计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(4):171−179 [doi: 10.15888/j.cnki.csa.008427]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合注意机制和多尺度卷积的 YOLO 行人检测①

孙家慧, 葛华勇, 张哲浩

(东华大学 信息科学与技术学院, 上海 201620)
通信作者: 葛华勇, E-mail: gehuayong@dhu.edu.cn

摘　要: 为提高行人检测的检测性能, 本文结合 SqueezeNet、注意力机制、空洞卷积和 Inception 等结构, 提出一种基于改进 YOLOv4 的行人检测算法. 改进 YOLO 在特征增强部分引入残差连接和结合空洞卷积的注意力模块 D-CBAM, 可以从提取到的特征中选择对目标检测重要的信息. 此外, 结合 SqueezeNet 的“squeeze-expand”结构和 Inception 网络的多尺度卷积思想提出 Inception-fire 模块用于替代网络中的连续卷积层, 通过增加网络的宽度达到提升算法性能的效果, 同时减少网络的参数. 最后, 根据行人检测任务的特点并结合 Focal loss 对损失函数进行改进, 分别对正负样本和难易样本添加权重因子, 强调对正样本和难分类样本的训练, 从而提高网络的检测能力. 改进的 YOLO 算法在 INRIA 行人数据集上的检测精度能够达到 94.95%, 相对原 YOLOv4 提高 4.25%, 同时参数量减少了 36.35%, 检测速度也获得 13.54% 的提升, 在行人检测中能够表现出更优秀的性能.

关键词: YOLOv4; 注意力机制; SqueezeNet; Inception; ResNet; 焦点损失; 深度学习; 目标检测

引用格式: 孙家慧,葛华勇,张哲浩.结合注意机制和多尺度卷积的 YOLO 行人检测.计算机系统应用,2022,31(4):171–179. http://www.c-s-a.org.cn/1003-3254/8427.html

YOLO Pedestrian Detection with Attention Mechanism and Multi Convolution Kernel

SUN Jia-Hui, GE Hua-Yong, ZHANG Zhe-Hao

(School of Information Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: To improve the pedestrian detection performance, this study proposes a pedestrian detection algorithm based on improved YOLOv4 by combining SqueezeNet, attention mechanism, dilated convolution, and Inception structure. An attention module named D-CBAM is proposed which is combined with dilated convolution. It is introduced to the feature enhancement part to select useful information from the extracted features. The residual connection is also used in this part to enhance feature reusability. In addition, an Inception-fire module is proposed by combining the “squeeze-expand” structure of SqueezeNet and the multi-scale convolution kernel structure of Inception, which replaces the continuous convolution layer in the network. Increasing the width of the network improves the performance of the algorithm and reduces network parameters. According to the characteristics of pedestrian detection and focal loss, the loss function is improved. The detection ability is enhanced through the addition of weights to the positive and negative samples and the hard and easy samples respectively and the strengthening of the training on positive samples and hard samples. The detection accuracy of the improved YOLO algorithm on INRIA person data set can reach 94.95%, which is 4.25% higher than that of YOLOv4. The parameters of the model are reduced by 36.35%, and the detection speed is improved by 13.54%. In short, the improved algorithm shows better performance in pedestrian detection than YOLOv4.

Key words: YOLOv4; attention mechanism; SqueezeNet; inception; ResNet; focal loss; deep learning; target detection

① 收稿时间: 2021-07-04; 修改时间: 2021-07-30; 采用时间: 2021-08-12; csa 在线出版时间: 2022-03-22

行人检测是目标检测的一个重要分支, 是识别行人目标和分析行人行为的基础, 在自动驾驶、安全监控等方面都有着广泛应用. 快速且高效的检测在行人检测任务中有着重要的意义. 随着卷积神经网络研究的深入, 基于深度学习的检测成为目标检测的主要方法, 可以分为 Two-stage 和 One-stage 两种. 在基于候选区域的 Two-stage 方法中, R-CNN[1−3] 系列算法采用卷积神经网络进行目标检测并表现出非常高的检测精度, 虽然之后的 Fast R-CNN[2]、Faster R-CNN[3] 进行了不同的改进来减少计算冗余, 但是在检测速度方面仍然不能满足实时检测. One-stage 方法中具有代表性的是 YOLO[4] 算法, 它采用单个卷积神经网络直接预测物体边界框和其所属类别的概率, 将目标检测问题转化为回归问题, 在检测速度上大幅提升, YOLO 系列算法能够同时在检测精度和速度上达到较高的性能, 能够满足行人检测的基本需求.

注意力机制是通过筛选当前任务较为关键的信息来提高网络检测能力的方法. Hu 等人提出的基于通道注意力的 SENet (squeeze-and-excitation networks)[5] 通过学习每个输入通道的权重对信息进行选择. Woo 等人提出 CBAM (convolutional block attention module)[6], 将空间和通道上的注意力结合, 能够更有效地筛选重要信息. 除了在网络深度上进行改进, 增加网络宽度也能达到提高网络性能的效果. GoogleNet[7]利用多尺度的卷积核进行特征提取, 提高了特征提取的多样性. 此外, 检测模型的大小也是一个重要的指标, SqueezeNet[8] 和 MoblieNet[9] 是具有代表性的轻量级网络. SqueezeNet 能够在网络参数量为 AlexNet 的 1/50的条件下达到与其相当的检测精度.

在相关的目标检测研究中, 李勇等人提出的结合通道注意 SENet 的 YOLOv3 算法能够有效提高网络的检测能力[10]; 方韦等人将 SqueezeNet 结构引入到 Tiny-YOLOv3 中, 将网络模型降为原网络的 1/4, 检测速度得到了提升[11]; 姜建勇等人提出的 PD-CenterNet 对样本进行加权, 通过平衡正负样本的损失提高了模型的检测能力[12].

为得到更好的检测性能, 本文在 YOLOv4 的基础上进行了研究和改进. (1) 引入一种结合空洞卷积的混合域注意力机制 D-CBAM, 并结合残差连接, 在网络的特征增强部分对有用特征进行筛选; (2) 利用 SqueezeNet 中的 squeeze-expand 思想和 Inception 中的多尺度的卷积“并连”的结构, 提出一种 Inception-fire 模块, 达到加宽网络的同时减少模型参数的效果; (3) 分别对正负样本、难易样本添加权重因子改进损失函数, 增强网络对正样本和难分类样本的训练, 以提高网络的检测能力.

1 目标检测网络

1.1 YOLOv4 目标检测

YOLOv4[13] 的网络结构分为主干网络部分、特征增强部分和预测部分, 如图 1. 主干网络采用 Mish 函数激活的 CSPDarknet53[14] 结构, 由 1×1 和 3×3 卷积构成, 利用残差连接和 CSPnet 划分通道的思想构建.主干网络进行特征提取的思想是:对于一个输入, 先对其进行通道的划分, 将分割后的一部分输入到残差块中进行运算, 提取到图像特征; 另一部分不做处理, 和残差块部分的输出进行通道上的级联, 输入到下一层中. 结合 CSP 结构的残差卷积使得 1/2 通道的特征图不参与计算, 可以将计算量减少一半左右.

YOLOv4 利用了空间金字塔池化结构 (spatial pyramid pooling, SPP) 和 PANet (path aggregation network) 对特征进行增强. SPP 将特征图用 4 个大小分别为{1, 5, 9, 13}的池化窗口进行并行的最大池化, 增加感受野, 以分离出最重要的上下文特征; PANet 结构包含了自上而下和自底向上两条路径上的特征聚合, 可以对已经提取到的 3 个尺度上的特征进行增强, 提高检测的能力[13]. 网络的预测部分仍然采用 YOLOv3[15] 的“head”结构, 分别在 3 个不同尺度上通过两层卷积运算对最终结果进行预测.

1.2 相关检测网络

(1) SqueezeNet: SqueezeNet 是一种轻量级网络, 它提出了一种 fire 结构进行特征提取. Fire 模块可以分为“squeeze”和“expand”两部分. “Squeeze”部分由 1×1 卷积实现: 来自上层的特征图先输入到一个 1×1 卷积中进行通道的压缩, 然后将其结果输入到一个由 1×1 卷积和 3×3 卷积组成的“expand”部分进一步处理. 利用两个不同卷积核对同一个特征图做卷积运算, 一方面可以提取到更丰富的特征, 另一方面, 在输入上进行通道压缩, 减少了输入到 3×3 卷积的特征图的通道数, 能够减少网络的参数[6].

(2) Inception 网络: GoogleNet 开创性的在增加网络宽度的角度上进行探索, 其主要组成部分为 Inception 结构, 该结构使用了多种不同大小的卷积核对图像进

行特征提取. 此外, Inception 借鉴了 Network-in-Network 思想, 使用 1×1 的卷积核实现降维操作来减少网络的参数量[7]. 在 Inception 结构中, 分别使用了 1×1、3×3、5×5 大小的卷积核以及一个 3×3 的最大池化层. 通过用不同核大小的卷积运算提取信息, 并将这些特征进行通道级联, 可以获得各个感受野下的特征. 由于同一层的多个卷积都对来自上一层的输入进行计算, 在输入通道数很大时会产生大量的参数, 在每次卷积运算之前引入 1×1 卷积进行通道压缩, 可以减少大卷积核产生的参数. Inception 这种网络结构参数数量仅为 AlexNet 网络的 1/12, 能够降低计算量的同时达到较高的检测精度.

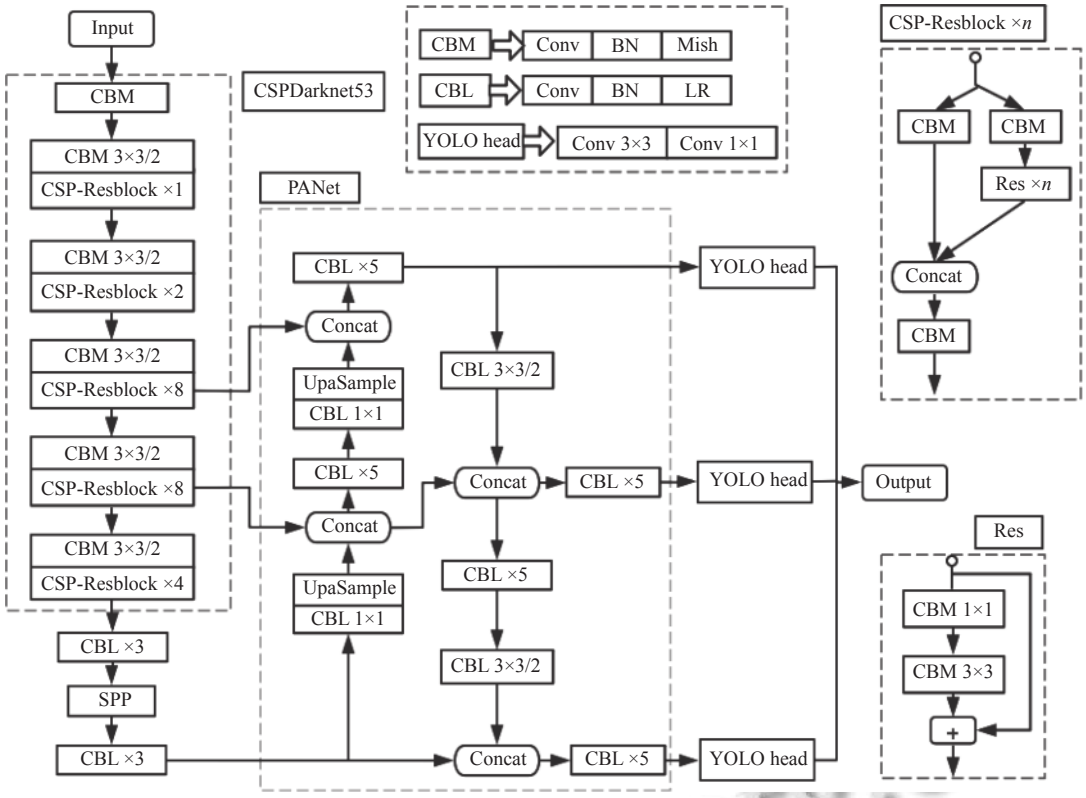


图 1 YOLOv4 网络结构

(3) 注意力机制: 注意力机制是基于人类视觉选择性注意机制提出的一种能够从众多信息中提取到有用信息的方法, 从作用域上可以分为 3 种: 空间注意机制、通道注意机制和混合注意机制. CBAM 是一种混合注意机制, 通过将通道注意和空间注意进行顺序上的连接, 实现双维度上的特征选择. 在 CBAM 的通道注意模块和空间注意模块的实现中, 都同时采用了全局平均池化和全局最大池化, 以提取到某一维度上的更全面的信息. 在目标检测网络中添加注意力机制, 能够显著增强特征中的重要信息, 对物体预测有着重要的作用[6].

2 改进算法的网络结构

2.1 网络整体结构

改进 YOLO 网络的结构如图 2 所示. 主干网络 CSPDarknet53 对输入图像进行特征提取, 分别得到 52×52, 26×26, 13×13 三个不同尺度的有效特征图, 并采用 LeakyReLU 函数激活, 相较于 Mish 激活能够一定程度上减少计算量. 在 13×13 大小的特征图输入到 SPP 模块之前, 先利用由多尺度卷积核和注意力机制构成的 Attention-I-F 模块对其进行处理, 以提取到更全面、更重要的信息. 在 SPP 之后, 同样采用 Attention-I-F 模块对输出进行处理, 对不同池化窗口得到的特征进一步添加注意力, 获取到重要信息. YOLOv4 的 PANet 将 3 个不同尺度的特征图进行自上而下和自下而上两条路径上的融合. 改进的 YOLO 中, 将融合之后的特征图输入到一个添加注意力机制的 Res-D-CBAM 模块, 从深层特征和浅层特征融合之后的信息中选择对目标检测有用的信息进行增强, 抑制那些无用的信息.

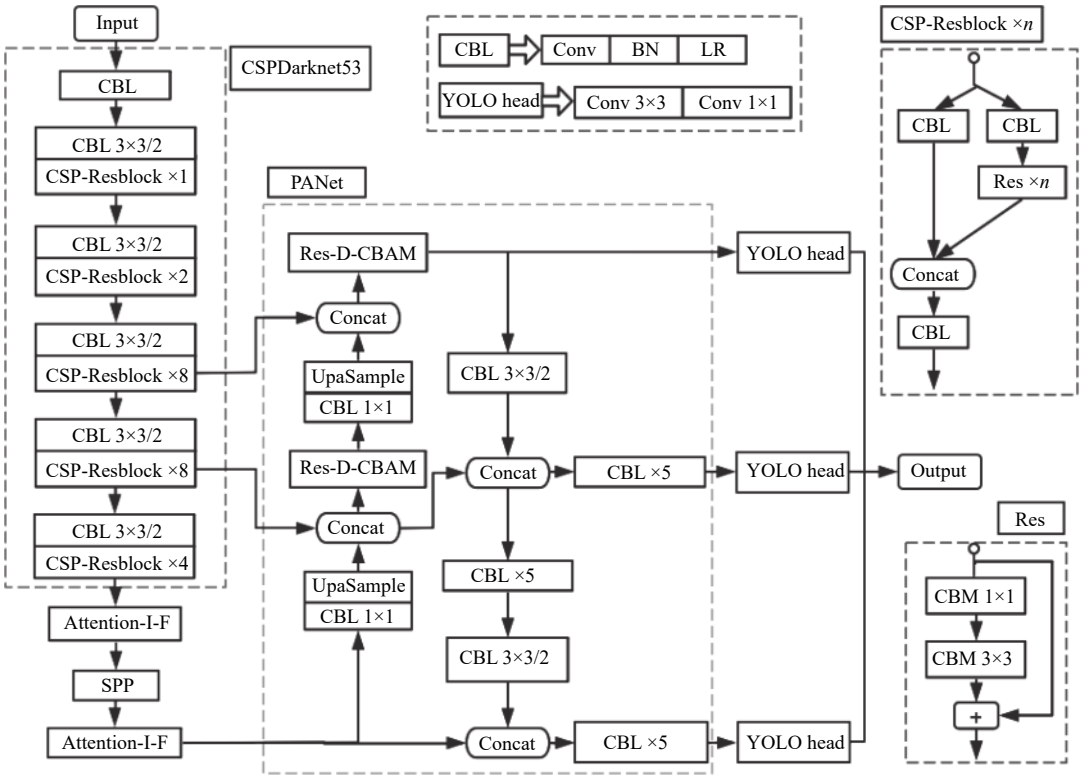


图 2 改进的 YOLO 网络结构

2.2 注意机制与残差连接的结合

在 YOLOv4 中, PANet 对提取到的不同尺寸的特征图反复融合, 并将融合后的特征进行连续的卷积处理, 其结果直接作为 YOLO-head 的输入进行预测. 这一过程虽然能够将提取到的 3 个尺度上不同的特征信息相互补充, 但此过程中也会将大量冗余信息和无用特征重复叠加. 同时, 在多通道级联后的特征图上使用大量连续卷积运算, 会产生较多的参数, 增大计算量, 降低网络的检测速度.

本文对 YOLOv4 中的 PANet 结构进行了改进, 提出一种结合残差连接、注意力机制和空洞卷积的网络模块—Res-D-CBAM, 其组成如图 3 所示.

其中“CBL $n\times n$”表示进行 $n\times n$ 卷积运算、批归一化 (batch normalization, BN) 和 LeakyReLU 激活. 将残差连接引入到连续的卷积中, 一方面可以增强特征的复用, 另一方面避免了深层网络中学习效率和准确率无法提升的问题[16]. 此外, 使用注意力机制对不同尺度上的特征信息分别进行通道域和空间域上的学习, 通过加权的方式有选择的增强本尺度特征图中对物体检测有用的信息, 并抑制不重要的信息, 能够减少后续对冗余特征的重复传递和运算.

Res-D-CBAM 模块中的注意力机制采用了如图 4 所示的 D-CBAM 结构, 此结构可分为通道注意和空间注意两个模块.

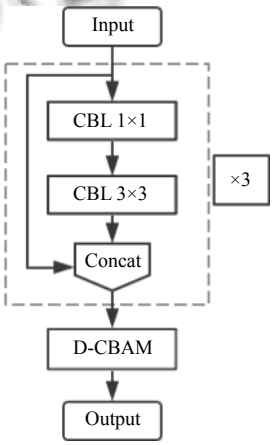


图 3 Res-D-CBAM 模块

对于一个 $H\times W\times C$ 大小的输入, 通道注意模块首先对其进行全局最大池化和平均池化, 得到两个 $1\times 1\times C$ 大小的输出, 然后利用一个共享的多层感知机

(multi-layer perceptron, MLP) 学习各通道信息的重要性, 经过 Sigmoid 函数获得 0 到 1 范围的权重$M_c(F)$, 如式 (1), 其中“+”表示对应像素点相加, σ代表 Sigmoid 激活. 最后将通道权重与输入 F 进行对应通道的加权, 即得到通道注意的结果F_1, 如式 (2), 其中⊗表示对应像素点相乘. MLP 部分采用了两层 1×1 卷积, 首先用一个输出通道数为 C/r 的 1×1 卷积获得对通道压缩后的特征图, 然后再经过一个输出通道数为 C 的 1×1 卷积, 这样两层卷积在控制卷积参数的同时能够学习到各个通道上的特征重要程度.

$$M_c(F)=\sigma(MLP(MaxPool(F))+MLP(AvgPool(F))) \tag{1}$$

$$F_1=M_c(F)\otimes F \tag{2}$$

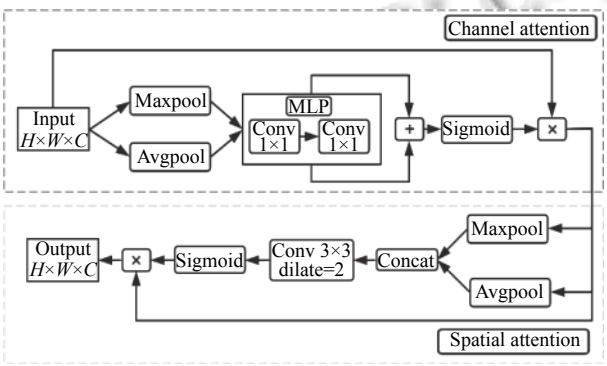


图 4 D-CBAM 结构

将通道注意的结果进一步进行空间权重的提取, 如式 (3). 首先将输入分别进行通道维度上的平均池化和最大池化, 得到两个 $H\times W\times 1$ 大小的输出, 并将二者进行通道上的级联, 得到一个 $H\times W\times 2$ 的特征图. 然后经过一个卷积核为 3×3, 膨胀系数为 2 的空洞卷积以及 Sigmoid 激活函数, 获得在空间上各点的权重$M_s(F_1)$. 最后将权重$M_s(F_1)$与输入F_1进行对应点的相乘, 得到 D-CBAM 模块的输出F_2, 如式 (4).

$$M_s(F_1)=\sigma(f^{3\times 3/2}(MaxPool(F_1),MLP(AvgPool(F_1)))) \tag{3}$$

$$F_2=M_s(F_1)\otimes F_1 \tag{4}$$

相比 CBAM 中采用的 7×7 卷积, 在 D-CBAM 中利用了扩张率为 2 的 3×3 空洞卷积来学习空间维度上特征信息的权重. 空洞卷积能够在得到较大感受野的条件下产生相对较少的卷积运算参数, 且不需要使用池化运算压缩特征图来增大感受野, 避免了分辨率降低造成的不可逆转的信息丢失[17].

2.3 多尺度卷积和 SqueezeNet 的结合

为加宽网络的同时减少连续卷积堆叠产生的参数, 本文提出一种结合 SqueezeNet 和 Inception 思想的 Inception-fire 结构, 如图 5 所示.

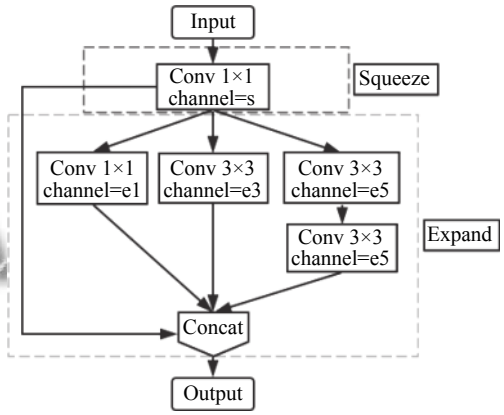


图 5 Inception-fire 模块

在此结构中, 首先对输入进行“squeeze”运算, 将通道数压缩为 s, 这一操作通过一个 1×1 卷积完成. 然后将 1×1 卷积的输出进行“expand”运算, 即将其分别输入到 1×1, 3×3, 5×5 三个尺寸的卷积中, 并将输出通道数分别设置为: e_1、e_3 和 e_5. 通过设置相应卷积的输出通道数, 可以控制网络产生的计算量. 借鉴残差网络的思想, 将“expand”层的 3 个卷积的输出和“squeeze”层的输出进行通道维度上的相加, 以增强特征的复用, “Concat”的结果作为整个模块的输出. 由于较大的卷积核在计算过程中会产生比较多的参数, 在本模块中, 将 5×5 卷积分解为两个连续的 3×3 卷积, 可以在减少参数量的同时保证卷积运算能够获得相同感受野.

将 Inception-fire 结构和图 4 的注意力模块 D-CBAM 结合, 得到增强注意的 Inception-fire 结构—Attention-I-F, 如图 6 所示.

本文提出的改进 YOLO 网络在主干网络和空间金字塔池化结构之后都分别添加了 Attention-I-F 模块, 用于进一步处理提取到的深层特征. 整体实现流程是: 将输入特征分别进行 3 层 Inception-fire 网络的特征提取, 然后利用一个 1×1 卷积进行通道整合, 最后结合注意力模块 D-CBAM 进行有用特征的选择. 其中 Inception-fire 通过多个大小不同的卷积核对深层特征处理, 能够减少过拟合. 同时, 通过增加卷积运算的多样性可以获

得不同感受野下的语义信息. 将注意力机制运用在上层提取到的丰富的特征信息上, 可以对其中重要的特征信息进行增强, 减少后续特征融合过程中深层语义信息的丢失, 以提高网络的预测能力.

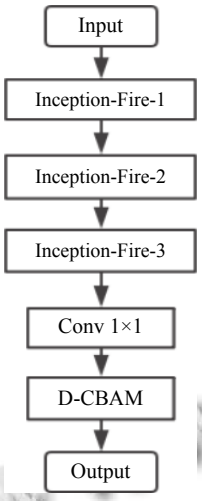


图 6 Attention-I-F 模块

3 损失函数

3.1 Focal loss

YOLOv4 中为定位物体会首先生成大量的先验框—anchor box, 但在实际的行人检测中, 多数情况下一幅图像中仅存在少量的目标, 因此会有大量的 anchor box 产生在背景区域. YOLO 算法会直接对这些正负样本不均匀的 anchor box 进行分类, 且后续使用交叉熵计算分类损失以及置信度损失, 如式 (5)、式 (6), 其中 p 表示预测类别的概率. 这一过程对所有类别无区别的对待, 忽略了正负样本不平衡问题. 为解决这一问题, 平衡交叉熵损失在每个类别前增加了一个权重因子 α_t 来协调类别不平衡, 如式 (7).

$$CE = -\log(p_t) \tag{5}$$

$$p_t = \begin{cases} 1-p, & p < 0 \\ p, & p \geqslant 0 \end{cases} \tag{6}$$

$$Balance\_CE = -\alpha_t \log(p_t) \tag{7}$$

样本中除了正负样本之外, 还存在易分类样本和难分类样本. 为了提高网络的检测能力, 在训练中应该对难分类的样本着重考虑, 而平衡交叉熵损失中仅增加一个权重因子平衡正负样本, 并没有考虑难易样本的区分. 为解决此问题, 焦点损失函数 (focal loss, FL) 在平衡交叉熵损失的基础上增加了一个调节因子 $(1-p_t)^{\gamma}$, 用来降低易分类样本权重, 聚焦于难分类样本的训练. FL 的表示如式 (8), 其中 γ 为聚焦参数, 可以调节权重 $(1-p_t)^{\gamma}$ 的降低程度, γ 越大则权重降低的程度就越大. 当 p_t 很小即表示难分类的样本, 此时调节因子 $(1-p_t)^{\gamma}$ 趋近 1, 损失函数中样本的权重不受影响; 当 p_t 很大时即表示样本较容易分类, 这时调节因子趋近 0, 该样本在损失函数中的权重下降很多, 以此达到增强对难分类样本的训练[18].

$$FL_{loss} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{8}$$

3.2 引入 focal loss 的损失函数

结合焦点损失函数和行人检测的特点, 本文对损失函数进行了改进. 改进的损失函数可分为两部分: 回归损失和置信度损失, 如式 (9).

$$loss = l_{\text{loc}} + l_{\text{conf}} \tag{9}$$

$$l_{\text{loc}} = \sum_{i=0}^{S\times S}\sum_{j=0}^{N} I_{ij}^{\text{obj}} l_{CIOU} \tag{10}$$

$$l_{CIOU} = 1 - IOU + \frac{\rho^2(b,b^{\text{gt}})}{c^2} + \alpha v \tag{11}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{\text{gt}}}{h^{\text{gt}}} - \arctan\frac{w}{h}\right)^2 \tag{12}$$

$$\alpha = \frac{v}{(1-IOU)+v} \tag{13}$$

其中, l_{loc} 为预测框的回归损失, 由 $CIOU$ 计算损失值, 如式 (10). $S\times S$ 表示将输入图像划分成 $S\times S$ 的网格, N 表示每个网格上产生的先验框个数. 若待检测物体的中心落在第 i 个网格的第 j 个先验框中, 则 I_{ij}^{obj}=1, I_{ij}^{noobj}=0, 否则 I_{ij}^{obj}=0, I_{ij}^{noobj}=1. C_i 和 $\hat{C}_i$ 分别表示预测得到的置信度和实际的置信度值. l_{CIOU} 为 $CIOU$ 损失, 如式 (11), 12, 13. c 为真实框和预测框的最小闭包的对角线距离, $\rho^2(b,b^{\text{gt}})$ 表示预测框和真实框中心点之间的欧氏距离, w^{gt} 和 h^{gt} 分别为真实框的宽和高, w 和 h 分别为预测框的宽和高. v 用于衡量长宽比的一致性, α 是用于权衡的参数.

l_{conf} 为预测结果的置信度损失, 如式 (14). 改进的置信度损失添加了控制因子 β, 用于控制正负样本在损失中所占比重, 减小大量负样本的损失对总的损失值的主导作用, 使得网络反向传播时聚焦于正样本的训练. 此外, 引入调节因子 $(1-C_i)^{\gamma}$ 用于调节难分类样本

和易分类样本的权重, 增加难分类样本的权重, 使得总的损失值更偏向于难分类样本的损失, 便于在反向传播中进一步对难分类样本进行训练, 增强网络对难分类样本的分类能力. 参考文献 [17] 并结合实验对比, 在实验中设置β= 0.25, γ= 1.8 可以较好的平衡检测精度和误检率.

$$\begin{aligned} l_{\text{conf}} = &-\sum_{i=0}^{S\times S}\sum_{j=0}^{N} I_{ij}^{\text{obj}}[(\beta(1-C_i)^{\gamma}\hat{C}_i\log(C_i) \\ &+(1-\beta)C_i^{\gamma}(1-\hat{C}_i)\log(1-C_i)] \\ &-\sum_{i=0}^{S\times S}\sum_{j=0}^{N} I_{ij}^{\text{noobj}}[(\beta(1-C_i)^{\gamma}\hat{C}_i\log(C_i) \\ &+(1-\beta)C_i^{\gamma}(1-\hat{C}_i)\log(1-C_i)] \end{aligned} \tag{14}$$

4 实验与结果分析

4.1 数据集处理和网络训练

为验证本文所提出的改进 YOLO 的性能, 在 INRIA 行人数据集上进行了训练和测试. INRIA 是目前使用最多的静态行人检测数据集, 图像中的行人姿态和光照条件等丰富多变, 存在单个行人以及拥挤遮挡人群的情况, 适合用于行人检测. 训练集中有 614 张图像, 测试集 288 张图像. 为避免训练过程中出现过拟合, 采用随机添加噪声、调整亮度、旋转、裁剪、平移以及 cutout 等方法对训练集图像进行数据增强, 将训练集扩充到了 3 070 张图像.

训练前在 INRIA 数据集上采用 k 均值聚类选择合适的先验框个数和尺寸大小. 根据聚类个数和平均 IOU 的曲线, 实验中选择对数据集进行聚类大小为 9 的先验框聚类, 得到的 9 个先验框的尺寸如表 1 所示.

表 1 先验框大小

编号	1	2	3	4	5	6	7	8	9
宽度	19	26	35	39	51	60	81	88	137
高度	68	100	123	159	186	244	310	205	287

使用 Python 语言在 PyTorch 框架下实现算法. 将训练集中 80% 的图像用于训练, 20% 用于验证. 采用 Adam 优化器, 权重衰减设置为 0.000 5. 学习率采用余弦退火衰减, 周期 T=5. 初始学习率设为 0.001, 最小学习率 0.000 01, batch size=8, 进行 3 000 次迭代后, 改用初始学习率 0.000 1, 最小学习率 0.000 01 的余弦退火衰减继续训练.

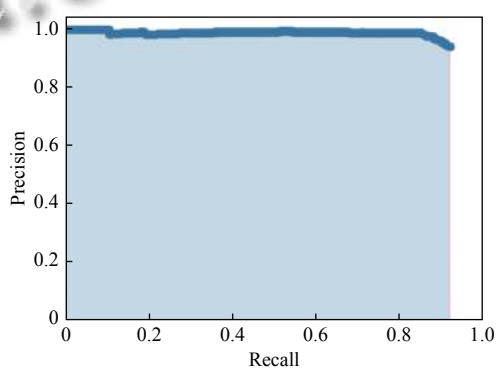
训练结束后获得改进网络权重参数, 将其和原网络的参数模型大小对比, 如表 2. 改进后的网络参数从 6.5 千万减少到了 4 千万, 降低了 37.34%.

表 2 参数量对比

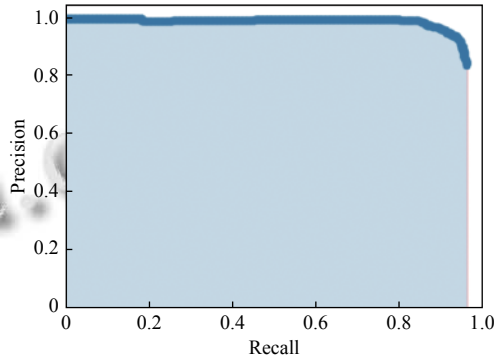
模型	Model size (MB)	参数数量 (万)
YOLOv4	244.29	6 504
本文的改进YOLO	155.48	4 075

4.2 测试结果

将训练后的网络在 INRIA 测试集上进行测试, 获得了如图 7 所示的召回率-精确度曲线, 其中曲线下围成的面积即为平均检测精度 (AP).



(a) Class: 90.70%=person AP YOLOv4



(b) Class: 94.95%=person AP 改进算法

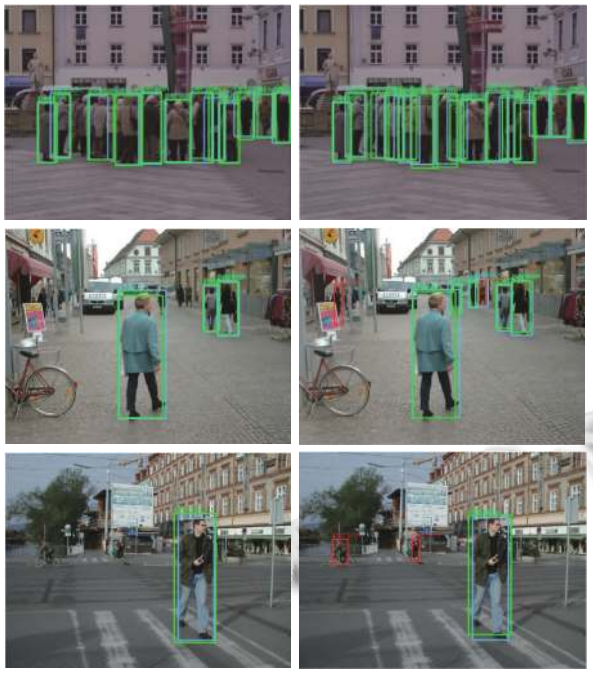
图 7 检测召回率-精度曲线对比

对改进前后的网络性能进行了对比, 如表 3. 改进的 YOLO 算法能够达到 94.95% 的平均检测精度, 比原 YOLOv4 算法高出 4.25%. 对比其检测时间, 可以发现改进的 YOLO 算法在检测速度上提高了 13.54%.

表 3 性能对比

模型	AP (%)	时间 (s)
YOLOv4	90.70	22.67
本文的改进YOLO	94.95	19.60

为了更直观的发现改进网络检测能力的提升, 实验进一步获取了改进前后的测试结果, 如图 8 所示.



(a) YOLOv4 检测结果 (b) 改进 YOLO 检测结果

图 8 检测结果对比

检测结果中蓝色框表示数据集中标注的 ground-truth, 绿色框表示正确的检测结果, 红色框代表假正例, 即网络预测结果没有匹配的 ground-truth, 判定为误检. 对比 YOLOv4 和改进算法的检测结果, 改进后的网络对于图像中尺寸较小的行人目标的检测能力具有明显的提高. 对于图像中人群出现大量遮挡的情况, 使用 YOLOv4 会产生部分的漏检, 而改进的网络能够检测出更多的存在部分遮挡的目标. 虽然改进网络中存在一些判定为假正例的检测结果, 但可以发现在实际图像中, 该检测是正确的, 这种误检的原因是数据集中标注的不完全.

4.3 不同改进策略的实验对比

为了验证文中所提出的改进策略对网络检测性能的影响, 本文开展了一系列的对比实验. 在原 YOLOv4 网络的基础上, 分别进行改进. (1) 将主干网络激活函数修改为 LeakyReLU 函数; (2) 在 (1) 的基础上, 将 Res-D-CBAM 模块添加到改进的网络中, 如图 2 中的 PANet 部分; (3) 在 (2) 的基础上将 Attention-I-F 模块引入到改进网络中, 如图 2 中的“Attention-I-F”; (4) 在 (3) 的改进基础上, 使用第 3.2 节提出的损失函数计算训练过程中的损失值. 对以上几个改进的策略在同样的环境下分别进行训练和测试, 得到如表 4 所示的结果.

表 4 不同改进策略的实验结果对比

算法	YOLOv4				改进YOLO
激活函数	Mish	LR	LR	LR	LR
Res-D-CBAM	—	—	√	√	√
Attention-I-F	—	—	—	√	√
损失函数	BCE	BCE	BCE	BCE	FL
时间 (s)	22.67	22.48	22.81	19.60	**19.60**
AP (%)	90.70	90.60	**91.63**	**93.87**	**94.95**

根据表 4 中数据的对比, 可以发现: 主干网络采用 LakyReLU 激活函数可以减少部分的计算量, 使网络总的检测时间有所减少, 且检测精度变化不大; 在此基础上, 在网络中添加 Res-D-CBAM 模块之后, 得到的检测精度增加到了 91.63%, 相较添加之前提高了 1.03%, 同时检测耗时仅增加 1.47%, 证明了注意力机制和残差连接结构对于特征信息的选择性增强具有较好的效果. 结合以上改进, 进一步将 Attention-I-F 模块添加到改进网络中, 从检测结果中可以发现, Attention-I-F 结构的引入使得网络的检测精度 91.63% 提高到了 93.87%, 提升了 2.24%, 同时检测时间从 22.81 s 减少到了 19.60 s, 降低了 14%. 此结果表明在网络中使用多尺度卷积核提取到的更丰富的特征对检测结果有重要作用, 且利用 1×1 卷积减小特征图的厚度, 能够对检测速度的提高产生较大的影响. 最后, 使用改进的损失函数对网络进行训练, 可以得到检测精度上 1.08% 的提升. 这一结果表明通过对正负样本、难易样本的损失分别进行加权, 提高损失函数中正样本和难分类样本的权重, 可以在网络反向传播时增强正样本和难分类样本的训练, 使得最终训练得到的网络模型能够有更高的检测能力.

5 结束语

本文提出了一种结合 SqueezeNet、Inception 结构、残差连接以及注意力机制的网络结构, 并基于 YOLOv4 的损失函数, 结合 focal loss 改进了行人检测中的损失函数, 使得网络整体得到了 4.25% 的精度提升. 结合 SqueezeNet 的“squeeze-expand”思想以及 Inception 中多尺度卷积核的运用, 提出的 Inception-fire 结构能够很大程度上同时提高检测的速度和精确度. 改进的注意力模块结合残差连接能够在增加很少

量参数的条件下有效提高网络检测精度. 在损失函数中通过调整正负样本、难易样本的权重, 可以在一定程度上提高网络的检测能力. 通过观察改进网络的检测结果, 可以发现改进网络对于密集和遮挡的人群具有较好的检测效果, 对图像中较远处小目标的检测能力也有所提升. 但由于现有的行人数据集对较小或不明显的目标物体标注的不完全, 测试时会将检测结果中的部分正确的检测判定为假正例. 在接下来的研究中, 会改进数据集问题, 降低误检率, 同时研究如何进一步提高网络的检测速度.

参考文献

1 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

2 Girshick R. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1440–1448.

3 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

4 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]

5 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

6 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 3–19.

7 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2014. 1–9. [doi: 10.1109/CVPR.2015.7298594]

8 Iandola FN, Moskewicz MW, Ashraf K, *et al*. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1 MB model size. arXiv: 1602.07360, 2016.

9 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861, 2017.

10 Li Y, Lv C. SS-YOLO: An object detection algorithm based on YOLOv3 and ShuffleNet. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Chongqing: IEEE, 2020. 769–772. [doi: 10.1109/ITNEC48623.2020.9085091]

11 Fang W, Wang L, Ren PM. Tinier-YOLO: A real-time object detection method for constrained environments. IEEE Access, 2020, 8: 1935–1944. [doi: 10.1109/ACCESS.2019.2961959]

12 姜建勇, 吴云, 龙慧云, 等. 基于 Center Net 的实时行人检测模型. 计算机工程, 2021, 47(10): 276–282. [doi: 10.19678/j.issn.1000-3428.0059043]

13 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv: 2004.10934, 2020. 1–17.

14 Wang CY, Liao HYM, Wu YH, *et al*. CSPNet: A new backbone that can enhance learning capability of CNN. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle: IEEE, 2020. 1571–1580. [doi: 10.1109/CVPRW50498.2020.00203]

15 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

16 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778.

17 宦海, 陈逸飞, 张琳, 等. 一种改进的 BR-YOLOv3 目标检测网络. 计算机工程, 2021, 47(10): 186–193. [doi: 10.19678/j.issn.1000-3428.0059234]

18 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318–327. [doi: 10.1109/TPAMI.2018.2858826]