

改进的 RetinaNet 目标检测算法

于 敏^{1,2}, 屈 丹², 司念文²

(1. 郑州大学 软件学院, 郑州 450000; 2. 战略支援部队信息工程大学 信息工程学院, 郑州 450000)

摘 要: 针对经典一阶段目标检测算法 RetinaNet 难以充分提取不同阶段特征、边界框回归不够准确等问题, 提出一个面向目标检测的改进型 RetinaNet 算法。在特征提取模块中加入多光谱通道注意力, 将输入特征中的频率分量合并到注意力处理中, 从而捕获特征原有的丰富信息。将多尺度特征融合模块添加到特征提取模块, 多尺度特征融合模块包括 1 个路径聚合模块和 1 个特征融合操作, 路径聚合模块通过搭建自底向上的路径, 利用较浅特征层上精确的定位信号增强整个特征金字塔的信息流, 特征融合操作通过融合来自每个阶段的特征信息优化多阶段特征的融合效果。此外, 在边界框回归过程中引入完全交并比损失函数, 从边界框的重叠面积、中心点距离和长宽比这 3 个重要的几何因素出发, 提升回归过程的收敛速度与准确性。在 MS COCO 数据集和 PASCAL VOC 数据集上的实验结果表明, 与 RetinaNet 算法相比, 改进型 RetinaNet 算法在 2 个数据集上的平均精度分别提高了 2.1、1.1 个百分点, 尤其对于 MS COCO 数据集中较大目标的检测, 检测精度的提升效果更加显著。

关键词: 深度学习; 目标检测; 多光谱通道注意力; 多尺度特征融合; 完全交并比

开放科学(资源服务)标志码(OSID):



中文引用格式: 于敏, 屈丹, 司念文. 改进的 RetinaNet 目标检测算法[J]. 计算机工程, 2022, 48(8): 249-257.

英文引用格式: YU M, QU D, SI N W. Improved RetinaNet algorithm for object detection[J]. Computer Engineering, 2022, 48(8): 249-257.

Improved RetinaNet Algorithm for Object Detection

YU Min^{1,2}, QU Dan², SI Nianwen²

(1. School of Software, Zhengzhou University, Zhengzhou 450000, China;

2. School of Information Systems Engineering, Strategic Support Force Information Engineering University, Zhengzhou 450000, China)

[Abstract] Based on the problems that the classical one-stage object detection algorithm RetinaNet is difficult to fully extract and fuse different stage features, while the bounding box regression is not sufficiently accurate, an improved RetinaNet algorithm for object detection is proposed. First, the algorithm adds multispectral channel attention to the feature extraction module, which incorporates more frequency components in the input features into the attention processing to capture the original rich information of the features. Thereafter, the multiscale feature fusion module is added after the feature extraction module, and the multiscale feature fusion module includes a path aggregation module and a feature fusion operation. The path-aggregation module enhances the information flow of the entire feature pyramid by building bottom-up paths and using accurate positioning signals on shallower feature layers. The feature fusion operation further enhances the fusion effect of multistage features by fusing the feature information from each stage. Finally, the Complete Intersection over Union (CIoU) loss function is introduced in the bounding box regression process. The loss function starts from three important geometric factors, namely, the overlapping area of the bounding box, the distance between the center points, and the aspect ratio to improve the convergence speed of the regression process and accuracy. The experimental results on the MS COCO and PASCAL VOC datasets show that, compared with the RetinaNet algorithm, the average accuracy of the improved RetinaNet algorithm on the two datasets is increased by 2.1 and 1.1 percentage points, especially for the MS COCO data set. For the detection of large targets, improving the detection accuracy is more significant.

[Key words] deep learning; object detection; multi-spectral channel attention; multi-scale feature fusion; Complete Intersection over Union (CIoU)

DOI: 10.19678/j.issn.1000-3428.0062134

0 概述

目前, 基于深度学习的目标检测算法主要分为一阶段目标检测算法和两阶段目标检测算法两大类。一

阶段目标检测算法是基于回归的目标检测方法, 能同时对图像进行分类和候选框参数的回归, 摒弃了多次回归的步骤; 两阶段目标检测算法是基于候选区域的

基金项目: 国家自然科学基金(62171470, 61673395)。

作者简介: 于 敏(1997—), 女, 硕士研究生, 主研方向为智能信息处理、目标检测; 屈 丹, 教授、博士; 司念文, 讲师、博士。

收稿日期: 2021-07-20 修回日期: 2021-08-26 E-mail: qudanqudan@sina.com

目标检测方法,该算法使用先选取候选区域,后对候选区域进行分类和回归的策略。相比于两阶段目标检测算法,一阶段目标检测算法无需候选区域分类回归步骤即可直接对目标进行分类预测。因此,一阶段目标检测算法不仅降低了计算复杂性,提高了时间效率,而且对实时目标检测具有更大的适用性,应用范围更加广泛。

在一阶段目标检测算法中,RetinaNet^[1]是一种基于Focal损失函数的经典网络,其继承了之前一阶段目标检测算法检测速度快的特点^[2],且基本克服了训练过程中类别不平衡问题。到目前为止,RetinaNet仍被作为基础网络进行改进^[3-4],或者作为主流基线网络与目前最新方法进行比较^[5],被广泛用于计算机视觉领域^[6-7]。然而,RetinaNet仅关注分类损失函数来解决类别不平衡问题,忽略了网络本身和边界框回归也是训练过程优化的重点,依旧存在难以充分提取与融合不同阶段特征,以及边界框回归不够准确的问题。传统的RetinaNet算法通过深度卷积网络后会输出不同阶段尺度大小不一致的特征图,深层特征对应的下采样率通常比较大,容易造成小目标在特征图上的有效信息较少,不利于小目标的检测。而浅层特征分辨率较高,往往学习到的是细节特征,不利于大目标的检测。同时,RetinaNet算法中的特征金字塔(Feature Pyramid Networks, FPN)^[8]虽然试图通过横向连接进行特征集成,但FPN中的顺序方式使集成特征更多地关注相邻层特征,而较少关注其他层特征。上述问题均可归结为不能充分提取与融合不同阶段特征的问题。此外,在边界框回归过程中,RetinaNet算法中的边界框回归损失函数无法判断预测框和目标框是如何相交的,如

果预测框和目标框没有重叠,那么损失函数将不起作用。上述问题归结为边界框回归不够准确的问题。

本文提出一种改进型RetinaNet算法,在特征提取模块的深度残差网络^[9]中引入多光谱通道注意力(Multi-spectral Channel Attention, MCA)模块^[10],该模块在ImageNet数据集^[11]上可达到最佳水平,能够提取不同阶段特征的丰富信息。此外,在特征提取模块后添加多尺度特征融合(Multi-scale Feature Fusion, MFF)模块,该模块包括1个具有自底向上路径的路径聚合模块^[12]和1个特征融合操作^[13],通过使用不同深度集成的特征来增强多层特征融合。将RetinaNet算法中的边界框回归损失函数替换为完全交并比(Complete Intersection over Union, CIOU)损失函数^[14],提高边界框在回归过程中的收敛速度。最后,在MS COCO数据集^[15]和PASCAL VOC数据集^[16]上验证改进型RetinaNet算法的检测性能。

1 改进型 RetinaNet 算法

改进型RetinaNet算法的整体架构如图1所示,首先输入一张图片,在加入MCA模块的ResNet-FPN特征提取模块中提取图像特征,通过MFF模块来融合多阶段特征,输出5层多尺度特征图,并在每层特征上设置锚框,其对应输入图像按固定长度进行平移。然后,生成的全部锚框覆盖了相对于输入图像的尺度范围,并设置交并比(Intersection over Union, IoU)阈值对锚框进行筛选。最后将其送入分类与边界框回归模块,分类分支和回归分支都是全卷积网络,分类分支预测了每个锚框上 K 个类别的概率,回归分支预测了锚框和目标框之间的相对偏移。

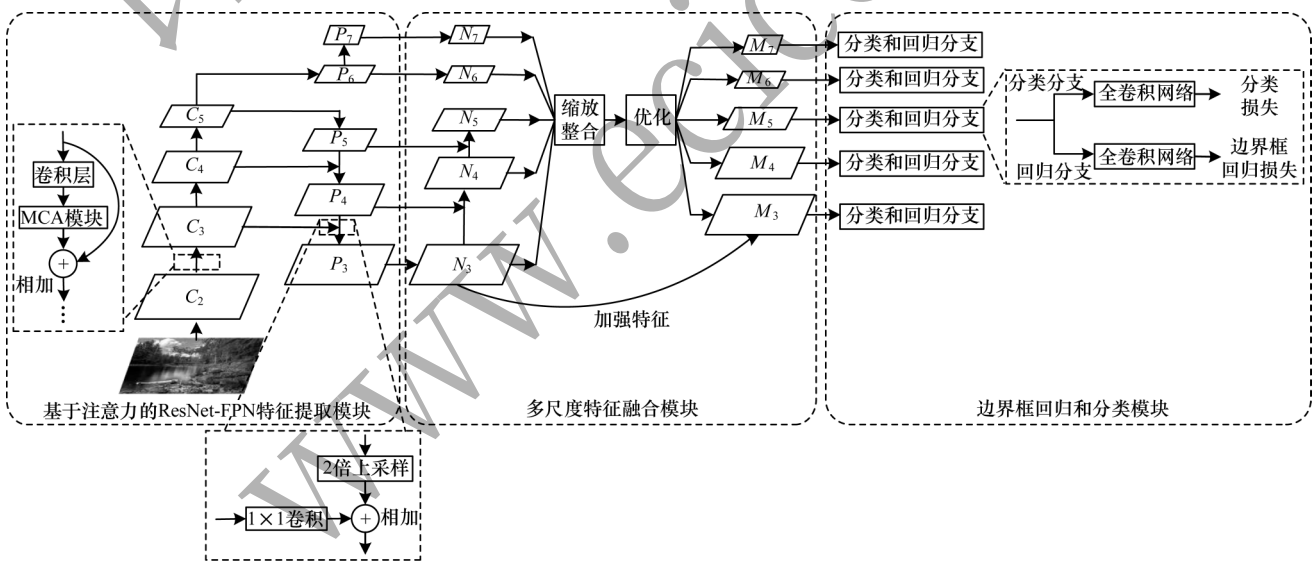


图1 改进型 RetinaNet 算法的整体架构

Fig.1 Overall architecture of improved RetinaNet algorithm

1.1 基于多光谱通道注意力的特征提取模块

如图1的左侧部分所示是基于多光谱通道注意力的ResNet-FPN特征提取模块。ResNet-FPN作为RetinaNet算法的特征提取模块,使用ResNet每个残差阶段中的最后一个残差块输出的特征图,即 C_2 、 C_3 、 C_4 和 C_5 这4层特征构成自底向上路径。而 P_2 、 P_3 、 P_4 和 P_5 这4层特征是经过横向连接和2倍上采样得到的。为

了减少计算量,不使用高分辨率特征图 P_2 。 P_7 、 P_6 是经过步幅为2的 3×3 卷积得到的,最后 P_7 、 P_6 、 P_5 、 P_4 和 P_3 这5层特征构成了自顶向下路径。由于ResNet-FPN特征提取模块存在不能充分提取图片特征信息的问题,因此,本文算法加入了多光谱通道注意力(Multi-spectral Channel Attention, MCA)模块^[10]来帮助有效地提取丰富的特征信息。

本文将MCA模块插入在深度残差网络的多个残差块中,以ResNet-50网络为例,该网络主要有5个阶段,第1个阶段包括7×7卷积和3×3最大池化层,后4个阶段分别包括3、4、6、3个残差块,每个残差块又包括1×1、3×3和1×1共3个卷积层,ResNet-50网络结构如表1所示,其中64、256、128、512、1 024、2 048等数字代表通道数。

表1 ResNet-50 网络结构

Table 1 Network structure of ResNet-50

子层名	输出尺寸/像素	具体配置
conv1	112×112	7×7,64, stride 2
		3×3 max pooling, stride 2
conv2	56×56	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
conv5	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$

以ResNet-50网络为例,在该网络的每个残差块中均插入一个注意力模块,具体操作如图2所示,总共插入了16个注意力模块。

多光谱通道注意力模块是基于离散余弦变换(Discrete Cosine Transform, DCT)^[17]提出的,二维DCT可定义为:

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} \underbrace{\cos\left(\frac{\pi h}{H}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(j + \frac{1}{2}\right)\right)}_{\text{DCT weights}}, \quad (1)$$

s.t. $h \in \{0, 1, \dots, H-1\}, w \in \{0, 1, \dots, W-1\}$
 $f^{2d} = 2 \text{DDCT}(x^{2d})$

其中: $f^{2d} \in \mathbb{R}^{H \times W}$ 为二维DCT的频谱; $x^{2d} \in \mathbb{R}^{H \times W}$ 为输入; $f_{h,w}^{2d}$ 是 f^{2d} 的一个元素; $x_{i,j}^{2d}$ 是 x^{2d} 的一个元素; H 为输入 x^{2d}

的高度; W 为输入 x^{2d} 的宽度,后面余弦部分为DCT的权重。

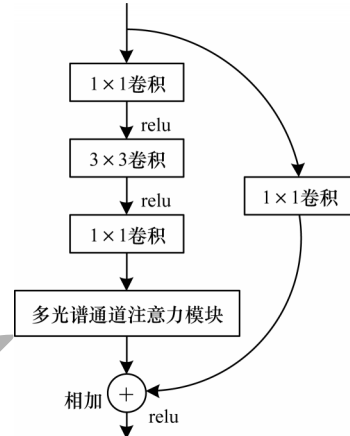


图2 插入多光谱通道注意力模块的残差块

Fig.2 Residual block inserted in the multi-spectral channel attention module

因此,二维离散余弦逆变换可定义为式(2)所示:

$$x_{i,j}^{2d} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} f_{h,w}^{2d} \underbrace{\cos\left(\frac{\pi h}{H}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(j + \frac{1}{2}\right)\right)}_{\text{DCT weights}},$$

$$\text{s.t. } i \in \{0, 1, \dots, W-1\}, j \in \{0, 1, \dots, W-1\} \quad (2)$$

为简单起见,使用 B 表示频率分量,即二维DCT的权重分量, B 的元素 $B_{h,w}^{i,j}$ 定义如式(3)所示:

$$B_{h,w}^{i,j} = \cos\left(\frac{\pi h}{H}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(j + \frac{1}{2}\right)\right) \quad (3)$$

根据式(3),可以将二维DCT的逆变换重写为式(4):

$$x_{i,j}^{2d} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} f_{h,w}^{2d} \cos\left(\frac{\pi h}{H}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(j + \frac{1}{2}\right)\right) = f_{0,0}^{2d} B_{0,0}^{i,j} + f_{0,1}^{2d} B_{0,1}^{i,j} + \dots + f_{H-1,W-1}^{2d} B_{H-1,W-1}^{i,j} \quad (4)$$

s.t. $i \in \{0, 1, \dots, H-1\}, j \in \{0, 1, \dots, W-1\}$

由式(4)可知,图像特征可表示为不同频率分量的组合。为了使用频谱表征所有特征信息,引入了MCA模块。

多光谱通道注意力模块的整体流程如图3所示。

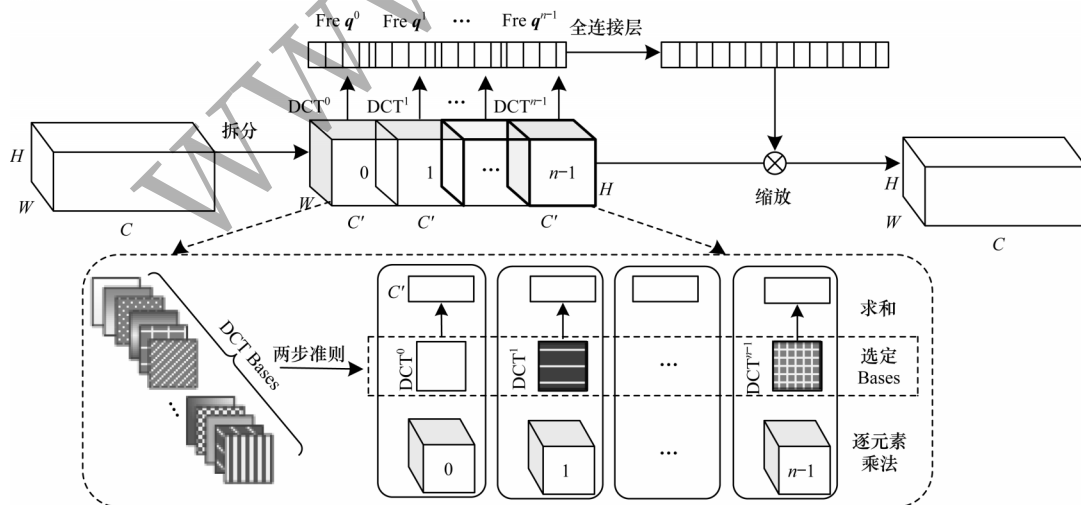


图3 多光谱通道注意力模块的整体流程

Fig.3 Overall procedure of multi-spectral channel attention module

从图3可以看出,将输入 X 沿着通道维度划分为 n 部分,用 $[X^0, X^1, \dots, X^{n-1}]$ 表示 n 个部分,其中: $X^i \in \mathbb{R}^{C' \times H \times W}, i \in \{0, 1, \dots, n-1\}; C' = \frac{C}{n}, n$ 必须能被通道数 C 整除。对于每个部分,分配相应的二维DCT频率分量,其结果可作为通道注意力的预处理结果,即:

$$\text{Fre } q^i = 2 \text{DDCT}^{u,v}(X^i) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X_{h,w}^i B_{h,w}^{u,v}, \text{ s.t. } i \in \{0, 1, \dots, n-1\} \quad (5)$$

其中: $\text{Fre } q^i \in \mathbb{R}^{C'}$ 是预处理后的 C' 维向量; $[u, v]$ 是对应于 X^i 的频率分量指数。实验中,选定Top- k 个性能最佳的频率分量后,通过 $u = r \lfloor H/7 \rfloor$ 可得出 u 的值, r 为选定频率分量中的一个值,计算 v 值同理。

最后,将各部分的预处理向量合并起来:

$$\text{Fre } q = \text{cat}([\text{Fre } q^0, \text{Fre } q^1, \dots, \text{Fre } q^{n-1}]) \quad (6)$$

其中: cat 表示向量级联; $\text{Fre } q \in \mathbb{R}^{C'}$ 是获得的多光谱向量。

整个MCA模块可以定义为:

$$\text{ms_att} = \text{sigmoid}(fc(\text{Fre } q)) \quad (7)$$

对于尺寸为 $H \times W$ 的特征,利用2维DCT将其分解为 $H \times W$ 个频率分量,得到总的频率分量为 $C \times H \times W$ 。由于测试所有组合的计算成本很高,因此通过2步准则来选择MCA模块中的频率分量。其主要思

想是先逐个计算每个频率分量的结果,然后再根据所得结果筛选出Top- k 个性能最佳的频率分量。

由于MCA模块可以将包含不同信息的频率分量合并到注意力处理中,从而提取出更多的特征信息,因此有效缓解了RetinaNet算法中存在的难以充分提取不同阶段特征的问题。

1.2 多尺度特征融合模块

MFF模块是受路径聚合网络^[12]和平衡特征金字塔^[13]的启发而构建的,其包含1个具有自底向上路径的路径聚合模块和1个特征融合操作。

1.2.1 路径聚合模块

图4为多尺度特征融合模块的结构。如图4路径聚合模块所示,在ResNet-FPN特征提取模块的后面连接具有自底向上路径的路径聚合模块。 P_3, P_4, P_5, P_6, P_7 层特征图不经过任何操作,直接作为 N_3, N_4, N_5, N_6, N_7 层特征图。以 P_4 层的特征图为例, P_4 与经过2倍下采样的 N_3 特征图根据元素相加,得到的特征图再经过 3×3 卷积后生成特征图 N_4 ,以此类推,生成特征图 N_5 ,最后生成5层通道数均为256的特征。 N_3, N_4, N_5, N_6 和 N_7 则构成了具有自底向上路径的路径聚合模块。该模块通过扩充自底向上路径,在较低特征层上用精确的定位信号增强了整个特征金字塔的信息流。

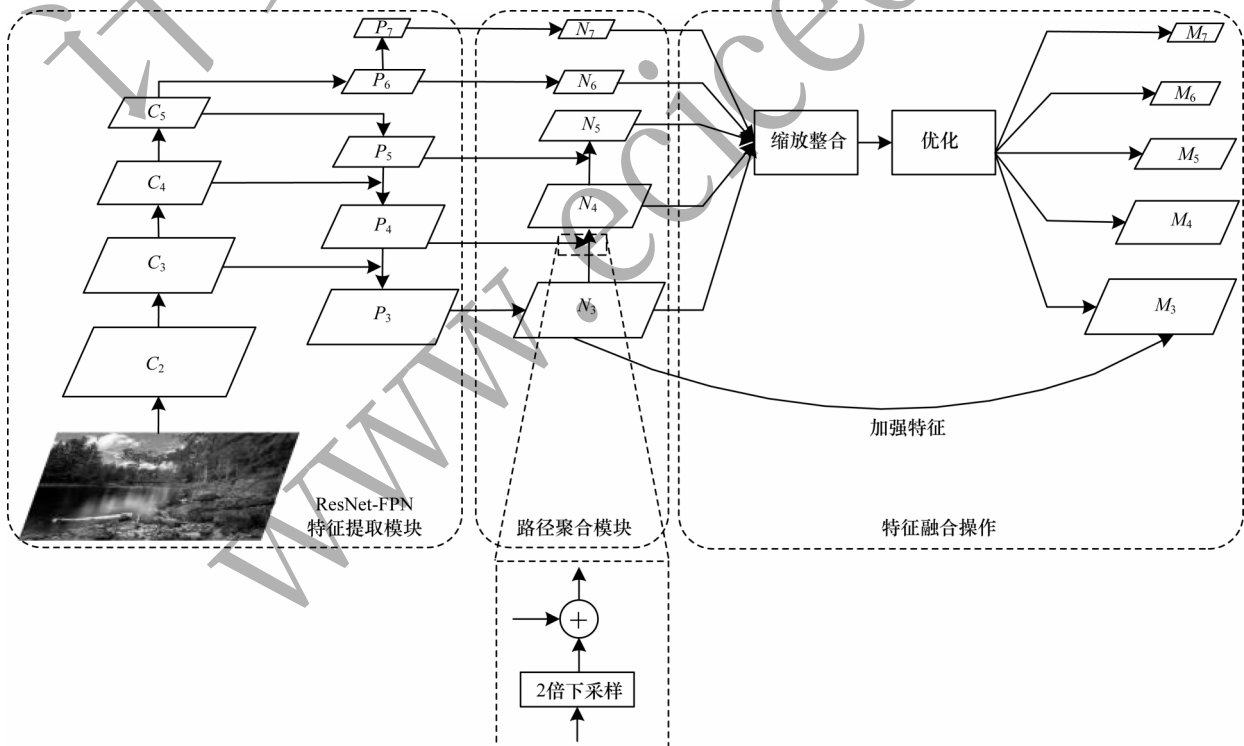


图4 多尺度特征融合模块的结构

Fig.4 Structure of multi-scale feature fusion module

1.2.2 特征融合操作

特征融合操作主要分为缩放整合、优化、加强特征这3个步骤,如图4特征融合模块所示。具体操作如下:

步骤1 调整特征图尺寸并平均融合后的特征。现有 $\{N_3, N_4, N_5, N_6, N_7\}$ 5层特征,由于低层特征分辨率高,一般只能学习到细节特征,高层特征分

分辨率低, 学习到的是语义特征。因此, 要先把这5层特征图的尺寸调整到中间层次 N_4 特征图的尺寸, 并进行融合。采取的操作是, 将 N_3 特征图下采样, 将 N_5 、 N_6 、 N_7 特征图进行上采样, 对 N_4 特征图不进行其他操作, 最后再做简单的相加取平均操作, 如式(8)所示:

$$N = \frac{1}{L} \sum_{l=\min}^{l_{\max}} N_l \quad (8)$$

其中: L 表示特征层的层数; N_l 表示第 l 层特征。

步骤2 将平均后的特征图进一步优化, 使特征图具有更强的辨别力, 使用 embedded Gaussian non-local 模块^[18]进行优化操作, 该操作的定义如下:

$$M_i = \frac{1}{C(N)} \sum_{\forall j} f(N_i, N_j) g(N_j) \quad (9)$$

$$C(N) = \sum_{\forall j} f(N_i, N_j) \quad (10)$$

$$f(N_i, N_j) = e^{\theta(N_i)^T \phi(N_j)}$$

其中: M 和 N 是尺寸相同的特征图; i 是特征图的一个像素位置; j 是所有可能位置的索引; g 是一元输入函数, 一般采用 1×1 卷积, 目的是进行信息变换; f 是配对计算函数, 计算第 i 个位置和其他所有位置的相关性; θ 和 ϕ 都是 1×1 卷积操作, T 设置为1; $C(N)$ 是归一化函数, 能够保证变换前后整体信息不变。

步骤3 将优化后的特征分散成多层特征 $\{M_3, M_4, M_5, M_6, M_7\}$, 并与 $N_3 \sim N_7$ 特征相加融合, 其中: M_3 特征是通过将优化后的特征进行上采样得到的; M_4 特征是直接输出得到的; M_5 、 M_6 和 M_7 特征是通过将优化后的特征进行下采样得到的。

上述过程是特征融合的3个步骤。通过添加路径聚合模块和特征融合操作, 有效缓解了 RetinaNet 算法难以充分融合不同阶段特征的问题。

1.3 边界框回归和分类模块

本文算法的边界框回归网络与分类网络使用的是 RetinaNet 算法的回归与分类网络。其中, 边界框回归网络是附加在输出特征每一层的全卷积网络^[19]。该网络使用4层通道数为256的 3×3 卷积, 每层卷积接一个 ReLU 激活层, 然后接1个通道数为36的 3×3 卷积层, 最后的输出预测了锚框和目标框之间的相对偏移。

得到预测值和目标值后, 便可以计算边界框回归损失, 但 RetinaNet 算法的边界框回归损失函数存在2个问题:

- 1) 如果目标框和预测框没有重叠, 那么损失函数将不起作用;
- 2) 如果两对预测框和目标框的大小均相同, 而且这2对框的相交值也相同, 那就不能确定这2对框是如何相交的。

针对上述问题, 本文算法使用 CIoU 损失函数^[14]替换了 RetinaNet 算法的边界框回归损失函数。CIoU 损失函数根据边界框回归中的重叠面积、中心点距离和长宽比这3个因素, 直接最小化预测框和目标框之间的归一化距离, 以达到更快的收敛速度。同时, 当预测框与目标框没有重叠, 或者预测框与目标框有重叠甚至有包含关系时, 该损失函数能使边界框回归更加准确。

交并比 (Intersection over Union, IoU) 的定义如式(11)所示:

$$I_{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (11)$$

其中: $B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$ 是目标框; $B = (x, y, w, h)$ 是预测框; x, y, w, h 分别是框的中心点坐标和宽高; $|B \cap B^{gt}|$ 表示目标框与预测框重叠部分的面积; $|B \cup B^{gt}|$ 表示目标框与预测框2个框包围的总面积, 如图5所示。

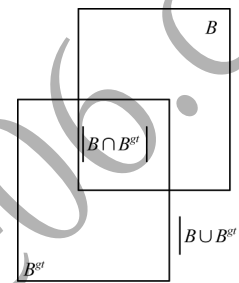


图5 IoU的定义

Fig.5 Definition of IOU

因此, CIoU 损失函数的定义如下:

$$L_{CIoU} = 1 - I_{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \quad (12)$$

其中: b 和 b^{gt} 分别表示预测框 B 和目标框 B^{gt} 的中心点; $\rho(\cdot)$ 是2个中心点的欧几里得距离; c 是同时包含预测框和目标框的最小封闭框对角线长度; α, v 是影响因子, α 是用来平衡长宽比的系数, v 是用来衡量预测框和目标框之间的长宽比一致性。 α 和 v 的定义如下:

$$\alpha = \frac{v}{(1 - I_{IoU}) + v} \quad (13)$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \quad (14)$$

参数 c 和 d 的示意图如图6所示。

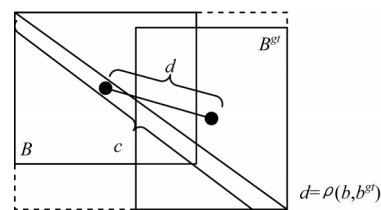


图6 参数 c 和 d 的示意图

Fig.6 Schematic diagram of parameters c and d

2 实验结果与分析

2.1 数据集和评价指标

选用MS COCO^[15]和PASCAL VOC^[16]两大公共数据集。其中MS COCO数据集包含80个类别,其中用于训练的图片有118 287张,用于验证的图片有5 000张,用于测试的图片有20 000张。令本文算法在test-dev 2017数据集上进行实验,并与最新的目标检测算法相比较,然后使用val 2017数据集进行消融实验。实验结果使用平均精度(Average Precision, AP)指标进行表征, AP表示IoU从0.5开始,每隔0.05作为阈值,直到取到0.95得到的平均精度再进行平均的结果。例如AP₅₀表示IoU阈值为0.5时的平均精度,AP₇₅表示IoU阈值为0.75时的平均精度,其它同理。AP_s、AP_m、AP_l分别表示小、中、大目标的平均精度。PASCAL VOC数据集包含20个类别,其中训练图片来自trainval 2007数据集和trainval 2012数据集,共22 136张;测试图片来自test 2007数据集,共4 952张,实验结果遵循VOC数据集的最终评价指标,即平均精度均值(mean Average Precision, mAP),其中类别精度表示该类别在IoU阈值为0.5时的平均精度。

2.2 实验参数设置

在COCO数据集中,先将输入图像的短边调整为800像素,然后使用随机梯度下降(Stochastic Gradient Descent, SGD)优化所有算法,权重衰减为 1×10^{-4} , batch size为4(2个GPU,每个GPU每训练一次选取2张图像)。学习率被初始化为0.002 5,共训练

12个epoch,并分别在第8和第11个epoch时将学习率降低10倍。在VOC数据集中,将输入图像的短边调整为600像素,在第9个epoch时将学习率降低10倍,其他设置与COCO数据集相同。

本文实验在PyTorch 1.7深度学习框架^[20]下进行,操作系统为Ubuntu 18.04,使用2个NVIDIA GeForce RTX GPU训练,显卡内存为11 GB。本文的基线算法Baseline即为RetinaNet算法,在超参数(如权重衰减、batch size、学习率、epoch等)设置均相同的情况下,对RetinaNet算法进行了重新实验,实验结果优于提出该算法的原始论文中的结果,本文的RetinaNet算法的实验结果与原始论文结果相比提高了超过1个百分点。

2.3 结果分析

2.3.1 对比实验与结果可视化

本文在COCO test-dev 2017数据集和PASCAL VOC测试集上评估了本文所提改进型RetinaNet算法的性能。在COCO test-dev 2017数据集上,实验主要分为2个部分,将主干网络为ResNet-50和ResNet-101的改进型RetinaNet算法分别与其他主干网络为ResNet-50或ResNet-101的最新一阶段、两阶段目标检测算法进行比较,实验结果如表2所示。由表2可知,主干网络为ResNet-101的改进型RetinaNet算法的AP值为40.9%,与RetinaNet算法相比性能得到显著提高。在主干网络相同的条件下,改进型RetinaNet算法与表2中的其他目标检测算法相比均达到了最佳结果。

表2 不同目标检测算法在COCO test-dev 2017数据集上的实验结果对比

Tabel 2 Comparison of experimental results of different object detection algorithms on COCO test-dev 2017 data set %							
算法	主干网络	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
RetinaNet(Baseline)算法	ResNet-50	36.9	56.3	39.3	20.6	39.9	46.5
Libra RetinaNet ^[13] 算法		37.8	56.9	40.5	21.2	40.9	47.7
FCOS ^[21] 算法		36.7	55.8	39.2	21.0	40.7	48.4
RetinaNet w/AugFPN ^[22] 算法		37.5	58.4	40.1	21.3	40.5	47.3
RetinaNet w/PISA ^[23] 算法		37.3	56.5	40.3	20.3	40.4	47.2
本文算法		39.0	58.4	41.6	22.1	42.2	48.8
Faster R-CNN ^[24] 算法	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN ^[25] 算法		38.2	60.3	41.7	20.1	41.1	50.2
LRF ^[26] 算法		37.3	58.5	39.7	19.7	42.8	50.1
RDSNet ^[27] 算法		38.1	58.5	40.8	21.2	41.5	48.2
RetinaNet w/SABL ^[28] 算法		40.5	59.3	43.6	23.0	44.1	51.3
本文算法		40.9	60.6	43.7	23.2	44.3	51.7

在PASCAL VOC测试集上,将改进型RetinaNet算法(主干网络为ResNet-50)的各个类别的精度与RetinaNet算法进行对比,结果如图7所示。由图7可知,RetinaNet算法的mPA为78.3%,改进型RetinaNet算法的mPA为79.4%,且改进型RetinaNet算法在大多数类

别上的精度高于RetinaNet算法,只在少部分类别如bird、boat、bus、sofa上的类别精度低于RetinaNet算法。其原因在于VOC训练集图片数量较少,随着网络层数的加深及参数量的增加,算法的训练效果稍低,且测试集中含有的小目标或重叠目标的图片不易被检测。

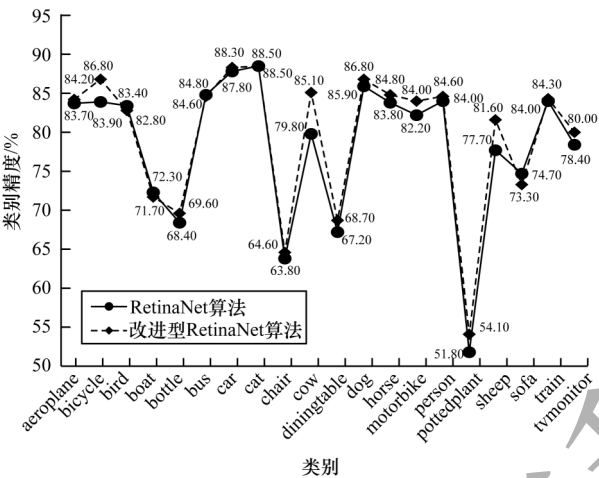


图7 不同算法在 PASCAL VOC test 2007 数据集上的结果对比

Fig.7 Comparison of results of different algorithms on PASCAL VOC test 2007 data set

从 MS COCO 数据集中随机选取一些图片进行可视化,本文选取 2 对具有代表性的检测结果进行对比,结果如图 8 所示。图 8(a)为 RetinaNet 算法的可视化结果,图 8(b)为改进型 RetinaNet 算法(主干网络为 ResNet-50)的可视化结果,由图 8 可知,改进型 RetinaNet 算法的检测结果具有更高的准确率,检测的边框更加准确。



图8 改进前后的 RetinaNet 算法在 COCO 数据集上的可视化结果对比

Fig.8 Comparison of visualization results of RetinaNet algorithm before and after improvement on the COCO data set

2.3.2 消融实验

本文所有消融实验均在 COCO val 2017 数据集上进行。实验结果均与基线算法 Baseline 进行对比,Baseline 是主干网络为 ResNet-50 的 RetinaNet 算法。

1)MCA 模块中频率分量数量的性能分析

分析不同数量的频率分量对 RetinaNet 算法的影响,也就是说在 Baseline 算法上加入具有不同频率

分量的 MCA 模块。本文选择了性能最高的前 k 个频率分量, k 分别为 4、8、16、32。

由表 3 可知,具有多光谱通道注意力的 RetinaNet 算法与基线算法相比,实验结果均存在明显差距,这验证了在通道注意力中使用多个频率分量的正确性。此外,由表 3 还可以发现当频率分量的数量为 8 时,RetinaNet 算法的 AP 值最大。

表3 不同数量的频率分量对 RetinaNet 算法的影响

Table 3 Influence of different number of frequency components on RetinaNet algorithm

频率分量数量	AP/%
4	37.0
8	37.5
16	37.3
32	37.1

2)MFF 模块中组成部分的性能分析

表 4 是在 RetinaNet 算法上加入特征融合模块各个部分的对比结果,由表 4 可知,路径聚合模块及特征融合操作分别验证了较低层特征的信息有用性、融合多层特征的有效性,将路径聚合模块与特征融合操作相结合可获得最佳性能。

表4 在 RetinaNet 算法上加入特征融合模块各个部分的对比结果

Table 4 The comparison results of each part of the feature fusion module added to the RetinaNet algorithm

加入的模块	AP/%
路径聚合模块	37.0
特征融合操作	37.6
MFF 模块	37.7

3)CIoU 损失函数不同损失权重的性能分析

在 Baseline 上加入具有不同损失权重值的 CIoU 损失函数,分析不同的损失权重对 RetinaNet 算法的影响。损失权重值分别设置为 1、2、3,实验结果如表 5 所示。由表 5 可知,CIoU 损失函数有效改善了 RetinaNet 算法中存在的边界框回归问题。当 CIoU 的损失权重值设置为 2 时,网络可获得最佳性能。

表5 不同损失权重值对 RetinaNet 算法的影响

Table 5 Influence of different loss weight values on RetinaNet algorithm

损失权重值	AP/%
1	36.9
2	37.5
3	37.1

4)3 个改进部分不同组合方式的比较

对 MCA 模块、MFF 模块、CIoU 损失函数 3 个改进部分以不同的方式进行组合,结果如表 6 所示,其中“×”代表不添加,“√”表示添加。由表 6 可知,在

基线算法上单独加入1个模块,或者加入其中2个模块都不能达到最佳性能效果,因为每个模块的作用各不相同。对于目标检测算法整体而言,特征提取、特征融合和边界框回归都是很重要的部分,因此在改进目标检测算法时,不能只着眼于部分网络,而要分析整体网络所存在的问题,再针对这些问题进行解决和改进。因此,这3个改进部分的结合不仅改善了难以充分提取和融合多层特征的缺陷,而且缓解了边界框回归不准确问题,验证了该改进算法的有效性。如表6所示,改进型RetinaNet算法在COCO val 2017数据集上的AP值比RetinaNet算法高出了2.4个百分点,性能得到显著提高。

表6 3个改进部分不同组合方式的对比实验结果
Table 6 Comparative experimental results of different combinations of three improved parts %

MCA 模块	MFF 模块	CIOU 损 失函数	AP	AP ₅₀	AP ₇₅	AP _s	AP _M	AP _L
×	×	×	36.5	55.5	38.8	20.5	40.0	48.3
√	×	×	37.5	56.8	40.2	21.5	41.1	49.5
×	√	×	37.7	57.4	40.1	22.3	41.4	49.0
×	×	√	37.5	55.9	40.1	21.1	41.3	48.9
√	√	×	37.9	57.7	40.6	22.1	42.2	49.7
√	×	√	37.9	56.8	40.3	21.6	41.8	49.4
×	√	√	38.0	56.9	40.4	21.9	41.9	49.3
√	√	√	38.9	58.1	41.2	22.9	42.8	51.0

3 结束语

本文针对RetinaNet算法难以充分提取及融合不同阶段特征、边界框回归不准确等问题,提出一种改进型RetinaNet算法。在特征提取模块中引入多光谱通道注意力模块,将路径聚合模块与特征融合操作相结合以构成多尺度特征融合模块,并在边界框回归过程中引入CIOU损失函数。在MS COCO和PASCAL VOC两大公共数据集上的实验结果表明,与RetinaNet算法相比,改进型RetinaNet算法的检测性能得到了显著提高。但目前所提算法尚未应用到两阶段目标检测中,下一步将通过调整该算法的网络结构或具体参数,将本文算法应用到两阶段目标检测中的特征提取与融合部分、边界框回归部分,使两阶段目标检测算法在提高检测精度的同时保持检测速度,提高本文算法的适用性。

参考文献

[1] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2017: 2999-3007.
[2] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA:

IEEE Press, 2017: 6517-6525.
[3] ZHANG H K, CHANG H, MA B P, et al. Cascade RetinaNet: maintaining consistency for single-stage object detection[EB/OL]. [2021-06-10]. <https://arxiv.org/abs/1907.06881>.
[4] LI Y X, REN F B. Light-weight RetinaNet for object detection[EB/OL]. [2021-06-10]. <https://arxiv.org/abs/1905.10011>.
[5] SUN P Z, ZHANG R F, JIANG Y, et al. Sparse R-CNN: end-to-end object detection with learnable proposals[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2021: 14449-14458.
[6] 吴华运, 任德均, 吕义钊, 等. 基于改进的RetinaNet医药空瓶表面气泡检测[J]. 四川大学学报(自然科学版), 2020, 57(6): 1090-1095.
WU H Y, REN D J, LÜ Y Z, et al. Bubble detection on the surface of medical empty bottles based on improved RetinaNet[J]. Journal of Sichuan University (Natural Science Edition), 2020, 57(6): 1090-1095. (in Chinese)
[7] 闫建伟, 张乐伟, 赵源, 等. 改进RetinaNet的刺梨果实图像识别[J]. 中国农机化学报, 2021, 42(3): 78-83.
YAN J W, ZHANG L W, ZHAO Y, et al. Image recognition of Rosa roxburghii fruit by improved RetinaNet[J]. Journal of Chinese Agricultural Mechanization, 2021, 42(3): 78-83. (in Chinese)
[8] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2017: 936-944.
[9] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2016: 770-778.
[10] QIN Z Q, ZHANG P Y, WU F, et al. FcaNet: frequency channel attention networks[EB/OL]. [2021-06-10]. <https://arxiv.org/abs/2012.11879>.
[11] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
[12] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2018: 8759-8768.
[13] PANG J M, CHEN K, SHI J P, et al. Libra R-CNN: towards balanced learning for object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2019: 821-830.
[14] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: faster and better learning for bounding box regression[EB/OL]. [2021-06-10]. <https://arxiv.org/abs/1911.08287>.
[15] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of Conference on Computer Vision. Berlin, Germany: Springer, 2014: 740-755.

- [16] EVERINGHAM M, GOOL L, WILLIAMS C K I, et al. The pascal visual object classes challenge [J]. International Journal of Computer Vision, 2010, 88(2): 303-338.
- [17] AHMED N, NATARAJAN T, RAO K R. Discrete cosine transform[J]. IEEE Transactions on Computers, 1974, 23(1): 90-93.
- [18] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[EB/OL]. [2021-06-10]. <https://arxiv.org/abs/1711.07971>.
- [19] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2015: 3431-3440.
- [20] PASZKE A, GROSS S, CHINTALA S, et al. Automatic differentiation in Pytorch[EB/OL]. [2021-06-10]. <https://openreview.net/forum?id=BJJsrnfCZ>.
- [21] TIAN Z, SHEN C H, CHEN H, et al. FCOS: fully convolutional one-stage object detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2019: 9626-9635.
- [22] GUO C X, FAN B, ZHANG Q, et al. AugFPN: improving multi-scale feature learning for object detection [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2020: 12592-12601.
- [23] CAO Y H, CHEN K, LOY C C, et al. Prime sample attention in object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2020: 11580-11588.
- [24] 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.
- [25] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]// Proceedings of IEEE International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2017: 2980-2988.
- [26] WANG T C, ANWER R M, CHOLAKKAL H, et al. Learning rich features at high-speed for single-shot object detection [C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2019: 1971-1980.
- [27] WANG S R, GONG Y C, XING J L, et al. RDSNet: a new deep architecture for reciprocal object detection and instance segmentation [EB/OL]. [2021-06-10]. <https://arxiv.org/abs/1912.05070>.
- [28] WANG J Q, ZHANG W W, CAO Y H, et al. Side-aware boundary localization for more precise object detection[C]// Proceedings of Conference on Computer Vision. Berlin, Germany: Springer, 2020: 403-419.

编辑 赖玉玲