DOI:10.3969/j.issn.1003-5060.2023.09.008

基于改进 YOLO v4 的轻量化烟梗识别方法

郑银环， 林晓琛， 吴 飞， 金圣洁， 吴傲男
(武汉理工大学 机电工程学院,湖北 武汉 430070)

摘 要:为完成烟叶精选工艺流程中打叶复烤后破碎烟叶的进一步去梗,实现破碎烟叶中烟梗的自动化检测,文章提出基于改进 YOLO v4 的轻量化烟梗识别方法。在 YOLO v4 基础模型上先后进行通道剪枝和层剪枝,大幅简化模型结构,改进后模型存储空间下降了 93.77%,模型平均精度均值(mean average precision, mAP)和前向运算时间与基础模型持平。与同类别算法相比,模型精度平均提升 8.7%,模型参数量大幅缩减。实验结果表明该实验剪枝模型更具轻量化,识别效果更好,能够满足实际生产需求。

关键词:烟梗识别;YOLO v4;通道剪枝;层剪枝;轻量化
中图分类号:TP391 **文献标志码**:A **文章编号**:1003-5060(2023)09-1196-08

Lightweight tobacco stem identification method based on improved YOLO v4

ZHENG Yinhuan, LIN Xiaochen, WU Fei, JIN Shengjie, WU Aonan
(School of Mechanical and Electrical Engineering, Wuhan University of Technology, Wuhan 430070, China)

Abstract: In order to realize further stem removal of broken tobacco leaves after leaf beating and re-roasting in tobacco leaf selection process and realize automatic detection of tobacco stems in broken tobacco leaves, a lightweight tobacco stem identification method based on improved YOLO v4 was proposed. In YOLO V4 model, channel pruning and layer pruning were carried out successively to simplify the model structure greatly. The storage space of the improved model is reduced by 93.77%, and the mean average precision(mAP) of the model and the forward computing time are the same as those of the basic model. Compared with algorithms of the same category, the model accuracy is improved by 8.7% on average, and the number of model parameters is greatly reduced. The analysis shows that the pruning model is more lightweight, has better identification effect, and meets the actual production demand.

Key words: tobacco stem identification; YOLO v4; channel pruning; layer pruning; lightweight

0 引 言

在烟叶制丝工艺中,首先需要将大片烟叶进行打叶复烤,通过打叶设备使叶片和烟梗分离,将整片烟叶破碎成小块。经打叶复烤后虽绝大多数烟梗已分离,但其中仍夹杂着一定量的烟梗。烟叶打叶复烤后应符合严格的质量标准[1],因此需将复烤后的破碎烟叶进行松散回溯处理,即通过120 ℃热蒸汽提升烟叶的温度与水分,达到烟叶松散和舒展的效果,随后通过振动筛装置使烟叶铺开减少烟叶堆叠情况,再进一步去梗。由于生产质量标准严格,该工序一直都是人工完成,带来生产效率低、耗时长、人工费用高等生产实际问题。

针对烟叶中烟梗识别与剔除的问题,国内外学者进行了大量的研究。文献[2]研究了低能 X

收稿日期:2022-09-19;**修回日期**:2023-01-03
基金项目:国家自然科学基金资助项目(52005376)
作者简介:郑银环(1974—),女,湖北钟祥人,博士,武汉理工大学副教授,硕士生导师;
吴 飞(1973—),男,河南叶县人,博士,武汉理工大学教授,硕士生导师.

射线透射成像，结合形态学滤噪、灰度阈值分割、归属判断等图像实时分析处理手段，实现了打叶片烟中烟梗的在线识别及剔除；文献[3]设计了一种以基于现场与编程门阵列（field programmable gate array，FPGA）的高速图像处理设备为核心的烟梗在线检测系统，该系统采用 X 光源透射皮带上的烟叶以及烟梗，通过探测器接受 X 射线并通过高速图像处理设备对图像进行算法处理，进而实现烟梗识别；文献[4]利用 X 射线的透射性原理，结合 X 射线穿过物质后的衰减规律，对片烟中烟梗质量进行拟合，从而计算出叶中含梗率；文献[5]采用高功率红外线作为透射光源，以高清连续摄录机为图像获取平台，以灰度算法为核心程序，实现烟梗面积自动化测算系统的开发。

上述研究利用烟叶和烟梗在不同光源下反映出的特性不同作为判据依据进行烟梗识别及分离，而在实际生产中，经松散回溯后的破碎烟叶通常为散乱堆叠状态且烟梗特征不一，造成上述方法在实际生产中的识别精度不高。

深度学习技术在计算机视觉领域中的应用发展迅速。根据算法原理不同主要分为 2 类：① 先进行区域生成，再通过卷积神经网络（convolutional neural network，CNN）进行样本的分类，其中最具代表性的算法有 R-CNN、SPP-Net、Fast R-CNN、Faster R-CNN、R-FCN 等[6-10]；② 不进行区域生成，直接在网络中提取特征来预测物体分类和位置，常见算法有 OverFeat[11]、SSD[12]、YOLO 系列算法[13-14]等，此类算法因具备可同时预测物体种类和位置的特性，相比于第①类算法在目标检测过程中更具快速性。而 YOLO v4 作为检测速度和检测精度兼备的检测算法，在工业现场应用广泛。

基于深度学习的目标检测技术虽然识别精度高、可移植性好，但往往计算量大，模型结构复杂，占用空间大，而工业现场的工控机等主机设备往往储存空间有限，使得检测算法的现场部署实施困难。

本文以 YOLO v4 为基础网络，建立了烟叶回溯振动铺开后烟梗的轻量化检测模型。利用通道剪枝和层剪枝相结合的方法对模型进行压缩，通道剪枝大大减小了模型参数量和计算量，降低了模型对资源的占用。层剪枝可以进一步减小计算量，并大大提高模型推理速度。两者相结合，可以大幅度压缩模型的深度和宽度。

1　数据集

1.1　数据采集

本文在数据集制作过程模拟真实工业场景，综合考虑了真实生产线上烟叶可能出现的情况。实验对象为松散回溯工序后的破碎烟叶，且已由振动筛装置铺开无明显堆叠情况，人工将破碎叶片随机摆在黑色背景的实验拍摄平台，使用英特尔 RealSense d435i 型号相机采集烟叶数据集。实验拍摄平台如图 1 所示。

实验采用设备灯光进行拍照，光照条件设置为曝光度 640、对比度 50、饱和度 64。图像输出尺寸为 1 920 像素×1 080 像素，帧率为 8，共拍摄 1 000 张烟叶照片。

(a) RealSense d435i相机

(b) 拍摄三脚架

图 1　图像采集平台

1.2　数据预处理

为丰富数据集，提高模型泛化能力，进一步提升模型性能，本文采用 Mosaic 数据增强方法，随机读取 4 张烟叶照片，分别对 4 张照片进行翻转、缩放等操作，随之将 4 张照片按照左上、左下、右下、右上的顺序排布，最后利用矩阵的方式将 4 张图片在固定位置截取下来并拼接成 1 张新的图片。原始烟叶照片及 Mosaic 数据增强后的照片如图 2 所示。

(a) 初始烟叶照片　　(b) Mosaic 数据增强照片

图 2　破碎烟叶图像

最终采集到烟叶数据集共 2 000 张,其中训练集与测试集比例划分为 7∶3。

2 烟梗快速识别方法

松散回溯后振动铺开的烟叶形状不规则,烟叶中烟梗特征较不明显。将标注好的烟叶烟梗数据集输入到 YOLO v4 网络中进行基础训练,得到基础模型。针对 YOLO v4 基础训练模型进行稀疏训练,随后将稀疏训练模型进行通道剪枝和层剪枝,修剪掉模型中的冗余通道、Shortcut 层和卷积层,大幅度减小了模型的参数量和模型大小,提高了模型推理速度,但模型精度会有所下降。最后对剪枝后的模型进行微调,使模型精度重新回升至较高水平。

2.1 基于 YOLO v4 的烟梗检测模型

YOLO v4 以 CSPDarknet53 作为主干网络实现图像特征的提取。CSPDarknet53 将 YOLO v3 的主干网络 Darknet53 与 CSPNet 进行结合形成新网络主干,其中 CSPNet 将基本层的特征图分成两部分,然后通过一个跨阶段的层次结构合并它们,使新的网络主干能够实现更丰富的梯度组合,同时减少计算量;以 PANet(path aggregation network)代替了 YOLO v3 中的特征金字塔网络(feature pyramid network,FPN)作为参数聚合,针对不同主干层进行参数聚合;同时继承了 YOLO v3 的多尺度预测模块,输出层的锚框机制与 YOLO v3 相同,主要改进了训练时的损失函数 CIoU_Loss 以及预测框筛选的 DIoU_nms,进一步提升了算法的检测精度。

检测烟叶中烟梗的 YOLO v4 网络模型结构如图 3 所示。

该检测模型主要由 CBM(convolution,batch normalization and mish)、CBL(convolution,batch normalization and Leaky ReLU)、CSP*X*(center and scale prediction)、SPP(spatial pyramid pooling)和 ResUnit 等基础组件组成。

CBM 是 YOLO v4 网络结构中的最小组件,由 Conv+BN+Mish 激活函数组成,Conv 是卷积层,用来提取特征,BN 用于归一化处理,Mish 为激活函数,允许更好的信息深入神经网络,从而得到更好的准确性和泛化;CBL 由 Conv+BN+Leaky_ReLU 激活函数组成,与 CBM 不同的是采用了 Leaky_ReLU 激活函数,在反向传播过程中,其输入小于 0 的部分,也可以计算得到梯度,避免了梯度方向出现锯齿的问题;ResUnit 借鉴网络中的残差结构,可以构建更深的网络,解决了深度神经网络中给网络叠加更多层后,网络深度在增加性能却快速下降的问题;CSP*X* 借鉴 CSPNet 网络结构,由 3 个卷积层和 *X* 个 ResUnit 模块 Concat 而成,在减少计算量的同时实现更丰富的梯度组合;SPP 采用 1×1、5×5、9×9、13×13 的最大池化方式进行多尺度融合,大大增加了感受野。

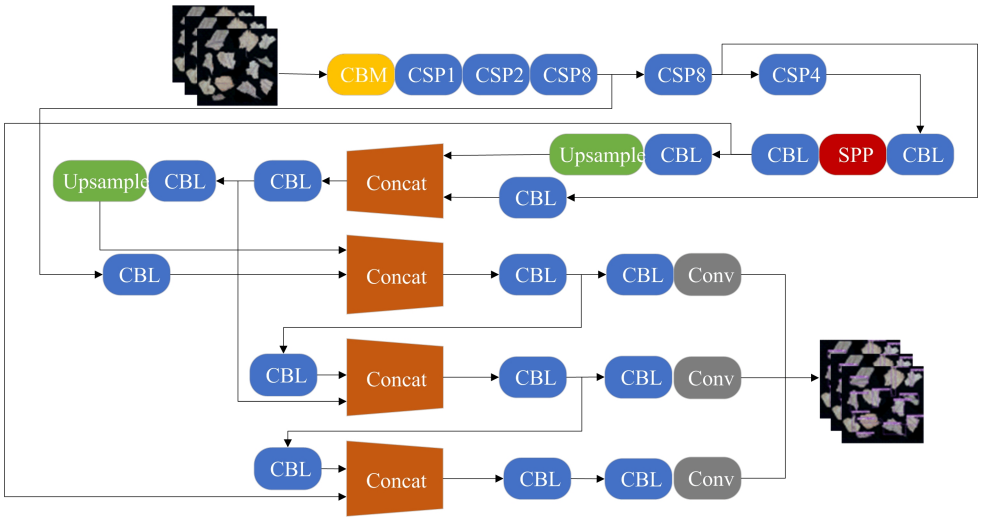


图 3 YOLO v4 网络模型结构图

2.2 损失函数

损失函数在模型中起到关键性作用,可以衡量模型预测的好坏,用来表现预测与实际数据之间的差距程度。基于 YOLO v4 的烟叶中烟梗检测模型的损失函数由回归框损失函数($loss_{CIoU}$)、置信度损失函数($loss_{conf}$)和分类损失函数

($\text{loss}_{\text{class}}$)组成。损失函数计算公式为：

$$\text{loss} = \text{loss}_{\text{CIoU}} + \text{loss}_{\text{conf}} + \text{loss}_{\text{class}} \tag{1}$$

回归框损失值为：

$$\text{loss}_{\text{CIoU}} = 1 - I_{\text{oU}} + d^2/f^2 + \alpha\nu \tag{2}$$

$$\alpha = \frac{\nu}{(1 - I_{\text{oU}}) + \nu} \tag{3}$$

$$\nu = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{\text{gt}}}{h^{\text{gt}}} - \arctan\frac{\omega}{h}\right)^2 \tag{4}$$

置信度损失值为：

$$\begin{aligned}\text{loss}_{\text{conf}} = &-\sum_{i=0}^{A^2}\sum_{j=0}^{B} I_{ij}^{\text{obj}}[\hat{e}_i^j \ln e_i^j + \\ &(1-\hat{e}_i^j)\ln(1-e_i^j)] + \\ &\lambda_{\text{noobj}}\sum_{i=0}^{A^2}\sum_{j=0}^{B} I_{ij}^{\text{noobj}}[\hat{e}_i^j \ln e_i^j + \\ &(1-\hat{e}_i^j)\ln(1-e_i^j)]\end{aligned} \tag{5}$$

分类损失值为：

$$\begin{aligned}\text{loss}_{\text{class}} = &\sum_{i=0}^{A^2} I_{ij}^{\text{obj}} \sum_{c\in\text{classes}} [\hat{p}_i^j(c)\ln(p_i^j(c)) + \\ &(1-\hat{p}_i^j(c))\ln(1-p_i^j(c))]\end{aligned} \tag{6}$$

其中：α 为权重系数；ν 为长宽相似比；A 为网格数量；B 为每个网格中先验框数量；I_{ij}^{obj}、I_{ij}^{noobj} 为判断权重；ω、h 分别为预测框宽、高；ω^{gt}、h^{gt} 分别为真实框宽、高；d 为预测框与真实框中心点间欧氏距离；f 为预测框与真实框所组成最小闭包区域对角线距离；$\hat{p}_i^j(c)$ 为第 i 个网格的第 j 个先验框物体为类别 c 的概率真实值；$p_i^j(c)$ 为第 i 个网格的第 j 个先验框物体为类别 c 的概率预测值；$\hat{e}_i^j$ 为第 i 个网格的第 j 个先验框实际类别；e_i^j 为第 i 个网格的第 j 个先验框预测类别；I_{oU} 为预测框与真实框交集与并集比值。

2.3　基于改进 YOLO v4 的烟梗检测模型

基于 YOLO v4 的烟梗检测模型可实现松散回溯后破碎烟叶中烟梗的识别，但基础训练模型储存空间较大，模型参数量较多，且模型推理速度较慢。基于上述问题，本文提出了基于通道剪枝和层剪枝相结合的模型轻量化剪枝方案，将 YOLO v4 训练后得到的模型进行裁剪，该操作能够大大减小模型存储量和计算量，显著提高模型推理速度，且能保证模型精度。

通道剪枝本质上是指在模型经过稀疏训练后，消除 BN 层 γ 值较小的通道，即输入通道中贡献率较低的通道。通道剪枝在保证网络结构完整的情况下降低了模型参数量，减小了模型存储空间，且不需要特定的网络框架或硬件支持，使其更方便部署在移动设备上。

通道剪枝算法原理如图 4 所示。烟梗识别模型采用通道修剪 γ 值来评估其通道的重要性。假设 γ 值接近于 0，输出通道也接近于 0，因此 γ 值所在的通道对整个网络的影响很小，可以进行修剪。

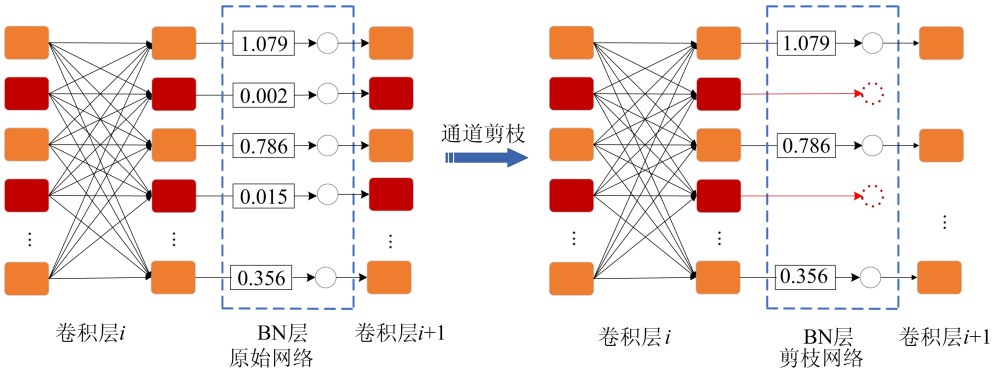


图 4　通道剪枝算法原理图

对冗余通道进行剪枝，首先找到要剪枝的 BN 层的指标值，然后将 γ 的绝对值提取成一个列表，从最小到最大值排序，最后定义一个 p 值来控制 BN 层的剪枝率，并定义一个全局阈值 $\hat{\gamma}$ 来确定待剪枝 γ 值的上限，其大小为列表中所有 γ 值的第 p 个百分位。此外，对每个通道的 γ 值进行排序，取最大的 γ 值作为 γ 值剪接的上限，避免了对所有信道进行剪接，保证了网络的完整性。当修剪通道的 γ 值等于 0 时，修剪模型的性能与稀疏训练模型相似，几乎等于基本训练模型。

在对烟梗特征提取网络进行通道修剪后，进一步去除多余的残差块，如图 5 所示。在修剪冗余残差块时，首先对所有 Shortcut 层的前一个 CBL 的 γ 值的均值进行排序。若它接近于 0，则该残留块没有学习到有用的信息，那么对其进行剪枝不会影响网络性能，因此 γ 值均值较小的残

差块将被修剪。为保证网络的完整性,每剪一个Shortcut结构,会同时剪掉一个 Shortcut 层及其前面的 2 个卷积层。本文一共剪掉 16 个 Shortcut,共计剪掉了 48 个层,精度下降较少。

图 5 修剪冗余残差块流程

稀疏训练后,许多 γ 值收敛到 0,因此对模型的冗余通道和网络层进行剪枝,减少了模型的参数、大小和推理时间。

2.4 模型剪枝具体步骤

本研究所用计算机型号为 GPU 1080Ti、CPU E5-2678v3、内存大小 62 GiB、显存大小 11 GiB,开发环境为 Python 3.6+PyTorch 1.9.0+openc v4.5.5+CUDA 11.4。模型剪枝具体步骤模型如下:

1) 基础训练。剪枝前,先对数据集进行基础训练,利用 YOLO 官网给出的初始权重训练数据集,初始学习率(learning rate)设置为 0.002 3,批样本量(batch size)为 8,迭代次数(epoch)为 200,基础训练情况如图 6 所示。

从图 6a 可以看出,随着迭代次数的推进,基础训练趋于 150 次时损失值趋于平缓,最终模型训练 200 次 loss 值稳定在 0.5 左右,模型收敛。从图 6b 可以看出,基础训练模型精度在前 40 次迭代稳步上升,到达 40 个 epoch 时达到 0.850,在接下来的 160 次迭代中平均精度均值(mean average precision,mAP)平缓上升,模型最终收敛精度达到 0.941。

2) 稀疏训练。烟梗识别模型经过基础训练后包含大量冗余通道,且通道数量在不同 BN 层间存在显著差异。本文采用稀疏训练方法来解决通道数量差异较大的剪枝问题,为后续模型剪枝作准备。通过将 BN 层 γ 值的 L1 归一化项添加到损失函数中,使 γ 值变得稀疏,L1 归一化消除最终没有为输出提供有用信息的通道。稀疏因子 s 设置为 0.001,批样本量(batch size)为 8,迭代次数(epoch)为 500,初始学习率(learning rate)为 0.002 3,在稀疏训练迭代次数达到 70%和 90%的 2 个阶段进行 γ 值为 0.1 的学习率衰减,使模型精度有所回升。

从图 6a 可以看出,基础训练得到的模型进行稀疏训练后,整体 loss 曲线呈先上升后下降趋势,整体存在较大波动,在迭代次数为 220 时达到峰值,随后下降且最终趋于平缓,loss 值最终停靠在 1.0 左右且存在细微波动。从图 6b 可以看出,稀疏训练模型精度继承基础训练的最终精度 0.940且在稀疏训练的前 120 次的迭代中趋于稳定,在迭代次数为 120 后模型精度发生了较大跨越,最终在迭代次数为 360 后趋于平缓,mAP 值在 0.700 左右且存在细微波动。

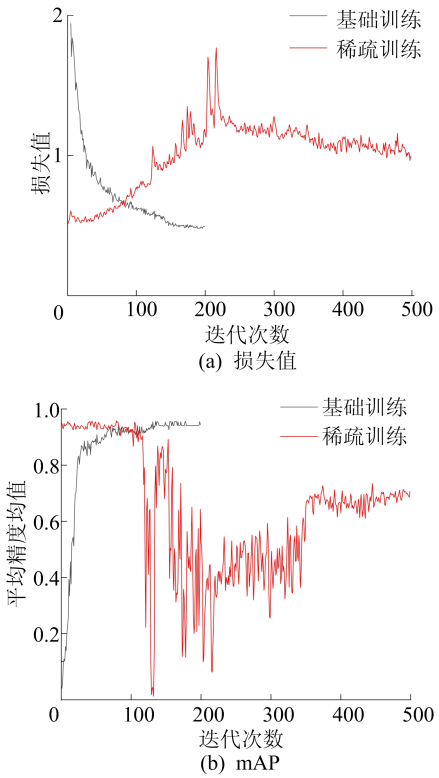


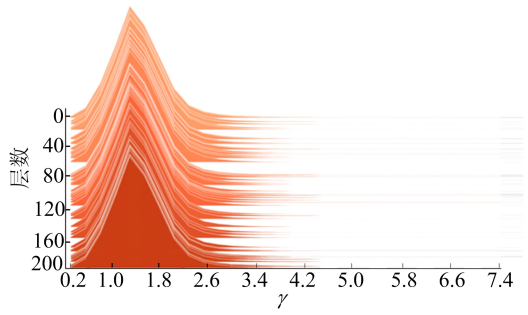
图 6 基础训练、稀疏训练模型损失值和精度值变化曲线

模型训练过程中权值分布直方图如图 7 所示,直方图表示模型训练过程中神经网络中每一层的权重。图 7 中:纵坐标值的增加表示模型 BN 层的递增,用来反映模型训练过程的推进;横坐标 γ 反映模型的稳定性和稀疏性,随着训练迭代次数即纵坐标值的增加,γ 值分布范围逐渐相似,说明模型趋于收敛,γ 值趋于 0 说明模型结构逐渐稀疏。

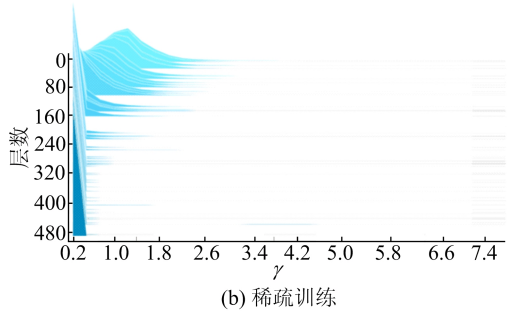
图 7a 为 YOLO v4 基础训练模型各 BN 层 γ 值分布中心的变化趋势图。从图 7a 可以看出,γ 值随着训练迭代次数的增加始终分布在 1 左右,γ 值始终趋于稳定,说明基础训练一直处于稳定状态。

图 7b 为 YOLO v4 稀疏训练模型各 BN 层 γ 值分布中心的变化趋势图。从图 7b 可以看出,γ

值在开始迭代时分布在 1 附近，随着训练迭代次数的增加逐渐趋于 0，说明 γ 值逐渐变得稀疏，当迭代次数达到 500 左右时 γ 值趋于稳定，说明此时稀疏训练已处于稳定状态。



(a) 基础训练



(b) 稀疏训练

图 7　YOLO v4 模型基础训练、稀疏训练 γ 值直方图

3) 模型剪枝及微调。对稀疏训练后的模型设置 80%的通道剪枝全局比例和 0.01 的每层最低保持通道数比例，剪掉 12 个 Shortcut 及其前面 24 个 CBL，总共剪掉 36 层。

剪枝结果见表 1 所列。

表 1　剪枝前后模型参数

模型	参数量/个	存储空间/MB	mAP/%	检测时间/s
原模型	63 937 686	244.46	94.10	0.021 74
通道剪枝模型	3 009 775	11.05	72.10	0.021 46
层剪枝模型	2 881 039	10.02	72.30	0.016 76
微调模型	22 746 510	15.22	93.00	0.210 00

从表 1 可以看出，模型经过通道与层剪枝后参数量下降了 95.49%，存储空间减少了 234.44 MB，模型检测时间减少了约 0.005 s，模型 mAP 降低了 21.8%。鉴于模型剪枝后精度出现下跌，对剪枝模型进行 100 次迭代基础训练的微调，批样本量(batch size)设置为 8，以补偿模型性能的下降。微调后模型存储空间略有增加，但控制在较小范围内。

模型精度回升到与剪枝前相近，且检测速度仍能维持在模型剪枝前水平。结果表明，对模型进行剪枝及微调可在保证模型精度及检测速度的前提下大幅简化模型。

3　实验结果与分析

3.1　烟梗检测效果分析

基于本文剪枝微调后的 YOLO v4 优化模型烟梗检测效果如图 8 所示。

从图 8 可以看出，在实验背景下能够准确地识别出不同破碎烟叶片中存在的烟梗，识别精度高，检测速度快，适用于卷烟自动化工厂中识别剔除带梗烟叶工艺步骤，能代替现阶段烟厂工人经验化识别剔除步骤，减少劳动力，提高识梗去梗准确率和成品烟丝品质。

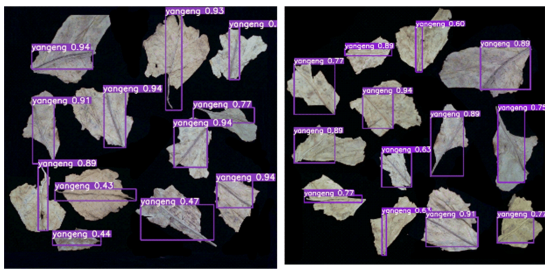


图 8　模型检测效果图

3.2　模型性能评价指标

选用 mAP、精确率 P、召回率 R、$F1$ 值、模型大小以及模型检测时间等指标对不同模型进行性能评估。

3.3　不同模型烟梗检测效果比较

基于 YOLO v4 剪枝模型对烟叶图像进行测试，精确率为 90.98%，召回率为 95.46%，$F1$ 值为 93.17%，mAP 为 93.00%，检测单张图片时间达到 0.021 00 s。模型识别精度高、识别速度快，且占用空间相比同类模型得到极大压缩，为烟厂片烟精选工作提供有利的技术支持。

与同类别的目标检测模型相比，本文使用的方法更具备优势。常用的目标检测模型有 SSD[12]、YOLO v3[13]、YOLO v4[14]、YOLO v4-tiny[15]、YOLO v5[16]等，在同样的硬件及开发环境下，使用这 5 种模型分别进行测试，但得结果见表 2 所列。

比较几种重要的模型评价指标可以看出，本文 YOLO v4 剪枝模型的召回率和 $F1$ 值均最高，其精确率较 Sota 方法的 YOLO v5 模型虽下降 0.21%，但模型 mAP 相较剪枝前下降了 1.10%，变化较小，但仍保持在较高的精度水平，模型检测

速度略有提升,同时模型存储大小得到了极大压缩,甚至比 YOLO v4-tiny 轻量化模型还要小 7.25 MB,应用于工业现场可极大减小部署设备储存空间。剪枝模型检测速度虽相较于 SSD、YOLO v3、YOLO v4-tiny 略有下降,但最大检测时间差控制在 0.008 00 s 左右,对工业应用现场影响不大,且与此 3 种模型相比,本文方法的模型精度分别提升了 9.31%、4.06%、13.64%。

表 2 不同模型性能评估

模型	P/%	R/%	$F1$ 值/%	mAP/%	存储空间/MB	检测时间/s
SSD	82.97	86.32	84.61	83.69	121.82	0.012 93
YOLO v3	88.53	92.86	90.64	88.94	302.52	0.019 76
YOLO v4-tiny	78.47	83.64	80.97	79.36	22.47	0.015 37
YOLO v4	90.23	91.76	90.99	94.10	244.46	0.021 74
YOLO v5	91.19	93.64	92.32	94.89	263.78	0.021 94
YOLO v4 剪枝	90.98	95.46	93.17	93.00	15.22	0.021 00

综合比较,本文模型剪枝方法与同类目标检测模型相比,在极大压缩模型存储空间的基础上,仍能保持较高的模型精度以及模型检测速度,表明本文方法在片烟精选工业现场中的烟梗识别剔除应用中具有很强的应用前景。

3.4 不同剪枝率和层数模型效果对比

本文方法的模型剪枝过程中,涉及到人为设置通道剪枝全局比例以及层剪枝 Shortcut 层数这 2 个重要参数,不同参数下模型剪枝效果存在差异。本文比较了不同剪枝参数下模型的性能,结果见表 3 所列。以通道剪枝全局比例 85%和 80%以及层剪枝层数 12 和 16 分别对 YOLO v4 基础模型进行剪枝操作,分析模型剪枝结果。

在模型精度方面,剪枝后精度均出现一定幅度的下降,但整体降幅不大,全局比例为 80%,层剪枝层数为 12 时,模型精度下降幅度最小,相比于原模型,mAP 仅下降 1.10%,说明较小的通道剪枝比例和层剪枝层数对模型精度影响更小。

表 3 不同剪枝率和剪枝层数模型性能对比

全局比例/%	层剪枝 Shortcut 层数	mAP/%	参数量/个	存储空间/MB	检测时间/s
85	12	92.10	1 978 411	10.59	0.200 00
80	12	93.00	2 881 039	15.22	0.210 00
85	16	90.30	1 791 471	9.88	0.015 00
80	16	91.30	2 733 567	14.48	0.014 00

在模型存储空间方面,剪枝后相比于原模型均得到大幅度压缩,且通道剪枝比例和层剪枝层数越大,对模型的压缩力度也越大,模型最小能压缩到 9.88 MB,仅占原模型空间的 4.0%。在模型检测速度方面,不同参数组合下的剪枝,模型检测均比原始模型有所提升,相比原始模型检测时间最快缩短了 0.080 00 s。最终本文选用 80%的通道剪枝全局比例及 12 个 Shortcut 的层剪枝层数作为 YOLO v4 基础模型的剪枝参数,在实现 93.00%模型高精度的同时仍能维持较高的检测速度,且极大降低了模型存储空间。

4 结 论

本文基于通道剪枝和层剪枝相结合的模型压缩方案,提出了一种基于改进 YOLO v4 的轻量化烟梗识别方法。在保证模型精度和检测速度的前提下,对模型进行大幅度压缩,简化了模型结构,减小了占用空间。

本文使用英特尔 RealSense d435i 型号相机采集原始烟叶数据集 100 张,利用 Mosaic 数据增强的方法拓展数据集到 200 张,运用 Labelimg 对数据集进行标注,随后运用 YOLO v4 算法对数据集进行基础训练,得到基础模型后进行稀疏训练、模型剪枝、微调,最终得到优化后的烟梗识别模型。经实验验证,优化模型的存储空间下降了 93.77%,模型精度达到 93.00%,检测速度达到 47.62 帧/s。实验结果表明,剪枝算法在保证检测速度快、精度高的同时极大地压缩了模型的大小,为算法模型在工业现场的部署提供了优势,有效解决了传统工控机存储空间有限的问题,适用于片烟精选中松散回溯后烟梗识别的工艺流程。

(下转第 1253 页)

上海交通大学出版社，2015：39-41.
[14] 赵鑫，宋英强，胡月明，等. 基于多源开放数据的城乡居民点空间布局优化[J]. 广西师范大学学报(自然科学版)，2020，38(1)：26-40.
[15] LI T，JING P，LI L C，et al. Revealing the varying impact of urban built environment on online car-hailing travel in spatio-temporal dimension：an exploratory analysis in Chengdu，China[J]. Sustainablity，2019，11(5)：1336.
[16] LAWSON A B. Bayesian disease mapping：hierarchical modeling in spatial epidemiology[M]. 2nd ed. London：Taylor & Francis，2013：13-15.
[17] 孙酩皓，申力，高博轩，等. 基于 GTWR 模型的陕西省 HFRS 发病影响因素分析[J]. 现代预防医学，2020，47(23)：4230-4234，4280.
[18] YANG H T，LU X Z，CHERRY C R，et al. Spatial variations in active mode trip volume at intersecions：a local analysis utilizing geographically weighted regression[J]. Journal of Transport Geography，2017，64：184-194.
[19] CAMPBELL A A，CHERRY C R，RYERSON M S，et al. Factors influencing the choice of shared bicycles and shared electric-bikes in Beijing[J]. Transportation Research Part C：Emerging Technologies，2016，67：399-414.
[20] 卢宾宾，葛咏，秦昆，等. 地理加权回归分析技术综述[J]. 武汉大学学报(信息科学版)，2020，45(9)：1356-1366.
[21] 谢蔚翰，周素红. 建成环境对出租车出行需求影响的时空间分异模式[J]. 现代城市研究，2018(12)：22-29.

（责任编辑　张淑艳）

(上接第 1202 页)

[参　考　文　献]

[1] 中国烟叶公司. 烟叶 打叶复烤 工艺规范：YC/T 146—2010[S]. 北京：中国标准出版社，2011.
[2] 朱文魁，刘斌，毛伟俊，等. 基于低能 X 射线透射成像的打叶片烟中烟梗在线检测[J]. 烟草科技，2015，48(2)：69-74.
[3] 席建平，易浩，刘斌，等. 基于 FPGA 的烟梗在线检测系统设计[J]. 中国烟草学报，2016，22(5)：50-54.
[4] 刘赐德. 基于 X 射线透射成像的叶中含梗率在线检测技术研究[D]. 昆明：昆明理工大学，2019.
[5] 汤龙文. 基于透光性差异的烟梗检测分析及算法实现[J]. 景德镇学院学报，2017，32(6)：26-29.
[6] DAI J，LI Y，HE K，et al. R-FCN：object detection via region-based fully convolutional networks[C]//Conference on Neural Information Processing Systems. [S. l.：s. n.]，2016：1-9.
[7] GIRSHICK R. Fast RCNN[C]//IEEE International Conference on Computer Vision. [S. l.]：IEEE，2015：1440-1448.
[8] HE K，GKIOXARI G，DOLLAR P，et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，42(2)：386-397.
[9] 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.
[10] REN S，HE K，GIRSHICK R，et al. Faster RCNN：Towards real-time object detection with region proposal networks[C]//Conference on Neural Information Processing Systems. [S. l.：s. n.]，2015：1-9.
[11] SERMANET P，EIGEN D，ZHANG X，et al. Overfeat：Integrated recognition，localization and detection using convolutional networks[EB/OL]. (2014-02-24)[2023-01-02]. https://arXiv：1312. 6229v4.
[12] LIU W，ANGUELOV D，ERHNA D，et al. SSD：Single shot multibox detector[C]//European Conference on Computer Vision. [S. l.：s. n.]，2016：21-37.
[13] BOCHKOVSKIY A，WANG C Y，LIAO H Y M. YOLOv4：optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2023-01-02]. https://arXiv：2004. 10934.
[14] REDMON J，FARHADI A. YOLOv3：an incremental improvement[EB/OL]. (2018-04-08)[2023-01-02]. https://arXiv：1804. 02767.
[15] JIANG Z，ZHAO L，LI S，et al. Real-time object detection method based on improved YOLOv4-tiny[EB/OL]. (2020-11-09)[2023-01-02]. https://arXiv：2011. 04244.
[16] JOCHER G. YOLOv5[EB/OL]. (2020-06-10)[2023-01-02]. https://github. com/ultralytics/yolov5.

（责任编辑　李　凯）