

文章编号:1007-2780(2019)08-0825-09

基于注意力掩模融合的目标检测算法

董潇潇,何小海*,吴晓红,卿鄰波,滕奇志

(四川大学 电子信息学院,四川 成都 610065)

摘要:在计算机视觉任务中,平衡目标检测的精度与速度对于后续的实际应用如目标跟踪和识别起到关键作用。基于此,提出了一种基于注意力掩模融合的目标检测算法。首先,通过 VGG 网络提取特征,经过初步二分类和回归后得到一系列预选框;然后,将上述预选框输入到特征金字塔结构中,通过构建注意力掩模模块自适应地学习有效特征,同时融合特征金字塔结构与注意力掩模模块得到更具表征性的特征;最后经过多分类和回归得到多尺度的检测结果。在 PASCAL VOC 2007 和 PASCAL VOC 2012 数据集上展开了实验,测试集结果显示,在交集并集比(IOU)为 0.5 的条件下,对于 320×320 的图片输入,平均精度均值(mAP)分别为 81.0% 和 79.0%,检测速度为 60.9 fps。本文算法将注意力信息结合到目标检测中,实现了通用目标检测的精度和速度均衡。

关键词:计算机视觉;目标检测;注意力掩模;特征金字塔;多尺度检测

中图分类号:TP391 **文献标识码:**A **doi:**10.3788/YJYXS20193408.0825

Object detection algorithm based on attention mask fusion

DONG Xiao-xiao, HE Xiao-hai*, WU Xiao-hong, QING Lin-bo, TENG Qi-zhi

(College of Electronics and Information Engineering, Sichuan University,
Chengdu 610065, China)

Abstract: In computer vision tasks, balancing the accuracy and speed of object detection plays a significant role in subsequent practical applications such as object tracking and recognition. An object detection algorithm based on attention mask fusion is proposed. Firstly, the VGG network is used to extract features, and a series of preselected boxes are obtained after preliminary regression and binary classification. Then, the preselected boxes are input into the feature pyramid structure, learning effective features adaptively by constructing the attention mask module, and more representational features are gotten by integrating the feature pyramid structure and the attention mask module. Finally, the multiscale detection results are obtained by multiple classification and regression. Experiments are conducted on the data sets of PASCAL VOC 2007 and PASCAL VOC 2012. Test set results show that under the condition that Intersection over Union(IOU) is 0.5, the mean average precision(mAP) for the image input of 320×320 is 81.0% and 79.0% respectively, and the detection speed is 60.9 fps, realizing the balance between precision and speed. In this paper, the attention information is integrated

收稿日期:2019-03-20; **修订日期:**2019-04-29.

基金项目:成都市产业集群协同创新项目(No.2016-XT00-00015-GX);国家自然科学基金(No.61871278)

Supported by Chengdu Industrial Cluster Collaborative Innovation Project(No.2016-XT00-00015-GX); National Natural Science Foundation of China(No.61871278)

* 通信联系人, E-mail: nic5602@scu.edu.cn

into object detection to achieve the balance of accuracy and speed of generic object detection.

Key words: computer vision; object detection; attention mask; feature pyramid; multiscale detection

1 引言

目标检测是结合了目标定位和识别 2 个任务的一项基础性计算机视觉任务,它的目的是在图像的复杂背景中找出若干目标,给出一个精确的目标边框(Bounding box),并判断该边框中目标所属的类别^[1]。目标检测在人工智能和信息技术的诸多领域如机器人视觉、消费电子、自动驾驶、人机交互、基于内容的图像检索、智能视频监控^[2-3]等有着广泛的应用。

目前,基于深度学习的目标检测算法主要分为两类,一是单阶段方法,二是两阶段方法。两阶段方法将检测问题划分为两个过程,首先产生区域建议(Region proposals),然后对候选区域进行分类和边界框回归,这类算法的典型代表是基于区域建议的 R-CNN 系列算法,如 R-CNN^[4]、SP-Net^[5]、Fast R-CNN^[6]、Faster R-CNN^[7]、Mask R-CNN^[8]等。一阶段方法采用基于回归的思想,跳过区域建议阶段,直接产生物体的类别概率和位置坐标值,经过单次检测即可得出最终的检测结果,故在检测速度上有更好的结果,其代表性算法有:YOLO^[9]、SSD^[10]、Focal Loss^[11]等。SSD 结合了 YOLO 中的回归思想和 Faster-RCNN 中的 Anchor 机制,使用多尺度的特征图谱进行回归,既保持了 YOLO 速度快的特性,也保证了 Faster-RCNN 精度高的优点。随后为了提高 SSD 的精度和对小目标的检测能力,各种不同的方法被提出来。Fu 等^[12]为了提高浅层特征的代表能力,将基础网络由 VGG 替换为 ResNet,并借助反卷积模块以增加各层之间的上下文信息,提高了对小目标的检测能力,但是速度相对于 SSD 却大幅下降。Liu 等^[13]从人类视觉的感受野结构出发,提出了感受野模块(RFB),通过构建多分支卷积层和随后的空洞卷积层以加强网络学习到的深层特征。Zhang 等^[14]综合一阶段和两阶段目标检测的技巧,一方面引入两阶段中对包围框由粗到细的思想,一方面结合特征融合用于检测网络,取得了很好的效果。Liang 等^[15]利用多层卷积特征进行目标检测,其中利用

浅层卷积特征实现目标定位,利用深层卷积特征实现目标识别,以充分发挥各层卷积特征的作用。

近年来,注意力机制在计算机视觉中的应用逐渐增加。现有的视觉注意模型可分为软注意模型和强注意模型。软注意模型以确定性的方式预测注意区域,由于它是可微的,故可以使用反向传播进行训练^[16],让神经网络学到每一张新图片中需要关注的区域,因而形成了注意力。强注意模型随机预测图像的注意点,通常通过强化学习或最大化近似变分下界来训练。一般来说,软注意模型比强注意模型更有效,因为强注意模型需要抽样进行训练,而软注意模型可以进行端到端训练。Hu 等^[17]设计了目标关系模块,其注意力权重主要由外观特征权重和空间权重组成,通过联合其他目标的信息从而允许对它们之间的关系进行建模。Wang 等^[18]提出了一种基于 Attention 的残差学习方式,设计了一种可堆叠的网络结构,并从可堆叠的基本模块中引入了注意力机制,不同的注意力特征图谱可以捕捉到不同类型的注意力。Hu 等^[19]重点关注通道关系,显示地建模特征通道之间的依赖关系,自适应地重新校准通道特征响应,从而增强有用的特征而抑制冗余特征。综上,为了让目标检测重点关注有效信息,本文将注意力模型融入到目标检测中去,在 SSD 和 RefineDet 基础上提出了基于注意力掩模融合的改进的特征金字塔算法,主要用于改进当前目标检测算法缺乏视觉注意力信息的缺点。实验结果表明,本文算法进一步改进了目标检测的精度,同时也能保持高效的速度。

2 RefineDet

对于目标检测,两阶段方法(如 Faster RCNN)可以获得较高的精度,而一阶段方法(如 SSD)可以获得较快的速度,为了达到两者平衡,RefineDet 在 SSD 的基础上引入了两阶段中类似区域建议网络的结构并结合特征金字塔融合网络,既获得了较高的精度又保持了高效率。RefineDet 网络框架如图 1 所示,主要由 3 个模块组成,锚点精细模块(ARM)和目标检测模块

(ODM)以及传输连接模块(TCB)。ARM旨在消除消极的锚点(Anchor),以减少分类器的搜索空间,同时粗调锚点的位置和大小,为随后的回归提供更好的初始化;ODM将ARM产生的锚点作为输入,旨在回归精确的目标位置和预测多级标签;最后设计了一个TCB来传输ARM中的特

征,以预测ODM中对象的位置、大小和类标签。

RefineDet的损失函数主要包含ARM和ODM两方面。ARM包含二进制分类损失 L_b 和回归损失 L_r ,ODM包含多类别分类损失 L_m 和回归损失 L_r ,ARM和ODM两部分的损失函数一起前向传播,损失函数如下:

$$L(\{p_i\}, \{x_i\}, \{c_i\}, \{t_i\}) = \frac{1}{N_{\text{arm}}} (\sum_i L_b(p_i, [l_i^* \geq 1]) + \sum_i [l_i^* \geq 1] L_r(x_i, g_i^*)) + \frac{1}{N_{\text{odm}}} (\sum_i L_m(c_i, l_i^*) + \sum_i [l_i^* \geq 1] L_r(t_i, g_i^*)) \quad (1)$$

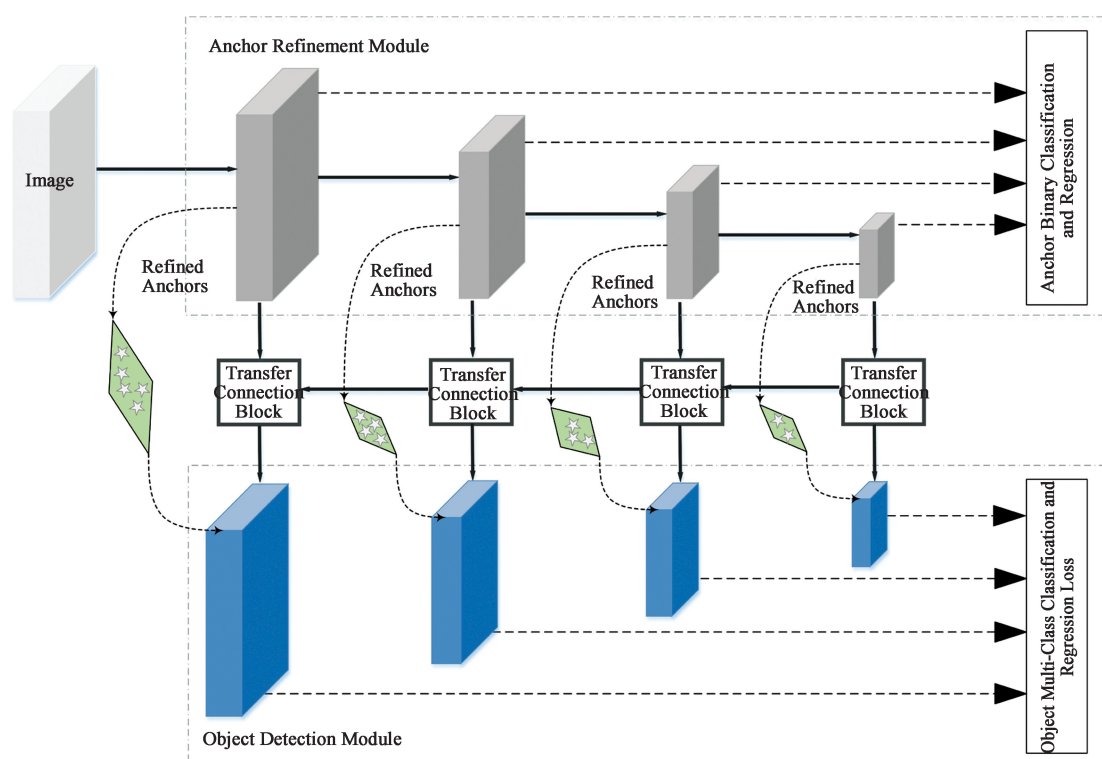


图1 RefineDet 网络结构

Fig.1 Network of RefineDet

3 本文算法

3.1 网络结构

算法的整体网络结构如图2所示。以输入尺寸为 320×320 的图像为例:

(1)使用基准网络VGG16^[20]提取特征,经过卷积操作和6次下采样操作后,得到多尺度的卷积特征层。

(2)使用基本的卷积层conv4_3,conv5_3,fc7以及conv6_2进行初步的二分类和包围框回归操作,以得到预选框,减少负样本搜索。

(3)将注意力掩模模块Attention Mask Module(AMM)和特征金字塔结构结合起来,即将顶层特征同底层特征融合后,再辅之以注意力信息AMM,得到更加有效的特征P1,P2,P3,P4。

(4)最后将上述特征直接用于多分类和回归操作,不断训练迭代模型,得到最终的检测结果。

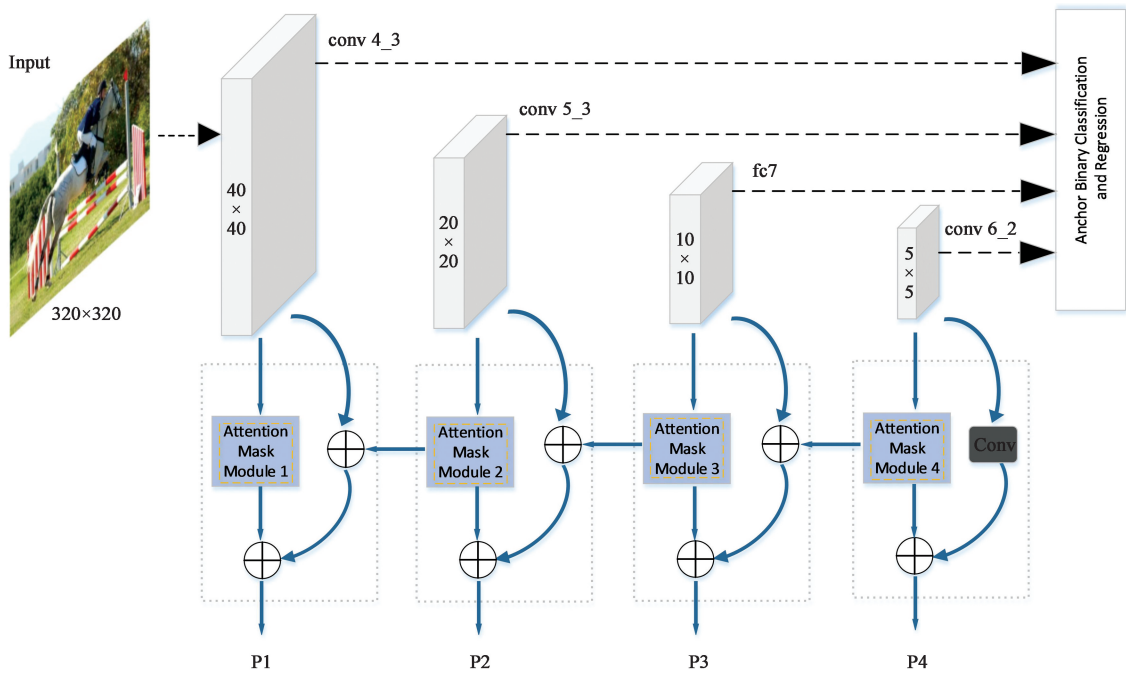


图 2 本文算法网络结构

Fig.2 Network of proposed algorithm

3.2 注意力掩模模块

观察一幅图片时,人脑并不均衡地关注整幅图片,而是带有一定权重分布的。对于深度学习中的目标检测任务而言,想要检测出相关的人或者特定物体,将其与背景天空等区分开来,同样需要有一定的选择性侧重。为此,引入文献[18]中的注意力模型,将原本的残差结构改为基本的神经网络结构,并去除了一些卷积层,以便更加适应本文的网络训练。

注意力掩模模块(AMM)的结构图如图3所示。在本文中,使用基本的卷积函数(conv),激活函数(relu)和池化函数(pool)来构建基本网络。每个注意力掩模模块分为两个分支:掩模分支和主干分支。给定输入 x ,主干分支经过卷积后得到输出 $G(x)$,掩模分支经过卷积和反卷积操作得到相同尺寸的输出 $M(x)$ 。掩模分支中,首先经过两次conv,relu,pool操作,然后经过两次deconv操作得到特征图谱 $S(x)$,最后经过sigmoid函数得到输出的掩模:

$$M(x) = \text{sigmoid}(S(x)) = \frac{1}{1 + e^{-S(x)}}. \quad (2)$$

输出掩模类似于 Highway Network^[21]中的门控单元,控制着主干分支的神经元,可以自适应地学习到有效的特征并且抑制噪声。主干分支

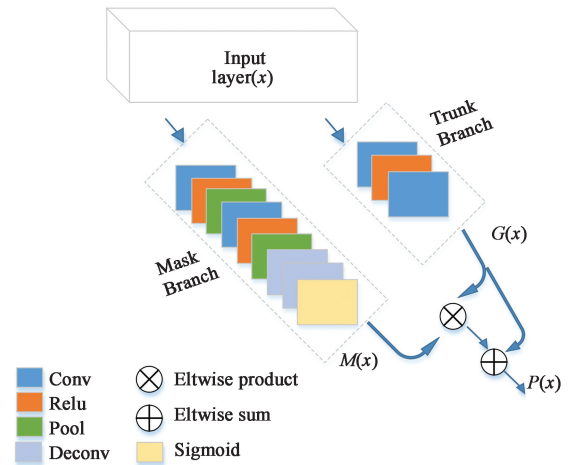


图 3 注意力掩模模块

Fig.3 Attention mask module

中,经过两次conv函数得到主干输出。最后输出的注意力模型 $P(x)$ 公式如下所示:

$$P_{i,c}(x) = (1 + M_{i,c}(x)) * G_{i,c}(x), \quad (3)$$

式中: i 代表神经元的空间位置, c 代表神经元的通道位置。类似于残差网络中的恒等映射, $M(x)$ 的输出范围为 $[0,1]$,当 $M(x)$ 接近0时, $P(x)$ 的输出接近原始的特征 $G(x)$ 。

在注意力掩模模块中,注意力掩模不仅在向前传播中充当特征选择器,在反向传播中也能进

行梯度更新。特征掩模的梯度如下所示:

$$\frac{\partial(1+M(x,\theta)T(x,\theta))}{\partial\varphi} = (1+M(x,\theta)) \frac{\partial T(x,\theta)}{\partial\varphi}, \quad (4)$$

其中: θ 是掩模分支参数, φ 是主干分支参数。

3.3 改进的特征金字塔融合

SSD 网络虽然取得了较好的效果,但由于其只在若干个特征图谱上进行检测,并没有结合不同尺寸的特征,导致低层的特征图谱缺乏语义信息,这样不利于小目标的检测。而 RefineDet 网络采取侧边连接的方法,将高层语义信息同低层定位信息融合在一起,可以进行多种尺度的检测,仍然缺乏注意力信息。本文提出了融合不同特征图谱与注意力掩模的结构,不仅能进行多尺度检测,还可以结合注意力模型关注感兴趣区域,从而有效地检测出所有尺度物体。具体网络结构如图 4 所示。

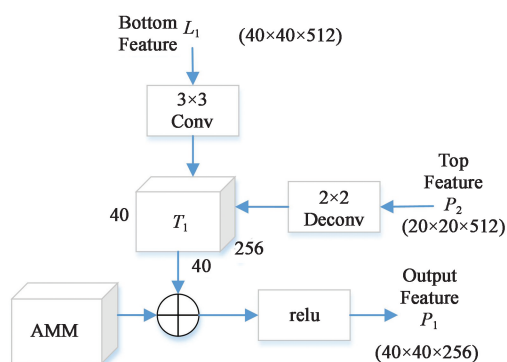


图 4 特征金字塔融合结构

Fig.4 Feature pyramid fusion structure

输入的特征图谱为 $(H \times W \times C)$, 其中 H 和 W 代表高和宽, C 代表通道的个数。对于底部特征 L_1 , 首先经过 3×3 的卷积层进行特征降维, 将维度从 512 降到 256, 对于检测特征 P_2 , 经过 2×2 的反卷积操作, 在扩大特征图谱同时也进行了降

维的操作, 得到的输出为 $(40 \times 40 \times 256)$; 然后将上述输出进行 eltwise sum 操作, 得到融合后的特征图谱 T_1 亦为 $(40 \times 40 \times 256)$; 最后将注意力掩模模块同 T_1 融合以增加注意力信息, 经过 Relu 层后得到的检测特征 P_1 可应用于检测系统。

若图片中存在多种尺度的目标, 低层的特征如 P_1 和 P_2 可以用来检测小目标, 而高层的特征如 P_3 和 P_4 具有更大的感受野, 负责检测大目标。

4 实 验

本次实验在以下两个公开数据集中开展: PASCAL VOC^[22] 2007 和 PASCAL VOC 2012。PASCAL VOC 数据集包含 20 类物体, 包括飞机、自行车、鸟等。本文的实验平台为: Ubuntu 16.04.5, Nvidia GTX 1080Ti GPU, Intel(R) Xeon(R) CPU E5-2686。采用的深度学习框架为 Caffe, 采用的精度评价指标为 mAP (Mean average precision), 速度评价指标为 fps (Frames per second)。

4.1 VOC2007 数据集

VOC2007 数据集的训练集有 5 011 张, 测试集有 4 952 张。本次实验在 VOC2007 trainval 和 VOC2012 trainval 上进行训练, 在 VOC2007 test 数据集上测试。所有的方法在 VGG16 基准网络中进行预训练。如图 2 所示, 使用 conv4, conv5, fc7 以及新增的 conv6 层进行定位和置信度的预测。使用 Multistep 方式来调整学习率, 即在前 80k 次迭代中学习率设置为 10^{-2} , 随后 20k 和 40k 次迭代中学习率分别设置为 10^{-3} 和 10^{-4} 。输入图片尺寸为 320×320 时, 训练时 batch size 设置为 16, GPU 数量为 1; 当图片尺寸为 512×512 时, 由于一块 GPU 显存不够, 故采用两块 GPU 进行联合训练。测试时 batch size 设置为 1。

表 1 不同算法在 VOC2007 和 VOC2012 数据集上的测试结果

Tab.1 Test results of different algorithms on VOC2007 and VOC2012 datasets

Method	Backbone	Input Size	FPS	GPU	mAP@IOU=0.5 (%)	
					VOC2007	VOC2012
Faster RCNN ^[7]	VGG-16	$\sim 1\,000 \times 600$	7	Titan X	73.2	70.4

续 表

Method	Backbone	Input Size	FPS	GPU	mAP@IOU=0.5 (%)	
					VOC2007	VOC2012
Faster RCNN ^[23]	ResNet-101	~1 000×600	2.4	K40	76.4	73.8
HyperNet ^[24]	VGG-16	~1 000×600	5	Titan X	76.3	71.4
R-FCN ^[25]	ResNet-101	~1 000×600	5.9	K40	80.5	77.6
CoupleNet ^[26]	ResNet-101	~1 000×600	8.2	Titan X	82.7	80.4
YOLO ^[9]	GooleNet	448×448	45	Titan X	63.4	57.9
YOLOv2 ^[27]	DarkNet-19	544×544	40	Titan X	78.6	73.4
SSD300 ^[10]	VGG-16	300×300	46	Titan X	74.3	72.4
SSD512 ^[10]	VGG-16	512×512	19	Titan X	76.8	74.9
RefineDet320 ^[14]	VGG-16	320×320	40.3	Titan X	80.0	78.1
RefineDet512 ^[14]	VGG-16	512×512	24.1	Titan X	81.8	80.1
RefineDet320 ^[14]	VGG-16	320×320	74.6	1080Ti	80.0	78.1
Proposed320	VGG-16	320×320	60.9	1080Ti	81.0	79.0
Proposed512	VGG-16	512×512	35.0	1080Ti	82.2	80.8

表 2 不同算法在 VOC2007 数据集上 20 类别的 AP 比较

Tab.2 AP comparison of 20 categoriesfor different algorithms on VOC2007 dataset

method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster RCNN ^[23]	76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	89.8	56.7	87.8	69.4	88.3	88.9	80.9	78.4	41.7	78.6	79.8	85.3	72.0
HyperNet ^[24]	76.3	77.4	83.3	75.0	69.1	62.4	83.1	87.4	87.4	57.1	79.8	71.4	85.1	85.1	80.0	79.1	51.2	79.1	75.7	80.9	76.5
R-FCN ^[25]	80.5	79.9	87.2	81.5	72.0	69.8	86.8	88.5	89.8	67.0	88.1	74.5	89.8	90.6	79.9	81.2	53.7	81.8	81.5	85.9	79.9
SSD300 ^[10]	74.3	75.5	80.2	72.3	66.3	47.6	83.0	84.2	86.1	54.7	78.3	73.9	84.5	85.3	82.6	76.2	48.6	73.9	76.0	83.4	74.0
SSD512 ^[10]	76.8	82.4	84.7	78.4	73.8	53.2	86.2	87.5	86.0	57.8	83.1	70.2	84.9	85.2	83.9	79.7	50.3	77.9	73.9	82.5	75.3
RefineDet320 ^[14]	80.0	83.9	85.4	81.4	75.5	60.2	86.4	88.1	89.1	62.7	83.9	77.0	85.4	87.1	86.7	82.6	55.3	82.7	78.5	88.1	79.4
RefineDet512 ^[14]	81.8	88.7	87.0	83.2	76.5	68.0	88.5	88.7	89.2	66.5	87.9	75.0	86.8	89.2	87.8	84.7	56.2	83.2	78.7	88.1	82.3
Proposed320	81.0	84.8	85.8	81.4	76.0	60.6	88.0	88.3	88.9	63.8	85.5	78.7	86.3	87.5	87.3	82.8	58.4	83.6	82.3	88.7	80.6
Proposed512	82.2	89.4	86.5	81.9	77.2	70.0	89.1	89.1	89.5	67.6	88.0	75.5	87.3	88.4	88.0	84.9	59.0	83.5	78.6	87.7	82.7

如表 1 所示,将本文的算法与其他主流的目标检测算法进行对比。对于 320×320 的输入,mAP 达到了 81%,与 RefineDet 320 相比提高了 1%,与 SSD 相比提高了 6.7%。由于本文测试所使用的 GPU 不同于 RefineDet,故将 RefineDet 中在 VOC07+12 数据集上训练的 320×320 的模型在 1080Ti 上进行测试,观察得到,本文算法的速度同其相比降低了 18.4%。在速度和精度的均衡下,本文算法仍能取得较好的效果。对于

512×512 的输入,mAP 为 82.2%,比 RefineDet 的精度提高了 0.4%,这也证明了本文算法的有效性。同精度更高的基于区域的目标检测方法进行比较,可以看到,本文算法在尺寸小于 CoupleNet 的情况下,mAP 几乎能够达到 CoupleNet 的效果。

表 2 中列举出各类算法中 20 类目标的 AP (Average precision),可以看到大部分类别的平均精度都有提升。

4.2 VOC2012 数据集

VOC2012 数据集的训练集有 11 540 张,测试集有 10 991 张。本次实验在 VOC2007 trainval 和 VOC2012 trainval 上进行训练,在 VOC2012 test 数据集上测试。使用 Multistep 方式来调整学习率,即在前 160k 次迭代中学习率设置为 10^{-2} ,随后 40k 和 80k 次迭代中学习率分别设置为 10^{-3} 和 10^{-4} 。对于 320×320 和 512×512 的输入,训练的 batch size 设置为 16,并且使用两块 GPU 进行训练。其余实验设定和 4.1 中相同。

如表 1 所示,Proposed320 mAP 为 79.0%,与 RefineDet 320 相比提高了 0.9%,与 SSD 相比提高了 6.6%。同二阶段方法相比,本文算法比精度最高的 CoupleNet 低了 1.4%。Proposed512 mAP 为 80.8%,与 RefineDet 320 相比提高了 0.7%,且精度超过了 CoupleNet,为所提算法中最高。

4.3 定性结果

图 5 给出了本文算法模型在 VOC2007 和 VOC2012 数据集上的具体测试结果。

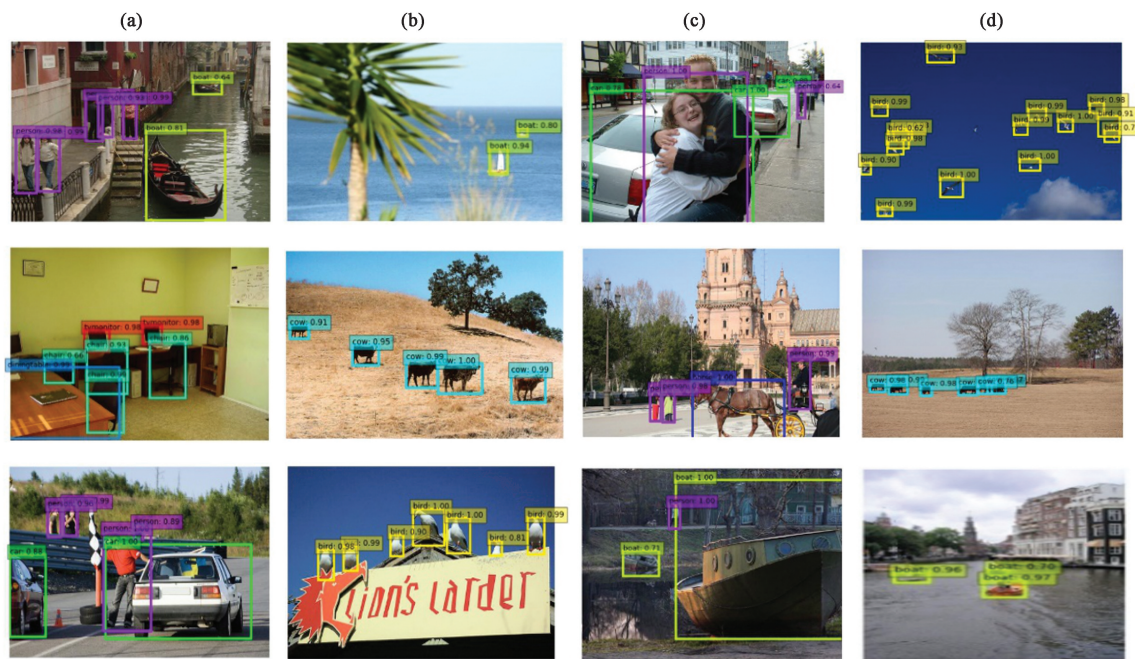


图 5 VOC2007 和 VOC2012 数据集的测试结果。(a) VOC2007 测试结果,强调检测的多尺度效果;(b) VOC2007 测试结果,重点检测小目标;(c) VOC2012 测试结果,强调检测的多尺度效果;(d) VOC2012 测试结果,重点检测小目标。

Fig.5 Test result on VOC2007 and VOC2012 dataset. (a) Test result on VOC2007, emphasizing the multi-scale effect of detection; (b) Test result on VOC2007, focusing on small targets; (c) Test result on VOC2012, emphasizing the multiscale effect of detection; (d) Test result on VOC2012, focusing on small targets.

5 结 论

本文提出了一种基于注意力掩模融合的目标检测算法,通过神经网络提取特征得到多尺度卷积特征图谱,初步进行二分类和回归,在特征金字塔结构中结合注意力掩模模块进行多尺度的检

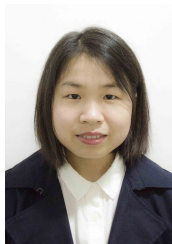
测。实验结果证明:对于 320×320 的图片输入,本文算法在 VOC2007 和 VOC2012 数据集上达到了 81.0% mAP 和 79.0% mAP,检测速度为 60.9 fps;对于 512×512 的图片输入,本文算法在 VOC2007 和 VOC2012 数据集上达到了 82.2% mAP 和 80.8% mAP,检测速度为 35.0 fps,实现了精度和速度的均衡要求。

参 考 文 献:

- [1] 黄继鹏, 史颖欢, 高阳. 面向小目标的多尺度 Faster-RCNN 检测算法[J]. 计算机研究与发展, 2019, 56(2): 319-327.
HUANG J P, SHI Y H, GAO Y. Multi-scale Faster-RCNN algorithm for small object detection [J]. *Journal of Computer Research and Development*, 2019, 56(2): 319-327. (in Chinese)
- [2] 龙思源, 张葆, 宋策, 等. 基于改进的加速鲁棒特征的目标识别[J]. 中国光学, 2017, 10(6): 719-725.
LONG S Y, ZHANG B, SONG C, et al. Object detection based on improved speeded-up robust features [J]. *Chinese Optics*, 2017, 10(6): 719-725. (in Chinese)
- [3] LIU L, OUYANG W L, WANG X G, et al. Deep learning for generic object detection: a survey [J]. *arXiv preprint arXiv:1809.02165*, 2018.
- [4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, 2014: 580-587.
- [5] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904-1916.
- [6] GIRSHICK R. Fast R-CNN [C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Santiago, Chile: IEEE, 2015: 1440-1448.
- [7] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.
- [8] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE, 2017: 2980-2988.
- [9] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016: 779-788.
- [10] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multiBox detector [C]//*Proceedings of 14th European Conference on Computer Vision*. Amsterdam, The Netherlands: Springer, 2016: 21-37.
- [11] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE, 2017: 2999-3007.
- [12] FU C Y, LIU W, RANGA A, et al. DSSD: deconvolutional single shot detector [J]. *arXiv preprint arXiv:1701.06659*, 2017.
- [13] LIU S T, HUANG D, WANG Y H. Receptive field block net for accurate and fast object detection [C]//*Proceedings of 15th European Conference on Computer Vision*. Munich, Germany: Springer, 2018: 404-419.
- [14] ZHANG S F, WEN L Y, BIAN X, et al. Single-shot refinement neural network for object detection [C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 4203-4212.
- [15] 梁华, 宋玉龙, 钱锋, 等. 基于深度学习的航空对地小目标检测[J]. 液晶与显示, 2018, 33(9): 793-800.
LIANG H, SONG Y L, QIAN F, et al. Detection of small target in aerial photography based on deep learning [J]. *Chinese Journal of Liquid Crystals and Displays*, 2018, 33(9): 793-800. (in Chinese)
- [16] ZHAO B, WU X, FENG J S, et al. Diversified visual attention networks for fine-grained object classification [J]. *IEEE Transactions on Multimedia*, 2017, 19(6): 1245-1256.
- [17] HU H, GU J Y, ZHANG Z, et al. Relation networks for object detection [C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 3588-3597.
- [18] WANG F, JIANG M Q, QIAN C, et al. Residual attention network for image classification [C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 6450-6458.
- [19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.

- [20] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C]//*Proceedings of the International Conference on Learning Representations*. Banff, Canada, 2015.
- [21] SRIVASTAVA R K, GREFF K, SCHMIDHUBER J. Highway networks [J]. *arXiv preprint arXiv*:1505.00387, 2015.
- [22] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, *et al.* The pascal visual object classes (VOC) challenge [J]. *International Journal of Computer Vision*, 2010, 88(2): 303-338.
- [23] HE K M, ZHANG X Y, REN S Q, *et al.* Deep residual learning for image recognition [C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016: 770-778.
- [24] KONG T, YAO A B, CHEN Y R, *et al.* HyperNet: towards accurate region proposal generation and joint object detection [C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016: 845-853.
- [25] DAI J F, LI Y, HE K M, *et al.* R-FCN: object detection via region-based fully convolutional networks [C]//*Proceedings of the 30th International Conference on Neural Information Processing Systems*. Barcelona, Spain: Curran Associates Inc., 2016: 379-387.
- [26] ZHU Y S, ZHAO C Y, WANG J Q, *et al.* CoupleNet: coupling global structure with local parts for object detection [C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE, 2017: 4146-4154.
- [27] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 6517-6525.

作者简介:



董潇潇(1994—),女,湖北黄冈人,硕士研究生,2016年于湖北工业大学获得学士学位,研究方向:数字图像处理、目标检测与识别。E-mail: alyssa21@163.com



何小海(1964—),男,四川绵阳人,教授,博士生导师,2002年于四川大学获得博士学位,研究方向:图像处理与网络通信、机器视觉与智能系统等。E-mail: nic5602@scu.edu.cn