DOI:10.19651/j.cnki.emt.2209349

基于改进 YOLOv4 的轻量化目标检测算法*

宋中山[1,2] 肖博文[1,2] 艾 勇[1,2] 郑 禄[1,2] 帖 军[1,2]
(1.中南民族大学计算机科学学院 武汉 430074;2.湖北省制造企业智能管理工程技术研究中心 武汉 430074)

摘 要: 为解决 YOLOv4 目标检测网络结构复杂、参数多、训练所需的配置高以及实时检测图片的传输帧数低,难以实现工业上的应用普及等问题,提出一种基于 YOLOv4 改进的轻量化算法 SL-YOLO。在原始的 YOLOv4 网络上进行改进和优化,使用 ShuffleNetv2 轻量级网络替换 YOLOv4 原始骨干网络,将 SENet 模块融入 ShuffleNetv2,降低网络计算复杂度,在网络层中加入 Swish 激活函数,使模型收敛效果更好;同时用简化后的加权双向特征金字塔结构改进原模型的特征融合网络,优化目标检测精度;通过消融实验判定各通道的重要性,对冗余剪枝,将模型进行压缩。在 PASCAL VOC 和 MS COCO 数据集上进行对比实验,改进后的模型与原始 YOLOv4 相比,模型内存减少 89.4%,浮点运算量下降 88.4%,检测速度提升了近 2 倍。实验结果表明,改进后的 YOLOv4 模型能够在保持较高的精度下有效减少模型推理计算量,大大提升模型检测速度。

关键词: 目标检测;轻量化网络;特征金字塔;ShuffleNetv2;YOLOv4
中图分类号: TP391.4 **文献标识码:** 文献标识号:A **国家标准学科分类代码:** 520.6040

Improved lightweight YOLOv4 target detection algorithm

Song Zhongshan[1,2] Xiao Bowen[1,2] Ai Yong[1,2] Zheng Lu[1,2] Tie Jun[1,2]
(1. College of Computer Science, South-Central Minzu University, Wuhan 430074, China;
2. Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprises, Wuhan 430074, China)

Abstract: In order to solve the problems of structurally complex, numerous parameters, high configuration required for training, low transmission frames of real-time detection pictures and difficult to achieve industrial application popularization of YOLOv4 target detection network, a lightweight target detection network SL-YOLO based on YOLOv4 is proposed. It improve and optimizes the original YOLOv4 network, and replaces the original backbone network of YOLOv4 with ShuffleNetv2 lightweight network, integrates SENet module into ShuffleNetv2, reduce the network computing complexity, add Swish activation function to the network layer to make the model convergence effect better; at the same time, the simplified weighted bidirectional feature pyramid structure is used to replace the feature fusion network of YOLOv4, aims to optimize the target detection accuracy; the importance of each channel was determined, thus the redundant pruning was performed, and the model was compressed. The result of a comparative experiment on open data set PASCAL VOC and MS COCO shows that the memory of the model is compressed by 89.4%, the amount of floating-point operations of the model is reduced by 88.4%, and the detection speed of the model is increased by nearly two times, which indicates the SL-YOLO lightweight network can effectively reduce the amount of model reasoning calculation and improve the model detection speed simultaneously, and greatly improve the speed of model detection.

Keywords: object detection;lightweight network;feature pyramid;shuffleNetv2;YOLOv4

0 引 言

随着深度学习理论的研究和软硬件性能的显著提高,深度卷积神经网络在计算机视觉领域的应用和发展得到了强力推进,其中目标检测技术是该领域的研究热点。计算机视觉技术的传统方式,是由人工设计的特征与滑动窗口

收稿日期:2022-03-21
* 基金项目:湖北省科技重大专项(2020AEA011)、武汉市科技计划应用基础前沿项目(2020020601012267)、中南民族大学 2022 年研究生学术创新基金(3212022sycxjj331)项目资助

以及分类算法共同作用实现,特征需要人为设置,这样会导致计算量太大。为提取更高层次的特征信息,神经网络模型的层数也逐渐加深,结构更加复杂,从而使得模型的参数量越来越多,计算量也越来越大,导致检测时间过长并且检测精度较低。

当前,基于深度学习的目标检测算法根据网络结构主要分为两类:第一类是基于深度学习的两阶段目标检测算法,例如 R-CNN[1]算法、Fast R-CNN[2]算法以及 Faster R-CNN[3]算法等。由于传统方法在更为复杂的背景下目标识别的困难逐步降低,更高精度的网络结构也被提出,从图像中获取候选区域,然后通过深度卷积神经网络提取特征,最后使用分类器进行分类与回归。虽然精度与速度都得到了巨大的提升,但这类算法会带来的大量参数,使得检测过程非常耗时,计算复杂度无法满足实时检测的要求。于是端到端的一阶段目标检测算法被提出,以 YOLO 系列[4-8]和 SSD[9-10]系列为代表,这类算法通过获取原图进行学习直接将其视为回归问题,预测类别概率和位置坐标,所以这类检测算法速度更快,但是准确率较为低下,对于需要进行快速目标检测的场景,更适合作为轻量网络结构的 YOLO 算法开始受到重视。目前,基于 Darknet 的 YOLOv4 是最受欢迎的一种检测框架,基于所追求的综合轻量化和高 mAP 以及高 FPS 算法需求考虑,本文选择 YOLOv4 算法进行轻量化改进。

原始 YOLOv4 模型将 Darknet-53 作为主干网络,由于网络的结构层次加深,结构更加复杂导致模型占用内存量较大,计算复杂度又非常高,造成了对硬件的计算能力要求过大。为了实现轻量化网络,文献[11]利用深度可分离卷积与逆残差结构改进 YOLOv4,该方法有效减少模型计算量,但模型整体结构复杂。文献[12]采用精简骨干网络,改变目标检测头结构虽然推理速度有提升但牺牲了模型的检测精度。文献[13]采用网络剪枝去掉特征层,但是优化效果不稳定。文献[14]在 Tiny-YOLOv3 算法中融入多个 1×1 卷积增强语义特征,引入空洞卷积方法,虽然改善了检测效果,但需要消耗大量算力搜索最优结构。文献[15]在 YOLOv4 中加入通道注意力机制与深度可分离卷积,改进模型体积,但是检测速度过慢。文献[16]使用小波-中值滤波处理图像,改进密集连接网络增强了模型的识别能力,但是检测速度慢。文献[17]采用 K-means 算法改进 YOLOv3 网络,虽然达到了聚集检测目标,加快推理速度的效果,但是精度不高。文献[18]设计了新的特征提取模块,增强了特征传递,减少了参数量。文献[19]使用 GhostNet 轻量化网络与 GELU 激活函数重构模型,虽然减少了内存的消耗,但模型的鲁棒性较差。文献[20]提出重构 YOLOv3 的特征金字塔机制,虽然实现了感受野的增强效果,提升了检测精度,但网络结构也变得复杂。上述文献都难以在模型大小大幅减少的情况下,稳定保持算法的准确率。

针对上述问题,为了实现轻量化网络,本文提出一种轻量化网络结构 SL-YOLO。在特征提取层,本文将轻量级网络 ShuffleNetv2[21]进行修改融合,利用深度卷积融合削减了参数量和计算量,然后对网络结构进行改进,使其更加贴合于 YOLOv4 的检测层。ShuffleNetv2 是一种轻量级网络模型,在保持速度快的同时具有高精度的优点。为了进一步将改进后的模型大小进行优化,本文利用网络中批归一化层的缩放因子来标识各个通道的重要性,然后通过正则化操作将缩放因子过低的通道进行剪枝处理,得到一个更轻量化的网络;并且为了验证 SL-YOLO 算法的合理性,针对网络改进点做了一系列消融实验,对重要性过低的通道进行冗余剪枝,通过批归一化的作用减轻过拟合问题。最后将得到参数量和计算量大幅减小且检测速度大幅提升的轻量化模型。

1 相关工作

1.1 YOLOv4 介绍

YOLOv4 算法与同时期的目标检测网络相比,在速度与精度上有着明显的优势。YOLOv4 在以往的 YOLO 检测架构的基础之上在许多方面都进行了优化。如图 1 所示,在网络的特征融合部分,尺寸为 13×13 的特征图进入空间金字塔池化结构中,其原理是将所得到的新的特征图与输入网络之前的特征图进行堆叠和卷积后,输出到路径聚合网络 PANet 中。PANet 能够使不同特征层得到有效地融合,通过最大池化将 3 个不同尺度的卷积层进行拼接,形成一个一维的向量,从而达到对输入图片尺寸没有要求的目的,可以使得输入的图像高度与宽度比例和大小任意,YOLOv4 算法将调整候选框的坐标以及宽高生成最终的预测框。

1.2 ShuffleNetv2 介绍

ShuffleNetv2 是在 ShuffleNetv1[22]的基础上经过大量实验后提出的加强改进版本。ShuffleNetv2 中做了大量实验证明了使用 FLOPs 作为计算复杂度的唯一度量是不够的,还要考虑内存访问成本 MAC、实验平台硬件、算法复杂度等因素。因此结合理论和实践提出了四条轻量化网络通用的性能度量指标:1)当输入输出通道数相同时,可以使内存访问量 MAC 最小;2)当分组数过大时,进行分组卷积操作会增加 MAC;3)碎片化操作对并行加速造成影响;4)逐元素操作的时间消耗远远大于 FLOPs 中引用的值,因此逐元素操作应该最小化。如图 2 所示,基于上述四个轻量化网络的设计原则,为提高网络的运行速度,ShuffleNetv2 在 ShuffleNetv1 的基础上进行了大量证明试验后重构了网络结构。将 channel split 结构引入 ShuffleNetv1 网络结构中,目的是将网络的特征通道分成两部分,从而达到通道内部分别计算的作用。ShuffleNetv2 可以容纳更多的特征通道,因为经过通道拆分操作后,每次卷积计算都是在部分特征通道上进行的,计算量和参数也相应减少。同时,将网络

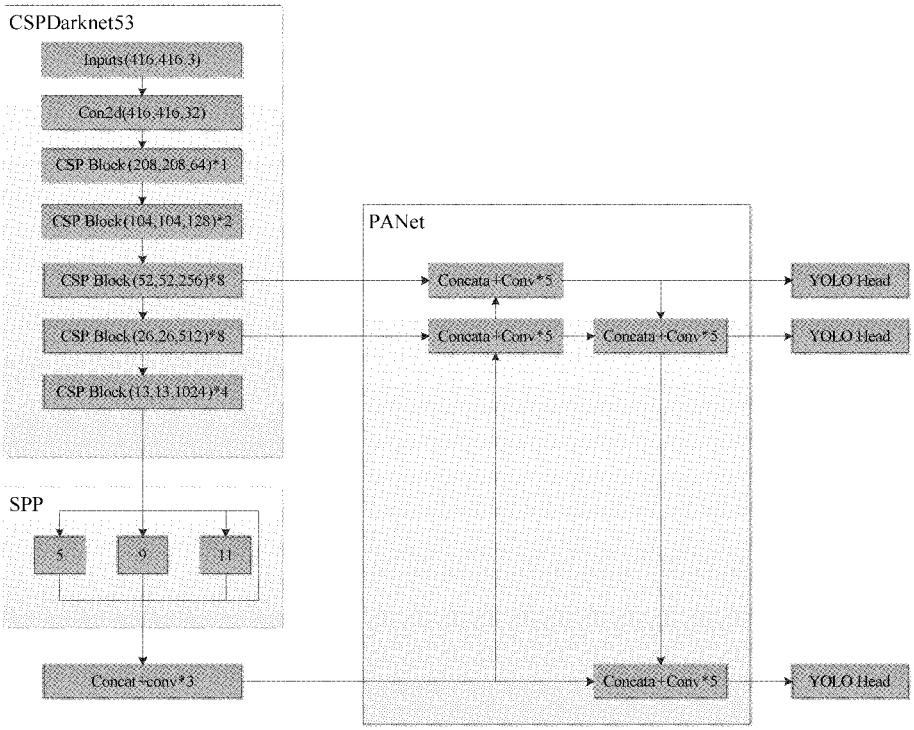
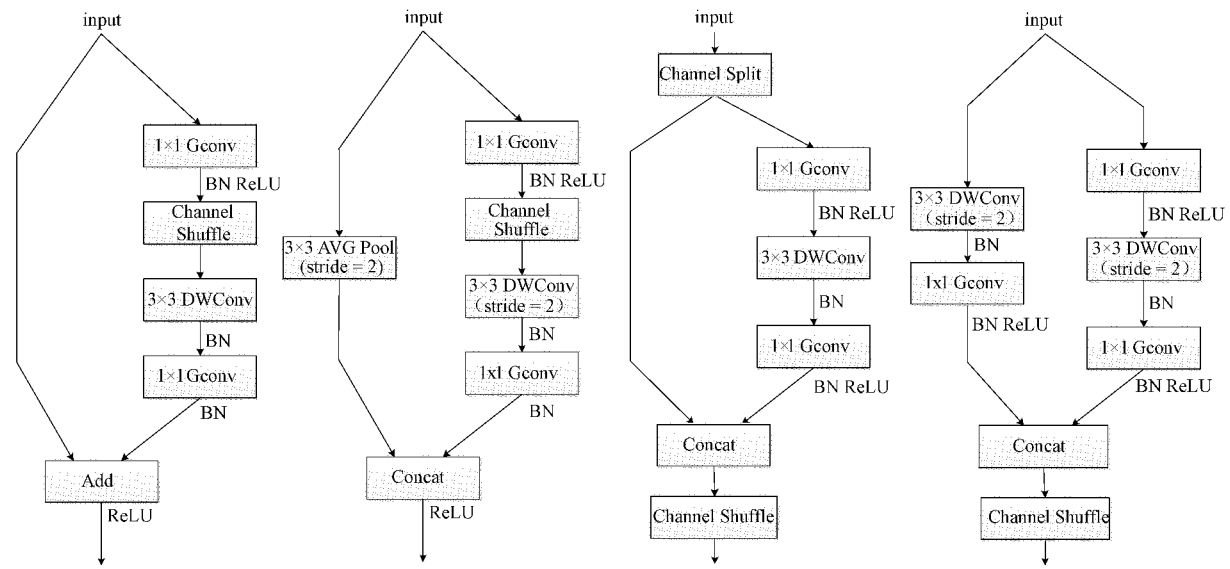


图 1　YOLOv4 网络框架



(a) ShuffleNetv 1基本单元　(b) 空间下采样的ShuffleNetv1单元(2x)　(c) ShuffleNetv2基本单元　(d) 空间下采样的ShuffleNetv2单元(2x)

图 2　ShuffleNetv1 与 ShuffleNetv2 对比图

模块中的特征通道直接送到下一个模块，不需要进行卷积计算，就可以实现特征复用。

2　轻量化网络结构设计

2.1　轻量级网络 SL-YOLO 设计

1）BiFPN

BiFPN 是 EfficientDet[28]提出的基于多尺度特征融合的高效网络，多尺度融合是指用来将不同分辨率的特征聚集起来，使得网络能够快速进行多尺度融合，是一种简单高效的网络，为了学习不同输入特征的对于整个网络重要性，该网络提出通过引入权重的方法，可以达到平衡不同尺度特征信息的作用。如图 3 所示，BiFPN 是在 PANet 的基础上进行改进，采用双向输入思想，将 PANet 网络中只有一条输入边的节点删除，同时增加跨层连接，形成简化

的双向网络,这样做的目的是用来弹性控制 FPN 的大小,在不增加损耗的同时添加额外的边融合更多的特征,一定程度上简化了 FPN 的结构。

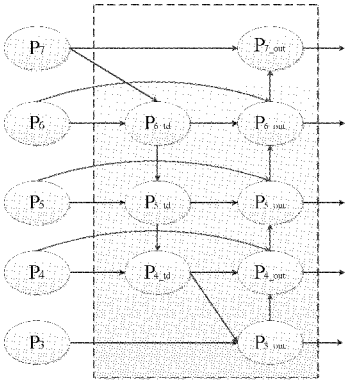


图 3 BiFPN 结构图

本文为了解决不同特征具有不同分辨率从而导致特征贡献不同,在 ShuffleNetv2 网络中引入权重进行训练使网络能够学习到不同分辨率输入特征的重要性。在权重选择方面,BiFPN 使用了快速归一化的优化策略,达到 softmax 的优化效果相似的情况下可以提升 30%的速度。快速归一化表达式(1)如下:

$$O=\sum_i \frac{W_i}{\varepsilon+\sum_j W_j}\cdot \boldsymbol{I}_i \tag{1}$$

i、j 分别为特征融合节点输入的特征图数;$\boldsymbol{I}_i$ 为输入特征图矩阵;ε 为常数 10^{-4};W_i、W_j 分别为输入特征图的权重,这些权重将通过 ReLU[24](rectified linear unit)激活函数用以确保数值的稳定。最终在每个特征融合节点,输入的特征图将获得使目标检测算法效果最好的权重,实现了高效可拓展的效果。

2)激活函数

Swish 函数[25]是一种简单高效的新型激活函数,表达式如式(2)所示。

$$f(x)=x\cdot sigmoid(\beta x) \tag{2}$$

其中,β 是可训练的参数。

观察函数图像,如图 4 所示,可知:

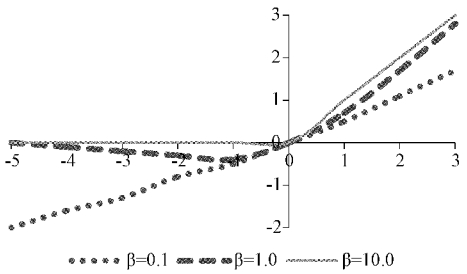


图 4 Swish 激活函数

(1) Swish 激活函数在负值时仍然保存其值,能够更好的保证信息流入;

(2) Swish 激活函数在梯度下降方面的效果更好,从而可以尽可能的保证每一个点的平滑;

(3) Swish 激活函数无边界,可以防止慢速训练期间梯度归零,避免梯度饱和。

总结上文所述 Swish 函数的特点为:具有平滑特性,并且是有下界无上界、非单调凸函数,融入这样的激活函数能够降低整个网络的推理成本。本文决定将 ShuffleNetv2 网络模型中原本 1×1 分组卷积层的 ReLU 函数替换为 Swish 激活函数。

3)网络模型最小组件

ShuffleNetv2 网络的通道拆分虽然提供了通道间的信息交换,但会导致融合特征的丢失。为解决这个问题,本文提出一种新型模块 SL 模块,将 MobileNetv3[26] 提出的 SE 模块与 ShuffleNetv2 的通道模块进行融合形成轻量级注意力机制模型。

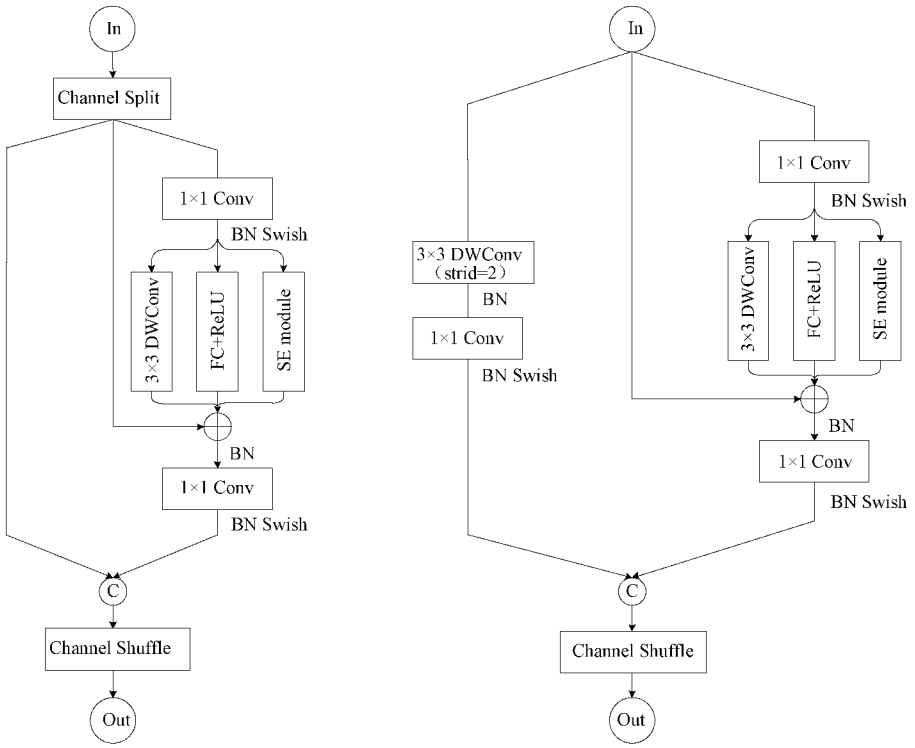
由于在经过 ReLU 函数激活后,低纬度的特征图极其容易造成部分信息的丢失,而经过高维度的特征图可以有效减少信息的丢失,所以本文在 ReLU 激活函数后加入一个全连接层,用 Swish 替代全连接层所有 ReLU 激活函数,从而提高训练后的质量,这样为了平衡特征图通道的权重,对原模型的尾部结构进行了一定的调整,减少了计算量的同时可以将更好的参数生成更多的特征图,提高网络的学习能力。

图 5 中,在改进后 ShuffleNetv2 网络中 SL 模块构使用金字塔内核大小来捕获更多的特征信息,而不是仅仅使用 3×3 内核大小的深度分离卷积,然后在 1×1 卷积之前合并金字塔卷积的所有输出。对于基本单元,通过通道拆分操作将输入拆分为两个独立分支,一个不做任何操作,另一个由两个具有相同通道的 1×1 卷积间的金字塔卷积组成。采用通道重排操作实现卷积后的两个重新连接的独立分支之间的信息通信。此外左分支由一个 1×1 步长为 2 的深度分离卷积和一个 1×1 卷机组成,右分支与左分支基本单元保持相同,其中特征提取网络的卷积层中融入轻量级注意力机制 SE 结构,采用线性瓶颈层作为基本设计单元,借鉴通道重排的效果,通过并列对称特征融合和通道拆分结合对卷积层的结构修改进行修改,形成一种新的网络结构;同时使用 swish 激活函数的门控机制进行建模,实现各个通道之间的特征重标定。将 SENet[27]模块与 ShuffleNetv2 网络进行融合,可以大大提升轻量级物体检测模型提取特征的能力。

卷积层通道数上,去掉最后一层卷积,在特征融合后不做降维操作。网络中每个阶段都会经过一个下采样单元和多个改进后的基本单元,同时使用全局平均池化用来减少计算参数,防止出现过拟合现象。下表为改进后的 ShuffleNetv2 网络,具体结构如表 1 所示。

4)网络模型架构

本文提出的 SL-YOLO 轻量化模型主要进行了两方面



(a) ShuffleNetv2 单元1　　(b) ShuffleNetv2 单元2

图 5　改进后的 ShuffleNetv2 网络最小基本组件

表 1　改进后的 ShuffleNetv2 网络结构

层	输出大小	卷积核	步长	SE	重复	输出通道(1×)
Image	224×224					3
Conv1	112×112	3 * 3	2		1	24
MaxPool	56×56	3 * 3	2			
Stage2	28×28		2	√	1	116
	28×28		1	√	3	
Stage3	14×14		2	√	1	232
	14×14		1		3	
Stage4	7×7		2	√	1	464
	7×7		1	√	3	
Conv5	7×7	1 * 1	1	√	1	1 024
GlobalPool	1×1	7 * 7				
FC						1 000

改进，一是对 ShuffleNetv2 网络进行修改并将其替换 YOLOv4 的主干网络，二是将 BiFPN 的结构改进为 L-FPN 并替换 YOLOv4 的特征融合网络，网络架构如图 6 所示。

ShuffleNetv2 网络中的通道重排可以提供通道间的信息交换，但会导致融合特征的丢失。SE 模块在对网络通道进行加权以获得更好的功能，将 SE 模块加入特征提取网络形成新的 SL 模块，可以增强提高网络的学习能力。

由于原本的 BiFPN 是将原图片连续下采样 5 次之后分别进入特征融合网络进行特征融合的结构，但 YOLOv4 算法经过主干特征提取网络之后只有 3 种尺度的特征图输出，并且 13×13 尺度的特征图难以继续进行下采样。为了使 BiFPN 能够匹配 YOLOv4 模型，L-FPN 减少了 BiFPN 中的两个特征层，只对 3 个尺度的特征层进行特征融合。等同于在原本 PANet 的基础之上增添一条残差边，同时在多节点间融入注意力机制。L-FPN 将特征融合节点输入的特征图乘以 ω 权重，然后对这些权重使用快速归一化进行训练。P_{5_in} 直接对 P_{4_td} 进行上采样，然后与 P_{4_in}

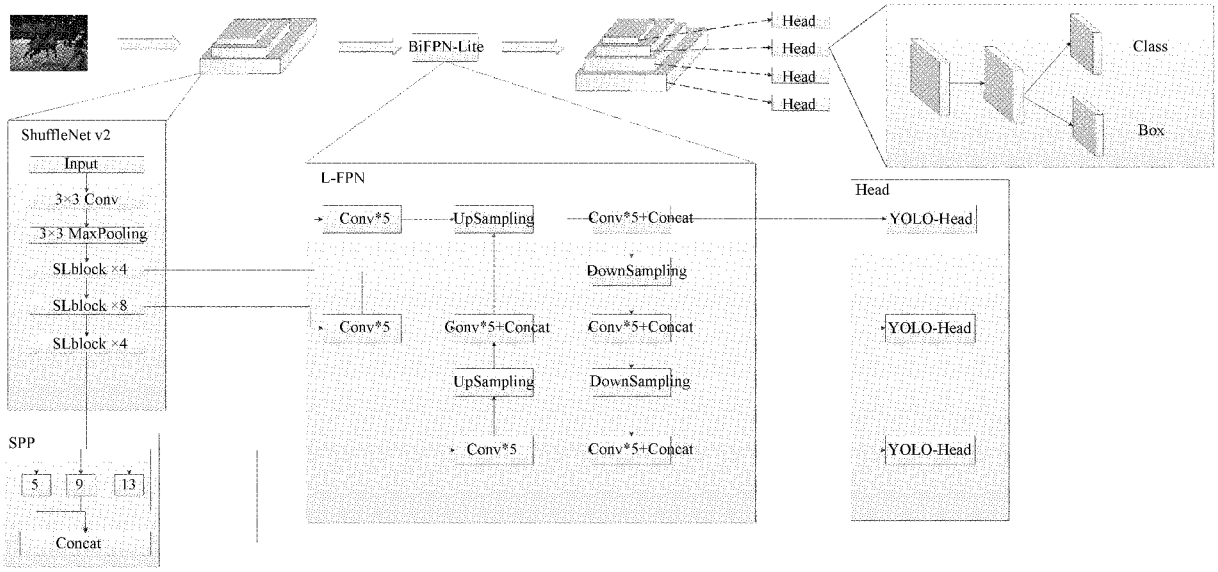


图 6 改进的 YOLOv4 网络模型

进行特征融合;P_{4_out} 与 P_{4_in}、P_{4_td}、P_{3_out} 下采样的结果进行特征融合。将 P_{4_td}、P_{4_out} 节点输入的特征矩阵带入式(1)得到式(3)、(4)如下:

$$P_{4\_td}=Conv\left(\frac{\omega_1\cdot P_{4\_in}\cdot Resize(P_{5\_in})}{\omega_1+\omega_2+\varepsilon}\right) \tag{3}$$

$$P_{4\_out}=Conv\left(\frac{\omega_1\cdot P_{4\_in}+\omega_2\cdot P_{4\_td}+\omega_3\cdot Resize(P_{3\_out})}{\omega_1+\omega_2+\omega_3+\varepsilon}\right) \tag{4}$$

$Conv$ 表示卷积操作,为了抽取特征;$Resize$ 表示上下采样过程,为了对其特征图尺寸;ω 表示各个特征图进行训练的权重,并且权重的初始值为 $0<\omega<1$ 范围内的随机数,通过每一轮训练得到合理的权重数值;ω 为常数 10^{-4};分母的加号代表不同特征图的堆叠。

最后将分别输出 3 个检测头,然后通过 YOLOv4 的解码算法将这 3 个检测头生成最后的预测框。

2.2 稀疏训练

将 YOLOv4 的主干网络替换为修改后的 ShuffleNetv2 网络后,虽然整个深度学习网络的计算量得到了有效降低,但是为了具有高效的拟合能力保持好的检测效果,卷积运算后的冗余结构会导致模型体积仍然较大,这会导致网络推理速度减慢,增加计算推理时长。为了解决这些问题,可以采用模型压缩的方法,常用的模型压缩方法有多种,本文结合层剪枝和模型剪枝的方法,在极大压缩模型体积的同时保持检测的高精度。

通过批归一化对神经网络进行优化,式(5)所示为批归一化原理,在提高训练效率的同时可以防止梯度爆炸,在对原模型中部分权值参数进行训练的情况下就能达到本体网络相近甚至是更好的网络性能。

$$\hat{x}_i=\frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\varepsilon}};\ y_i=\gamma\hat{x}_i+\beta \tag{5}$$

式中:μ_B 与 σ_B 分别表示每批次数据的均值与方差;γ 与 β 为可训练参数,分别对因子进行评议和缩放,$\hat{x}_i$ 为归一化结果,y_i 为输出结果。

基于批归一化层原理,将 ShuffleNetv2 每个卷积层后的批归一化层的缩放系数 γ 作为缩放因子,用来保留归一化操作前的特征,以此作为评估通道重要程度的指标因子。如果权重分别不稀疏,需要通过正则化操作对缩放因子进行强制稀疏化,这种方式可能会造成精度的损失,因此通过正则化得到的结果需要与通道的输出相乘,经过稀疏网络的训练-模型剪枝-微调网络的迭代这一系列过程并联合权重与缩放因子训练网络,损失的精度可以通过微调进行补偿。

2.3 通道剪枝

稀疏训练后可以得到相对稳定的网络结构,然而还需要进行进一步修建。首先需要将全局剪枝率对应的缩放因子 γ 的绝对值进行排序,所对应的阈值为 θ_1。为了防止剪枝误差造成损失,需要设定剪枝比例为 50%,并设定一个保护阈值 β,用来保护一定比例的通道,所对应的每层阈值为 θ_2。当满足 $\gamma<\theta_1$ 且 $\gamma<\theta_2$ 时,可以对通道进行剪枝,为了防止剪枝率过大造成精度的损失,本文设置 40%的剪枝比例。剪枝过程如图 7 所示,通过迭代执行稀疏训练、通道剪枝、微调的过程,可以在保持精度基本无损的同时实现模型最大限度的压缩,并且可以使模型的前向推导过程得到很大程度的提升。

3 实验结果与分析

3.1 实验环境介绍

训练所用的实验平台硬件配置为 Intel(R)Core(TM) i7-6700cpu,系统为 WIN10。实验操作系统为 CentOS Linux 7.7,采用的集成开发环境是 Anaconda3,深度学习

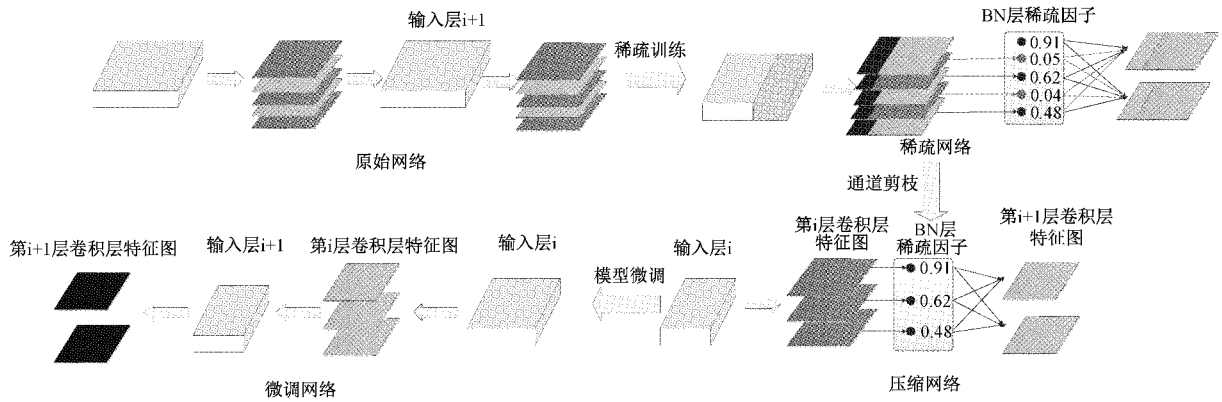


图 7　通道剪枝过程

框架为 PyTorch。GPU 为 NVIDIA Tesla P40，22919Mib。

训练时批量操作大小为 16，初始学习率为 0.002、正常迭代次数为 10 000 训练 300 个 epoch，稀疏训练和微调的初始学习率为 0.001 训练 100 个 epoch；借助余弦函数的特性来调整学习率；在迭代优化的前期，梯度下降速度随着学习率的减小而加快；在迭代中后期，学习率的减小速度会随着迭代次数的增加而变慢，这将有助于算法的收敛，直到 Loss 不再变化自动终止训练。学习率变化曲线如图 8 所示。

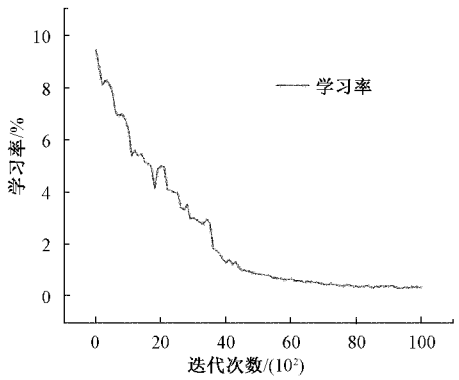


图 8　学习率变化曲线

3.2　模型性能评价指标

模型评估标准主要评估模型的参数量、召回率、交并比、计算量以及平均精度（average precision，AP）、平均精度均值（mean average precision，mAP）来衡量不同分辨率下的性能，mAP 越大整体精度越高，FPS 来衡量每秒能够处理的图片个数。为了在不减少精度或是在可接受范围内的前提条件下，最大化的压缩模型尺寸，减少模型计算量（FLOPs），提高帧率（FPS）。

3.3　消融实验

为了验证改进 YOLOv4 算法的优越性，本实验采用 PASCAL VOC 与 MS COCO 公共数据集使验证集具有全面性，将本文所提出的 SL-YOLO 模型与常用的目标检测算法从不同方面进行消融实验，下面是实验分析的详细说明。

1）PASCAL VOC 数据集实验分析

本文所提出的 SL-YOLO 模型在保持训练参数一致的前提下与 Faster R-CNN、SSD、YOLOv3、YOLOv4、YOLOv4tiny 等目前主流模型在 PASCAL VOC2007 测试集中的检测精确度对比结果，对比如图 9、表 2、3 所示。

其中，SL-YOLO（without AF）表示 SL-YOLO 模型采用稀疏训练与通道剪枝，不采用 swish 激活函数；SL-YOLO（without SC）表示只采用 swish 激活函数；SL-YOLO（without SP）表示只采用稀疏训练。如图 10，将各模型参数量进行对比，得到以下结论：

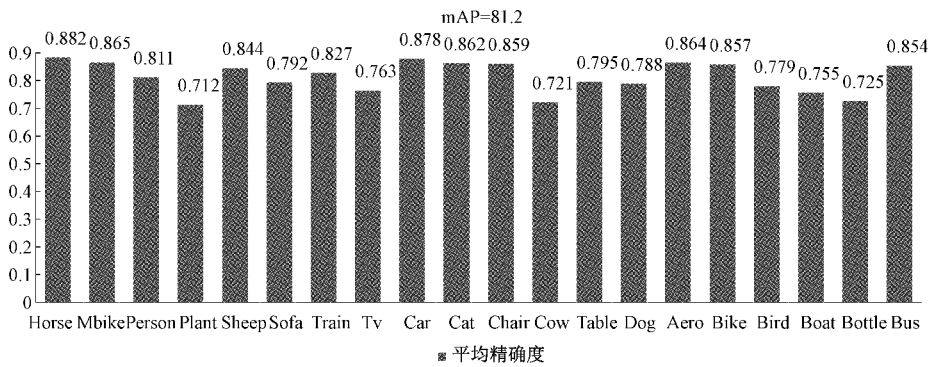


图 9　SL-YOLO 在 VOC 测试集 mAP

表 2 SL-YOLO 与其他算法在 VOC2007、2012 上的结果对比

模型(输入尺寸)	年份	骨干网络	召回率	mAP/%
Faster RCNN	2015	VGG16	0.538	71.5
SSD(300)	2016	VGG16	0.574	72.1
SSD(300)	2016	MobileNetv2	0.542	75.2
SSDLite(300)	2017	MobileNet	0.618	70.7
YOLOv3(416)	2018	Darknet53	0.725	76.3
YOLOv4(416)	2020	Darknet53	0.817	83.4
YOLOv4		MobileNet	0.761	75.4
YOLOv4		GhostNet	0.783	72.9
EfficientDet	2020	EfficientNet	0.774	69.8
YOLOv4tiny(416)	2020	CSPDarknet53_Tiny	0.797	71.1
SL-YOLO(without AF)	2021		0.799	69.2
SL-YOLO(without SC)	2021		0.814	79.1
SL-YOLO(without SP)	2021		0.831	77.5
SL-YOLO	2021		**0.836**	**81.2**

表 3 SL-YOLO 与 YOLO 模型对比

模型	模型大小/MB	计算量/(FLOPs)	检测速度/(帧/s)
YOLOv4	244	160	6.48
YOLOv4tiny	23.9	15.72	21.57
SL-YOLO	26.1	18.58	17.79

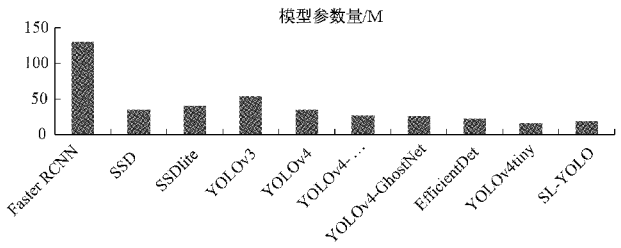


图 10 各模型参数量对比

(1) SL-YOLO 与 YOLOv4 相比,参数量和计算量更少,模型的检测速度的得到大幅提升同时兼备高精确度。通过替换激活函数后的模型虽然在检测速度上没有进行提高,但能够确保整个网络保持较高的检测精度。此外,采用稀疏训练以及通道剪枝,可以压缩模型大小、加快检测速度,缩短推理时间。

(2) SL-YOLO 与 YOLOv4 相比,SL-YOLO 模型大小为 26.1 MB,压缩了近 9 倍,有效降低了存储成本,适合在资源受限的平台上进行部署。

(3) 总体来看,改进后模型的整体性能有很大提升,模型内存大小和运算速度的综合性能优于其他网络。虽然剪枝后损失了部分精度,但通过迭代微调是整体结构达到了一定的平衡,在可接受误差范围内,模型的检测速度相比 YOLOv4 提升了 4.4 倍,并且与 YOLOv4tiny 相比检测的精确度提高了 10.1%。

2)MS COCO 数据集实验分析

本文继续使用所提出的 SL-YOLO 模型在 MS COCO 测试集中进行训练,在测试过程中,本文采用不同的交并比值对性能进行检测,4 种模型的测试结果对比如表 4 所示。

表 4 SL-YOLO 与 YOLO、SSD 在 COCO 上的结果对比

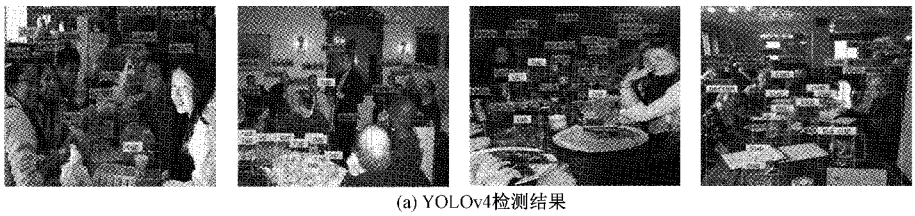
模型	AP,IoU			AP,Area		
	0.5…0.95	0.5	0.75	S	M	L
SSD	22.3	41.1	23.5	5.6	23.9	39.4
GhostNet	25.7	44.6	26.9	8.3	27.1	42.8
EfficientDet	27.4	46.7	28.6	9.6	29.6	39.7
YOLOv4	40.2	60.8	42.3	19.4	42.4	54.0
YOLOv4tiny	29.6	40.2	31.4	9.8	33.5	41.8
SL-YOLO	**36.8**	**51.3**	**37.3**	**11.7**	**37.7**	**48.4**

其中 AP 表示检测的准确率。Area 表示 3 种尺度 S、M、L 分别对应小、中、大目标。Iou 表示不同的交并比值。

由表 4 可知,改进的 SL-YOLO 轻量化模型在 MS COCO 公共数据集上的检测精度略低于 YOLOv4,原因在于 SL-YOLO 模型通过稀疏训练的剪枝操作与压缩后,对参数量与计算量进行了极大程度压缩后会造成网络层数和通道数一定的损失,无法完整保留分类后的图片特征,并且 MS COCO 数据集分类较为细致,从而导致部分误差。

分别将 YOLOv4、YOLOv4tiny、SL-YOLO 3 种模型对 MS COCO 测试集上个别图像识别的结果对比如图 11 所示。

对比实验结果与图 11 可以看到,SL-YOLO 与 YOLOv4


(a) YOLOv4检测结果


(b) YOLOv4tiny检测结果

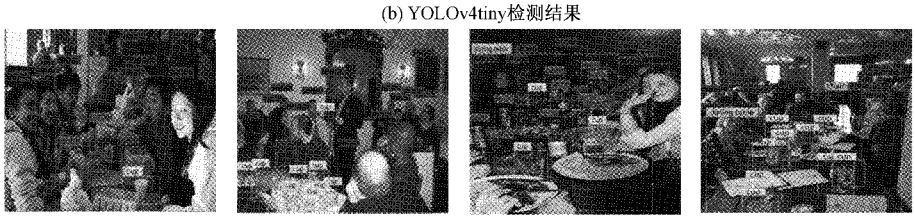

(c) SL-YOLO检测结果

图 11　在 MS COCO 数据集上的检测效果比较

的检测效果近乎相同，只是对于一些小目标的检测结果有误差，不过在牺牲精度可接受范围内大幅提升检测速度，这种方法是可以接受的。相较于 YOLOv4tiny 这种非常轻量化的算法，SL-YOLO 虽然模型大小稍大，但 mAP 有大幅度提升，并且计算速度相差不大，并且 YOLOv4tiny 的检测结果中漏检现象很严重。综合模型大小、计算速度、mAP 来看，SL-YOLO 确实有较好的性能。

3.4　轻量级目标检测系统

为验证本文方法的实际应用效果，将设计一个基于云端和移动端开发的应用系统对现实生活中的物体进行测试。本系统主要是基于 Android 技术来实现手机界面的开发。系统的核心算法部分主要包括由目标检测算法对物体进行定位与分类，以及轻量化模型实现对物体的高效识别。

本系统基于 PASCAL VOC 数据集和 MS COCO 数据集，测试系统可以对用户通过移动端拍照或上传的图片进行处理，或者是对现实中的人或物体图像上传至云端进行处理，云端基于 SL-YOLO 轻量化目标检测算法对目标进行识别，最后将识别结果反馈至移动端。移动端测试结果界面如图 12 所示。通过测试结果可以看出，系统对于物体的检测准确度较高，并且识别速度适用于终端部署。

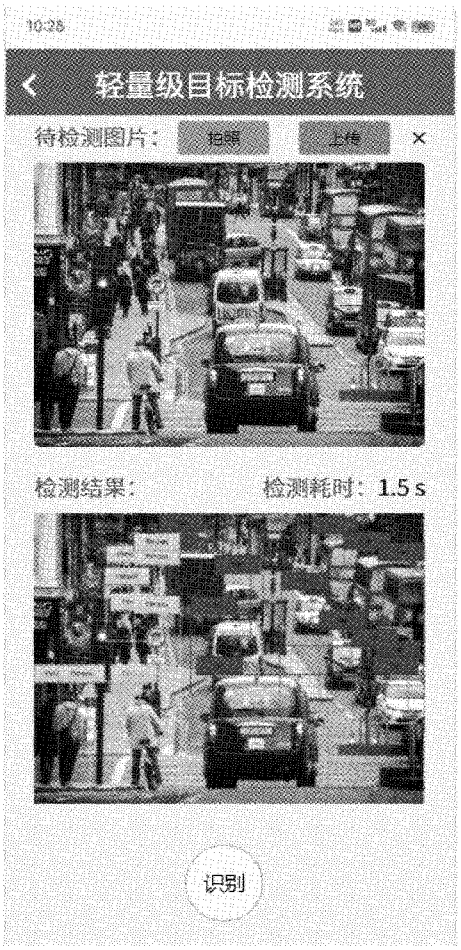


图 12　测试系统识别结果

4 结 论

为了提升模型在进行目标检测时的检测速度,针对 YOLOv4 的缺陷,本文提出了一种基于改进 YOLOv4 的轻量化目标检测算法 SL-YOLO。通过引入 SE 模块与 ShuffleNetv2 网络进行融合,达到在保持高检测精度的同时,极大限度压缩模型体积的目的。同时用简化后的加权双向特征金字塔(L-FPN)结构替换 YOLOv4 的特征融合网络,简化后的 BiFPN 结构继承了 PANet 双向特征融合特点的同时,适当减少了结构中的节点数,同时加入注意力机制,在减少卷积次数的同时提高了不同层间特征融合的鲁棒性,增强了模型的检测效果。实验结果表明,相比于其他目标检测算法,SL-YOLO 算法的检测精度与检测速度能满足实际应用需求。在未来的研究工作中,将会进一步优化 SL-YOLO 的轻量化改进方法,在尽可能保证降低计算量、参数量和模型大小的前提下,将检测精度与检测速度进行更好的提升,实现工业上的普及。

参考文献

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

[2] GIRSHICK R. Fast R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[3] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[5] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7263-7271.

[6] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. ArXiv Preprint, 2018, ArXiv: 1804.02767.

[7] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. ArXiv Preprint, 2020, ArXiv:2004.10934.

[8] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. Scaled-yolov4: Scaling cross stage partial network[C]. Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition, 2021: 13029-13038.

[9] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]. European Conference on Computer Vision. Springer, Cham, 2016: 21-37.

[10] FU C Y, LIU W, RANGA A, et al. Dssd: Deconvolutional single shot detector[J]. ArXiv Preprint, 2017, ArXiv:1701.06659.

[11] 张明路,郭策,吕晓玲,等. 改进的轻量化 YOLOv4 用于电子元器件检测[J]. 电子测量与仪器学报,2021,35(10):17-23.

[12] 方仁渊,王敏. 基于改进型 YOLO 网络的商品包装类型检测[J]. 电子测量技术,2020,43(7):108-112.

[13] FANG W, WANG L, REN P. Tinier-YOLO: A real-time object detection method for constrained environments[J]. IEEE Access, 2019, 8: 1935-1944.

[14] 化嫣然,张卓,龙赛,等. 基于改进 YOLO 算法的遥感图像目标检测[J]. 电子测量技术,2020,43(24):87-92.

[15] 彭继慎,孙礼鑫,王凯,等. 基于模型压缩的 ED-YOLO 电力巡检无人机避障目标检测算法[J]. 仪器仪表学报,2021,42(10):161-170.

[16] 李庆党,李铁林. 基于改进 YOLOv3 算法的钢板缺陷检测[J]. 电子测量技术,2021,44(2):104-108.

[17] 李云鹏,侯凌燕,王超. 基于 YOLOv3 的自动驾驶中运动目标检测[J]. 计算机工程与设计,2019,40(4):1139-1144.

[18] 冯宇平,管玉宇,杨旭睿,等. 融合注意力机制的实时行人检测算法[J]. 电子测量技术,2021,44(17):123-130.

[19] 石晨宁,周春,靳鸿,等. 基于卷积神经网络的农作物病害识别研究[J]. 国外电子测量技术,2021,40(9):93-99.

[20] 薛瑞晨,郝媛媛,张振,等. 基于改进 YOLOv3 的头盔佩戴检测算法[J]. 电子测量技术,2021,44(12):115-120.

[21] MA N, ZHANG X, ZHENG H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]. Proceedings of the European Conference on Computer Vision(ECCV), 2018: 116-131.

[22] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6848-6856.

[23] TAN M, PANG R, LE Q V. Efficientdet: Scalable

and efficient object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781-10790.

[24] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks [C]. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011: 315-323.

[25] RAMACHANDRAN P, ZOPH B, LE Q V. Searching for activation functions[J]. ArXiv Preprint, 2017, ArXiv:1710.05941.

[26] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3 [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1314-1324.

[27] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

作者简介

宋中山，副教授，主要研究方向为深度学习理论与方法、图像处理。

E-mail:songzs@mail.scucc.edu.cn

肖博文，硕士研究生，主要研究方向为计算机视觉、目标检测。

E-mail:392176217@qq.com

艾勇（通信作者），讲师，主要研究方向为人工智能、机器学习。

E-mail:aiy_scucc@qq.com

郑禄，讲师，主要研究方向为深度学习、图像处理。

E-mail:lu2008@mail.scucc.edu.cn

帖军，教授，主要研究方向为模式识别、图像处理。

E-mail:tiejun@mail.scuec.edu.cn