计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(9):215−218 [doi: 10.15888/j.cnki.csa.005943]

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

基于卷积神经网络的实时行人检测方法①

龚　安, 李承前, 牛　博

(中国石油大学(华东) 计算机与通信工程学院, 青岛 266580)

摘　要: 近年来, 卷积神经网络在行人检测领域取得了同其他方法相似甚至更好的检测成绩, 然而缓慢的检测速度远不能满足现实需求. 针对这一问题, 本文提出一种实时的行人检测方法, 将分散的检测过程整合成单一的深度网络模型, 被检测图片通过模型的计算可以直接输出检测结果. 使用扩充的 ETH 数据集进行训练测试, 实验结果表明, 在保证准确率的情况下, 该方法检测速度极快, 可以满足实时检测的目的.

关键词: 行人检测; 目标检测; 卷积神经网络; 图像处理; 深度学习

引用格式: 龚安,李承前,牛博.基于卷积神经网络的实时行人检测方法.计算机系统应用,2017,26(9):215–218. http://www.c-s-a.org.cn/1003-3254/5943.html

Real-Time Pedestrian Detection Method Based on CNNs

GONG An, LI Cheng-Qian, NIU Bo

(College of Computer &Communication Engineering, China University of Petroleum(East China), Qingdao 266580, China)

Abstract: In recent years, the convolution neural networks in the field of pedestrian detection have achieved similar and even better results, compared to other methods. However, the slow detection speed can't meet the realistic demand. To solve this problem, a real-time pedestrian detection method is put forward. The scattered detection processes are integrated into a single depth network model. Images which can be calculated through the model can directly output detection results. The extended ETH dataset is used for training and testing the model. The experimental results show that the method is very fast and can achieve the goal of real-time detection with the guaranteed accuracy.

Key words: pedestrian detection; object detection; convolution neural networks; image processing; deep learning

行人检测作为目标检测在行人领域内的一部分, 由于其检测目标的特殊性、广泛的应用前景及商业价值, 成为国内外学者及相关从业者研究的热点. 多年来, 科研人员研究设计了众多行人检测方法, 典型的有: P. Viola 等人设计使用的行人检测模型, Dalal 提出的基于 HOG 特征的行人检测方法, Felzenswalb 等人提出的 DPM 模型等. 在经典模型的基础上, 目前研究人员设计改进的算法[1-3]取得了较好的行人检测效果.

近年来, 卷积神经网络发展迅速, 在物体分类, 行为识别, 物体检测等领域取得成功, 研究人员开始尝试将卷积神经网络应用到行人检测上来[4-6], 其中文献[6]是目前准确率最好的检测方式之一. 然而, 这些方法检测速度较慢, 以文献[6]为例: 作者在 R-CNN[7]模型上进行修改, 主要是将 region proposal 方法由原来的 selective search 方法[8]改为 Katamari 方法[9], 实验取得很好的成绩. 然而, region proposal 运行过程消耗大量的时间, 故很难达到实用目的.

根据卷积神经网络在目标检测领域取得的最新进展[10-12], 本文提出了一种基于卷积神经网络的实时行人检测方法. 该检测方式非常简单: 图片作为输入无需任何预处理, 经过本文设计的网络模型计算, 直接输出检测到的行人位置. 相比一般的行人检测方法, 该模型

① 收稿时间: 2016-12-22; 采用时间: 2017-01-18

具有以下优点: 首先, 模型是一个完整的神经网络, 不需要对分散的模块逐一设计分析, 训练及运行方法简单; 其次, 模型以原始图片作为输入, 进行整体的训练调优, 可以减少分部处理导致的信息丢失, 更好的获取图片内的上下文联系; 实验证明, 在保证准确率的情况下, 模型可以达到实时处理的需求.

1 基于卷积神经网络的实时行人检测

文献[12]提出的 SSD 模型在物体检测领域取得 state-of-art 的检测成绩, 并可以达到实时的检测效果. SSD 模型作为完整的卷积神经网络, 包含特征提取、不同尺度的物体检测和输出层等 3 个部分. 输入图片通过去顶的 VGG-16 模型[13]进行特征提取, 经由 6 种不同尺度的检测模块实现卷积特征到物体检测的映射. 受到 SSD 模型的启发, 考虑到行人检测领域的特殊性, 本文设计了一种基于卷积神经网络的实时行人检测模型.

1.1 模型设计

图 1 是本文设计的行人检测模型, 由统一的卷积神经网络实现. 模型前端是用来提取图片特征向量的卷积神经网络, 后端通过卷积生成不同大小的特征图. 在不同的特征图上, 计算不同候选框的形状偏移量和该区域为行人的概率得分, 通过非极大值抑制方法输出行人检测结果.

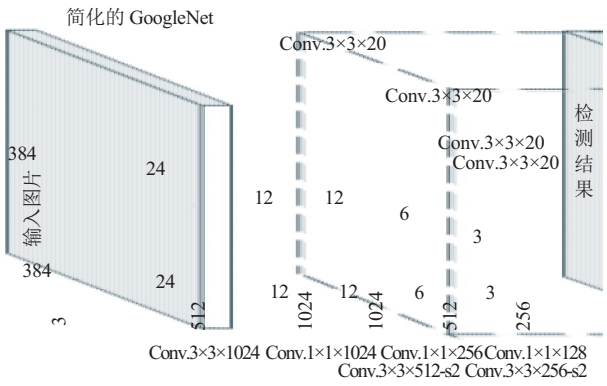

图 1 基于卷积神经网络的实时行人检测模型

GoogleNet[14], VGG 等作为当前流行的深度卷积神经网络模型, 能够很好的提取图片特征. 但这两种模型过于庞大, 很难训练. 文献[11]提出的特征提取模型作为 GoolgeNet 的精简版, 取得类似 GoogleNet 的训练成绩, 故采用其方法实现模型前端的特征提取.

考虑到行人检测的特殊性, 本文设计了如图 2 所示的多候选框行人检测模板. 模板包含 4 个不同大小的宽高比为 1:3 的候选框, 用来预测不同大小的行人. 其中, 为了防止两个相邻的较小行人像素的漏检, 本文为模板设计了两个相同的最小候选框. 对于每一个候选框, 我们预测其形状偏移量(t_x,t_y,t_w,t_h)及行人概率 p. 训练阶段, 根据输入图片内的不同 Ground truth 的尺寸, 选取特定特征图上的特定候选框与之匹配, 通过计算(t_x,t_y,t_w,t_h,p)的损失函数对模型进行训练.

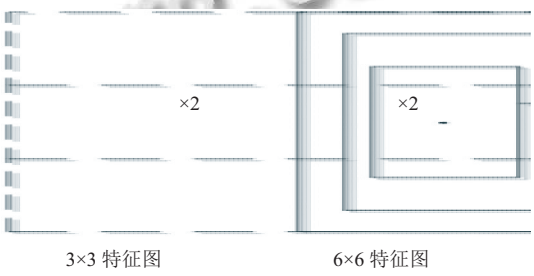

图 2 多候选框行人检测模板

本文在模型前端的特征提取部分后添加不同的卷积操作, 得到 3 个大小不同的特征图, 加上模型前端 24×24×512 的特征图, 用于不同尺度的行人检测. 对于大小为 m×m×p 的特征图, 其任意位置为中心 3×3×p 的特征向量, 负责计算 4 个不同候选框的形状偏移量和行人概率.

1.2 训练方法

设(x,y,w,h)、(x_a,y_a,w_a,h_a)、(x^*,y^*,w^*,h^*)表示预测框、候选框、Ground truth 的中心和宽高, (t_x,t_y,t_w,t_h)、$(t_x^*,t_y^*,t_w^*,t_h^*)$表示预测框和 Ground truth 相对于候选框的偏移量, 其对应关系如下:

$$\begin{aligned} &t_x=(x-x_a)/w_a \quad t_y=(y-y_a)/h_a \\ &t_w=\log(w/w_a) \quad t_h=\log(h/h_a) \\ &t_x^*=(x^*-x_a)/w_a \; t_y^*=(y^*-y_a)/h_a \\ &t_w^*=\log(w^*/w_a) \quad t_h^*=\log(h^*/h_a) \end{aligned} \tag{1}$$

模型以候选框的偏移量 t 和其为行人的概率 p 作为输出.

对于任意一个 Ground truth, 都有唯一的候选框与之匹配. 另外, 对于任意的候选框, 若其与某个 Ground truth 相交比例大于 0.7, 则该候选框也匹配该 Ground truth. 将上述两种情况作为正样本. 若某个候选框与任意的

Ground truth 相交比例小于 0.3, 则将其作为负样本. 采用上述正负样本对模型进行训练, 目标损失函数如下:

$$\begin{aligned}&\mathrm{L}(\{p_i\},\{t_i\})\\&=\frac{1}{N}\left(\sum_i L_{conf}\left(p_i,p_i^*\right)+\lambda\sum_i p_i^* L_{loc}\left(t_i,t_i^*\right)\right)\end{aligned}\tag{2}$$

其中, i 是候选框的索引, p_i代表第 i 个候选框为行人的概率, 对于正样本, p_i^*=1, 否者p_i^*=0; N 代表训练样本的个数, λ为平衡参数. 行人分类损失函数L_{conf}的计算使用 Logistic 函数实现, 对于行人定位的损失函数L_{loc}计算方法如下:

$$L_{loc}(t,t*)=\sum_{j\in\{x,y,w,h\}} Smooth_{L1}(t_j-t_j*)\tag{3}$$

$$Smooth_{L1}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{其他}\end{cases}\tag{4}$$

2 实验与分析

本文搭建了 Linux+CUDA+NVIDIA GTX Titan 显卡的实验环境. 为加速整个行人检测模型的拟合, 在 ImageNet[15]数据集上预训练模型前端的特征提取层, 预训练同时可以防止模型过度拟合. 模型后端的卷积层使用 Xavier 方法[16]进行初始化.

2.1 实验数据

现今主流的行人检测数据集有 INIRA, ETH, Caltech, KITTI 等. 考虑到 INIRA、ETH 包含的数据集较少, Caltech 提供的图像清晰度较差等因素, 本文以 ETH 数据集为基础, 从其他数据集中抽出图像作为扩充, 制作了实验数据 ETH+. 其中选取 5000 张图片(约 3×10^4个行人标签)作为训练集, 1000 张图片作为测试集.

2.2 实验结果

实验阶段对模型进行 120 次的训练, 其中前 80 次训练的学习率为10^{-3}, 后 40 次的学习率调整到10^{-4}. 同时采用随机缩放和水平翻转等方法进行数据增强, 防止模型过度拟合. 实验取得了漏检率(MR)约为 25%, 检测速度高达 100 帧每秒的成绩.

2.3 不同算法对比

如表 1 所示, 在相同数据集下本文进行了不同行人检测方法的实验对比. 其中, 方法 1、2、3 分别代表本文的模型、文献[6]和基于 HOG 的行人检测方法. 可以看出, 在漏检率较低的情况下, 本文算法的检测速度取得了极大的提升.

表 1 不同方法的实验对比

方法	MR(%)	速度(fps)
1	25.6	>100
2	24.9	<10
3	46.2	<1

2.4 候选框对结果的影响

如 1.1 所述, 多候选框行人检测模板包含 4 个候选框, 为了防止两个相邻的较小行人像素的漏检, 其中有两个相同的最小候选框. 作为对比, 当仅使用一个最小候选框, 实验结果显示 MR 上升了 8% 左右. 图 3 上图为使用单最小候选框, 下图为使用双最小候选框的实验结果对照.

图 3 候选框对实验结果影响图

3 结语

本文对卷积神经网络用于行人检测领域进行了研究. 为了加快行人检测的速度, 满足现实使用的需求, 设计实现了一种实时的行人检测模型. 本文同时扩展了 ETH 数据集, 使训练样本更加充分. 实验结果表明, 该模型在取得高准确率的同时, 检测速度超过 100 帧

每秒, 可以用于实时检测.

参考文献

1 Benenson R, Mathias M, Tuytelaars T, *et al*. Seeking the strongest rigid detector. 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA. 2013. 3666–3673.

2 Park D, Zitnick CL, Ramanan D, *et al*. Exploring weak stabilization for motion feature extraction. 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA. 2013. 2882–2889.

3 Yan JJ, Zhang XC, Lei Z, *et al*. Robust multi-resolution pedestrian detection in traffic scenes. 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA. 2013. 3033–3040.

4 Sermanet P, Kavukcuoglu K, Chintala S, *et al*. Pedestrian detection with unsupervised multi-stage feature learning. 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA. 2013. 3626–3633.

5 Luo P, Tian YL, Wang XG, *et al*. Switchable deep network for pedestrian detection. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 899–906.

6 Hosang J, Omran M, Benenson R, *et al* B. Taking a deeper look at pedestrians. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4073–4082.

7 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, OH, USA. 2014. 580–587.

8 Uijlings JRR, van de Sande KEA, Gevers T, *et al*. Selective search for object recognition. International Journal of Computer Vision, 2013, 104(2): 154–171. [doi: 10.1007/s11263-013-0620-5]

9 Benenson R, Omran M, Hosang J, *et al*. Ten years of pedestrian detection, what have we learned? In: Agapito L, Bronstein M, Rother C, eds. Computer Vision-ECCV 2014 Workshops. Cham, Switzerland. 2014. 613–627.

10 Tan M, Hu ZF, Wang BY, *et al*. Robust object recognition via weakly supervised metric and template learning. Neurocomputing, 2016, (181): 96–107. [doi: 10.1016/j.neucom.2015.04.123]

11 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. Las Vegas, NV, USA. 2016. 779–788.

12 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. In: Leibe B, Matas J, Sebe N, *et al*, eds. Computer Vision-ECCV 2016. Cham, Switzerland. 2016. 21–37.

13 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409. 1556 2014.

14 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9.

15 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

16 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research, 2010, (9): 249–256.