高技术通讯 2024 年 第34 卷 第3 期:233-247

doi:10.3772/j.issn.1002-0470.2024.03.002

基于注意力与密集重参数化的目标检测算法①

陈志旺②[*,**] 雷春明[*] 吕昌昊[***] 王 婷[*] 彭 勇[****]

([*]燕山大学智能控制系统与智能装备教育部工程研究中心 秦皇岛 066004)

([**]燕山大学工业计算机控制工程河北省重点实验室 秦皇岛 066004)

([***]燕山大学河北省电力电子节能与传动控制重点实验室 秦皇岛 066004)

([****]燕山大学电气工程学院 秦皇岛 066004)

摘 要 针对目标检测任务中背景复杂、目标尺寸差异大等因素导致目标检测结果较差的问题,本文提出基于注意力和密集重参数化的目标检测算法。首先,基于 CSP-DarkNet 提出高效的特征提取网络,主要包括密集重参数化模块和 CASA 模块 2 个设计。前者利用密集连接保留浅层特征,又通过重参数化结构降低网络复杂度;后者 CASA 模块用于获取需要的目标信息。其次,特征融合在特征金字塔(FPN)和路径聚合网络(PAN)的基础上,引入内容感知特征重组(CARAFE)进行上采样,有效解决了邻近插值法等未能捕捉丰富语义信息的问题;提出更高效的 C3-G 模块,获取丰富的梯度信息,增强模型表达能力和感知能力;同时,引入深度可分离卷积提升运算效率。最后,检测输出采用在更大范围上跨领域正负样本匹配策略扩充正样本数量,提升检测效果。该算法在 MS COCO 和 PASCAL VOC 数据集上的 mAP@0.5 分别达到了 57.5% 和 83.0%,充分说明了本文算法的先进性。

关键词 目标检测;重参数化;注意力机制;特征融合;上采样;正负样本匹配

目标检测属于计算机视觉领域中的一个关键问题,主要结合目标定位和识别检测两项技术,在给定图像中实现目标边框的准确定位并检测出该目标所属的具体类别。根据目标检测技术的发展历程,目标检测可以分为基于手工提取特征和基于深度学习提取特征两大类。然而,基于手工提取特征的方法步骤繁琐且窗口冗余问题严重,同时存在着检测速度慢和检测精度低等问题[1]。随着深度学习技术应用范围逐渐增大,研究人员开始将卷积神经网络(convolutional neural networks,CNN)[2]用于目标检测领域方面的研究,这种基于深度特征的目标检测技术已经成为当前最具代表性的方法之一。

相较于基于手工提取特征的目标检测方法,基于深度学习提取特征的目标检测方法能更准确地学习到图像更深层的语义特征,特征表达能力更强且目标识别精度更高,已经成为目标检测领域的主流方式。

尽管目标检测技术已经取得了很大的进展,但在实际目标检测任务中仍然存在着多种挑战。其中包括复杂背景下噪声干扰的问题、对不同尺寸目标准确检测的问题,以及在检测精度和速度之间平衡的困境。为了应对这些问题,研究者们提出了多种解决方案,如增加训练数据集的数量、采用多尺度训练(multi scale training,MST)[3]、深层特征与浅层特征融合、优化网络结构(如 ResNet101、ResNet152)[4]等方法。但应用这些方法时需要注意,过大规模的

① 国家自然科学基金(61573305)和河北省自然科学基金(F2022203038,F2019203511)资助项目。

② 男,1978 年生,博士,副教授;研究方向:多旋翼飞行控制,计算机视觉;E-mail:czwaaron@ysu.edu.cn。

(收稿日期:2023-08-16)

数据集会增加训练和评估模型的计算资源需求,从而导致训练时间变长和成本增加。更深的网络模型可能会带来过拟合和可解释性差等问题。

针对上述问题,本文提出基于注意力和密集重参数化的一阶段轻量化级目标检测算法,该算法网络由特征提取、特征融合和检测输出 3 个部分组成。其中,特征提取采用 CSP-Darknet 网络,并引入密集重参数化结构和 CASA(coordinate and spatial attention)模块。CSP-Darknet 是一种高效的特征提取网络。密集重参数化在浅层特征提取过程中避免了目标特征丢失,并显著减少推理过程中的计算量。CASA 模块对不同尺度的特征图进行信息筛选。特征融合采用特征金字塔(feature pyramid network,FPN)结构与路径聚合网络(path aggregation network,PAN)结构的组合。为了进一步提高网络表达能力和感知能力,本文提出了梯度流更加丰富的 C3-G 模块。同时引入深度可分离卷积加快网络推理速度,并采用内容感知特征重组(content-aware reassembly of features,CARAFE)模块[5],有效地保留上采样过程中目标的细节信息和位置信息,以提升网络检测效果。检测输出方面,通过卷积层在不同尺度特征图上的预测结果,采用范围更广的跨领域正负样本匹配策略,扩充正样本数量,提升训练效果。

综上所述,本文的主要贡献总结如下。

(1) 基于 CSP-DarkNet 提出一种高效的特征提取网络,该网络的密集重参数化结构既能充分提取和保留特征又保证了模块大小在一个合适的范围,同时对不同尺寸的特征图采用 CASA 模块过滤背景信息。

(2) 特征融合采用 FPN 结构和 PAN 结构为框架,提出 C3-G 高效模块进一步提高网络表达能力和感知能力,同时通过更有效的上采样,缓解了特征信息丢失的问题。

(3) 检测输出采用多尺度预测方式,同时为了提升训练效果,采用更合理的正负样本匹配策略。

1 相关工作

1.1 基于深度学习的目标检测算法

基于深度学习的目标检测算法主要分为两阶段(two-stage)的目标检测算法和一阶段(one-stage)的目标检测算法。

两阶段目标检测算法的核心思想是逐步细化,首先从输入图像中生成大量可能存在目标的候选区域,然后对这些候选区域的特征进行目标定位和分类[6],从而实现目标检测。其中,具有代表性的两阶段目标检测算法有区域卷积神经网络(region convolutional neural network, R-CNN)[7]、Faster R-CNN[8]等。R-CNN 通过使用选择性搜索(selective search)的方法生成候选框,采用卷积神经网络对候选区域进行特征提取,训练支持向量机(support vector machine,SVM)分类器进行分类。但 R-CNN 仍然存在训练步骤繁琐、检测精度低等问题。Faster R-CNN[8]提出区域生成网络(region proposal network, RPN),能够自动生成候选区域,进一步提升算法效率。

与两阶段算法相比,一阶段目标检测算法结构简单,对输入的图像直接预测目标位置和类别,避免了候选区域产生[9]。其中具有代表性的一阶段目标检测算法有 YOLO(you only look once)[10]系列算法和 SSD (single shot detection)[11] 算法等。YOLOv1[10]将目标检测视为回归问题,并采用统一的框架从图像中提取特征,直接预测出目标的位置和类别,实现端到端检测。随后的 YOLOv2 采用 Darknet-19 作为特征提取网络,降低模型复杂度,并使用批归一化(batch normalization, BN)[12]、多尺度训练和卷积预测边界框等策略,进一步提升目标检测精度。YOLOv3[13]则采用特征提取能力更强的主干网络 Darknet-53,并引入 FPN 结构以提升网络空间表征能力,使其在小目标检测方面效果更好。YOLOv4[14]通过结合加权残差连接(weighted residual connections)和自对抗训练(self-adversarial training)[14]等技术,进一步提高了目标检测精度。随后,YOLOv5 引入自适应 Anchor 计算和 GIoU(generalized intersection over union)[15]度量预测框损失等方法,提升了目标检测的速度和精度。YOLOv6[16]通过优化特征提取网络结构,并设计了高效的解耦头,从而提升了模型的收敛速度。YOLOv7[17]采用高效聚合网络作为特征提取器,并引入辅助训练模

块来指导标签分配策略[18]。YOLOv8 在 YOLOv5 的基础上将 C3 结构更换为梯度流更丰富的 C2f 结构,并对不同尺度的模型进行通道数的调整,以适应不同类型的检测任务。SSD 借鉴了 Faster R-CNN 的 Anchor 技术并引入先验框(prior box)[8]技术,此外,还基于 FPN 结构在不同尺度的特征图上进行目标预测。一阶段目标检测方法具有检测速度快的特点,因此,受限于计算资源,本文所提算法采用一阶段目标检测方法。

1.2 注意力机制

卷积神经网络中的注意力机制(attention mechanism)[19]通过对全局特征图赋予一个权重,增大不同特征之间的梯度,使得网络在后续的特征提取过程中重点关注到含有目标信息的区域,有效增强模型学习特征的能力。现有的目标检测算法倾向于在全局中提取特征,而忽略了局部的细节信息。而注意力机制可以更好地选择性地关注某些区域。Hu 等人[20]在卷积神经网络中引入通道注意力机制,通过对空间维度上的压缩后再膨胀的操作增强通道上的重要特征,从而让提取到的特征指向性更强。WOO 等人[21]通过将通道注意力与空间注意力串联组合的方式加强了特征提取网络在通道和空间上的提取特征能力。

1.3 轻量化网络

轻量化网络是在保证算法模型精度的基础上进一步降低算法复杂度。轻量化网络既包含了对网络结构削减优化的探索,又有诸如知识蒸馏、剪枝、结构重参数化等模型压缩方面的应用。

在网络结构削减设计方面,Han 等人[22]通过采用小卷积核、削减特征图通道数、延迟下采样等办法降低网络的可学习参数量和整体计算量。Andrew 等人[23]则采用深度可分离卷积代替原始卷积,在不改变原始感受野大小、输入输出不变的情况下降低了网络计算量。

Hinton 等人[24]提出知识蒸馏的概念,其核心是使用一个已经训练好的大模型(teacher 模型)来指导训练一个小模型(student 模型),并使得 student 模型能够近似地复现 teacher 模型的输出结果。Liu 等人[25]应用剪枝技术,针对网络中存在的大量接近 0 的冗余参数且去掉后不影响模型表达能力的问题,通过采用对网络通道稀疏化的方法达到模型快速推理的效果,同时通过将网络的归一化层与通道缩放因子相结合,既不增加额外的计算开销又能达到通道剪枝减小网络模型的目的。

Ding 等人[26]在 2021 年提出结构重参数化(structural rep-param)结构,其主要思想是首先构造一系列用于训练的结构,并将其参数等价转换为另一组用于推理或者部署的参数,从而将这一系列结构等价转换为另一系列结构。

1.4 正负样本匹配策略

正负样本匹配策略是解决正负样本不均衡导致训练效果不佳的有效方法。目标检测领域的正负样本匹配策略主要分为 2 大类:固定标签分配(fixed label assignment)和动态标签分配(dyanmic label assignment)。

固定标签分配主要是基于距离、交并比(intersection over union,IoU)等先验知识设置固定阈值区分正负样本。Faster R-CNN 中计算所有预测框和每个真实框的 IoU[8],对于每个预测框,找到与它最匹配的真实框对应的最大 IoU,若该最大 IoU 超过正样本阈值,则被指定为正样本;若该最大 IoU 低于负样本阈值,则被指定为负样本。YOLOv5 中采用预设 anchor box 和真实框的宽高比匹配度作为划分规则[15],同时根据真实框中心点的位置引入跨领域网格策略增加正样本数量。

动态标签分配则根据不同策略动态设置阈值选择正负样本。YOLOv6 中根据每个真实框的大小确定负责的预测框个数[16],计算每个真实框对所有负责的预测框的代价函数,并选取该真实框的代价函数值最小的前 10 个位置的预测框作为候选框,再计算所有候选框与真实框的交并比,并求和取整数 x,将交并比最大的前 x 个预测框作为正样本。

2 算法整体结构

如图 1 所示,本文算法分为特征提取、特征融合、检测输出 3 个部分。其中特征提取在 CSP-DarkNet 网络基础上,引入了密集重参化结构和 CASA 模

块,提升网络的特征提取能力。特征融合采用 FPN 结构和 PAN 结构,提出 C3-G 高效模块进一步提高模型性能。通过使用深度可分离卷积,提升算法检测的实时性。同时,采用了内容感知特征重组的上采样方法,避免在上采样过程中过多的语义信息丢失。检测输出通过卷积分别对不同尺寸特征图输出预测结果,再通过正负样本匹配策略进行筛选和损失计算。

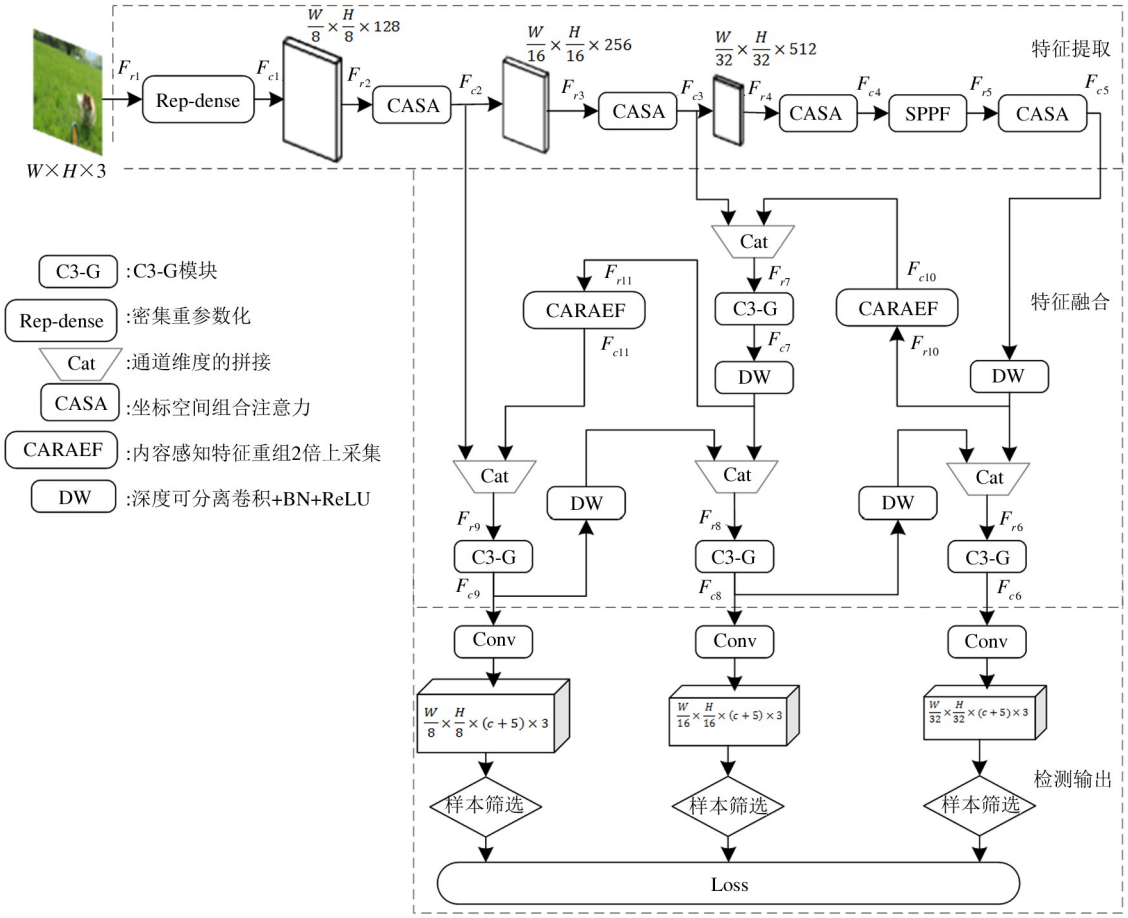


图 1　本文算法整体结构图

2.1　特征提取

CSP-DarkNet 网络使用 CSP 结构可以避免不同层学习到重复的梯度特征[13],同时,CSP-DarkNet 通过 5 次下采样减少网络计算量和增大感受野。本文在其基础上通过引入密集重参数化结构聚合浅层特征,CASA 模块对下采样中的不同尺寸特征图进行背景信息筛选,获取特征的通道和空间信息。

2.1.1　密集重参数化

密集连接结构[27]具有比残差连接结构[4]更强的特征提取能力、更少的参数以及更少的计算量,但由于密集连接结构高内存访问成本和能耗导致推理速度很慢。算法推理速度除了网络模型大小的影响以外,还受到内存访问成本和图形处理器(graphics processing unit,GPU)计算速度的制约。Ma 等人[28]通过对卷积层内存访问成本的计算分析得出,当计算量不变时,输入输出通道的比值越趋近 1:1 时内存访问成本越小,卷积层速度越快,由此可以推断出当输入输出通道比值为 1:1 时,卷积的内存访问成本最低。

如图 2 所示,本文采用只在模块最后输出聚合前面所有卷积层的输出结果,同时将中间卷积层的数量由 4 个减少至 3 个。这既避免了密集连接造成特征冗余,又能通过控制卷积输入输出通道的比值,使内存访问成本最低,加快推理速度。此外,为了提升网络特征提取能力引入 RepConv 作为密集连接的基本结构。

与原始 RepConv[26] 相比,本文改进的 RepConv 仍然分为训练和测试推理 2 种方式。在训练阶段,考虑到计算资源充足,采用多分支结构,包括 2 个 3×3 卷积和一个 1×1 卷积,这种多分支结构在训练时可以有效地提取到图像中的特征信息,从而提高模型的性能和精度。在测试推理阶段,仍然采用单路模型。通过这种方式,可以在训练阶段充分利用计算资源,同时在测试推理阶段保证模型的轻量化和高效性。此外,为了增强特征图通道间信息交互和增加网络非线性特性,在每个 RepConv 结构输出端引入一个 1×1 卷积。

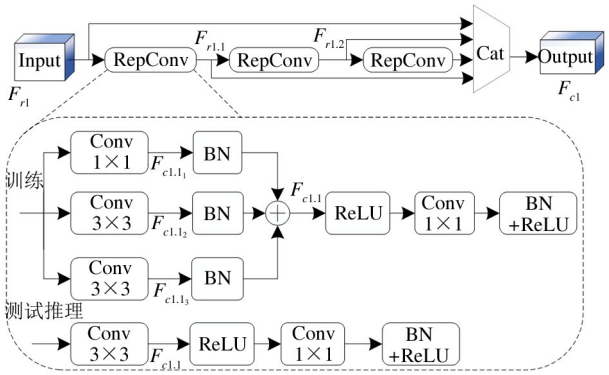


图 2　密集重参数化整体结构与 RepConv 构成细节

如图 2 所示,以下以特征图 $F_{r1} \in R^{C\times H\times W}$ 为输入来介绍改进后的 RepConv 的运算过程,其余 2 个 RepConv 结构的运算过程均参考如下流程。对于中间输入特征图 F_{r1},在训练阶段经过 3 个并行卷积分支得到输出特征图 $F_{c1.1} \in R^{C\times H\times W}$,这个过程也可采用等效的卷积层实现,本文重参数化实现可表示为式(1)。

$$\begin{aligned} F_{c1.1} &= BN_1(F_{r1}\odot K_1 + B_1) + BN_2(F_{r1}\odot K_2 + B_2) \\ &\quad + BN_3(F_{r1}\odot K_3 + B_3) \\ &= F_{r1}\odot(BN_1\times K_1 + BN_2\times K_2 + BN_3\times K_3) \\ &\quad + (BN_1\times B_1 + BN_2\times B_2 + BN_3\times B_3) \end{aligned} \tag{1}$$

其中,⊙表示卷积运算方式,K_1、K_2、K_3 分别表示不同卷积核的权重,B_1、B_2、B_3 分别表示不同卷积运算中的偏置,BN_1、BN_2、BN_3 为批归一化。

实现上述不同卷积层的融合分为 3 个步骤,首先将小尺寸的 1×1 卷积核进行 0 填充,扩张为 3×3 大小。确保所有参与卷积层融合的卷积核的数量和尺寸完全一致。其次,将卷积层与 BN 层融合为一个中间卷积层,其过程可表示为

$$BN_i = \gamma_i \frac{F_{c1.1_i} - \mu_i}{\sqrt{\delta_i^2 + \varepsilon}} + \theta_i \tag{2}$$

$$F_{c1.1_i} = F_{r1}\odot K_i \tag{3}$$

其中,μ_i 和 δ_i^2 分别为特征图的均值和方差,ε 是防止分母为 0 的极小值,θ_i 为可训练参数,综合式(2)、(3)可得:

$$BN_i = F_{r1}\odot \frac{\gamma_i}{\sqrt{\delta_i^2 + \varepsilon}} \times K_i + \left(\theta_i - \frac{\gamma_i \times \mu_i}{\sqrt{\delta_i^2 + \varepsilon}}\right) \tag{4}$$

其中,$F_{c1.1_i}$ 表示为训练阶段 RepConv 第 i 个分支的卷积输出结果。ε 是一个防止分母为 0 的极小值常数。式(4)表示一个卷积核权重为 $\frac{\gamma_i}{\sqrt{\delta_i^2 + \varepsilon}} \times K_i$、偏置为 $(\theta_i - \frac{\gamma_i \times \mu_i}{\sqrt{\delta_i^2 + \varepsilon}})$ 的卷积计算。

最后,如图 3 所示,将不同分支融合后得到的新卷积的权值进行对应位置相加得到推理阶段的等效卷积,可表示为式(5)。

$$\begin{aligned} F_{c1.1} &= F_{r1}\odot\left(\sum_{i=1}^{3} \frac{\gamma_i}{\sqrt{\delta_i^2 + \varepsilon}} \times K_i\right) \\ &\quad + \left(\sum_{i=1}^{3} \theta_i - \frac{\gamma_i \times \mu_i}{\sqrt{\delta_i^2 + \varepsilon}}\right) \end{aligned} \tag{5}$$

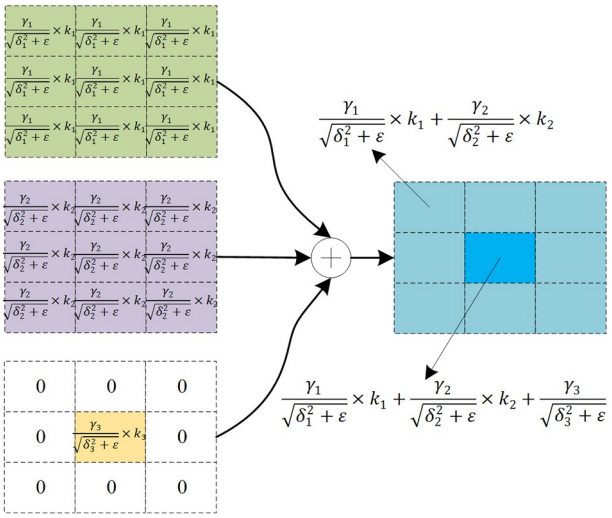


图 3　多分支卷积核融合

本文通过改进的 RepConv 不仅保留了原模块的优点,又将其与密集连接实现了更好的结合。

2.1.2　CASA 模块

注意力机制通过赋予空间中不同通道或区域不同的权重,使得网络可以在全局范围专注于重要信息的提取。本文提出的 CASA 模块如图 4 所示,由坐标注意力和空间注意力 2 部分并行连接组成,分别捕获输入特征图的通道和空间信息。以下以特征图 F_{r2} 为输入来介绍 CASA 模块的过程,其余 3 个 CASA 模块均参考如下流程。

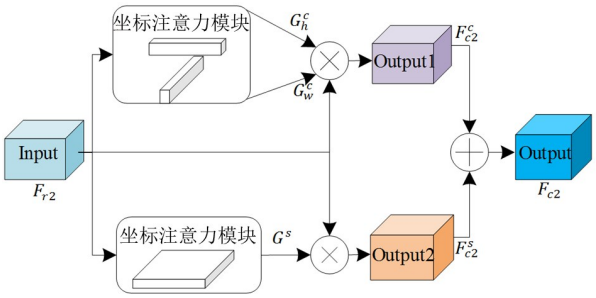


图 4　CASA 模块

将输入特征图 $F_{r2} \in R^{C\times W\times H}$ 同时经过坐标注意力(coordinate attention,CA)模块和空间注意力模块。经过坐标注意力模块得到 2 个二维的坐标注意力权重 $G_w^C \in R^{C\times 1\times W}$ 和 $G_h^C \in R^{C\times H\times 1}$;经过空间注意力模块得到 1 个二维的空间注意力权重 $G^S \in R^{1\times W\times H}$。将得到的注意力权重分别作用于输入特征图在相应维度上进行特征筛选,即可输出经过特征筛选的特征图 F_{c2}^C 和 F_{c2}^S,最后将输出特征图中的元素通过对应位置求和,得到 CASA 模块的输出特征图 F_{c2}。整个注意力过程可表示为式(6)、(7)、(8)。

$$F_{c2}^C = F_{r2} \otimes G_w^C(F_{r2}) \otimes G_h^C(F_{r2}) \tag{6}$$

$$F_{c2}^S = G^S(F_{r2}) \otimes F_{r2} \tag{7}$$

$$F_{c2} = F_{c2}^C \oplus F_{c2}^S \tag{8}$$

其中,⊗表示元素乘法,在乘法过程中使用广播[29],坐标注意力模块的值采用沿着空间方向广播,空间注意力模块的值采用沿着通道方向广播;⊕表示逐元素相加;F_{c2}^C 表示坐标注意力输出结果;F_{c2}^S 表示空间注意力输出;F_{c2} 表示最终整个模块的输出结果。

坐标注意力模块是将特征图的位置信息嵌入到通道注意力中,使网络可以在更大的区域上进行注意[30]。坐标注意力模块可以看作是一个用来增强特征表示能力的计算单元,它将特征图 $F_{r2} \in R^{C\times H\times W}$ 作为输入并输出一个有增强表示能力的同样尺寸的输出 $F_{c2}^C \in R^{C\times W\times H}$。如图 5 所示,坐标注意力模块主要由 2 个步骤完成:坐标信息嵌入和坐标注意力生成[31]。坐标信息嵌入是沿特征图的竖直和水平方向分别执行一维全局平均池化操作,生成 2 个单独的方向特征图,进而捕获特征的精准位置信息的空间长程依赖关系,可表示为式(9)和(10)。

$$Z_h^C = \frac{1}{W}\sum_{0\leqslant i\leqslant W} F_{r2}(H,i) \tag{9}$$

$$Z_w^C = \frac{1}{H}\sum_{0\leqslant j\leqslant H} F_{r2}(j,W) \tag{10}$$

其中,F_{r2} 为输入特征图,维度为 $C\times H\times W$。Z_h^C 和 Z_w^C 分别为在竖直方向和水平方向感知注意力特征。

坐标注意力生成过程首先级联 Z_h^C 和 Z_w^C 这 2 个特征,然后使用 1 个共享的 1×1 卷积 $Conv_{1\times 1}$ 促进通道间的信息交流,表示为式(11)。

$$I^C = \sigma(Conv([Z_h^C, Z_w^C])) \tag{11}$$

其中,$I^C \in R^{C/r\times(H\times W)}$ 是包含水平和竖直方向位置信息的中间特征图,r 表示下采样比例,实现控制模块的大小,设置为 16;[·,·]表示沿空间维度上的连接操作;σ 表示非线性激活函数。将中间特征图 I^C 沿空间维度拆分成 2 个单独的具有强位置指向性的一维特征图 $I_h^C \in R^{C/r\times H}$ 和 $I_w^C \in R^{C/r\times W}$,再利用 2 个 1×1 卷积 $Conv_h$ 和 $Conv_w$ 将特征图 I_h^C 和 I_w^C 变换到和输入 F_{r2} 同样的通道数,过程表示为式(12)、(13)。

$$G_h^C = \sigma(Conv_h(I_h^C)) \tag{12}$$

$$G_w^C = \sigma(Conv_w(I_w^C)) \tag{13}$$

其中,G_h^C 和 G_w^C 即为坐标注意力权重,坐标注意力模块的最终输出可以表示为式(14)。

$$F_{c2}^C(i,j) = F_{r2}(i,j) \times G_h^C(i) \times G_w^C(j) \tag{14}$$

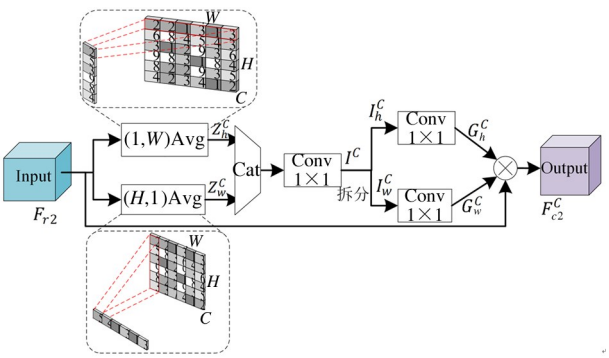


图 5　坐标注意力模块结构

空间注意力主要聚焦于输入特征图中含有丰富有效信息的区域,可以在一定程度上弥补坐标注意力模块的不足。如图 6 所示,空间注意力模块首先对输入特征图 $F_{r2} \in R^{C\times H\times W}$ 沿通道方向同时进行二维全局平均池化操作和二维全局最大池化操作,分别得到 2 个仅包含空间信息的二维特征图 $Z^S_{\mathrm{Avg}} \in R^{1\times H\times W}$ 和 $Z^S_{\mathrm{Max}} \in R^{1\times H\times W}$。之后,将得到的 2 个特征图在通道方向上进行拼接操作,并通过 1 个 1×1 大小的卷积层,融合 Z^S_{Avg} 和 Z^S_{Max} 上的空间信息,生成有效的空间注意力权重 $G^S \in R^{W\times H}$。生成空间注意力权重可表示为式(15)、(16)。

$$G^S = \sigma(Conv_{7\times7}([Avg(F_{r2}), Max(F_{r2})])) \tag{15}$$

$$G^S = \sigma(Conv_{7\times7}([Z^S_{\mathrm{Avg}}, Z^S_{\mathrm{Max}}]) \tag{16}$$

其中,$Conv_{1\times1}$ 表示卷积核为 1×1 的卷积,Avg 和 Max 分别表示在通道上的全局平均池化操作和全局最大池化操作。

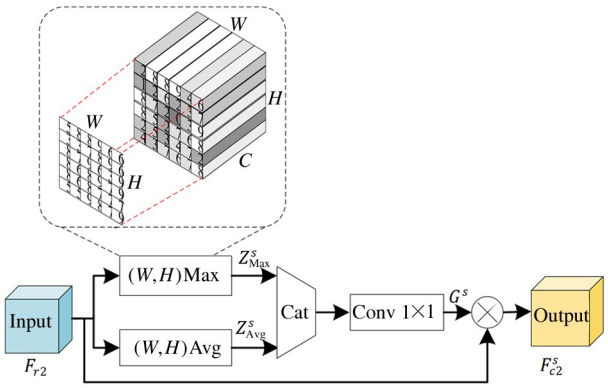


图 6　空间注意力模块结构

对于坐标注意力和空间注意力 2 个模块的连接,由于 2 个模块对于输入的图像中的特征关注存在互补,本文通过对不同组合方式和排列顺序调整的实验表明,对 2 种注意力模块采取任何一种组合排列方式都要比单独的注意力模块对算法整体性能提升效果更好。在引入模块参数量和计算量相同情况下,采用并行排列组合结构比采用顺序排列组合结构对算法性能提升效果更好。详见 3.2 节。

2.2　特征融合

为了适应不同尺寸目标检测任务,本文特征融合采用 FPN 和 PAN 的双向特征金字塔结构框架,并引入深度可分离卷积以提升算法的检测实时性和降低模型复杂度。此外,基于 C3 结构,本文引入 Ghost module 替换该模块中的普通卷积,实现更高效的特征融合。通过内容感知特征重组模块实现上采样,提升模型的准确率和泛化能力。

2.2.1　深度可分离卷积

深度可分离卷积可拆分为逐通道卷积(depth-wise convolution)和逐点卷积(pointwise convolution)2 个步骤完成[23]。其中逐通道卷积输入特征图的通道与卷积核的数量一一对应,这个过程产生的特征图通道数和输入特征图的通道数完全一样。逐点卷积与普通卷积相似,利用 1×1 卷积在通道方向上进行加权组合,融合不同的通道信息生成最终的特征图。与普通卷积相比,使用深度可分离卷积可以实现神经网络模型轻量化和提高网络模型推理速度。此外,深度可分离卷积还可以有效地缓解过拟合现象。

当输入输出特征图尺寸相同时,在不损失模型精度的情况下,使用 3×3 深度可分离卷积的参数量和计算量大约为普通 3×3 卷积的 1/9。

2.2.2　C3-G 模块

如图 7 所示,本文提出了 C3-G 结构,该结构采用多梯度流的设计,旨在获取丰富的梯度信息。结构整体由左右 2 个分支构成,其中左路分支由一个 1×1 卷积和 n 个 Ghost module 组成,右路分支仅包含一个 1×1 卷积。这种设计使得左路分支能够通过 Ghost module[32] 进行有效的特征提取和信息增强,而右路分支则进一步精炼特征的作用。通过这样的多梯度流结构,C3-G结构能够充分利用梯度信

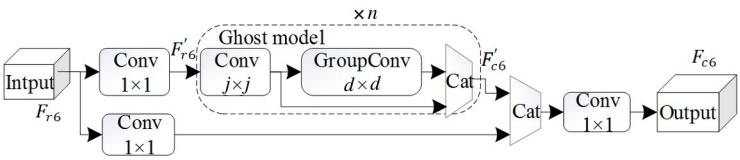


图 7　C3-G 结构图

息,从而提升模型的性能和表达能力。

此外,Ghost module 还具有轻量化模型的效果,在保持模型参数较少的情况下,能够有效提升模型的表达能力和感知能力。通过串联 n 个 Ghost module,能够将特征图进行更深层次的变换和扩展。Ghost module 通过 1 × 1 卷积的方式将输入特征图分为 2 个子特征图,然后利用共享权重的方式将其中一个子特征图进行复制和扩增。

以下以 F_{r6} 为输入来介绍 C3-G,其余 3 个 C3-G 均参考如下流程。如图 8 所示,给定输入特征图 $F_{r6} \in R^{C\times H\times W}$,将分别经过 2 个分支的处理生成的特征图在通道维度上进行拼接,最后通过 1 × 1 卷积对特征图的通道进行调整。其中,Ghost module 的主要方法是先通过一个常规卷积生成一定数量的常规特征图,再将生成的特征图通过一个与输入通道数相同分组的分组卷积生成 Ghost 特征图,再将上一步得到的常规特征图与 Ghost 特征图在通道维度上拼接,得到最终需要的特征图。

对于输入的中间特征图 $F_{r6} \in R^{C\times H\times W}$,经过 C3-G 模块生成特征图 $F_{c6} \in R^{C\times H\times W}$ 的过程可表示为式(17)。

$$\begin{cases} F'_{r6} = Conv_{1\times1}(F_{r6}) \\ F_{c6} = Conv_{1\times1}[Conv_{1\times1}(F_{r6}),\ Gm(F'_{r6})] \end{cases} \tag{17}$$

其中,$Conv_{1\times1}$ 表示 1 × 1 卷积运算。Gm 为 Ghost module 模块,[· , ·]表示沿通道方向的拼接。

输入为 $F'_{r6} \in R^{C\times H\times W}$、经过卷积核大小为 $j\times j$ 的卷积计算输出为 $F'_{c6} \in R^{N\times H\times W}$ 的参数量 $P1$ 和计算量 $F1$ 可表示为式(18)、(19)。

$$P1 = N \times j \times j \times C \tag{18}$$

$$F1 = N \times j \times j \times C \times H \times W \tag{19}$$

其中,C 为特征图 F'_{r6} 的通道数,H、W 和 N 为特征图 F'_{c6} 的高、宽和通道数。假设对于相同输入 $F'_{r6} \in R^{C\times H\times W}$ 采用 Ghost module 生成与普通卷积相同的结果 $F'_{c6} \in R^{N\times H\times W}$,且有 L 个中间特征图是使用卷积核尺寸为 $j\times j$ 的普通卷积生成,L 与 N 的关系有:$s = N/L$,$s > 1$。Ghost module 的运算过程可表示为式(20)。

$$F'_{c6} = [Conv_{j\times j}(F'_{r6}), GConv_{d\times d}(Conv_{j\times j}(F'_{r6}))] \tag{20}$$

其中,$GConv_{d\times d}$ 为卷积核大小为 $d\times d$ 的分组卷积且 $d\times d$ 与 $j\times j$ 大小相似。[· , ·]表示沿通道方向的拼接。Ghost module 的参数量 $P2$ 和计算量 $F2$ 可表示为式(21)、(22)。

$$P2 = L \times j \times j \times C + L \times d \times d \times (s-1) \tag{21}$$

$$F2 = L \times j \times j \times C \times H \times W + L \times d \times d \times (s-1) \times H \times W \tag{22}$$

在输入与输出相同的情况下,普通卷积参数量与 Ghost module 参数量的比值 r_c 可表示为式(23)。

$$r_c = \frac{P1}{P2} \approx \frac{C\times s}{s+C-1} \approx s \tag{23}$$

在输入与输出相同的情况下,普通卷积计算量与 Ghost module 计算量的比值 r_s 可表示为式(24)。

$$r_s = \frac{F1}{F2} \approx \frac{C\times s}{s+C-1} \approx s \tag{24}$$

通过以上分析,本文提出的 C3-G 模块不仅能提升模型的表达能力和感知能力,还可以在一定程度上促进网络模型轻量化,提升网络检测推理速度。

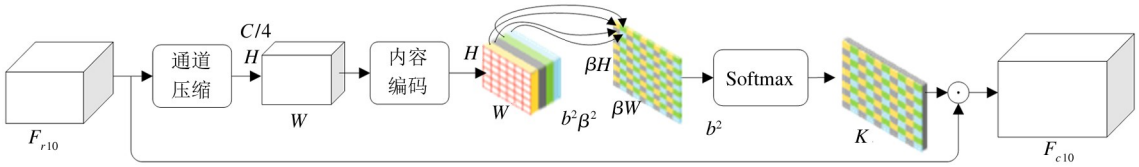


图 8　CARAFE 结构图

2.2.3　上采样

本文采用内容感知特征重组模块[5]替换特征融合结构中的最邻近插值法上采样模块,与上述的上采样方式相比 CARAFE 具有更大的感受野,能更好利用周围信息,并且上采样核与特征图语义信息相关。这些特性使 CARAFE 更好地获取特征图的

丰富信息,从而提高模型的准确率和泛化能力。如图 8 所示,CARAFE 模块输入特征图 F_{rl0} 的大小为 $C \times H \times W$, 上采样倍率为 β。CARAFE 模块主要包括 2 个步骤。

第 1 步是对特征图中每个目标位置的内容进行预测,得到一个重组核。首先,先经过一个 1×1 卷积将通道压缩为原来的 1/4,不仅可以减少后续步骤的参数和计算成本,还能提高 CARAFE 的效率。其次,使用内核大小为 $b \times b$ 的卷积层根据输入特征的内容生成重组核,可以得到更大的感受野,在更大区域上利用上下文信息。最后,将得到的特征图经过一个通道维在空间维展开的操作后再进行归一化处理,得到权重和为 1 的上采样核 $K \in R^{b^2 \times \beta H \times \beta W}$。

第 2 步是使用预测的重组核对特征进行重组。将输出特征图中的每个位置映射回输入特征图,取出以该位置为中心的 $b \times b$ 区域,并将其与上一步中预测得到的上采样核 K 进行点积运算,得到对应的输出值。在这个过程中,每个位置对应的上采样核都是不相同的,这种设计可以更好地提取特征图的丰富信息。与此同时,相同位置的不同通道共享同一个上采样核,这种共享机制可以有效地减少计算量和参数量,提高 CARAFE 模块的效率。通过这种方式,CARAFE 模块能够更好地实现特征上采样操作,提高模型的准确率和泛化能力。

2.3 检测输出

为了适应不同大小的目标检测任务,本文检测输出采用在多个尺度的特征图上预测的方式,通过分别在特征融合的不同输出层抽取特征图进行预测,再经过正负样本分配,筛选出与真实值接近的预测结果作为正样本。

如图 9 所示,本文采用基于 anchor base 的方法,在输出的每个特征图中每个特征点预设 3 种不同尺度的 anchor box。算法的预测结果回归分为中心点偏移和 anchor box 缩放 2 部分。中心点偏移如式(25)所示,anchor box 缩放如式(26)所示。

$$\begin{cases} b_x = 3\sigma(t_x) - 1 + c_x \\ b_y = 3\sigma(t_y) - 1 + c_y \end{cases} \tag{25}$$

$$\begin{cases} b_h = p_h(2\sigma(t_h))^2 \\ b_w = p_w(2\sigma(t_w))^2 \end{cases} \tag{26}$$

其中, b_x、b_y 为预测框中心点在特征图上的坐标, b_h、b_w 分别为预测框的高和宽; σ 为 sigmoid 函数, t_x、t_y、t_w、t_h 分别为网络模型在特征图上预设 anchor box 的中心点坐标偏移量和边框缩放倍数, c_x、c_y 为预设 anchor box 中心点在特征图上的坐标, p_h、p_w 分别为预设 anchor box 的高和宽。

由式(25)和(26)提出筛选正样本的 2 个条件,一是预设 anchor box 的宽和高与真实框的宽和高比值在 1/4 和 4 之间,二是包括真实框中心点所在的网格和以该网格为中心的周围 8 个网格所对应的预设 anchor box。同时满足这 2 个筛选条件下的预设 anchor box 所回归生成的预测框为正样本,其余为负样本。

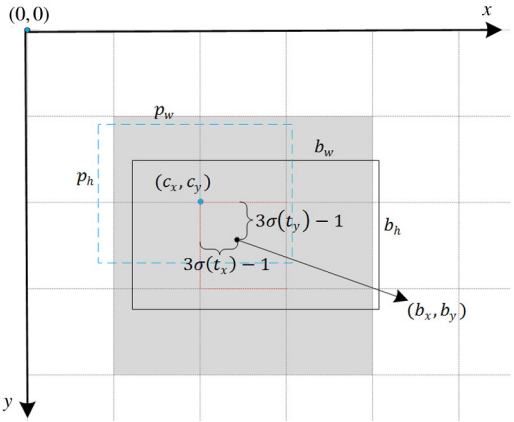


图 9 预测框回归

网络模型会对不同尺度输出特征图上的所有像素点预设的 anchor box 进行预测。对于每个 anchor box,模型会输出矩形框的位置、置信度和分类概率 3 个指标。在训练过程中,通过定义损失函数来度量预测结果与真实标签之间的距离。本文中,损失函数包含了 3 个方面的损失,分别是位置损失(localization loss)、置信度损失(confidence loss)和分类损失(classification loss)。本文采用 CIoU 损失[33]作为位置损失的度量标准,CIoU 损失定义如下:

$$CIoU = IoU - \frac{\rho^2}{c^2} - \alpha v = DIoU - \alpha v \tag{27}$$

$$v = \frac{4}{\pi^2}\left(\tan^{-1}\frac{w_{gt}}{h_{gt}} - \tan^{-1}\frac{w_p}{h_p}\right)^2 \tag{28}$$

$$\alpha = \frac{v}{1 - IoU + v} \tag{29}$$

$$L_{\text{local}} = \sum_{t=1}^{T}\sum_{cell=1}^{3}\sum_{n=1}^{anchors} 1^{\text{obj}}(1 - CIoU) \tag{30}$$

其中，w_{gt} 和 h_{gt} 表示标签中真实框的宽和高，w_p 和 h_p 表示预测框的宽和高，T 表示目标个数；*cell* 为检测头个数；1^{obj} 表示为正样本；*anchors* 表示参与计算的预测框个数；ρ 为真实框中心点与预测框中心点之间的距离；c 为真实框与预测框的最小包围矩形的对角线长度；v 用于度量真实框与预测框长宽比的一致性，取值范围为 0～1，当真实框与预测框的宽高比相等时 v 为 0，当真实框与预测框的宽高比相差无限大时 v 取 1；α 为 v 的影响因子。

分类损失采用二元交叉熵损失，分类损失只考虑正样本损失，可以表示为

$$L_{\text{class}} = \sum_{t=1}^{T}\sum_{cell=1}^{3}\sum_{n=1}^{anchors} 1^{obj} \sum_{cls=1}^{classes}(-\hat{p}\ln(p) - (1-\hat{p})\ln(1-p)) \tag{31}$$

其中，$\hat{p}$ 为标签真实值，p 为预测的类别概率，*classes* 为总的目标类别数。

置信度损失采用二元交叉熵损失，由正样本损失和负样本损失 2 部分组成。其中，正样本的目标标签根据 CIoU 的值采用标签平滑，生成软标签，表示为

$$L_{\text{obj+}} = \sum_{t=1}^{T}\sum_{cell=1}^{3}\sum_{n=1}^{anchors} 1^{obj}(-\hat{q}\ln(q) - (1-\hat{q})\ln(1-q)) \tag{32}$$

$$\hat{q} = (1-gr) + gr \times CIoU \tag{33}$$

$$L_{\text{obj-}} = \sum_{t=1}^{T}\sum_{cell=1}^{3}\sum_{n=1}^{anchors} 1^{\text{noobj}}(-\ln(1-q)) \tag{34}$$

$$L_{\text{obj}} = L_{\text{obj+}} + L_{\text{obj-}} \tag{35}$$

其中，$\hat{q}$ 为平滑标签，根据正负样本比例设置具体值；q 分别为预测置信度；gr 为设置的定值；1^{noobj} 为负样本。总损失表示为

$$Loss = \lambda_1 L_{\text{local}} + \lambda_2 L_{\text{class}} + \lambda_3 L_{\text{obj}} \tag{36}$$

其中，λ_1、λ_2、λ_3 分别为位置损失、分类损失和置信度损失的权重系数。

3 实验验证

3.1 实验设置

为了验证本文方法的有效性，本文在 2 个常见的公开标准数据集 MS COCO 与 PASCAL VOC 数据集上进行实验。

本文实验结果验证所用硬件条件：CPU 为 6 核的 Intel Corei7-8700K，主频为 3.7 GHz，内存为 32 GB。GPU 为 GeForce RTX 1080Ti，GPU 显存为 12 GB。程序运行的操作系统版本为 Ubuntu16.04，算法模型使用 Pytorch1.8 框架实现。

图像预处理阶段采用图片自适应缩放将输入图片调整为 640×640 固定大小，训练一共 300 个迭代（epoch），以保证算法能充分学习到数据集的特征。其中前 290 个迭代使用 Mosaic 数据增强辅助训练，增加数据训练量，提升算法泛化能力和鲁棒性。

在训练阶段优化器使用 Adam 优化器进行参数更新与优化，初始学习率设置为 0.01，采用余弦退火进行学习率衰减，训练批次大小（batch size）根据实验设备设置为 16、32 或者 64。损失函数的权重系数采用文献[34]中网格搜索的超参数搜索方法得到，且 $\lambda_1 = 0.05$、$\lambda_2 = 0.50$ 和 $\lambda_3 = 1.00$。

3.2 消融实验结果分析

为了评估本文设计方法的合理性，本文使用具有挑战性的 MS COCO 数据集进行消融实验。所有实验的参数设置都相同。

本文设计 2 组消融实验，第 1 组为了探究 CASA 模块结构的合理性分别对单独注意力模块、串行连接结构和并行连接结构进行实验验证，如表 1 所示。第 2 组为了探究所提出特征融合方法的有效性，在 FPN 结构和 PAN 结构的基础上依次采用本文所提方法，结果如表 2 所示。

表 1 不同注意力模块对算法性能的影响

方法	参数量 /×10⁶	计算量 /×10⁹	mAP @0.5/%	mAP@0.5 -0.95/%	FPS /(帧/秒)
CSP-DarkNet	7.3	17.0	55.4	36.7	**206.0**
+CA	7.3	17.0	56.1	36.8	112.0
+Spatial	7.3	17.0	55.8	36.7	126.9
+CA&Spa	7.3	17.1	56.5	37.1	109.8
+Spa&CA	7.3	17.1	56.7	37.3	109.8
+CASA	7.3	17.1	**57.1**	**37.8**	111.0

表 1 中的实验结果表明，引入注意力模块前后

网络模型的参数量和计算量基本都维持在 7.3×10^6 和 17.0×10^9 附近,说明注意力模块对网络模型参数的影响比较小。坐标注意力比空间注意力对算法性能提升更明显,mAP@0.50 和 mAP@0.50-0.95分别提高0.3%和0.1%。坐标注意力与空间注意力任意组合的模块对算法精度提升的影响大于采用其中单一注意力模块。这也表明,通过将不同类型的注意力模块进行组合,可以更好地利用它们各自的优势,从而提高网络的性能。

表1中的实验结果还展示了坐标注意力与空间注意力采用不同连接方式对网络性能的影响。并行连接坐标注意力和空间注意力的结构对算法精度提升最大,mAP@0.50 和 mAP@0.50-0.95 分别达到57.1%和37.8%。其次,空间注意力模块后接坐标注意力模块的顺序连接结构比注意力模块后接空间注意力模块的顺序连接结构对算法精度的提升大,mAP@0.50 和 mAP@0.50-0.95 分别提高0.2%。这些发现对于在实际应用中选择合适的注意力模块和连接方式提供了有益的指导和参考。

表2 不同方法对算法性能的影响

方法	参数量 /$\times10^6$	计算量 /$\times10^9$	mAP@0.50/%	mAP@0.50-0.95/%	FPS/(帧/秒)
FPN+PAN	7.3	17.0	55.4	36.7	**206**
+CARAFE	7.3	17.0	56.2	37.1	116
+CARAFE+DW	6.5	15.5	56.4	37.0	135
+CARAFE+DW+C3-G	6.7	15.7	**56.6**	**37.4**	140

根据表2的实验结果可以发现,CARAFE 上采样相较于原结构中的最邻近插值法上采样虽然检测速度有所下降,但算法的精度显著提升,mAP@0.50从55.4%提升到56.2%,mAP@0.50-0.95 从36.7%提升到37.1%,这意味着在目标检测任务中应用CARAFE 上采样具有很大优势。

采用深度可分离卷积,网络模型参数量和计算量分别降低了 0.8×10^6 和 1.5×10^9,在检测精度不下降的同时,检测速度也从每秒116帧提升到了每秒135帧。说明深度可分离卷积在不影响算法精度的前提下,有助于提升检测的实时性。

更进一步看,C3-G 模块不仅提升了算法精度,使 mAP@0.50 达到56.6%,mAP@0.50-0.95 达到37.4%,同时,网络推理速度也从每秒135帧提升到了每秒140帧。

3.3 与其他算法对比

为了充分证明本文方法的有效性,本文将所提方法在 MS COCO 与 VOC 数据集上的实验结果与目前常见的基于深度学习的目标检测算法进行对比。为了保证公平比较,选用算法的 mAP 评价指标数据和推理速度 FPS 指标均采用 GeForce RTX 1080Ti GPU 设备进行推理测试得到。

表3列举了部分常见目标检测算法和本文所改进算法在 MS COCO 数据集上的实验结果。从表中可以看出,本文算法 mAP@0.50 达到57.5%,mAP@0.50-0.95 达到38.5%,仅低于 DETR 算法[35],优于大部分常见算法。同时,得益于本文对特征融合部分的轻量化设计,模型推理速度达到了每秒106帧,能保持较好的实时性。

表3 与常见算法在 COCO 数据集上对比

模型	主要结构	mAP@0.50/%	mAP@0.50-0.95/%	FPS/(帧/秒)
SSD512	VGG	46.5	26.8	22
SSD300	VGG	41.2	23.2	59
Faster R-CNN	ResNet-101	55.7	34.9	4
RetinaNet	ResNet-101	57.5	37.8	24
EfficientDet0	Efficient	52.2	33.8	35
DETR	Transformer	**62.4**	**42.0**	12
YOLOv7-tiny	--	55.2	37.4	127
YOLOv6n	--	51.2	35.9	**309**
YOLOv5s	CSPDarkNet	55.4	36.7	206
YOLOv4-tiny	DarkNet53	42.1	24.9	82
本文算法	—	57.5	38.3	106

为了更加全面、准确地评估本文方法,本文也采用 VOC 数据集进行评估,并对比其他常见算法在该数据集上的结果。如表4所示,实验结果表明,本文算法在 VOC 数据集上的精度保持在较高水平,mAP@0.50达到83%,mAP@0.50-0.95 达到59.5%。此外,本文算法的实时性是上述常见算法中最好的。

为了对本文算法在不同目标类别上的检测效果

表 4　与常见算法在 VOC 数据集上对比

模型	主要结构	mAP@0.50/%	mAP@0.50-0.95/%	FPS/（帧/秒）
SSD500	VGG	76.8	--	22
Faster R-CNN	ResNet-101	73.2	49.5	2
RetinaNet	ResNet-101	80.5	58.0	24
EfficientDet0	EfficientNet	**84.0**	**66.7**	35
CenterNet	ResNet-101	73.1	47.0	40
本文算法	—	83.0	59.5	**106**

有更深入的了解，本文评估了 VOC 测试数据集中 20 个目标类别的具体检测精度如图 10 所示。可以发现，pottdplant、chair、bottle 和 boat 等几个类别的精度明显较低，其主要由以下 2 个原因导致。首先，训练集中的目标数量较少，导致算法在这些类别上的识别能力受到限制。其次，一些类别的目标与背景相似度过高，导致算法难以准确区分。例如，在室内场景下，椅子和桌子等类别的目标与背景颜色、纹理等特征相似，容易造成误判。

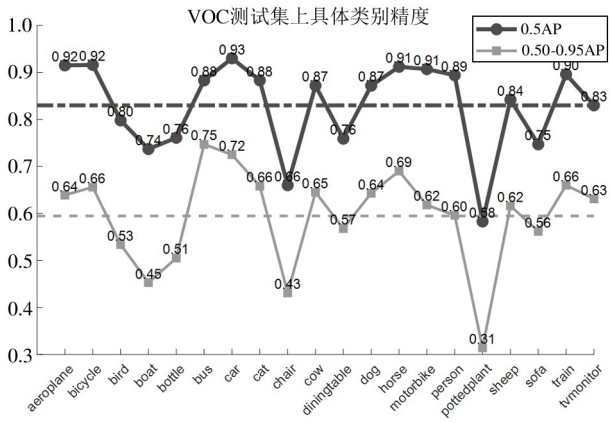
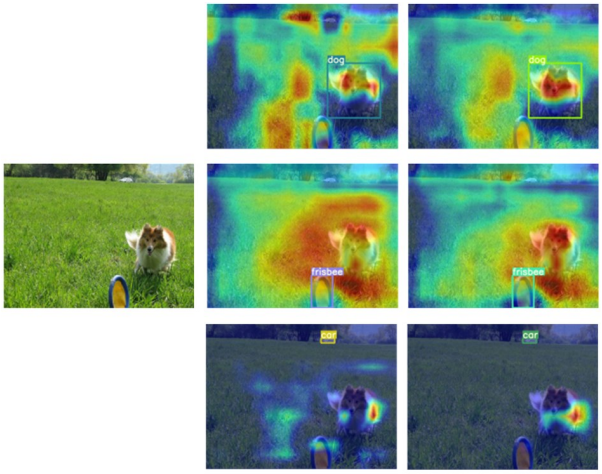


图 10　VOC 数据集检测结果

3.4　可视化分析

对于相同输入图像，在图 1 中的 F_{c5} 处对输出特征图进行可视化。如图 11 所示，未添加注意力的算法虽然也能预测出目标在图像中的位置，但从图 11(b)中能明显发现，其高亮区域更广，且包含的背景区域大于目标区域，说明特征图中包含的冗余信息太多，这不仅增加了后续处理成本，还会影响模型的性能和精度。同时，由于缺乏有效的注意力机制，高亮区域分布比较散乱，对目标的边界和定位存在偏差，这导致模型在目标检测和定位任务中的表现不尽如人意。

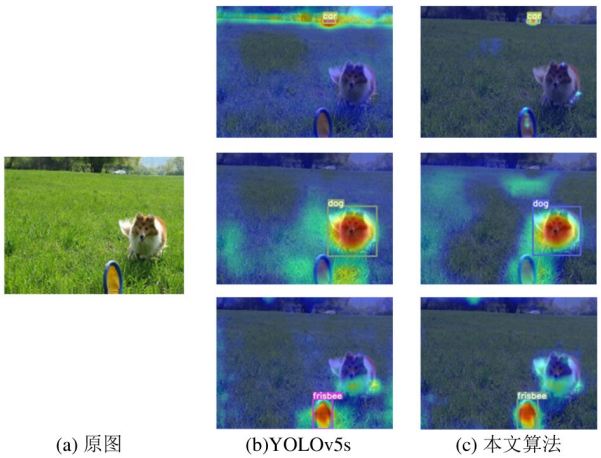
与图 11(b)相比，添加 CASA 模块后的热力图高亮区域更加集中于目标所在位置，且包含的背景区域小于目标区域。这说明 CASA 模块可以有效地过滤掉冗余背景信息，提取与目标相关的特征。同时，图 11(c)中的特征图的高亮区域基本能反映出目标位置和大小，表明 CASA 模块有助于更加准确定位目标的边界。综上所述，CASA 模块是一种有效的注意力机制，可以为深度学习模型提供更加准确、有效的特征表示。



(a) 原图　(b) 添加注意力前　(c) 添加注意力后

图 11　注意力可视化对比

本文还比较了表 4 中所列的 YOLOv5s 算法和本文算法的特征图可视化结果。如图 12 所示，采用相同的输入，从结果可视化对比中可以发现，图 12(b)中的特征图对小目标检测包含的背景信息明显要多于图12(c)中的特征图。这表明本文算法对小目标



(a) 原图　(b)YOLOv5s　(c) 本文算法

图 12　检测结果可视化对比

的检测效果更好。同时,通过图 12(c)可以发现,热力图的高亮区域能完全集中在目标位置,特征图可以提取到目标的特征信息。这表明本文所提算法的准确性和有效性更高。

4 结 论

针对目标检测任务中的挑战,如目标与背景间的复杂区分和目标尺寸的显著差异,本文提出基于注意力和密集重参数化的一阶段目标检测算法。该算法综合考虑了特征提取、特征融合与检测输出这 3 大核心环节。在特征提取部分,采用 CSP-DarkNet 网络为基础,进一步融合了密集重参数化结构和 CASA 模块。这样的设计策略旨在强化网络对于复杂背景下目标特征的提取能力,确保在各种背景条件下目标都能被准确识别。特征融合部分,以 FPN 和 PAN 结构为基础,并融合了 C3-G 模块、深度可分离卷积以及 CARAFE 等技术,增强模型对同一图像中不同尺寸目标的识别能力,尤其是容易被漏检的小目标。检测输出环节,采纳了多尺度检测的策略,并创新性地提出了一种正负样本匹配策略,从而提高模型的训练效果和泛化能力。

为了验证所提出算法的有效性和其结构设计的合理性,本文进行了详尽的消融实验,并提供了清晰的可视化结果。最终,在公开 MS COCO 数据集上检测,本文算法 mAP@0.50 达到了 57.5%,mAP@0.50-0.95 达到 38.3%;而在 VOC 数据集上,mAP@0.50 达到了 83%,mAP@0.50-0.95 达到了 59.5%的检测精度,这些成果均明显优于当前大部分常见目标检测算法,充分展现了本文所提算法在目标检测任务上的强大潜力和应用价值。

参考文献

[1] CHEN Y, JIANG H, LI C. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016,54(10):6232-6251.

[2] 窦慧, 张凌茗, 韩峰, 等. 卷积神经网络的可解释性研究综述[J]. 软件学报, 2024,35(1):159-184.

[3] SINGH B, NAJIBI M, DAVIS L S. Sniper: efficient multi-scale training[J]. Advances in Neural Information Processing Systems, 2018,31:9333-9343.

[4] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[EB/OL]. (2015-12-10)[2023-07-26]. https://arxiv.org/pdf/1512.03385.pdf.

[5] WANG J, CHEN K, XU R, et al. Carafe: content-aware reassembly of features[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019:3007-3016.

[6] 李鹏芳, 刘芳, 李玲玲. 嵌入标签语义的元特征再学习和重加权小样本目标检测[J]. 计算机学报, 2022, 45(12):2561-2575

[7] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014:580-587.

[8] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015,39(6): 1137-1149.

[9] WANG H, TIAN Y, DU Y, et al. A survey of aerial image target detection based on single-stage series algorithms[C]// The 3rd International Symposium on Computer Engineering and Intelligent Communications. Xi'an, China: ISCEIC, 2022:394-398.

[10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 779-788.

[11] WOMG A, SHAFIEE M J, LI F, et al. Tiny SSD: a tiny single-shot detection deep convolutional neural network for real-time embedded object detection[C]// 2018 15th Conference on Computer and Robot Vision (CRV). Windsor, Canada: IEEE, 2018:95-101.

[12] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017:6517-6525.

[13] MAO Q C, SUN H M, LIU Y B, et al. Mini-YOLOv3: real-time object detector for embedded applications[J]. IEEE Access, 2019,7:133529-133538.

[14] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. Scaled-yolov4: scaling cross stage partial network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021: 13029-13038.

[15] TANG B H. ASFF-YOLOv5: Multielement detection method for road traffic in UAV images based on multiscale feature fusion[J]. Remote Sensing, 2022, 14(14), 3498.

[16] YUNG N D T, WONG W K, JUWONO F H, et al. Safety helmet detection using deep learning: implementation and comparative study using YOLOv5, YOLOv6, and YOLOv7[C]//2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST). Miri Sarawak, Malaysia: IEEE, 2022:164-170.

[17] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023:7464-7475.

[18] LI J, CHENG B, FERIS R, et al. Pseudo-IOU: improving label assignment in anchor-free object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021:2378-2387.

[19] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452:48-62.

[20] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:7132-7141.

[21] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: ECCV, 2018:3-19.

[22] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[EB/OL]. (2016-11-04) [2023-07-26]. https://arxiv.org/pdf/1512.03385.pdf.

[23] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:4510-4520.

[24] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7):38-39.

[25] LIU Z, LI J, SHEN Z, et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017:2736-2744.

[26] DING X, ZHANG X, MA N, et al. RepVGG: making VGG-style convnets great again[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021: 13733-13742.

[27] ZHU Y, NEWSAM S. Densenet for dense flow[C]//2017 IEEE International Conference on Image Processing (ICIP). Beijing, China: IEEE, 2017:790-794.

[28] MA N, ZHANG X, ZHENG H T, et al. Shufflenet v2: practical guidelines for efficient CNN architecture design[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: ECCV, 2018:116-131.

[29] 斋藤康毅,深度学习入门:基于 Python 的理论与实现[M]. 北京: 人民邮电出版社. 2018:13-72.

[30] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021: 13713-13722.

[31] YI S, LI J, LIU X, et al. CCAFFMNet: dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module[J]. Neurocomputing, 2022,482:236-251.

[32] HAN K, WANG Y, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020:1580-1589.

[33] DU S, ZHANG B, ZHANG P, et al. An improved bounding box regression loss function based on CIOU loss for multi-scale object detection[C]//2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML). Chengdu, China: IEEE, 2021:92-98.

[34] LI B, WU W, WANG Q, et al. SiamRPN ++ : evolution of Siamese visual tracking with very deep networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019:4277-4286.

[35] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]// European Conference on Computer Vision. Berlin, Germany: Springer International Publishing, 2020:213-229.

Object detection algorithm based on attention and dense reparameterization

CHEN Zhiwang[*,**], LEI Chunming[*], LV Changhao[***], WANG Ting[*], PENG Yong[****]

([*]Engineering Research Center of the Ministry of Education for Intelligent Control System and Intelligent Equipment, Yanshan University, Qinhuangdao 066004)

([**]Key Laboratory of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao 066004)

([***]Key Laboratory of Power Electronics for Energy Conservation and Drive Control of Hebei Province, Yanshan University, Qinhuangdao 066004)

([****]School of Electrical Engineering, Yanshan University, Qinhuangdao 066004)

Abstract

This paper presents an object detection algorithm using attention and dense reparameterization to tackle challenges posed by complex backgrounds and variations in object sizes, which can adversely affect detection results. The proposed algorithm consists of two key components within an efficient feature extraction network based on CSP-DarkNet: the dense reparameterization module and the coordinate and spatial attention (CASA) module. The former leverages dense connections to retain shallow features while reducing network complexity through reparameterization structures, while the CASA module captures necessary target information. Feature fusion is performed using feature pyramid network (FPN) and path aggregation network(PAN), and upsampling is achieved through content-aware reassembly of features (CARAFE), addressing the issue of insufficient capture of rich semantic information. To enhance model capabilities, a more efficient C3-G module is introduced to obtain gradient information, and depthwise separable convolution is employed to improve computational efficiency. Lastly, the detection output is enhanced by employing a cross-domain positive-negative sample matching strategy on a larger scale, augmenting positive samples and improving detection performance. Experimental results showcase the algorithm's advancements, achieving mAP@0.50 scores of 57.5% and 83.0% on the MS COCO and PASCAL VOC datasets, respectively.

Key words: object detection, reparameterization, attention mechanism, feature fusion, upsampling, positive and negative sample matching