此,Mask R-CNN 易于实现,具有有希望的实例分割和目标检测结果。总之,Mask R-CNN 是一个灵活高效的实例级别识别框架,可以很容易地推广到其他任务(例如人体姿势估计),只需进行最小的修改。

2.2 基于回归/分类的框架

基于区域提议的框架由几个相关阶段组成,包括区域提议生成、使用卷积神经网络进行特征提取、分类和边界框回归,这些通常是分开训练的。即使在最近的端到端模块 Faster R-CNN 中,仍然需要进行替代训练以获得 RPN 和检测网络之间共享的卷积参数。因此,在处理不同组件时所花费的时间成为实时应用中的瓶颈。

基于全局回归/分类的一步框架,直接将图像像素映射到边界框坐标和类别概率,可以减少时间开销。我们首先回顾一些先驱的卷积神经网络模型,然后重点介绍两个重要的框架,即 YOLO 和 SSD。

2.2.1 先前经验

在 YOLO 和 SSD 之前,许多研究人员已经尝试将目标检测建模为回归或分类任务。

Szegedy 等人将目标检测任务定义为基于 DNN 的回归,为测试图像生成二进制掩码,并使用简单的边界框推断提取检测结果。然而,该模型难以处理重叠的物体,并且通过直接上采样生成的边界框远非完美。

Pinheiro 等人提出了一个具有两个分支的 CNN 模型:一个生成类别无关的分割掩码,另一个预测以物体为中心的给定补丁的可能性。由于类别分数和分割可以在单个模型中获得,并且大部分 CNN 操作是共享的,因此推断是高效的。

Erhan 等人和 Szegedy 等人提出了基于回归的 MultiBox,用于生成评分的类别无关

的区域提议。引入了统一的损失函数,偏置了多个组件的定位和置信度,以预测类别无关边界框的坐标。然而,最终层引入了大量的额外参数。

Yoo 等人采用了迭代分类方法来处理目标检测,并提出了一种令人印象深刻的端到端 CNN 架构,名为 AttentionNet。从图像的左上角和右下角开始,AttentionNet 通过生成量化的弱方向来指向目标对象,并通过迭代预测的集合收敛到准确的物体边界框。然而,在使用渐进的两步过程处理多个类别时,该模型变得相当低效。

Najibi 等人提出了一种无需提议的迭代网格型目标检测器(G-CNN),将目标检测建模为从固定网格到紧密包围目标的边界框之间的路径查找。从固定的多尺度边界框网格开始,G-CNN 训练一个回归器,通过迭代地将网格元素移动和缩放到对象上。然而,G-CNN 在处理小型或高度重叠的对象时存在困难。

2.2.2 YOLO

Redmon 等人提出了一种名为 YOLO 的新颖框架,它利用整个最顶层的特征图来预测多个类别的置信度和边界框。

YOLO 将输入图像分成一个 S×S 的网格,每个网格单元负责预测该网格单元中心的物体。每个网格单元预测 B 个边界框及其对应的置信度分数。形式上,置信度分数被定义为 Pr(Object)∗IOUtruth pred,它表示存在物体的可能性 (Pr(Object) ≥ 0) 并显示其预测的置信度 (IOUtruth pred)。同时,每个网格单元还应该预测 C 个条件类别概率 (Pr(Classi|Object))。需要注意的是,只计算包含物体的网格单元的贡献。

YOLO 由 24 个卷积层和 2 个全连接层组成,其中一些卷积层通过 1×1 降维层后跟 3×3 卷积层构建了 Inception 模块的集合。该网络可以以每秒 45 帧的速度实时处理图像,而简化版本的 Fast YOLO 则可以达到每秒 155 帧,并且比其他实时检测器具有更好的结果。此外,YOLO 在背景上产生的误报较少,这使得与 Fast R-CNN 的合作成为可能。后来提出了改进版本 YOLOv2,采用了几种令人印象深刻的策略,如 BN,锚框,尺寸聚类和多尺度训练。

2.2.3 SSD

YOLO 在处理小物体组时存在困难,这是由于边界框预测所施加的强空间约束。同时,由于多次下采样操作,YOLO 在处理具有新/非常规长宽比/配置的对象时难以泛化,并且产生相对粗糙的特征。

针对这些问题,Liu 等人提出了 SSD,其受到了 MultiBox 、RPN 和多尺度表示中采用的锚点的启发。给定特定的特征图,SSD 不采用 YOLO 中的固定网格,而是利用一组具有不同长宽比和尺度的默认锚点框来离散化边界框的输出空间。为了处理不同尺寸的对象,网络将来自具有不同分辨率的多个特征图的预测融合在一起。

在给定 VGG16 骨干网络结构的基础上，SSD 在网络末尾添加了几个特征层，负责预测具有不同尺度和长宽比的默认锚点框的偏移量及其相关的置信度。网络使用定位损失（如平滑 L1 损失）和置信度损失（如 Softmax）的加权和进行训练，这与 (1) 类似。通过在多尺度优化后的边界框上进行非极大值抑制（NMS），可以获得最终的检测结果。

SSD 结合了困难负样本挖掘、数据增强和更多精心选择的默认锚点，相对于 Faster R-CNN，在 PASCAL VOC 和 COCO 数据集上的准确性显著优于后者，同时速度提高了三倍。SSD300（输入图像尺寸为 300×300）的运行速度为 59 fps，比 YOLO 更准确高效。然而，SSD 在处理小物体方面并不擅长，可以通过采用更好的特征提取器骨干（例如 ResNet101），添加具有跳跃连接的反卷积层以引入额外的大尺度上下文，以及设计更好的网络结构（例如干细胞块和稠密块）来缓解这个问题。

三　总结

本文主要介绍了目标检测相关的技术, 主要分为基于区域提议的框架和基于回归/分类的框架两类。

在基于区域提议的框架中, 我们介绍了几种经典的方法。首先是 R-CNN, 它通过选择性搜索生成候选区域, 并使用卷积神经网络对每个候选区域进行分类和边界框回归。接着是 SPP-Net, 它通过共享卷积特征来提高计算效率。Fast R-CNN 在此基础上进一步优化了训练和测试的速度。R-FCN 则通过使用位置敏感的 ROI 池化层来提高准确性。最后, Mask-CNN 在 Fast R-CNN 的基础上增加了对目标实例分割的支持, 可以生成目标的精确掩码。

在基于回归/分类的框架中, 我们介绍了先前经验、YOLO 和 SSD 这几种方法。先前经验是指利用先验知识来设计目标检测模型, 例如使用滑动窗口和手工特征。YOLO (You Only Look Once) 是一种端到端的目标检测方法, 通过将目标检测任务转化为回归问题来提高检测速度。SSD (Single Shot MultiBox Detector) 则是另一种经典的单阶段目标检测方法, 它通过在不同层级的特征图上预测多个不同尺度和比例的边界框来实现多尺度目标检测。

综上所述, 目标检测是计算机视觉领域的重要任务, 不同的方法在速度和准确性上有所取舍。基于区域提议的框架更加准确, 但计算量较大; 而基于回归/分类的框架在速度上更有优势, 但准确率可能稍低。研究人员可以根据具体应用的需求选择适合的方法。

参考文献

[1] Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey[J/OL]. Proceedings of the IEEE, 2023, 111(3): 257-276. DOI: 10.1109/JPROC.2023.3238524.

[2] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916. DOI: 10.1109/TPAMI.2015.2389824.

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[A]. 2014. arXiv: 1311.2524.

[4] Girshick R. Fast r-cnn[C/OL]//2015 IEEE International Conference on Computer Vision (ICCV). 2015: 1440-1448. DOI: 10.1109/ICCV.2015.169.

[5] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C/OL]//Cortes C, Lawrence N, Lee D, et al. Advances in Neural Information Processing Systems: volume 28. Curran Associates, Inc., 2015. https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

[6] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[C/OL]//Lee D, Sugiyama M, Luxburg U, et al. Advances in Neural Information Processing Systems: volume 29. Curran Associates, Inc., 2016. https://proceedings.neurips.cc/paper_files/paper/2016/file/577ef1154f3240ad5b9b413aa7346a1e-Paper.pdf.

[7] Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[8] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C/OL]//2017 IEEE International Conference on Computer Vision (ICCV). 2017: 2980-2988. DOI: 10.1109/ICCV.2017.322.

[9] Erhan D, Szegedy C, Toshev A, et al. Scalable object detection using deep neural networks[A]. 2013. arXiv: 1312.2249.

[10] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[A]. 2016. arXiv: 1506.02640.

[11] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[M/OL]. Springer International Publishing, 2016: 21–37. http://dx.doi.org/10.1007/978-3-319-46448-0_2.

[12] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger[A]. 2016. arXiv: 1612.08242.

[13] Fu C Y, Liu W, Ranga A, et al. Dssd : Deconvolutional single shot detector[A]. 2017. arXiv: 1701.06659.

[14] Zhou L Y, Wei D, Ran Y B, et al. Reclining public chair behavior detection based on improved yolov5 [J/OL]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2023, 27(6): 1175-1182. DOI: 10.20965/jaciii.2023.p1175.