

# 密集交通场景的目标检测算法研究

李 轩,李 静,王海燕

(沈阳航空航天大学 电子信息工程学院,辽宁 沈阳 110136)

**摘 要:**在现实的交通场景中,行人和车辆经常聚集在一起,形成相互遮挡的现象,给交通场景的目标检测带来了极大的挑战。针对交通场景中目标密集,位置接近造成的目标漏检、同一检测框中包含多个目标的问题,提出一种针对性遮挡回归损失函数 Occlusion Loss。Occlusion Loss 有两个作用:一是指导神经网络学习检测框和真实框匹配程度得到更为准确的位置信息;二是在学习到位置信息后尽可能减少一个检测框有多个被检测目标的情况。将提出的 Occlusion Loss 应用到 YOLOv3 目标检测算法上,经过实验证明改进后的 YOLOv3 在密集的交通场景中有更准确的检测结果,能够有效防止目标漏检现象,定位更加准确,具有很强的鲁棒性。在重新划分的交通场景数据集 KITTI 中准确率和召回率均有所提高,平均准确率达到 92.67%,优于其他目标检测算法。

**关键词:**目标密集;回归损失函数;匹配程度;位置信息;YOLOv3;目标检测

中图分类号:TP391.4

文献标识码:A

文章编号:1673-629X(2020)07-0046-05

doi:10.3969/j.issn.1673-629X.2020.07.011

## Research on Object Detection Algorithm in Dense Traffic Scenes

LI Xuan, LI Jing, WANG Hai-yan

(School of Electronic Information Engineering, Shenyang Aerospace University, Shenyang 110136, China)

**Abstract:** In the actual traffic scene, pedestrians and vehicles often gather together to form a mutual occlusion phenomenon, which brings great challenges to the object detection in traffic scenes. Aiming at the problem that the object is dense in the traffic scene, the object is missed, and the same detection frame contains multiple objects, we propose a targeted occlusion regression loss function Occlusion Loss which has two functions. The first is to guide the neural network to learn the matching degree of detection box and the groundtruths for more accurate position information. The second is to reduce the number of detected objects in one detection box as much as possible after learning the position information. The proposed occlusion loss is applied to the YOLOv3 object detection algorithm. It is proved by experiments that the improved YOLOv3 has more accurate detection results in dense traffic scenarios, which can effectively prevent object miss detection, with more accurate positioning and stronger robustness. In the re-divided traffic scene dataset KITTI, the accuracy and recall rate are improved, and the average accuracy rate is 92.67%, which is better than other object detection algorithms.

**Key words:** dense object; regression loss function; matching degree; position information; YOLOv3; object detection

## 0 引言

自深度学习发展以来,目标检测领域已经取得了很大进展<sup>[1-4]</sup>,基于区域建议和回归思想的目标检测算法相继提出,在准确率和检测速度上都有较大的提升。但密集场景的目标检测一直是计算机视觉领域中的难点问题。在现实的交通场景中由于人为因素造成的目标遮挡阻塞情况十分常见。造成密集场景检测困难的原因分为两种,一种是由同类目标相互遮挡造成的类内遮挡,另一种是由不同类之间相互遮挡造成的类间遮挡。在常用的交通场景数据集中类内遮挡较为常见,在 City Persons 数据集<sup>[5]</sup>中类内遮挡率达

48.8%,在 KITTI 数据集中遮挡和截断率也接近 50%,因此在检测过程中极容易出现漏检、重复检测等情况。

近年来在密集目标检测领域中已经取得了一些成果。例如,Wang 等<sup>[6]</sup>提出了一种排斥力损失函数,在有效吸引真值框和预测框相互接近的同时排斥由其他物体真值的影响,但在现实操作过程中很难掌握吸引和排斥之间的平衡。文献[7]从正负样本量入手,通过 Focal Loss 重新定义交叉损失熵,改变难分样本占总 loss 的权重,提高对负样本的判断能力,改进了 one-stage 算法对密集目标的检测能力。文献[8]使用

收稿日期:2019-07-30

修回日期:2019-11-29

基金项目:辽宁省教育科学技术研究项目(社会服务类)(L201715)

作者简介:李 轩(1967-),男,博士,研究方向为图像处理;通讯作者:李 静(1994-),女,研究生,研究方向为计算机视觉。

Jaccard index 作为评估检测质量分数,重新定义了 IoU Loss,提出新型的 EM-Merge 单元,致力于解决检测框重叠歧义问题。尽管在密集场景中的检测已经取得了一些进展,但如何在复杂的遮挡环境下准确地对目标进行定位仍然是交通场景检测中的难点。

## 1 基本原理

文中在实验过程中发现由于人为因素造成的目标密集遮挡问题,使面积相对较大,包含特征因素更多的检测框往往有更高的置信度,这容易在非极大值抑制时使真正更为准确的检测框被抑制掉造成目标的漏检。如图 1 所示,在检测过程中,行人 A(左)和行人 B(右)之间位置接近,在判断特征得分时极容易出现特征的混淆和重叠,因此图中实线预测框相比虚线预测框有更高的置信度,在后置的非极大值抑制中导致定位更为准确的预测框被当作同一目标的冗余框而被抑制掉。造成这种目标错检可能存在两种情况,一是特征提取不够完善,二是学习过程没有针对性。解决第一种情况需要学习更强壮的特征信息,扩增数据集训练样本,文中采用 mixup<sup>[9]</sup>算法重构数据集,将任意两张图像以一定比例融合,构成新的样本集,神经网络学习的样本数量成倍增多;针对第二种情况,文中提出一种新的损失函数 Occlusion Loss, Occlusion Loss 中包含两项内容即 IoG Loss 和 UoG Loss。IoG Loss 用于衡量图 1 中虚线检测框和阴影之间的相似度, UoG Loss 用于降低实线检测框对整体检测的影响。

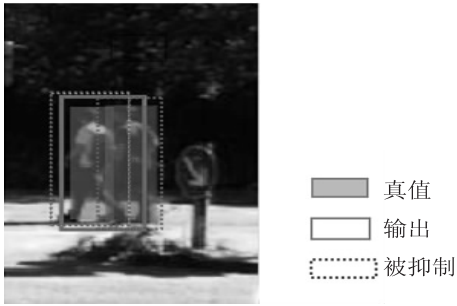


图 1 相互遮挡造成的错误检测情况

## 2 相关工作

### 2.1 Occlusion Loss 损失函数

假设预测框的集合表示为  $P = \{x_1^p, y_1^p, x_2^p, y_2^p\}$ , 真值框为  $G = \{x_1^g, y_1^g, x_2^g, y_2^g\}$ 。  $(x_1, y_1)$  表示目标框的左上角坐标,  $(x_2, y_2)$  表示右下角坐标。通常使用 IoU 衡量真实框  $G$  和预测框  $P$  之间的相关度  $\text{IoU}(p_i, g_i) = \frac{p_i \cap g_i}{p_i \cup g_i}$ 。文献[10]将  $-\ln(\text{IoU})$  作为损失函数惩罚定位不准确的检测框。但当  $-\ln(\text{IoU})$  定义为损失函数时,那么只要预测框足够小,在极限情况下 IoU Loss

可以降低,这并不利于对目标的定位。在精确目标定位的函数必须满足以下两个条件:

(1)  $p \cap g$  尽可能大,保证检测框和真值框尽可能匹配;

(2) 在满足条件 1 的情况下  $p$  尽可能小,防止出现一个检测框对多个目标的情况。

基于以上两点考虑,文中使用 IoG 和 UoG 作为损失函数,而不是传统的 IoU。IoG 定义为:

$$\text{IoG}(p_i, g_i) = \frac{p_i \cap g_i}{g_i} \quad (1)$$

在 IoG 的定义中,用真值框  $g$  取代了  $p \cup g$ ,  $g$  是待检测样本的自然属性,在整个检测过程中保持不变,希望神经网络能够通过增大  $p \cap g$  的值不断拟合预测框和真值框的相似程度,并且在这个学习过程中损失函数能不断减小,因此定义  $L_{\text{IoG}}$  损失函数为:

$$L_{\text{IoG}}(p, g) = \sum_{p_i \in |P|, g_i \in |G|} \exp\left(-\frac{p_i \cap g_i}{g_i}\right) \quad (2)$$

在 IoG 函数中已经能够匹配位置相对准确的框, UoG 函数用来抑制在同等匹配情况下检测面积更大的预测框, UoG 定义如下:

$$\text{UoG}(p_i, g_i) = \frac{p_i \cup g_i}{g_i} \quad (3)$$

在  $L_{\text{UoG}}$  定义中,同样要求真值框保持不变,损失函数通过不断学习降低  $p \cup g$  的值得到定位更为准确的预测框。为了防止由于预测框过大,使整体损失函数花更多的精力在  $L_{\text{UoG}}$  上,而忽略了学习其他特征如位置回归或置信度损失值的影响,利用 sigmoid 函数 ( $\text{sigmoid}(\text{UoG}) = \frac{1}{\exp(-\text{UoG})}$ ) 对  $L_{\text{UoG}}$  进行归一化。

sigmoid 函数对中央区的信号增益较大,对两侧区的信号增益较小,在信号的特征空间映射上,符合  $L_{\text{UoG}}$  的基本要求。并且 sigmoid 函数在定义区间内可导,利于反向传播。最终 UoG 的损失函数定义为:

$$L_{\text{UoG}}(p, g) = \sum_{p_i \in |P|, g_i \in |G|} \text{sigmoid}\left(\frac{p_i \cup g_i}{g_i}\right) \quad (4)$$

$L_{\text{UoG}}$ 、 $L_{\text{IoG}}$  算法流程如下:

Algorithm: IoG, UoG 回归框损失

输入: 预测框  $P = \{x_1^p, y_1^p, x_2^p, y_2^p\}$  和真值框  $G = \{x_1^g, y_1^g, x_2^g, y_2^g\}$ , 输出:  $L_{\text{IoG}}$ ,  $L_{\text{UoG}}$

对于预测框  $P$ , 确定  $x_2^p > x_1^p, y_2^p > y_1^p$ :

$\hat{x}_1^p = \min(x_1^p, x_2^p), \hat{x}_2^p = \max(x_1^p, x_2^p)$

$\hat{y}_1^p = \min(y_1^p, y_2^p), \hat{y}_2^p = \max(y_1^p, y_2^p)$

计算  $P$  面积

$A^p = (\hat{x}_2^p - \hat{x}_1^p) \times (\hat{y}_2^p - \hat{y}_1^p)$

计算  $G$  面积

$$A^g = (x_2^g - x_1^g) \times (y_2^g - y_1^g)$$

计算  $P, G$  相交区域面积

$$x_1^r = \max(x_1^p, x_1^g), x_2^r = \min(x_2^p, x_2^g)$$

$$y_1^r = \max(y_1^p, y_1^g), y_2^r = \min(y_2^p, y_2^g)$$

$$\tau = \begin{cases} (x_2^r - x_1^r) \times (y_2^r - y_1^r) & x_2^r > x_1^r, y_2^r > y_1^r \\ 0 & \text{其他} \end{cases}$$

计算

$$\text{IoG} = \frac{\tau}{A^g}$$

$$\text{UoG} = \frac{A^p + A^g - \tau}{A^g}$$

$$L_{\text{IoG}} = \exp(-\text{IoG}), L_{\text{UoG}} = \text{sigmoid}(\text{UoG})$$

## 2.2 YOLOv3 检测框架

文中采用 YOLOv3 算法作为主体框架进行实验。YOLOv3 网络主体为 Darknet53 残差网络,并结合多尺度检测,已经成为目前为止最优秀的目标检测算法之一,常用在交通、工业等多种场景中<sup>[11-14]</sup>。在检测过程中首先将图片划分为  $S \times S$  个小格,当目标中心落在某个小格时该小格负责预测这个物体。YOLOv3 借鉴了类似 FPN 网络<sup>[15]</sup>的金字塔结构对特征图进行了上采样和融合做法,分别在  $13 \times 13, 26 \times 26, 52 \times 52$  三个尺寸上进行检测。每个单元格首先借助 anchor box 预测 3 个检测框,每个检测框预测 4 个相对坐标值即中心坐标  $(x, y)$  与目标宽  $w$  和高  $h$ ,再根据卷积神经网络输出的坐标对 anchor box 进行修正,修正过程公式如下:

$$b_x = \sigma(t_x) + c_x \quad (5)$$

$$b_y = \sigma(t_y) + c_y \quad (6)$$

$$b_w = p_w e^{t_w} \quad (7)$$

$$b_h = p_h e^{t_h} \quad (8)$$

其中,  $c_x, c_y$  为目标中心所在小格相对于左上角的偏移量,  $t_x, t_y, t_w, t_h$  为网络学习输出,  $p_w, p_h$  为预设锚点的宽高,坐标的损失函数采用平方误差损失函数。

在类别预测方面,YOLO 原有的 softmax 层假设一张图片或者一个待检测物体都只属于一个类别,但这并不适用于复杂的场景,在复杂的场景中一个待检测的物体可能分属于不同的类别,存在多个标签,因此 YOLOv3 采用逻辑回归层进行类别预测。在逻辑回归层加入 sigmoid 函数对输出类别概率值进行限制,概率大于 0.5,就表示属于该类。类别公式计算如下:

$$\text{Pr}(\text{object}) * \text{IoU}(b, \text{object}) = s(t_0) \quad (9)$$

在损失函数上,YOLOv3 整合了均方误差调整目标相对单元格的宽高,二值交叉损失熵调整置信度得分,类别信息和中心坐标可抽象表达为:

$$\text{loss} = \sum \text{coord\_Err} + \text{conf\_Err} + \text{class\_Err} \quad (10)$$

YOLOv3 并没有将检测匹配度作为学习的模块,十分适合文中的算法移植。

文中算法检测流程如图 2 所示,首先经过 DarkNet 53 进行特征提取,利用多尺度特征的原理在  $13 \times 13, 26 \times 26, 52 \times 52$  三个尺度的特征图上分别对同一目标进行不同尺度的预测。在训练过程中加入了  $L_{\text{IoG}}$  指导神经网络学习更准确的定位,  $L_{\text{UoG}}$  防止检测框中多个特征的情况,得到最终输出结果。

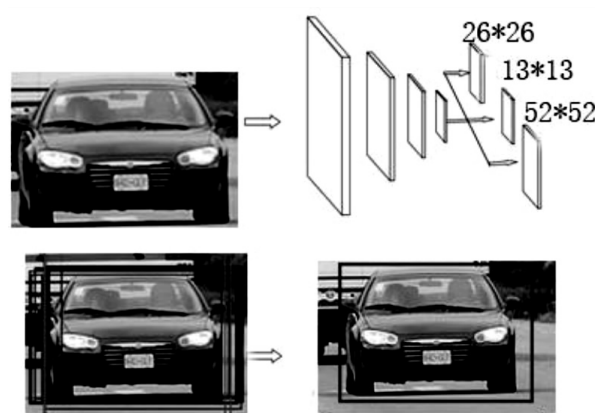


图 2 YOLOv3 算法检测过程

## 2.3 数据集

文中算法在自动驾驶数据集 KITTI 上进行测试。KITTI 数据集是目前交通场景中公认的大规模算法测试数据集,数据量达 12 G。存在行人、车辆的互相遮挡截断情况,反映了现实中复杂的交通场景。文中定义两个目标重合度 IoU 在 0.5 以上为遮挡阻塞情况,车辆阻塞和截断情况占总数的 40% 以上,行人间的遮挡占总数的 30% 左右。

文中检测主要针对类间遮挡情况,将 KITTI 原数据集重新整理,‘Van’, ‘Truck’, ‘Tram’ 标签都重新标定为 ‘Car’ 类, ‘Person\_sitting’ 标定到 ‘Pedestrian’ 类中,最终检测类别为 ‘Car’, ‘Cyclist’, ‘Pedestrian’。另外将 KITTI 训练集重新划分得到新的训练集 4 000 张,验证集 400 张,测试集 3 081 张,评价指标基于重新划分数据集的结果。表 1 展示了经过处理后的数据集各类别样本的数量情况。

表 1 KITTI 数据集样本数量统计

集合	Car	Van	Truck	Tram	Pedestrian	Cyclist	Car(合并)
Train	15 209	1 567	575	283	12 855	1 010	17 634
Test	13 533	1 347	519	228	11 850	611	15 627

### 3 分析与讨论

文中算法在开源框架 Keras 上实现,电脑配置为 Intel(R) Core(TM) i3-4170CPU@3.70 GHz,运行内存 8 G,显卡 1050Ti,操作系统为 Windows 10。

#### 3.1 训练方法

训练阶段参数设定动量为 0.9,权重衰减率为 0.005,初始学习率为 0.001,采用 Adam 优化方法。模型在 100 K 左右收敛,达到最优值 Loss=0.254 8,相比

改进之前有更强的特征表达能力。

#### 3.2 实验结果

将改进后的算法与原 YOLOv3 进行对比,对比结果如表 2 所示。在平均准确率上,文中算法较原算法提升了 2.12%,在‘Pedestrian’类别上提升明显,可见由于行人体积小,并且容易成群结队是交通场景检测困难的主要因素。

表 2 改进算法和原算法性能对比

方法	输入	MAP/%	Recall	Car	Pedestrian	Cyclist
YOLOv3	416×416	90.49	95.36	93.93	85.49	92.07
文中方法	416×416	92.67	95.53	95.70	88.43	93.88

图 3 展示了文中算法的部分检测结果,可以看出文中算法对于不同场景的车辆和行人的密集场景检测

表现良好,能够更加准确地定位,有效减少了漏检情况,具有很好的鲁棒性。



图 3 原 YOLOv3(左)与改进后的 YOLOv3(右)检测结果对比

图 4 显示了文中算法和原 YOLOv3 算法在不同类别上的 PR 曲线,通过比较曲线下的面积可以看出,在提出的损失函数指导学习后的 YOLOv3 在‘Car’,

‘Cyclist’, ‘Pedestrian’三个类别上均获得了优于原算法的性能。

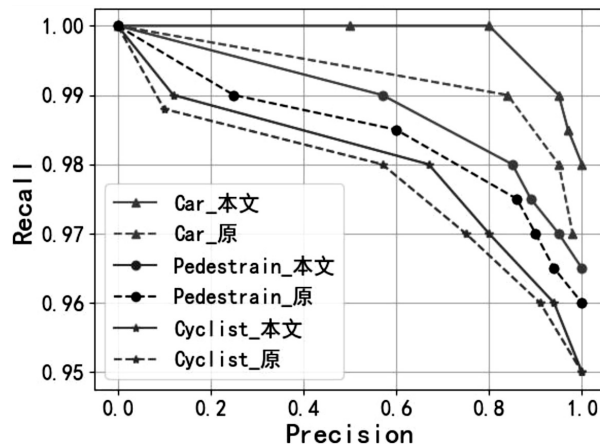


图 4 不同类别的 PR 曲线

#### 4 结束语

文中提出了针对密集场景检测的 Occlusion Loss。Occlusion Loss 包含两项内容:第一项是负责更准确定位的 IoG Loss;第二项是负责调整一个检测框对应多个目标的 UoG Loss。将该损失函数移植到 YOLOv3 网络中,在 KITTI 数据集上获得了更好的表现,不仅能准确匹配到目标位置,而且有效抑制了目标漏检,在准确率和召回率上都有更好的表现。在后续工作中,将该算法移植到多种框架中,以实现密集交通场景的车辆多类别识别任务。

#### 参考文献:

- [1] DAI J, LI Y, HE K, et al. R-FCN: object detection via region-based fully convolutional networks[J]. Advances in Neural Information Processing Systems, 2016, 29(14): 379-387.
- [2] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.
- [3] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 42(2): 386-397.
- [4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//IEEE conference on computer vision and pattern recognition. Las Vegas; IEEE, 2016: 779-788.
- [5] ZHANG S, BENENSON R, SCHIELE B. Citypersons: a diverse dataset for pedestrian detection[C]//IEEE conference on computer vision and pattern recognition (CVPR). Honolulu, HI, USA; IEEE, 2017: 3213-3221.
- [6] WANG X, XIAO T, JIANG Y, et al. Repulsion loss: detecting pedestrians in a crowd[C]//IEEE conference on computer vision and pattern recognition (CVPR). Salt Lake City; IEEE, 2017: 7774-7783.
- [7] LIN T, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//IEEE international conference on computer vision (ICCV). Piscataway, NJ; IEEE, 2017: 2980-2988.
- [8] GOLDMAN E, HERZIG R, EISENSCHTAT A, et al. Precise detection in densely packed scenes[C]//IEEE conference on computer vision and pattern recognition (CVPR). Rosten; IEEE, 2019: 2125-2135.
- [9] ZHANG H, CISCÉ M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization[C]//International conference on learning representations (ICLR). Vancouver; International Conference on Learning Representations, 2018: 112-125.
- [10] YU J, JIANG Y, WANG Z, et al. Unitbox: an advanced object detection network[C]//Proceedings of the 24th ACM international conference on multimedia. New York; ACM, 2016: 516-520.
- [11] 李云鹏, 侯凌燕, 王超. 基于 YOLOv3 的自动驾驶中运动目标检测[J]. 计算机工程与设计, 2019, 40(4): 1139-1144.
- [12] 吴涛, 王伟斌, 于力, 等. 轻量级 YOLOV3 的绝缘子缺陷检测方法[J]. 计算机工程, 2019, 45(8): 275-280.
- [13] 戴伟聪, 金龙旭, 李国宁, 等. 遥感图像中飞机的改进 YOLOv3 实时检测算法[J]. 光电工程, 2018, 45(12): 180350.
- [14] 邵义浩, 高志权, 张明月, 等. 基于 YOLOV3 和 KCF 的高速公路监控视频交通事件检测[J]. 中国交通信息化, 2019(S1): 197-201.
- [15] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//IEEE conference on computer vision and pattern recognition. Honolulu; IEEE, 2017: 936-944.