计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(4):266−271 [doi: 10.15888/j.cnki.csa.007321]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 SENet 改进的 Faster R-CNN 行人检测模型①

李克文, 李新宇

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 李新宇, E-mail: m17071008@s.upc.edu.cn

摘　要: 随着无人驾驶和智能驾驶技术的发展, 计算机视觉对视频图像检测的实时性和准确性要求也越来越高. 现有的行人检测方法在检测速度和检测精度两个方面难以权衡. 针对此问题, 提出一种改进的 Faster R-CNN 模型, 在 Faster R-CNN 的主体特征提取网络模块中加入 SE 网络单元, 进行道路行人检测. 这种方法不仅能达到相对较高的准确率, 用于视频检测时还能达到一个较好的检测速率, 其综合表现比 Faster R-CNN 模型更好. 在 INRIA 数据集和私有数据集上的实验表明, 模型的 mAP 最好成绩能达到 93.76%, 最高检测速度达到了 13.79 f/s.

关键词: 行人检测; 卷积神经网路; Faster R-CNN; SENet

引用格式: 李克文,李新宇.基于 SENet 改进的 Faster R-CNN 行人检测模型.计算机系统应用,2020,29(4):266–271. http://www.c-s-a.org.cn/1003-3254/7321.html

Pedestrian Detection Model Based on Improved Faster R-CNN with SENet

LI Ke-Wen, LI Xin-Yu

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Computer vision is an important branch of machine learning at present, which requests much higher instantaneity and accuracy as the driverless and SI-Drive development. To optimize the current methods, the Faster Region-based Convolutional Neural Network (Faster R-CNN) is upgraded by adding SENet to it in this study. The upgraded Faster R-CNN model is applied in pedestrian detection. The new model does not only bring higher accuracy but also accomplish a better detection rate. To verify the new method, an examine was done in INRIA set and our set. The result shows that the upgraded model has a better detection performance on both accuracy and rate which can meet the related specifications of real-time pedestrian detection basically. Finally, the method was tested in the NVIDIA GTX1080Ti GPU. The results show that the mAP of upgraded model can achieve up to 92.7%, while the detection rate is up to 13.79 f/s under a relatively plain experimental condition. On the whole, the new model performs better than the traditional Faster R-CNN model.

Key words: pedestrian detection; CNN; Faster R-CNN; SENet

引言

行人检测是通过计算机视觉技术确定图像或视频中是否包含行人并标记行人的具体位置[1,2]. 目前, 行人检测算法主要分成 3 类: 基于背景建模, 基于模板匹配和基于统计学习. 基于背景建模的算法可以检测到具有光流变化的移动着的人, 但是它无法检测处于静态的人. 基于模板匹配的方法利用图像中的轮廓、纹理和灰度信息来检测目标. 该方法比较简单, 因为它是在原始图像上进行操作, 不需要进行复杂的特征提取, 缺点是我们需要各种行人模板才能取得较好的效果, 并

① 收稿时间: 2019-08-02; 修改时间: 2019-09-09; 采用时间: 2019-09-18; csa 在线出版时间: 2020-04-05

© 中国科学院软件研究所 http://www.c-s-a.org.cn

且模板匹配的方法花费时间较长. 为了更好地描述行人特征, 达拉尔等人提出了梯度直方图 (HOG) 特征, 结合简单的线性支持向量机 (SVM), 取得了很好的效果[3]. 之后, Felzenswalb 通过梳理 HOG 特征提出了可变形组件模型 (DPM). DPM 通过解决行人遮挡问题, 进一步提高了检测精度[4]. 传统的行人检测过程主要包含五个部分: 图像预处理、窗口滑动、特征提取、特征分类和后处理. 传统的方法存在以下缺点: (1) 特征粗糙; (2) 分类误差高; (3) 一种特征只能适用于特定的场景, 在另一种情况下很难取得好成绩[5,6].

近年来, 深度学习发展迅速, 人工智能领域研究的格局也随之变化, 计算机视觉方向尤其得到广泛关注. 2006 年, Hinton 等提出一种基于深度学习的行人检测算法, 利用卷积神经网络 (CNN) 从行人数据集中学习具有高表示性的特征. 与传统特征相比, 高级特征更丰富, 表现力更强, 行人检测性能更好. 2012 年, Hinton 使用 CNN 在 2012 年 ILSVRC 中获得第一名, 分类任务 Top-5 错误率为 15.3%[7]. CNN 在图像识别领域的成功应用, 使得越来越多的人开始关注 CNN. 在物体检测领域, Girshick 提出了 R-CNN 模型[8], 应用选择性搜索算法选择图像中相同大小的几个候选区域, 然后通过 CNN 提取高级特征并通过 SVM 进行分类. 为了提高 R-CNN 模型的准确性和计算速度, Girshick 提出了 Fast R-CNN 模型[9]. 而后 Ren 等基于 Fast R-CNN 提出了 Faster R-CNN 模型, 该模型使用 RPN 网络生成目标候选区域. 基于 Faster R-CNN 的目标检测过程包含在整个深度卷积神经网络中, 旨在加速候选框的提取并克服手工特征的鲁棒性问题[10,11]. 本文在经典 Faster R-CNN 的基础上提出了一种改进的行人检测方法, 使用嵌入 SENet 单元的 VGG-16 网络作为原有模型的特征提取网络. 该方法在 INRIA 行人数据集上进行了训练并在 INRIA 行人数据集和自制的私有的数据集上进行了联合测试. 实验证明该方法提高了模型的检测性能.

1 相关工作

1.1 SENet 简介

SENet (Squeeze-and-Excitation Networks)[12], 是一种网络原子模型, 由 Hu 等人提出, 并在 ImageNet2017 竞赛 Image Classification 任务中获得冠军. 近几年来, 卷积神经网络在计算机视觉领域取得巨大突破. 卷积神经网络的核心部件是卷积核, 卷积核可以看作是空间信息和特征维度信息的聚合体, 就好比人的眼睛在一幅画面中的局部感受视野中的信息. 深度卷积神经网络有一系列的卷积层、池化层、非线性层和归一化层组成, 这种结构使得网络能够捕获图像的全局特征.

目前大部分的卷积神经网络模型都是在空间维度上提升网络性能, VGG 结构和 Inception 模型表明, 增加网络的深度可以显著提高网络学习特征的质量. SENet 则侧重考虑特征通道之间的关系, 对特征通道之间的相互依赖关系进行显式建模. 具体的说, 就是采用一种对特征通道进行重新标定的方法, 网络在学习特征的过程中, 同时学习了每个通道对总体特征的贡献值, 然后依照这个贡献值来提升有用的特征并抑制对当前任务贡献不大的特征.

给定一个特征通道数是 c_1 的输入 x, 对 x 进行一系列卷积操作变换得到一个特征通道数为 c_2 的特征集. 然后在此基础上进行 Squeeze、Excitation、Reweight 等一系列操作, 最终得到一个具有通道权重分配的特征集 $\tilde{x}$.

1.2 Faster R-CNN 简介

Faster R-CNN 是一种通用的目标检测算法, 采用 Two-Stage 策略, 输入图像分别通过卷积层和区域提议网络 (RPN), 最后经过一层池化层和全连接层, 得到最终的分类得分和边框回归. 该算法的主要思想是设计 RPN 网络提取所提出的区域并利用卷积神经网络生成所提出的区域. 用于生成区域的卷积神经网络 (CNN) 的卷积层参数被共享给用于分类的 CNN. 该方法使得算法不再依赖于单独的模块来生成所提出的区域. 然后对生成的提议区域进行分类并计算边界框回归. 用 RPN 替换选择性搜索以缩短区域提议的时间, 同时大大减少了模型在检测网络上花费的时间.

RPN 采用滑动窗口选择的方法, 在共享卷积网络的最后一层输出的特征图上生成区域提议. RPN 的输入是卷积特征图的 $n\times n$ 滑动窗口. 对于每个滑动窗口来预测 k 个锚点的对象区域建议, 每个锚点具有相应的比例. 卷积特征图中的每个点都是一个锚点中心, 其中有 k 个对应的锚点. 对于 $w\times h$ 卷积特征图, 存在 $w\times h$ 个锚点. 每个窗口同时作为低维特征向量传递到分类网络和回归网络中. 分类网络输出每个锚属于对象的概率. 对于每个窗口, 有 $2k$ 个得分输出. 边界回归网络的输出是每个锚点的平移和缩放值, 每个窗口输出 $4k$ 个坐标.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Faster R-CNN 网络结构如图 1 所示, 其中特征提取网络可以使用 ZFNet、VGG-16、ResNet 等. ZF 网络由 5 层卷积层和两层全连接层组成, 相对于 AlexNet, ZF 网络只是卷积核个数和步长发生了变化, 网络结构并没有明显的进步. 无论从其结构和检测结果上都无法达到深度网络的要求.

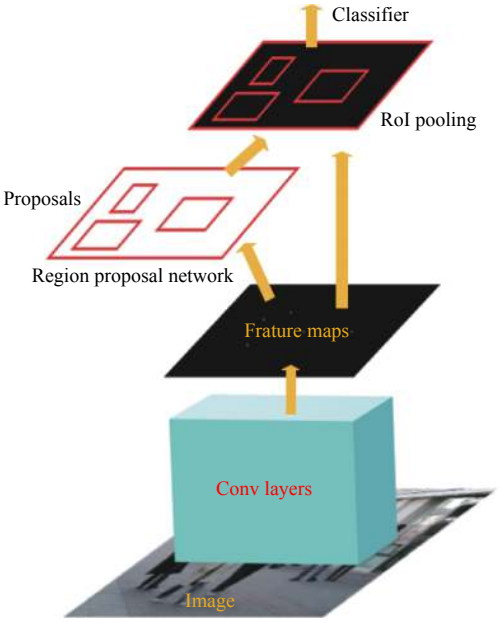

图 1 Faster R-CNN 结构图

2 改进的 Faster R-CNN 模型

2.1 改进的网络结构及可行性分析

提升网络性能一般是提升网络的训练速度和准确率这两个指标. 许多方法研究表明, 适当增加网络的深度可有效增强网络提取特征的质量, 然而网络加深参数也随之增多, 计算量增大导致运行成本增加, 速度变慢. 因此要在准确率和运行速度两个指标权衡的基础上对网络进行优化.

VGG-16 网络是一个由 13 层卷积层、3 层全连接层、5 层最大池化层组成的网络结构, 结构图如图 2 所示.

由于 SE 块中的操作就是池化、全连接这样的基本操作, 因而具备一定的灵活性, 能够直接嵌入到含有 skip-connections 的模块中, 例如 ResNet、Inception 等结构. 本文正是基于这一特点改进 VGG-16 网络结构, 在每一层最大池化层之前加上一层 SE 网络层, 用于处理池化层之前的卷积特征集. 卷积层改进如图 3 所示.

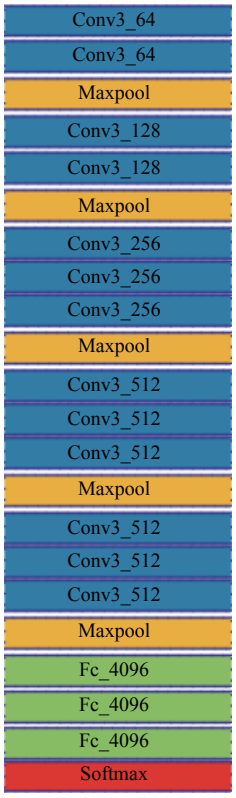

图 2 VGG-16 网络结构

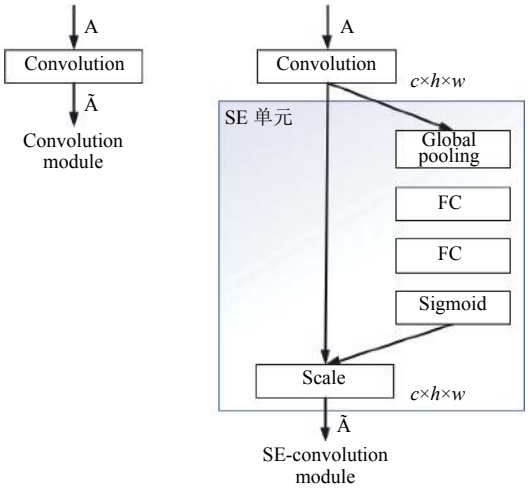

图 3 VGG 卷积模块及改进

目前几乎所有检测模型的特征提取网络都是几种经典的网络, 如基于 GooleNet 的 YOLO 模型, 基于 Darknet-53 的 YOLOv3, 基于 VGG16 的 SSD 网络模型, 以及基于 ZF 的 Faster R-CNN 模型等. 传统的 Faster R-CNN 模型使用经典的 ZF 网络进行特征提取. 用于道路行人检测时, 尤其是有遮挡和小目标检

© 中国科学院软件研究所 http://www.c-s-a.org.cn

测的情况下, 原始网络结构检测精度和检测速率相对较低. 本文提出一种改进的 Faster R-CNN 方法, 使用嵌入 SE 单元的 VGG-16 作为新的特征提取网络, 使用 K-means 聚类和 RPN 相结合设计分类网络, ReLU 作为激活函数, 来设计检测网络, 最终的输出结果是一个 Bounding box(以下统称 bbox) 回归和归一化分类得分. 通过本文方法, 给卷积提取的特征通道之间赋予相应的权重, 提升了特征表示的质量, 从而提升了检测的精度. 另一方面优秀的特征表示能够使网络加快收敛, 从而提高了检测速率. 实验证明该模型可以用于道路行人实时检测. 改进的网络结构图如图 4 所示.

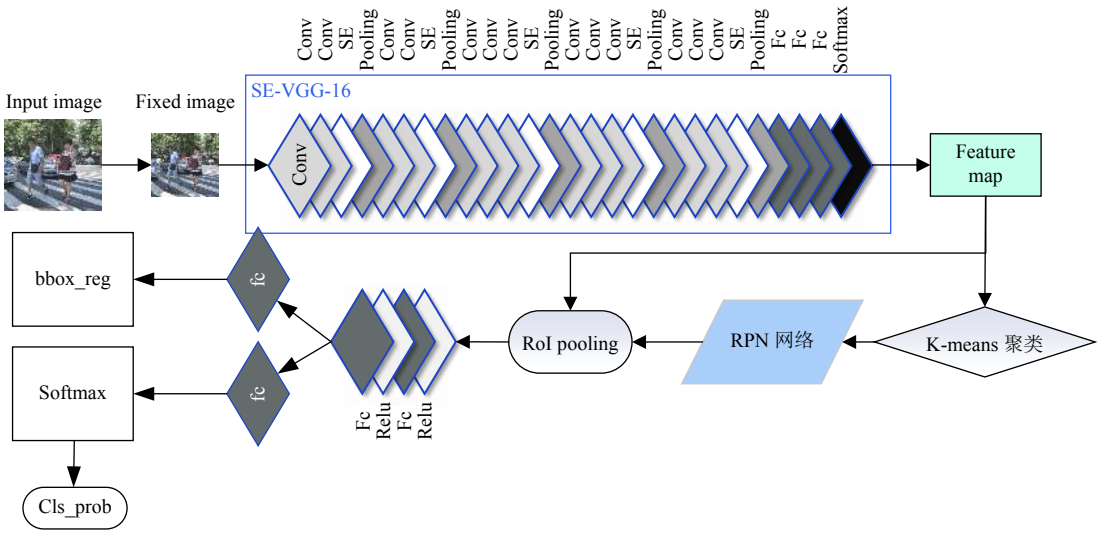

图 4 改进的网络结构图

网络检测主要流程如下:

(1) 将原始数据预处理为 $M\times N$ 大小的图像作为网络输入.

(2) 通过特征提取网络 SE-VGG-16 提取特征.

(3) 将提取的特征集分成两路, 一路输入到 RPN 网络, 一路传到特有卷积层以获取更高维的特征.

(4) 经过 RPN 网络处理的特征图会产生一个对应的区域得分, 然后通过最大值抑制算法得到区域建议.

(5) 把步骤 (3) 得到的高维特征和 (4) 中得到的区域建议同时输入到 RoI 池化层, 提取对应区域建议的特征.

(6) 将得到的区域建议特征输入到全连接层, 得到区域的分类得分以及回归后的 bbox.

2.2 训练

在网络训练阶段, 需要设置候选区域的规格和数量. 随着迭代次数的增加, 候选区域参数被连续调整, 最终接近真实的行人检测区域. 为了加快收敛速度, 使用 K-means 方法聚类与图像中的行人相似的候选区域. K-means 聚类应用欧氏距离来测量两点之间的距离, 其聚集了单位网格的宽度和长度的比率. *IoU* 是反映候选区域和真实待检测区域之间差异的重要指标. *IoU* 值越大, 两个区域之间的差别越小. K-means 聚类函数为: $min\sum_N\sum_M(1-IoU(Box[N],Truth[M]))$. 其中, N 指聚类的类别, M 指聚类的样本集, $Box[N]$指候选区域的宽和高, $Box[M]$指实际行人区域的宽和高.

在检测阶段, 利用检测网络进行汇集操作, 通过 bbox 分类网络对区域进行分类, 并通过 bbox 回归网络预测行人的边界框. 在检测网络中有两个并行的输出层, 分类层的输出是每个候选区域在行人和背景上的概率分布. 每个候选区域对于行人和背景这两个类别的概率分布是 $p=(p0,p1)$. 回归网络的输出是行人边界框坐标的参数: $t^k=\left(t_x^k,t_y^k,t_w^k,t_h^k\right)$, 其中 k 代表类别. 边界框回归网络和边界分类网络通过如下联合损失函数进行训练: $L(p,u,t^u,v)=L_{\text{cls}}(p,u)+\lambda[u\geq 1]\cdot L_{\text{reg}}(t^u,v)$. $L_{\text{cls}}(p,u)=-\log(p_u)$指的是实际类别的对数损失. L_{reg} 只有当检测到的是行人时才会激活.

RPN 和检测网络共享特征提取的卷积网络, 即 SE-VGG 模型, 其通过计算图像特征大大减少了计算

© 中国科学院软件研究所 http://www.c-s-a.org.cn

时间. 首先对 SE-VGG 网络进行训练, 然后训练 RPN, 通过 RPN 提取的区域建议训练检测网络, 然后对检测网络的参数进行 RPN 训练. 也就是说, RPN 和检测网络被联合训练, 重复进行直到实现收敛.

3 实验设计及结果分析

3.1 实验环境及参数设计

本实验的训练样本取自 INRIA 数据集, 包括 15 560 个正样本和 6744 个负样本, 分别选择 5000 个正样本和负样本作为训练样本. 测试集中有 1000 张图片, 其中包括来自 INRIA 的 500 张图片和 500 张个人收集制作的测试图片. 实验的硬件配置采用 Intel Core i7 处理器, 16 GB 内部存储器, NVIDIA GTX1080Ti. 在训练阶段, 训练集中的每个行人都需要用矩形框标记. 在测试中, 如果识别出的行人和标记的矩形框的重叠部分达到标记矩形框的 90% 以上, 则认为是成功检测到的.

该实验选用 Tensorflow 框架实现卷积神经网络模型, 分别做了原始的 Faster R-CNN 模型 (基于 ZF 网络)、基于 VGG-16 的 Faster R-CNN 模型和基于 SE-VGG-16 的 Faster R-CNN 模型的实验, 其中特征提取网络都是在 ImageNet 上进行预训练的. 参数设计包含 dropout、最大迭代次数、候选区域框的大小和 nms (非极大值抑制) 阈值. 这些值的设置都会对 mAP 值产生一定的影响, 所以优化这些参数以获得更好的输出也是实验要做的一部分. 在本实验中, 最大迭代次数设置为 8000, 初始候选区域大小为 256, 边界框的大小为 128.

3.2 实验结果分析

首先在原始网络上进行实验, 以选取最佳参数值. 结果如表 1. 可以得到如下结论: 当 *dropout* 值从 0.2 增加到 0.6 时, *mAP* 随之增加, 从 0.6 增加到 0.8 时, *mAP* 值随之减小. 当 *dropout* 为 0.6 时, *mAP* 获得最大值. 因此本实验选取 *dropout*=0.6.

表 2 中的结果是在 *dropout* 为 0.6 且最大迭代次数为 8000 的条件下得到的, 该结果表明候选区域的大小不同会导致不同的 *mAP* 值, 候选区域越小, *mAP* 值越大. 因为候选框的选取会影响检测速率, 理论上候选框大小设定的越小, 检测精度越高, 速率越慢. 本文综合考虑检测速率和精度选取候选框大小为 128.

表 1 *dropout* 对 *mAP* 的影响

dropout	Size of region proposal	Size of bounding box	*mAP*
0.2	256	128	0.806
0.3	256	128	0.812
0.4	256	128	0.845
0.5	256	128	0.828
0.6	256	128	0.905
0.7	256	128	0.805
0.8	256	128	0.795

表 2 区域框大小对 *mAP* 的影响

Size of region proposal	Size of bounding box	*mAP*
256	128	0.905
128	64	0.913
64	32	0.917
32	16	0.921

一般用准确率、检测速率来评价模型的质量. 除此之外, 召回率也是衡量模型好坏的一个指标, 在图片中待检测的目标较多时, 应该尽量使模型做到“一个不漏”. 表 3 中的结果是固定 dorpout 值为 0.6 和区域框大小为 128 时, 选取不同的共享网络得到的准确率、检测速率和召回率的结果. 我们做了 3 个对比实验: 经典的 Faster R-CNN 模型 (特征提取网络为 ZF 网络); 基于 VGG-16 的 Faster R-CNN 模型 (特征提取网络使用 VGG-16 网络模型); 基于 SE-VGG-16 的 Faster R-CNN 网络模型 (特征提取网络使用本文提出的模型). 从以表 3 数据可以看出, 我们的模型在准确率、检测速度和召回率上都有一个较好的结果.

表 3 选取不同网络时的结果对比

Faster R-CNN	*mAP*	FPS	Recall
+ZF	0.913	23	0.87
+VGG-16	0.920	21	0.79
+SE-VGG-16	0.937	25	0.91

4 结束语

本文从行人检测的背景和意义出发, 对现有检测方法做了综合阐述, 得知传统的行人检测方法或是以牺牲准确率来提高检测效率以达到实时性的要求, 或是以牺牲检测时间和空间为代价加深网络来获得高准确率, 我们的研究通过对 Faster R-CNN 模型进行改进, 使用添加 SE 单元的 VGG 网络作为 Faster R-CNN 的特征提取网络, 通过赋予特征图相应权重来提升有用特征并抑制无用特征. 理论和实验证明这种方法很好的权衡了网络检测速率和鲁棒性, 网络整体性能得到了提升.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

参考文献

1 Li HL, Wu ZD, Zhang JW. Pedestrian detection based on deep learning model. International Congress on Image and Signal Processing, Biomedical Engineering and Informatics. Datong, China. 2017. 796800.

2 Hattori H, Lee N, Boddeti VN, *et al*. Synthesizing a scene-specific pedestrian detector and pose estimator for static video surveillance. International Journal of Computer Vision, 2018, 126(9): 1027–1044. [doi: 10.1007/s11263-018-1077-3]

3 Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA. 2005. 886893.

4 Li Y, Ding WL, Zhang XG, *et al*. Road detection algorithm for autonomous navigation systems based on dark channel prior and vanishing point in complex road scenes. Robotics and Autonomous Systems, 2016, 85: 1–11. [doi: 10.1016/j.robot.2016.08.003]

5 Conte D, Foggia P, Percannella G, *et al*. Counting moving persons in crowded scenes. Machine Vision and Applications, 2013, 24(5): 1029–1042. [doi: 10.1007/s00138-013-0491-3]

6 Wang ZQ, Liu J. A review of object detection based on convolutional neural network. 2017 36th Chinese Control Conference (CCC). Dalian, China. 2017. 6.

7 Zhao HY, Kim O, Won JS, *et al*. Lane detection and tracking based on annealed particle filter. International Journal of Control, Automation and Systems, 2014, 12(6): 1303–1312. [doi: 10.1007/s12555-013-0279-2]

8 Gupta S, Girshick R, Arbeláez P, *et al*. Learning rich features from RGB-D images for object detection and segmentation. European Conference on Computer Vision. Zurich, Switzerland. 2014. 345–360.

9 Girshick R. Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. 2015. 1440–1448.

10 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]

11 Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Computer Vision-ECCV 2014. Springer International Publishing, 2014.

12 Hu J, Shen L, Sun G, *et al*. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017: 99.

www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn