Computer Science and Application 计算机科学与应用, 2021, 11(9), 2333-2341
Published Online September 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.119238

Hans汉斯

基于轻量化网络的YOLOv4检测算法研究

张小燕，汪华章

西南民族大学电气工程学院，四川 成都
Email: cxd3405@163.com

收稿日期：2021年8月14日；录用日期：2021年9月10日；发布日期：2021年9月17日

摘　要

针对YOLOv4算法网络模型大、运算量大、运行效率低的问题，本文提出一种基于GhostNet的YOLOv4轻量化模型，使用GhostNet代替原网络中的CSPDarknet-53，通过一些简单的线性运算代替部分卷积，减少卷积运算，从而减少参数量以及浮点运算量。然后，在GhostNet中引入CBAM模块，结合通道注意力机制和通道注意力机制，增强模型对有效特征关注。另外使用HDC代替原始网路中的SPPNet，减少浅层网络特征的丢失。最后，将改进的算法与基于其他轻量化网络的YOLOv4算法对比。实验结果证明，与原始YOLOv4相比，在精度损失较小的情况下，基于GhostNet以及注意力机制的YOLOv4轻量化模型的大小得到了压缩，检测速度有了明显的提升。

关键词

目标检测，YOLOv4，轻量化模型，GhostNet，CBAM模块，HDC

Research on YOLOv4 Detection Algorithm Based on Lightweight Network

Xiaoyan Zhang, Huazhang Wang

School of Electrical Engineering, Southwest Minzu University, Chengdu Sichuan
Email: cxd3405@163.com

Received: Aug. 14th, 2021; accepted: Sep. 10th, 2021; published: Sep. 17th, 2021

Abstract

In view of the problems of large network model, large amount of computation and low operation efficiency of YOLOv4 algorithm, this thesis proposes a lightweight model of YOLOv4 based on GhostNet. The CSPDarknet-53 in the original network is replaced by GhostNet. And partial convo-

文章引用: 张小燕, 汪华章. 基于轻量化网络的 YOLOv4 检测算法研究[J]. 计算机科学与应用, 2021, 11(9): 2333-2341.
DOI: 10.12677/csa.2021.119238

lution is replaced by some simple linear operations to reduce convolution operation, thus reducing the number of arguments and floating point operations. Then, CBAM module is introduced into GhostNet, combining channel attention mechanism and channel attention mechanism, to enhance the model's attention to effective features. In addition, HDC is used instead of SPPNet in the original network to reduce the loss of shallow network features. Finally, the improved algorithm is compared with YOLOv4 algorithm based on other lightweight networks. Experimental results show that, compared with the original YOLOv4, the size of the lightweight model of YOLOv4 based on GhostNet and attention mechanism is compressed and the detection speed is significantly improved under the condition of less accuracy loss.

Keywords

Target Detection, YOLOv4, Lightweight Model, GhostNet, CBAM Module, HDC

1. 引言

随着深度学习的技术不断攀升，在目标检测方面应用基于深度卷积神经网络的一系列算法之后，在速度和精度上都取得了极大的进步。卷积神经网络实现了在特征提取、边界框的以及目标分类的判别。

目前基于卷积神经网络的目标检测算法大体分为两类：两阶段检测算法，例如，R-CNN [1]、Fast R-CNN [2]和 Faster R-CNN 等算法；单阶段检测算法，例如 YOLO 算法和 SSD 等算法[3]。两阶段目标检测算法有两个步骤，先是生成候选区域，然后对候选区域进行分类、回归，产生最后结果。由此可见，这类目标检测算法步骤多，训练困难，虽然准确度高，但是检测速度慢[4]。于是学者们将两个阶段融合为一个阶段，即单阶段目标检测算法。该类算法通过将目标定位转换为回归，可以一步获取目标的类别概率以及坐标位置，不再使用候选框[5]。这类算法虽然提高了检测速率，但是检测的准确率也降低了。但无论是两阶段目标检测算法还是单阶目标检测算法，为了更好地进行目标特征提取，网络层数不断深化，但是由此也导致了参数量剧增，极大地增加了浮点运算的运算量，并且对内存资源的占用也是极大的负担[6]。

于是学者们开始对网络进行轻量化设计，2016 年，Iandola 等在文献[7]中提出 SqueezeNet 轻量化网络，该网络是基于 Fire module 模块构建的，与 Inception 的网络构建一致，参数量有所减少。2017 年 Howard 等在文献[8]中提出 Mobilenet_v1 轻量化网络，基本组件就是引入的深度可分离卷积，同时引入两个超参数，经过改进模型参数量得到进一步降低。同年，Zhang 等人提出了 ShuffleNet v1 轻量化网络，该算法使用分组卷积代替 1 × 1 卷积，并进行 Channel Shuffle，达成通道间信息交互，保证网络性能。2018 年，Mark [9]等在文献[9]中提出了 Mobilenet_v2 轻量化网络，该网络是在 Howard 的基础上引入了倒残差网络以及线性瓶颈结构，用 Relu6 激活函数代替 Narrow layer 后的 ReLU，因为 Relu6 在增加高维空间非线性方面更加有效，这样能够保持特征多样性，增强网络表达能力。同年 Ma [10]等人在文献[10]中提出 ShuffleNet v2 轻量化网络，该网络先划分通道，一个分支不操作，一个分支卷积，然后先拼接再进行 Channel Shuffle 操作。2019 年 Grace [11]等在文献[11]中提出 Mobilenet_v3 轻量化网路，该网络是对 Mobilenet_v2 的改进，在 MobileNet_v3 中先进行平均池化[12]，再使用 1 × 1 卷积，保留了高维特性，还减少了延迟，与 Mobilenet_v2 的瞬息相反，此外，Mobilenet_v3 主要使用了 NAS (Neural Network Search)，

并引入轻量级注意力模型(SE)、h-swish 激活函数，使用 5 × 5 的卷积[13]。但是这些轻量化网络都存在同一个问题，网络为全卷积网络，浮点运算量非常大。

本文以 YOLOv4 算法为基础，对 YOLOv4 进行轻量化设计，减少浮点运算，从而提高检测速度。将 GhostNet 引入 YOLOv4 算法，代替原网络中 CSPDarknet-53，达到使用更少的参数数据生成更多有效特征的目的。并与其他轻量化网络进行对比，验证 GhostNet 的有效性。为保证目标检测的精度，在 GhostNet 网络中引入混合注意力机制(CBAM) [14]，使得网络模型更加关注有效的信息，提取出有效的特征信息，使用 HDC (混合空洞卷积)代替原网络中的 SPPNet，保留浅层网络特征，从而保证检测精度。

2. 算法概述

2.1. YOLOv4 算法

YOLOv4 算法作为最近两年目标检测的热门算法，是在原有的 YOLO 算法的检测架构上从数据处理、主干网络、网络训练、激活函数以及损失函数等各个方面做出了改进[15]，其网络结构如图 1 所示。

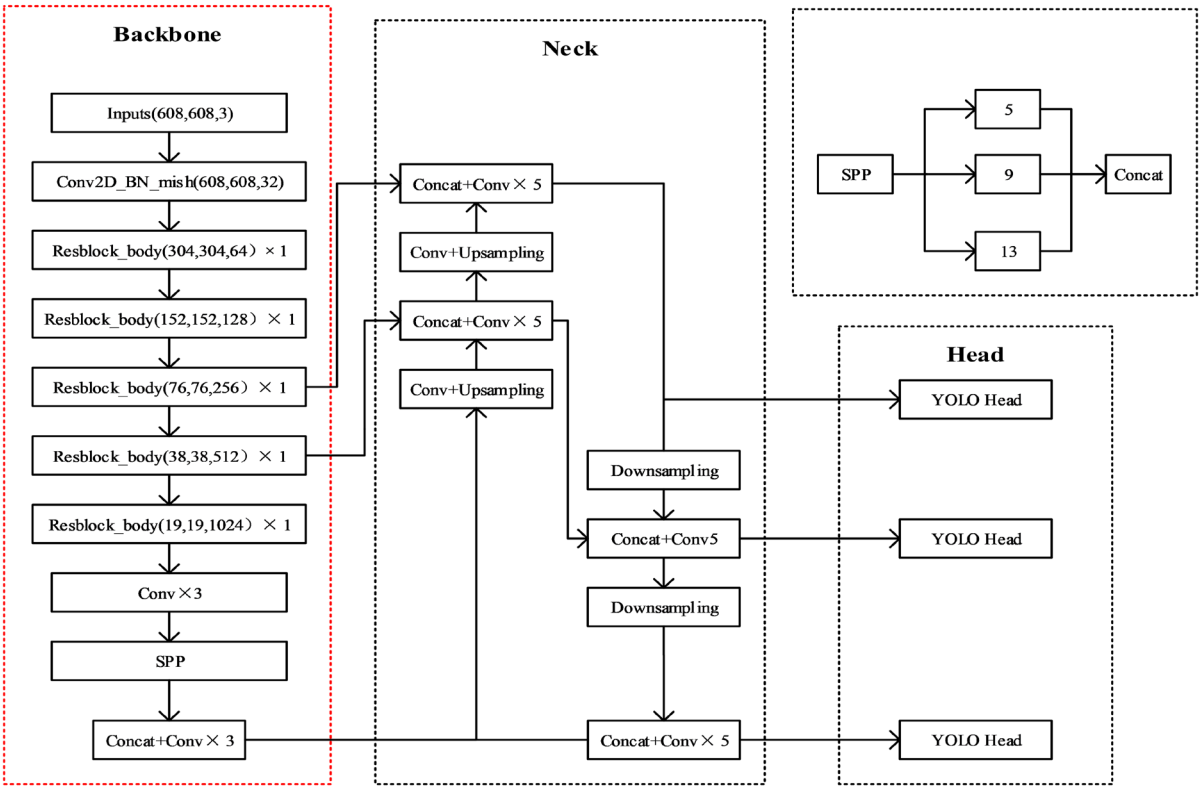

Figure 1. The Structure diagram of YOLOv4
图 1. YOLOv4 结构图

从上图可以看出，其主干网络采用了 CSPDarknet53，该部分主要是一系列的残差网络构成，都是全卷积，所以浮点运算量很大。相较于 Darknet53，它增加 SPPnet (Spatial Pyramid Pooling network)，就是对特征图进行不同程度的池化，目的是增加感受野，但是这样会降低所提取的特征的分辨率，丢失掉一些细节信息。其次在 neck 部分采用 PANet (Path Aggregation Network for Instance Segmentation)，充分融合特征[16]。原始的 YOLOv4 算法在梯度消失方面做了弥补，模型的学习能力有所提高，但是模型的参数量和计算量非常大。

2.2. 算法改进

本文针对上节所述不足对 YOLOv4 算法进行改进。

1) 针对原始 YOLOv4 算法浮点运算量大的问题，本文使用 GhostNet [17]代替 CSPDarknet53，使用简单的线性运算代替部分卷积神经网络，以此减少参数量以及浮点运算量。

通过标准卷积得到的特征图如图 2(a)所示，特征图均由卷积运算生成，有很多冗余信息，其浮点运算量为：

$$Q_1 = n \times h' \times w' \times c \times k \times k \tag{1}$$

本文在 YOLOv4 的特征提取网络引入 GhostNet 网络，对输入的图片包括两次卷积操作，第一次通过卷积生成一部分特征图，假设输出通道数为 $m \times s$，则生成 m 个特征图；第二次是用第一次的特征图映射生成 $s-1$ 个新特征图，从而生成 $m \times (s-1)$ 个特征图，最后把两次卷积的到的特征图拼接起来作为输出，最后的通道数是 $m \times s$。

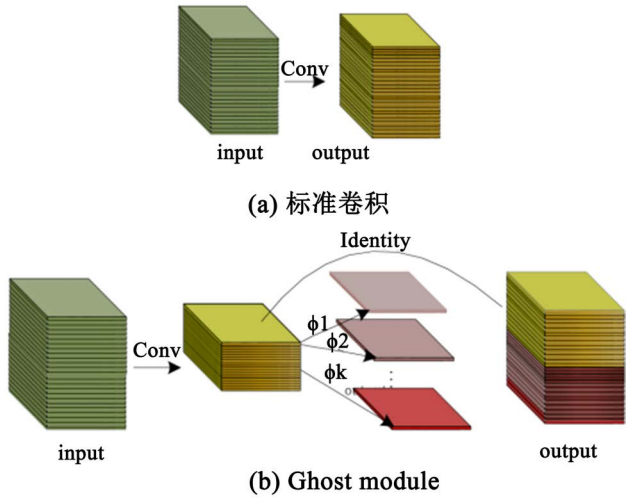

Figure 2. Ordinary convolution and Ghost module
图 2. 普通卷积与 Ghost module

如图 2(b)所示，Ghost module 在进行运算时，Identity 部分作为标准卷积部分，输入数据 $X \in Rc \times h \times w$，其中 c 为输入通道数，h 是输入数据的长，w 是输入数据的宽，$f \in Rc \times k \times k \times m$，$k$ 是卷积核尺寸，m 是标准卷积部分的卷积核数量，$m \le n$ 于是得到获取特征图的公式：

$$Y = X * f + b \tag{2}$$

其中，$Y \in Rm \times h' \times w'$ 是输出特征图，h' 和 w' 分别表示输出特征图的长和宽，b 表示偏置。那么利用线性操作得到的特征图为：

$$y_{ij} = \phi_{i,j}(y_i') \tag{3}$$

其中，y_i' 是 Y 中的第 i 个特征图，$\phi_{i,j}$ 是对第 y_i' 个特征图的第 j 次线性运算，得到第 y_{ij} 个特征图。当 $j = s$ 时，得到 $m \times s$ 个特征图，通卷积的通道数保持一致。

通过这种方法的浮点运算量

$$Q_2 = \frac{1}{s} \times Q_1 + (s-1) \times \frac{1}{s} \times h' \times w' \times d \times d \tag{4}$$

于是能够得到理论加速比：

$$R=\frac{Q_1}{Q_2}=\frac{s\times c}{c+s-1} \tag{5}$$

其中 $d\times d$ 是线性运算的内核大小，其幅度与 $k\times k$ 的幅度相似，当 $s\ll c$ 时：

$$R\approx\frac{s\times c}{c+s-1}=\frac{s}{1+\frac{s-1}{c}}\approx s \tag{6}$$

当 $d\times d=k\times k$ 时，使用 Ghost Module 的速度理论上能比原本提升 s 倍。

用两个 Ghost Module 进行堆叠，构成 Ghost Bottleneck，和残差结构有些类似，第一个 Ghost Module 是为了增加特征维度，第二 Ghost Module 是为了减少特征维度，使其通道数与短接过来的通道数一致。

为保证目标检测的精度，相对于原始的 GhostNet 网络，本文综合空间注意力和通道注意力的优势，在 GhostNet 特征提取网络中加入了 CBAM 模块，如图 3 所示。在通道和空间上都使用 attention。第一步，对输入数据进行全局最大池化和全局平均池化，得到两个特征映射，这里基于通道注意力机制进行；第二步，将两个特征图拼接后降维为 1 个通道；第三步，通过 sigmoid 生成空间注意力特征图；第四步，利用第三步的特征图乘以该模块的输入特征图生成最终特征图。

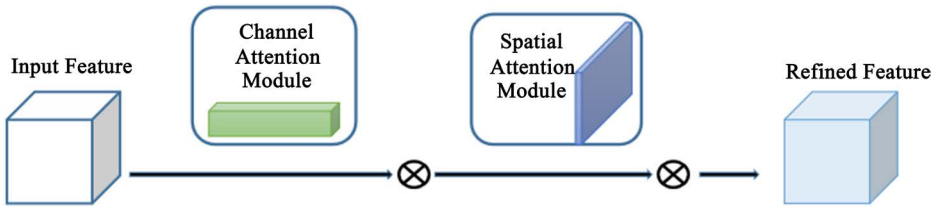

Figure 3. CBAM
图 3. 混合主力机制

2) 针对使用普通池化所存在的缺陷，本文使用 HDC 代替原始 YOLOv4 网络中的 SPPNet。

空洞卷积就是在卷积核中填充 0，使用不同大小的空洞率，同样达到增大感受野的目的，还使得获取的特征信息来自不同尺度。如图 4 所示，是空洞率为 2 的空洞卷积。不直接使用空洞卷积是因为空洞卷积虽然增大了浅层特征的感受野，但是容易导致丢失局部信息。

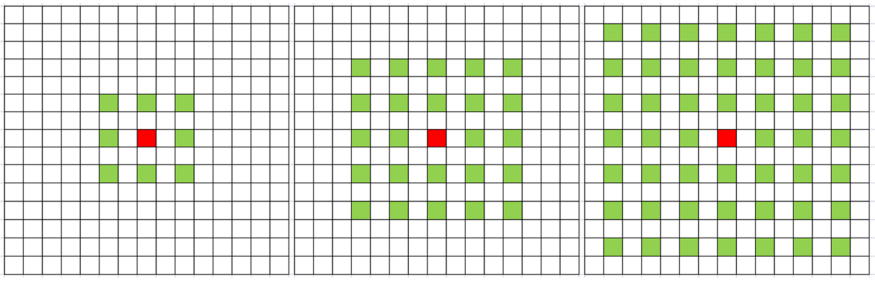

Figure 4. The rate of atrous convolution = 2
图 4. 空洞率 = 2 的空洞卷积

3. 实验与分析

3.1. 实验平台数据

本次实验算法基于 CUDA 10.2 以及 pytorch1.2.0 编程语言实现，使用了 NVIDIA Geforce GTX 1080ti 11G 进行训练。超参数设置如下：Batch size = 4，初始学习率为 0.001，Epoch = 100，输入图片尺寸为 416 × 416。

为了验证本次实验的有效性，本次实验的所有对比试验均在同等的环境下，采用的数据集为公开数据集 PASCAL VOC2007 + 2012，该数据集一共 11,530 张标注过的图片，共 27,450 个物体被标定，一共 20 个类别。用 Mobilenetv2、v3 以及 ShuffleNet 分别代替原网络中的目标提取网络作为对比实验。使用本文改进的 YOLOv4 进行网络模型训练。

3.2. 实验结果分析

本次实验采用 P-R 曲线作为模型性能的测评标准，P-R 曲线以召回率(Recall, R)为横坐标，以精确率(Precision, P)作为纵坐标。经过实验测试，得到各目标的检测性能。由于所用数据集类别较多，选取其中 car 类别和 bicycle 类别的 PR 曲线做对比如图 5 和图 6 所示，该曲线所围成的阴影部分的面积即为平均精确度值。

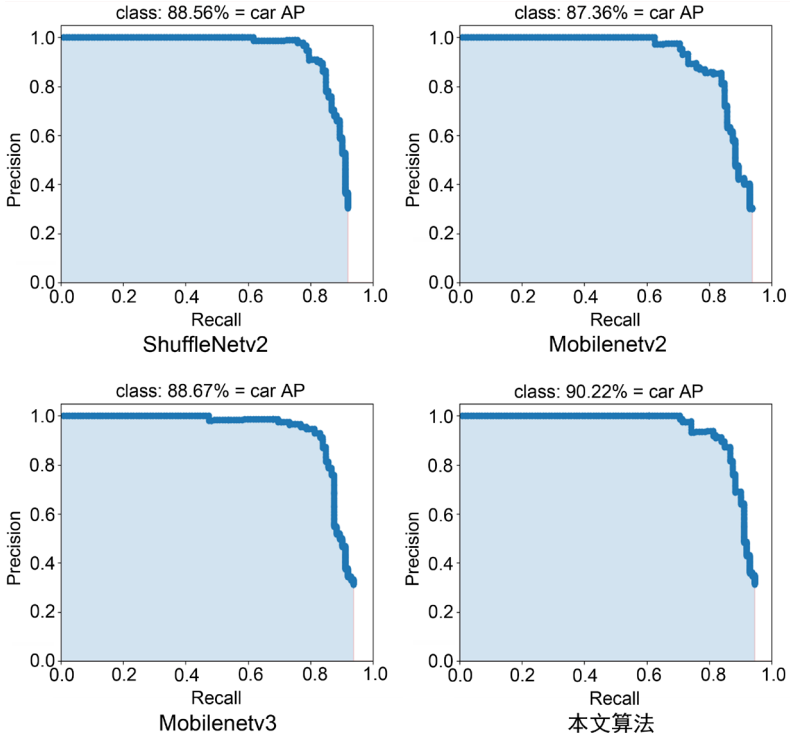

Figure 5. P-R curves of car class under different lightweight models
图 5. 不同轻量化模型下 car 类的 P-R 曲线

如表 1 所示，各个轻量化网络的参数量都有所减少，如图 7 所示，是该数据集在不同轻量化模型下部分类别的 mAP，本文算法 mAP 有所提升，速度提高了很多。

Table 1. Comparison of each lightweight model
表 1. 各个轻量化模型对比

Model (基于 YOLOv4)	Parms	浮点运算量	mAP/%	FPS
ShuffleNetv2	48.42M	120.46	81.96	22.62
Mobilenet_v2	44.74M	115.83	82.73	23.34
Mobilenet_v3	43.58M	107.48	87.86	26.71
本文算法	19.53M	62.56	89.54	44.63

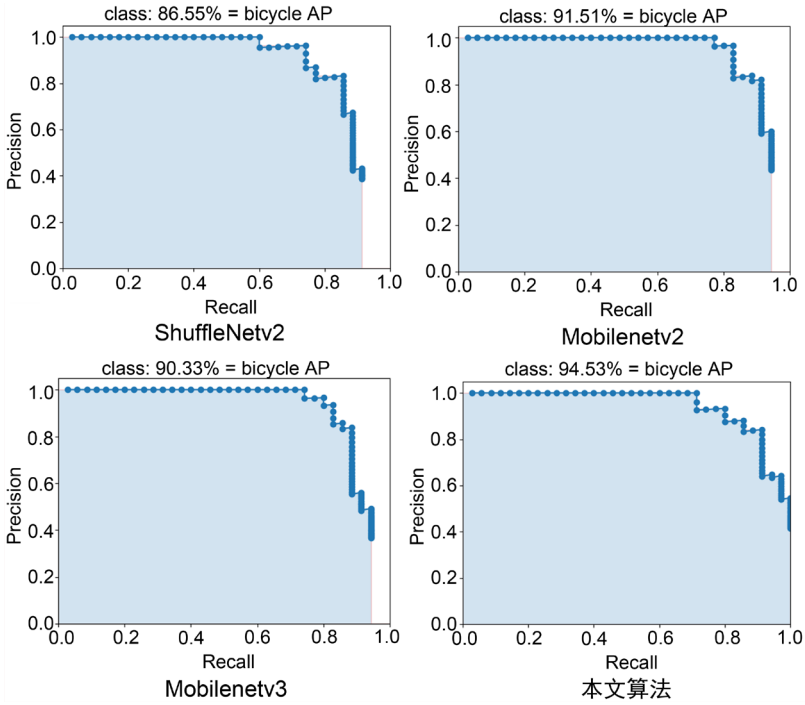

Figure 6. P-R curves of bicycle class under different lightweight models
图 6. 不同轻量化模型下 car 类的 P-R 曲线

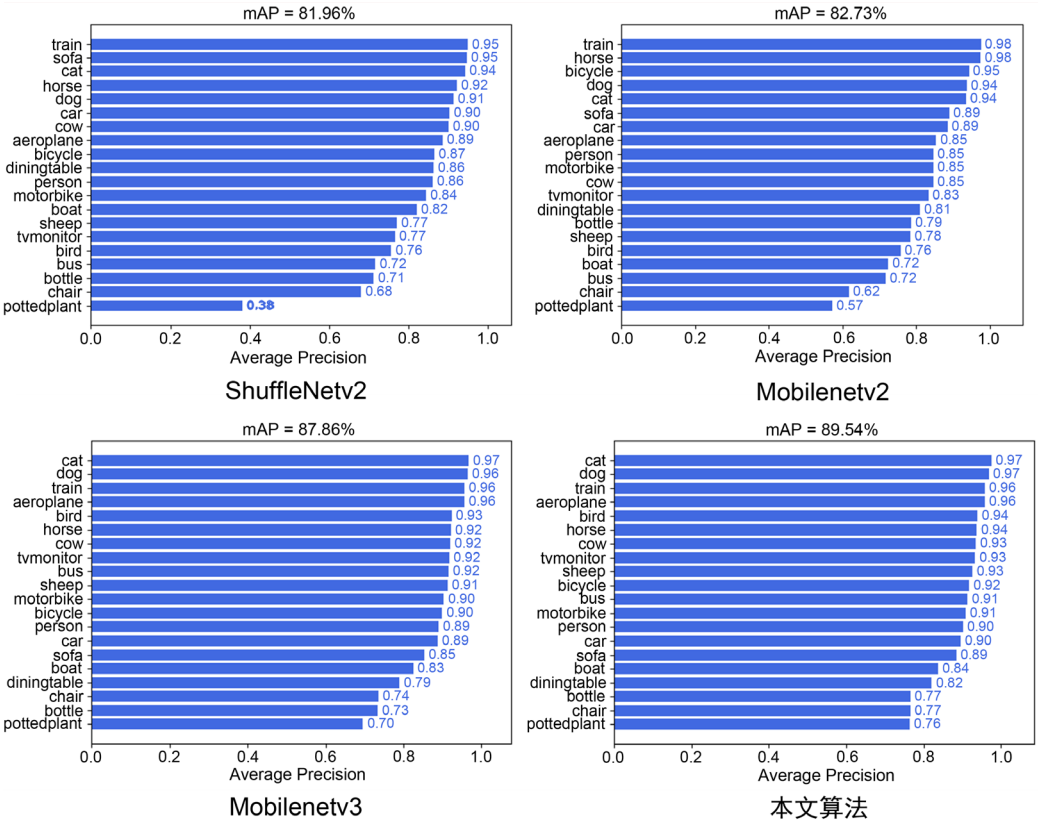

Figure 7. mAP comparison of each lightweight model
图 7. 各个轻量化模型的 mAP

如图 8 所示，是部分图片的目标识别的可视化图，图 9 是自取数据测试的可视化图。

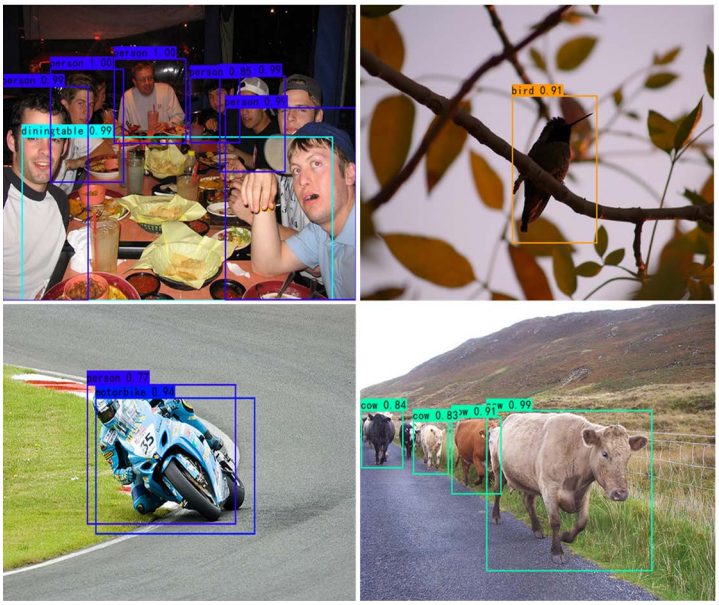

Figure 8. Visualization of the test section of the dataset
图 8. 数据集测试部分可视化图

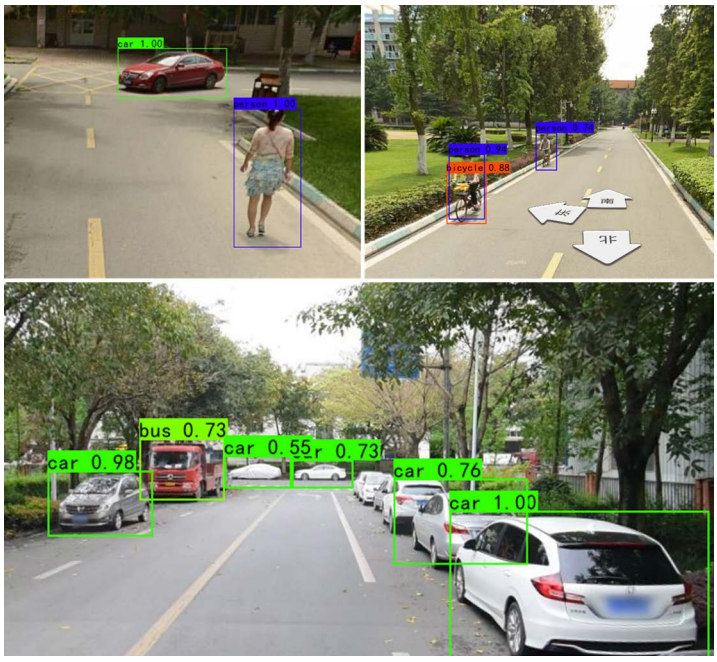

Figure 9. Visualization of self-fetching data test
图 9. 自取数据测试可视化图

4. 结论

针对目前 YOLOv4 网络模型过大，导致检测效率低下，本文在 YOLOv4 的基础上，对网络模型进行轻量化改造。使用 GhostNet 通过用标准卷积产生部分特征图，再使用这部分特征图通过线性运算产生相似特征图，减少了卷积操作，从而减少浮点运算量，达到提升检测速率的效果。使用 HDC 代替原网络中

SPP 网络，增大感受野的同时，减少了浅层网络的特征丢失，也增加了不同尺度的特征信息。为了提高模型的泛化能力，在引入 Ghost 的基础上加上了 CBAM 模块，使得实验结果在提升速度的情况下保证精度。通过几种轻量化网络的对比实验，实验结果证明，本文算法有效地减少了浮点运算量，对模型进行了有效的压缩，并且在保证精度的情况下，在速度上得到了很大的提升，但本文的 HDC 中，空洞率不宜过大，过大时会丢失图片的全局信息，反而适得其反。

基金项目

西南民族大学研究生创新型科研项目(项目编号：CX2021SP103)。

参考文献

[1] Girshick, R., Donahue, J., Darrell, T., *et al.* (2016) Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 142-158. https://doi.org/10.1109/TPAMI.2015.2437384

[2] Ren, S.Q., He, K.M., Girshick, R., *et al.* (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[3] 帅泽群, 李军. 基于深度学习的目标检测研究[J]. 汽车工程师, 2021(5): 11-14.

[4] Deng, J., Xuan, X.J., *et al.* (2020) A Review of Research on Object Detection Based on Deep Learning. *Journal of Physics*: *Conference Series*, **1684**, Article ID: 012028. https://doi.org/10.1088/1742-6596/1684/1/012028

[5] Liu, J. and Wang, X.W. (2020) Tomato Diseases and Pests Detection Based on Improved Yolo V3 Convolutional Neural Network. *Frontiers in Plant Science*, **11**, 898. https://doi.org/10.3389/fpls.2020.00898

[6] Wan, J.X., Jian, D.F. and Yu, D.Z. (2021) Research on the Method of Grass Mouse Hole Target Detection Based on Deep Learning. *Journal of Physics: Conference Series*, **1952**, Article ID: 022061. https://doi.org/10.1088/1742-6596/1952/2/022061

[7] Iandola, F.N., Han, S., Moskewicz, M.W., *et al.* (2016) SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <0.5 MB Model Size. ArXiv:1602.07360

[8] Howard, A.G., Zhu, M.L., Chen, B., *et al.* (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv:1704.04816

[9] Mark, S., Andrew, H., Zhu, M.L., *et al.* (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. *The IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018, 4510-4520.

[10] Ramin, S., Javad, F. and Farshad, K. (2021) A New Damping Strategy of Levenberg-Marquardt Algorithm with a Fuzzy Method for Inverse Heat Transfer Problem Parameter Estimation. *International Communications in Heat and Mass Transfer*, **126**, Article ID: 105433. https://doi.org/10.1016/j.icheatmasstransfer.2021.105433

[11] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., *et al.* (2019) Searching for MobileNetV3. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 Ocober-2 November 2019, 1314-1324. https://doi.org/10.1109/ICCV.2019.00140

[12] 邵伟平, 王兴, 曹昭睿, 白帆. 基于 MobileNet 与 YOLOv3 的轻量化卷积神经网络设计[J]. 计算机应用, 2020, 40(S1): 8-13.

[13] 刘晋, 邓洪敏, 徐泽林, 杨洋. 面向目标识别的轻量化混合卷积神经网络[J/OL]. 计算机应用: 1-8. http://kns.cnki.net/kcms/detail/51.1307.tp.20210622.0848.002.html, 2021-09-14.

[14] Fu, H.X., Song, G.Q. and Wang, Y.C. (2021) Improved YOLOv4 Marine Target Detection Combined with CBAM *Symmetry*, **13**, 623. https://doi.org/10.3390/sym13040623

[15] 孔维刚, 李文婧, 王秋艳, 曹鹏程, 宋庆增. 基于改进 YOLOv4 算法的轻量化网络设计与实现[J/OL]. 计算机工程: 1-10. https://doi.org/10.19678/j.issn.1000-3428.0060948, 2021-09-14.

[16] 杨玉敏, 廖育荣, 林存宝, 倪淑燕, 吴止锾. 轻量化卷积神经网络目标检测算法综述[J]. 舰船电子工程, 2021, 41(4): 31-36.

[17] Han, K., Wang, Y.H., Tian, Q., *et al.* (2020) GhostNet: More Features from Cheap Operations. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, 13-19 June 2020, 1577-1586. https://doi.org/10.1109/CVPR42600.2020.00165