doi:10.3969/j.issn.1001-4616.2019.01.012

基于轻量化 SSD 的车辆及行人检测网络

郑　冬[1],李向群[1],许新征[1,2]

(1.中国矿业大学计算机科学与技术学院,江苏 徐州 221116)
(2.数据科学与智能应用福建省高校重点实验室,福建 漳州 363000)

[**摘要**]　近年来,基于深度学习的目标检测算法发展迅速. 但是由于深度网络规模过大,导致其还不能在嵌入式平台上进行广泛应用. 本文针对 SSD(Single Shot Multi-box Detector)模型的规模进行优化,引入了轻量化卷积神经网络 MobileNetv2,对比了 SSD 和其轻量化版本 SSDLite 的网络结构,在此基础上提出了基于轻量化 SSD 的车辆及行人检测模型 LVP-DN(Lightweight Vehicle and Pedestrian Detection Network). 首先,通过 MobilNetv2 替代 VGG 作为基础网络进行特征提取. 然后,用轻量化的 SSD 版本 SSDLite 替代 SSD,从而达到减少模型大小、加快检测速度的目的. 进一步通过优化默认候选框的比例,提高了网络对行人的检测精度. 最后,在 KITTI 和 PASCAL VOC 数据集上分别对比了不同基础网络、输入图像尺寸及是否使用预训练模型这 3 个因素对网络性能的影响. 实验结果表明,相比其他流行的目标检测模型,本文所提出的车辆及行人检测模型在精度、速度和模型大小等评价标准上取得了较好的效果.

[**关键词**]　目标检测,卷积神经网络,轻量化神经网络,SSD,MobileNetv2
[**中图分类号**]TP193　[**文献标志码**]A　[**文章编号**]1001-4616(2019)01-0073-09

Vehicle and Pedestrian Detection Model Based on Lightweight SSD

Zheng Dong[1], Li Xiangqun[1], Xu Xinzheng[1,2]

(1.School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)
(2.Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou 363000, China)

Abstract: In recent years, the object detection algorithm based on deep learning has developed rapidly. However, it can't be widely used in embedded platforms because the network is too large. This paper optimized the model size of SSD (Single Shot Multi-box Detector) network, introduced the lightweight convolutional neural network—MobileNetv2, analyzed the inverted residual and linear bottleneck structure in MobileNetv2, and compared SSD and its lightweight version—SSDLite. We proposed a lightweight vehicle and pedestrian detection model which named LVP-DN(Lightweight Vehicle and Pedestrian Detection Network). First, the MobilNetv2 was used to instead of VGG as the basic network to perform feature extraction. Then, the SSDLite was used to replace the original structure, in order to reduce the model size and speed up the detection process. It is improved that the accuracy of network detection for pedestrians by optimizing the ratio of the default box. We compared the impact of three factors on network performance on the KITTI and PASCAL VOC datasets. The factors are the input image size, different basic network and whether used the pre-training models. The experimental results show that compared with other popular object detection models, the vehicle and pedestrian detection models proposed in this paper have achieved good results in the evaluation standards such as accuracy, speed, and model size.

Key words: object detection, convolutional neural network, lightweight neural network, SSD, mobileNetv2

随着深度学习的不断发展,卷积神经网络在目标检测[1-2]任务上有着非常卓越的表现,但是当前的研究重点在于如何构建更深的网络以达到提高检测精度的目的. 这样导致了网络模型过于庞大,大部分表现优秀的网络仅能运行在高性能的图形处理器(GPU)上. 为了将深度学习模型更广泛地应用在嵌入式平

收稿日期:2018-08-19.
基金项目:国家自然科学基金(61672522)、数据科学与智能应用福建省高校重点实验室开放课题(D1804).
通讯联系人:许新征,博士,副教授,研究方向:机器学习和模式识别. E-mail:xuxinzh@163.com

台上,如智能车视觉系统、无人驾驶系统,构建轻量化的网络能够有效地降低硬件成本,提高网络的运行效率.

目前,大部分研究主要集中在设计越来越复杂的卷积神经网络以提高目标检测的精度,如 Faster-RCNN[3],R-FCN[4],Mask-RCNN[5]以及其扩展网络为代表的两阶段网络;以及 SSD[6],YOLO[7]及其衍生的网络为代表的一阶段网络.

以 R-CNN[8]为基础的两阶段网络,在精度上达到了很高的标准,但是其检测速度缓慢,网络参数非常庞大;SSD 和 YOLO 实现了端到端训练,在牺牲一定精度的前提下提高了网络的检测速度,但是也没有解决其模型参数过大的问题.

由于计算和内存的限制,这样的网络难以在嵌入式平台中应用. 因此,针对无人驾驶、无人机等实时嵌入式场景下的目标检测任务,本文基于目前优秀的轻量化卷积网络 MobileNetv2[9],对 SSD 网络进行优化,构建了轻量化的车辆及行人检测网络 LVP-DN. LVP-DN 网络通过优化传统的卷积操作,利用分解卷积降低模型参数,提高运算效率. 本文分别在基础网络和检测网络中进行对比实验,验证了分解卷积的有效性,并在行人和车辆检测中加以应用.

1 相关工作

在嵌入式场景,基于深度学习的目标检测的挑战主要有 3 个方面:检测速度、检测精度和模型大小,其研究的重点在于如何设计出能够有效权衡这 3 个指标的深度学习模型.

早期的深度学习方法主要是以 R-CNN 为基础,不断改进优化的两阶段检测网络,这类算法的检测精度最高,但是 R-CNN 存在训练步骤繁琐,时间和内存消耗过大,测试过程过于缓慢的问题,且网络结构较为复杂不适用于嵌入式场景;Fast-RCNN[10]改进了训练步骤,不需要额外保存特征,并提出了 Roi-Pooling,使得可以输入任意尺寸的图片,回归分类器同网络一起训练用 Softmax 取代 SVM 分类器;Faster-RCNN 提出 RPN 网络取代 Roi-Pooling,进一步提高网络的训练速度;Mask-RCNN 则使用 Roi-Align 替换 Roi-Pooling,并引入全卷积网络将目标检测任务扩展到实例分割;RFCN 则是提出了位置敏感分数图,把目标的位置信息融合进 Roi-Pooling. 但是这些两阶段网络仍存在网络模型过于庞大的问题.

目前较为先进的一阶段网络可以达到实时检测的要求. Redmon 等人提出了 YOLO,YOLOv2[11]和 YOLOv3[12],YOLO 把 Faster-RCNN 中对候选框进行分类和识别的两阶段任务进行了结合,大大提高了检测速度,能够在高端 Nvidia Titan X GPU 上实现实时检测. 目前,YOLOv3 使用更好的基础网络 ResNet[13]和特征融合网络 FPN[14]达到了最好的效果,可以通过改变模型大小权衡速度和精度,对于大多数嵌入式设备而言,YOLOv3(237 MB)仍过于庞大. 此外,在嵌入式芯片上运行时,检测速度大大降低. 为了解决这个问题,Redmon[7]在 YOLO 的基础上开发了 Tiny YOLO(60.5 MB). Liu[6]等人提出了 SSD 多尺度特征图的目标检测网络,在 YOLO 的基础上进一步提高了检测精度,是目前较为优秀的目标检测算法. Alexander[15]等人基于 Fire Module[16]结构提出了 Tiny SSD(2.3 MB),在 VOC2007[17]数据集上的平均检测率达到 61.3%,Tiny SSD 在精度和模型大小上没有达到很好的权衡,其检测率还有待提升.

目标检测网络主要由基础网络层和回归检测层组成,基础网络的任务是对输入图像进行特征提取,对网络检测的精度、速度和网络大小都有很大的影响. Francois Chollet[18]等人提出了深度分解卷积(Depthwise Separable Convolution),这项成果极大地缩减了基础网络参数的数量;Andrew G[19]等人提出了轻量化卷积神经网络 MobileNet,MobileNet 使用深度分解卷积和 3×3 卷积核,极大地减少了计算量,同时性能没有明显的降低,相比其他模型压缩方法有很大的优势;Mark Sandler[9]等人对 MobileNet 进行了改进,提出了 MobileNetv2,并改进了 ResNet 的残差结构,提出了倒残差块及线性瓶颈层,网络结构相比 MobileNet 更深,网络参数更少.

本文主要研究轻量化的目标检测网络在嵌入式无人驾驶场景中应用的可能,主要用来识别过往其他车辆、行人、非机动车等目标. 数据集使用目前国际上最大的无人驾驶算法评测数据集 KITTI[20-21]. KITTI 包含市区、乡村和高速公路等场景采集的真实图像数据,部分图像中多达 15 辆车和 30 个行人,还有各种程度的遮挡与截断. 通过对 MobileNetv2 和 SSD 模型进行改进,在稍微降低精度的情况下,减少模型的参数数量,使其达到能够在嵌入式平台实时运行,其检测效果如图 1 所示.

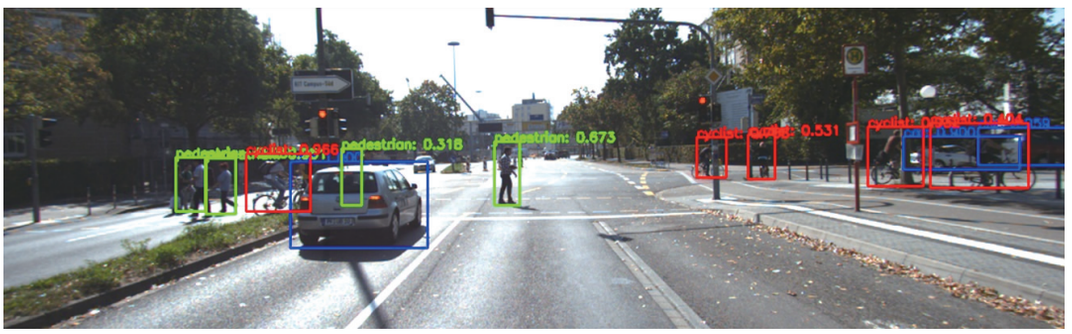


图 1　LVP-DN 检测结果

Fig. 1　The test results of LVP-DN

2　轻量化网络 MobileNetv2

MobileNetv2 是一种轻量化的卷积神经网络,主要是在 MobileNet 网络的基础上进行两点改进. 首先,MobileNetv2 借鉴了 ResNet 的残差结构,在 ResNet 中残差块对输入图像先降维、卷积、再升维,而 MobileNetv2 则是对输入图像先升维、深度分解卷积、再降维,这个过程被称作倒残差. 其次,MobileNetv2 对深度分解卷积进行了改进,在深度分解卷积之前添加了一个 1×1 卷积,保证卷积过程是在高维度进行. 为了不破坏特征,去掉了第二个 1×1 卷积后的激活函数. 这种做法被称为线性瓶颈,可以减少激活函数在低维卷积中对特征的破坏.

2.1　深度分解卷积

深度分解卷积是轻量化网络的主要结构,其主要作用是缩减网络参数,加快网络运行速度. 标准卷积[22]的过程是通过使用多个与输入数据深度相同的卷积核,对其进行卷积运算之后求和得到结果. 深度分解卷积[18]将标准卷积的过程分解成两步:第一步使用单通道卷积核对输入数据的每个通道进行卷积;第二步使用 N 个深度与输入数据相同的 1×1 卷积核对上一步分离的结果进行组合生成新的结果(图 2). 深度分解卷积的输出与标准卷积的输出维度相同,这种分解形式的卷积大幅度减少了模型的计算量.

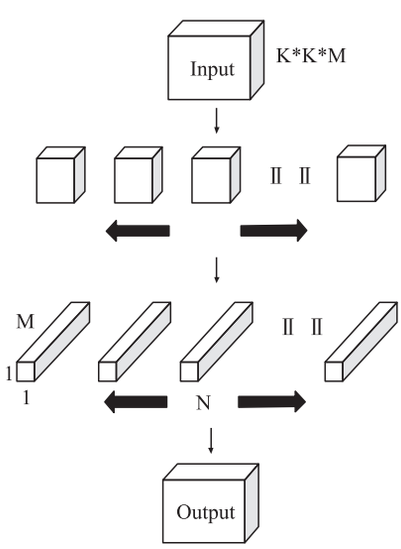


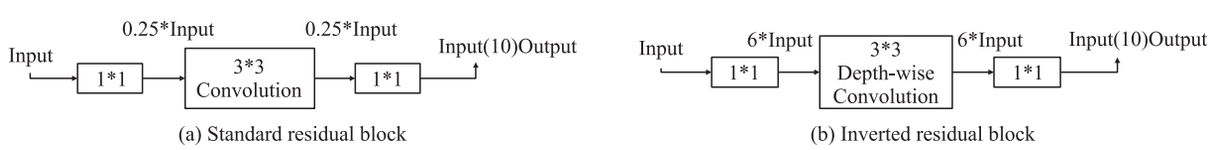
图 2　深度分解卷积

Fig. 2　Depth-wise convolutional filters

2.2　倒残差块

倒残差块[9]是 MobileNetv2 对 ResNet 的残差块的改进. 标准残差块的过程如图 3(a)所示,输入图像先经过一个 1×1 卷积 0.25 倍的降维后,通过标准卷积然后再用 1×1 卷积升维. 倒残差块的执行过程如图 3(b)所示,先经过一个 1×1 卷积进行 6 倍升维后,通过深度分解卷积再降维.

因为深度分解卷积在特征提取的过程中通常期望在高维度的输入中进行,这样能够提高模型的表达能力,所以倒残差模块的作用就是将输入数据变换到高维度后再经过深度分解卷积提取特征.



(a) Standard residual block　　(b) Inverted residual block

图 3　残差卷积模块

Fig. 3　Residual convolution module

2.3　线性瓶颈

由于深度分解卷积没有改变输入数据通道的能力,其提取特征所在的维度取决于上一层输入的数据,

所以 MobileNetv2 在深度分解卷积之前添加了一个 1×1 卷积用于改变输入数据的维度,使得深度分解卷积可以在高维空间提取特征.

通常认为激活函数在高维空间能够有效地增加非线性,而在低维空间则会破坏非线性,深度卷积后的第二个 1×1 卷积用于降维,故线性瓶颈[9]的做法是去掉第二个 1×1 卷积后的激活函数,使用线性激活函数以保持卷积提取到的特征.

2.4　网络结构

MobileNetv2 网络结构的参数如表 1 所示,其中 conv2d 是标准卷积,Inverted Residual 是由倒残差块组成的卷积层,avgpool 是平均池化,通道扩张系数是倒残差块中通道升维的倍数. 该网络共 19 层,中间各层用于提取特征,最后一层用于分类. 倒残差卷积层结构如图 4 所示,其中 1×1 卷积层被称作点卷积.

表 1　MobileNetv2 网络参数

Table 1　MobileNetv2 network parameters

卷积层	通道	重复次数	通道扩张系数	步长	输入分辨率
Conv2d	32	1	–	2	224×224×3
Inverted Residual	16	1	1	1	112×112×32
Inverted Residual	24	2	6	2	112×112×16
Inverted Residual	32	3	6	2	56×56×24
Inverted Residual	64	4	6	2	28×28×32
Inverted Residual	96	3	6	1	28×28×64
Inverted Residual	160	3	6	2	14×14×96
Inverted Residual	320	1	6	1	7×7×160
Conv2d 1×1	1280	1	–	1	7×7×320
Avgpool	–	1	–	–	7×7×1280
Conv2d 1×1	k	–	–	–	1×1×k

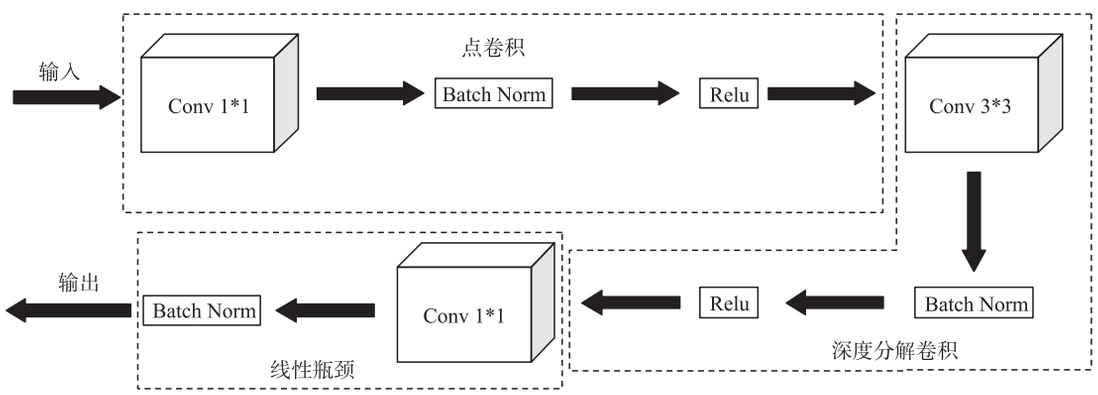


图 4　倒残差卷积层

Fig. 4　Inverted residuals convolution layer

3　SSD 网络

SSD[6]是一种直接预测目标类别和位置的目标检测网络. 与传统的图像金字塔检测方法类似,SSD 在不同尺度的特征图上分别检测,然后将检测结果结合起来使用非极大值抑制算法得到最终的物体框. SSD 的网络结构参数如表 2 所示,网络使用 VGG16[23]作为基础网络,把全连接层替换成卷积层,并添加了 4 个额外的卷积层来输出不同尺度的特征图. 对每一层的结果分别使用 2 个 3×3 大小的卷积核进行卷积,一个用于分类,一个用于回归检测. 在回归任务中,每个特征图输入预测层输出物体的候选框,对输出的结果进行合并后传入损失函数.

在网络的训练阶段,通过将候选框和真值标签进行匹配,划分正负样本. 根据损失函数的值对负样本进行排序,使得正负样本的比例保持在 1∶3. 在预测阶段,得到默认框的偏移及目标类别相应的置信度,通过非极大值抑制算法去除多余的候选框,最后输出物体位置和相应类别的置信度.

SSD 的损失函数由分类损失和位置损失两部分组成,如式(1)所示.

$$L(x,c,l,g)=\frac{1}{N}(L_{\text{conf}}(x,c)+\alpha L_{\text{loc}}(x,l,g)). \tag{1}$$

式中,N 是匹配的候选框个数,α 是权重参数,交叉验证时 α 设为 1.

分类损失函数是 c 类的 Softmax 损失函数,如式(3)所示. 式中 p 表示目标类别,x_{ij}^{p}表示第 j 个匹配到的正样本.

$$\hat{c}_i^p=\exp(\hat{c}_i^p)/\sum_p \exp(\hat{c}_i^p). \tag{2}$$

$$L_{\text{conf}}(x,c)=-\sum_{i\in \text{Pos}}^{\text{N}} x_{ij}^{p}\log(\hat{c}_i^p)-\sum_{i\in Neg}\log(\hat{c}_i^0). \tag{3}$$

位置损失是 Smooth L1[24] 损失函数,如式(4)所示.

$$L_{\text{loc}}(x,l,g)=\sum_{i\in Posm\in\{cx,cy,w,h\}}^{N} x_{ij}^{k}S_{L1}(l_i^m-\hat{g}_j^m), \tag{4}$$

$$\hat{g}_j^{cx}=(\hat{g}_j^{cx}-d_i^{cx})/d_i^w \quad \hat{g}_j^{cy}=(\hat{g}_j^{cy}-d_i^{cy})/d_i^h,$$

$$\hat{g}_j^w=\log(g_j^w/d_i^w) \quad \hat{g}_j^h=\log(g_j^h/d_i^h).$$

式中,$\hat{g}_j^{cx}$,$\hat{g}_j^{cy}$,$\hat{g}_j^w$,$\hat{g}_j^h$ 分别表示物体的真值标签(x,c,w,h)和物体的高度 h,宽度 w, d_i^{cx},d_i^{cy},d_i^w,d_i^h 表示物体的候选框坐标以及高度和宽度. 其网络结构参数如表 2 所示.

表 2　SDD 网络参数

Table 2　SSD network parameters

卷积层	卷积核	通道	重复次数	步长	输出分辨率	卷积层	卷积核	通道	重复次数	步长	输出分辨率
Conv1	3×3	64	2	1	300×300	Fc6	3×3	1 024	1	1	19×19
Pool1	2×2	1	1	2	150×150	Fc7	1×1	1 024	1	1	19×19
Conv2	3×3	128	2	1	150×150	Conv6_1	1×1	256	1	1	19×19
Pool2	2×2	1	1	2	75×75	Conv6_2	3×3	512	1	2	10×10
Conv3	3×3	128	3	1	75×75	Conv7_1	1×1	128	1	1	10×10
Pool3	2×2	1	1	2	38×38	Conv7_2	3×3	256	1	2	5×5
Conv4	3×3	512	3	1	38×38	Conv8_1	1×1	128	1	1	5×5
Pool4	2×2	1	1	2	19×19	Conv8_2	3×3	256	1	1	3×3
Conv5	3×3	512	3	1	19×19	Conv9_1	1×1	128	1	1	3×3
Pool5	3×3	1	1	1	19×19	Conv9_2	3×3	256	1	1	1×1

SSD 不管是检测速度还是精度上都是目前表现非常优秀的 CNN 网络,但是由于传统卷积操作的限制,其参数量过于庞大. Tiny SSD 的模型参数虽然很小,但是在精度上没有达到很好的标准. 因此,本文基于 MobileNetv2,对 SDD 网络进行轻量化改进,得到基于轻量化 SSD 的车辆行人检测模型 LVP-DN.

4　轻量化的车辆及行人检测网络

SSDLite[9] 网络是使用倒残差卷积(图 4)优化的轻量化 SSD 网络. 与 SSD 相比,SSDLite 使用倒残差卷积层替换了 SSD 回归检测层中的所有标准卷积层,减少了网络预测过程的计算量. SSDLite 使用 MobileNetv2 作为基础网络,相比其他目标检测算法,SSDLite 在检测速度、精度和模型大小上都有很大的优越性.

对于本文所使用的数据集和应用场景,提出了对 SSDLite 的改进方法,通过改变默认候选框的比例来提高目标的检测率,构建了轻量化的车辆行人检测模型 LVP-DN.

LVP-DN 的网络结构参数如表 3 所示,去掉了 MobileNetv2 的全连接层和池化层,并添加了用倒残差块组成的额外的四层卷积层(Extras IRblock_1~4),将其中 6 层卷积层生成的特征图通过 2 个 3×3 卷积核,其中一个输出目标分类结果,另一个结果输入到回归检测层中预测目标的候选框(6 层卷积层包括 Inverted Residual 32×32, Inverted Residual 16×16, Extras IRblock_1 8×8, Extras IRblock_2 4×4, Extras IRblock_3 2×2, Extras IRblock_4 1×1). 与 SSD 网络相比,LVP-DN 将输入图像分辨率扩大到了 512×512. 本文通过实验对比了不同输入图像分辨率对网络性能的影响,实验结果表明,这种方法可以有效提高网络的检测精度,但是存在增加网络训练时间的问题.

表 3　LVP-DN 网络参数

Table 3　LVP-DN network parameters

卷积层	通道	重复次数	步长	输出分辨率	卷积层	通道	重复次数	步长	输出分辨率
Conv2D	32	1	2	256×256	Inverted Residual	160	3	2	16×16
Inverted Residual	16	1	1	256×256	Inverted Residual	320	1	1	16×16
Inverted Residual	24	2	2	128×128	Extras IRblock_1	512	1	2	8×8
Inverted Residual	32	3	2	64×64	Extras IRblock_2	256	1	2	4×4
Inverted Residual	64	4	2	32×32	Extras IRblock_3	256	1	2	2×2
Inverted Residual	96	3	1	32×32	Extras IRblock_4	128	1	2	1×1

LVP-DN 的检测过程如图 5 所示,检测部分使用的特征图尺寸和默认候选框的设置如表 4 所示,原 SSD 网络默认候选框的形状对行人这类小目标并不敏感,本文根据行人的形状,在分辨率较高的特征图中额外添加了{3,1/3}默认候选框的纵横比. 对于 m 个特征图,默认候选框的大小按式(4)计算:

$$s_k=s_{\min}+\frac{s_{\max}-s_{\min}}{m-1}(k-1)\quad k\in[1,m]. \tag{5}$$

表 4　特征图尺寸及纵横比

Table 4　Feature map scale and aspect ratio

特征图尺寸	纵横比	默认候选框个数	预测框数量
32×32	{1,2,3,1/2,1/3}	6	6144
16×16	{1,2,3,1/2,1/3}	6	1536
8×8	{1,2,3,1/2,1/3}	6	384
4×4	{1,2,3,1/2,1/3}	6	96
2×2	{1,2,1/2}	4	16
1×1	{1,2,1/2}	4	4

式中,$s_{\min}$是特征图尺寸最小的候选框尺寸,设为 0.2;$s_{\max}$是特征图尺寸最大的候选框尺寸,设为 0.9. 每个候选框的宽为 $w_k^a=s_k\sqrt{a_r}$,高为 $h_k^a=s_k/\sqrt{a_r}$. 当纵横比为 1 时会添加额外的候选框 $s_k'=\sqrt{s_ks_k+1}$.

检测阶段对输入特征图的每个像素生成该层设置的 k 个预测框,然后将各层特征图生成的预测框(共 8 180 个)进行非极大值抑制,最后生成每个物体的物体框和分类置信度.

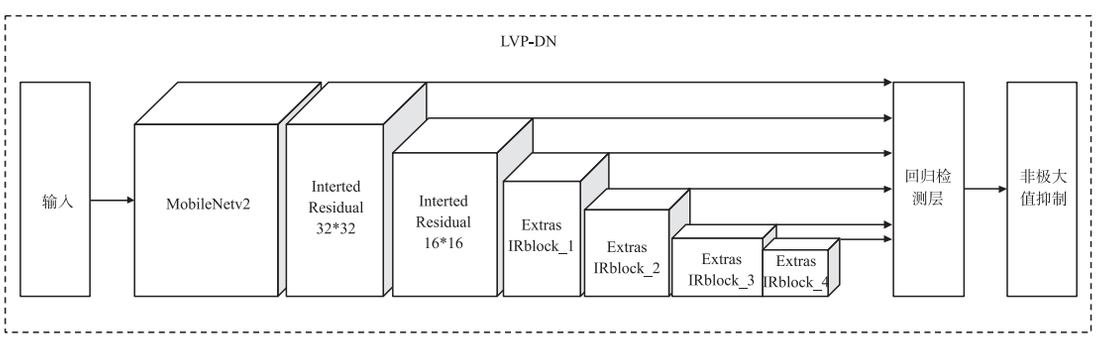


图 5　LVP-DN 检测过程

Fig. 5　LVP-DN detection process

在原始 SSD 中,输入图像尺寸为 300×300,检测过程使用的最大特征图尺寸(38×38)经过 3 次采样得到. 根据文献[25]的结果,引入 75×75 尺寸的特征图可以有效地提高网络的检测精度,但是牺牲了过多的检测速度. LVP-DN 将输入图像尺寸扩大到 512×512,使用的最大特征图尺寸为(32×32). 相比 SSD,LVP-DN 通过扩大输入分辨率增加特征图语义信息,弥补了由于网络缩减导致的语义信息的损失.

5　实验与分析

5.1　实验准备

本文实验平台为 Intel® Core I7-8700k 处理器,NVIDIA GTX 1080 8G 显存,软件环境为 Ubuntu16.04,CUDA8.0,OpenCV3.0 和 Pytorch 深度学习框架,使用 KITTI 数据集对模型进行训练和评估.

本文把 KITTI 数据集的 car,van,truck 三类数据样本合并为 car 类,pedestrian 和 person sitting 合并为 pedestrian 类,数据集由 car,pedestrian,cyclist 和 background 四类组成. 在 7 400 张带标签图片中,5 940 张作为训练集,740 张作为训练验证集,740 张作为测试集.

对比其他流行算法时,主要在 PASCAL VOC2007 和 2012 数据集集上进行评估. PASCAL VOC 数据集包括 20 个类别. 本文把 VOC2007+2012 的训练数据集一起作为训练集,VOC2007 的测试集用来测试. 数

据增强策略采用与 SSD 中相似的方法:

(1)随机剪裁:对输入图随机采样,使裁剪出的部分与检测的目标重叠率为{0.1,0.3,0.5,0.7,0.9},采样后的图片尺寸设为 512×512;

(2)对输入图片采用随机水平翻转.

5.2　评价标准

根据应用的侧重点不同,目标检测算法有很多评价标准. 本文使用检测精度、检测效率和模型大小对目标检测模型进行评估.

采用 mAP(mean average precision)来评估目标检测模型在数据集上的所有类别的性能好坏,使用每秒帧检测数(frames per second,FPS)来评估检测效率,使用 MB(MByte)评估模型的大小. 通过实验权衡这 3 个性能指标,得到适用于嵌入式平台运行的轻量级目标模型.

5.3　实验结果与分析

(1)本节实验通过控制变量,分别使用不同的基础网络和 SSD 结构,对比不同结构组合模型的效果. 基础网络使用 MobileNet 和 MobileNetv2,分别与 SSD 和 SSDLite 结构进行组合. 输入图像尺寸为 300×300,在 KITTI 数据集上进行实验.

实验结果如表 5 所示,对比第 1、2 和 3、4 行结果可知,以 MobileNetv2 作为基础网络可以有效减少网络模型的参数(模型减少 8.2~8.9 MB),加快模型的检测速度(加速 7FPS),而检测精度基本没有影响. 对比第 1、3 和 2、4 行的结果可以看到,检测网络采用 SSDLite 结构时各评价标准均有提升(模型减少 4.2~4.9 MB,检测速度提升 13FPS,mAP 提升 1.3%~1.6%). 综上,本文提出 LVP-DN 网络采用的 MobileNetv2+SSDLite 结构的性能取得了最优效果.

(2)本节通过不同尺度的输入图像,权衡检测精度、检测速度和模型大小.

实验结果如表 6 所示,不同输入尺寸的图像基本不会影响网络模型的大小与检测速度. 而提高输入图像分辨率可以有效地提高网络的检测精度,其原因是随着图像分辨率的增加,图像中目标的尺寸也相应增大,提高了网络的检测精度,这也与文献[25-26]的实验结果一致.

表 5　不同基础网络和检测结构的实验结果

Table 5　The results on different base networks and detection structures

基础网络	模型大小/MB	速度/FPS	精度/mAP
MobileNet+SSD	26.3	43	55.1%
MobileNetv2+SSD	18.1	50	55.3%
MobileNet+SSDLite	22.1	56	56.7%
LVP-DN(300)	13.2	63	56.6%

表 6　不同输入尺度的 LVP-DN 网络实验结果

Table 6　The results on LVP-DN networks with different input scales

网络模型	模型大小/MB	速度/FPS	精度/mAP
LVP-DN(300)	13.2	63	56.6%
LVP-DN(512)	13.2	63	75.5%
LVP-DN(800)	13.2	63	83.6%

(3)为了对比目前其他流行的轻量化目标检测网络,本节在 PASCAL VOC2007+2012 数据集上对模型进行训练,VOC2007 测试集上进行评估. 其中,Tiny SSD 和 Tiny YOLO 是目前较为流行的两种轻量化目标检测网络.

实验结果如表 7 所示,本文提出的网络结构 MobileNetv2+SSDLite300 在模型大小上相比 SSD 和 Tiny YOLO 有很大的优越性,在检测精度上要优于 Tiny SSD 和 Tiny YOLO. Tiny SSD 的模型参数最小,但是其检测精度较低. 权衡这两个检测标准,本文提出的网络结构性能最好.

表 7　不同目标检测网络对比

Table 7　Different target detection network comparison

网络模型	数据集	模型大小/MB	速度/FPS
SSD300	VOC0712	105.2	77.4%
Tiny YOLO	VOC0712	60.5	57.1%
Tiny SSD	VOC0712	2.3	61.3%
MobileNet+SSDLite300	VOC0712	31.5	72.7%
LVP-DN(300)	VOC0712	18.5	73.2%

(4)前文基于 SSDLite 网络训练的车辆及行人检测模型均使用了在 PASCAL VOC2007+2012 数据集预训练的模型. 为了验证预训练模型对网络训练收敛速度的影响,本节实验在 KITTI 数据上,对比了是否使用预训练模型对网络收敛速度的影响,实验共执行 300 个 step,每个 step 训练整个训练集. 结果损失图如图 6 所示,损失值由 conf_loss(图 6(a))分类损失和 loc_loss(图 6(b))回归检测损失组成. 结果表明,使用在 PASCAL VOC 数据集上预训练的网络模型可以加快网络收敛速度,降低网络的损失值,增加网络的检测精度.

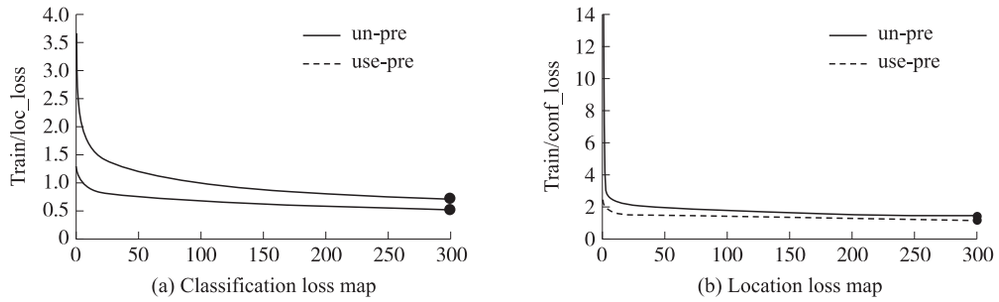


图 6　网络训练损失值图

Fig. 6　Network training loss map

6　结束语

本文针对嵌入式无人驾驶场景提出了轻量化车辆行人检测模型,实现了基于深度学习的轻量化实时目标检测模型. 模型基于 SSDLite[9] 在 PASCAL VOC 数据集训练的基础上,使用预训练网络在 KITTI 数据集上再训练. 实验结果表明,本文提出的轻量化车辆行人检测网络 LVP-DN 在输入尺度为 512 * 512 时得到的训练结果在各项评估指标中表现更为优秀. 同时与其他轻量化目标检测模型相比,本文所提出的网络结构,综合精度、速度和模型大小,其实验结果也是最好的. 在未来的工作中将会在嵌入式平台中进行进一步的实时测试.

[参考文献]

[1] SZEGEDY C,TOSHEV A,ERHAN D. Deep neural networks for object detection[C]//International Conference on Neural Information Processing Systems. USA:MIT Press,2013,26:2553-2561.

[2] SERMANET P,EIGEN D,ZHANG X,et al. OverFeat:integrated recognition,localization and detection using convolutional networks[J]. Eprint Arxiv,2013:1312.6229.

[3] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:towards real-time object detection with region proposal networks[C]//International Conference on Neural Information Processing Systems. Canada:MIT Press,2015:91-99.

[4] DAI J,LI Y,HE K,et al. R-FCN:Object detection via region-based fully convolutional networks[J]. Eprint Arxiv,2016:1605.06409.

[5] HE K,GKIOXARI G,DOLLáR P,et al. Mask R-CNN[C]//IEEE International Conference on Computer Vision. Italy:IEEE,2017:2980-2988.

[6] LIU W,ANGUELOV D,ERHAN D,et al. SSD:single shot multibox detector[C]//Computer Vision-ECCV 2016. Amsterdam,the Netherlands:Springer International Publishing,2016:21-37.

[7] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:unified,real-time object detection[C]//Computer Vision and Pattern Recognition. USA:IEEE,2016:779-788.

[8] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. USA:IEEE Computer Society,2014:580-587.

[9] SANDLER M,HOWARD A,ZHU M,et al. MobileNetV2:inverted residuals and linear bottlenecks[J]. Eprint Arxiv,2018.

[10] GIRSHICK R. Fast R-CNN[C]//IEEE International Conference on Computer Vision. USA:IEEE Computer Society,2015:1440-1448.

[11] REDMON J,FARHADI A. YOLO9000:better,faster,stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition. Italy:IEEE Computer Society,2017:6517-6525.

[12] REDMON J,FARHADI A. YOLOv3:an incremental improvement[J]. Eprint Arxiv,2018:104.02767.

[13] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. USA:IEEE Computer Society,2016:770-778.

[14] LIN T Y,DOLLAR P,GIRSHICK R,et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Italy:IEEE Computer Society,2017:936-944.

[15] WONG A,SHAFIEE M J,LI F,et al. Tiny SSD:a tiny single-shot detection deep convolutional neural network for real-time

embedded object detection[C]//Conference on Computer and Robot Vision. Foronto:IEEE,2018(15):95-101.

[16] IANDOLA F N,HAN S,MOSKEWICZ M W,et al. SqueezeNet:AlexNet-level accuracy with 50×fewer parameters and<0.5 MB model size[J]. Eprint Arxiv,2016:1602.07360.

[17] EVERINGHAM M. The PASCAL visual object classes challenge[J]. Lecture notes in computer science,2005,111(1):98-136.

[18] CHOLLET F. Xception:deep learning with depthwise separable convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition. Italy:IEEE Computer Society,2017:1800-1807.

[19] HOWARD A G,ZHU M,CHEN B,et al. MobileNets:efficient convolutional neural networks for mobile vision applications[J]. Eprint Arxiv,2017.

[20] GEIGER A,LENZ P,URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//IEEE Conference on Computer Vision and Pattern Recognition. USA:IEEE Computer Society,2012:3354-3361.

[21] GEIGER A,LENZ P,STILLER C,et al. Vision meets robotics:the KITTI dataset[J]. International journal of robotics research,2013,32(11):1231-1237.

[22] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[23] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//International Conference on Leoorning Representations. USA:IEEE,2014.

[24] GIRSHICK R. Fast R-CNN[C]//IEEE International Conference on Computer Vision. USA:IEEE Computer Society,2015:1440-1448.

[25] KIM H,LEE Y,YIM B,et al. On-road object detection using deep neural network[C]//IEEE International Conference on Consumer Electronics-Asia. Korea:IEEE,2016:1-4.

[26] HUANG J,GUADARRAMA S,MURPHY K,et al. Speed/accuracy trade-offs for modern convolutional object detectors[C]//IEEE International Conference on Computer Vision. USA:IEEE Computer Society,2016:3296-3297.

[责任编辑:黄　敏]