探究平均准确度 AP 指标的缺陷及其影响

窦勇*, 乔鹏, 靳若春

国防科技大学并行与分布式处理重点实验室, 长沙 410073

* 通信作者. E-mail: yongdou@nudt.edu.cn

收稿日期: 2018-05-25; 修回日期: 2018-09-27; 接受日期: 2018-12-17; 网络出版日期: 2019-10-15

国家重点研发计划 (批准号: 2018YFB1003405) 和国家自然科学基金 (批准号: 61732018) 资助项目

摘要 平均准确度 AP 是计算机视觉领域物体检测算法的重要性能评价指标, 长期用于相关学术研究成果的定量比较. 我们在实践中发现 AP 指标存在缺陷不能准确反映物体检测算法的数量准确性. 我们具体分析了 AP 存在缺陷的两种表现, 首先是 AP 定义和计算方法允许虚警出现时召回率不变, 出现一个召回率对应多个准确度的情况, AP 值计算仅和该召回率对应的最大准确度有关, 使得准确度降低的虚警检测被 AP 忽略或者称为被 AP 隐藏. 其次, 我们发现在高召回率区域, AP 对于准确度的增量和对于召回率的增量偏好不同, 且偏向于召回率增加, 这意味着检测算法增加虚警对增加 AP 值更有益. AP 值评价指标的缺陷有很大的危害, 首先导致该领域研究存在人为反复调整阈值从而获得高 AP 值的可能, 其次能够增强物体检测数量准确性的研究方法被长期抑制, 阻碍了该领域的创新发展, 最后导致追求高 AP 值的学术研究与追求数量准确性的实际应用严重分离, 很多学术成果在实际运用中无法达到应用的要求. 基于上述分析, 我们呼吁对 AP 指标进行批判式的改进, 以弥合学术研究和实际需求的鸿沟, 促进技术创新和产业运用良性互动.

关键词 物体检测, 平均准确度, 数量准确性, 评价指标, 深度学习

1 引言

近年来随着深度学习技术在计算机视觉的图片分类^[1~8]和物体检测^[9~22]领域取得突破性进展, 提升了人们对人工智能技术在实际应用中的期望. 比如在视频监控领域, 计算机自动对室内视频图像进行人体检测, 可以及时发现异常情况, 大幅度降低监控值班人员工作强度. 这项工作可以概括为对室内视频图像自动进行人体检测, 需要报告室内人数、每个人的位置. 在对人体数量和位置进行准确检测的基础上, 可以对人个体身份进行确认、对人的个体行为进行检测, 进一步对人与人之间的关系进行分析. 所以基于计算机视觉的人工智能技术在智能监控应用领域具有非常广泛的应用前景. 根据

引用格式: 窦勇, 乔鹏, 靳若春, 等. 探究平均准确度 AP 指标的缺陷及其影响. 中国科学: 信息科学, 2019, 49: 1369–1382, doi: 10.1360/N112018-00133
Dou Y, Qiao P, Jin R C, et al. Exploring the defects of the average precision and its influence (in Chinese). Sci Sin Inform, 2019, 49: 1369–1382, doi: 10.1360/N112018-00133

图 1 (网络版彩图) **Faster R-CNN** 在 **Brainwash** 测试集上每幅图人头检测框数量和正确标注框数量差的分布
Figure 1 (Color online) Distribution of the difference between the head number of **Faster R-CNN** detection output and the head number of groundtruth in **Brainwash** testset

应用的需求, 物体检测是指给定图片, 准确地检测到图片中的物体, 准确性包括物体的类别、位置和数量等 [23].

但是在实际应用中采用平均准确度 (average precision, AP) [24] 衡量的多种主流物体检测算法, 远远没有达到实际应用的预期. 当前主流的通用物体检测方法和专门的人头检测方法, 一直是计算机视觉领域非常活跃的研究热点, 每年都有新方法不断刷新物体检测的性能. 其中绝大多数算法都采用 AP 值来衡量物体检测的性能, 例如 2017 年 CVPR 会议中的 41 篇文章, ICCV 会议中的 34 篇文章都采用了 AP 值来评价各类算法的性能. 以目前典型的物体检测算法 **Faster R-CNN** [11] 为例, 针对咖啡厅数据集 [25] 进行人头检测实验, 首先用训练集进行人头检测训练, 然后用测试集进行测试, 其 AP 指标可以达到 0.81, 是一个非常不错的性能. 但是将检测结果可视化出来后, 发现每张图像上除了一些人头检测不到的情况外 (漏检), 还有相当数量的检测框并没有框在人头上, 而是框在其他物体上 (虚检). AP 指标所反映的检测性能似乎与目标的数量准确性没有正相关关系 [23]. 如图 1, 横轴是检测框数量和正确标注框数量差, 纵轴是对应差值的图片数量. 在总计 484 张测试图片中, 只有 10 张图片完全准确地报告出了人头的数量和位置, 其余图片中均有不同程度的虚检或者漏检. 如果需要统计安全敏感区域中人头数量来辅助预警, 图 1 所示结果意味着仅有 10 帧的数量结果是可用的, 将数量差异放松到 5 个, 结果可用的帧也只有不到 190 帧. 使用其他物体检测模型, 如单次多框检测器 (single shot multibox detector, SSD) [14] 或者单次检测器 (you only look once, YOLO) [13], 通用物体检测数据集 VOC [24, 26, 27] 和 COCO [28], 以及车辆检测数据集 KITTI [29], 上述现象都普遍存在, 即 AP 指标所体现的物体检测性能和数量准确性并没有直接的关系.

更恶劣的情况是当检测方法能够提高物体数量检测准确性时, AP 指标反而有可能下降. 当前主流的物体检测方法, 都采用了深度卷积神经网络技术, 图 2 给出了两种主要技术路线. 这类方法大致分为基于候选框的 2 步检测方法 (图 2 中浅橙色中蓝色框 [9~12]) 和不基于候选框的 1 步检测方法 (图 2 中浅橙色中绿色框 [13, 14]). 以人头检测为例, 无论是哪个技术路线, 检测流程都是将检测图片输

图 2 (网络版彩图) 基于卷积神经网络的物体检测方法流程图

Figure 2 (Color online) Flow chart of object detection methods based on convolutional neural networks

入检测网络, 得到若干检测框 (检测框包含得分和框的位置及大小), 通过非极大抑制 (non-maximum suppression, NMS)^[30] 去除重叠的检测框, 得到最终检测结果. 为了消除漏检和虚检, 我们改进以 NMS 为代表的后处理部分 (本文旨在阐述 AP 指标存在的问题, 引起学术界的讨论. 对于改进方法的描述以及和相关工作的比较超出了本文的范畴, 故不做具体介绍). 令人惊讶的是, 我们观察到了一个很令人困惑的现象, 即改进方法去除了 90% 以上的虚检框, 但是其 AP 指标反而降低了, 即在 AP 指标下, 所做改进是无益的.

我们产生了一个疑问, 到底是我们的改进方法真的无益, 还是 AP 指标没有正确评价我们的改进方法呢? 我们倾向于认为我们的改进是有益的, 因此进一步分析了 AP 指标, 发现 AP 指标除了不能展现检测器检测能力的细节这一问题外^[31], 还存在设计上的缺陷, 即不能反映物体检测结果在数量上的准确性. 本文后续从实际案例入手, 具体分析了 AP 指标潜在的缺陷, 并在理论上证明了该缺陷的存在性, 进而分析 AP 指标的危害性.

2 物体检测评价方法

为了更好地分析 AP 是如何失效的, 本节结构如下. 首先, 参照 VOC2012^[27] 评测方法, 依次介绍评测中涉及的概念, 如真阳性 (TP)、假阳性 (FP)、准确度 (Precision)、召回率 (Recall) 和平均准确度 (AP). 然后, 简要回顾 AP 指标本身的提出与改进历史. 最后, 以物体检测为例, 回顾 AP 对于物体检测技术发展的作用.

2.1 基本概念

给定一个图片 I 作为测试样本, 其上包含若干个待检测物体. 采用人工方法, 对图片 I 进行标注, 每个物体被一个矩形框所标识 (GroundTruth), 该矩形框的信息包含正确的物体类别、位置和范围. 图片上所有物体框形成一个序列 B_{gt} , 称为物体框基准序列, 该序列表示了有关每个物体在图片上的正确信息. 针对一种物体检测算法, 采用图片 I 作为测试样本进行测试, 在实施一次测试后, 可以得到一个物体框检测序列 B_p , 其按照物体类别得分的降序进行排列, 如图 3 中序列, 得分如图中数字所示. 为了显示效果, 从上到下 4 个一组, 如此向后直到序列结束. 在此基础上, 采用如下过程, 计算该算法在此次测试中的物体检测性能.

(1) 匹配检测框序列 B_p 和人工标注框序列 B_{gt} , 得到真阳性检测和假阳性检测序列, 分别记为 TP_p 和 FP_p .

图 3 (网络版彩图) Brainwash 数据集 11-13-27000 检测框示例. 其中绿色星号离散点代表 TP, 红色圆圈离散点代表 FP. 蓝色线代表准确度曲线, 黑色曲线代表召回率曲线. 图下部四排数字为检测框对应的得分, 从上到下 4 个一组, 依次从左到右.

Figure 3 (Color online) Detection result of sample 11-13-27000 in Brainwash testset. Green stars indicate TPs. Red circles indicate FPs. Blue curve indicates the precision curve. Black curve indicates the recall curve.

(2) 计算给定 B_p 中前 i 个检测框时的准确度和召回率, 得到准确度和召回率序列, 分别记为 Precision_p 和 Recall_p .

(3) 以 Recall_p 为 x 轴, Precision_p 为 y 轴, 绘制 PR 曲线.

(4) 计算 PR 曲线与 x 轴围成的面积, 即为平均精确度, AP_p .

2.1.1 匹配检测框序列 B_p 和标注框序列 B_{gt}

根据规则, 依次确定 B_p 和 B_{gt} 之间的匹配情况. 匹配过程如下:

(1) 计算 B_p 中第 i 框 $b_{(p,i)}$ 和 B_{gt} 中框 $b_{(\text{gt},j)}$ 的重合度 IoU;

(2) 当满足如下判定条件时, 称框 $b_{(p,i)}$ 和框 $b_{(\text{gt},j)}$ 为匹配上的, 记为真阳性检测, 此时在 TP_p 序列的对应位置上标记为 “1”, 在 FP_p 序列的对应位置上标记为 “0”. 否则, 记为假阳性检测, 此时在 TP_p 序列的对应位置上标记为 “0”, 在 FP_p 序列的对应位置上标记为 “1”. 判定条件, 检测框重叠比例 IoU 最大的且大于阈值, 同时 $b_{(\text{gt},j)}$ 没有被 B_p 中 i 之前的框匹配过. 注意, B_{gt} 中的每个框被 B_p 中多个框匹配时, 仅第 1 次匹配为真阳性, 其他均视为假阳性, 也称为虚警. 后面论述中不加说明, 假阳性和虚警视为等同概念.

根据上面的策略遍历 B_p 后, 得到真阳性和假阳性序列, 分别记为 TP_p 和 FP_p . 如图 3 所示, TP_p 和 FP_p 序列情况如图中绿色星号和红色圆圈所示.

图 4 (网络版彩图) (a) 准确度 - 召回率曲线 (蓝色) 和插值后的曲线 (红色); (b) 11 点插值准确度 - 召回率曲线
Figure 4 (Color online) (a) Precision-recall curve (blue) and interpolated precision-recall curve (red); (b) interpolated precision-recall curve for 11 recall levels

2.1.2 Precision_p 和 Recall_p

根据 TP_p 和 FP_p , 计算给出 B_p 中前 i 时, 正确检测的数量和错误检测的数量, 然后按照如下定义, 计算当前的准确度和召回率如下:

$$\text{Precision}_{(p,i)} = \frac{\#\text{TP}_{(p,i)}}{\#\text{TP}_{(p,i)} + \#\text{FP}_{(p,i)}}, \quad (1)$$

$$\text{Recall}_{(p,i)} = \frac{\#\text{TP}_{(p,i)}}{\#B_{\text{gt}}}. \quad (2)$$

遍历 B_p , 得到准确度和召回率序列, 分别记为 Precision_p 和 Recall_p . 然后, 以召回率为 x 轴, 准确度为 y 轴, 按照 B_p 序列的顺序, 逐一添加检测框信息, 绘制出准确度 - 召回率曲线 (PR curves). 如图 3 中蓝色实线和黑色虚线所示, 根据准确度和召回率的定义, 可以看到, 随着正确检测数量的增加, 召回率随之增加. 但是准确度和正确检测和错误检测数量都有关.

2.1.3 计算平均准确度 AP_p

在确定了准确度 - 召回率曲线之后, 计算准确度 - 召回率曲线与 x 轴围成的面积, 即为平均精确度. 给定如图 4(a) 中蓝色线所示的准确度 - 召回率曲线, 不难发现, 准确度 - 召回率曲线不是光滑的, 而是可能存在锯齿形部分. Text REtrieval Conference (TREC) 中引入 AP 概念时, 提出对准确度 - 召回率曲线进行插值, 得到一个单调不增的曲线, 如图 4(a) 中红色曲线. 并明确了一些测试点的缺省值, 即添加召回率等于 1 和 0 的定义. 如果召回率等于 1 的准确度不存在, 准确度赋值为 0. 召回率等于 0 的准确度不存在, 准确度赋值为 0. 明确缺省值定义后, 从召回率等于 1 向召回率等于 0 依次计算, 遍历准确度和召回率序列.

$$\text{Precision}_{\text{interp},i} = \max(\text{Precision}_{\text{recall}@i}, \text{Precision}_{\text{recall}@i+1}), \quad (3)$$

图 5 (网络版彩图) (a) Brainwash 测试集时间戳为 11-13-27000 的样本的 Faster R-CNN 检测结果; (b) (a) 中检测结果的准确度 - 召回率曲线

Figure 5 (Color online) (a) Detection result of sample 11-13-27000 in Brainwash testset using Faster R-CNN; (b) the precision-recall curve of the detection result in (a)

其中 i 为 Precision_p 和 Recall_p 第 i 个元素下标. 基于插值的准确度 - 召回率曲线, TREC 使用 11 点插值准确度计算 AP, 如图 4(b). 这 11 点依次为召回率水平为 0.0, 0.1, ..., 1.0 时的插值准确度. VOC2005 评测物体检测方法时, 借鉴了 TREC 中的 AP 计算方法, 即 11 点插值 AP. VOC2012 提出了改进的 AP 计算方法, 并被沿用至今. 修改后 AP 计算方法为, 计算插值后的准确度 - 召回率曲线与 x 轴围成的面积. 在准确度不变对应的召回率区间内, 区间内的面积等于, 以召回率区间长度为宽, 对应的准确度为高的矩形面积. 注意, 插值后的准确度 - 召回率曲线是一个单调递减函数, 召回率区间内的准确度值唯一, 除了区间端点处. AP 为各个区间的矩形面积之和.

AP 是准确度和召回率之间的权衡. 理想情况下, AP 用于比较多种检测方法的性能, AP 大的好于 AP 小的方法. 其主要目的是, 能够对多种检测方法进行定量比较.

2.1.4 计算示例

以图 3 中样本为例, 具体说明一下上述计算过程. 物体检测器均为 Faster R-CNN. 检测框去重均为 NMS, 前置阈值 0.05, IoU 阈值 0.3, 后置阈值 0.0. 测试数据集均为 Brainwash 的测试集. 评测采用 VOC2012 方法, IoU 阈值为 0.5.

图 5 中标注框 B_{gt} 中框的数量 $\#B_{gt} = 18$ (图 5(a) 中绿色框), 检测序列 B_p 中框的数量 $\#B_p = 40$ (图 5(b) 中 PR 曲线上的数值为检测框的人头得分). 依次比较 B_p 和 B_{gt} 中的框, 得到 TP 序列为 [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0]. FP 序列为 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1]. 图 5 中白色为正确检测, 其数量为 $\#TP_p = 17$, 红色为错误检测, 其数量为 $\#FP_p = 23$, 漏检 $\#B_{gt} - \#TP_p = 1$. 根据 TP_p 和 FP_p , 计算得到 Precision_p 和 Recall_p , 绘制成图 5(b) 中所示的准确度 - 召回率曲线. 由于 B_p 中前 15 个检测框都是正确检测, 因此召回率线性增加, 准确度一直等于 1, 表现为图 5(b) 中的平行于 x 轴的直线段. 第 16 个检测框为错误检测, 此时正确检测还是 15 个, 错误检测 1 个, 召回率没有变化, 准确度降低, 表现为图 5(b) 中第 1 个垂直于 x 轴的沿 y 轴负方向的下降. B_p 中最后 21 个检测框均为错误检测框, 表现为图 5(b) 中垂直于 x 轴沿 y 轴负方向的连续下降. 根据 VOC2012 的计算规则, 计算准确度 - 召回率曲线和 x 轴围成的面积, 即为 $AP = 0.933$.

2.2 AP 指标的发展

AP 最早是被用于评价文本检索结果. 后来被 VOC 数据集提出方采用, 来评价物体检测结果.

2.2.1 文本检索中使用的 AP

在文本检索中, 检索方法会给出每个检索结果和被检索主题之间的相关度得分, 因此检索结果是一个基于得分序的检索集合. 为了更好地评价基于得分序的检索集合, 文本检索会议 (TREC) 使用 AP 作为指标.

该方法将检索结果根据得分降序排列, 依次判断每个检索结果和主题是否相关. 如果相关, 召回率和准确度增加; 否则, 召回率不变, 但是准确度降低, 如图 4(a) 中蓝色曲线. 图中曲线有很多锯齿形部分. TREC 对准确度 - 召回率曲线插值后, 得到如图 4(a) 中红色曲线. 关于插值的合理性, 文献 [32] 中给出的解释为 “The justification is that almost anyone would be prepared to look at a few more documents if it would increase the percentage of the viewed set that were relevant (that is, if the precision of the larger set is higher)”. 即为了看到更多相关的检索结果, 我们不介意看更多几页检索结果. 前提是多看一些检索结果时准确度更高一些. 就是说, 召回率比准确度重要, 隐藏不相关的检索结果是 AP 应有之意.

2.2.2 物体检测中的 AP

VOC2005 物体检测比赛中, 检测方法给出一个检测框序列, 其中每个检测框除了具有位置和大小外, 还具有属于某个类别的置信度得分. 检测框序列可以看成是一个基于得分序的检测结果. 因此, VOC2005 物体检测比赛采用 TREC 中提出的 AP 来评价物体检测方法.

2.3 AP 对于物体检测技术发展的历史作用

如图 6, 我们回顾了 VOC 测试集中 “person” 类别的准确度 - 召回率曲线的变化. 在 2005 年 VOC 比赛刚开始的时候, 物体检测方法的性能很差, 召回率不高, 准确度更是无从谈起, 如图 6 中的标记为 VOC2005 的蓝绿色曲线. 随着技术的发展, 特别是梯度直方图 (histograms of oriented gradients, HOG) 描述子 [30] 和形变部件模型 (deformable part model, DPM) [33] 等模型的使用, 检测方法的性能得到了大幅提升, 如图 6 中的深蓝色和黄色曲线, 分别是 VOC2007 和 VOC2012 的准确度 - 召回率曲线. 虽然检测方法在召回率上没有提升, 但是准确度得到了很大的改善.

2012 年 ImageNet 比赛中, 深度网络获得了较传统方法更好的性能提升, 实现了图片分类的突破性进展. 之后基于深度网络的物体检测方法也开始出现, 如 R-CNN, Fast R-CNN, Faster R-CNN, SSD 和 Yolo 等. 通过观察 R-CNN, Fast R-CNN, Faster R-CNN 的准确度 - 召回率曲线, 如图 6 中的绿色、蓝色和紫色曲线, 不难发现基于深度网络的检测方法的性能提升来自于两个主要方面, 一方面得益于特征的提升, 一方面得益于候选框方法的提升, 如 2013 年出现的 selective search [34] 和 2015 年的 RPN [11].

3 分析 AP 评价指标的缺陷

本节如无特殊说明, 物体检测器和数据集等设置同第 2 节. 本章以 Brainwash 数据集为代表的类别物体检测问题为例, 展开论述. 所得到的证明过程和结论同样适用于分析 AP 在 VOC [24, 26, 27] 和 COCO [28] 等多类别检测数据集.

图 6 (网络版彩图) VOC 测试集中“person”类别历年准确度 - 召回率趋势曲线
Figure 6 (Color online) The precision-recall curve of person class in VOC testset in recent years

3.1 AP 评价指标可能掩盖虚警, 不能准确刻画数量准确性

本小节先从两个示例入手, 说明虚警和 AP 的计算方法对于 AP 的影响, 试图说明 AP 和数量准确性之间不是严格对应的, 即 AP 不能准确刻画数量准确性. 然后形式化的推导虚警对于 AP 的影响.

首先介绍一个人工构造的案例: 假设有两个检测方法, 分别为 A 和 B, 对同一图像进行人头检测, 方法 A 检测结果如图 7(a) 所示, 正确标注的为绿圆框总计有 6 个, 方法 A 正确标记了 5 个用黑色方框表示, 还有 4 个红色的六边形框为虚警框. 图 7(b) 为方法 B 检测的结果, 正确标记了 4 个目标用黑色方框表示, 有 2 个红色的六边形虚警框.

首先, 在该示例中虽然虚警会改变准确度 - 召回率曲线的性状, 但是出现虚警不一定影响 AP 值. 如图 7(c) 中方法 A 对应的准确度 - 召回率曲线, 第 3 个检测框是一个虚警. 此时正确检测的数量还是 2, 因此 Recall 不变. 但是错误检测数量为 1, 准确度从 $1 (\frac{2}{2})$ 降低到 $0.667 (\frac{2}{3})$. 考虑前 3 个检测框, 召回率等于 $0.333 (\frac{2}{6})$ 时对应两个准确度, 即 1 和 0.667. 根据插值后的准确度 - 召回率曲线, 计算得到 $AP = 0.4$, 第 3 个框产生虚警并没有影响 AP. 虚警出现时召回率不变, 准确度降低, 准确度 - 召回率曲线平行准确度轴向其负方向移动. 此时出现一个召回率对应多个准确度的情况, 根据 AP 的定义, AP 仅和该召回率对应的最大准确度有关. 结果就是, 使得准确度降低的虚警检测被 AP 忽略或者称为被 AP 隐藏.

其次, AP 计算中的插值方法也会隐藏虚警, 进一步弱化了虚警对 AP 的影响. 如图 7(c) 中蓝色虚线, 还是分析方法 A 对应的准确度 - 召回率曲线. 第 3 个框是虚警, 第 4 和 5 个框是正确检测, 如果直接用准确度 - 召回率虚线计算线下面积, 前 5 个检测框的 $AP = 0.58$. 采用插值后的准确度 - 召回率曲线, 前 5 个检测框的 $AP = 0.60$. 按照 AP 计算方法, 方法 A 和 B 的 AP 依次为 0.69 和 0.67. 从 AP 看两种方法难分上下, 但是方法 A 虚警了 4 个框, 漏检了 1 个框. 方法 B 虚警了 2 个框, 漏检了

图 7 (网络版彩图) AP 计算过程 (VOC2012) 隐藏虚警示意图. (a) 和 (b) 分别是方法 A ($AP = 0.69$) 和 B ($AP = 0.67$) 的检测结果. 绿色圆圈表示实例, 方框表示检测结果. 其中黑色表示正确检测, 红色表示错误检测 (即虚警). (c) 为准确度 - 召回率曲线, 蓝色虚线为 AP 计算时平滑后的准确度 - 召回率曲线轮廓.

Figure 7 (Color online) Diagram of how AP hides the false alarms in the computation process. (a) and (b) show the detection results of method A ($AP = 0.69$) and B ($AP = 0.67$) respectively. Green circles indicate the instances. Boxes indicate the detection results. The black boxes indicate the true positives, while the red ones indicate the false positives (i.e., the false alarms). (c) Precision-recall curve. Blue dotted line is the interpolated precision-recall curve used in AP calculation.

2 个框.

在人工合成示例分析的基础上, 再看 2.1.4 小节中所展示的真实示例. 从 AP 值看, 0.933 不算低, 但是虚警数量高达 23 个, 标注框的数量 $\#B_{gt}$ 才 18 个. 综合 Brainwash 数据集上 Faster R-CNN 的 AP 指标, 以及图 1 所示的数量准确性的统计, 上述情况在 Brainwash 数据集中大量存在. 注意, 这一现象同样会发生在其他物体检测模型, 以及其他物体检测数据集.

在上述实例分析的基础上, 我们发现了一个规律, 即在某种特定情况下, AP 值不能刻画物体检测数量的准确性. 我们可以将其形式化为, 存在注入位置 i^* , 使得无论注入虚警数量 m 如何变化, AP 值保持不变. 因此 AP 指标存在缺陷, 不能刻画数量准确性.

3.2 AP 偏好召回率, 高召回率比高精度更能提高 AP 值

我们发现在高召回率区域, AP 对于准确度的增量和对于召回率的增量偏好不同, 且偏向于召回率增加. 这意味着检测算法提高召回率对增加 AP 值更有益. 为了说明这一点, 我们以 2.1.4 小节中的示例为基础, 对 B_p 的人头得分进行阈值处理来控制虚警. 取人头得分阈值为 0.8, 记后处理后的检测框序列为 B_{p1} , 其中检测框数量 $\#B_{p1} = 18$. 得到 TP 序列为 [1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0].

FP 序列为 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1]. 正确检测的数量为 $\#TP_{p1} = 16$, 错误检测的数量为 $\#FP_{p1} = 2$, 漏检 $\#B_{gt} - \#TP_{p1} = 2$. 根据 TP_{p1} 和 FP_{p1} 序列, 计算 $Precision_{p1}$ 和 $Recall_{p1}$, 最后得到 $AP = 0.886$.

对比“后处理”操作的前后结果, 准确度从 0.425 提升到 0.889, 提高了 109%. 召回率从 0.944 降低到 0.889, 降低了 6%. 但是, AP 值从 0.933 降低到 0.886, 降低了 5%. 从案例分析可见, 在高召回率区域, 大幅度提升准确度的检测方法与提升召回率方法相比, 提升召回率对提高 AP 值的作用更大.

综合图 6 中物体检测方法的准确度 - 召回率曲线的变化, 不难发现, 在基于深度网络的物体检测方法的推动下, 物体检测进入了高召回率区域. 如上示例分析, 我们发现在高召回率情况下, 提升召回率和提升准确度带来的 AP 提升不同. 根据分析, 我们发现在高召回率情况下, AP 对召回率的提升有明显的偏向. 因此, 在 AP 指导下, 仅仅提高准确度, 没有提高召回率的技术创新在 AP 值方面没有突出表现, 几乎很难发表在好的会议或者期刊上. AP 的偏好对于学术工作进一步表现出了逆向选择性.

3.3 即使 $AP = 1$, 也不意味着数量准确

假设存在检测器 $D1$, 其给出的检测序列 B_{p1} 可以覆盖所有 B_{gt} , 并且对于所有的 B_{gt} 给出的得分都很高. 在 B_{p1} 尾部存在若干低分检测框.

从召回率和准确度的定义看, 召回率和检测器给出的检测框数量无关, 仅和其中真阳性检测数量和正确标注目标数量有关. 同时, 准确度和检测器给出的检测框数量和真阳性检测数量有关. 当上述情况出现时, B_{p1} 中的低分检测框虽然在拉低准确度, 但是他们对对应同一个召回率, 即召回率等于 1. 根据 3.1 小节的证明, 在计算 AP 时, 这一部分虚警不影响 AP, 因此 $AP = 1$. 因此, 即使 $AP = 1$, 也不意味着数量的准确性是 100%.

本小节我们结合示例和形式化的证明, 从 AP 会隐藏虚警导致 AP 不能刻画数量准确性, AP 在高召回区域偏向于提高召回率, $AP = 1$ 也不意味着数量准确这 3 个方面, 展示了 AP 存在的问题. 综合这 3 种情况, 我们认为 AP 不能反映检测方法在数量准确性上的表现. 导致 AP 对数量准确性不敏感的原因主要归因于 AP 的定义、计算方法.

4 AP 缺陷的来源及危害

4.1 采用 AP 方法作为评价指标是一个历史发展的过程, 具有其本身固有的局限性

结合文本检索的任务看, AP 的设计初衷是合理而且有益的, 前提是检索结果的数量不是我们核心关注的评测指标. 很显然, 物体检测应用对数量准确性是有要求的, 并不是像上面文本检索那样, 只是覆盖更多我们关注的主题这么简单. 虽然 AP 在被引入到物体检测之初, 在推动领域发展上起到了积极的作用, 但是从物体检测的长远发展看, 尤其是在评价数量准确性方面, AP 设计上的不足和不适用显现得越来越明显. 因此, 我们认为, 需要根据物体检测的任务需求, 即类别、定位、数量的准确性, 重新设计评价指标.

4.2 现阶段仍然采用 AP 方法作为评测物体检测算法指标, 会误导研究方向

如图 2 所示, 在物体检测框架中, 检测方法总体分为 3 大阶段. 很多文章聚焦在第 1 和 2 阶段. 第 3 阶段通常采用 NMS 方法, 通过设置阈值, 抑制掉多余的重叠框, 达到控制虚警的目的. 在以 AP 值为评价指标的情况下, 这个阶段存在人为控制的因素, 不能反映技术的进步.

首先, NMS 阶段需要设置一个判断阈值, 研究人员可以调整这个阈值有选择地报告高 AP 值结果. 以 2.1.4 小节中的数据集和检测器为例. 在 Brainwash 数据集上完成调优后, 在验证集上对前置阈值进行选择, 然后再在测试集上测试. 在验证集上调整前置阈值时, 研究人员会带有主观感情, 即提高 AP 值. 阈值选低一些, NMS 留下来的检测框多一些, 召回率可能会提高. 阈值选高一些, NMS 留下来的检测框少一些, 准确度虽然会高一些, 但是召回率会降低. 结合 3.2 小节中的分析, AP 的偏好会影响研究人员对于前置阈值的选择偏好. 研究人员在论文中倾向于报告经过反复调整阈值得到的 AP 值, 而不会报告其选择参数的过程.

在评价物体检测算法的适应性过程中, 针对不同的数据集, 研究人员可以在 NMS 阶段选择不同的判断阈值, 报告各自最好的 AP 值. 将 NMS 阶段的研究变成了一个技巧性工作, 或者是投机取巧性的工作, 导致研究人员忽略了 NMS 阶段的本质学术问题. 再者, AP 值有倾向性, 有效改进 NMS 效果的技术不一定能够提高 AP 值, 导致这些方法无法体现其优越性. 这个方向的研究一直没有得到应有的重视和实质性提高. 虽然文献 [23] 中指出了使用 AP 评价技术准确性存在的问题, 并提出了新的指标来评价计数准确性, 但该指标没有惩罚虚警. 因此, 该指标也不能准确评价计数准确性. 我们可以看到有关第 2 阶段的检测网络技术从 2014 年 R-CNN 之后, 有了突飞猛进的发展, 论文分布在 CVPR, ICCV, NIPS, ECCV 等领域内顶级会议. 相对的, 第 3 阶段工作 greedyNMS^[30] 方法从 2005 年被用于物体检测就没有被改动过, 仅有的几个相关工作^[15] 也没有被主流研究采用. 最近, 如 CVPR2017 的 learning NMS^[35]、CVPR2018 的 relation networks^[36]、ICCV2017 的 softNMS^[37] 等少数工作, 才对 NMS 过程有所涉及.

最后, 追求高 AP 值的学术研究与追求高数量准确性的实际应用严重分离. 由于 AP 设计上的问题, 导致其不能刻画数量准确性, 而这是在实际应用中很重要的指标. 比如, 安全敏感区域的监控, 我们需要根据检测结果, 尤其是物体、人员数量, 进行决策. 再比如, 无人超市, 我们需要根据检测结果, 不仅是物品的种类, 还有物品的数量, 进行购买行为的结算、供货方面的决策等. 尽管实际应用有着数量准确性的高要求, 但是从目前的学术研究看, 追求高 AP 似乎是一个普遍被接受的概念. 论文报告的工作如果 AP 值比不过 baseline, 那么基本可以断定该工作是不值得发表的. 而如本文第 3 节论述的, AP 不能刻画数量准确性, 且存在偏好等缺陷, 这样会使得学术研究越来越偏离实际需求.

5 总结

对 AP 指标的质疑源于我们在实际工作中运用物体检测算法进行人头数量统计的实践. 我们发现尽管我们试用了顶级会议的多种最新学术成果, 但是其在统计数量准确性方面与实际需求差距甚远. 为此, 我们深入分析了 AP 指标的来源、定义和计算方法, 发现在最初引进 AP 指标时, 其本身就有当时的历史局限性. 进一步我们发现 AP 指标在特定区域对于虚警不敏感, 甚至虚警有益于 AP 值提升. 对于 AP 指标的这些缺陷虽然在业界大家都习以为常, 甚至出现刷 AP 值的学术研究, 但是随着物体检测算法逐步走向实际应用, 对于数量、类别和位置的准确性愈发重要, AP 指标虽然之前对于技术发展有着积极作用, 但是如本文分析的, 在当前技术情况下, AP 已经成为阻碍技术进步的重要因素. 一方面它误导学术研究的发展方向, 掩盖了真实的挑战问题, 另一方面导致学术研究脱离实际需求. 因此, 我们认为非常有必要对 AP 指标进行批判式的改进, 以弥合学术研究和实际需求的鸿沟, 让技术转化成更好的产品服务大众, 同时让技术创新本身得到更好的发展.

参考文献

- 1 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521: 436–444
- 2 Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. In: *Proceedings of Advances in Neural Information Processing Systems*, 2012. 1097–1105
- 3 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
- 4 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 770–778
- 5 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015
- 6 Chen Y P, Li J N, Xiao H X, et al. Dual path networks. In: *Proceedings of Advances in Neural Information Processing Systems*, 2017. 4467–4475
- 7 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018
- 8 Sabour S, Frosst N, Hinton G. Dynamic routing between capsules. In: *Proceedings of Advances in Neural Information Processing Systems*, 2017. 3856–3866
- 9 Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 580–587
- 10 Girshick R. Fast R-CNN. In: *Proceedings of IEEE International Conference on Computer Vision*, 2015. 1440–1448
- 11 Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Proceedings of Advances in Neural Information Processing Systems*, 2015. 91–99
- 12 Dai J F, Li Y, He K M, et al. R-FCN: object detection via region-based fully convolutional networks. In: *Proceedings of Advances in Neural Information Processing Systems*, 2016. 379–387
- 13 Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 779–788
- 14 Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: *Proceedings of European Conference on Computer Vision*, 2016. 21–37
- 15 Wan L, Eigen D, Fergus R. End-to-end integration of a convolution network, deformable parts model and non-maximum suppression. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 851–859
- 16 Cao J L, Pang Y W, Li X L. Learning multilayer channel features for pedestrian detection. *IEEE Trans Image Process*, 2017, 26: 3210–3220
- 17 Lin T, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2117–2125
- 18 Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017
- 19 He K M, Gkioxari G, Dollár P, et al. Mask R-CNN. In: *Proceedings of IEEE International Conference on Computer Vision*, 2017. 2980–2988
- 20 Wang W G, Shen J B, Shao L. Video salient object detection via fully convolutional networks. *IEEE Trans Image Process*, 2018, 27: 38–49
- 21 Wang W G, Shen J B, Yang R G, et al. Saliency-aware video object segmentation. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40: 20–33
- 22 Wang W G, Shen J B. Deep visual attention prediction. *IEEE Trans Image Process*, 2018, 27: 2368–2378
- 23 Chen S, Fern A, Todorovic S. Person count localization in videos from noisy foreground and detections. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1364–1372
- 24 Quiñonero-Candela J, Dagan I, Magnini B, et al. *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*. Berlin: Springer, 2006
- 25 Stewart R, Andriluka M, Ng A. End-to-end people detection in crowded scenes. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2325–2333
- 26 Everingham M, van Gool L, Williams C, et al. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>
- 27 Everingham M, van Gool L, Williams C, et al. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

- <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>
- 28 Lin T, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of European Conference on Computer Vision, 2014. 740–755
 - 29 Andreas G, Philip L, Raquel U. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012. 3354–3361
 - 30 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005. 886–893
 - 31 Hoiem D, Chodpathumwan Y, Dai Q Y. Diagnosing error in object detectors. In: Proceedings of European Conference on Computer Vision, 2012. 340–353
 - 32 Christopher D, Prabhakar R, Hinrich S. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008
 - 33 Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008
 - 34 Uijlings J R R, van de Sande K E A, Gevers T, et al. Selective search for object recognition. *Int J Comput Vis*, 2013, 104: 154–171
 - 35 Hosang J, Benenson R, Schiele B. Learning non-maximum suppression. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4507–4515
 - 36 Hu H, Gu J Y, Zhang Z, et al. Relation networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3588–3597
 - 37 Bodla N, Singh B, Chellappa R, et al. Improving object detection with one line of code. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 5561–5569

Exploring the defects of the average precision and its influence

Yong DOU*, Peng QIAO & Ruochun JIN

Science and Technology on Parallel and Distributed Laboratory, National University of Defense Technology, Changsha 410073, China

* Corresponding author. E-mail: yongdou@nudt.edu.cn

Abstract The average precision (AP) is an important evaluation index for object detection algorithms in computer vision and has long been used for quantitative comparisons of relevant research results. In practice, we found that the AP index has defects, i.e., it does not accurately measure the quantity accuracy of object detection algorithms. We particularly analyze two manifestations of the AP defects. First, in the AP definition and calculation, there is no change in the recall rate when there is a false alarm and the precision rate decreases. Consequently, one recall rate may correspond to multiple precision rates. The AP value only corresponds to the maximum precision at a certain recall rate; therefore, false alarms reducing the precision are ignored or hidden by the AP. Second, in a high recall rate area, the AP is more sensitive to the increase in the recall rate than that in the precision rate, indicating that AP may favor detection algorithms that increase the number of false alarms. AP defects are very harmful. First, there is a possibility that researchers will artificially adjust the threshold value to obtain a high AP value. Second, a research method that can improve the object detection quantity accuracy may be suppressed, which will hinder development in the field. Finally, this situation may result in a serious divergence between academic research pursuing high AP values and practical applications that require quantitative accuracy. This may result in a failure to use many academic studies in practical applications. Based on the above analysis, we call for a critical improvement in the AP index to bridge the gap between academic research and practical applications, promoting positive interactions between technological innovation and industry use.

Keywords object detection, average precision, quantity accuracy, evaluation metric, deep learning

Yong DOU was born in 1966. He received his Ph.D. degree in computer science and technology from the National University of Defense Technology, Changsha, in 1995. Currently, he is a professor and Ph.D. supervisor in the School of Computer at the National University of Defense Technology. His research interests include artificial intelligence, deep learning, high performance computer architecture, and reconfigurable computing.

Peng QIAO was born in 1988. He is now pursuing his Ph.D. degree in computer science and technology at the National University of Defense Technology. His research interests include computer vision, image processing, machine learning, and deep learning.

Ruochun JIN was born in 1993. He is a Ph.D. student in the School of Informatics at the University of Edinburgh. His research interests include databases, machine learning, and deep learning.