基于 YOLO v4 的夜间车辆检测模型轻量化研究

徐 丽,刘星星,屈立成
(长安大学 信息工程学院,陕西 西安 710000)

摘 要:针对夜间车辆检测模型的实时性要求,以 YOLO v4 模型为基础,将主干特征提取网络更改为灵活性强且易于实现的 MobileNet V2,并将加强特征提取网络里面的普通卷积全部更改为深度可分离卷积,同时模型给每个通道引入缩放因子,并与该通道输入相乘。然后将缩放因子正则项和权重损失函数联合进行稀疏正则化训练,此时选择较小的缩放因子进行通道剪枝,剪枝后模型的部分通道缺失,检测性能会降低,因此通过模型微调来弥补精度损失,并经过性能评估后再进行修剪迭代。最后得到一个轻量化的车辆检测模型,使其检测速度更快,更能满足夜间车辆检测的实时性需求。经过在 UA-DETRAC 数据集的实验分析可知:轻量化夜间车辆检测模型的检测精度可达 98.29%,同时每秒处理帧数高达 42 帧图像。

关键词:夜间车辆检测;YOLO v4;MobileNet;深度可分离卷积;通道剪枝
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)03-0084-06
doi:10.3969/j.issn.1673-629X.2022.03.014

Research on Lightweight of Night Vehicle Detection Model Based on YOLO v4

XU Li,LIU Xing-xing,QU Li-cheng
(School of Information Engineering,Chang'an University,Xi'an 710000,China)

Abstract:In response to the real-time requirements of the night vehicle detection model,based on the YOLO v4 model,the backbone feature extraction network is changed to MobileNet V2,which is flexible and easy to implement,and changes all the ordinary convolutions in the enhanced feature extraction network to deep separable convolutions. At the same time,the model introduces a scaling factor to each channel and multiplies it with the channel input. Then the scaling factor regular term and the weight loss function are combined for sparse regularization training. At this time,a smaller scaling factor is selected for channel pruning. After pruning,some channels of the model are missing,and the detection performance will be reduced,so the model is fine-tuned to compensate for the loss of accuracy,and after the performance evaluation,the pruning iterations are performed. Finally,a lightweight vehicle detection model is obtained,which makes the detection speed faster and can better meet the real-time requirements of night vehicle detection. The experimental analysis on the UA-DETRAC data set shows that the detection accuracy of the lightweight night vehicle detection model can reach 98.29%,and the number of frames per second can be as high as 42 images.

Key words:night vehicle detection;YOLO v4;MobileNet;depth separable convolution;channel pruning

0 引 言

在 YOLO v3[1]算法结构的基础上 YOLO v4[2]融合了最近几年在深度神经网络中有明显优势的算法模型思想和训练技巧,达到了很高的检测精度。主干特征提取网络在 Darknet 基础上,结合 CSPNet 算法的思想,构成 CSPDarknet53,并使用 Mish[3]激活函数、Dropblock 正则化方式[4];加强特征提取网络由 YOLOv3 采用的特征金字塔网络(FPN)修改为加入空间金字塔池化层[5](SPP)和路径聚合网络[6](PANet)的组合;预测网络使用 YOLO v3 中的 YOLO-Head,预测框筛选使用 DIOU_NMS。最终形成"CSPDarknet53+SPP-PANet+YOLO-Head"的模型结构。

YOLO v4 网络相较于传统机器学习算法有着明显的优势,其优势主要体现在更多的参数和更大更深的网络模型。复杂的模型结构能够增强网络的非线性拟合能力,并且海量数据能够增强模型的泛化能力[7]。然而在实际使用时,YOLO v4 网络模型需要面临计算资源的限制及速度的严格要求,故该文主要对基于

收稿日期:2021-04-24 **修回日期**:2021-08-25
基金项目:陕西省自然科学基础研究计划资助项目(2020JM-258)
作者简介:徐 丽(1977-),女,博士,副教授,研究方向为图像处理;通讯作者:刘星星(1995-),女,硕士,研究方向为图像处理。

YOLO v4 的夜间车辆检测模型进行轻量化,使其在检测精度降低少许的基础上最大程度地缩减参数量及模型体积,使检测速度更快,更能满足夜间车辆检测的实时性需求。

1　轻量化夜间车辆检测算法

夜间车辆检测模型的轻量化过程如图 1 所示。

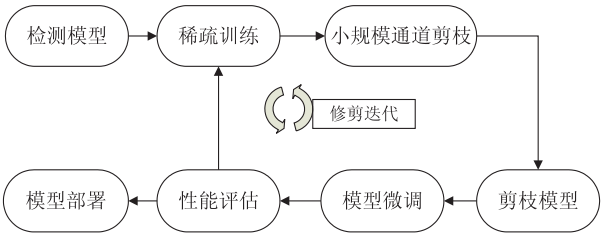


图 1　夜间车辆检测模型轻量化过程

1.1　算法流程

轻量化夜间车辆检测的问题也是转化成一个回归问题求解,即将整张图像用作网络输入,直接在输出层回归边框位置及其所属类别,流程如图 2 所示。

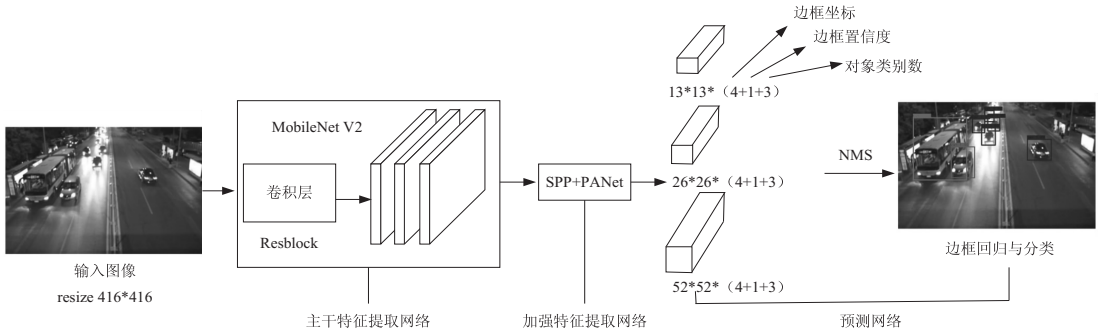


图 2　夜间车辆检测算法过程

(1)将输入的原始图像 resize 到 416 * 416;(2)利用更改为 MobileNet V2 的主干特征提取网络,使得网络参数量与计算量较少,可以更快地提取待检测目标的有效特征;(3)在加强特征提取网络以 SPP+PANet 的模型结构分别进行最大池化、卷积、上下采样、特征层融合的操作,实现将多层语义特征以及多尺度感受野进行融合,同时将加强特征提取网络里面的 3 × 3 普通卷积全部更改为深度可分离卷积;(4)在预测网络利用加强特征提取网络处理后的 3 个特征层进行结果预测,判断先验框里面是否存在待检测对象以及对象类别,并采用非极大值抑制 NMS 处理和先验框调整来确定最终的预测框,得到车辆种类及坐标信息。

训练前,对数据集进行白平衡处理[8]以减少路灯颜色干扰进而增强图像画质,并采用 K-means++算法针对数据集聚类出合适的先验框,其中距离度量使用交并比距离。训练时,首先采用 Mosaic 算法进行数据增强,有利于区别检测物体的背景、前景,丰富了数据集中的图像特征,尤其是随机缩放出现了很多小目标车辆,增强了网络对小目标车辆的特征提取能力。最后采用构建的夜间车辆检测模型对数据集进行稀疏正则化训练。

1.2　特征提取网络

YOLOv4 夜间车辆检测算法的特征提取网络包括主干特征提取网络和加强特征提取网络两部分,相比预测网络参数量在这两部分较多。而 MobileNet 即 MobileNet V1 是 2017 年谷歌提出的一种轻量化的 CNN 模型,目前共有 3 个版本,其结构是以深度可分离卷积为基础的,除第一层是普通卷积,其余卷积层都由深度可分离卷积组成,计算量与参数量非常少,简单的网络定义,就可获得检测性能优异的网络,因此常作为轻量化的特征提取网络,故将主干特征提取网络 CSPDarknet53 分别更改为三个版本的 MobileNet,得到的模型参数量和体积如表 1 所示。

表 1　不同主干特征提取网络的模型参数量和体积

主干特征提取网络	参数量/M	模型体积/MB
CSPDarknet53	64.0	245.0
MobileNet V1	40.6	155.6
MobileNet V2	38.7	148.0
MobileNet V3	39.6	151.8

从表 1 可以看出,主干特征提取网络使用了 MobileNet 后其参数量和模型体积还是相当庞大,而夜间车辆检测模型的大多数参数在加强特征提取网络居多,其参数量主要集中在 3 × 3 的卷积上面,故尝试将加强特征提取网络里面的 3 × 3 普通卷积全部更改为在模型轻量化方面有优异特性的深度可分离卷积,此时可得到模型参数量和体积的情况如表 2 所示。

表 2　修改 PANet 后不同主干网络的模型参数量和体积

主干特征提取网络	参数量/M	模型体积/MB
MobileNet V1	12.3	47.6
MobileNet V2	10.5	40.2
MobileNet V3	11.4	43.6

从表 2 可以看出,在主干特征提取网络使用 MobileNet 的基础上,同时把加强特征提取网络里面的

普通卷积全部更改为深度可分离卷积,此时参数量和模型体积明显缩减了。同时可以看到 MobileNet V2 的模型参数量和体积少于 MobileNet V1 和 MobileNet V3,故决定将 CSPDarknet53 更改为 MobileNet V2 作为模型的主干特征提取网络,同时将加强特征提取网络里面的 3 × 3 普通卷积全部更改为深度可分离卷积。

1.3 稀疏训练

深层模型的通道稀疏性有利于通道剪枝,且能获得有可能会被剪枝的重要性低的通道数量。按照通道修剪原理,稀疏训练过程中得为每个卷积层的每个通道分配一个缩放因子 γ(γ 作为通道剪枝的重要依据[9],将其正则项和权重损失函数联合优化,网络本身就可识别重要性低的通道并将其安全剪除,而不会影响其泛化性能),将其与该通道的输入相乘,使得每层的各通道提取的特征产生不一样的作用效果,其中缩放因子的绝对值就表示通道重要性。因为夜间车辆检测网络中每个卷积层后都增加了批量归一化(BN)层,BN 层的公式如(1)所示:

$$y = \gamma \times \frac{x - \bar{x}}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{1}$$

其中,γ 为可训练的比例因子,β 为偏差,$\bar{x}$ 和 σ^2 分别为输入特征的均方和误差。可以看出 BN 层存在将标准线性激活转换为各种尺度的可能性。故为了不给模型增加额外的参数,该文的模型通道剪枝的缩放因子使用 BN 层中的可训练比例因子 γ 参数。

L1 正则化是以 L1 范数为基础,即把参数的 L1 范数和项与目标函数相加,如公式(2):

$$C = C_0 + \frac{\lambda}{n}\sum_{w}|w| \tag{2}$$

其中,C_0 是原始的代价函数,n 是样本数,λ 是权衡正则项与 C_0 项比重的正则项系数,后面一项即 L1 正则项。计算时 w 梯度的变化如公式(3)所示:

$$\frac{\partial C}{\partial w} = \frac{\partial C_0}{\partial w} + \frac{\lambda}{n}\mathrm{sgn}(w) \tag{3}$$

其中,sgn 是符号函数,那么使用下面公式(4)对参数进行更新:

$$w = w + \alpha\frac{\partial C_0}{\partial w} + \beta\frac{\lambda}{n}\mathrm{sgn}(w) \tag{4}$$

从上式可以看出,L1 正则项是将原本接近零(即 $|w| \approx 0$)的那些参数 w 移到零,从而使某些参数为零,这意味着 L1 正则化能够发挥选择通道的作用,使得重要性降低的通道 γ 尽可能靠近 0。故为有效地区分通道的重要与否,该文通过 L1 正则化来进行通道的稀疏训练。与此同时,依据网络中参数的变化,引入与 γ 有关的惩罚项,稀疏训练损失函数如公式(5)所示:

$$L = \sum_{(x,y)} l(f(x,w),y) + \lambda\sum_{\gamma\in\Gamma} g(\gamma) \tag{5}$$

其中,(x,y) 表示网络的输入、输出,w 表示权重,第一项表示原始网络训练时的损失,第二项表示关于 γ 的 L1 正则,$g(\gamma)$ 表示缩放因子的稀疏诱导惩罚,λ 表示超参数,作用是将正常训练损失和通道缩放因子惩罚项损失的比例进行平衡,Γ 表示缩放因子 γ 的取值集合。通常 L1 正则化即 $g(\gamma) = |\gamma|$ 为通道的重要性,被广泛应用于实现稀疏化。

1.4 通道剪枝

模型剪枝是复杂模型轻量化时一种广泛使用的方法。最早是用来去除模型中的多余参数,减小复杂度,进而提升泛化性能。模型剪枝可以从不同剪枝粒度上实现,例如权重级、通道级或层级。细粒度剪枝(如权重级)比较灵活常用,而且剪枝率高,但它一般得使用特殊设备来辅助实现。相反粗粒度剪枝并不需要特殊设备,但是要对某些整层进行修剪,所以并不灵活,而且只有当深度足够深时,例如多余 50 层的网络[10],剪枝才是有效果的。相比之下通道级别的剪枝能够很好地权衡灵活性与易于实现程度,且能够应用于任何 CNN 或全连接的网络,由此生成一个轻量化的网络,使得传统的 CNN 能够在任何平台上快速有效地运行。

实现通道级剪枝需要修剪所有与通道相关的输入和输出,但是对已预训练好的模型做通道剪枝效率不高,因为不可能通道所有输入或输出的权重都有近似 0 的值。实验结果表明,在保证相对准确率时,对预训练好的 ResNet 做通道剪枝,仅仅可以减少 10% 的参数量[11]。

通道剪枝的实现流程如图 3 所示。

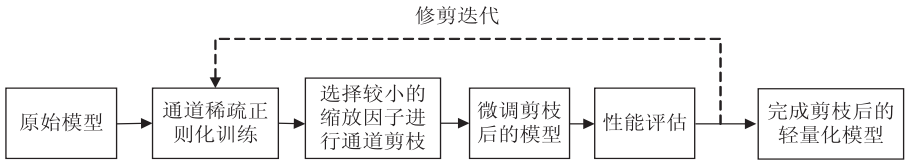
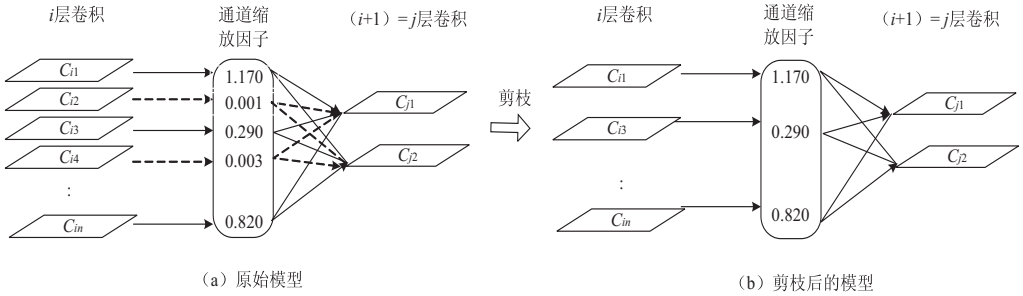


图 3 通道剪枝流程

(1)对每个通道引入缩放因子 γ,与该通道输入相乘;(2)将 γ 正则项和权重损失函数联合进行稀疏正则化训练;(3)剪去较小缩放因子的通道;(4)微调剪枝后的模型;(5)经过性能评估后再进行修剪迭代,

最后得到一个轻量化的车辆检测模型。

在通道进行稀疏正则化训练后,可以得到一个许多缩放因子接近于零的模型,然后通过移除它们的所有进出连接和相应的权重来修剪缩放因子接近于 0 的通道。剪枝比例是所有缩放因子值的一个比例,由设定的全局阈值 $\hat{\gamma}_g$ 来决定,剪去小于该阈值的通道,即删除所有该通道上的输入输出连接和卷积核[12]。在代码中设定百分比 $n_{\text{prune}-g}$(即剪枝率)可确定阈值 $\hat{\gamma}_g$,剪去比例为 $n_{\text{prune}-g}$ 的通道,即阈值 $\hat{\gamma}_g$ 为动态值。例如要剪去原始网络中 60% 的通道,就选由小到大排序中前 60% 位置的缩放因子作为阈值,由此可获得一个内存占用率少且体积较小的轻量化网络。在剪枝过程中,为防止部分用于特征提取的卷积层被过度修剪,使得模型检测效果欠拟合,故添加卷积层的安全阈值即局部阈值 $\hat{\gamma}_p$,在第 i 个 CNN 层中,只剪去缩放因子低于 $\hat{\gamma}_p$ 的通道,与全局阈值 $\hat{\gamma}_g$ 的确定方式相同,$\hat{\gamma}_p$ 由局部百分比 $n_{\text{prune}-p}$ 计算确定。故为保证模型的结构完整性,在剪枝过程中只去除比例因子同时小于 $\hat{\gamma}_g$ 和 $\hat{\gamma}_p$ 的通道。



(a) 原始模型　　(b) 剪枝后的模型

图 4　通道剪枝

稀疏训练后获得的通道缩放因子如图 4(a)所示,当前的缩放因子 γ 是稀疏分布的,第 i 层卷积层中的通道 2 和 4 中缩放因子近似为 0,这代表着经过训练的模型认为该通道所提取的特征对目标的识别和分类基本没有任何作用,通过剪枝操作后,会删去该通道,里面的卷积核及其参数将不会被保存,达到轻量化模型的效果。剪枝之后的轻量化模型如图 4(b)所示。

1.5　模型微调

在细粒度目标检测任务中,模型检测性能通常会受通道剪枝的影响,当模型修剪比例过高时,会使模型的一些通道缺失,因此模型精度会降低,但是可以通过后续的模型微调来弥补精度损失。模型微调的过程实际上是对剪枝后的模型进行再训练[13],需要对剪枝后的模型权重参数 $\bar{W}$ 进行微调得到性能良好的权重参数 W',计算如公式(6)所示,其中 argmin 表示使目标函数取最小值时的变量值。在微调过程中的学习率 learning_rate 和迭代次数 epoch 均比预训练时的小。

$$W' = \underset{\bar{W}}{\operatorname{argmin}} \operatorname{Loss}(\bar{W}) \tag{6}$$

2　实验结果及分析

2.1　实验环境与参数设置

实验的硬件配置为 Intel(R) Core(TM) i7-8700K CPU @ 3.70 GHz,64 GB 内存与 NVIDIA GeForce GTX 1080 Ti 独立显卡,软件环境为 Windows 10 系统,算法基于 Keras 框架,CUDA 版本为 10.0。在实验中,选择 Adam 作为优化器,early_stopping 用于设定早停,val_loss 多次不下降自动结束训练,表示模型基本收敛,batch_size = 16,学习率设置为 0.001,并在 110 epoch 后调整为 0.000 1,共训练 220 epoch。将实验里面的数据划分为 68% 的训练集、12% 的验证集与 20% 的测试集,其中训练验证集占到 80%。

2.2　数据集介绍与评价指标

使用的数据集是 UA-DETRAC 数据集。该数据集是由佳能摄像机在天津和北京 24 个不同地点拍摄的 10 小时的视频组成,图像分辨率为 960 × 540 像素。该数据集将车辆分为四类,即 car、bus、van 和 others,由于只有车辆被检测,故只取前三类 car、bus、van。因为要研究的是夜间场景下的车辆检测算法,所以选取了 16 000 张夜间图像作为该文的研究对象。

常用的车辆检测算法评价指标有:精确率(Precision,P)、召回率(Recall,R)、平均精度 AP、全部类别目标平均精度 mAP、每秒处理帧数 FPS 等。

2.3　模型轻量化实验结果分析

基于 YOLO v4 检测模型,将 CSPDarknet53 更改为 MobileNet V2 作为模型的主干特征提取网络,并将加强特征提取网络里面的普通卷积全部更改为深度可分离卷积后的车辆检测模型(M1)进行训练。对模型 M1 进行 220 次迭代的稀疏性训练,稀疏因子 λ 采用默认值 0.000 1,剪枝率设为 50%,可以发现剪枝后模型(M2)的检测精度下降了 4.9%,所以对模型进行微调,最终得到微调后的检测模型(M3)。模型轻量化

过程中各个模型的检测结果如表 3 所示。

表 3 模型轻量化过程中各个模型的检测结果

模型	参数量/M	模型体积/MB	剪枝率/%	mAP/%
原始 YOLO v4	64.0	245.0	0	98.97
M1	10.5	40.2	0	98.53
M2	5.2	20.0	50	93.63
M3	5.2	20.0	50	98.29

从表 3 可知,通道剪枝可有效剪除网络冗余,并且利用模型微调对剪枝后的模型进行再训练可弥补剪枝带来的检测精度损失,在参数量及模型体积缩减一半以上的情况下,检测精度只损失了 0.68%。因此该方法能够对实验中的检测模型进行有效的轻量化。该轻量化夜间车辆检测模型的 car、bus、van 的 AP 和 mAP 的变化如图 5 所示。

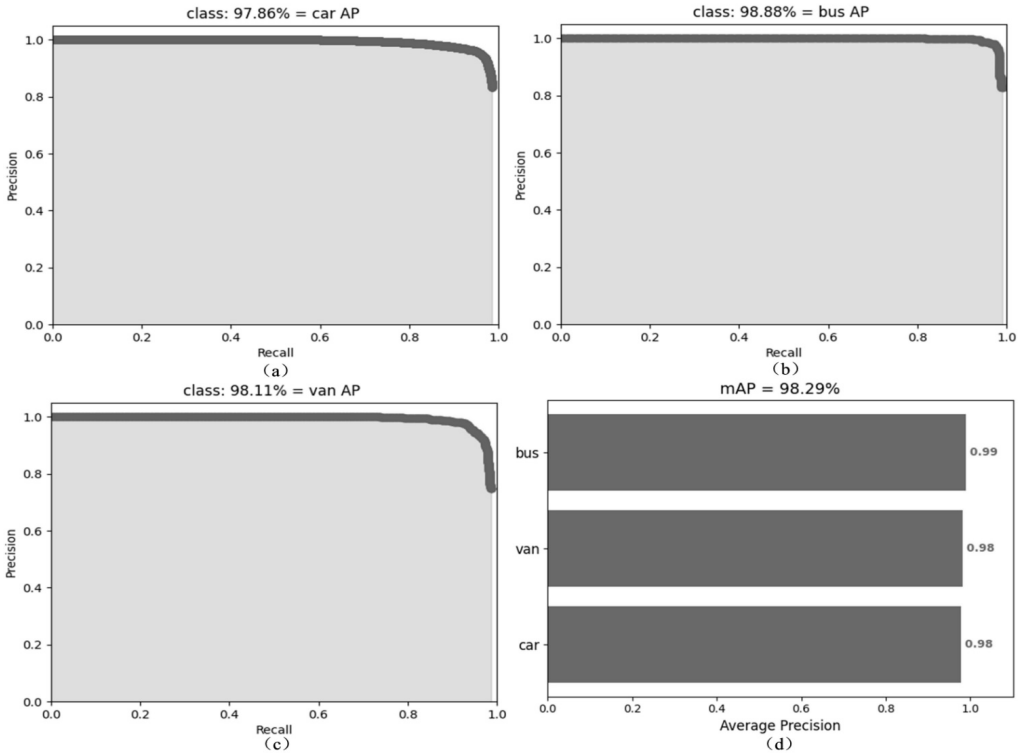


图 5 AP 和 mAP 变化图

损失函数曲线如图 6 所示。

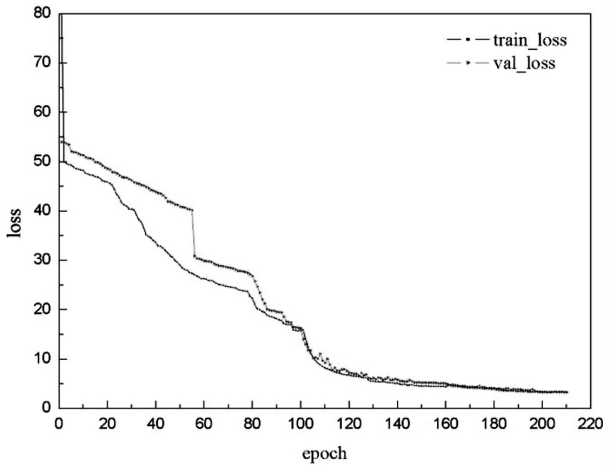


图 6 loss 变化曲线

由图可知网络在迭代训练 200 次后损失曲线不再明显波动,模型更快地趋于收敛,曲线相当平滑,整体训练效果良好,且在迭代训练 210 次后网络发生了早停,停止训练。

对于嵌入式及移动终端设备来说,因计算性能、内存有限,不能满足深度神经网络的运行需求时,就会追求极致的剪枝率[14],但是对于模型轻量化来说,超过一定的剪枝率,就会使得检测精度直线下降。因此,该文对模型剪枝率方面进行了实验与对比分析,使得模型检测精度在基本保持不变的情况下能够去除更多的网络冗余,使检测速度更快,更能满足夜间车辆检测的实时性需求。

由表 4 可以看出,剪枝率越大,参数量和模型体积越小,检测精度会随之有一定的下降。当剪枝率在 75% 时,剪枝后模型的参数量和模型体积降为大约原来的 5%,FPS 提升为大约原来的 2 倍,同时 mAP 只降

低了 1.6%。当剪枝率达到 90% 时,检测精度会大幅降低。故该文采用剪枝率为 50% 时的轻量化模型,此时检测精度 mAP 为 98.29%,且每秒处理帧数 FPS 为 42 帧图像。

表 4　剪枝率对比

模型	参数量/M	模型体积/MB	剪枝率/%	FPS	mAP/%
原始 YOLO v4	64.0	245.0	0	24	98.97
M3	10.5	40.2	0	38	98.53
M3	5.2	20.0	50	42	98.29
M3	3.1	10.6	75	46	97.37
M3	2.9	6.1	90	49	80.56

模型轻量化后的检测效果如图 7 所示。

图 7　车辆检测轻量化模型的检测效果

为了验证轻量化模型对小目标车辆的检测效果,从 UA-DETRAC 数据集中专门挑选了 3 200 张未经过预处理的夜间车辆图像,其特点为存在小目标车辆。经过测试,在 YOLO v4 算法中的 mAP 为 98.39%,而在该文轻量化模型的 mAP 达到 98.04%,证明所提出的轻量化夜间车辆检测模型对小目标车辆也具有很好的鲁棒性。

3　结束语

该文构建了基于 YOLO v4 的轻量化夜间车辆检测模型,其检测精度和实时性都可满足夜间车辆检测的需求,而且对小目标车辆也有很高的检测精度。然而仅实现了夜间车辆的类别与位置检测,对检测到的车辆进行实时跟踪[15]以及预测其接下来准确的行驶方向也具有重要意义。

参考文献:

[1] REDMON J, FARHADI A. YOLOv3:an incremental improvement[J]. arXiv:1804.02767,2018.

[2] BOCHKOVSKIY A,WANG C Y,LIAO H Y M. YOLOv4:optimal speed and accuracy of object detection[J]. arXiv:2004.10934,2020.

[3] MISRA D. Mish:a self regularized non-monotonic neural activation function[J]. arXiv:1908.08681,2019

[4] GHIASI G,LIN T Y,LE Q V. DropBlock:a regularization method for convolutional networks[C]//Proceedings of the 32nd international conference on neural information processing systems. Montreal, Canada: Curran Associates Inc., 2018:10750-10760.

[5] HE K,ZHANG X,REN S,et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1904-1916.

[6] LIU S,QI L,QIN H,et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:8759-8768.

[7] 雷　杰,高　鑫,宋　杰,等. 深度网络模型压缩综述[J]. 软件学报,2018,29(2):251-266.

[8] 黄成强. 结合深度卷积神经网络的智能白平衡研究[J]. 光电子·激光,2020,31(12):1278-1287.

[9] 刘　源. 卷积神经网络的稀疏约束与剪枝方法研究[D]. 深圳:深圳大学,2019.

[10] WEN W,WU C P,WANG Y,et al. Learning structured sparsity in deep neural networks[C]//Advances in neural information processing systems. Barcelona:[s. n.],2016:2074-2082.

[11] LI H,KADAV A,DURDANOVIC I,et al. Pruning filters for efficient ConvNets[J]. arXiv:1608.08710),2016.

[12] 宋叶帆,王国书,盛步云. 一种混合阈值剪枝的稀疏化训练图像识别算法[J]. 科学技术与工程,2021,21(2):638-643.

[13] 季繁繁,杨　鑫,袁晓彤. 基于深度神经网络二阶信息的结构化剪枝算法[J]. 计算机工程,2021,47(2):12-18.

[14] 黄　聪,常　滔,谭　虎,等. 基于权值相似性的神经网络剪枝[J]. 计算机科学与探索,2018,12(8):1278-1285.

[15] 熊昌镇,车满强,王润玲. 自适应卷积特征选择的实时跟踪算法[J]. 中国图象图形学报,2018,23(11):1742-1750.