

基于特征映射贡献度的轻量化路面损伤检测方法

王悦铭 吴泱序 常志宇 陈平

Lightweight pavement damage detection method based on feature mapping contribution degree

WANG Yueming, WU Yangxu, CHANG Zhiyu, CHEN Ping

引用本文:

王悦铭, 吴泱序, 常志宇, 等. 基于特征映射贡献度的轻量化路面损伤检测方法[J]. 应用光学, 2025, 46(3): 652–662. DOI: 10.5768/JAO202546.0302005

WANG Yueming, WU Yangxu, CHANG Zhiyu, et al. Lightweight pavement damage detection method based on feature mapping contribution degree[J]. Journal of Applied Optics, 2025, 46(3): 652–662. DOI: 10.5768/JAO202546.0302005

在线阅读 View online: <https://doi.org/10.5768/JAO202546.0302005>

您可能感兴趣的其他文章

Articles you may be interested in

基于RT-DETR的轻量化交通标志检测

Lightweight traffic sign detection based on RT-DETR

应用光学. 2025, 46(2): 300–308 <https://doi.org/10.5768/JAO202546.0202002>

基于NPU的实时深度学习跟踪算法实现

Implementation of real-time deep learning tracking algorithm based on NPU

应用光学. 2022, 43(4): 682–692 <https://doi.org/10.5768/JAO202243.0402003>

基于知识蒸馏的夜间低照度图像增强及目标检测

Nighttime low-light image enhancement and object detection based on knowledge distillation

应用光学. 2023, 44(5): 1037–1044 <https://doi.org/10.5768/JAO202344.0502004>

基于轻量化卷积神经网络的光伏电池片缺陷检测方法研究

Defects detection method of photovoltaic cells based on lightweight convolutional neural network

应用光学. 2022, 43(1): 87–94 <https://doi.org/10.5768/JAO202243.0103003>

轻量化铝反射镜拓扑优化与加工试验研究

Topological optimization and machining experiment of lightweight aluminum mirror

应用光学. 2024, 45(5): 1034–1041 <https://doi.org/10.5768/JAO202445.0505001>

基于改进YOLOv5s模型的红外弱小目标检测方法

Infrared dim target detection method based on improved YOLOv5s model

应用光学. 2024, 45(5): 975–981 <https://doi.org/10.5768/JAO202445.0502005>



关注微信公众号, 获得更多资讯信息

文章编号: 1002-2082 (2025) 03-0652-11

引用格式: 王悦铭, 吴泱序, 常志宇, 等. 基于特征映射贡献度的轻量化路面损伤检测方法 [J]. 应用光学, 2025, 46(3): 652-662.

WANG Yueming, WU Yangxu, CHANG Zhiyu, et al. Lightweight pavement damage detection method based on feature mapping contribution degree[J]. Journal of Applied Optics, 2025, 46(3): 652-662.



在线阅读

基于特征映射贡献度的轻量化路面损伤检测方法

王悦铭, 吴泱序, 常志宇, 陈 平

(中北大学 信息探测与处理山西省重点实验室, 山西 太原 030051)

摘 要: 道路损伤检测在道路维护过程中至关重要, 针对目前的道路损伤检测依赖于大型车载检测设备导致其成本过高且普适性差的问题, 提出了一种应用于边缘设备的超轻量化路面裂纹检测模型。首先构建轻量化的道路损伤检测网络, 在此基础上基于特征映射贡献度剪枝策略, 删除网络中的冗余参数, 保证在边缘设备中进行实时监测的同时不降低检测精度。通过实验对提出的模型进行了验证, 结果表明本文所提出的模型大小仅有 2.06 M, 在帧率 40 的情况下准确率达到 68.6%, 相较于原始的 YOLOv5s 模型, 在准确率仅降低 3.4% 的情况下, 参数量降低了 70.8%, 在边缘平台推理延时降低 175 ms。综上所述, 所提出的模型在能够更有效地检测道路特征的同时, 可以保持高水平的准确性, 对于道路损伤检测具有良好的鲁棒性。

关键词: 深度学习; 目标检测; 剪枝; 模型压缩; 轻量化

中图分类号: TN206; TP183

文献标志码: A

DOI: 10.5768/JAO202546.0302005

Lightweight pavement damage detection method based on feature mapping contribution degree

WANG Yueming, WU Yangxu, CHANG Zhiyu, CHEN Ping

(Shanxi Key Laboratory of Signal Capturing and Processing, North University of China, Taiyuan 030051, China)

Abstract: Road damage detection is of vital importance in the process of road maintenance. In order to solve the problem that the current road damage detection relies on large vehicle-mounted detection equipment, resulting in high cost and poor universality, an ultra-lightweight surface crack detection model applied to edge equipment was proposed. Firstly, a lightweight road damage detection network was constructed, and on this basis, redundant parameters in the network were deleted based on the feature mapping contribution pruning strategy, so as to ensure real-time monitoring in edge devices without reducing the detection accuracy. Finally, the experimental results show that the size of the proposed model is only 2.06 M, and the accuracy rate reaches 68.6% at the frame rate of 40. Compared with the original YOLOv5s model, when the accuracy is only reduced by 3.4%, the number of parameters is reduced by 70.8%, and the inference delay on the edge platform is reduced by 175 ms. In summary, the proposed model can detect road characteristics more effectively while maintaining a high level of accuracy, and has good robustness for road damage detection.

收稿日期: 2024-07-22; 修回日期: 2025-01-09

基金项目: 国家自然科学基金 (62122070); 国家自然科学基金青年基金资助项目 (62103384); 山西省研究生教育创新计划 (2024KY608); 中北大学研究生科技立项 (20231928)

作者简介: 王悦铭 (2001—), 女, 硕士研究生, 主要从事深度学习、边缘计算技术研究。E-mail: 835822146@qq.com

通信作者: 陈平 (1983—), 男, 博士, 教授, 博士生导师, 主要从事信号与信息处理、图像处理与重建、人工智能技术研究。

E-mail: pc0912@163.com

Key words: deep learning; object detection; pruning; model compression; lightweight

引言

高速公路发展已进入“建设与养护”并重时期, 养护技术的科学化和数字化已逐步成为高速公路养护工作的关注重点。目前, 高速公路养护监测仍以人工路面巡查为主, 但是对于复杂路段和突发灾害的监测, 养护快速响应和施工安全防护逐渐成为有待突破的难题。巡查车辆的监测与防控是实现该场景下信息感知与决策的最有效手段。目前, 大部分道路损伤检测主要依赖专用车辆搭载大型检测设备进行, 但是大型检测设备不具备便携性与普适性。随着边缘计算与图像处理技术的日益发展, 采用低功耗的嵌入式设备^[1-3]来自动识别路面状态具有重要意义, 而低功耗设备的计算能力通常低于大型服务器, 内存、处理能力和电池寿命有限, 因此需要设计轻量化的算法, 以确保低功耗设备在资源受限的环境中仍能提供高质量的服务。

传统的道路损伤检测多依赖数字图像处理算法, 包括基于拉普拉斯算子的边缘检测方法^[4]、形态学方法^[5]和区域生长的分割算法^[6], 然而这些传统方法需要人工提取病害特征, 在复杂的自然道路场景中难以检测多种类型的路面损伤, 导致实用性较差。而随着 ResNet^[7]、VGG^[8]、YOLO^[9]等大量深度学习算法的提出, 为道路损伤的检测和识别提供了新的思路。WANG W^[10]等利用经典的双阶段 Faster R-CNN 网络对裂缝进行识别。SU B^[11]等从训练角度出发, 认为路面的个别损伤类型的数据稀少, 提出了基于小样本学习的三阶段路面损伤检测算法。YIN T^[12]等提出跨层注意力机制, 在更深的特征层中检测裂纹。基于深度学习的研究方法^[10-12]目前在道路损伤检测和识别方面已取得了较高的精度。然而, 以上的目标检测网络倾向于使用更深层和大型的架构, 直接导致了模型参数数量的增加, 以及较高的存储需求和复杂的计算量, 难以部署在内存和运算能力相对有限且对实时性要求较高的嵌入式设备和移动端。尽管已有研究如 YUAN Y C^[3]等提出的道路损伤检测的边缘云框架 EcRD 将检测过程下放到边缘端, GONAL A^[13]等基于 YOLOv5 模型对路面缺陷进行识别分类, 并将模型部署在 Android 设备上, 但是未能充分利用边缘端和终端资源, 导致检测精度降低。因此, 在保证检测精度的同时降低模

型复杂度, 是当前研究的热点与难点。

针对上述问题, 采用量化^[14]、剪枝^[15]、蒸馏^[16]等模型压缩策略, 旨在提升模型的性能和效率。其中, 剪枝技术通过直接删除模型中冗余或不必要的参数和连接, 能够显著减小模型的大小, 使其更适合在资源受限的设备上部署。这种直接减小模型大小的方式, 相较于其他方法(量化、知识蒸馏), 更为灵活和高效。现有剪枝方法^{[15][17]}基于卷积核的 L1 范数、L2 范数, 将其作为剪枝标准, 针对卷积核进行排序, 但是压缩模型的过程中忽略了中间激活层的大小。因此, LUO J H^[18]等提出一种基于熵值的裁剪方式, 利用熵值判断卷积核的重要性, 同时下一层相应的卷积核也被删除, 但仍可能误删重要通道。YE H^[19]等提出多任务 CNN 模型全局通道剪枝框架, 考虑了层内和层间的卷积核对多任务模型压缩性能的共同影响。以上的剪枝算法都是基于简单的分类任务完成的, 且剪枝过程缺乏可解释性。因此, 针对特定任务的压缩部署, 需进一步优化剪枝策略以增强其实用性与可解释性。

本文针对在低功耗边缘设备部署道路损伤检测模型时面临的低效率、高时延的问题, 对目标检测模型剪枝过程中选取剪枝准则的合理性以及有效性进行研究。主要贡献如下:

1) 提出基于特征映射贡献度的剪枝策略, 选取均值、方差、秩、中位数这四种评价指标作为特征映射的贡献度, 从神经网络可视化角度对特征贡献度进行展示, 用信息熵判别贡献度的优劣性, 选取平均秩作为特征映射的贡献度。

2) 基于单阶段的目标检测框架设计轻量化的网络, 采用轻量化 MobileNetV3^[20]网络作为主干特征提取网络, 将经过剪枝后的模型的特征融合部分引入改进的 SE 注意力机制(advanced squeeze-and-excitation networks, ASE)。

3) 通过对比实验和消融实验验证了本文所提出算法的有效性和优越性, 并成功将轻量化模型部署在 Atlas 200DK A2 设备上, 实现了道路损伤的实时高效检测。

1 建模方法

本文提出的轻量化路面损伤检测算法是端到端的神经网络模型, 采用通道剪枝技术对模型进行压缩, 需要经过训练、剪枝、微调 3 个过程, 最终

得到高精度且轻量化的算法网络。本文的神经网络模型整体框架如图1所示,首先,构建轻量化的目标检测网络,但是其计算量仍不足以在嵌入式设备中实时运行,为此,本文提出了基于特征映射

贡献度进行模型剪枝的策略,通过分析特征映射的秩识别特征图中的冗余特征,并利用注意力机制保留其关键特征,实现目标检测网络轻量化的同时提高检测精度。

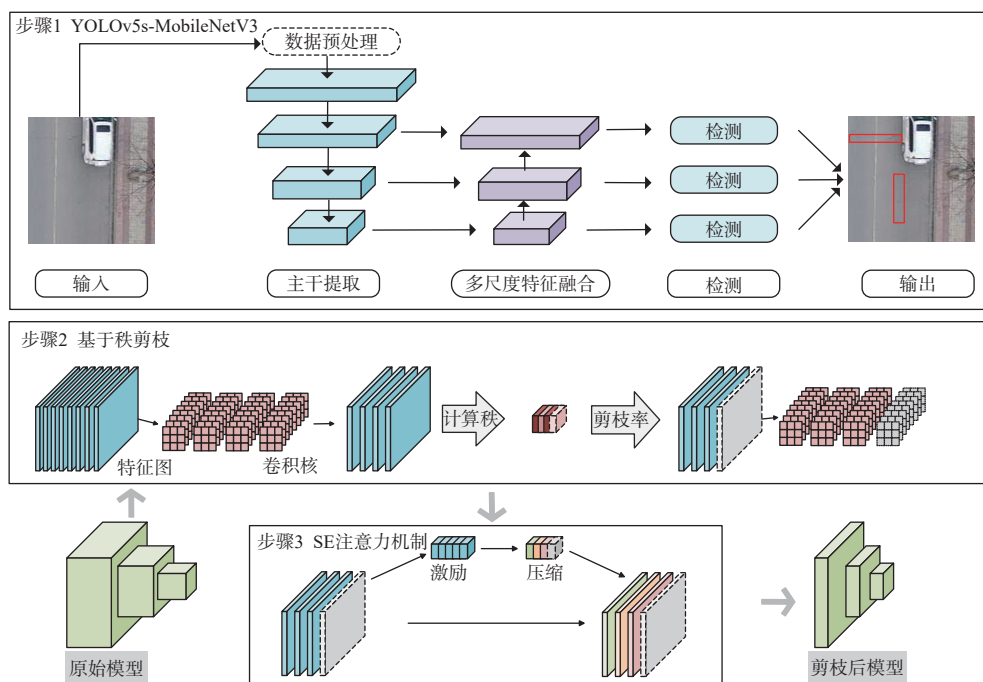


图1 模型整体结构图

Fig. 1 Overall structure diagram of model

1.1 轻量化道路损伤检测模型

1.1.1 MobileNetV3

如图2所示, MobileNet系列通过引入深度可分离卷积,将卷积操作进行分解。首先对通道进行卷积操作,各通道与卷积核之间相互对应,但此时未改变特征图的输出通道,各特征图在深度维度上缺乏信息交流。随后通过逐点卷积,实现特征通道数的改变,完成通道之间的信息交流。

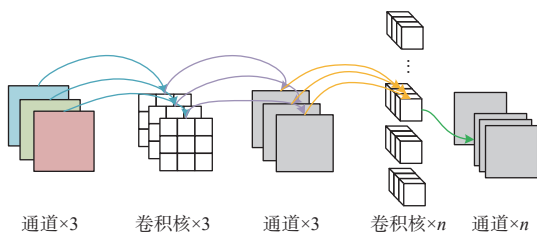


图2 MobileNetV3中的深度可分离卷积

Fig. 2 Depthwise separable convolution in MobileNetV3

如图3所示, MobileNetV3的核心在于优化卷积操作的同时引入了逆残差结构。在逆残差结构中,首先对输入图像进行维度提升,这有助于捕捉

更丰富的特征信息,完成后将高维特征恢复为低维特征,并将两个低维特征图通过恒等路径进行连接,避免了深度卷积核的参数较多为0的情况。同时,引入 ReLU6 与 Hard-sigmoid 激活函数,促进稀疏特征学习并加速推理,简化计算复杂度,其计算公式分别如式(1)、式(2)所示:

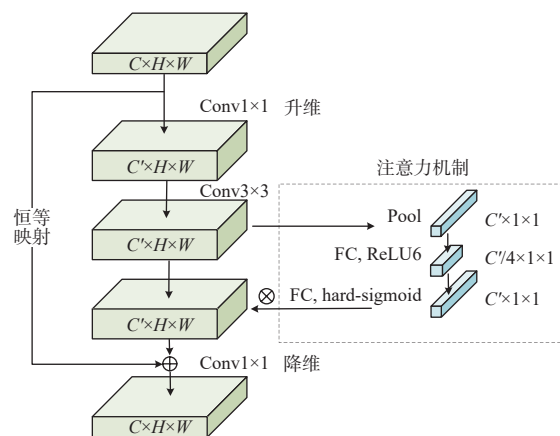


图3 MobileNetV3 结构图 M_block

Fig. 3 M_block: Structure diagram of MobileNetV3

$$\text{ReLU6}(x) = \begin{cases} 6, & x \geq 6 \\ x, & 0 \leq x < 6 \\ 0, & x < 0 \end{cases} \quad (1)$$

$$h_{\text{sigmoid}}(x) = \frac{\text{ReLU6}(x+3)}{6} \quad (2)$$

式中 x 是输入的特征图。

1.1.2 YOLOv5s-MobileNetV3

文献[21,22]说明在嵌入式平台使用YOLOv5s网络性能表现更好, YOLOv5s网络相比于YOLO最新系列网络, 结构上更为紧凑, 易于在资源受限的低功耗设备上部署。因此, 本文所提出的轻量

化路面损伤检测算法基于YOLOv5s网络设计。YOLOv5s的网络结构包括输入(Input)、主干提取(Backbone)、特征融合(Neck)和预测(Head)4个部分。由于模型在构建过程中大量使用了跨阶段的局部结构(cross stage partial, CSP), 通过特征图分割与融合机制, 实现特征快速降维, 有效降低模型参数量并加速检测, 但仍难以满足嵌入式设备的低内存要求。为此, 在原有网络的基础上, 采用MobileNetV3网络替换原有的主干(Backbone)网络, 完整的轻量化目标检测网络结构如图4所示。

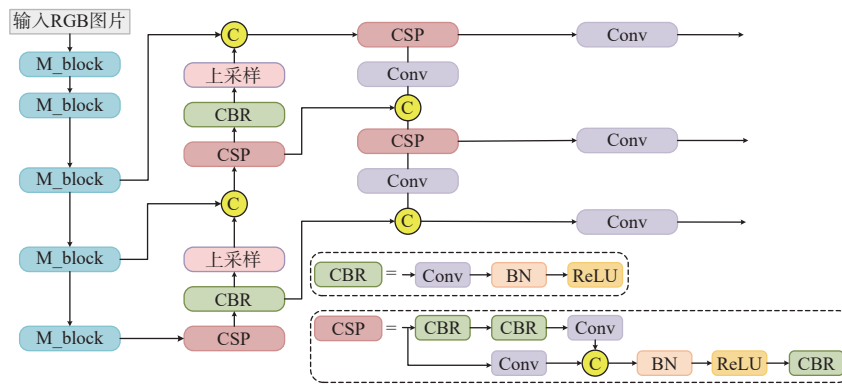


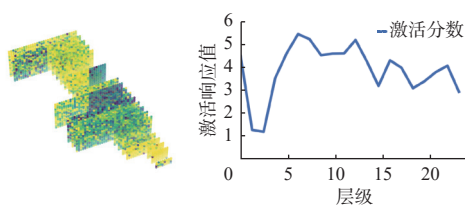
图4 轻量化道路损伤检测网络

Fig. 4 Lightweight road damage detection network

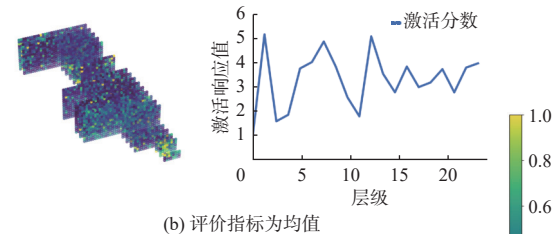
1.2 基于特征映射的剪枝

1.2.1 基于特征映射贡献度的剪枝准则

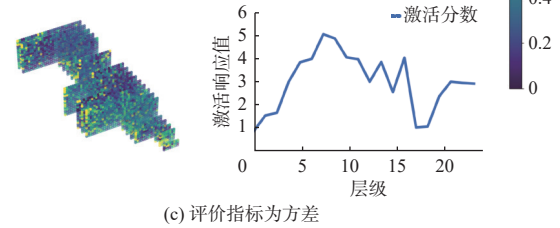
神经网络在剪枝的过程中, 不同的剪枝准则会影响压缩后模型的性能, 因此选取合适的剪枝准则至关重要。为了比较不同评价指标 E 是如何影响判断网络冗余性的, 本文选取平均秩、均值、方差和中位数4个指标对激活输出进行可视化, 体现模型压缩过程中激活值的筛选过程。图5中, 左图为可视化的YOLOv5s网络整体结构图, 给定网络的输入 $X \in \mathbb{R}^{c \times w \times h}$, $\mathbb{R}$ 为特征图, w 和 h 分别表示特征图的宽和高, c 表示特征图数量, 一个模块的激活输出为 $Y \in \mathbb{R}^{c \times w \times h}$, 使用评价指标 E 将输出 Y 压缩为 $y' \in \mathbb{R}^{c \times 1 \times 1}$, 激活输出的变换公式如式(3)所示:



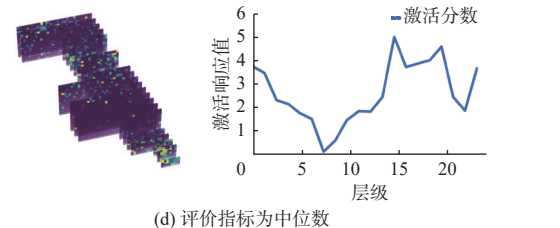
(a) 评价指标为平均秩



(b) 评价指标为均值



(c) 评价指标为方差



(d) 评价指标为中位数

图5 YOLOv5s-MobileNetV3 通道权重可视化及其激活响应图

Fig. 5 Channel weight visualization and activation response diagrams on YOLOv5s-MobileNetV3

$$y' = E(Y) \quad (3)$$

将激活输出 y' 进行归一化处理,消除不同指标的量纲,用公式表示为:

$$y = \frac{y' - y'_{\min}}{y'_{\max} - y'_{\min}} \quad (4)$$

认为 y 趋于1表示该组卷积核表示的信息量更多, y 趋于0则相反。激活输出 $A = [y_0, y_1, y_2, \dots, y_c]$ 表示经过一个模块后的激活输出,用散点平面表示。网络的整体结构用 $N = [A_0, A_1, A_2, \dots, A_n]$ 表示,其中 n 表示网络中模块的数量。图5(a)中的 y 落在0或1上,而图5(b)~(d)中的 y 大部分落在0~1之间。图5(a)中低层级的信息熵值较低,通过计算网络满秩的情况得出低层级的满秩通道更多,因此,特征映射的秩能够更直观的反映一组卷积核的重要性。图5(a)基于平均秩的剪枝准则去判断网络的冗余性,能够保留大多数对模型性能有重要贡献的卷积核,这样可以确保在减少模型复杂度的同时,最小化对模型性能的影响。而均值、方差和中位数等指标,虽然能在一定程度上反映

激活输出的特性,但是在判断网络冗余方面不如平均秩直观和有效。

信息熵 $H(X)$ 的计算公式如下:

$$H(X) = - \sum_{i=1}^n p(E_{C_{[i][j]}}) \log p(E_{C_{[i][j]}}) \quad (5)$$

式中: $E_{C_{[i][j]}}$ 表示第 i 层,第 j 个通道的评价指标; $p(E_{C_{[i][j]}})$ 表示评估特征映射的过程中对数据的响应情况和特征的多样性。一个层中的特征图信息熵较高,表示不同通道的特征图响应较为多样化,具有更丰富的特征表示能力,反之,信息熵较低则表示特征图的响应相对集中,可能存在较弱的多样性和表征能力。通过对比图5中右侧4幅图可以看出,基于平均秩的剪枝准则对应的信息熵相比于其他三种评价指标更能表现网络的特征多样性。因此,本文选用特征映射的平局秩作为剪枝准则。

1.2.2 基于特征映射秩的剪枝

基于特征映射秩的剪枝策略如图6所示,本文所提出的剪枝方法流程为:计算特征映射的平均秩并进行排序,为网络的每层设定个性化阈值。

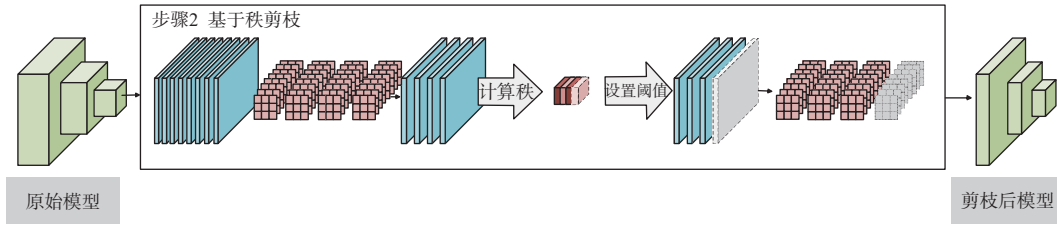


图6 基于特征映射秩的剪枝过程

Fig. 6 Pruning process based on feature mapping rank

卷积核剪枝的目的是识别并去除网络中不太重要的卷积核,可定义为一个优化问题:

$$\min_{\delta_j} \sum_{j=1}^C \delta_j L(v_j) \quad (6)$$

式中: C 表示一个卷积层中卷积核的个数;用 I 表示保留的卷积核; N 表示移除的卷积核,其中将每一层的卷积核 V 分为两个部分, $I = [v_{I_1}, v_{I_2}, \dots, v_{I_{n_1}}]$, $N = [v_{N_1}, v_{N_2}, \dots, v_{N_{n_2}}]$, 其中 $n_1 + n_2 = C$, $I \cap N = \emptyset$, $I \cup N = V$; δ_j 的值取1和0,1表示当通道定义为重要通道,0表示该通道为移除的通道; $L(\cdot)$ 表示由卷积核生成特征图的参数的冗余性。本文用秩衡量通道的重要性:

$$\begin{aligned} \min_{\delta_j} \sum_{j=1}^C \delta_j (v_j) \sum_{t=1}^g \text{Rank}(o_j(t, :, :)) \\ \text{s.t.} \sum_{j=1}^C \delta_j = I_{n_1} \end{aligned} \quad (7)$$

式中: $\text{Rank}(\cdot)$ 表示特征图的秩; $o_j(t, :, :)$ 表示特征图。基于1.2节所阐述的剪枝准则,本算法对网络每一层设定了个性化的剪枝阈值,以此实现神经网络模型的有效压缩,并赋予该压缩过程以高度的可解释性。

1.3 SE 注意力机制

SENet^[23]是典型的通道注意力网络,由于在神经网络中所提取到的特征重要程度不同,通过增强重要特征、抑制一般特征可提高模型效果。

将特征映射为低秩的卷积核移除后,在Neck部分的ReLU激活层后添加SE注意力机制,其示意图如图7所示,激活权重作为特征图中每个通道的重要性的评价指标,考虑到特征图中的元素有正有负,为了尽可能多地保留特征图中的语义信息,充分展现每个特征图的关键程度,将特征图的

所有元素先取绝对值, 再进行全局平均池化操作。给定的特征图 $X = [x_1, x_2, \dots, x_c]$, 其中 $x_c \in \mathbb{R}^{w \times h}$, 将网络中每一幅特征图的像素值取绝对值之后再

求平均得到 z_c , 通过下式计算:

$$z_c = \frac{1}{w \times h} \sum_{i=1}^w \sum_{j=1}^h |x_c(i, j)| \tag{8}$$

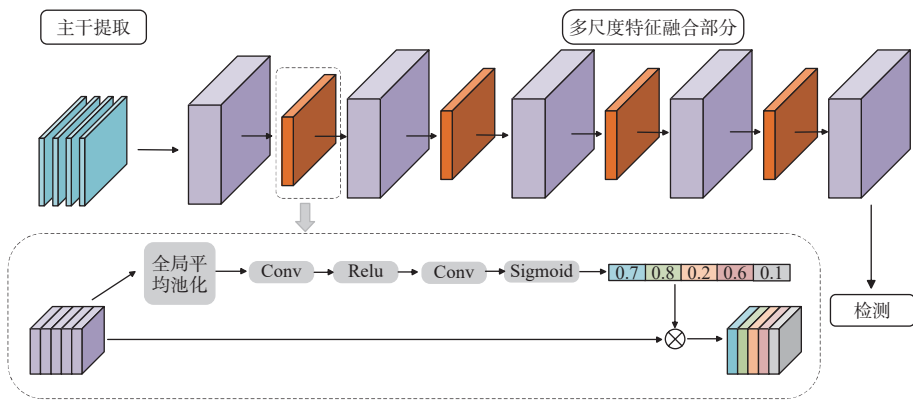


图 7 Neck 部分每个卷积层添加 SE 注意力机制

Fig. 7 Adds SE attention mechanism to each convolutional layer on Neck section

2 实验

2.1 实验设置和数据集

1) 数据集

本文选取 IEEE 在 2022 年大数据杯上发布的 RDD2022^[24] 数据集, 该数据集涉及印度、日本、捷克共和国、挪威和中国 6 个国家的路面损伤图像。中国的路面损伤图像是通过两种拍摄角度获取的, 分别是将相机搭载到摩托车和无人机上进行拍摄。选用无人机角度所拍摄的数据集 China_Drone, 共 2477 张图片, 尺寸大小为 512×512 像素, 且包含五种类型的损伤, 其详细信息如表 1 所示, 由于 China_Drone 数据集不包含测试集, 因此, 本文按照 7 : 2 : 1 的比例划分数据集为训练集、验证集和测试集。

表 1 RDD2022 数据集中 China_Drone 数据集的损伤定义及其数量统计

Table 1 Damage definition and quantity statistics of China_Drone dataset in RDD2022 dataset

损伤类型	名称	标签数量
纵向裂纹(Longitudinal crack)	D00	1426
横向裂纹(Transverse crack)	D10	1263
龟状裂纹(Alligator crack)	D20	293
坑洞(Potholes)	D40	86
修复过路面(Repaired road)	Repair	769

2) 实验设置及其评价参数

本实验环境为 Ubuntu18.04 操作系统, 基于 Pytorch1.8 框架, 显卡为 RTX 4090D, CPU 为 Intel(R)

Core(TM)i7-9750H, 内存为 8.0 GB。输入图像设置为 512×512 像素、使用 SGD 优化器、输入网络的批次大小为 16、初始学习率为 0.01、训练周期 300 轮、微调周期 30 轮。为保证实验结果的可靠性和准确性, 在训练阶段均不使用预训练权重。当网络完成裁剪之后, 选择继承原始网络的卷积核权重和偏置等参数, 通过微调恢复剪枝网络的性能。本实验采用与 YOLOv5s 算法相同的数据增强算法。使用每秒浮点运算数(FLOPs)、网络参数量(Parameters)、精度(mAP)、以及推理速度(FPS)作为评价指标对目标检测网络进行对比。

3) 模型部署

本文使用的嵌入式设备是华为 Atlas 200DK A2, 搭载昇腾 AI 处理器, 使用的推理方式是离线推理。在 PC 端对 pt 格式模型转换为通用的神经网络交换格式(open neural network exchange, ONNX), 在边缘段使用昇腾张量编译器(ascend tensor compiler, ATC)将 ONNX 转换为适配昇腾 AI 处理器的离线模型(*.om 文件)。

2.2 实验结果

为了验证所提出方法的有效性, 本文设计了 6 组实验。第一组对比实验验证 MobilenetV3 作为 YOLOv5s 的主干提取网络相比于其他网络的性能提升; 第二组实验验证本文所提出的基于特征映射贡献度的剪枝方法, 第三组实验比较不同的剪枝率对网络性能的影响; 第四组实验验证不同插入方式的 ASE 注意力机制在特征融合中对网络性能的影响; 第五组实验对比不同的剪枝方法的性

能;第六组实验对比所设计的轻量化网络与当前主流的网络的性能。

1) 实验一

为了评估提出的目标检测算法性能,将目前主流的轻量化主干网络算法 ShuffleNetV2^[25]、EfficientNet^[26]、MobileNetV3^[20] 以及原始网络的 4 种模型进行对比实验,结果如表 2 所示,用 MobileNetV3 替换掉原始 YOLOv5s 中的主干提取部分, MobileNetV3 结构中的深度可分离卷积更易于在低功耗设备上部署推理,不仅显著减少了参数量(降低约一半),还极大提高了模型的运行推理速度。YOLOv5s-MobileNetV3 网络相较于原始网络精度降低了 2%,但是参数量降低了 52%,在边缘平台 Atlas 200DK A2 上运行模型,推理延时降低了 162 ms,推理速度大幅提升。

表 2 不同主干网络的检测性能

Table 2 Detection performance of different backbone networks

模型	Parameters/M	GFLOPs	mAP	FPS	边缘 FPS
YOLOv5s	7.02	15.8	0.72	13	5
YOLOv5s-ShuffleNetV2	3.84	6.8	0.62	28	15
YOLOv5s-EfficientNet	4.09	7.7	0.63	23	19
YOLOv5s-MobileNetv3	3.37	6.0	0.70	35	26

2) 实验二

为验证所选择的平均秩作为剪枝准则的可行性,用平均秩、均值、方差和中位数对单层特征映射过程进行修剪,固定剪枝率为 0.2,统计对应层级中所包含的信息熵的大小,如表 3 所示,基于平均秩的剪枝标准所统计的信息熵值呈现较高水平,表明基于平均秩的剪枝策略在识别并保留特征方

面具有优势,进一步证实了将其作为剪枝准则的可行性和有效性。

表 3 不同剪枝标准对特定层级剪枝后的信息熵统计值

Table 3 Information entropy statistics of different pruning standards for specific level

剪枝标准	层级						
	0	1	2	5	10	15	20
平均秩	4.47	1.25	1.43	3.97	3.91	2.91	3.00
均值	3.21	0.83	0.45	0.32	0.31	0.52	0.51
方差	2.36	1.22	1.23	0.67	0.52	0.33	0.54
中位数	1.14	0.22	0.35	0.16	0.36	0.38	0.31

3) 实验三

为了深入研究剪枝对网络模型体积和检测精度之间的关系,针对 YOLOv5s-MobileNetV3 网络,对不同剪枝比例的网络性能进行了测试。为了验证不同剪枝率对模型性能的影响,以 YOLOv5s 为基线进行实验,实验环境及配置参数不变,结果如表 4、表 5 所示。其中表 4 的前 4 行表示全局剪枝率,后 2 行表示分段剪枝率,网络模型总共有 23 个层级。

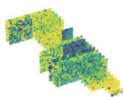
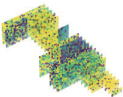
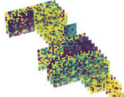
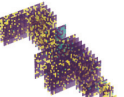
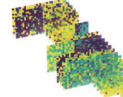
表 4 不同剪枝率的检测性能

Table 4 Detection performance of different pruning rates

剪枝率	Parameters/M	GFLOPs	mAP	FPS
0.1	3.08	6.1	0.643	37
0.2	2.77	4.9	0.612	40
0.5	1.97	3.1	0.546	42
0.8	0.60	1.3	0.462	50
0.0×7+0.1×16	2.11	4.6	0.676	39
0.0×7+0.3×16	1.82	2.9	0.657	44

表 5 不同剪枝率的可视化

Table 5 Visualization of different pruning rates

未剪枝	0.1	0.2	0.5	0.0×7+0.1×16	0.0×7+0.3×16
					

实验结果表明,随着剪枝率的增加,模型的参数量和 FLOPs(每秒浮点运算数)逐渐减少,这直接导致了模型体积的减小和计算复杂度的降低。然而,模型的精度(mAP)也呈现下降趋势,但剪枝率较低时(如 0.1 和 0.2),精度下降幅度相对较小,仍保持在较高水平。同时,剪枝后的模

型在帧率(FPS)上有所提升,表明剪枝操作有助于提升模型的推理速度。此外,基于特征映射贡献度的分段剪枝策略依据图 5(a)设计,低层级所包含的信息量较多,则剪枝率较小。分段剪枝率(0.0×7+0.3×16)在保持较高精度的同时,也实现了推理速度的提升,表明分段剪枝可以在一定程度

上平衡精度和速度。基于特征映射贡献度的剪枝策略能够识别并去除网络中冗余或不必要的参数和连接,从而减小模型体积并提升推理速度。同时,该策略保留了对模型性能有重要贡献的卷积核,确保了剪枝后的模型仍能保持较高的精度。

用固定剪枝率和分段的剪枝率对网络进行全局剪枝,用式(5)计算网络在不同剪枝率下的激活响应值,实验结果如图8所示,在固定剪枝率为0.5的修剪过程中,网络中信息熵值下降较为明显;在固定剪枝率为0.1和0.2的修剪过程中,网络中信息熵值相比于原始网络下降较少;在分段剪枝率的情况下,修剪结果介于以上两种情况之间,能够更好地表征网络中所含的信息量。

4) 实验四

改进后的YOLOv5s-MobileNetV3网络参数量明显降低,但是其计算复杂度也较高,而剪枝后的网络精度下降较为明显,因此在网络Neck部分添加注意力机制,设计消融实验。对比添加注意力机制前后的性能变化,结果如表6所示,可以看到,

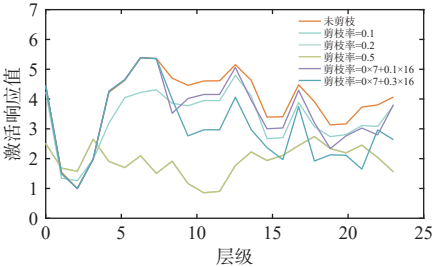


图8 不同剪枝率下的激活响应值

Fig. 8 Activation response values under different pruning rates

添加ASE注意力机制后,模型的精度有所提升,同时保持了较高的推理速度。ASE注意力机制能够自适应地调整特征图中每个通道的重要性,从而增强模型对路面损伤特征的表征能力。此外,ASE注意力机制的引入并未显著增加模型的参数量和计算复杂度,因此能够在保持较高推理速度的同时提升模型的精度。在网络的Neck部分的不同位置添加SE注意力机制,通过消融实验证明在卷积层和C3结构层添加注意力机制能够提高模型的检测精度。

表6 注意力机制的性能对比

Table 6 Improved performance comparison of attention mechanisms

Pruned-YOLOv5s-MobileNetV3	ConvSE	C3SE	ConvASE	C3ASE	Parameters/M	GFLOPs	mAP	FPS
√					1.83	2.9	0.657	44
√	√				2.09	4.3	0.663	26
√		√			2.04	3.8	0.666	40
√			√		2.09	4.3	0.668	26
√				√	2.05	3.8	0.672	40
√			√	√	2.07	4.1	0.686	40

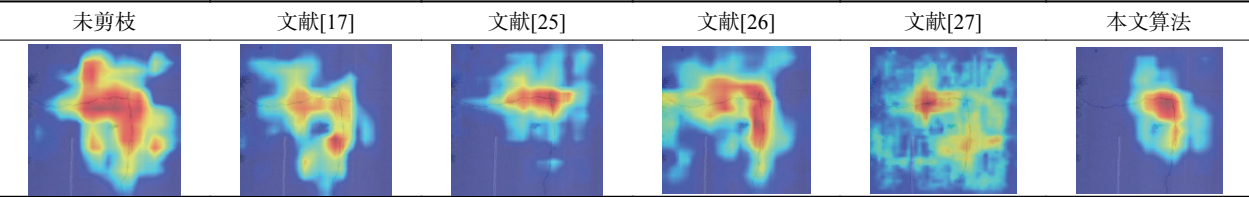
5) 实验五

为了更直观地对比其他剪枝算法与本文所提出的基于特征映射贡献度与权重向量的剪枝算法

的不同,用如表7所示的热力图形式可视化检测结果,结果表明本文算法能够集中关注路面损伤的部分。

表7 不同剪枝算法的热力图展示

Table 7 Heat map presentation of different pruning algorithms



同时本文针对剪枝算法覆盖特征的范围及其有效性进行了验证,结果如图9所示。固定剪枝率为0.3的情况下对轻量化模型进行剪枝,X轴表示

网络的层级,Y轴表示统计经过剪枝后的通道中经ReLU激活处理后的值大于所设阈值(0.5)的通道比例,可以看到,本文所提出的基于特征映射度的

剪枝算法能够保持通道中更高的激活值,实验结果验证了算法在精简模型参数、提升运算效率方面的优势。

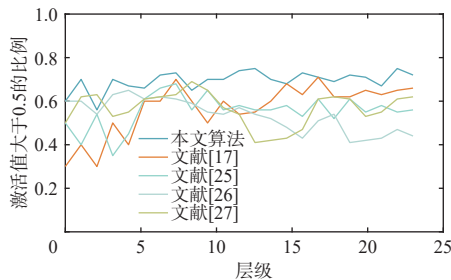


图9 统计经过 ReLU 激活处理后大于所设阈值(0.5)的比例
Fig.9 Count proportion of ReLU activation that is greater than set threshold (0.5)

6) 实验六

同类检测模型的对比实验结果如表8所示,检

测结果如图10所示。可以看到,本文所提方法能够在保证检测精度较高的情况下,使模型得到充分压缩并且显著提高检测速度。本文提出的检测算法参数量显著降低,延时仅有39 ms, FPS 可达到40,相较于YOLOv4 Tiny和YOLOx-Tiny,推理时延分别降低37.5 ms和46 ms,推理速度则分别提升60%和64%,推理速度大幅度上升。因此,改进后轻量化网络具有良好的性能。

表8 各模型推理时间对比

模型	Parameters/M	mAP	FPS
YOLOv4-Tiny	6.06	0.66	16
YOLOx-Tiny	5.06	0.67	14
FCOS-Tiny	2.08	0.65	34
本文算法	2.06	0.686	40

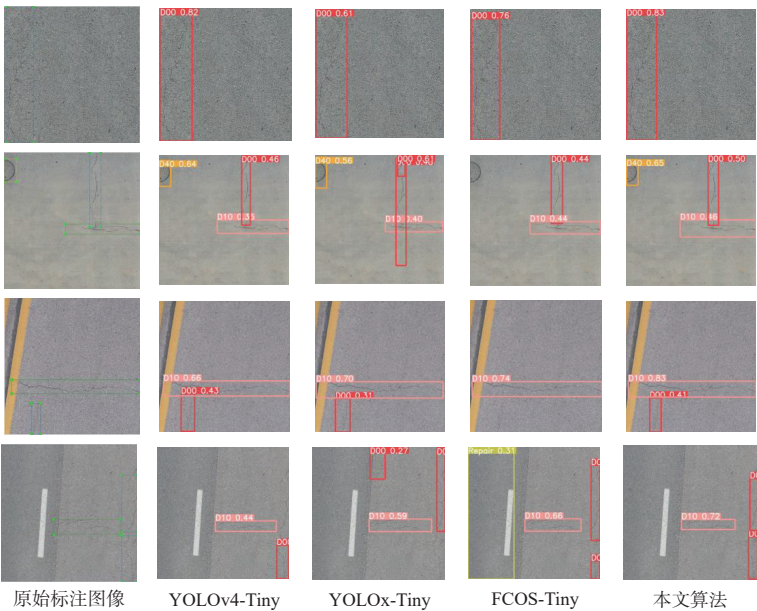


图10 本文算法在 China_Drone 测试集中的检测结果

Fig.10 Detection results of proposed algorithm on China_Drone test set

3 结论

为了解决道路损伤检测模型部署在资源受限的低功耗设备上存在的模型参数量和计算量都较大的问题,本文提出了一种基于特征映射贡献度的轻量化路面损伤检测方法,利用剪枝技术对原有网络的复杂结构进行简化,并对剪枝准则进行可视化,使模型压缩过程具有高度的可解释性。对剪枝后的网络融合注意力机制,减轻通道剪枝对模型的性能影响,利用Atlas 200DK A2的边缘计算优势,以低内存消耗提供高质量的服务。通过

边缘计算进行路面状态识别的方法不仅降低了道路养护对专业人员的依赖程度,同时嵌入式设备具有便携性,可进行无人机挂载,从而使得“天地一体”的道路养护成为可能。

通过与主流算法的对比实验以及不同剪枝率的对比实验,可以看到,改进后的网络参数量仅有2.06 M,推理速度达到40 FPS,相较于原始网络,在精度降低3.4%的情况下,参数量降低了70.8%,推理时延降低了51ms,速度提升了67.5%,有效提高了模型对低功耗设备的适应性。新的轻量化模型

结构更加精简, 计算复杂度显著降低, 模型的检测速度加快, 采用计算和存储能力有限的边缘设备实现了道路损伤的实时检测。

参考文献:

- [1] KHAN M W, OBAIDAT M S, MAHMOOD K, et al. Real-time road damage detection and infrastructure evaluation leveraging unmanned aerial vehicles and tiny machine learning[J]. *IEEE Internet of Things Journal*, 2024, 11(12): 21347-21358.
- [2] XIE Q W, HU X Y, REN L, et al. A binocular vision application in IoT: realtime trustworthy road condition detection system in passable area[J]. *IEEE Transactions on Industrial Informatics*, 2023, 19(1): 973-983.
- [3] YUAN Y C, ISLAM M S, YUAN Y, et al. EcRD: edge-cloud computing framework for smart road damage detection and warning[J]. *IEEE Internet of Things Journal*, 2021, 8(16): 12734-12747.
- [4] LIM R S, LA H M, SHENG W H, et al. A robotic crack inspection and mapping system for bridge deck maintenance[J]. *IEEE Transactions on Automation Science and Engineering*, 2014, 11(2): 367-378.
- [5] LI P, WANG C, LI S M, et al. Research on crack detection method of airport runway based on twice-threshold segmentation[C]//2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control. Qinhuangdao, China: IEEE, 2015: 1716-1720.
- [6] ZHANG D J, LI Q Q, CHEN Y, et al. An efficient and reliable coarse-to-fine approach for asphalt pavement crack detection[J]. *Image and Vision Computing*, 2017, 57: 130-146.
- [7] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.
- [8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2024-07-15]. <https://arxiv.org/abs/1409.1556v6>.
- [9] BOCHKOVSKIY A, WANG C Y, LIAO H M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2024-07-15]. <https://arxiv.org/abs/2004.10934v1>.
- [10] WANG W, WU B, YANG S, et al. Road damage detection and classification with faster R-CNN[C]//2018 IEEE International Conference on Big Data (Big data). Seattle, USA: IEEE, 2018: 5220-5223.
- [11] SU B, ZHANG H, WU Z, et al. FSRDD: An efficient few-shot detector for rare city road damage detection[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(12): 24379-24388.
- [12] YIN T, ZHANG W, KOU J, et al. Promoting automatic detection of road damage: a high-resolution dataset, a new approach, and a new evaluation criterion[J]. *IEEE Transactions on Automation Science and Engineering*, 2025, 22: 2472-2484.
- [13] GONAL A, BALIGA B C. Road damage detection and classification using YOLOv5 on an Android device[J]. *International Journal for Research in Applied Science & Engineering T*, 2022, 10(8): 333-339.
- [14] RASTEGARI M, ORDONEZ V, REDMON J, et al. XNOR-net: imagenet classification using binary convolutional neural networks[C]// Computer Vision – ECCV 2016. Amsterdam, Netherlands: Springer International Publishing, 2016: 525-542.
- [15] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient convnets[C]//International Conference on Learning Representations. Toulon, France: ICLR, 2017: 1-13.
- [16] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2024-07-15] <https://arxiv.org/abs/1503.02531v1>.
- [17] HE Y, KANG G L, DONG X Y, et al. Soft filter pruning for accelerating deep convolutional neural networks[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI, 2018: 2234-2240.
- [18] LUO J H, WU J X. An entropy-based pruning method for CNN compression[C]//Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017.
- [19] YE H, ZHANG B, CHEN T, et al. Performance-aware approximation of global channel pruning for multitask CNNs[J]. *IEEE Trans Pattern Anal Mach Intell*, 2023, 45(8): 10267-10284.
- [20] ANDREW H, MARK S, CHEN B, et al. Searching for MobileNetV3[C] //International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019: 1314-1324.
- [21] WANG Y L, FENG W Q, JIANG K, et al. Real-time

- damaged building region detection based on improved YOLOv5s and embedded system from UAV images[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023, 16: 4205-4217.
- [22] CHANG R, LI B Y, DANG J P, et al. Real-time intelligent detection system for illegal wearing of on-site power construction worker based on edge-YOLO and low-cost edge devices[J]. *Applied Sciences*, 2023, 13(14): 8287.
- [23] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2020, 42(8): 2011-2023.
- [24] DEEKSHA A, HIROYA M, KUMAR S G, et al. RDD2022: A multi-national image dataset for automatic road damage detection[J]. *Geoscience Data Journal*, 2024, 11(4): 846-862.
- [25] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]// *Computer Vision – ECCV 2018*. Munich, Germany: Springer International Publishing, 2018: 122-138.
- [26] TAN M X, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks[C]// *International Conference on Machine Learning*. Long Beach, USA: ICML, 2019: 6105-6114.