DOI:10.19651/j.cnki.emt.2108726

基于 YOLOv5 的轻量化交通标志检测方法

张 上[1,2] 王恒涛[1,2] 冉秀康[3]
(1.三峡大学湖北省建筑质量检测装备工程技术研究中心 宜昌 443002;
2.三峡大学计算机与信息学院宜昌 443002; 3.三峡大学电气与新能源学院 宜昌 443002)

摘 要:针对目前交通标志检测算法存在网络复杂度高、计算量大、边缘端部署难度高。提出一种基于 YOLOv5 的轻量化交通标志目标检测算法。通过增加注意力机制,使用 CBAM 和 CA 融合的方式,强化检测模型抗干扰能力;通过 FPGM 剪枝,对模型进行了压缩,降低计算量、提高推理速度;通过软硬件融合设计,实现 YOLOv5s 模型与硬件融合,形成一整套完整的移动智能交通标志目标检测系统;结果表明,增加多种注意力机制后,模型精度提高了 2.8%。在极限剪枝的情况下,模型仅有 0.54 MB。在 Jetson Nano(20 W)的环境下,检测速度达 21 帧/s,满足实时的交通标志检测。

关键词:目标检测;注意力机制;模型剪枝;软硬件融合;YOLOv5;FPGM
中图分类号:TP391.4 **文献标识码**:A **国家标准学科分类代码**:510.8020

Lightweight traffic sign detection algorithm based on YOLOv5

Zhang Shang[1,2] Wang Hengtao[1,2] Ran Xiukang[3]
(1. Hubei Province Engineering Technology Research Center for Construction Quality Testing Equipment, China Three Gorges University, Yichang 443002, China; 2. College of Computer and Information, China Three Gorges University, Yichang 443002, China; 3. College of Electrical and New Energy, China Three Gorges University, Yichang 443002, China)

Abstract: Aiming at the shortcomings of traffic sign detection algorithm, such as high network complexity, large amount of calculation and difficult to be applied at the edge. A lightweight traffic sign target detection algorithm based on YOLOv5 is proposed. By increasing the attention mechanism and using the fusion of CBAM and CA, the anti-interference ability of the detection model is strengthened; Through FPGM pruning, the model is compressed to reduce the amount of calculation and improve the reasoning speed; Through the integration design of software and hardware, YOLOv5s model and hardware are integrated to form a complete set of mobile intelligent traffic sign target detection system; The results show that the accuracy of the model is improved by 2.8% after adding multiple attention mechanisms. In the case of extreme pruning, the model is only 0.54 MB. Under the environment of Jetson Nano (20 W), the detection speed is up to 21 frames/s, which meets the real-time traffic sign detection.

Keywords: target detection; attention mechanism; model pruning; integration of software and hardware; YOLOv5;FPGM

0 引 言

交通标志的检测和识别是自动驾驶和高级驾驶辅助系统(advanced driving assistance system,ADAS)的重点研究内容之一,其对驾驶安全具有至关重要的作用[1]。

目前,针对交通标志的检测方法主要分为两大类:基于传统特征提取的方法和基于深度学习的方法。基于传统特征提取的方法主要从颜色、轮廓、形状等方面考虑。Yang 等[2]提出了一种 Ohta 空间颜色概率模型,通过绘制颜色概率图来进行交通标志的检测。Wang 等[3]在德国交通标志大赛中,使用 HOG(方向梯度)和 SVM 分类器来检测交通标志,获得了不错的成绩。Xiao 等[4]将 HOG 融入到 BCNN 中,对 GTSDB[5]进行交通标志的识别。传统的算法泛化能力较弱,在复杂环境下不能满足检测需求。基于深度学习的目标检测算法有两类,一类是 two-stage,模型为两个阶段,先产生候选区域,后进行分类和定位,典型的算法有 R-CNN[6]、Fast RCNN[7]、Mask-RCNN[8]等;另一类是 one-stage,模型没有候选区域产生的阶段,直接进行类

收稿日期:2021-12-30

别的分类与定位，以 YOLO[9] 与 SSD[10] 算法为代表。单阶段目标检测算法的实时性高于基于区域候选框的算法，检测精确度也具有优势。同时，单阶段模型也更容易嵌入到嵌入式端。王文胜等[11]将 YOLOv5 嵌入到 Jetson Nano 开发板中，具有良好的表现。张明路等[12]在网络中增加了注意力机制，提高了网络模型的精度。

以上所总结的算法存在计算量较大、检测速度慢、检测小目标困难、边缘端移植难度大、泛化能力弱等缺陷。因此，为了减小计算量、提高检测效果，本文采用 YOLOv5s 模型作为基准模型，融入多种注意力机制，过滤掉冗余特征信息，保留重要特征，提高检测精度。对优化的模型进行剪枝，减小模型体积、提高检测速度。本文贡献如下：

1)将 CBAM 注意力机制与 CA 注意力机制进行融合，并将其融合结果部署到模型中。提高模型的检测精度。

2)对模型进行剪枝，使用 FPGM 剪枝算法对 YOLOv5 模型进行剪枝，降低模型的体积、计算量和检测速度。

3)将模型部署到嵌入式端，实现边缘处理。

4)在 CCTSDB 数据集上进行本文提出的改进算法验证，并与多个模型进行对比。

通过实验，本文提出的算法相较于原始 YOLOv5s 检测算法，能够在检测精度和推理时间上做到明显的提升。并且，改进后的模型嵌入到嵌入式端能够比原模型有较大的性能提升。在极限剪枝的情况下，模型可压缩至 0.54 MB，模型体积降低 96.2%，检测速度提升 41.7%。

1　YOLOv5 网络结构

YOLOv5 算法由四部分组成，第一部分为输入端，输入端需要完成 Mosaic 数据增强、自适应锚框计算和自适应图片缩放，对数据进行处理，增加图像的处理能力。第二部分为主干网络(Backbone)，由 Focus 结构和 CSP 结构组成，主要用于提取目标特征。第三部分为颈部网络(Neck)，由 FPN＋PAN 结构组成，用于收集组合目标特征。第四部分为检测层(Prediction)，由损失函数和预测框筛选函数组成。用于预测信息损失部分。

YOLOv5 总共有四个版本，分别为 YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x。其中 YOLOv5s 为深度和宽度最小的网络，其仅有 14.4 MB，基于体积小、计算速度快的优势，YOLOv5s 非常容易部署到嵌入式设备中。

2　基于 YOLOv5 的轻量化快速检测算法

探究 YOLOv5 加入注意力机制、模型剪枝、软硬件融合的方法。

2.1　注意力机制

注意力机制来源于人类大脑处理图像信息。通过观察图像的全局信息，人类可以凭借注意力来锁定重点关注的候选区域，自动屏蔽部分背景和冗余信息，能够快速的锁定焦点。

CA[13]注意力机制(coordinate attention)的主要思想是将位置信息嵌入到通道注意力。CA 不仅能捕获通道的信息，而且可以捕获方向和位置的信息，可以使模型更加精准的定位和识别重要信息。如图 1 所示，CA 注意力机制主要分为两个步骤，分别为坐标信息嵌入和坐标信息特征图生成。图像信息输入后，池化层分别沿着水平方向和垂直方向对通道进行编码，得到两个方向的特征图。将特征图拼接后，利用 1×1 卷积函数得到中间特征图，然后沿着空间维度分解两个张量，分别沿水平和垂直方向进行卷积和激活函数的处理，最后将两个方向的输出结果进行融合。

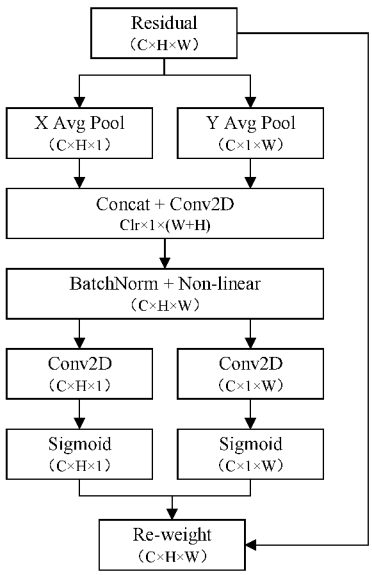


图 1　CA 注意力机制

通过初步实验，使用 CA 注意力机制替换 YOLOv5 中的 C3 层不仅有利于网络特征信息的提取，而且可以降低网络的体积。

卷积运行通过跨通道和空间信息混合在一起来进行特征信息的提取，通过通道注意力机制可以给定图像信息类别，空间注意力机制决定聚焦的位置。CBAM[14](convolutional block attention module)最大优点是结合了空间和通道两个方面的注意力机制模块。如图 2 所示，CBAM 第一步通道注意力机制进行图像类别的分析，着重于全局的信息。分别使用最大池化和平均池化提取通道信息，然后经过过滤、激活和归一化，提高通道信息的提取能力。第二步空间注意力机制进行信息的聚焦，着重于局部的信息。通过池化进行信息过滤，然后通过卷积提取重要信息。

为了精确地识别和定位交通标志，提高准确率。本文在 YOLOv5 网络中主要做了两大改进。改进后 YOLOv5 网络结构如图 3 所示，使用 CA 注意力机制代替部分 C3 结构，不仅降低了网络的体积和计算量，而且有利于提取检测网络的特征信息。在 Neck 部分增加 CBAM 注意力机制，可以在预测前再进行通道和空间的特征提取。

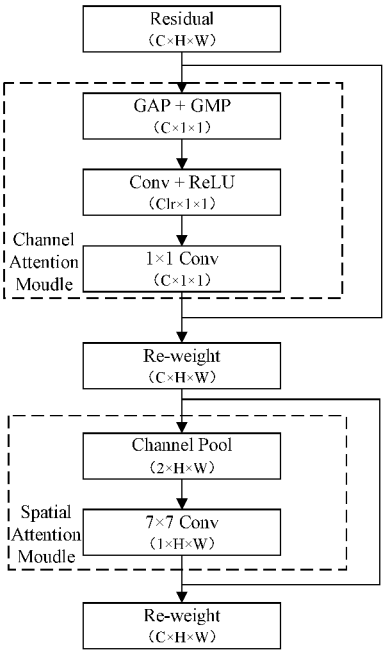


图 2　CBAM 注意力机制

2.2　FPGM 剪枝

FPGM[15]（基于几何中值的卷积神经网络滤波器剪枝）通过修建掉冗余的滤波器来进行模型的压缩。FPGM 实现了从“相对不重要”到“可替代性”的转变。FPGM 不再仅考虑低范数滤波器的裁剪，而是根据过滤器的可替代性来确定裁剪规则。如图 4 所示，相较于基于范式的剪枝方式，FPGM 保留了更多的特征信息。

FPGM 算法描述如表 1 所示，结合 YOLOv5 训练过程，YOLOv5 实现 FPGM 具体流程如下：

1）加载模型参数到 YOLOv5 中。

2）对于 YOLOv5 的每个卷积层，

（1）对该层的每个卷积核，计算该卷积核和其它所有卷积核的欧式距离（L2）之和（共得到 N 个欧式距离之和）。

（2）对得到的 N 个欧式距离之和，从小到大排序，剪裁掉前 N * R 个最小值对应的卷积核。

3）对剪枝过的网络进行训练。在训练过程中，被剪掉的卷积核的梯度必须强制为零。

4）不断迭代 2、3 两步，等到模型收敛，就可以得到待去零的剪枝模型。

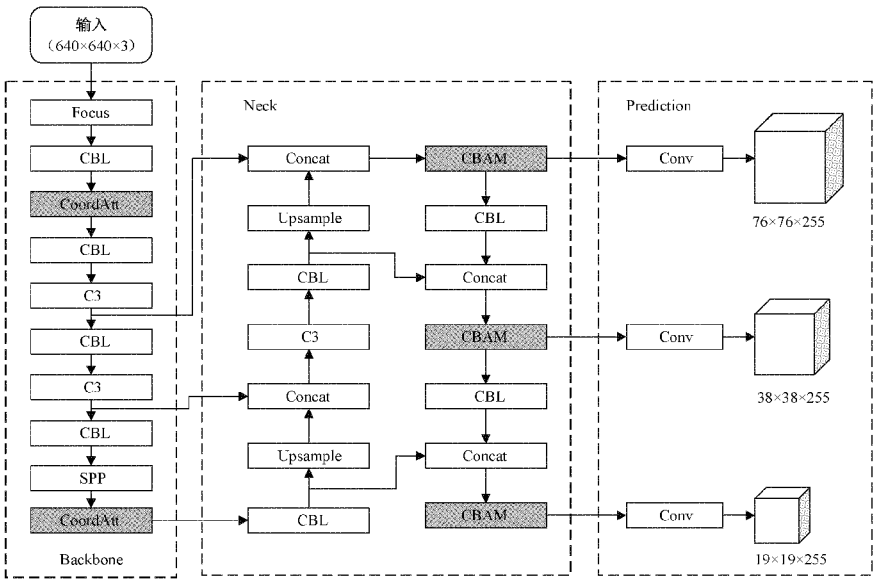


图 3　改进后 YOLOv5 网络结构

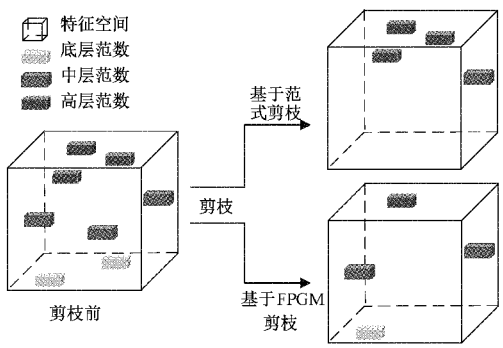


图 4　FPGM 剪枝原型图

5）待模型收敛后进行去零操作（该步骤在推理过程中单独完成）。

（1）去掉全零的卷积核。

（2）去掉卷积核中的冗余通道。

（3）去掉 BN 层参数冗余数值。

6）得到剪枝且去零的压缩模型。

2.3　软硬件融合及加速

为了能够将本文算法落实到具体的环境中，本文在嵌入式设备中进行了嵌入。实验环境为 Jetson Nano 开发板，如表 2 所示，为 Jetson Nano 具体的配置参数。

表 1　FPGM 算法

算法 1 FPGM 算法描述
准备工作：完成数据集和训练参数的设定：X
设定剪枝率：R.
1：初始化：模型参数 W＝{W(i)，0 ≤ i ≤ L}
2：for epoch＝1；epoch≤epoch_max；epoch＋＋do
3：　更新模型参数：W
4：　　for i＝1；i ≤ L；i＋＋do
5：　　　计算 N 个欧氏距离之和
6：　　end for
7：　　得到 N ＊ R 个相关度最低的滤波器
8：　　将相关度低的滤波器梯度置零
9：　end for
10：根据 W 模型获得待去零模型 W_1
11：去除 W_1 模型中的零参数，获得去零模型 W_2
输出：剪枝并去零后的模型与参数

表 2　Jetson Nano 配置参数

参数	配置
GPU	NVIDIA 128-core Maxwell
CPU	Quad-core ARM Cortex-A57
Memory	4GB 64 bit LPDDR4 25.6 GB/s
Storage	16 GB eMMC 5.1
Display	HDMI2.0 and eDP1.4
USB	4×USB3.0，USB2.0 Micro-B
加速环境	CUDA10.2
camera	Raspberry Pi v2

Jetson Nano 部署 YOLOv5s 过程如图 5 所示，将 YOLOv5 模型嵌入到嵌入式端需要在 PC 或服务器端进行模型的训练，对模型进行初步的整理后生成权重文件，使用 gen_wts.py 转化成 wts 文件格式的权重文件。然后在 Jetson Nano 环境下，将权重文件转化为 engine 文件，使用 DeepStream 进行模型的部署，实现 TensorRT 的硬件加速。最后调用 cmi 摄像头（或 USB 摄像头）实现最终模型的实时化处理。

3　实验及结果分析

本实验训练过程在 Windows10、CUDA10.2、环境下进行，GPU 配置：NVIDIA RTX1660ti，6 GB 显存，调用 GPU 进行训练。

3.1　交通标志数据集

为验证本文模型压缩和改进算法的效果，本文采用的数据集为 CCTSDB[16]（CSUST chinese traffic sign detection benchmark）。CCTSDB 为长沙理工大学中国交通标志检测数据集。是当前中国交通标志公认的数据集之一。其数据集图片的拍摄角度和分类都十分的规范，故作为本文算法实现的数据集。

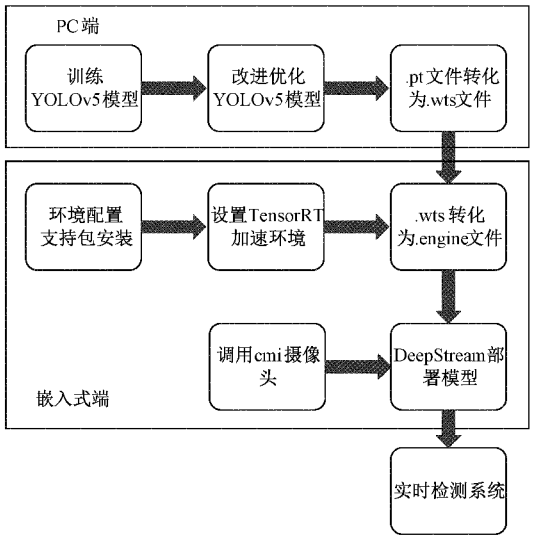


图 5　Jetson Nano 部署 YOLOv5 过程

CCTSDB 数据集共包含 3 大类交通标志，选取数据集中的 5 880 张图片，分别为“指示标志”、“禁止标志”、“警告标志”。将标注信息转化为对应的 XML 格式的目标区域信息，再经过转换程序统一转换为 TXT 格式，使其能够在 YOLOv5 训练中读取图像标注信息。按照 VOC2007 的数据集格式进行调整，按照 3∶1 的比例将数据集分为训练集和验证集，并且使用原数据集提供的测试集来进行结果的验证，详细数据集划分说明如表 3 所示。

表 3　数据集划分

	训练集	验证集	测试集
图片数量	4 785	1 596	401
标注框数量	6 459	2 154	—

3.2　评测指标

图像目标检测的重要评价指标主要有平均准确率 mAP(mean average precision)、准确率 P(precision)、召回率 R(recall)、F1 值和每秒处理帧率(fps)。P 为预测为正的样本中实际为正的比例；R 为正样本中被正确预测的比例。其中，准确率、召回率、平均准确率的计算公式如下：

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{AP} = \int_0^1 P(R) \tag{3}$$

$$\text{mAP} = \frac{\sum_{j=1}^{c} AP_j}{c} \tag{4}$$

TP(true positive)为模型检测为真的正样本个数；FN(false negative)为模型预测非真的正样本个数；FP(false

positive)为模型预测为真的负样本个数;mAP 为所有类别的平均准确率均值;为类别数。

3.3　模型剪枝

模型剪枝数据统计表如表 4 所示,从数据上看,模型的体积、计算量和参数量随着剪枝率的增加均沿线性下降,但模型的检测速度并不呈现为线性下降,当剪枝到 60%后,模型的检测速度降低十分有限。为此,在精度、检测速度和模型体积的均衡考虑下,本文后续实验使用 60%的剪枝率。

表 4　模型剪枝计算量统计表

Model	计算量(GFLOs)	参数量	体积/MB	推理时间/ms	计算量变化/%	参数量变化/%	模型体积变化/%	推理时间变化/%
YOLOv5s	16.4	7 068 936	14.4	8.803	—	—	—	—
0.1prunc	13.9	5 866 417	11.9	7.798	↓(15.24)	↓(17.01)	↓(17.36%)	↓(11.41%)
0.2prunc	11.6	4 744 456	10.0	7.096	↓(29.26)	↓(32.88)	↓(30.05%)	↓(19.39%)
0.3prune	9.4	3 756 400	8.0	6.459	↓(42.68)	↓(46.86)	↓(44.44%)	↓(26.62%)
0.4prune	7.5	2 867 215	6.2	5.863	↓(54.26)	↓(40.56)	↓(56.94%)	↓(33.39%)
0.5prune	5.6	2 088 216	4.6	5.305	↓(65.85)	↓(70.46)	↓(68.05%)	↓(39.73%)
0.6prune	4.2	1 457 571	3.3	5.170	↓(74.39)	↓(79.38)	↓(77.08%)	↓(41.27%)
0.7prune	2.8	918 040	2.1	5.149	↓(82.93)	↓(87.01)	↓(85.41%)	↓(41.50%)
0.8prune	1.7	500 432	1.2	5.132	↓(89.63)	↓(92.92)	↓(91.66%)	↓(41.70%)
0.9prune	0.8	193 679	0.54	5.129	↓(95.12)	↓(97.26)	↓(96.25%)	↓(41.73%)

3.4　实验结果与分析

为了进一步分析每个改进点对 YOLOv5s 算法的影响,本文进行了消融实验,如表 5 所示,在模型中加入 CA 与 CBAM 注意力机制均可提高网络的精度,两者的融合可进一步提高模型的检测精度。使用 FPGM 进行剪枝,模型的精度有所降低,但在剪枝率为 50%时,模型的精度仍高于 YOLOv5s。

表 5　消融实验

算法	CA	CBAM	剪枝	P/%	R/%	mAP/%
YOLOv5s				88.4	93.8	92.9
A	√			90.3	95.2	94.6
B		√		89.3	94.9	94.4
C	√	√		91.0	96.6	95.7
D	√	√	50	89.2	94.1	93.2
E	√	√	90	83.8	87.4	85.3

改进后模型剪枝率结果对比如表 6 所示,模型随着剪枝率的提高,平均精度有所下降,并且平均精度下降速度也在提高。在 90%剪枝率的情况下,模型平均精度下降最大。

表 6　各剪枝率结果对比

剪枝率/%	平均精度/(mAP/%)
无	92.91
20%	94.89
40%	93.82
60%	92.64
70%	91.32
80%	89.63
90%	85.05

在 Jetson Nano 嵌入式环境下,进行模型的移植。Jetson Nano 环境运行结果如表 7 所示,相比于服务器端,Jetson 计算能力极低,为此使用 TensorRT 进行推理加速。实验结果显示,模型随着剪枝率的提高,检测速度在不断加快,但在剪枝率超过 60%后,检测速度的提升放缓。

表 7　Jetson Nano 环境运行结果表

	推理时间/ms	检测速度/fps
PC(服务器)	8.8	69.4
Jetson Nano	43.0	14.2
Nano+0.3 剪枝	32.6	18.7
Nano+0.6 剪枝	28.6	21.3
Nano+0.9 剪枝	28.2	21.6

算法对比如表 8 所示,在公开数据集 CCTSDB 上,本文改进算法在平均精度与模型体积上均有提高,在仅考虑精度的情况下(无剪枝),本文平均精度可达 95.73%,相比 SSD、YOLOv4、YOLOv5s 算法,平均准确度分别提高了 7.2%、1.6%、2.8%。在仅考虑模型体积的情况下(模型

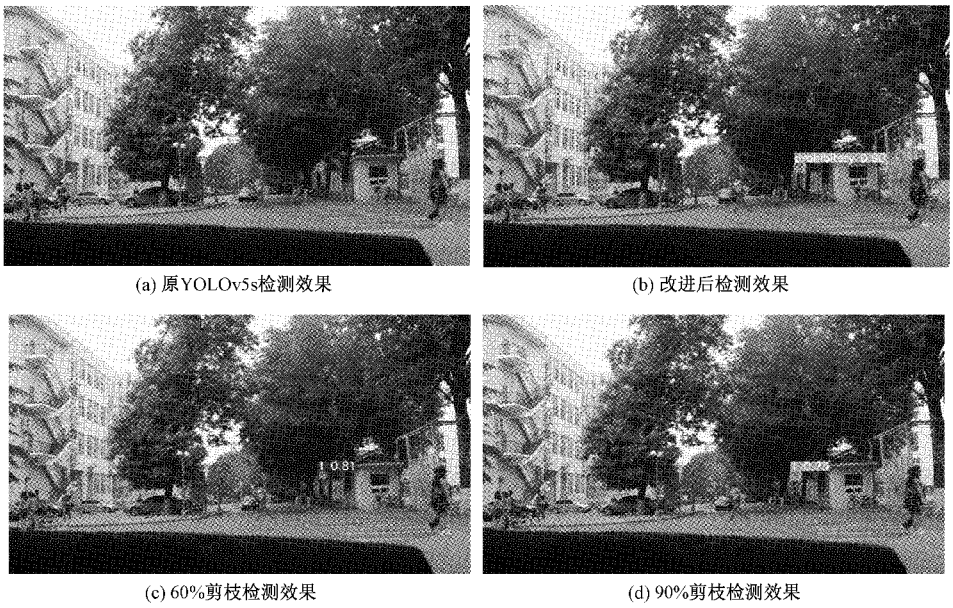
表 8　算法对比

算法	体积/MB	平均精度/(mAP/%)	检测速度/fps
YOLOv5s	14.4	92.91	69.4
YOLOv4	245.9	94.12	24.6
SSD	98.3	88.53	54.5
本文(精度)	12.2	95.73	73.2
本文(体积)	0.54	85.3	108.9

剪枝率为 90%)，本文模型可压缩至 0.54 MB，为 SSD 模型的 0.54%、YOLOv4 的 0.22%、YOLOv5s 的 3.75%，模型体积大幅降低，为嵌入式移植奠定基础。同时，从检测速率上来看，本文算法的检测速率优于 YOLOv5s、YOLOv4 和 SSD 算法，分别提高了 36.3%、77.4%、50.0%。

3.5　实验效果与分析

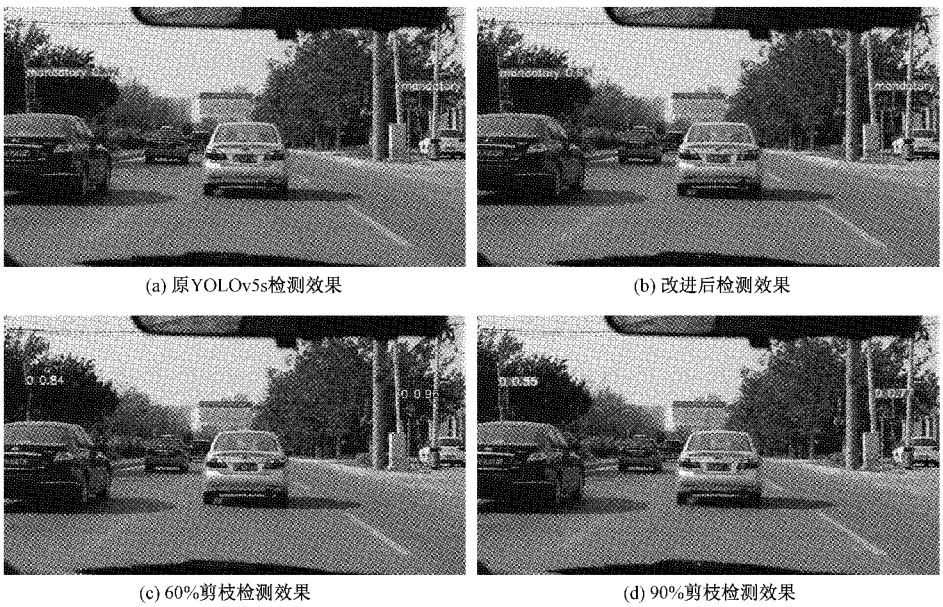
小目标检测结果如图 6 所示，由图 6(a)与图 6(b)对比可得使用 CA 与 CBAM 注意力机制能有效提高模型的检测效果，可以检测出原 YOLOv5s 不能检测出来的小目标，且置信度高。模型经过 60%剪枝后，模型检测结果置信度有所降低，但置信度仍高于 80%。经过 90%的剪枝后，模型仍能够保持较好的检测效果。异形目标检测如图 7 所示，图片左侧的交通标志为侧面拍摄。由图 7 中的图片对比可得，本文改进算法能够较好对其进行检测，即使在剪枝率为 90% 的情况下，模型检测的置信度仍高于原 YOLOv5s 模型。



(a) 原YOLOv5s检测效果　(b) 改进后检测效果

(c) 60%剪枝检测效果　(d) 90%剪枝检测效果

图 6　小目标检测结果



(a) 原YOLOv5s检测效果　(b) 改进后检测效果

(c) 60%剪枝检测效果　(d) 90%剪枝检测效果

图 7　异形检测结果对比

4　结　　论

交通标志的检测和嵌入式移植是自动驾驶和汽车辅助驾驶等方向的基础，本文对此进行了研究，综述了当前交通标志常用的检测方法，在 YOLOv5s 目标检测网络的基础上提出了一种轻量化交通标志检测算法。

首先,在 YOLOv5s 模型上增加注意力机制,通过对比多种注意力机制的融合实验,得到相对合适的注意力机制融合方案。然后,使用 FPGM 对模型进行剪枝,在剪枝率为 90%时,模型的大小仅为 0.54 MB,相较于原模型降低了 96.25%,这将更加有利于嵌入式部署。最后,将模型嵌入到 Jetson Nano 开发板中,实现软硬件融合。

通过实验,在 CCTSDB 数据集上验证了本文的轻量化交通标志检测算法具有实时性高、性能稳定、高效、鲁棒性高等特点。相较于大部分交通标志目标检测检测算法,本文的算法能够保证在精度、模型体积和检测速度等方面具有较高的成绩。

参考文献

[1] 陈红,王相超,陈志琳. 自然场景下的交通标志检测与识别[J]. 电子测量技术, 2021,44(12):102-109.

[2] YANG Y, WU F. Real-time traffic sign detection via color probability model and integral channel features [C]. Chinese Conference on Pattern Recognition. Springer, Berlin, Heidelberg, 2014: 545-554.

[3] WANG G, REN G, WU Z, et al. A robust, coarse-to-fine traffic sign detection method [C]. The 2013 Unternational Joint Conference on Neural Networks (IJCNN), IEEE, 2013: 1-5.

[4] XIAO Z, YANG Z, LEI G, et al. Traffic sign detection based on histograms of oriented gradients and boolean convolutional neural networks [C]. 2017 International Conference on Machine Vision and Information Technology (CMVIT), IEEE, 2017: 111-115.

[5] HOUBEN S, STALLKAMP J, SALMEN J, et al. Detection of traffic signs in real-world images: The German traffic sign detection benchmark [C]. International Joint Conference on Neural Networks, IEEE, 2013:1-8.

[6] 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.

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

[8] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

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

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

[11] 王文胜,李继旺,吴波,等. 基于 YOLOv5 交通标志识别的智能车设计[J]. 国外电子测量技术,2021,40(10):158-164,DOI:10.19652/j.cnki.femt.2102913.

[12] 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.

[13] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713-13722.

[14] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module [C]. Proceedings of the European Conference on Computer Vision(ECCV), 2018: 3-19.

[15] HE Y, LIU P, WANG Z, et al. Filter pruning via geometric median for deep convolutional neural networks acceleration[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4340-4349.

[16] 李旭东,张建明,谢志鹏,等. 基于三尺度嵌套残差结构的交通标志快速检测算法[J]. 计算机研究与发展,2020,57(5):1022-1036.

作者简介

张上,工学博士,副教授,研究方向为物联网技术、计算机应用技术。

E-mail:3011408157@qq.com

王恒涛(通信作者),硕士研究生,研究方向为目标检测。

E-mail:1248558938@qq.com