【文章编号】 2096－2835(2018)04－0393－05 DOI:10.3969/j.issn.2096－2835.2018.04.008

基于多尺度特征融合的Faster－RCNN道路目标检测

陈 飞,章东平
(中国计量大学 信息工程学院,浙江 杭州 310018)

【摘 要】 道路目标检测在智慧城市建设中扮演着重要角色,而Faster-RCNN是目前主流的目标检测网络结构算法.本文在Faster-RCNN卷积神经网络结构基础上增加了特征金字塔网络层,并采用关注损失函数替代了原有的交叉熵损失函数.其中增加的特征金字塔特征融合层可以提取到检测图片中更具鲁棒性和一般性的前背景特征,而通过关注损失函数则能起到缓解检测图片中的正负样本不均的情况.最后,在公开数据集KITTI上实验证实,改进的目标检测算法能实现提高原有的Faster-RCNN目标检测准确率.

【关键词】 目标检测;特征融合;卷积神经网络;Faster-RCNN算法
【中图分类号】 TP391 【文献标志码】 A

Road object detection based on multi－scale merged feature Faster-RCNN

CHEN Fei, ZHANG Dongping
(College of Information Engineering, China Jiliang University, Hangzhou 310018, China)

Abstract: Road object detection plays a vital role in intelligent city construction; and Faster-RCNN is one of the main stream object detection algorithms. The paper proposed a concatenated feature pyramid network layer on the original Faster-RCNN feature extraction layer and substituted the cross entropy loss function with the focal loss function. The concatenated feature pyramid network could extract the enriched feature maps that were more robust and generalized to diverse situations. The adopted focal loss function could alleviate the sample inhomogeneity in the detected picture. The algorithm was verified by using open KITTI datasets. The result shows that the updated Faster-RCNN algorithm can improve detection precision.

Key words: object detection; merged feature; CNN; Faster-RCNN

目标检测作为计算机领域的一个重要分支,在自动驾驶、智能视频监控、安防领域都有广泛的应用.目标检测主要解决的是图片或视频帧中目标的what与where问题,即是否存在目标,并对

【收稿日期】 2018-09-20 《中国计量大学学报》网址:zgjl.cbpt.cnki.net
【基金项目】 浙江省自然科学基金项目(No. LY15F020021),浙江省公益技术应用研究计划项目(No. 2016C31079).
【第一作者简介】 陈 飞(1992-),男,江西省上饶人,硕士研究生,主要研究方向为深度学习与图像处理. E-mail:1459101904@qq.com
通信联系人:章东平,男,教授. E-mail:silenttree_zju@cjlu.edu.cn

已判断出的目标给定对应的边界框. 目标检测对人类来说并不困难,人眼对图片的不同目标的颜色模块,人眼能很容易的进行定位并给出相应类别. 但对于计算机而言,由于图片的目标物体都是RGB像素矩阵,很难直接得到目标物体的确切位置,再加上有时是多个物体的混叠以及一些背景信息的干扰,以致增大了目标检测难度. 传统机器学习通过对目标提取人工选择特征. 如 HOG[1] (histogram of oriented gradient)、SIFT[2] (scale invariant feature transform)等方法,然后将提取特征输入到分类器,如 SVM[3] (support vector machine)、AdaBoost[4]等进行分类识别[5]. 人工特征构造复杂且泛化性差而滑动窗口又有检测速度慢等缺点. 近年来卷积神经网络在图像识别目标检测等领域取得了突破性进展,掀起了新的研究热潮[6].

1 相关工作

物体结构能很好的反映出目标信息,因此,将目标结构作为目标特征用于检测,可以较为准确的实现目标检测目的. TRIGGS 和 DALAR[7] 采用梯度方向直方图 HOG 特征描述算子在行人检测目标任务中取得了较好的结果,FELZENSZWALB[8] 等人在目标识别中使用了 part-based 模型来表示目标,并在 TRIGGS 和 DALAR 等人基础上采用了混合多模板模型,每个模板含有可移动变形部分,并结合 Latent SVM 分类器进行训练. 实验证明该模板比单一模板具有更好效果. 但由于基于滑动模板的目标检测大多采用密集采用方法,这种方法严重影响了目标检测速度. Dollar[9] 等人在 2014 年提出了一种新的快速特征金字塔计算方法,首先对原始图像进行稀疏采用,然后给出特征金字塔相邻层间的幂指运算,并使用统计方法进行运算.

随着计算机硬件技术的快速发展,特别是 GPU 计算性能的开发和提升以及数据量的几何式增长,促进了深度学习领域的蓬勃发展. 2014 年 Ross Girshick[10] 首先把深度学习方法 RCNN (区域卷积神经网络)应用于目标检测,由于采用了 Selective Search(选择性搜索方法),直接对原图像提出 2 000 多个候选框,并直接对 2 000 个候选框输入 CNN(卷积神经网络)提取相应的特征,然后通过 SVM(支持向量机)对目标进行分类,并采用边框回归(Bounding Box Regression)确定和校准目标位置. 由于每张图片要做 2 000 多次的前向传播,导致检测速度过慢. 鉴于此,Ross Girshick 团队提出了其改进版 Fast-RCNN,Fast-RCNN 采用了直接对整张图片做 CNN 特征提取,并采用了深度学习常见的 Soft-max 分类器来替代 SVM 分类器,以提升其检测速度. 由于 Fast-RCNN 仍旧采用 Selective Search 方法来提取候选框,影响了检测速度. 所以在 Fast-RCNN 改进版 Faster-RCNN 提出了 RPN(区域提议网络),改进了之前的 Selective Search 方法,通过在 RPN 生成候选框,并在 RPN 中对相关边框做出是背景还是目标的初步过滤处理,从而改进了原始 Selective Search 生成提议窗口过慢的缺点.

YOLO[11] 算法的解决办法是通过把目标检测问题转化为回归问题求解,采用一个卷积神经网络来直接获得要预测的 bounding box 及对应的类别概率. 算法首先把输入图像均等划分为 $S\times S$ 个 grid cell(框格),然后对划分的每个框格预测 N 个 bounding box,而预测取得的每个 bounding box 将包含 5 个预测值:x、y、w、h 以及 confidence(置信度分数). 其中 x、y 是 bounding box 的中心坐标点,而 w、h 指的是 bounding box 的宽高值. 每个 bounding box 对应一个置信度分数,若检测到的框格中没有待检目标物体,其置信度值设为 0,若有,其值就是预测 bounding box 与 ground truth 的 IOU(交并比)值. 至于判断框格是否有目标物体,若一个物体的 ground truth 的中心点坐标在一个框格中,那么由该框格负责这个目标检测.

SSD[12] 方法为了避免利用太低层特征,SSD 从后的 conv4_3 开始,又往后增加了几层,分别抽取每层特征,并在每层特征上分别使用了 Soft-max 做背景和目标类别分类处理,并采用边框回归对目标进行定位. 由于 SSD 对高分辨率的底层特征没有再利用,可是这些底层特征对小目标的检测具有重要作用.

FPN 利用 CNN 的金字塔层次结构性质(从低到高的语义特征)构建从低到高的语义特征金字塔. CNN 前馈计算是从下到上的,特征图通过 CNN,一般特征图是越来越小的,也存在同样大小的,此时称为相同网络阶段(same network stage). 在 FPN 中,每个阶段定义一个金字塔级

别，由于每个阶段最深层具有最强的表示特征，所以选择每个阶段的最后一个输出层作为参考层，并实现特征融合效果.

2　多特征融合的特征金字塔结构

本文针对原 FPN(图 1)的特征提取融合结构上进行改进(图 2)，原有特征结构进行增加，以实现丰富语义特征的融合. 特征融合的路径自下向上，通过卷积核运算，对每一次卷积核的输出构成一个特征金字塔. 并对每一特征金字塔同一网络阶段做特征激活输出，确保获得每个阶段的最强特征. 对相关模块分别作相对原图像大小的 2，4，8，16 像素步长. 对自上至下路径，则将原本获得的特征输出做 1×1 的卷积操作实现横向连接. 相邻网络特征层的特征大小成 2 倍比例.

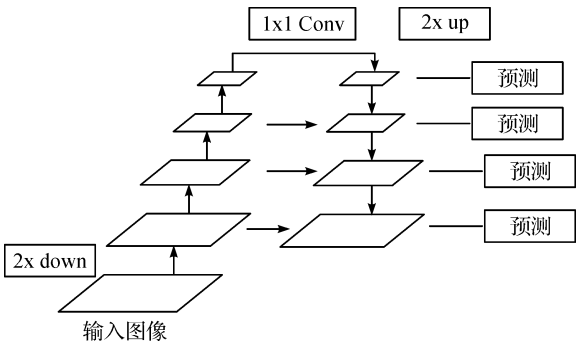

图 1　特征金字塔结构

Figure 1　Feature pyramid architecture

3　特征融合目标检测网络结构描述

本文提出的多特征融合目标检测算法在原有检测算法中增加了 FPN 层，实现多层特征融合，充分利用各个卷积层的特征图取得的不同特性. 低卷积层的高分辨率有利于检测小物体目标，而高卷积层拥有较大感受野，可以用来检测大目标物体. 本文主网络结构采用 Faster-RCNN，Faster-RCNN 在原 Fast-RCNN 目标检测方法上增加了 RPN，对每个像素点生成锚框，并对每个锚框做出背景与目标的判断，过滤掉背景锚框，并对锚框做边框回归处理，然后再把选择有目标锚框输入到 ROI Pooling 层以减小物体检测的区域敏感性. 其中锚框的生成大小满足以下要求：$\{(w\times h),(\alpha w\times\alpha h),(w\gamma,h/\gamma)\}$.（其中锚框的大小 w，h 表示初始边框的宽高，α，γ 表示边框缩放比例，且 $\alpha\in(0,1]$，$\gamma>0$，当生成 n 个大小，m 个比例时，则共产生有 $m\times n$ 个锚框）. 由特征融合的网络结构图描述，将输入的 RGB 三通道彩色图片做四次卷积运算，构成自下到上的特征金字塔结构 Conv1_1 至 Conv4_1，然后从 Conv4_2 至 Conv1_2 构成自上到下的特征金字塔结构. 按特征金字塔结构图描述，由原来的自下到上的同一网络阶段做 1×1 卷积核操作实现特征融合，并对同一网络阶段的融合特征在做 2×2 卷积核操作以解决上采样造成的混叠效应. Conv4_2，Conv3_2，Conv2_2，Conv1_2 的输出结果分别输入到 RPN. 虽然看上去较为复杂，但由于 RPN 思想更为直观. 首先提议预先配置好的一些区域. 然后通过神经网络来判断这些提议区域是否是感兴趣的，是的话，则在最后输出时再进行预测，得到一个更加准确的边框，这样我们能有效降低搜索边框的代价.

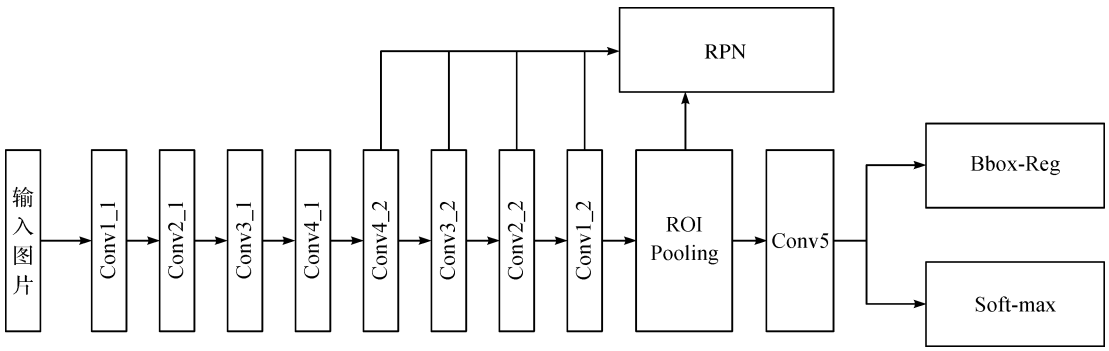

图 2　特征融合网络结构图

Figure 2　Feature merged network architecture

4 训练方法

KITTI[13]数据集由德国卡尔斯鲁理工大学和丰田美国研究院联合制作，是目前世界上应用于自动驾驶相关的计算机视觉算法评估的重要数据集. 包含市区、乡村、高速公路等真实图像数据，一张图片最多包含 15 辆车，且存在遮挡、截断等复杂路况. 本文采用 KITTI 数据集来训练完成车辆检测任务. 神经网络训练通过前向传播，联合数据真实标签和损失函数，获取训练样本的类别预测值的概率值和边框值，再通过反向传播，不断迭代更新带训练参数值，直到获取对应的类别参数和边框位置参数的损失函数输出值收敛. 数据预处理阶段对待训练数据做亮度、色调、裁剪等数据增强操作，获得更多训练数据. 鉴于 ImageNet 是一个拥有千万图像数据集，而本文 KITTI 数据样本较少，采用 fine tune 来训练车辆检测模型. 两个集合的相似度常用的判别方法是 Jaccard 距离. 在图像上，被称为 IOU(Intersection over Union). $J(S_1,S_2)=\frac{|S_1\cap S_2|}{|S_1\cup S_2|}$，值越大，表示两个框相似度越高. 本文取 $J(S_1,S_2)>0.7$ 为正样本，$J(S_1,S_2)<0.3$ 为负样本. 对于分类问题，常见的损失函数是交叉熵损失函数. 不同于交叉熵 $\log(p_i)$，这里采用了 focaloss[14] 的 i 是真实的类别，且 p_i 是相应的类别预测概率. 给定 γ，λ，作为调节曲线权重的陡度，均是待调节超参数，本文中 λ 值设为 2，γ 设置为 0.5，并给出相应的模型训练的关注损失函数的定义为

$$\mathrm{Loss}_{ds}=-\gamma(1-p_i)^{\lambda}\log(p_i) \tag{1}$$

常见边框回归损失函数一般使用平方损失函数 $\mathrm{Loss}=x^2$. 但该损失函数对较大的误差惩罚过高，可以通过采用绝对损失函数来降低惩罚 $\mathrm{Loss}=|x|$. 由于零点处左右导数不相等[15]. 通过增加平方项使其变得更为平滑.

$$\mathrm{Loss}_{loc}=\begin{cases}(\alpha x)^2/2, & \text{if } x<1/\alpha^2;\\ |x|-0.5/\alpha^2, & \text{otherwise.}\end{cases} \tag{2}$$

其中，$x=\sum_{i\in(x,y,x+w,y+h)}(m_i-\hat{m}_i)$，其中 i 代表对应边框的左上点坐标(x,y)，右下点坐标$(x+w,y+h)$. 由联合损失函数 $\mathrm{Loss}_{\mathrm{joint}}=\mathrm{Loss}_{\mathrm{loc}}+\mathrm{Loss}_{\mathrm{cls}}$. 联合上述公式，迭代训练使 $\mathrm{Loss}_{\mathrm{joint}}$ 取得最小值，可以获得对应的边框位置值和分类置信度值.

5 结果分析

本文相关实验采用的服务器 CPU 配置为 Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz，GPU 配置为 NVIDIA TITAN X 所采用深度学习框架为 mxnet. 训练使用的 KITTI 数据集共有 7 481 张训练图片，7 581 张测试图片，采用分离验证方法，将数据分成训练集 7 481 张，验证集3 800 张和测试集 3 781 张图片. 首先用训练集获得训练模型，再用验证集评估模型性能，如图 3、图 4，并用测试集获得目标检测的测试性能，如表 1[16].

图 3 基于特征融合的 Faster-RCNN 检测效果实例

Figure 3 Detection performance based on merged feature Faster-RCNN

图 4 基于 Faster-RCNN 的检测效果实例

Figure 4 Detection performance based on original Faster-RCNN architecture

表 1　不同网络结构对分类性能对比

Table 1　Comparation of various network structures on classification performance

网络结构	平均检测时间/s	行人 mAP	车辆 mAP
Faster-RCNN	1.75	75.26	79.51
Faster-RCNN＋3 层特征金字塔	1.83	77.84	80.12
Faster-RCNN＋4 层特征金字塔	1.89	78.36	80.30
Faster-RCNN＋5 层特征金字塔	1.94	78.53	80.46
Faster-RCNN＋6 层特征金字塔	1.98	78.43	79.80

观察上述实验验证结果，当增加卷积特征层层数，Faster-RCNN 的检测结果有相应的提高，当特征层选择为 5 时，在此次实验中能取到较好的结果. 在卷积特征层的的特征融合部分增强了局部和全局的特征信息，获得检测目标更为丰富的语义信息，从而提升了检测精度. 但更多的特征融合层可能会增加过拟合风险，导致后续的结果检测率下降. 而从最终的检测效果可以看到，改进的算法能检测到较小的目标物体，更具有实用性.

6　结语

本文提出了一种基于特征融合的神经网络目标检测方法，根据在原 Faster-RCNN 的结构基础上结合能实现丰富语义特征的特征金字塔网络结构. 在公开 KITTI 数据集上的对比实验结果表明增加特征金字塔后对目标检测率能有一定程度的提高，特别是对小目标物体的检测效果有一定的提升. 同时，采用关注损失函数替代之前的深度学习的交叉熵损失函数. 最后的实验结果表明对原有的 Faster-RCNN 对比效果有一定的提升. 由于对遮蔽面积较大的目标车辆检测效果还不明显. 后续工作还需要对上述的算法缺陷进行针对性的改进和优化.

【参　考　文　献】

[1] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// **2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. San Diego: IEEE, 2005: 886-893.

[2] MA X Y, GRIMSON W E L. Edge-based rich representation for vehicle classification[C]//**Proc of 10th IEEE International Conference on Computer Vision**. Beijing: IEEE, 2005: 1185-1192.

[3] KAZEMI F M, SAMADI S, POURREZA H R, et al. Vehicle recognition using curvelet transform and SVM[C]// **Proc of 4th International Conference on Information Technology**. Las Vegas, USA: IEEE, 2007: 516-512

[4] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of online learning and an application to boosting[J]. **Journal of Computer and System Sciences**, 1997, 55(1): 119-139.

[5] **宋焕生，张向清，郑宝峰，等. 基于深度学习方法的复杂场景下车辆目标检测[J]. 计算机应用研究**，2018，35(4)：1270-1273.
SONG H S, ZHANG X, ZHEN B F, et al. Vehicle detection based on deep learning in complex scene[J]. **Application Research of Computers**, 2018, 35(4): 1270-1273.

[6] **周俊宇，赵艳明. 卷积神经网络在图像分类和目标检测应用综述[J]. 计算机工程与应用**，2017，53(13)：34-41.
ZHOU J Y, ZHAO Y M. Application of convolution neural network in image classification and object detection[J]. **Computer Engineering and Applications**, 2017, 53(13): 34-41.

[7] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// **IEEE Conference on Computer Vision and Pattern Recognition**. Boston: IEEE, 2005: 886-893.

[8] DOLLAR P, OMRAN M, HOSANG J, et al. Fast feature pyramids for object detection[J]. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, 2014, 36(8): 1532-1545.

[9] GIRSHICK R. Fast R-CNN[C]//**International Conference on Computer Vision**. New York: IEEE Computer Society, 2015: 1440-1448.

[10] GEIGER A. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//**IEEE Conference on Computer Vision and Pattern Recognition**. Boston: IEEE Computer Society, 2012: 3354-3361.

[11] JOSEPH R, SANTOSH D, ROSS G, et al. You only look once: Unified, real-time object detection[C]//**2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Las Vegas, USA: IEEE, 2016: 27-30.

[12] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector [M]//**European Conference on Computer Vision**. Berlin: Springer, 2016: 21-37.

[13] VAPNIK V, GOLOWICH S, SMOLA A, et al. Support vector method for function approximation, regression estimation, and signal processing[J]. **Advances in Neural Information Processing system**, 1997, 9: 281-287.

[14] FELZENSZWALB P F, GIRSHICKR B, MCALLESTER D. Visual object detection with deformable part models [C]// **2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. San Francisco: IEEE, 2010: 13-18.

[15] **王竣，王修晖. 特征融合的多视角步态识别研究[J]. 中国计量大学学报**，2017，28(2)：234-240，268.
WANG J, WANG X H. Research on multi-perspective gait recognition using feature fusion[J]. **Journal of China University of Metrology**, 2017, 28(2): 234-240, 268.

[16] **姚群力，胡显，雷宏. 深度卷积神经网络在目标检测中的研究进展[J]. 计算机工程与应用**，2018，54(17)：1-9.
YAO Q L, HU X, LEI H. Application of deep convolutional neural network in object detection[J]. **Computer Engineering and Applications**, 2018, 54(17): 1-9.