控制与决策
Control and Decision

Anchor-free的尺度自适应行人检测算法

邹逸群， 肖志红， 唐夏菲， 赖普坚， 汤松林， 张泳祥， 唐琎

引用本文:
邹逸群, 肖志红, 唐夏菲, 等. Anchor-free的尺度自适应行人检测算法[J]. 控制与决策, 2021, 36(2): 295-302.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0124

Anchor-free的尺度自适应行人检测算法

邹逸群[1], 肖志红[1], 唐夏菲[2], 赖普坚[1], 汤松林[1], 张泳祥[1], 唐 琎[1†]

(1. 中南大学 自动化学院，长沙 410083；2. 长沙理工大学 电气与信息工程学院，长沙 410004)

摘 要: Anchor作为行人检测算法中的初始框,可以解决行人平移问题和缓解行人尺度变化问题,目前的行人检测算法通常都基于anchor. 然而,使用anchor就需要精心调整对检测性能影响非常大的anchor超参数,如anchor的尺度和高宽比等. 为避免这一问题,提出一种基于anchor-free损失函数的行人检测算法,并通过融合特征金字塔网络(FPN)所有检测分支的特征,使anchor-free行人检测算法在训练过程中不需要为FPN的每个检测分支设置有效的训练尺度范围. 另外，还提出一个尺度注意力(scale attention, SA)模块,用于融合FPN所有检测分支特征的过程,使网络在检测某个尺度的行人时,能够自适应地为行人所对应的不同尺度的感兴趣区域(ROI)特征赋予合适的权重. 实验结果显示,所提出的行人检测算法不仅可以实现anchor-free,从而避免anchor的超参数调整问题,而且在性能上优于其他行人检测算法,在CityPersons数据集上取得了目前最好的效果9.19 % MR^{-2}.

关键词: 行人检测；卷积神经网络；anchor-free；注意力机制；尺度自适应

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0124 **开放科学(资源服务)标识码(OSID):**

引用格式: 邹逸群,肖志红,唐夏菲,等. Anchor-free的尺度自适应行人检测算法[J]. 控制与决策, 2021, 36(2): 295-302.

Anchor-free scale adaptive pedestrian detection algorithm

ZOU Yi-qun[1], XIAO Zhi-hong[1], TANG Xia-fei[2], LAI Pu-jiang[1], TANG Song-lin[1], ZHANG Yong-xiang[1], TANG Jin[1†]

(1. School of Automation，Central South University，Changsha 410083，China；2. School of Electrical & Information Engineering，Changsha University of Science & Technology，Changsha 410004，China)

Abstract: As the initial box of the pedestrian detection algorithm, anchor can solve the problem of pedestrian translation and alleviate the problem of pedestrian scale variation. At present, the pedestrian detection algorithms are usually based on the anchor. However, the usage of the anchor requires careful adjustment of the hyper-parameters of the anchor, such as the scale and aspect ratio of the anchor, which have a great impact on the detection performance. To circumvent this problem, we present a pedestrian detection algorithm based on an anchor-free loss function. By fusing the features of all detection branches of feature pyramid network (FPN), the algorithm does not need to set an effective training scale range for each detection branch of FPN in the training process. In addition, a SA (scale attention) module is proposed to fuse all the detection branch features of FPN, so that appropriate weights can be adaptively assigned to the region of interest (ROI) features of different scales corresponding to pedestrians when the network detects a certain scale of pedestrian. Experiment results show that the proposed pedestrian detection algorithm not only realizes anchor-free, thus circumvent the super-parameter adjustment problem of the anchor, but also outperforms other pedestrian detection algorithms, achieves 9.19 % MR^2 which is the best of state-of-the-art results on CityPersons dataset.

Keywords: pedestrian detection；CNN；anchor-free；attention mechanism；scale adaptive

0 引 言

行人检测具有广泛的应用场景,如安防、自动驾驶、移动机器人等,因此,行人检测在计算机视觉领域一直是个非常热门的研究方向.

在Faster R-CNN[1]提出之前,所有基于CNN (convolution neural network)的检测算法都是先使用传统区域建议算法(如selective search算法)生成建议框,再用CNN对建议框进行分类和回归. 由于传统区域建议算法计算量十分巨大,处理一张图像需要几十毫秒甚至几百毫秒,一直是实时检测的瓶颈. 为了减少建议框生成时间,Faster R-CNN方法提出anchor机制,使用RPN (region proposal network)在backbone输出的特征图上滑窗,以anchor为初始框直接生成建议框,不仅能够达到比传统区域建议算法更高的召

收稿日期: 2020-02-09；**修回日期:** 2020-06-23.

基金项目: 国家自然科学基金青年项目(61403427)；2020湖南省科技厅青年自然科学基金项目(1541).

†通讯作者. E-mail: tjin@csu.edu.cn.

回率,而且计算成本几乎为零,从而解决了建议框生成问题. anchor在建议框生成的过程中主要起两个作用: 1) 解决物体平移问题; 2) 缓解尺度变化问题. 具体如下.

1) 解决物体平移问题.

假设图像中有两个在不同位置但是高和宽相同的物体. 如果将回归子网络的预测目标设置为物体的绝对位置,如左上角坐标和右下角坐标,则回归子网络对于这两个物体的预测目标是不一样的. 由于CNN具有平移不变性,CNN从这两个物体提取的特征向量是相同的,但现在回归子网络却需要根据这两个相同的特征向量去学习不同的目标,这显然是矛盾且不可行的. 为了解决这一问题,Faster R-CNN通过引入anchor来解耦绝对位置,使回归任务变成相对于anchor进行局部相对位置的回归. 由于不同位置的物体与anchor之间的相对位置是相同的,从而很好地解决了物体平移问题.

2) 缓解尺度变化问题.

在Faster R-CNN被提出之初,还没有像SSD (single shot detector)[2] 和FPN (feature pyramid network)[3] 这种可以显示处理尺度变化问题的网络结构,如果只使用backbone最后一个特征图来预测所有尺度的物体,这显然是一件非常困难的事. 为了缓解尺度变化问题,FasterR-CNN通过显示枚举多种尺度和高宽比的anchor,然后根据anchor匹配策略来分配预测目标,使不同尺度和不同高宽比的anchor所对应的网络权重只需要负责学习一个比较小的尺度和高宽比范围即可,从而缓解了尺度变化问题.

自从anchor机制被提出之后,很多检测算法都是基于anchor的,如Replusion Loss[4]、ALFNet[5]等. 虽然anchor很有效,但同样也存在一些问题: 1) 基于anchor的检测算法对anchor的尺度和高宽比超参数非常敏感,要想获得理想的检测性能,就需要花费大量时间去调整这些超参数; 2) 为保证高召回率,通常需要在输入图像上密集地预定义大量anchor,这样会存在大量的冗余anchor,而且在训练过程中,绝大多数anchor会被标记为负样本,只有少部分anchor会被标记为正样本,从而导致在训练过程中正负样本极度不均衡.

为了解决anchor存在的这些问题,Yolov3方法[6]使用k-means算法对训练数据集的真实框进行聚类以得到anchor; Guided anchor方法[7]使用图像特征来指导anchor的生成. 虽然这些方法解决了anchor的超参设定问题,但仍未解决anchor的冗余问题. 近几年也有一些工作质疑anchor的必要性,并提出一些anchor-free检测算法[8-9]. 这些算法虽然摆脱了对anchor的依赖,但仍然存在一些问题: 1) 这些算法大多数是基于FPN结构的,如果没有anchor,则无法根据anchor匹配策略将不同尺度的物体分配到各个检测分支,所以需要人为地为每个检测分支设置有效的训练尺度范围[8],这样又引入了额外的超参数; 2) 当多个真实框的中心映射到特征图上的同一个点时,这些算法通常是让该特征向量只预测其中一个真实框而忽略其他所有真实框,这种做法如果用在行人密集的场景下,则会极大地降低召回率.

除了anchor存在的问题,由于行人框的尺度变化范围很大,而目前的CNN并不具备尺度不变性,如何解决尺度变化问题仍然是行人检测的一大难点. 为了解决尺度变化问题,SSD方法[2]使用多个具有不同感受野的特征图来检测不同尺度的物体,在一定程度上缓解了尺度变化问题,但由于低层特征图的语义信息不足,SSD对小目标的检测效果仍然不好. 为解决这一问题,FPN方法采用自上而下的结构,使用高层语义特征图来丰富低层特征图的语义信息,极大地提升了小目标的检测效果,进一步缓解了尺度变化问题. 由于FPN结构并没有引入太多的计算和参数,近年来的检测网络[1]基本都是基于FPN结构的. 与SSD一样,FPN也会将不同尺度的物体根据anchor匹配策略分配到各个具有不同感受野的检测分支. 这种做法隐含一个先验假设: 检测小物体只需要小感受野的特征,检测大物体只需要大感受野的特征. 但这种先验假设与近年来通过引入上下文信息来提升检测性能的思想[10]相违背. 为解决这个问题,最近提出的LapNe方法[11]将FPN各个检测分支的特征融合起来,使得检测某个尺度的物体时能够充分利用多个感受野的特征. 这种方法虽然对检测性能有所提升,但是在检测某个尺度的物体时,如何权衡来自各个检测分支的特征仍然没有得到解决.

针对尺度变化问题,本文在模型设计部分将FPN所有检测分支的特征进行融合,并提出一个SA模块用于该特征融合过程,使网络在检测某个尺度的行人时能够自适应地权衡来自FPN各个检测分支的ROI(region of interest)特征. 为了使行人检测算法摆脱对anchor的依赖,本文在损失函数设计部分提出一种anchor-free方法,不仅可以实现anchor-free,而且相比于其他基于anchor的行人检测算法,具有更快的检测速度和更低的误检率. 最后,通过实验验证了本文方法的有效性.

1 模型设计

1.1 网络整体结构

由于anchor-free属于损失函数设计的一部分,且本文实现anchor-free的方法也与网络结构有关,本文首先讲解网络结构的设计,然后再讲解anchor-free的损失函数设计.

图 1 为 anchor-free 的尺度自适应行人检测算法 (anchor-free scale adaptive pedestrian detection algorithm, AFSA) 的网络结构. AFSA 的网络结构共由 3 部分组成: 特征提取网络、SA 模块、分类和回归子网络. 接下来本文将分别对这 3 部分的设计方法进行详细讲解.

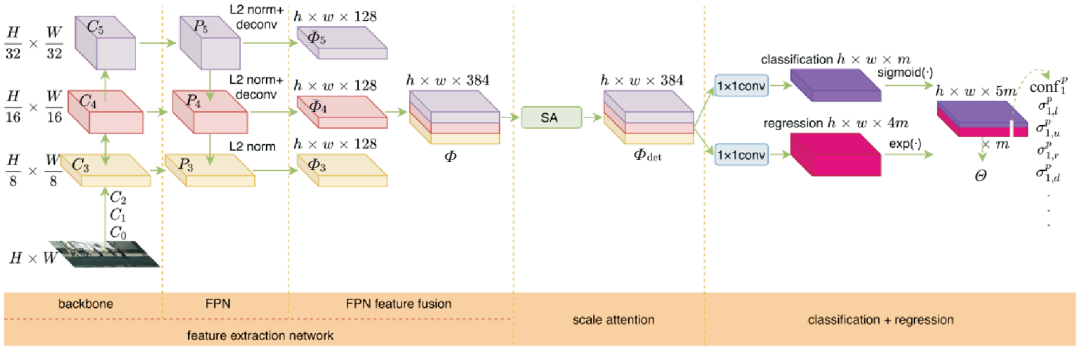

图 1　AFSA 的网络结构

1.1.1　特征提取网络设计

AFSA 的特征提取网络是在 Yolov3[6] 的 FPN 基础上改进的. 本文所有实验均是在 CityPersons 数据集[12] 上进行的, 而该数据集的图像尺度为 1 024×2 048, 受 GPU 显存的限制, AFSA 的训练 batch size 最大设置为 2. 文献 [13] 指出, 当 batch size 很小时, batch normalization[14] 的性能远不如 group normalization[13], 所以本文将 FPN 中的 batch normalization 层全部替换成 group normalization 层.

本文用 H 和 W 分别表示输入图像的高和宽, $P_3 \in R^{h\times w\times 128}$、$P_4 \in R^{h/2\times w/2\times 256}$、$P_5 \in R^{h/4\times w/4\times 512}$ 分别表示 backbone 中第 3、第 4、第 5 阶段所对应的 FPN 输出特征图, 其中 $h = H/8$ 和 $w = W/8$ 分别表示 P_3 特征图的高和宽. 因为本文基于 anchor-free, 所以无法根据 anchor 匹配策略将不同尺度的行人分配到相应的检测分支. 如果使用传统的 FPN 结构, 则需要像文献 [8] 一样为 FPN 的每个检测分支设置有效的训练尺度范围, 这样又会引入额外的超参数. 为了解决这一问题, 本文使用文献 [11] 中提出的方法: 将 FPN 所有检测分支的输出特征图沿通道级联起来. 首先对 P_3、P_4、P_5 进行 L_2 归一化[15], 得到 $\Phi_3 \in R^{h\times w\times 128}$、$\Phi_4 \in R^{h/2\times w/2\times 256}$、$\Phi_5 \in R^{h/4\times w/4\times 512}$; 然后使用转置卷积将 Φ_4 和 Φ_5 的尺度和通道调整到与 Φ_3 相同, 得到 $\Phi_4 \in R^{h\times w\times 128}$ 和 $\Phi_5 \in R^{h\times w\times 128}$; 最后, 再将 Φ_3、Φ_4、Φ_5 沿通道级联起来, 得到 $\Phi = \{\Phi_3, \Phi_4, \Phi_5\} \in R^{h\times w\times 384}$. 这样不仅可以避免为 FPN 的每个检测分支设置有效的训练尺度范围, 而且在检测某个尺度的行人时, 网络也可以利用多个感受野的特征.

1.1.2　SA 模块设计

如图 2 所示, 图中红色框标注的行人在 Φ_3、Φ_4、Φ_5 上分别对应一块 ROI 特征, 这 3 块 ROI 特征具有不同大小的感受野. 从图 2 中可以看出, 只有 Φ_4 上的 ROI 特征所对应的感受野与该行人最匹配, 所以分类和回归子网络在预测该行人时, 应该更注重 Φ_4 上的 ROI 特征. 但如果直接将 Φ 作为检测特征图送入分类和回归子网络进行预测, 则分类和回归子网络在预测该行人时, 将同等对待这些 ROI 特征, 这显然是不合理的.

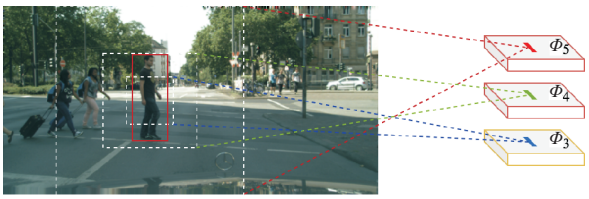

图 2　行人对应的 ROI 特征和 ROI 特征对应的感受野示意

为了让网络在检测某个尺度的行人时, 能够自适应地为该行人所对应的多个具有不同感受野的 ROI 特征赋予合适的权重, 本文设计一个 SA 模块, 其整体结构如图 3 所示.

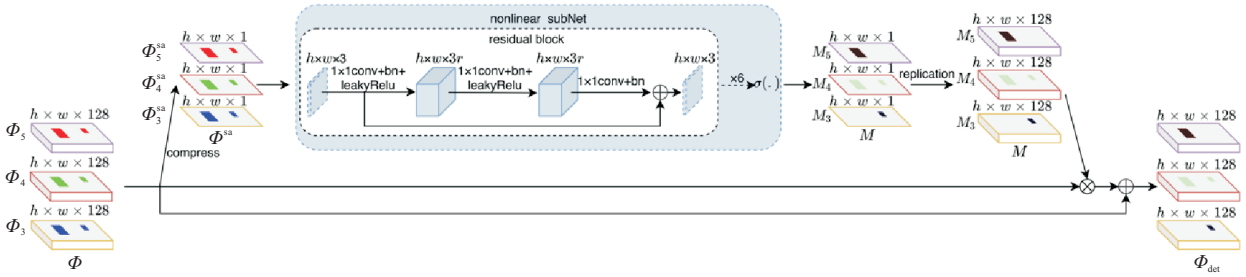

图 3　SA 模块

为了降低SA模块的复杂度,在设计SA模块时,首先使用下式分别对Φ_3、Φ_4、Φ_5进行降维:

$$\Phi^{\rm sa}_{(k=3,4,5),ij}=\frac{1}{128}\sum_{c=1}^{128}\Phi_{k,ijc},\tag{1}$$

然后将降维后得到的$\Phi^{\rm sa}_3$、$\Phi^{\rm sa}_4$、$\Phi^{\rm sa}_5$沿通道级联起来得到$\Phi^{\rm sa}\in R^{h\times w\times 3}$.

为了对不同尺度特征之间的相关性进行建模,在对Φ_3、Φ_4、Φ_5降维后,又设计了一个子网络. 这个子网络必须满足两个要求: 1) 它必须能够学习不同尺度特征之间的非线性相互作用; 2) 它必须学会一种非互斥的关系,因为SA模块是确保增强多个尺度的特征,而不是只增强其中某一个尺度的特征. 为了达到这些目的,本文参考SENet方法[16]实现自注意力机制的思想,设计了一个具有sigmoid激活门控机制的非线性子网络,即

$$M=\sigma(F(\Phi^{\rm sa})).\tag{2}$$

其中: $M\in R^{h\times w\times 3}$表示非线性子网络输出的多尺度特征注意力图; σ表示sigmoid函数; F表示一个非线性函数,由6个残差模块组成. 非线性子网络的输入特征图$\Phi^{\rm sa}$的通道非常少,且文献[17]指出,如果在低维特征向量中使用激活函数,则会损失特征向量中所包含的信息,而且损失的信息是无法恢复的. 因此,在设计非线性子网络中的残差模块时,首先使用1×1卷积层将低维输入特征图的通道扩张r倍,本文在实验中发现$r=32$时效果最好; 然后,在高维特征图后面使用leakyRelu激活函数来引入非线性,添加1×1卷积层以增强非线性子网络的建模能力; 最后在残差连接前,使用1×1卷积层把高维特征图降维至与残差模块的输入相同,且降维后的低维特征图不使用leakyRelu激活函数,以避免损失信息. 与文献[17]一样,残差模块中每个卷积层后面均使用了batch normalization层.

多尺度特征注意力图M的通道为3,即SA模块为特征图$\Phi_{i=3,4,5}\in R^{h\times w\times 128}$分别学得一个注意力图$M_{i=3,4,5}\in R^{h\times w\times 1}$. 为方便$M_i$与$\Phi_i$相乘,此处分别将$M_i$沿通道维复制128次,得到$M_{i=3,4,5}\in R^{h\times w\times 128}$. 因为$M$的值都在(0,1)范围内,所以如果直接将$M$乘到输入特征图$\Phi$上,则会弱化特征图中的所有特征[18]. 为了解决这个问题,与文献[18]一样,SA模块使用残差注意力学习方式,其计算方法为

$$\Phi_{\rm det}=(1+M)\times\Phi,\tag{3}$$

其中$\Phi_{\rm det}\in R^{h\times w\times 384}$表示SA模块输出的用于分类和回归子网络检测的特征图.

1.1.3 分类和回归子网络设计

由于设计分类和回归子网络不是本文的重点,为简单起见,本文的分类和回归子网络都只使用一个带偏置的1×1卷积层. 如果使用其他精心设计的分类和回归子网络,如文献[19],则会进一步提升AFSA的性能. 因为分类子网络输出的是行人置信度,而行人置信度的范围为(0,1),所以在分类子网络后面还添加了sigmoid函数. 此外,下一部分中的anchor-free将指出,回归层的回归目标都是大于0的,所以本文在回归层后面使用指数函数$\exp(\cdot)$,用来将回归层的输出限制在$(0,\infty)$范围内.

1.2 anchor-free的损失函数设计

1.2.1 anchor-free设计

如果要实现anchor-free,则需要在不使用anchor的情况下解决行人平移问题和尺度变化问题. 由于现在的FPN和SSD已经可以显示地处理尺度变化问题,本文使用的FPN结构可以替代anchor缓解尺度变化问题的作用. 对于行人平移问题,本文采用与文献[9]类似的方法,通过预测行人框中心映射到特征图上的点与行人框的左上角坐标和右下角坐标之间的偏差来解决. 接下来,将详细阐述本文实现anchor-free的方法.

本文用$\Theta\in R^{h\times w\times c}$表示分类和回归子网络最终的输出,$\Theta$上每个预测框对应的标签为$(1^{\rm valid},{\rm conf},\sigma_l,\sigma_u,\sigma_r,\sigma_d)$. 其中: $1^{\rm valid}$表示该预测框是否参与训练,conf表示行人置信度预测目标,$(\sigma_l,\sigma_u,\sigma_r,\sigma_d)$表示回归目标. 输入图像中的行人框用$\{B_i=(x^i_{\min},y^i_{\min},x^i_{\max},y^i_{\max})\}$表示,其中$(x^i_{\min},y^i_{\min})$和$(x^i_{\max},y^i_{\max})$分别表示行人框的左上角坐标和右下角坐标.

在分配正样本时,首先将行人框B_i的中心$(x^i_{\rm center},y^i_{\rm center})$映射到$\Theta$上,如果映射到$\Theta$的点$(x,y)$上,则点$(x,y)$所对应的预测框将负责预测行人框$B_i$. 但如果$\Theta$上的每个点只有一个预测框,则当多个行人框的中心映射到Θ上的同一个点时,将难以确定让预测框去预测哪个行人框. 文献[8-9]一般是选择预测其中尺度最小的行人框,而丢弃其他行人框,这种解决方法如果用在行人框中心重叠度很高的场景,则会严重降低召回率. 针对行人框丢弃的问题,本文根据数据集的行人密集情况,通过让Θ上的每个点预测多个行人框来解决,具体方法如下.

1) 计算训练数据集中的最大重叠次数m.

用$\{I_i,\ i=1,\ldots,n\}$表示训练数据集,首先将图像I_i的所有行人框中心映射到Θ上,然后计算Θ上的最大重叠次数o_i,这样每张图像I_i都会对应一个o_i,最后使用

$$m=\max_{i=1,2,\ldots,n}o_i\tag{4}$$

计算整个训练数据集的最大重叠次数m.

2) 通过Θ上的每个点预测m个行人框.

计算出整个训练数据集的最大重叠次数m后,将Θ的输出通道c设置为5 m,即Θ上的每个点有m个预测框. 在分配正样本标签时,如果行人框B_i的中心点映射到Θ的点(x,y)上,则点(x,y)所对应的m个预测框的标签首先全部设置为$(1,1,\sigma_l^i,\sigma_u^i,\sigma_r^i,\sigma_d^i)$,其中$(\sigma_l^i,\sigma_u^i,\sigma_r^i,\sigma_d^i)$的计算方法为

$$\sigma_l^i = x + 0.5 - \frac{x_{\min}^i}{8},\ \sigma_u^i = y + 0.5 - \frac{y_{\min}^i}{8},$$

$$\sigma_r^i = \frac{x_{\max}^i}{8} - x - 0.5,\ \sigma_d^i = \frac{y_{\max}^i}{8} - y - 0.5. \quad (5)$$

如果有第2个行人框B_j的中心点映射到点(x,y)上,则将点(x,y)所对应的第2个预测框的标签设置为$(1,1,\sigma_l^j,\sigma_u^j,\sigma_r^j,\sigma_d^j)$,以此类推. 这样,如果只有一个行人框$B_i$的中心映射到点$(x,y)$上,则点$(x,y)$所对应的$m$个预测框都将以$B_i$为预测目标. 因为在训练过程中这$m$个预测框的预测目标相同,所以在测试阶段这$m$个预测框的预测结果也会很接近,在NMS(non maximum suppression)时将会去掉其中$m-1$个预测结果而只保留其中置信度最高的预测框,并不会造成误检. 另一方面,这种方法在训练过程中不会丢弃任何行人框,因此即使在密集行人场景下也不会降低召回率.

当分配负样本时,如果简单地把所有非正样本的预测框全部设置为负样本,则会产生两个问题: 1) 极度的正负样本不均衡,以1 024 × 2 048的输入图像为例,假设该图像中有30个行人,则正负样本的比例将高达1:1 000,这种极端的正负样本不均衡会导致严重的漏检; 2) 正如文献[15]所述,如果行人框B_i的中心映射到Θ的点(x,y)上,则点(x,y)附近区域的点所对应的特征向量与点(x,y)所对应的特征向量很相似,如果将这些点所对应的预测框全部设置为负样本,则会严重影响正样本的学习,同样会造成漏检. 为了缓解这些问题,本文提出一种针对anchor-free的"负样本选择策略": 首先将所有非正样本的预测框全部设置为负样本,即标签设置为(1,0,0,0,0,0); 然后在计算损失时,计算标记为负样本的预测框与输入图像中所有行人框的IOU(intersection of union),如果最大IOU大于0.5,则认为该预测框已经能够相对准确地预测出行人框,不应属于负样本,因此,在训练过程中该预测框所对应的1^{valid}标签将被修改为0,即该预测框将不参与负样本损失计算.

1.2.2 损失函数

本文定义的损失函数如下:

$$L = \sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{c}(L_{ijk}^{\text{conf}} + L_{ijk}^{\text{reg}}),$$

$$L_{ijk}^{\text{conf}} = 1_{ijk}^{\text{valid}} \times w_{ijk}^{\text{conf}} \times \text{CE}(\text{conf}_{ijk}^{p}, \text{conf}_{ijk}),$$

$$w_{ijk}^{\text{conf}} = |\text{conf}_{ijk}^{p} - \text{conf}_{ijk}|^{\lambda},$$

$$L_{ijk}^{\text{reg}} = 1_{ijk}^{\text{positive}} \times w_{ijk}^{\text{reg}} \times (1 - \text{GIOU}(\sigma_{ijk}^{p}, \sigma_{ijk})),$$

$$w_{ijk}^{\text{reg}} = 2 - \frac{h_{ijk}^{\text{bbax}} w_{ijk}^{\text{bbax}}}{\text{HW}}. \quad (6)$$

其中: h、w、c分别是Θ的高、宽、通道; 1_{ijk}^{valid}表示该预测框是否参与损失计算,对应预测框的第1位标签值; conf_{ijk}^{p}表示预测的行人置信度; conf_{ijk}表示预测框所对应的行人置信度预测目标,对应预测框的第2位标签值; CE(cross entropy)表示交叉熵函数; 因为$1_{ijk}^{\text{positive}} = \text{conf}_{ijk}$,所以只有正样本才参与回归损失的计算; σ_{ijk}^{p}表示预测的偏差; σ_{ijk}表示预测框所对应的偏差预测目标,对应预测框的后4位标签值; GIOU (generalized intersection over union)是文献[20]提出的一种替代IOU的度量方法; w_{ijk}^{conf}表示置信度损失的权重,类似于focal loss,可用来缓解正负样本不均衡问题; λ在本文的取值为2; w_{ijk}^{rep}用于平衡不同尺度行人的回归损失.

2 实验研究

2.1 实验环境

2.1.1 数据集和评价指标

为了验证AFSA的有效性,本文在CityPersons行人检测数据集[12]上做了大量实验. 相比于经典的Caltech行人检测数据集,CityPersons的行人密集程度和遮挡程度更严重,是目前行人检测领域中最具挑战性的数据集. 由于CityPersons的测试集没有提供标签数据,本文与文献[12]一样,只在训练集中训练,在验证集中测试. 与文献[12]一样,本文采用MR^{-2} (miss rate over false positive per mage (FPPI) ranging in $[10^{-2}, 10^{0}]$)作为评估指标,并根据行人框的可视化程度和尺度范围,将验证集细分为reasonable、heavy、partial、bare、small、middle、large共7个子验证集,其中reasonable为主要验证集.

2.1.2 训练和测试设置

本文在CityPersons数据集上统计出来的m等于2,所以在实验中如果没有特殊说明,则默认将Θ的输出通道设置为10,即Θ上的每个点都有2个预测框. 为了增加训练数据的多样性,本文在训练期间采用一些简单的数据增强方法,如多尺度训练、随机裁剪、随机平移、随机翻转、mix up[21]等. 本文遵循文献[12]中的做法,将图像中标签为ignore和person group的区域填充为128,并去掉ignore和group标签,且在训练过程中去掉尺度小于5的行人框. backbone采用的是在ImageNet数据集上预训练过的darknet53,对于其他非backbone的权重采用均值为0、方差为0.01的高斯分布来随机初始化,所有偏置均初始化

为0. 在训练过程中,本文的计算资源为一块RTX 2080Ti, batch size设置为2. 网络优化器采用Adam,学习率衰减策略使用cosine learning rate[21],初始学习率为10^{-4},最终学习率为10^{-6},并在训练的前两个周期使用warm up[21]策略来稳定训练,最大训练周期为100个周期.

在测试期间,为了与文献[15]保持相同的测试环境,本文的测速GPU为GTX 1080Ti,置信度阈值设置为0.1,NMS的阈值设置为0.5,测试尺度为CityPersons的原图尺度$1\,024\times 2\,048$,并去掉所有尺度小于5的检测框.

2.2 anchor-free的实验

在进行anchor-free的实验之前,本文首先测试了在不使用SA模块的情况下,基于anchor的AFSA的性能,用于给anchor-free实验提供对比基准. 为了尽可能获取高质量的anchor,本文通过对CityPersons训练集中的真实框使用k-means聚类算法得到anchor. 聚类过程中使用的距离度量标准为

$$d(\text{box},\text{centriod})=1-\text{IOU}(\text{box},\text{centriod}). \quad (7)$$

其中: box表示真实框,centriod表示聚类中心. 由于anchor-free中Θ上的每个点默认预测2个框,为了对比公平,在聚类时也只聚类2个anchor. 最终聚类得到两个anchor分别为(22, 51)和(60, 144). 在训练过程中,使用与Yolov3相同的正负样本分配策略,最终在reasonable验证集上的MR^{-2}为9.78 %,如表1的第1行所示.

表 1 anchor-free消融实验表

method	负样本选择策略	Θ上每个点预测m个检测框	reasonable
anchor-based	—	—	9.78 (10.50)
anchor-free	—	—	23.16 (24.68)
anchor-free	✓	—	9.85 (10.15)
anchor-free	✓	✓	9.51 (9.92)

为了对比公平,与上述基于anchor的方法一样,接下来的anchor-free消融实验也都不使用SA模块. 如前所述,为解决anchor-free算法在分配负样本时存在模棱两可的问题,本文提出一种“负样本选择策略”. 从表1的实验结果来看,不使用“负样本选择策略”的anchor-free算法在reasonable验证集上的MR^{-2}为23.16 % (MR^{-2}越高,检测效果越差),远不如上述基于anchor的方法. 使用“负样本选择策略”之后,本文anchor-free算法的性能得以大幅提升,MR^{-2}降低至9.85 %,与上述基于anchor的方法基本一致. 因此“负样本选择策略”对于AFSA实现anchor-free是不可或缺的.

本文在实验中发现,CityPersons的验证集较小,仅有500张图像,具体到各个子验证集则更小,因此,在训练过程中即使模型已经收敛,其性能仍然存在波动. 为了保证实验数据的有效性和公平性,本文在实验中记录了模型的两种性能: 1) 获取整个训练期间在reasonable验证集上MR^{-2}最低的模型,然后记录该模型在所有子验证集上的性能; 2) 记录最后10个周期所对应的模型在所有子验证集上的平均性能,其中第2种性能记录在实验表格的括号中,在对比实验中,本文主要对比第1种性能,第2种性能对比只作为参考.

为解决因多个行人框的中心映射到Θ上重叠而带来的漏检问题,本文通过让Θ上的每个点预测m个行人框来解决. 如表1所示,该方法可以进一步将模型在reasonable验证集上的MR^{-2}降低至9.51 %. 但因为AFSA的默认网络步长为8,且CityPersons数据集中行人密集程度不高,所以即使Θ上的每个点只有1个预测框,在训练期间也只会丢弃$275/19\,654=1.4\,\%$的行人框. 为了进一步验证该方法在行人密集场景下的有效性,本文补充了另一组实验,该实验将AFSA的网络步长调整至32,这样,如果Θ上的每个点只有一个预测框,则在训练期间将会丢弃$2\,739/19\,654=13.9\,\%$的行人框,一定程度上可以模拟行人密集场景. 从表2的实验结果可以看出,在行人密集场景下,该方法可以有效地降低漏检.

表 2 网络步长为32时的对比实验

Θ上每个点预测m个检测框	丢失框	reasonable
1	2 739	19.15 (21.82)
4	0	15.90 (16.33)

2.3 尺度attention实验

为了验证SA模块的有效性,本文对比了bseline (不加SA模块)和baseline + SA,实验结果如表3所示. 可以看出,使用SA模块后,模型在reasonable验证集上的MR^{-2}进一步降低至9.19 %,在各个尺度范围上的MR^{-2}也均有降低,其中小目标的MR^{-2}降低最为明显,降低了1.63 % MR^{-2}.

使用SA模块后,会引入61 092个参数,为了进一步验证使用SA模块带来的性能提升不是因为引入了额外的参数,本文将SA模块替换为一个输入输出

表 3 SA模块对比实验

method	added paras	reasonable	small	medium	large
baseline	0	9.51 (9.92)	13.10 (13.80)	4.16 (4.16)	5.89 (6.17)
+conv	148 224	9.85 (10.30)	14.38 (14.55)	4.56 (4.45)	5.79 (5.87)
+SA	61 092	9.19 (9.54)	11.47 (12.47)	4.16 (3.99)	5.65 (5.57)

通道均为 384 的 1 × 1 卷积层, 且在该卷积层后面添加 batch normalization 层. 从表 3 所示的实验结果来看, baseline 添加卷积层之后, 共引入了 148 224 个参数, 其引入的参数是 SA 模块的 2.4 倍, 但性能反而下降了, 说明简单地增加网络参数并不会提升性能, 甚至会造成过拟合而降低模型性能. 因此使用 SA 模块能够提升性能, 并不是因为增加了模型的参数量.

2.4　与其他最先进算法的比对实验

表 4 对比了 AFSA 与之前在 CityPersons 数据集上取得过最好效果的算法. 可以看出, AFSA 在所有验证集上均取得了目前最好的效果, 在 reasonable 这个主要验证集上的 MR^{-2}, 比之前效果最好的 CSP 低 1.81 %. 值得关注的是, AFSA 在小目标上表现非常出色, 在 small 验证集上达到了 11.47 % MR^{-2}, 比其他算法低 4 % MR^{-2} 以上, 且在没有针对遮挡问题做特殊处理的情况下, AFSA 在 Heavy 验证集上的检测效果也大幅超越其他算法, 包括专门用于解决遮挡问题的文献 [22]. 不仅在性能上取得了最好的效果, AFSA 在速度上也表现得十分优异, 处理一张 1 024 × 2 048 的图像也只需要 214 ms, 优于文献 [15] 等检测算法.

图 4 是 CityPersons 验证集中典型图像的检测结果, 可以看出, AFSA 在各种复杂情况下均能准确地将行人检测出来, 包括密集场景、行人遮挡、行人重叠、极小行人等.

表 4　AFSA 与之前在 CityPersons 上取得过最好效果的算法进行对比

method	reasonable	heavy	partial	bare	small	medium	large	test time / (ms / img)
FRCNN[1]	15.4	—	—	—	25.6	7.2	7.9	—
RepLoss[4]	13.2	56.9	16.8	7.6	—	—	—	—
OR-CNN[22]	12.8	55.7	15.3	6.7	—	—	—	—
ALFNet[5]	12.0	51.9	11.4	8.4	19.0	5.7	6.6	270
CSP[15]	11.0	49.3	10.4	7.3	16.0	3.7	6.5	330
AFSA	9.19	43.65	8.51	6.06	11.47	4.16	5.65	214

图 4　CityPersons 验证集中典型图像的检测结果

3　结　论

本文首先分析 anchor 在行人检测算法中的作用; 然后设计了一种 anchor-free 方法, 并通过融合 FPN 所有检测分支的特征, 使 AFSA 在训练时不需要人为地将不同尺度的行人分配到各个检测分支; 最后, 提出了一个 SA 模块用于 FPN 的特征融合过程, 使网络在检测某个尺度的行人时, 能够自适应地为行人所对应的多个具有不同感受野的 ROI 特征赋予合适的权重, 以增强 AFSA 对行人尺度变化的鲁棒性. 通过在 CityPersons 数据集上的实验表明, 所提出的 AFSA 不仅实现了 anchor-free, 而且在速度和性能上均优于其他行人检测算法, 对遮挡和小目标问题也能处理得很好.

尽管本文设计的行人检测算法已经达到了非常不错的性能, 但仍然存在一些不足之处, 需要进一步研究, 具体包括以下几点:

1) 由于本文提出的行人检测算法是基于 CNN 的, 计算量较大, 要想实时检测, 需要使用 GPU 进行加速, 然而, 很多实际场景中的计算平台是没有 GPU 的. 因此, 要将本文算法广泛应用到实际场景中, 在后续工作中, 还需要对本文算法进行加速处理.

2) 由于实验室计算资源有限, 本文在实验中只使用了 CiytPersons 这一个数据集. 尽管与其他行人检测数据集相比, CityPersons 的数据更具多样性, 但一个数据集始终无法对模型的泛化能力进行评估. 因此, 在后续工作中, 还需要使用更多的行人检测数据集对本文方法进行交叉验证, 以评估本文方法的泛化能力.

参考文献 (References)

[1] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. Advances in Neural Information Processing

Systems. Montreal: IEEE, 2015: 91-99.

[2] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. European Conference on Computer Vision. Cham: IEEE, 2016: 21-37.

[3] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 936-944.

[4] Wang X L, Xiao T T, Jiang Y N, et al. Repulsion loss: Detecting pedestrians in a crowd[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7774-7783.

[5] Liu W, Liao S C, Hu W D, et al. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting[C]. Proceedings of the European Conference on Computer Vision. Munich: IEEE, 2018: 618-634.

[6] Redmon J, Farhadi A. Yolov3: An incremental improvement[EB/OL]. (2018-04-08)[2020-02-05]. https://arxiv.org/abs/1804.02767.

[7] Wang J Q, Chen K, Yang S, et al. Region proposal by guided anchoring[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2960-2969.

[8] Zhu C C, He Y H, Savvides M. Feature selective anchor-free module for single-shot object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 840-849.

[9] Tian Z, Shen C H, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]. Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 9626-9635.

[10] Li J N, Wei Y C, Liang X D, et al. Attentive contexts for object detection[J]. IEEE Transactions on Multimedia, 2017, 19(5): 944-954.

[11] Chabot F, Chaouch M, Pham Q C. LapNet: Automatic balanced loss and optimal assignment for real-time dense object detection[EB/OL]. (2019-11-04) [2020-02-05]. https://arxiv.org/abs/1911.01149.

[12] Zhang S S, Benenson R, Schiele B. CityPersons: A diverse dataset for pedestrian detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4457-4465.

[13] Wu Y X, He K M. Group normalization[C]. Proceedings of the European Conference on Computer Vision. Munich: IEEE, 2018: 3-19.

[14] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB/OL]. (2015-02-11) [2020-02-05]. https://arxiv.org/abs/1502.03167.

[15] Liu W, Liao S C, Ren W Q, et al. High-level semantic feature detection: A new perspective for pedestrian detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5182-5191.

[16] 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: IEEE, 2018: 7132-7141.

[17] Sandler M, Howard A, Zhu M L, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510-4520.

[18] Wang F, Jiang M Q, Qian C, et al. Residual attention network for image classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6450-6458.

[19] Li Z, Peng C, Yu G, et al. Light-head R-CNN: In defense of two-stage object detector[EB/OL]. (2017-11-20) [2020-02-05]. https://arxiv.org/abs/1711.07264.

[20] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 658-666.

[21] Zhang Z, He T, Zhang H, et al. Bag of freebies for training object detection neural networks[EB/OL]. (2019-02-11) [2020-02-05]. https://arxiv.org/abs/1902.04103.

[22] Zhang S F, Wen L Y, Bian X, et al. Occlusion-aware R-CNN: Detecting pedestrians in a crowd[C]. Proceedings of the European Conference on Computer Vision. Munich: IEEE, 2018: 637-653.

作者简介

邹逸群 (1985−), 男, 副教授, 博士, 从事系统辨识、目标定位等研究, E-mail: yiqunzou@csu.edu.cn;

肖志红 (1995−), 男, 硕士生, 从事图像处理、机器视觉的研究, E-mail: StinkyTofu95@gmail.com;

唐夏菲 (1984−), 女, 讲师, 博士, 从事非线性系统的噪音屏蔽、线性系统的模型辨识的研究, E-mail: xiafei.tang@csust.edu.cn;

赖普坚 (1993−), 男, 硕士生, 从事图像处理、机器视觉的研究, E-mail: laipujian0921@163.com;

汤松林 (1996-), 男, 硕士生, 从事图像处理、机器视觉的研究, E-mail: 992164786@qq.com;

张泳祥 (1994−), 男, 硕士生, 从事图像处理、机器视觉的研究, E-mail: 1213271098@qq.com;

唐琎 (1966−), 男, 教授, 博士生导师, 从事图像处理、机器视觉等研究, E-mail: tjin@csu.edu.cn.

(责任编辑：李君玲)