计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.008612]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 YOLOx 残差块融合 CoA 模块的改进检测网络①

安鹤男[1,2], 杨佳洲[1], 邓武才[2], 管　聪[1], 马　超[1]

[1](深圳大学 微纳光电子学研究院, 深圳 518054)
[2](深圳大学 电子与信息工程学院, 深圳 518060)
通信作者: 杨佳洲, E-mail: yangjiazhou2016@email.szu.edu.cn

摘　要: YOLOx-Darknet53 是以 YOLOv3 为基准增加各种技巧 (trick) 升级改进的检测网络, 但其仍然是以 Darknet53 为特征提取骨干网络 (backbone), 因此网络的特征提取能力仍有欠缺. 本文依据 CoTNet 中的注意力机制改进得到 CoA (contextual attention) 模块, 并将其替代 YOLOx 骨干网络残差块里的 3×3 卷积, 得到融合注意力后的新残差块, 加强了骨干网络的特征提取能力, 并在 Pascal VOC2007 数据集上进行对比实验, 融合 CoA 模块的网络比原网络的平均精度均值 AP@[.5:.95] 高 1.4, AP@0.5 高 1.4; 在改进骨干网络后的 YOLOx 检测头前加入无参 3D 注意力模块, 得到最终改进的检测网络, 进行上述对比实验, 结果表明比原网络的 AP@[.5:.95] 高 1.6, AP@0.5 高 1.5. 因此, 改进后的网络比原网络检测更加精准, 在工业应用中能达到更好的检测效果.

关键词: YOLOx; 骨干网络; 残差块; CoA 模块; 3D 注意力; 深度学习; 目标检测

引用格式: 安鹤男,杨佳洲,邓武才,管聪,马超.基于 YOLOx 残差块融合 CoA 模块的改进检测网络.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/8612.html

Improved Detection Network Based on YOLOx Residual Block Fusion CoA Module

AN He-Nan[1,2], YANG Jia-Zhou[1], DENG Wu-Cai[2], GUAN Cong[1], MA Chao[1]

[1](Institute of Microscale Optoelectronics, Shenzhen University, Shenzhen 518054, China)
[2](College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China)

Abstract: YOLOx-Darknet53 is an improved detection network integrating a basis of you only look once version 3 (YOLOv3) with various tricks added. Nevertheless, it still uses Darknet53 as the backbone network to extract features, so the feature extraction capability of the network is still insufficient. In this paper, we acquire a contextual attention (CoA) module by improving the attention mechanism in CoTNet and replace the 3×3 convolution in the residual block of the YOLOx backbone network with the module to obtain a new residual block after attention fusion and thereby strengthen the feature extraction capability of the backbone network. A comparison experiment is conducted on the Pascal VOC2007 data set. The mean average precision AP@[.5:.95] and the AP@0.5 of the network integrating the CoA module are both 1.4 higher than those of the original network. After the backbone network is improved, a non-parameter 3D attention module is added in front of the YOLOx detection head to obtain the final improved detection network. The results of another round of the above comparative experiment show that the AP@[.5:.95] and the AP@0.5 of the final network are respectively 1.6 and 1.5 higher than those of the original network. Therefore, the improved network is more accurate than the original network in detection and can achieve better detection effects in industrial applications.

Key words: YOLOx; backbone network; residual block; contextual attention (CoA) module; 3D attention; deep learning; object detection

① 收稿时间: 2021-10-30; 修改时间: 2021-11-29; 采用时间: 2021-12-13; csa 在线出版时间: 2022-04-18

随着计算机硬件设备以及各种深度学习框架的发展, 深度学习模型成为了当下人工智能领域研究的重要途径. 属于人工智能领域热点之一的计算机视觉, 其主要任务包括目标分类、目标检测、目标分割等, 其中不同时期内视觉图像处理的侧重点也有所不同. 随着卷积神经网络 (CNN) AlexNet[1] 在分类任务大放异彩, 再到 2015 年深度残差网络 ResNet[2] 诞生并一举刷新当时多项视觉任务的最佳, 标志着带有残差块的更深的卷积神经网络时代的到来, 随后计算机视觉任务的重点也从分类转到了检测和分割.

目标检测任务是目前计算机领域的一大热点, 主要就是找出图像中所有我们感兴趣的目标, 并确定他们的类别和位置大小, 即解决物体是什么和在哪里这两个问题. 如今, 凭借着大量可用数据集、快速先进的 GPU 以及更好的算法框架, 我们可以轻松用深度学习模型训练计算机以高精度检测出图像里的诸多目标. 目前主流的 CNN 目标检测算法主要分为两大类: 一种是以 RCNN[3] 和 Faster-RCNN[4] 等为主要代表的两阶段检测算法 (two-stage), 另一种是以 SSD[5] 及 YOLO[6] 系列等为代表的单阶段检测算法 (one-stage). Two-stage 算法多是先生成大量的候选瞄框 (anchors), 再进行分类和定位, 尽管在 Faster R-CNN 中, 生成瞄框网络 RPN (region proposal network) 和 CNN 分类定位网络融合在一起, 无论在速度上还是精度上都得到了不错的提高, 然而 Faster-RCNN 还是达不到实时的目标检测, 其中预先获取候选框, 然后再对每个候选框分类定位的计算量还是比较大. 而 one-stage 算法使用端到端的网络进行目标检测, 只用一个 CNN 网络便可实现目标的分类和定位, 在检测速度上更胜一筹, 其中的 YOLOv3[7] 系列算法兼具了速度和不错的精度, 因此在工业界都有广泛应用[8]. YOLOx-Darknet53[9] 则是在 YOLOv3 的骨干网络 (backbone) 基础上使用大量先进技巧 (trick), 因此比原网络有更高的检测精度, 更加兼具了工业应用中的精度和速度要求.

随着注意力机制[10] 在计算机视觉领域的广泛应用, 以及考虑到 YOLOx-Darknet53 骨干网络的特征提取能力仍有不足的问题, 本文结合 CoTNet[11] 的思想, 重新设计了 CoA (contextual attention) 模块, 并将其融合在 Darknet53 中的残差块, 这加强了骨干网络的特征提取能力; 并依据注意力模块 SimAM[12] 对检测头前的特征层加强空间和通道 (spatial and channel) 的关注, 即实现 3D Attention. 综上, 本文设计了一个全新的 YOLOx 检测网络, 改进后的网络具有更强的特征提取能力和目标检测能力, 对小中大的目标检测精度都有提升, 更加能满足工业界对不同大小物体检测精度的要求.

1 YOLOx-Darknet53 算法

YOLOx-Darknet53 算法[9] 就是在 YOLOv3 骨干网络的基础上进行消融实验, 尝试各种优化 trick 最终改进得到的优化算法. 主要的优化点有输入端的数据增强、检测端的解耦头 (decoupled head), 以及预测框的检测方式从 Anchor-based 变为 Anchor-free 等.

输入端通过 Mosaic 加上 MixUp[13] 的数据增强方式, 对训练集中随机抽取的图片进行随机裁剪、随机缩放、随机排布等方式进行拼接, 生成新融合的图片. 由于融合图片的复杂性和多目标性, 且一定程度上相当于扩大了 batch size, 因此网络模型可以不通过 ImageNet[14] 数据集预训练和迁移学习, 仅根据自有数据集从头开始训练便能达到很好的检测效果, 这在提升网络检测精度的同时也增强了网络的鲁棒性和泛化能力.

YOLOx-Darknet53 还将原来的 YOLO-head 改为 decoupled head (YOLOx-Head), 将原来由单个卷积统一得到的检测信息改为 3 个分支输出, 实现了将置信度 (conf)、瞄框偏移量 (x, y, w, h)、类别 (class) 信息解耦, 让分类和回归更加精确. 作者在权衡精度和速度后改进解耦头如图 1, 在仅增加一点参数的同时使得网络收敛速度更快, 精度也更高一些. YOLOx-Darknet53 网络的损失值也由 3 部分组成, 分别是目标置信度损失 (*loss_obj*)、目标定位损失 (*loss_iou*) 以及目标类别损失 (*loss_cls*). 网络的总损失 *Loss* 组成关系如下:

$$Loss=\alpha_1 loss\_iou+\alpha_2 loss\_obj+\alpha_3 loss\_cls \quad (1)$$

其中, α_1、α_2、α_3 为平衡系数 (即可自己调节的超参数, 原网络采用的是 5, 1, 1, 改进网络与原网络保持一致, 未对损失函数进行修改). 定位损失采用的是 IOU 损失函数, 目标损失和置信度损失均采用二值交叉熵损失函数 (binary cross entropy, BCE).

除此之外, 预测框的检测方式从 Anchor-based 变成了 Anchor-free, 大大减少了生成预测框数量及所需参数, 并结合标签分配中的初步筛选、SimOTA 方法[15] 将预测框和目标框关联起来, 筛选出更加合适的正样本进行训练, 不断的预测迭代更新参数, 让网络的检测

结果更加准确. 但这些 tricks 大多是对输入端和检测端进行改进, 并没有改变网络本身的骨干, 没有加强骨干网络的特征提取能力, 因此本文将从这一方面做主要改进, 进一步提升网络对不同大小目标的检测效果.

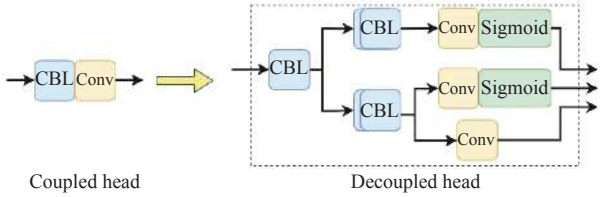

图 1 改进后的 Yolox-Head

2 方法与改进

2.1 CoA 模块

随着 Transformer[16] 开启注意力机制以来, 越来越多研究者将注意力结构迁移应用在计算机视觉任务中, 但大多还是通过 Self-Attention 中的查询 (query) 和键值 (key) 的交互来获得注意力矩阵, 并未考虑相邻 key 之间的丰富信息, 其大致结构如图 2 所示. 而 CoTNet (contextual transformer network)[11] 设计了新的注意力块, 充分利用输入键之间的上下文信息来生成动态注意力矩阵, 从而增强视觉表示能力. 本文的 CoA (contextual attention) 正是基于 CoTNet 改进的注意力模块, 其在 Darknet53 残差块中的结构如图 3 所示.

在 CNN 中常用的 Self-Attention 模块中 keys, query, value 都是由上一层的 input 通过 1×1 卷积映射而来的. 由于 Self-Attention 对输入特征的位置信息并不敏感, 所以也会在 query 层增加一个位置 (position) 信息, 然后与 query 和 keys 交叉生成的信息相互融合, 再经过归一化 (Softmax) 操作得到概率分布, 最后和 value 矩阵相乘得到注意力矩阵. 其中主要的注意力权重来自于 query-key pair 之间的交互, 而特征层 keys 由 1×1 卷积生成, 并没有包含 input 中丰富的相邻键特征, 没能充分利用 keys 之间的上下文信息.

在 CoA 模块中, input 经过 3 个不同的卷积块抽取出 keys, query 和 value.先采用 3×3 的卷积先对 input 的相邻键进行上下文编码获得的静态上下文特征 keys, 再用不同的 1×1 卷积提取特征层 query 和 value. 不同于 CoT 模块, 其中特征层 query 采用 1×1 卷积进行特征提取, 接着将有上下文特征的 keys 和加强特征层 query 拼接 (concat), 再通过两个连续的 1×1 卷积块并利用 Softmax 操作将结果归一化为概率分布, 得到注意力权重矩阵 A, 此时的 A 中每个空间位置都考虑了 query 和 keys 的全局特征. 权重矩阵 A 与 value 进行矩阵相乘后可以得到拥有更强注意力机制的全局动态特征层, 最后还要和静态上下文特征层 keys 进行融合, 融合之后再进行 1×1 卷积达到跨通道信息交融和升维目的, 最终得到了对突出点着重关注、同时不缺失对普通点关注的注意力输出特征层 (output). 公式如下:

$$output = Conv(keys + A \odot value) \tag{2}$$

CoA 中的每个卷积 (Conv) 都包含普通卷积 Conv2d、批量归一化 BN、激活函数 *Mish*[17]. *Mish* 是 2019 年提出用来代替 ReLU 函数的新型激活函数, 其函数表达式如下:

$$Mish = X \times \tanh(\ln(1 + e^{X})) \tag{3}$$

其中, X 是前面归一化层传过来的参数值. *Mish* 激活函数的图像如图 4 所示.

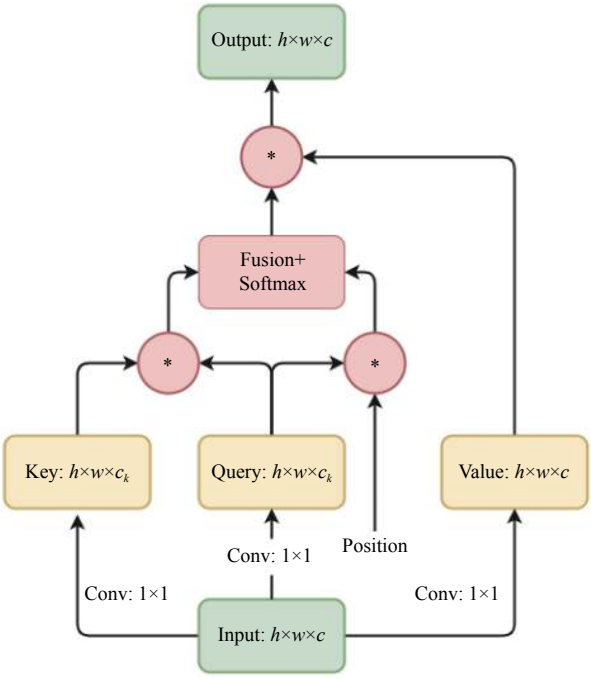

图 2 传统的 Self-Attention 模块

结合图像可以得知, *Mish* 激活函数在 X 为负值的情况下, 并没有完全截断信息流, 而是允许存在轻微的梯度流, 保留更多信息流入神经网络; 当 X 为正值时, 函数梯度逐渐趋近于 1, 相对 ReLU 更加平滑, 梯度下降效果更好, 从而得到更好的准确性和鲁棒性. 鉴于以

上优点, CoA 模块采用 *Mish* 激活函数替代 CoT 中采用的 ReLU 激活函数.

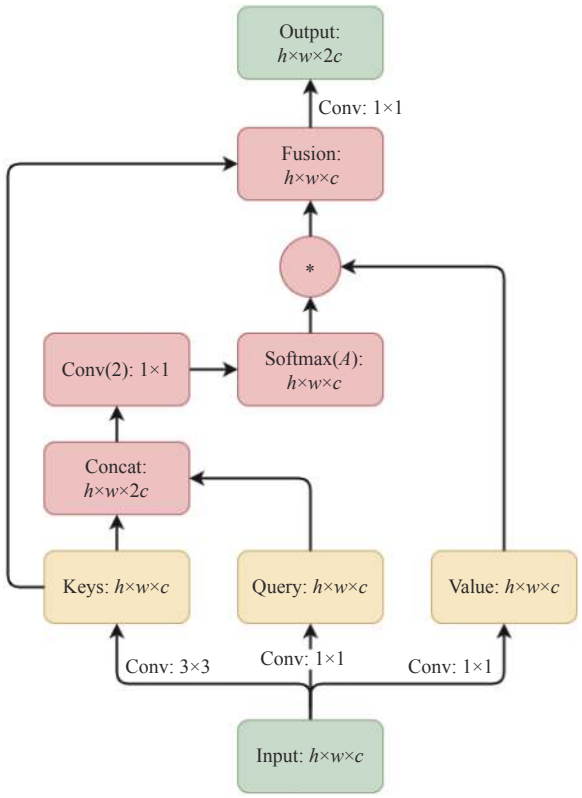

图 3 改进后的 CoA 模块

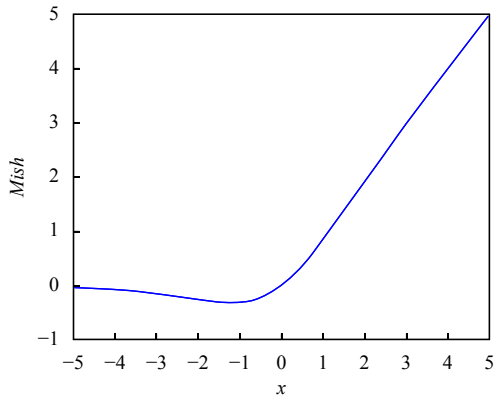

图 4 *Mish* 激活函数图像

CoA 模块能够根据骨干网络残差块中前后通道的不同进行适配, 弥补了 CoT 模块通道不变性的缺点, 同时改进特征层 query 和 *keys* 的获取, 相较于 Self-Attention 模块, 加强了对输入特征提取, 进而生成更强的注意力矩阵权重, 再结合 *value* 和 *keys* 获得更强的、拥有动静态结合的注意力矩阵, 最后通过 1×1 卷积实现跨通道信息交融和升维目的. 综上, 改进后的 CoA 模块既可以加深通道特征又能保留了原有的空间特征, 强化了网络对输入的特征提取能力.

2.2 最终改进后的 YOLOx 网络

注意力机制本身是模仿人的神经系统提出的概念. 中大课题组的 SimAM[12] 也受此启发, 其基于一些著名的神经科学理论, 构建了一个简单又非常有效的注意力模块, 无须其他参数便可通过十行代码计算解析解为特征图推导出 3D 注意力权值, 兼顾了对空间和通道特征的关注. SimAM 是一个轻量化又可以非常灵活地嵌入到卷积网络的视觉任务中, 在本文检测网络中, 将其加在输入 YOLOx-Head 之前的一层特征层上, 这能够简单地对特征层加强一些特征提取, 提升网络的检测效果.

本文最终改进的 YOLOx 网络如图 5, 将 CoA 模块替代普通的 3×3 卷积, 形成新的残差块融合到骨干网络中以及在检测头 YOLOx-Head 前增加 3D 注意力模块.

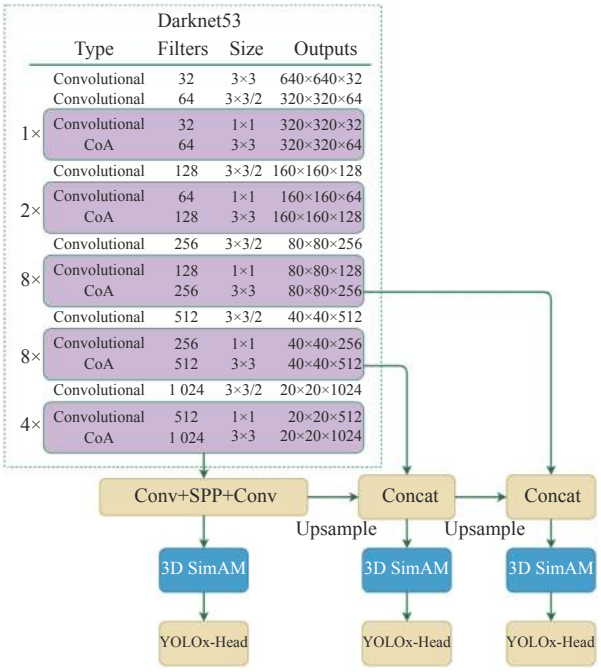

图 5 改进后的 YOLOx 检测网络

3 实验分析

3.1 实验数据集及环境介绍

为了较好的评估改进注意力模块的效果, 原网络与改进的网络均是在 Pascal VOC2007 公开数据集上

训练和预测的, 该数据集主要包含人、狗、车等常见的 20 个类别, 共 9963 张图片, 其中用于训练的图片为 5011 张 (trainval data), 测试集 (test data) 图片为 4952 张. 训练网络所使用的服务器环境配置如表 1.

表 1 服务器环境和参数

实验室服务器配置	
CPU	Intel Xeon E5-2620 v4
GPU	GTX1080 Ti ×4
操作系统	CentOS Linux release 7.6.1810 (Core)
模型使用框架	PyTorch 1.7, Python 3.7, cuda 10.1

3.2 实验评价指标

实验采用 COCO 目标检测评价指标, 能够更加全面的展示网络对不同尺寸物体的检测精度. 本文主要考虑的评价标准是平均精度均值 (mean average precision)AP@[.5:.95] 和 AP@0.5. 该精度值由 *P*-*R* 曲线积分求和获得, 其中 *P* 为查准率, 表示网络预测出的检测框中检测正确的比例; *R* 为查全率, 表示网络预测正确的框在所有标注框中的比例, 具体公式如下:

$$P = TP/(TP + FP) \tag{4}$$

$$R = TP/(TP + FN) \tag{5}$$

其中, *TP* (true positive) 为将正样本预测为正确的个数; *FP* (false positive) 为将负样本预测为正确的个数; *FN* (false negative) 为将正样本预测为错误的个数 (漏框). *P*-*R* 曲线是由 *R* 为横轴, *P* 为纵轴组成的曲线, AP 值则是曲线和坐标轴所围成的面积, 而 mAP 则是多类物体求得 AP 后再求和平均的精度值. 其中 AP@0.5 是指判断正负样本的 IOU 阈值设为 0.5 时求得的 mAP; AP@[.5:.95] 是 IOU 阈值从 0.5 以 0.05 为步进增加到 0.95 的十个阈值求得的 mAP 的均值, 相比 AP@0.5 更为严格, 更能作为检测网络精确度的评价指标. 往往 AP@[.5:.95] 越高, 越能代表网络具有更好的检测精度.

3.3 实验结果与分析

为了保证实验的公平性, 改进的 YOLOx 网络与原网络 YOLOx-Darknet53 采用相同的训练策略[9]. 在 VOC 数据集上从头开始训练 300 个 epochs, 所有超参数设置相同: 输入图片大小为 (640, 640), batch size 设置为 8, 初始学习率设为 0.01 且采用余弦退火 (cosine annealing) 方法来衰减学习率, 权重衰减系数为 5E–4, 以及最后在训练结束前 15 epochs 自动关掉数据增强继续训练至完成, 目的是让检测网络避开数据增强导致的不准确标注框的影响, 从而在自然图片的数据分布下完成最终的收敛.

图 6 和图 7 分别展示了原网络和改进网络训练过程中损失 Loss 和检测精度 AP@[.5:.95] 的变化. 从图中可以看出, 改进后的网络曲线相对而言都比较平滑, 训练过程损失波动较小, 收敛效果更好, 精度更高, 网络更加稳定. 正如上面所说, 在训练过程的最后 15 epoch 中去掉数据增强, 网络损失得以进一步下降, 收敛到更小的损失值, 使得网络检测精度进一步提升并最终收敛到一个稳定的峰值.

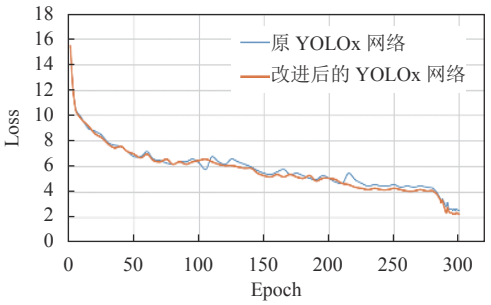

图 6 原网络与改进网络的损失值曲线

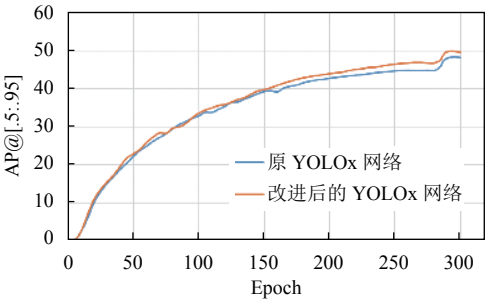

图 7 原网络与改进网络的检测精度曲线

使用 VOC2007 数据集对不同网络的训练所得到的具体检测结果如表 2 所示. 通过表 2 的实验结果可以看出, 将 CoA 模块与 Backbone 中的残差块融合后, 会增加大概 10M 的参数量, 但 AP@[.5:.95] 和 AP@0.5 都比原网络提升了 1.4 个百分点, 其中对大目标的检测提升 1.5 个百分点, 对较难检测的小目标也有 0.4 个百分点的提升. 由于小目标本身容易因为过深的网络深度从而丢失边缘信息导致检测精度下降, 且 CoA 模块本身一定程度加深了网络深度, 但融合后的 Backbone 对小目标的精度仍有小幅提升, 这正好说明了 CoA 模块确实加强了 Backbone 的特征提取能力.

在融合了 CoA 模块的 Backbone 后的 YOLOx-Head 检测头前加入了 SimAM模块, 对即将输入检测头前的不同尺寸特征层再进行一次简单地 3D 注意力关注, 形成了最终改进的 YOLOx 检测网络, 在不增加参数的同时, AP@[.5:.95]比只改进 Backbone 的网络提升 0.2 个百分点, 除此之外的指标也都有 0.2 左右的提升. 最终改进的 YOLOx 检测网络比原网络 YOLOx-Darknet53 的精度提升大多在 1.5 以上, 其中最主要的指标 AP@[.5:.95] 和 AP@0.5的提升分别为 1.6 和 1.5 个百分点. 上述实验结果也表明改进后的 YOLOx 网络不仅对大物体的特征关注明显, 也对小物体的边缘信息有所加强, 能够对不同大小的物体有更准确地检测.

表 2 不同网络在 VOC2007 测试集的实验结果

网络结构	AP@[.5:.95]	AP@0.5	AP@0.75	APS	APM	APL	参数 (M)
SSD-ResNet50	46.8	74.8	49.4	16.4	31.4	52.5	—
Yolox-Darknet53	48.3	72.8	53.0	16.8	35.8	55.8	63.68
融合CoA的Backbone	49.7	74.2	54.6	17.2	36.9	57.3	74.05
最终改进的YOLOx (图5)	49.9	74.3	54.8	17.3	37.1	57.6	74.05

SSD-ResNet50 检测网络同样是 Backbone 采用残差结构性能优异的改进版 one-stage SSD 网络. 由表 2 可知, SSD 的 AP@0.5 较高, 是因为其采用 Anchor-based 的多尺度大量瞄框方式, 当 IOU=0.5 时, 获取到的正样本预测框较多, 即 *TP* 多, 因此检测精度较高, 但当评价指标为更加严格的 AP@[.5:.95] 时, 获取到的大量预测框是负样本, 即 *FP* 多, 检测精度明显下降. 本文最终改进版的 YOLOx 对比改进版 SSD, 除 AP@0.5 和 APs 外的指标基本都大幅超越, 其中最重要的指标 AP@[.5:.95] 超过 3 个百分点, 说明改进后的 YOLOx 网络达到了高精度的检测水平.

图 8 是 SSD-ResNet50 网络与 YOLOx-Darknet53 原网络、最终改进的 YOLOx 网络的检测效果对比图, 其中置信度阈值设置为 0.4, 非极大值抑制 (NMS) 阈值设置为 0.45. 通过第 2 张对比图可以看出 SSD 网络的检测框较多, 虽然能找出正确检测目标, 但同时错检率较高, 检测准度有限. 综合测试图片的检测效果来看, 本文改进网络的检测效果比原网络和 SSD 网络都更加精准, 能够很好地改善漏检、错检的情况, 更加适合实际工业应用中高检测准度的要求.

4 结论与展望

本文根据注意力机制和 CoTNet 的优缺点, 基于 YOLOx-Darknet53 中残差块的特性改进得到 CoA 模块, 并将其融入于 Darknet53 中, 从而获得具有更强特征提取能力的骨干网络, 并结合最近的 SimAM 模块设计了一个全新的 YOLOx 检测网络, 检测精度比原网络大多提高了 1.5 个百分点, 比改进版 SSD 网络也有大幅提升, 总体检测效果有很大的提升, 更加符合工业界对不同大小物体检测准度的要求. 但由于 CoA 模块的引入, 改进后的网络也带来了参数量和计算量增加, 因此后续研究会基于此考虑尝试优化主干网络结构, 降低网络深度, 争取在减少参数量、计算量的同时进一步提升检测精度.

(a) 改进版 SSD 网络 (b) 原 YOLOx 网络 (c) 改进版 YOLOx 网络

图 8 检测效果对比图

参考文献

1 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

2 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

3 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

4 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

5 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot MultiBox detector. 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

6 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

7 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

8 王振, 邓三鹏, 祁宇明, 等. 基于 YOLO v3 的钢轨螺栓组件故障检测方法. 机器人技术与应用, 2021, (1): 34–36. [doi: 10.3969/j.issn.1004-6437.2021.01.009]

9 Ge Z, Liu ST, Wang F, *et al*. YOLOX: Exceeding YOLO series in 2021. arXiv: 2107.08430, 2021.

10 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

11 Li YH, Yao T, Pan YW, *et al*. Contextual transformer networks for visual recognition. arXiv: 2107.12292, 2021.

12 Yang LX, Zhang RY, Li LD, *et al*. SimAM: A simple, parameter-free attention module for convolutional neural networks. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 11863–11874.

13 Zhang HY, Cissé M, Dauphin YN, *et al*. mixup: Beyond empirical risk minimization. 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.

14 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255.

15 Ge Z, Liu ST, Li ZM, *et al*. OTA: Optimal transport assignment for object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 303–312.

16 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

17 Misra D. Mish: A self regularized non-monotonic activation function. 31st British Machine Vision Conference 2020. BMVA Press, 2020.

(校对责编: 牛欣悦)