文章编号:1671-6833(2022)05-0031-08

基于 YOLOv5s 的金属制品表面缺陷的轻量化算法研究

贾云飞, 郑红木, 刘闪亮

(中国民航大学 电子信息与自动化学院, 天津 300300)

摘 要: 为解决企业降低智能化成本的要求,运用低成本、低算力的硬件设备,通过深度学习中目标检测算法模型对产品进行缺陷检测。基于深度学习目标检测中的 YOLOv5s 网络,采用结构裁剪思想,并基于网络中的 BN 层对网络进行稀疏训练,将稀疏训练后的模型对应权重值较小的层进行裁剪,从而降低模型的计算参数数量以及模型文件大小,达到轻量化的效果。使用 NVIDIA 的加速推理框架 TensorRT 对训练好的裁剪模型进行层级融合,实现推理加速效果。实验结果表明:所提目标检测模型相对于原始 YOLOv5s 模型权重文件大小降低约 70%,同时在公开数据集 NEU-DET 上检测精度达到了 74.2%。在搭建的高性能实验台中单图推理速度相比原模型提升了 11.3%,且网络没有精度损失;在低性能实验台中,所提模型相比原网络模型推理速度提升了 165%,相比高性能实验台中的结果有了更显著的提升,说明所提模型在低算力硬件设备中表现优秀。再针对所提模型采用公开的潜水泵叶轮的俯视图数据集进行普适性测试,最后对所提模型采用推理加速框架 TensorRT 进行加速后,在高性能实验台上可以达到单图 5.8 ms 的推理时间。所提目标检测模型在低算力硬件设备上推理速度提升较大,可以帮助企业降低预算。

关键词: 表面缺陷; 目标检测; 轻量化; YOLOv5s; TensorRT

中图分类号: TP391.7 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2022.05.001

0 引言

当今,计算机视觉中目标检测技术飞速发展,这一技术在工业产品缺陷检测中的应用已经十分广泛,它不仅可以帮助企业进行产品质量把关,还可以降低人力成本,对企业的发展十分有利。目标检测网络主要分为 two-stage 网络和 one-stage 网络,其中 two-stage 网络中的经典网络如 SPP-NET[1]、R-CNN[2]、Faster R-CNN[3]等。这些网络通过算法生成一系列预选框,同时进行区域分类,然后对预选位置进行修正,精确目标位置。而 one-stage 网络如 YOLOv3[4]、EfficientDet[5]、SSD[6]等网络则可以联合区域提案和分类预测,对输入网络的图像特征进行提取并直接输出目标的类别和位置信息[7]。现阶段,由于 one-stage 网络的高效性,该类网络发展迅速。郝用兴等[8]通过对 Faster R-CNN 中提取网络,并在特征提取网络中对 VGG16 和 Resnet50 进行模型融合,在瑕疵大小分布不均匀的铝材表面瑕疵检测上更具优势;程婧怡等[9]针对金属表面目标尺寸小和特征不清晰等问题,通过在 YOLOv3 网络中加入 DIOU 边框回归损失和 *K*-means++聚类分析筛选最优回归框,该算法在定位金属板表面小缺陷目标的精度上更具优势;Li 等[10]通过引入 MobileNet 网络对 SSD 的特征提取网络进行改进,在电子产品表面缺陷检测上取得了不错的准确率以及检测速率。

上述研究分别对不同的目标检测网络进行改进并对工业产品或电器设备进行缺陷检测,达到了可观的效果,但没有针对企业内需要体积更小、计算参数更少的模型进行研究,这样的模型在成本预算低、算力相对不足的设备中也能达到不错的检测速度。基于此,本文结合深度学习目标检测的方法,在原有的轻量级 YOLOv5s 网络的基础上,通过结构裁剪的思想,基于网络的 BN(batch normalization)层进行模型裁剪,得到更轻量级模型的同时也具备了优异的检测效果,并对裁剪后的模型进行训练,并使用 TensorRT 进行推理加速,进一步提高了模型的检测速率。

收稿日期:2021-12-02;**修订日期**:2022-03-19

基金项目:中央高校基本科研业务费资助项目(3122019185)

作者简介:贾云飞(1979—),男,河北石家庄人,中国民航大学副教授,博士,主要从事人工智能研究,E-mail: yfjia@cauc.edu.cn。

1 YOLOv5目标检测模型

1.1 YOLO系列算法简介

目标检测算法YOLO由Redmon等[11]在2016年提出。其结构包含了24个卷积层和2个全连接层,输出层的激活函数为线性函数,其他激活函数为ReLu函数。其核心思想在于将整张图输入网络,输出层会直接输出对目标边界框的位置预测和类别回归。之后在此基础上提出了采用Darknet-19作为主干网络的YOLOv2算法模型,YOLOv2相比YOLOv1改进了准确度、预测对象广度并具有更快的速度[12]。而之后的YOLOv3则更是成为了YOLO系列的经典算法,其采用Darknet-53作为主干网络,包含53个卷积层,且层间残差结构连接,组成残差层[4]。YOLOv4和YOLOv5则是在YOLO系列原作者退出研究之后由后人继续研究完成,对比之前的YOLO系列网络改动相对较小,其中YOLOv5s模型大小仅有27 MB,相比之前的YOLO模型大幅度减小。

1.2 YOLOv5s网络模型

YOLOv5s的特征提取网络为CSPDarknet(cross stage partial darknet),由CSP结构和Darknet网络相结合构成。其网络结构主要分为3部分,即Backbone、Neck和Head,其中Backbone部分主要进行输入图像特征提取;Neck部分对图像特征进行融合;Head部分对图像特征进行结果预测并输出[13]。网络模型中的CBL结构主要由提取特征的卷积层、BN层和Leaky ReLu激活函数构成,如图1所示。YOLOv5s还设计了2种CSP网络结构,其中CSP1_N用于Backbone中,CSP2_N用于Neck部分,并且采取交叉结构连接,可以在减少计算量的同时保证运算的精度。

2 YOLOv5s模型裁剪

2.1 BN层概述

在卷积神经网络中,网络通过前向传播和反向传播共同作用来进行权值更新,在进行一次反向传播时,网络会对之前的前向传播训练得到的参数进行修正,此时各层网络参数权重会同时更新。由于每层的输入都会受到前面所有层输入的影响,网络参数微小的变化都会随着网络深入而放大,所以需要较低的学习率和初始参数,从而降低了学习效率。2015年,Google提出了BN层,通过这个算法对网络各层的输入进行归一化,将输入的批数据进行均值为0、方差为1的归一化处理[14],对各层输入数据分布进行了控制,但对网络上一层的输出数据经过归一化处理之后送入下层网络的数据,在上层网络学习的特征会因为数据的分布变化而遭到破坏,因此BN层添加了缩放参数γ和平移参数β,来恢复上层网络学习的特征。因此,BN层的总体计算过程大致如下:首先,对输入数据进行求均值和方差;其次,对数据进行标准化处理;最后,训练缩放参数和平移参数,输出通过参数线性变换得到新值。通过这种方法使得每层数据分布相同,能提高网络学习效率,具体公式如下:

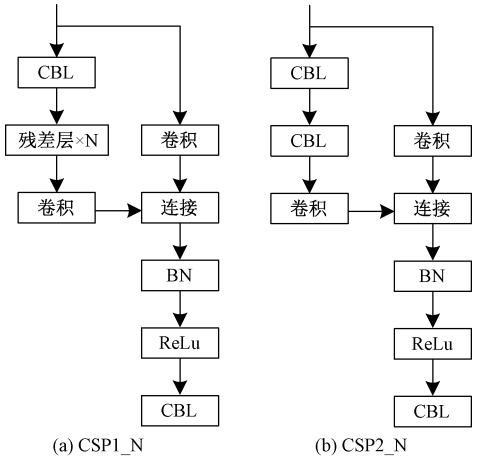


图1 YOLOv5s中的2种CPS结构

Figure 1 Two CSP structures in YOLOv5s

$$\mu = \frac{1}{m}\sum_{i=1}^{m} z^{[i]}; \tag{1}$$

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} (z^{[i]} - \mu)^2; \tag{2}$$

$$\tilde{z}^{[i]} = \frac{z^{[i]} - \mu}{\sqrt{\sigma^2 + \varepsilon}}; \tag{3}$$

$$y^{[i]} = \gamma \tilde{z}^{[i]} + \beta。 \tag{4}$$

式中:$z^{[i]}$和$y^{[i]}$分别表示batch中的第i个样本经过BN层的输入与输出;m为batch中的样本数;μ和σ^2分别为batch中输入样本的均值和方差;ε为常数;$\tilde{z}^{[i]}$为对输入的$z^{[i]}$归一化的结果;γ和β为BN层的训练参数。

2.2 网络模型裁剪

一般来说,神经网络中层数越深,网络结构越精细,通过训练网络得到的输出结果也越精准。但与此同时,也带来了计算参数更多、计算量更大的弊端,而且在训练数据量较少的情况下,模型相对数据集复杂度过高,容易产生过拟合现象,虽然产生的网络更深更精细,但最终训练模型的效果反而不佳。与ImageNet数据集中1 400多万张样

本数据相比,工业产品的表面瑕疵检测面临样本小、图像样本效果不佳等问题,所以在工业检测领域,通过模型剪枝不仅能够解决网络计算量大的问题,还可以解决网络过拟合现象,提升计算速度,在降低训练参数量大小的同时,提高网络模型的精确度。

神经网络经过训练后得到模型权重,通过前向传播和反向传播生成了一系列的权重参数,在使用验证数据进行推理时,输入数据进入神经网络并通过上述参数进行推理计算,而其中的参数有一部分权重占比小,对输出结果不起作用或是仅仅起到微小的作用,而这样的参数存在对计算结果虽然影响很小,但却加大了计算量,因此对其进行裁剪可以使得输入数据通过最有效率的参数空间路径得出输出结果,在提高计算效率的同时,也几乎不会降低网络推理效果,且对数据集规模小的数据还能降低训练过拟合带来的影响,提高网络的推理效果。神经网络的裁剪方法具体可分为非结构裁剪和结构裁剪。非结构裁剪是一种细粒度的裁剪方法,精度相对较高,但对硬件平台和特定算法依赖程度较高;而结构裁剪是针对层级的裁剪,是一种粗粒度的裁剪方法,对硬件平台和特定算法依赖程度低,可行性强,因此,本文采用结构裁剪的方法对 YOLOv5s 进行裁剪。

对原始 YOLOv5s 网络进行训练后,模型 BN 层的 γ 值一般情况下呈现近似正态分布,难以对训练后的网络模型进行裁剪,所以在模型裁剪之前需要先对 YOLOv5s 网络 BN 层中的损失函数的 γ 值添加 L1 正则化约束。通过 L1 正则化约束可以使得网络模型稀疏化,通常也称作稀疏训练,此时损失函数:

$$J = \sum_{(a,b)} J_0(f(a,w),b) + \alpha \sum_{i=1}^{n} |\gamma| 。 \quad (5)$$

式中:a 表示输入数据;b 表示输出数据;w 表示训练权重参数;J_0 表示网络每层的模型损失函数;$\alpha \sum_{i=1}^{n} |\gamma|$ 表示添加的 L1 正则化约束项;α 表示平衡损失系数;n 表示模型中全部 BN 层参数。

对于本文所研究的工业瑕疵检测这类问题的小样本数据集来说,稀疏训练可以使得网络中很多参数分布接近于 0,降低了训练后的模型复杂度,从而可以减少过拟合现象在训练数据集上的表现,提升模型在验证数据集上准确率。由文献[15]中的实验结果可以看出,当采用复杂度更高的 YOLOv5 模型时会在数据集上表现出过拟合现象,导致在验证集上模型预测准确率降低。文献[16]中的实验表明了稀疏训练的模型可以学习到更泛化的特征,具备鲁棒性和易用性。在对网络模型进行稀疏训练后,BN 层中的一部分 γ 值会趋向于 0 分布,这部分权值对推理结果影响甚微,此时可以对其所在的 BN 层以及相邻上层的卷积核和向下层输出的通道进行裁剪,以此降低模型参数数量,实现模型的轻量化,γ 裁剪过程如图 2 所示。

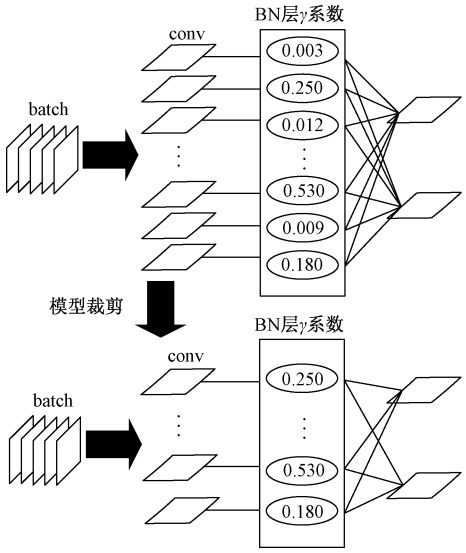


图 2　γ 裁剪过程

Figure 2　γ pruning process

3　基于 TensorRT 的推理加速

TensorRT(以下简称 TR)是 NVIDIA 基于深度学习加快推理的高性能平台。TR 可应用于图像分类、目标检测和图像分割等计算机视觉领域的相关推理任务中。TR 作为可编程的推理器,能够为训练好的模型提供推理加速[17]。TR 包含了构建和部署 2 个阶段,构建阶段是 TR 针对网络配置进行优化,产生一个优化决策并用于网络的前向传播,其对神经网络进行了一系列转换和优化,其中优化包括删除未使用的网络层以避免不必要的计算,对多个层进行融合形成单层。当训练好的模型进行网络推理时,模型网络的各个层都会进行函数的调用,被调用的函数将在硬件设备上执行,而这些函数大多计算速度快,当多次调用函数时,对推理速度起决定性影响的则是启动时间,函数的频繁调用启动会导致更多的内存传输,而 TR 将网络和张量进行了融合操作,融合包括垂直方向和水平方向 2 种方式,其中垂直融合将 conv 层、BN 层或偏移层、激活层整合为一个

CBR;水平融合则是使得几个输入相同(例如1×1CBR)融合在融合后网络层数减少,降低了网络的深度和宽度,并且可以减少对函数的调用,提升推理速度。除此之外,TR还可以将连接层移除,将结果输出到预先分配的缓冲区直接输入到下一网络层中。部署阶段则是设备通过长时间运行程序或服务,等待待处理数据的提交,当收到待处理数据时,执行构建阶段产生的优化决策并进行推理,将结果通过设备输出。

4 实验方案

4.1 实验数据来源

本文实验数据来源为东北大学采集并发布的NEU-DET钢表面6种缺陷的数据集,这6种缺陷分别为氧化皮、斑块、开裂、麻点状表面、杂质和划痕,共计1 800张图像,其中每一类缺陷包含300个样本。将1 800张图像按照8∶2的比例随机分为训练集和验证集,即训练集1 440张,验证集360张,样例如图3所示。

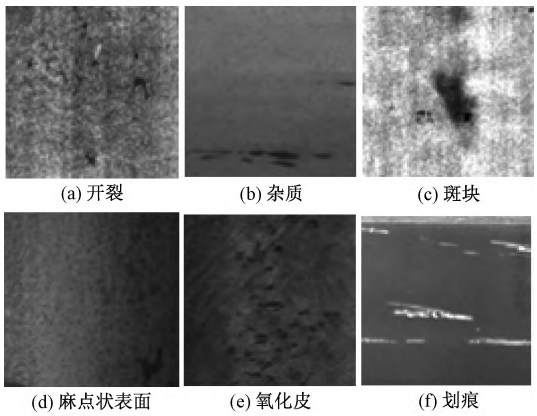


图3 数据集样例图

Figure 3 Sample dataset diagram

4.2 实验平台搭建

本实验采用高性能实验台和低性能实验台来进行实验,高性能实验台配置为ubuntu20.04LTS操作系统,硬件方面采用的处理器为Intel Xeon Gold 5218R,GPU为NVIDIA Quadro RTX5000。低性能实验台配置操作系统为windows10,处理器为Intel® Core™ i5-8265U,GPU为NVIDIA GeForce MX250。实验使用框架Pytorch(1.7版本),GPU加速计算方面采用CUDA10.1、CUDNN7.6.0和若干第三方Python库支持模型训练及运算。

4.3 实验流程设计

首先,需要对YOLOv5s网络中添加L1正则化约束;其次,在搭建好的训练平台进行模型稀疏训练,减小BN层的γ分布,针对接近0分布的γ值,对相应的BN层及其相邻卷积核和卷积核通道进行裁剪,得到裁剪后的新模型;再次,对该模型再进行训练,得到最终的训练模型并进行TR加速;最后,在上述搭建好的实验平台上进行模型推理测试,实验流程如图4所示。

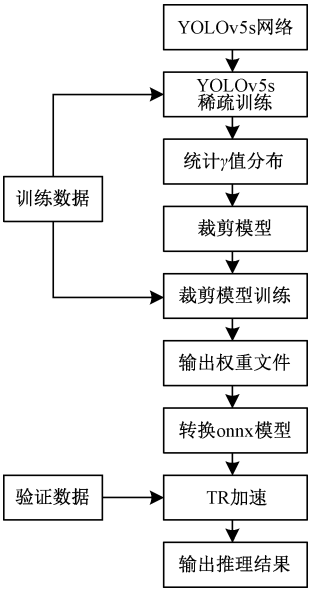


图4 本文模型实验流程图

Figure 4 Flow chart of the experiment of this model

5 实验结果与分析

5.1 评估指标及目标识别过程

目标检测作为机器视觉的一个任务分支,评估指标除了常规有监督机器学习中的精确率P和召回率R之外,还使用各类别平均精确度均值(AP)作为网络模型训练结果好坏的评估标准。类似于其他机器学习算法,目标检测需要正确预测出图像物体所属类别,但同时还需要对其进行位置预测,因此需要引入模型预测边框与真实标签边框的交并比指标IOU,如图5所示,IOU即为左图阴影面积和右图阴影面积比例大小[18]。

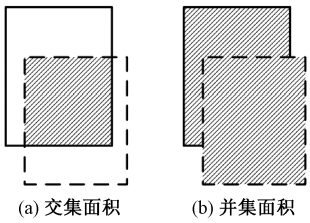


图5 锚框的交并集示意图

Figure 5 Intersection and union of the anchor box

在图5中,实线框为标注框,虚线框为模型预测框。当对预测情况进行判断时,常用IOU阈值为0.5,即当预测结果IOU大于阈值时,即认为该预测框瑕疵类别为标注框类别。本文数据集缺陷

预测的精确率和召回率计算公式如下：

$$P = \frac{TP}{TP + FP}; \tag{6}$$

$$R = \frac{TP}{TP + FN}。 \tag{7}$$

式中：TP 为模型预测为正样本而验证集中标签为正样本的数量；FN 为模型预测为负样本而验证集中标签为正样本的数量；FP 为模型预测为正样本而验证集中标签为负样本的数量；TN 为模型预测为负样本而验证集中标签为负样本的数量。

在模型对单个类别进行预测时，对 IOU 置信度取不同的阈值时，P 和 R 的值也会不断在 0～1 内变化，将改变置信度阈值得到的上述 2 个值在二维坐标系的变化曲线称为 PR 曲线，而 PR 曲线与坐标轴围成的面积即为单个类的 AP 值，对需要预测的 n 个类的 AP 值取平均值即为 mAP，公式如下：

$$mAP = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 P(r)\,\mathrm{d}r。 \tag{8}$$

5.2　稀疏训练以及裁剪训练 γ 分布结果

对原始 YOLOv5s 模型进行训练，得到 γ 值分布如图 6 所示，由图 6 可以看出，原始网络训练得到的结果接近正态分布。

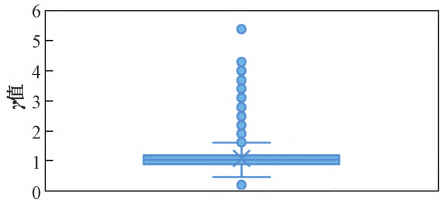


图 6　YOLOv5s 训练 γ 值分布图

Figure 6　γ value distribution of YOLOv5s training

对模型进行稀疏训练，得到 γ 值分布如图 7 所示，在稀疏训练后绝大多数 γ 值降低，有约 40%的 γ 值降低到 0 附近，可以对其进行裁剪。

将稀疏训练后得到的模型进行裁剪，在反复裁剪微调后，最终对 33%的接近 0 的 γ 值进行裁剪，并对裁剪后的模型进行训练，得到 γ 值分布如图 8 所示。在裁剪时为了保证 tensor 维度可加，对残差结构不进行裁剪，保留原始通道，否则会导致残差结构两端维度不同，前向传播无法进行[19]。同时在裁剪时为了保证网络结构的完整性，裁剪阈值不能超过各个网络层中 γ 参数最大值的最小值。

5.3　对比实验结果

对实验数据集使用几种网络进行训练，其中包括经典 two-stage 网络 Faster R-CNN、经典 one-stage 网络 SSD 和 YOLOv3-SPP，还有本文使用的 YOLOv5s 原始网络和裁剪后网络，对各个网络在搭建好的高配置实验台中进行训练，epoch 设置为 100，batch-size 设置为 16。网络训练过程中 mAP 稳定后的 15 个 epoch 如图 9 所示。由图 9 可以看出，本文模型 mAP 略高于原始的 YOLOv5s 模型，与经典 two-stage 网络 Faster-RCNN 几乎不相上下，本文网络相比其他网络的优势在于模型的轻量化，训练后权重文件小，可以在低算力硬件设备中展现出不错的推理速度。对每种模型采用验证集中的 10 张图片进行模型推理测试并计算出单张图片的平均推理时间，对图 9 中的 mAP 值取平均，结合模型训练最终的权重文件大小，得出实验结果如表 1 所示，模型识别效果如图 10 所示。

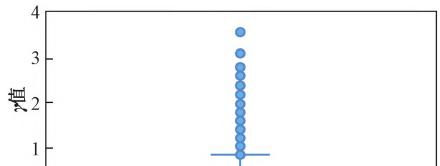


图 7　稀疏训练 γ 值分布图

Figure 7　γ value distribution of sparse training

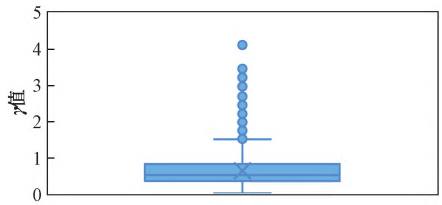


图 8　裁剪模型训练 γ 值分布图

Figure 8　γ value distribution of pruning model

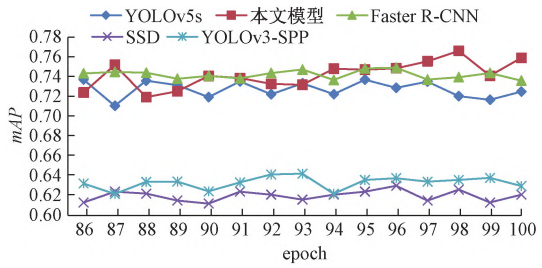


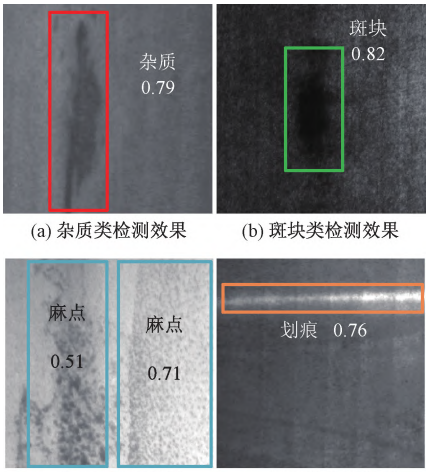
图 9　各网络训练情况对比

Figure 9　Comparison of each network training

表 1　各网络模型实验结果

Table 1　Experimental result each network model

网络模型	mAP	单图推理时间/ms	模型大小/MB
Faster R-CNN	0.74	45.69	315.3
SSD	0.62	27.74	95.7
YOLOv3-SPP	0.63	17.49	326.6
YOLOv5s	0.73	7.72	13.7
本文模型	0.74	6.85	4.4



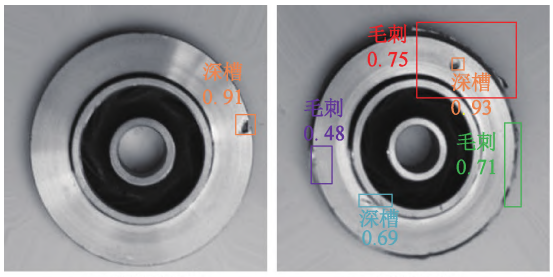
(a) 杂质类检测效果 (b) 斑块类检测效果

(c) 麻点类检测效果 (d) 划痕类检测效果

图 10 最终模型检测效果图

Figure 10 Final model detection effect

为了体现本文方法的普适性,选用 kaggle 数据集网站公开的潜水泵叶轮的俯视图数据集进行普适性实验,数据集共有 2 400 张图片。将数据集图像中缺陷种类分为深槽和毛刺类别并标注数据集,按 8∶2将数据集分为训练集和验证集。在高性能实验台上进行训练,最终采用 YOLOv5s 和本文模型分别在该数据集上得到了 72.7%和 73.4%的 *mAP*,证明了本文方法在其他数据集上也有优秀的表现。检测效果如图 11 所示。



(a) 单类效果图 (b) 多类效果图

图 11 普适性实验检测效果图

Figure 11 Universal experimental detection effect

由于 YOLOv5 网络在输入端采用 mosaic 数据增强,对输入数据集进行随机缩放、裁剪、排布的方式拼接,使其对小目标检测效果较 SSD 和 YOLOv3-SPP 网络有所提升,Faster R-CNN 模型采用双阶段的目标检测方式,在检测头部先进行候选框提取,再对候选框进行筛选得到预测框,虽然获得了一定的检测精度,但是速度较慢,难以满足流水线产品的快速检测需求,采用本文裁剪方法后的 YOLOv5s 模型减轻了原模型在本文小样本数据集上的过拟合现象,在保证精度的同时,减少了运算参数,提升了推理速度。由表 1 可以得出,本文裁剪后的模型在预测精度没有损失的情况下,模型文件大小相比原 YOLOv5s 模型减少了 67.9%,推理时间提升了 11.3%,检测效率得到了明显的提升。如表 2 所示,为展现本文裁剪模型在低算力硬件设备上的推理时间优势,在上述搭建的低性能实验台上对 NEU-DET 验证数据集中 10 张图片进行推理时间测试并取平均值,从结果可以看出本文模型在低算力硬件设备上推理时间提升更大。

表 2 单图推理时间对比实验结果

Table 2 Inference time of single figure comparison experiment result

网络模型	推理时间/ms	
	高性能实验台	低性能实验台
YOLOv5s	7.72	299.2
本文模型	6.85	112.7

最后对本文训练得到的模型文件转 onnx 格式并采用 TR 推理框架进行加速,在高性能实验台中测试表现可以达到单图 5.82 ms 的推理时间。本文方法与同样采用本文 NEU-DET 数据集的文献[9]、文献[20]相比,*mAP* 分别提升了 6.56%和 0.5%,且本文模型更轻量化,检测速率更快。

6 结论

本文从企业需要更低的硬件成本的角度出发,通过深度学习网络模型中结构裁剪的思想,提出一种基于 YOLOv5s 轻量化改进的网络并结合 TR 加速推理框架。该方法很好地实现了模型权重文件大小缩减,降低运算量,且保证精度没有损失,可以满足产品缺陷实时检测的需求。下一步将基于本文模型继续研究该模型在集成于低算力嵌入式设备和移动端设备的优势,并将本文模型与实际工业场景应用相结合。

参考文献:

[1] HE K M, ZHANG X Y, REN S Q, 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.

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

[3] REN S Q, HE K M, 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, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2021-09-01]. https://arxiv.org/abs/1804.02767.

[5] ZHOU P, NI B B, GENG C, et al. Scale-transferrable object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 528-537.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]// European Conference on Computer Vision. Cham: Springer, 2016:21-37.

[7] 程旭, 宋晨, 史金钢, 等. 基于深度学习的通用目标检测研究综述[J]. 电子学报, 2021, 49(7): 1428-1438.
CHENG X, SONG C, SHI J G, et al. A survey of generic object detection methods based on deep learning [J]. Acta electronica sinica,2021,49(7): 1428-1438.

[8] 郝用兴, 李泽坤, 张太萍, 等. 改进 Faster R-CNN 对铝型材表面瑕疵的检测[J]. 工具技术, 2021, 55(3): 76-80.
HAO Y X, LI Z K, ZHANG T P, et al. Detection of surface defect of aluminum profile by improved faster R-CNN [J]. Tool engineering, 2021, 55(3): 76-80.

[9] 程婧怡, 段先华, 朱伟. 改进 YOLOv3 的金属表面缺陷检测研究[J]. 计算机工程与应用, 2021, 57(19): 252-258.
CHENG J Y, DUAN X H, ZHU W. Research on metal surface defect detection by improved YOLOv3[J]. Computer engineering and applications, 2021, 57(19): 252-258.

[10] LI Y, XU J B. Electronic product surface defect detection based on a MSSD network[C]//2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference. Piscataway: IEEE, 2020: 773-777.

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

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

[13] 陈友升, 周介祺, 梁敏健, 等. 基于 YOLOv5 视觉感知的实时叉车驾驶操作行为识别方法[J]. 自动化与信息工程, 2021, 42(3): 21-26.
CHEN Y S, ZHOU J Q, LIANG M J, et al. Real time forklift driving operation behavior recognition method based on YOLOv5 visual perception[J]. Automation & information engineering, 2021, 42(3): 21-26.

[14] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[EB/OL]. (2015-02-11)[2021-09-16]. https://arxiv.org/abs/1502.03167.

[15] SHI X H, HU J, LEI X Y, et al. Detection of flying birds in airport monitoring based on improved YOLOv5 [C]//2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP). Piscataway: IEEE, 2021: 1446-1451.

[16] DETTMERS T, ZETTLEMOYER L. Sparse networks from scratch: faster training without losing performance [EB/OL]. (2019-06-13)[2021-09-14]. https://arxiv.org/abs/1907.04840.

[17] 周立君, 刘宇, 白璐, 等. 使用 TensorRT 进行深度学习推理[J]. 应用光学, 2020, 41(2): 337-341.
ZHOU L J, LIU Y, BAI L, et al. Using TensorRT for deep learning and inference applications[J]. Journal of applied optics, 2020, 41(2): 337-341.

[18] 张震, 李浩方, 李孟洲, 等. 改进 YOLOv3 算法与人体信息数据融合的视频监控检测方法[J]. 郑州大学学报(工学版), 2021, 42(1): 28-34.
ZHANG Z, LI H F, LI M Z, et al. Video surveillance detection method based on improved YOLOv3 algorithm and human body information data fusion[J]. Journal of Zhengzhou university (engineering science), 2021, 42(1): 28-34.

[19] LIU Z, LI J G, SHEN Z Q, et al. Learning efficient convolutional networks through network slimming[C]// 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2755-2763.

[20] 陈建强, 刘明宇, 符秦沈, 等. 基于深度学习的热轧钢带表面缺陷检测方法[J]. 自动化与信息工程, 2019, 40(4): 11-16, 19.
CHEN J Q, LIU M Y, FU Q S, et al. Hot rolled steel strip surface defect detection method based on deep learning[J]. Automation & information engineering, 2019, 40(4): 11-16, 19.

Lightweight Surface Defect Detection Method of Metal Products Based on YOLOv5s

JIA Yunfei, ZHENG Hongmu, LIU Shanliang

(School of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China)

Abstract: In order to reduce the intelligent cost in the enterprise, the hardware equipment with low cost and low computing power was used to detect the defects of products through the object detection algorithm model in deep learning. Based on the YOLOv5s network in target detection, this study adopts the idea of structure cutting, sparsely training the network based on the BN layer, and cuting the sparsely trained model corresponding to the layer with small weight value, so as to reduce the number of calculation parameters and the size of model file and to achieve the effect of lightweight. Finally, the trained pruning model was hierarchically fused using NVIDIA's accelerated framework TensorRT to realize the reasoning acceleration effect. The experimental results showed that the weight file size of this model was reduced by about 70% compared with the original YOLOv5s model, and the detection accuracy on the public dataset NEU-DET reached 74.2%. In the high-performance experimental platform built in this study, the single graph inference speed was improved by 11.3% compared with the original model, and the network had no accuracy loss. In the low-performance experimental platform compared with the original network model, the inference speed of this model increased by 165%, which was more significantly improved than the results in the high-performance experimental platform, indicating that this model perform well in low computing power hardware devices. Then the model was tested by using the open top view data set of submersible pump impeller. At last, the inference acceleration framework TensorRT is used to accelerate the model in this study, and the inference time of single figure 5.8 ms can be achieved on the high-performance experimental platform. The experimental results showed that the inference speed of this model could be greatly improved on low computing power hardware equipment, which could help enterprises reduce their budget.

Keywords: surface defect; object detection; light weight; YOLOv5s; tensorRT

(上接第7页)

Inverse Kinematic Parameters Calibration of 3-RPS Parallel Robot Based on Modified Differential Evolution

PENG Jinzhu, ZHANG Jianxin, ZENG Qingshan

(School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: Due to the complex structure of parallel robot, the joint torque solved by the design parameters cannot drive the parallel robot to achieve the ideal position and pose. However, the pose accuracy of parallel robot directly affects the work quality. To improve the model accuracy of the parallel robot, the kinematic error model was established for the designed 3-RPS parallel robot. In addition, on the basis of the traditional differential evolution (DE), a competitive multi-mutation differential evolution (CMDE) algorithm was proposed to calibrate the model parameters. In this algorithm, two populations were designed for the local exploitation and global exploration, where each population contained three mutation strategies. Moreover, a competitive system was developed in each population to select the better strategy in the calibration process, which could obtain the best optimal parameters. The kinematic parameters by calibration were used to modify the inverse kinematics model, and the accuracy of the modified model was verified by Adams software. The simulation results show that the proposed CMDE could achieve 50% faster convergence speed and smaller final convergence value in comparison to DE method. Also, compared with PSO and DE algorithms, the proposed CMDE had the strongest anti-interference ability in the evolution process. Moreover, compared with calibration before, the improvements of 3-RPS parallel robot with three degrees of freedom were 73.5%, 88.7% and 95.2%, respectively.

Keywords: parallel robot; CMDE; parameter calibration; inverse kinematics