Attention-YOLO:引入注意力机制的YOLO检测算法

徐诚极,王晓峰,杨亚东
上海海事大学 信息工程学院,上海 201306

摘 要:实时目标检测算法YOLOv3的检测速度较快且精度良好,但存在边界框定位不够精确、难以区分重叠物体等不足。提出了Attention-YOLO算法,该算法借鉴了基于项的注意力机制,将通道注意力及空间注意力机制加入特征提取网络之中,使用经过筛选加权的特征向量来替换原有的特征向量进行残差融合,同时添加二阶项来减少融合过程中的信息损失并加速模型收敛。通过在COCO和PASCAL VOC数据集上的实验表明,该算法有效降低了边界框的定位误差并提升了检测精度。相比YOLOv3算法在COCO测试集上的$mAP_{@IoU[0.5:0.95]}$提升了最高2.5 mAP,在PASCAL VOC 2007测试集上达到了最高81.9 mAP。

关键词:目标检测;YOLOv3算法;Attention-YOLO算法;通道注意力机制;空间注意力机制
文献标志码:A **中图分类号:**TP391.41 doi:10.3778/j.issn.1002-8331.1812-0010

徐诚极,王晓峰,杨亚东.Attention-YOLO:引入注意力机制的YOLO检测算法.计算机工程与应用,2019,55(6):13-23.
XU Chengji, WANG Xiaofeng, YANG Yadong. Attention-YOLO: YOLO detection algorithm that introduces attention mechanism. Computer Engineering and Applications, 2019, 55(6):13-23.

Attention-YOLO: YOLO Detection Algorithm That Introduces Attention Mechanism

XU Chengji, WANG Xiaofeng, YANG Yadong
College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

Abstract: YOLOv3 is a real-time object detection algorithm, its speed and accuracy reach good trade-off, but the disadvantages are that the boundary box positioning is inaccurate and it is difficult to distinguish overlapping objects. For the above problems, this paper proposes the Attention-YOLO algorithm based on the item-wise attention mechanism which embeds channel and spatial attention mechanism in the feature extraction network, uses the filtered weighted feature vector to replace the original residual fusion, and adds a second-order item to reduce the information loss in the process of fusion and accelerate the convergence of the model. Based on the experiments on COCO and PASCAL VOC datasets, the results show that the Attention-YOLO algorithm effectively reduces the boundary box positioning loss and improves the detection accuracy. Compared with YOLOv3, the Attention-YOLO improves at most 2.5 mAP@IoU[0.5:0.95] on COCO dataset, and reaches 81.9 mAP on PASCAL VOC 2007 test.

Key words: object detection; YOLOv3 algorithm; Attention-YOLO algorithm; channel attention; spatial attention

1 引言

目标检测[1]是计算机视觉领域中最具有挑战性的问题之一,同时它也被广泛应用于人脸检测[2]、自动驾驶[3-4]、行人检测[5]等许多领域。目标检测任务通常包括标记出所给图片中待检测物体的边界框,并且判断边界框中的物体属于哪一类别。传统的目标检测可以分为三个步骤:首先,选取感兴趣区域,考虑到待检测的物体可能出现在图片中的任一位置,同时其大小比例也不是固定的,通常采用多尺度、多长宽比的滑动窗口技术[6]来扫描整个输入图片。尽管这种技术可以较好地标记出所有可能出现待检测物体的位置,但是由于扫描时无差别地对待每一个区域,导致计算量巨大,并且会产生冗余的标记窗口。其次,从标记的区域中提取特征,常用的有SIFT[7]、HOG[8]以及Haar-like[9]等手工特征。但是由于外

基金项目:国家自然科学基金(No.61872231,No.61703267);上海海事大学研究生创新基金(No.2017ycx083)。
作者简介:徐诚极(1996—),男,硕士研究生,研究领域为深度学习与目标检测,E-mail:13761350550@163.com;王晓峰(1958—),男,博士生导师,教授,研究领域为人工智能,数据挖掘与知识发现等;杨亚东(1990—),男,博士研究生,研究领域为计算机视觉,图像处理。
收稿日期:2018-12-03 **修回日期:**2019-01-07 **文章编号:**1002-8331(2019)06-0013-11

形,光照条件以及背景的多样性,手工特征的鲁棒性较差,无法表征所有情况下的物体特征。最后,对所提取的特征进行分类[10-11],识别出检测框中物体的类别。

近几年来,随着深度卷积神经网络在计算机视觉领域的深入应用,以YOLO算法[12-14]以及SSD[15]算法为代表的实时目标检测模型在工业领域以及实际应用场景中[16-17]发挥了很好的检测效果。但是,由于这两种算法将目标检测过程视作回归问题来处理,不能很好地将前景区域与背景区域区分开,容易产生误检和漏检。而以Faster-RCNN[18]为代表的含有Region Proposal Networks(RPN)的检测算法能在感兴趣区域的生成阶段就将可能含有待检测物体的区域大致确定下来,在大大提升准确率的同时也能为之后的分类阶段节省处理时间。

目前对目标检测算法的改进主要包括:采用能够提取到更丰富特征的基础神经网络、融合多个尺度的特征来进行检测或是其他对于检测环节改进的方法。Fu等人[19]提出在SSD检测框架的基础上,采用更深的ResNet-101[20]网络来进行特征提取,并且采用了反卷积层,引入额外的大量语义信息,改进了SSD算法对于小物体的检测能力。Shen等人[21]同样在SSD的基础上借鉴了DenseNet[22]的稠密连接,提出了一种能提升检测准确率的Stem Block结构,改善了训练目标检测模型时往往需要加载预训练权重的情况。Bodla等人[23]针对非极大值抑制算法可能导致的漏检情况,提出了一种Soft-NMS算法,对于重叠部分超过阈值的得分框,降低其置信度,而不是直接进行抑制。该算法在不增加训练成本以及参数量的前提下,获得了平均1%的mAP提升。温捷文等人[24]对YOLOv2算法的特征提取网络进行批再规范化的处理,并且移除Dropout层,相比较原YOLOv2算法取得了更高的检测精度和更快的训练速度。Lin等人[25]利用了深度卷积神经网络的多尺度特征,提出了一个自上而下并且具有横向连接的特征金字塔网络结构,在不同的尺度上分别进行预测,并将多层的特征图进行融合。以上这些工作从不同角度提升了检测算法的性能。

在最近的研究中,Hu等人[26]将所提出的通道注意力模块分别应用于ResNet及ResNeXt[27]系列网络上,在ImageNet 2012数据集[28]上的top-1及top-5分类错误率降低了最多1.80%和1.11%,在COCO 2014数据集[29]上的$mAP_{IoU=0.5}$提升了1.6%。此外,Woo等人[30]发现,同时对卷积操作间的通道关系及空间关系进行建模加权,可以更好地筛选出所需要的特征。在YOLOv3检测算法中,所提取的卷积特征并未对卷积核中不同的位置进行加权处理,即同等对待整个特征图中的每个区域,认为每个区域对于最终检测的贡献是相同的。然而在实际的生活场景中,图中待检测物体的周围往往具有复杂且丰富的语境信息,对目标区域的特征加以权重,可以使之能更好地定位在待检测的特征之上,也能在不引入过多参数量的基础上提升网络的泛化性能。

本文在YOLOv3算法的基础上,提出基于注意力机制的Attention-YOLO算法。在特征提取网络的残差连接中加入注意力机制,使得具有注意力效果的梯度能流入更深的网络中。此外,改进了残差连接中单一直接的特征融合方式,加入二阶项及微小的偏置项。实验表明,在不引入过多额外参数量的前提下,Attention-YOLO算法相比原始的YOLOv3算法有一定的性能提升。

本文的主要贡献如下:提出引入注意力机制的Attention-YOLO算法,在特征提取网络中加入通道注意力及空间注意力机制,最终仅增加1.4%的参数量,在不影响实时性的前提下改进了其对于关键特征的提取筛选能力;修改特征提取网络中残差连接直接线性融合两层特征图的方法,加入二阶项来更好地保留提取到的深层及浅层特征并提升结构的非线性程度。其中,Attention-YOLO算法的纵向性能比较实验在不借助预训练权重及多尺度训练等技巧的基础上,取得了比原文方法更好的检测精度。

2 相关工作

2.1 注意力机制

在神经网络中,可以存储的信息量称为网络容量,网络容量和网络的复杂度成正比[31]。人脑在计算资源有限的情况下,不能对于过载的视觉信息同时处理每一位置的视觉图像,而是通过视觉的注意力机制(Attention mechanism)进行处理[32-33]。

按照注意力本身的形式分类,注意力机制可以分为软性注意力和硬性注意力。按照注意力作用的特征的形式,注意力机制可分为基于位置[34]的注意力和基于项[35]的注意力。基于位置的注意力的输入是具有空间维度的特征图,基于项的注意力的输入是包含明确的项的序列性数据。在计算机视觉领域中,基于位置的注意力是与任务较为相关、作用方法较为直接的注意力机制,其应用较为广泛。基于项的注意力在很多特殊的模型中也得到了应用[36],由于其可以直接嵌入目前流行的诸多卷积神经网络结构中,并且能够在不改动原有结构的前提下进行端对端训练,因此采用基于项的注意力来改进卷积神经网络是改动成本最低,且收益较好的选择。

当注意力机制用于图像描述[34]任务中时,注意力机制模块所要处理的信息包含明确的项序列 $\boldsymbol{a}=\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_L\},\boldsymbol{a}_i\in\mathbb{R}^D$,其中 L 代表特征向量的个数,D 代表的是维度空间。因此所采用的注意力机制需要计算出当前时刻 t 每个特征向量 $\boldsymbol{a}_i$ 的权重 $\alpha_{t,i}$,公式如下:

$$e_{ti}=f_{\text{att}}(\boldsymbol{a}_i,h_{t-1}) \tag{1}$$

$$\alpha_{t,i}=\frac{\exp(e_{ti})}{\sum_{k=1}^{L}\exp(e_{tk})} \tag{2}$$

其中，$f_{att}(\cdot)$代表多层感知机，e_{ti}代表中间变量，h_{t-1}代表的是上个时刻的隐状态，k代表特征向量的下标。

计算出权重后，模型就可以对输入的序列a进行筛选，得到筛选后的项序列$\hat{z}_t$，其中：

$$\hat{z}_t=\phi(\{\boldsymbol{\alpha}_i\},\{\boldsymbol{a}_i\}) \tag{3}$$

最终，注意力是硬性的或软性的取决于函数ϕ的选取。

当$\hat{z}_t$为线性加权函数时，注意力为软性注意力。而硬性注意力对L个特征向量进行离散选取，令s_t表示模型选取作为注意力关注点的位置，$s_{t,i}$表示独热编码向量，如果某个特征向量a_i被选中，则其对应的$s_{t,i}=1$。j代表某个小于t的时刻，将$a_{t,i}$视作概率，由其构成的多项式分布得到最终选择的$\hat{z}_t$，如下式所示：

$$p(s_{t,i}=1|s_{j<t},\boldsymbol{\alpha})=a_{t,i} \tag{4}$$

$$\hat{z}_t=\sum_i s_{t,i}\boldsymbol{\alpha}_i \tag{5}$$

在细粒度图像识别领域中，Fu等人[37]提出的RA-CNN网络按照由粗及细的过程，使用递归网络依照注意力提取重要区域，将其进行放大并作为下一级的输入图像。其中应用的注意力机制结合了硬性注意力和软性注意力，同时也属于基于位置的注意力方法，将产生的硬性注意力位置用k阶逻辑函数这样的阶梯型函数进行拟合，从而得到可导的注意力权重，进而构成可端对端训练的网络模型，在Stanford Dogs datasets[38]上达到了最高的87.3%的分类准确率，但由于其注意力模型复杂，速度上仍然低于回归型检测算法。

类似的，Hu等人[26]提出的挤压与激励网络(SENet)以及Woo等人[30]提出的卷积注意力模块(CBAM)分别在网络的特征通道维度以及特征空间维度上进行了特征压缩和生成权重并重新加权的操作，这两种方法可以看作在特征通道维度及特征空间维度上的基于项的注意力。本文选取了ResNet50分类网络及在此基础上加入了上述两种注意力机制后的分类网络作为对比，为直观说明注意力机制对分类结果的影响，选用了Grad-CAM方法[39]来进行分类结果依据的可视化。通常，卷积神经网络的最后一个卷积层具有最丰富的空间及语义信息，其输出维度与分类的类别数一致，Grad-CAM方法通过求解全局平均后的梯度来得到每个类别所对应的特征图所占的权重，最后将得到的权重与对应的特征图进行加权求和，在每个类别上都能得到一个可视化的热力图。如图1所示，红色部分是特征图中对应类别置信度较高的地方，也是分类网络在特征图中所集中关注的部分，其中P值为Softmax打分。得益于残差结构和较深的网络层数，ResNet50网络能较好地专注于目标类别所在的特征图区域，在此基础上，通道及空间注意力的作用使得分类网络能更好地区分无关特征，抑制影响分类结果的其他信息。在检测算法中，选用分类特征更加精确的特征提取网络将有助于之后的回归预测及分类训练。

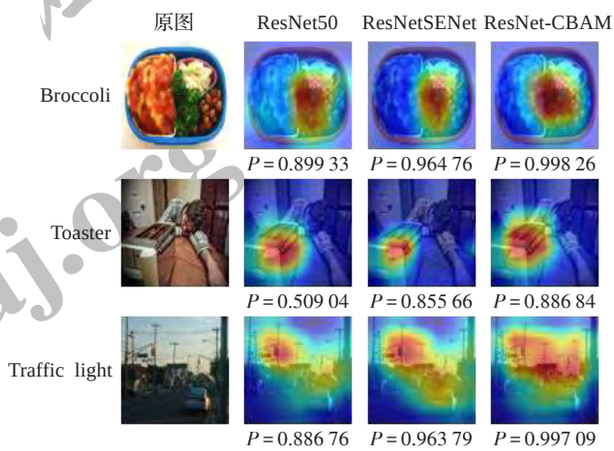


图1　不同注意力机制分类网络的类别热力图

2.2　YOLOv3目标检测模型

YOLO算法将整个检测环节作为一个回归及分类问题处理，并没有选择Select Search[40]、Edge Box[41]或者是RPN[18]这样的区域生成算法来完成感兴趣区域的初始标定，而是直接使用原始的输入图像及标注进行训练，节省了大量计算资源和耗时。

YOLOv2算法开创性地提出了一种联合训练目标检测数据集和图像分类数据集的方法，可以使模型预测没有分类标注过的数据。YOLOv3算法针对YOLOv2算法定位不准确及召回率较低等问题进行了改进，主要改进点有以下几点：类别预测阶段由原先的单标签分类改进为多标签分类，改善了其在类别复杂型数据集上的分类性能；采用了三个尺度上的特征进行预测，相比较YOLOv2仅仅使用13×13的特征图进行预测，大大地增加了特征图中保留的信息量；同时也采用了更深且具有残差连接的网络来进行特征提取。

2.2.1　网络结构

YOLOv3算法的网络结构主要由Darknet-53特征提取网络以及特征金字塔网络两部分组成。整个特征提取网络完全由卷积层组成，没有用到池化操作。

如图2所示，YOLOv3中所采用的特征金字塔结构(FPN)则是直接在原来的单一网络上做修改，在每个分辨率的特征图上引入后一分辨率缩放两倍的特征图并做相加(element-wise)操作。

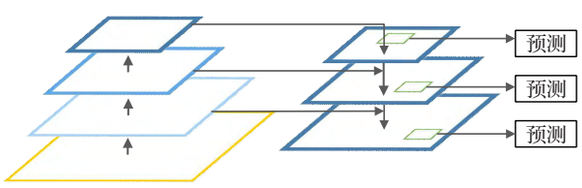


图2　YOLOv3的多尺度预测结构

通过这样的连接,每一层预测所用的特征图都融合了不同分辨率、不同语义强度的特征,融合的不同分辨率的特征图分别用作对应分辨率大小的物体检测。这样保证了每一层都有合适的分辨率以及强语义特征。同时,由于此方法只是在原网络基础上加上了额外的跨层连接,在实际应用中几乎不增加额外的时间和计算量。

2.2.2　检测过程

YOLOv3算法不需要预先生成感兴趣区域(ROI),而是直接以回归的方式来训练网络,同时对COCO 2014训练数据集使用*K*-means算法来进行训练样本边界框的聚类,最终分别在3个尺度大小上预设了3组预定义的边界框大小,之后的定位预测将基于这9种大小的边界框进行,如图3。首先通过特征提取网络在原始的416×416的输入图像上进行特征提取,接着将特征向量送入FPN结构,产生3个尺度上的网格区域,分别为13×13、26×26以及52×52,每个网格区域预测3个边界框,共产生(52×52+26×26+13×13)×3=10 647个边界框。接着在每一个边界框中预测一个向量 P ,向量 P 的组成如下式所示:

$$P=(t_x+t_y+t_w+t_h)+P_0+(P_1+P_2+\cdots+P_n) \tag{6}$$

向量 P 中前4个元素为与边界框有关的4个坐标,它们的关系如下式所示:

$$b_x=Sigmoid(t_x)+C_x \tag{7}$$

$$b_y=Sigmoid(t_y)+C_y \tag{8}$$

$$b_w=p_w\times e^{t_w} \tag{9}$$

$$b_h=p_h\times e^{t_h} \tag{10}$$

其中 C_x 和 C_y 表示的是该边界框所属网格相对于图片左上角的偏移, p_h 和 p_w 表示的是预定义边界框的长宽大小, b_x 和 b_y 表示的是最终预测结果边界框的中心距离图片左上角的位置, b_h 和 b_w 则是预测边界框的长与宽。

向量 P 的第5个元素 P_0 由下式表示:

$$P_0=Prob(object)\times IoU_{object}^{gt} \tag{11}$$

$Prob(object)$ 表示的是物体处于预测框中的概率, IoU_{object}^{gt} 表示的是预测框和真实边界框的交并比。当使用逻辑斯特回归对预测框进行的打分最高时,物体处于预测框中的概率为1,否则为0。向量 P 中剩余的 n 个值代表的是预测的物体属于 n 类中某一类的分数,经Sigmoid函数得出。最后,对产生的预测框进行非极大值抑制,得到最终的预测结果。整个检测过程如图3所示。

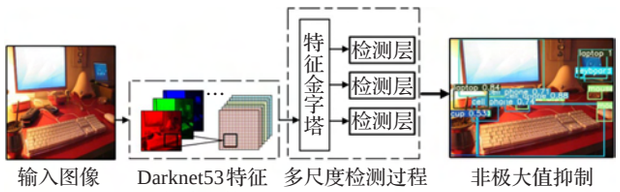


图3　YOLOv3算法检测流程图

3　Attention-YOLO检测算法

3.1　基于通道注意力机制的改进网络

YOLOv3算法的实时检测性能得益于其全卷积网络结构和较小的卷积核尺寸以及回归边界框的算法设计,为了在不加深网络深度的前提下提升检测精度,Attention-YOLO算法仅仅对网络中所有的残差连接进行替换,通过对于传递特征的筛选,使得残差融合时保留的信息更加有利于训练损失的降低,有利于定位及分类的准确。网络在这部分增加的计算量集中在全连接层部分,通过一定比例的降维可以权衡性能和检测速度的取舍。

通过对于神经网络中传递的特征通道加以不同的权重,网络可以更加重视权重较大的通道进行参数更新。直观来说,在前向传播的过程中,重要的特征通道将会占有更大的比重,在最终所呈现的输出图像中也能更加明显地展现出检测网络所重点关注的部分,更好地分辨出待检测物体"是什么"。

全局平均池化是一种特殊的池化,最初由Lin等人[42]所提出,常被用来聚集空间信息。全局平均池化是对整个特征图进行平均池化,一张特征图最终得出一个值。通道注意力模块的作用是通过对特征图的各个通道之间的依赖性进行建模以提高对于重要特征的表征能力。首先通过在各层特征图上的全局平均池化获得各个通道的全局信息。然后使用两个全连接层及ReLU非线性激活函数和Sigmoid激活函数来自适应地对各通道间的相关程度进行建模,最后再将原特征通道的信息与自适应学习建模后的权重进行加权处理,实现特征响应及特征重校准的效果。通过这样的结构,网络可以有选择性地加强包含重要信息的特征并抑制作用无关或较弱关联的特征。

设输入注意力模块的卷积核集合为 Y,且 $Y\in\mathbb{R}^{H'\times W'\times C'}$,其中 H'、W'、C' 分别表示的是特征图的长度、宽度和通道数。设卷积操作为 F_{conv} ,前一层的卷积核集为 $X=[x_1,x_2,\cdots,x_c]$,且 $X\in\mathbb{R}^{H\times W\times C}$,其中 H 、W 、C 同样表示的是特征图的长度、宽度和通道数。则 F_{conv} 表示 $X\to Y$ 的转变过程,用符号 $\otimes$ 表示。令 $K=[k_1,k_2,\cdots,k_c]$ 表示卷积过程中的参数, c 表示的是第 c 个特征通道, k_c 表示的是第 c 个特征通道上的卷积核参数, $Y=[y_1,y_2,\cdots,y_c]$ 表示卷积操作后的输出,则整个卷积过程由下式表示:

$$y_c=k_c\otimes X=\sum_{c=1}^{C'}k_c\otimes x_c \tag{12}$$

从公式(12)中可以看出,所有通过卷积核计算得出的通道信息都被直接叠加在一起,所有的特征图对于结果判断都占有同样的比重。

对于目标检测任务而言，特征提取网络对于不同的物体所关注的关键特征区域是不一样的，如果在训练初期就以同样的关注程度对待每一个特征图，会增加网络收敛所需的时间。由于定位精确和分类准确是彼此互相促进的，相比分类网络，检测算法更受益于精准分明的特征。同时，由于引入的参数量并不影响算法的实时性，并且获得了良好的mAP提升，因此选用通道注意力机制是个较好的选择。

Darknet53特征提取网络具有大量的残差连接，因此加入注意力模块时需要进行一定的结构调整。用于残差连接的通道注意力模块主要由两部分组成，如图4所示，分别是挤压(Squeeze)和激活(Excitation)操作，设挤压操作为 F_{sq}，即全局池化过程。保留输入注意力模块的卷积集 Y 为残差分支的输入之一，设输出结果为长度为 c 的一维数组 $z_c=[z_1,z_2,\cdots,z_c]$，(i,j) 表示的是在大小为 $H\times W$ 的特征图上横纵坐标分别为 i 和 j 的点，由此可以得到：

$$z_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}y_c(i,j)=F_{sq}(y_c) \tag{13}$$

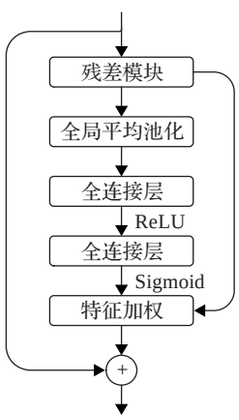


图4　用于残差连接的通道注意力机制结构

设激活操作为 F_{ex}，r 为降维的比例，当 r 越小时，可以更好地保留前一层传入的全局信息，但是相对会增加计算量，为了达到传播速度和检测准确率的平衡，参考SENet[26]中的取值 $r=16$。

令 $FC_1\in\mathbb{R}^{\frac{c}{r}\times c}$，$FC_2\in\mathbb{R}^{c\times\frac{c}{r}}$ 为两个不同的全连接层，设最终激活操作的输出参数为向量 e，则：

$$e=Sigmoid(FC_2\times ReLU(FC_1\times z_c)) \tag{14}$$

接着通过特征加权操作：

$$output_c=y_c\times e_c \tag{15}$$

得到重新筛选过后的特征图向量：

$$output_{ch}=[output_1,output_2,\cdots,output_c] \tag{16}$$

最终，整个通道注意力机制模块的输出结果为 $output_{ch}+Y$。

由于引入了新的池化层以及全连接层，在计算量上相比原YOLOv3有所增加，这也是降低部分检测速度的主要原因。

3.2　基于通道和空间注意力机制的改进网络

通道注意力机制在通道维度上对特征进行了筛选加权，提升了其检测性能。根据Woo等人[30]在CBAM中所提出的观点，除了全局平均池化外，全局最大池化同样也能对特征的筛选起到帮助，因此引入的通道注意力机制需要同时考虑两种池化操作。根据Zagoruyko等人[43]的研究发现，沿着通道方向使用池化操作可以突出有效区域中的信息。除此之外，Woo等人[30]认为，特征在空间上的关系同样可以用来进行建模，用以补充通道注意力机制无法较好获取的位置关系信息。在此基础上，进一步加入空间注意力机制，整个注意力模块同时对通道以及空间的特征信息进行筛选。

为了达到最佳的筛选效果，并且为了获得最佳的注意力模块组合顺序。Attention-YOLO算法参考了Hu[26]等人和Woo等人[30]所提出的不同的注意力模块排列顺序在分类网络ResNet50上的测试结果，分类实验在ImageNet进行训练测试。

由表1可知，不同的注意力机制模块都能提升分类的准确率，但性能最佳的组合顺序是CBAM中通道注意力模块直接连接空间注意力模块的方法。首先，这是由于CBAM模型在通道注意力模块中加入了全局最大池化操作，它能在一定程度上弥补全局平均池化所丢失的信息。其次，生成的二维空间注意力图使用卷积核大小为7的卷积层进行编码，较大的卷积核对于保留重要的空间区域有良好的帮助。

表1　注意力模块排列顺序对分类性能的影响

网络结构	Top-1错误率/%	Top-5错误率/%
ResNet	24.56	7.50
ResNet+channel(SENet)	23.14	6.70
ResNet+channel+spatial(CBAM)	22.66	6.31
ResNet+spatial+channel	22.78	6.42
ResNet+channel & spatial in parallel	22.95	6.59

首先改进原有的注意力机制，加入全局最大池化操作，两种池化操作完成后进行合并送入MLP进行通道信息筛选。接着，沿着通道维度进行平均池化和最大池化，将两者的输出合并后得到特征描述子。最后，使用卷积操作来进行编码，得到空间注意力图。以上改进不但能帮助网络进行更准确的分类，又能更精准地定位物体所在的位置。通道及空间注意力机制的结构如图5所示。

类似的，设输入该注意力结构的卷积集为 X，保留作为残差分支的输入之一，且 $X\in\mathbb{R}^{H\times W\times C}$，其中 H、W、C 同样表示的是特征图的长度、宽度和通道数。随后将

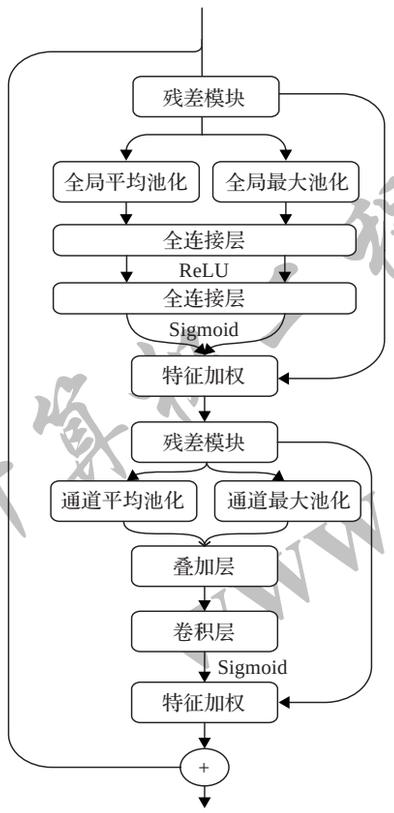


图5　用于残差连接的通道及空间注意力机制结构

其分别送入两个独立的分支进行两种不同类型的池化操作。

设全局平均池化过程为 F_{avg}，全局最大池化过程为 F_{max}。令 F_{avg} 和 F_{max} 的输出分别为 Att_{avg} 及 Att_{max}，且 $Att_{avg}\in\mathbb{R}^{1\times1\times C}$，$Att_{max}\in\mathbb{R}^{1\times1\times C}$。一维的权重序列 Att_{avg} 可以很好地筛选出目标物体的全局背景信息，同时 Att_{max} 可以很好地突出目标物体的显著特征。令 $X=[x_1,x_2,\cdots,x_c]$，其中 x_c 表示的是第 c 个卷积核的参数。则：

$$Att_{avg}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j)=F_{avg}(\boldsymbol{x}_c) \tag{17}$$

$$Att_{max}=\operatorname{argmax}\left(\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j)\right)=F_{max}(\boldsymbol{x}_c) \tag{18}$$

接着，令 $FC_3\in\mathbb{R}^{\frac{C}{r}\times c}$ 和 $FC_4\in\mathbb{R}^{c\times\frac{C}{r}}$ 为两个全连接层，训练时两条分支的输入共享全连接层的参数，其中 r 同样为全连接层的降维比例，参考 Woo 等人[29]在 CBAM 模块中的取值，取 $r=16$。设通道注意力模块部分的输出为：

$$output_{ch}=output_{avg}+output_{max} \tag{19}$$

其中两条分支的输出分别是：

$$output_{avg}=Sigmoid(FC_4\times ReLU(FC_3\times Atten_{avg})) \tag{20}$$

$$output_{max}=Sigmoid(FC_4\times ReLU(FC_3\times Atten_{max})) \tag{21}$$

接着通过矩阵乘法完成特征加权操作：

$$W=(\boldsymbol{x}_c,output_{ch})=\boldsymbol{x}_c\times output_{ch} \tag{22}$$

得到筛选过的通道特征 $W=[w_1,w_2,\cdots,w_c]$。

第一部分的通道特征筛选完成后，需要将 W 输入至新的空间注意力机制模块中。首先，输入的特征向量将分别经过 Att_{avg} 以及 Att_{max}，再沿着通道维度进行特征叠加，得到 $C_{con}\in\mathbb{R}^{1\times1\times2c}$。为了能得到二维的特征权重信息，还需要进行卷积操作来降维，令 $F_{3\times3}$ 表示的是输入通道数为2，输出通道数为1，卷积核大小为3×3的卷积操作，则最后特征加权后的输出向量为 $output_{chsp}=F_{3\times3}(C_{con})\times W$。

最终，整个通道和空间注意力机制模块的输出为 $output_{chsp}+X$，且输出的特征图维度与输入的维度一致，不需要对于网络结构进行较大的改动，并且根据 $r=16$ 的降维比例对全连接层参数进行压缩，能权衡性能与传播速度的平衡。由于引入了更复杂的结构，在速度上会略微慢于普通的通道注意力机制模块，多引入的计算时间主要集中于卷积层以及通道池化部分，但整个注意力结构的计算量最大的部分集中在通道注意力部分。

3.3　加入二阶项融合改进的残差连接

自 ResNet 及其变体网络[19,26]提出以来，许多性能优异的分类网络都相继借鉴了残差连接结构。由于非线性激活函数的作用，该结构能够很好地将浅层的特征逐渐送往深层的网络中，在网络层数较深的情况下，不仅能最大程度地保留浅层的全局特征，同时对于反向传播训练时的梯度稳定性也有帮助。

普通的残差连接如图6所示。

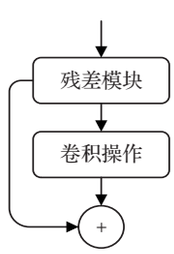


图6　残差连接示意图

令输入的特征向量为 X，经过卷积操作后的输出向量为 X'（图中的加号代表的是特征向量矩阵与特征向量矩阵之间的逐元素相加），最后融合输出的结果为 $f=X+X'$。可以看出，普通的残差连接方式为线性连接，线性结构会限制网络对于特征空间复杂分布的学习能力，而非线性结构无疑有助于网络的泛化性能。

YOLOv3 中所采用的特征提取网络为 Darknet53，是一个较深的全卷积神经网络，其采用了大量的残差连接来帮助浅层特征的传输及梯度的流动。由于 Attention-YOLO 算法采用了注意力模块来完成对特征的筛选，与原有的单一残差连接有较大的区别。为了缩短新结构和原有结构在训练过程中分布不一致的过渡时间，同

时为了进一步保留之前网络层中有效特征的传递。Attention-YOLO算法参考了Wang等人[44]所提出的Second-Order Response Transform(SORT)方法，该种方法对常见的分支结构及残差连接结构都适用，同时不需要对整个结构做很大改动，仅增加5%的前向传播时间，但能有效提升使用了残差结构或分支结构的网络的分类性能。考虑到注意力模块产生的筛选后的特征信息更加具有意义，在网络的残差连接中进一步加入一个二阶项 F'' 及一个小偏置量，增加整个结构的非线性程度，而非线性程度往往与网络的性能紧密相关。改进后的残差连接示意图如图7所示。

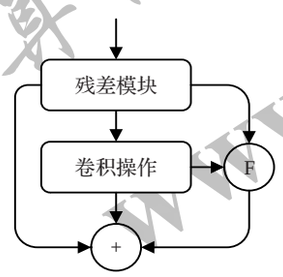


图7　加入二阶项的残差连接示意图

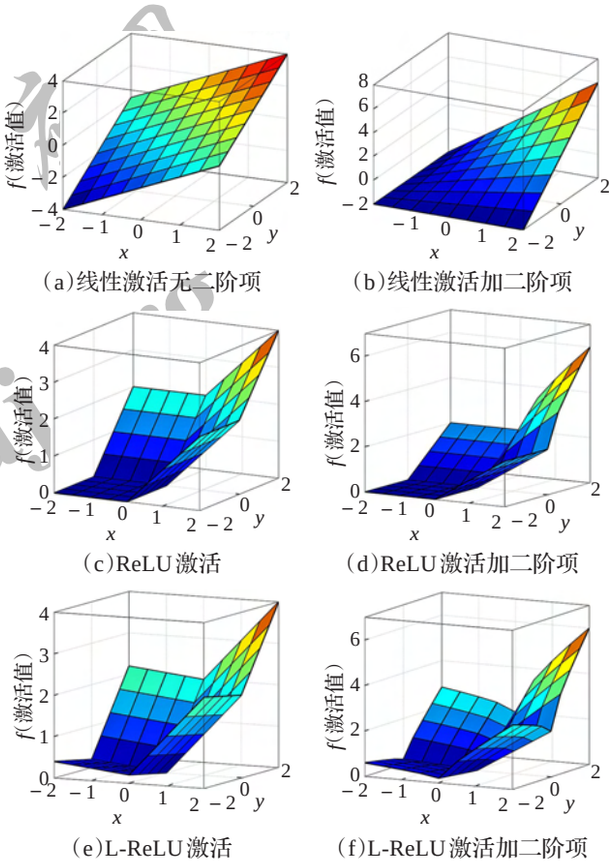
最终融合输出的结果为 $f=X+X'+F''$，其中 $F''=F(X\odot X'+\xi)$，$\odot$ 表示的是矩阵间的逐元素相乘，且 $F(\cdot)$ 必须是可微函数，这样才能够参与梯度求导并进行反向传播更新参数。参考SORT方法中的实验过程，采用平方根形式作为 $F(\cdot)$ 函数，并且取偏置量 $\xi=0.000\,1$，以维持反向传播时的梯度稳定性，最终改进后的残差连接输出为：

$$f=X+X'+\sqrt{X\odot X'+\xi} \tag{23}$$

由于 X 和 X' 特征向量矩阵中的元素可能会出现负数，因此，需要在进行乘积前进行非负处理。即：

$$f=X+X'+\sqrt{ReLU(X)\odot ReLU(X')+\xi} \tag{24}$$

如图8所示，仅仅考虑残差连接及激活函数映射在二维子空间的情况，图(a)、(b)是不使用非线性激活函数的残差连接，图(c)、(d)是使用了ReLU非线性激活函数的残差连接，图(e)、(f)是使用了Leaky-ReLU ($a=0.1$) 非线性激活函数的残差连接。其中(a)、(c)、(e)是不增加二阶项的残差连接，(b)、(d)、(f)则是增加了二阶项



激活方式	$X+X'$	$X+X'+F''$
线性激活	图(a)	图(b)
ReLU	图(c)	图(d)
Leaky-ReLU	图(e)	图(f)

图8　不同残差连接情况下的非线性程度对比图

后的残差连接。曲面的弯曲程度代表了该过程的非线性程度，可以看出，Attention-YOLO算法使用了Leaky-ReLU激活函数，并且使用了加入二阶项后的残差连接，在这几种结构中具有最好的非线性特性，当训练时，对于网络的泛化性能有很好的帮助。

3.4　训练和测试细节

3.4.1　实验准备与超参数设置

实验所用数据集为COCO(Common Objects in Context)2014和PASCAL VOC，并使用了常用的2007和2012两个版本进行训练和测试。由图9及图10可知，

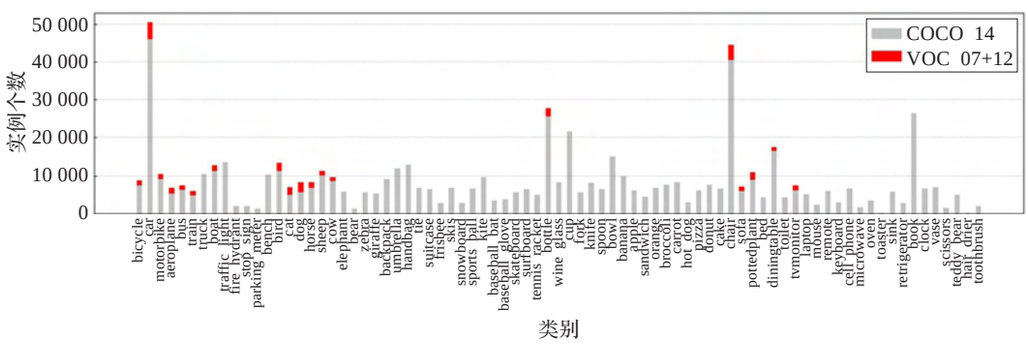
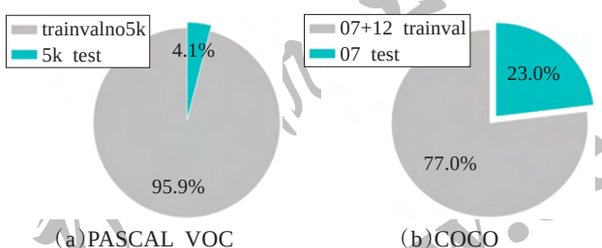


图9　COCO和VOC数据集中每个类别的目标数统计

COCO数据集中单幅图片所含的平均目标数是PASCAL VOC的3倍左右,且小目标居多。在两个数据集共有的20个类别中,COCO数据集单个类别的样本个数是PASCAL VOC的数倍,不论是规模还是检测难度,COCO数据集的训练及检测难度都大于PASCAL VOC数据集。除此之外,以上两种数据集在评估方法上也有较大差异。



(a)PASCAL VOC　(b)COCO

数据集	COCO 2014	PASCAL VOC 07+12
类别数	80	20
训练集	117 264	16 551
测试集	5 000	4 952
样本框总数	902 435	52 090
总样本框/总图片数	7.4	2.4

图10　数据集的训练和测试图片量对比

COCO数据集需要在IoU(交并比)为[0.5∶0.95]这10个取值上分别计算mAP,最后取平均得到最终结果。而VOC 2007测试集,只需要在IoU=0.5时计算相应的精度和召回率,最后得到相应的mAP。毫无疑问,COCO的多IoU评估方法更加能反映检测算法的综合性能,对检测算法的要求更高。

在此基础上,本文分别设计了用于纵向和横向比较的两组实验。纵向性能比较采用COCO 2014数据集,共80类。考虑到Attention-YOLO算法增加了新的全连接层和卷积层,通常要将新的特征提取网络在ImageNet数据集上重新进行预训练,整个过程需要耗费大量时间。但如果直接加载已有的预训练权重进行迁移学习,没有预训练权重且需要随机初始化的那部分层会与其他层的分布有一定的差异。因此,该组实验采取了He等人[45]所提出的MSRA Initialization初始化方法,该方法更加适用于使用了非线性激活函数的网络,由此也可以加速模型收敛及稳定梯度。由于没有借助预训练权重进行初始化,需要更加久的训练时间以及更加精细的超参数设置才能达到原有的效果[46]。该组实验训练迭代次数为160轮,由于没有使用多尺度变化训练的技巧,网络的输入大小固定为416×416,参数更新方法为Adam,初始学习率为0.000 1,权重衰减设置为0.000 5。

横向性能比较采用PASCAL VOC数据集,共20类,并使用了常用的2007和2012两个版本进行训练和测试。由于PASCAL VOC的训练集数量远远小于COCO,而在数据量较小时进行随机初始化训练是很难达到收敛的[46]。因此,使用部分已有的预训练权重来完成初始化,令其他新加入的部分在训练过程中自适应微调参数。该组实验训练迭代次数为120轮,其余超参数设置与上一组实验一致。

3.4.2　数据增强

在预处理阶段,对输入模型的图片进行平移、旋转,尺度变化、左右反转以及HSV饱和度及强度的变化,具体的设置如表2所示。

表2　训练时采用的数据增强

增强类型	参数
Translation	+/－20%
Rotation	+/－5 degrees
Shear	+/－3 degrees
Scale	+/－20%
Reflection	50% probability
HSV Saturation	+/－50%
HSV Intensity	+/－50%

3.4.3　网络训练

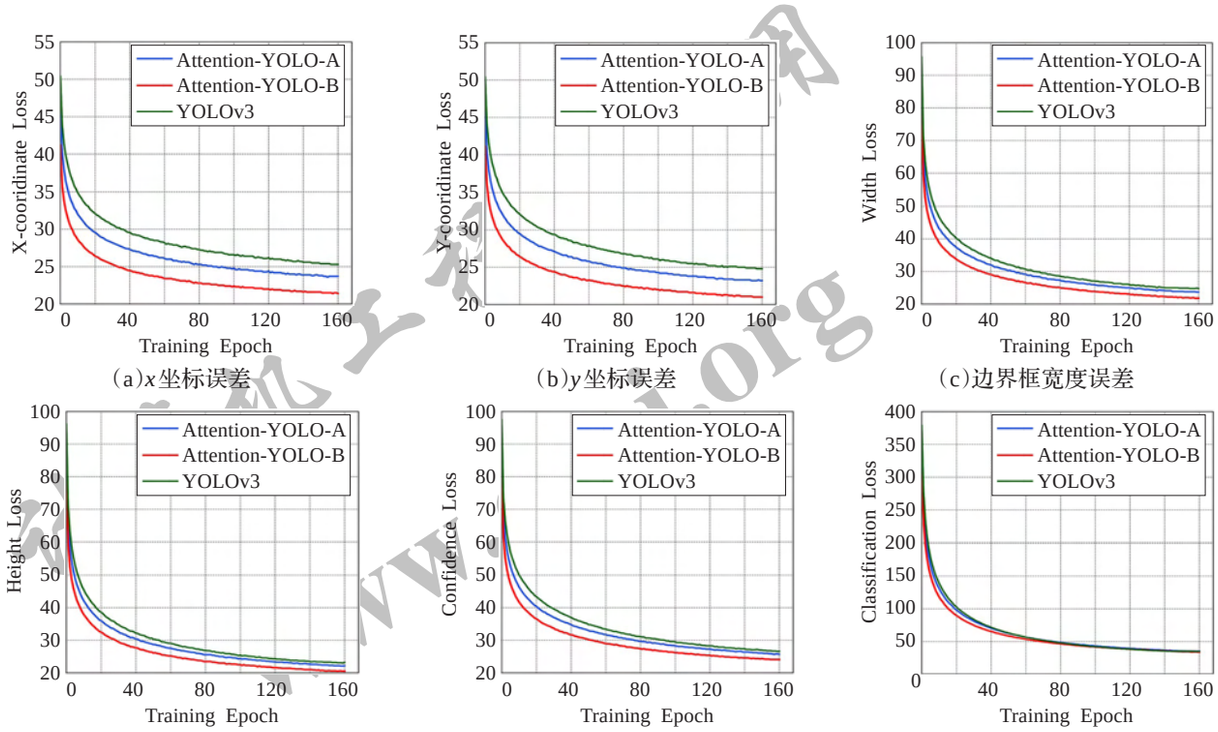
由于Attention-YOLO算法中加入了通道及空间注意力机制和残差二阶项,算法在判断某个网格;是否含有待检测物体时更加准确,从图11可以看出,特征提取网络在训练COCO数据集的过程中能够更好地定位可能存在物体的目标像素格。得益于坐标点的定位更加准确,边界框的长度及宽度损失也相应减少,定位的准确程度与分类的准确程度之间存在相辅相成的关系,因此分类损失与置信度损失随着定位损失的减小而相应减小。仅加入通道注意力机制对于原始算法的各个损失变化有一定的训练加速效果,而加入两种注意力机制并加入二阶项进行残差融合的检测模型具有最佳的收敛速度,也相应具有更好的检测性能。

4　实验结果分析

Attention-YOLO算法共有两种模型,为了简化名称,将第一种加入通道注意力机制的YOLOv3算法称为Attention-YOLO-A(以下简称模型A),将第二种同时加入两种注意力机制的模型称为Atten-tion-YOLO-B(以下简称模型B)。三种模型的具体指标汇总如表3所示。

4.1　纵向对比实验

在纵向对比实验训练的过程中,当模型的各项损失不再显著下降时,停止训练并进行测试寻找具有最佳性能的轮数。COCO 2014测试集共有5 000张图片,为了更好地衡量Attention-YOLO模型在各个尺寸各个比例边界框下的定位及分类精准度,采用mAP@[0.5∶0.95]来综合评估模型的检测性能,表4的测试结果表明:即使模型A和模型B没有采用预训练权重,但是最终仍能



(a)x坐标误差　(b)y坐标误差　(c)边界框宽度误差

(d)边界框长度误差　(e)置信度损失　(f)分类损失

图11　加入两种注意力机制前后的COCO训练过程对比

表3　三种模型的具体指标汇总

检测算法	YOLOv3	Attention-YOLO-A	Attention-YOLO-B
初始化方法	预训练权重	MSRA init/预训练权重	MSRA init/预训练权重
多尺度训练	有	无	无
参数量	$6.194\,9\times10^7$	$6.281\,6\times10^7$	$6.281\,7\times10^7$
FPS	35	26	25
Inference Time/ms	29	39	40

表4　三种模型在COCO 2014验证集上的$mAP_{@[0.5:0.95]}$测试结果

IoU取值	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.5:0.95
YOLOv3	55.30	53.40	48.80	44.01	39.03	33.84	23.43	10.62	5.00	0.52	31.0
Attention-YOLO-A	55.40	53.22	49.98	45.92	40.83	34.10	25.35	14.62	5.42	0.74	32.6
Attention-YOLO-B	55.94	53.63	50.62	46.66	41.60	35.19	26.62	16.09	6.54	0.93	33.5

达到超越原始YOLOv3算法的mAP，这是由于并不一定要使用预训练权重初始化才能将模型训练至收敛，如果要从随机初始化训练模型至收敛，需要采用诸如MSRA initialization这样的参数初始化方法，并且采用一定形式的数据增强和更久的训练迭代次数。预训练权重虽然可以帮助加速各个指标损失值的降低，但是不采用预训练权重初始化对本实验没有明显的影响。同时，原YOLOv3算法采用了多尺度训练方法，每经过10次迭代就随机从320至608这10个输入大小(以32为间隔)中选择一个新的数作为网络的输入进行训练，这种做法虽然可以得到不同分辨率下的具有强鲁棒性的模型，但是由于训练条件较为苛刻，适用性不强。由于网络中存在全连接层，Attention-YOLO算法的输入大小固定为416×416，由注意力机制的作用及对残差连接的改进来弥补多尺度训练带来的效果。实验结果表明，模型A和模型B都取得了比原YOLOv3算法更高的mAP。由于两种注意力机制的叠加使用，模型B算法的检测精度高出原YOLOv3算法2.5 mAP，高出模型A算法0.9 mAP，但在前向传播速度上仅仅增加10 ms，参数量仅增加1.4%，仍然是实时的目标检测算法。

根据图12和图13所示，在目前性能优异的几种检测算法中，Attention-YOLO在综合性能上取得了较好的表现。纵向比较来看，Attention-YOLO相比其他两种前代算法都有不同程度的提升。

4.2　横向对比实验

由于用作横向对比实验的数据集PASCAL VOC规模相比COCO较小，因此采用预训练权重来进行初始化。同时训练的轮数也设置较少，但是由于数据集的训

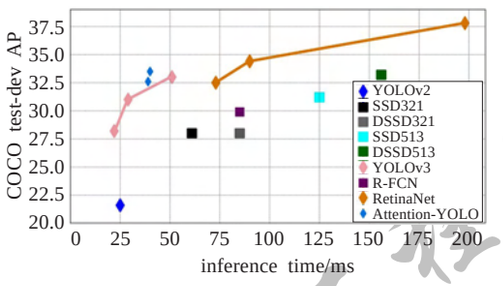


图12　COCO测试集上的time-AP对比图

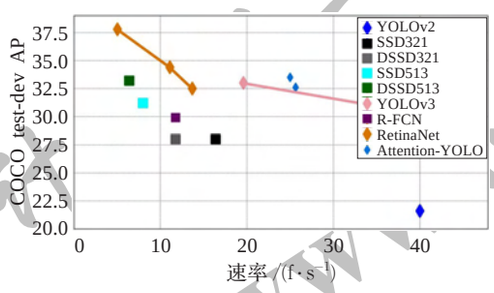


图13　COCO测试集上的速率-AP对比图

练难度低于COCO,理论上可以取得更显著的性能。PASCAL VOC 2007测试集的mAP计算采用的是11个点的精度-召回率插值法,最终Attention-YOLO-B算法在PASCAL VOC 2007测试集上取得了最高81.9 mAP的性能。表5的测试结果表明,Attention-YOLO算法在同类检测算法中取得了最好的检测精度,同时也较好地保证了算法的实时性。根据图14和图15中所示,得益于YOLOv3优秀的算法设计以及注意力机制和改进后

表5　在VOC 2007测试集上的测试结果对比

检测算法	mAP	速率/(f·s⁻¹)
Faster R-CNN(VGG-16)	73.2	7.0
Faster R-CNN(ResNet-101)	76.4	2.4
R-FCN(ResNet-101)	80.5	9.0
YOLOv1	63.4	45.0
YOLOv2	76.8	67.0
YOLOv2 544×544	78.6	40.0
SSD300	74.3	46.0
SSD321(ResNet-101)	77.1	11.2
SSD500	76.8	19.0
DSSD321	78.6	9.5
DSSD513	81.5	5.5
Attention-YOLO-A	81.7	26.0
Attention-YOLO-B	81.9	25.0

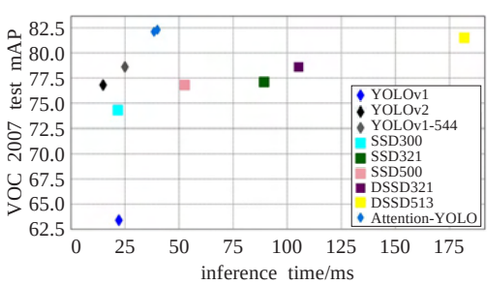


图14　VOC 2007测试集上的time-mAP对比图

的二阶残差连接,Attention-YOLO算法在检测精度上超过了同类型的回归型检测算法,在速度上快于SSD和DSSD系列的大多数算法。同时,Attention-YOLO算法在实时性和精确到了更好的平衡,相比起区域建议型检测算法具有更好的应用前景。

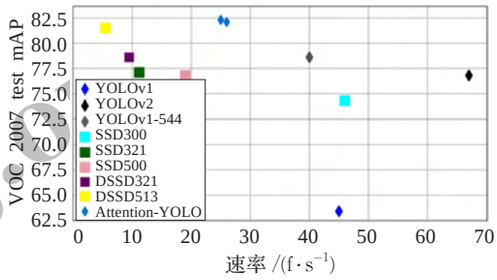


图15　VOC 2007测试集上的速率-mAP对比图

5　结论与展望

本文提出了一种引入注意力机制的YOLO检测算法:Attention-YOLO,该算法以较低的额外传播时间代价换取了检测精准度的提升。其主要思想是在保持回归型检测算法的能快速预测结果的前提下,通过对于所提取出的特征向量进行充分的利用和修正,同时修改残差连接结构中的单一连接方式,使整个网络能更好地筛选出有利于后续检测的特征向量。同时,Attention-YOLO中涉及的注意力机制和改进的特征融合方法可以迁移至其他具有残差连接的特征提取网络中,对于其他若干种注意力机制迁移至检测算法中的研究具有一定指导意义。其次,高效、低耗的检测算法能够更好地应用于智能设备或无人机之类的使用场景中,因此对回归型算法的改进研究也具有一定的现实意义。下一步工作将继续深入研究注意力机制的可视化机理,以更加直观的方式呈现模型内部的特征表示。

参考文献:

[1] 姚群力,胡显,雷宏.深度卷积神经网络在目标检测中的研究进展[J].计算机工程与应用,2018,54(17):1-9.

[2] 龙敏,佟越洋.应用卷积神经网络的人脸活体检测算法研究[J].计算机科学与探索,2018,12(10):1658-1670.

[3] Zhou Y,Tuzel O.VoxelNet:end-to-end learning for point cloud based 3D object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition,2018:4490-4499.

[4] LI B.3D fully convolutional network for vehicle detection in point cloud[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems,2017:1513-1518.

[5] 谢林江,季桂树,彭清,等.改进的卷积神经网络在行人检测中的应用[J].计算机科学与探索,2018,12(5):708-718.

[6] Sermanet P,Eigen D,Zhang X,et al.OverFeat:integrated recognition,localization and detection using convolutional

networks[C]//International Conference on Learning Representations,2014.

[7] Lowe D G.Distinctive image features from scale-invariant keypoints[J].International Journal of Computer Vision,2004,60(2):91-110.

[8] Wang X,Han T X,Yan S.An HOG-LBP human detector with partial occlusion handling[C]//IEEE International Conference on Computer Vision,2009:32-39.

[9] Viola P,Jones M.Rapid object detection using a boosted cascade of simple features[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2011:511-518.

[10] Felzenszwalb P F,Girshick R B,Mcallester D A.Cascade object detection with deformable part models[C]//IEEE Conference on Computer Vision and Pattern Recognition,2010:2241-2248.

[11] Chang C C,Lin C J.LIBSVM:a library for support vector machines[J].ACM Transactions on Intelligent Systems and Technology,2011,2(3):1-27.

[12] Redmon J,Divvala S,Girshick R,et al.You only look once:unified,real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:779-788.

[13] Redmon J,Farhadi A.YOLO9000:better,faster,stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017:6517-6525.

[14] Redmon J,Farhadi A.YOLOv3:an incremental improvement[J].arXiv:1804.02767,2018.

[15] Liu W,Anguelov D,Erhan D,et al.SSD:single shot multiBox detector[C]//European Conference on Computer Vision,2016:21-37.

[16] 高宗,李少波,陈济楠.基于YOLO网络的行人检测方法[J].计算机工程,2018,44(5):215-219.

[17] 吴天舒,张志佳,刘云鹏,等.基于改进SSD的轻量化小目标检测算法[J].红外与激光工程,2018,47(7):47-53.

[18] Ren S,He K,Girshick R,et al.Faster R-CNN:towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(6):1137-1149.

[19] Fu C Y,Liu W,Ranga A,et al.DSSD:deconvolutional single shot detector[J].arXiv:1701.06659,2017.

[20] He K,Zhang X,Ren S,et al.Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.

[21] Shen Z,Liu Z,Li J,et al.DSOD:learning deeply supervised object detectors from scratch[C]//IEEE International Conference on Computer Vision,2017:1919-1927.

[22] Huang G,Liu Z,Maaten L V D,et al.Densely connected convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017:2261-2269.

[23] Bodla N,Singh B,Chellappa R,et al.Soft-NMS—improving object detection with one line of code[C]//IEEE International Conference on Computer Vision,2017:5561-5569.

[24] 温捷文,战荫伟,凌伟林,郭灿樟.实时目标检测算法YOLO的批再规范化处理[J].计算机应用研究,2018,35(10):3179-3185.

[25] Lin T Y,Dollár P,Girshick R,et al.Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017:2117-2125.

[26] Hu J,Shen L,Sun G.Squeeze-and-excitation networks[J].arXiv:1709.01507,2017.

[27] Xie S,Girshick R,Dollar P,et al.Aggregated residual transformations for deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2017:1492-1500.

[28] Russakovsky O,Deng J,Su H,et al.ImageNet large scale visual recognition challenge[J].International Journal of Computer Vision,2014,115(3):211-252.

[29] Lin T Y,Maire M,Belongie S,et al.Microsoft COCO:common objects in context[J].arXiv:1405.0312,2014.

[30] Woo S,Park J,Lee J Y,et al.CBAM:convolutional block attention module[C]//The European Conference on Computer Vision,2018:3-19.

[31] Simonyan K,Zisserman A.Very deep convolutional networks for large-scale image recognition[J].arXiv:1409.1556,2014.

[32] Borji A,Itti L.State-of-the-art in visual attention modeling[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2013,35(1):185-207.

[33] Chikkerur S,Serre T,Tan C,et al.What and where:a Bayesian inference theory of attention[J].Vision Research,2010,50(22):2233-2247.

[34] Wang F,Tax D M J.Survey on the attention based RNN model and its applications in computer vision[J].arXiv:1601.06823,2016.

[35] Xu K,Ba J,Kiros R,et al.Show,attend and tell:neural image caption generation with visual attention[J].arXiv:1502.03044,2015.

[36] Vinyals O,Toshev A,Bengio S,et al.Show and tell:a neural image caption generator[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:3156-3164.

[37] Fu J,Zheng H,Mei T.Look closer to see better:recurrent attention convolutional neural network for fine-grained image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017:4476-4484.

(下转第125页)