

基于改进 ShuffleNetV2 与模型压缩的椰子幼苗病害识别方法研究

廖碧蓉 李睿 李雪宁 王家豪 黄闽英*

西南民族大学计算机与人工智能学院

DOI:10.32629/as.v9i5.3970

[摘要] 椰子幼苗易受叶片病害影响。随着深度学习发展,卷积神经网络已广泛应用于植物病害识别。然而,现有方法仍存在多尺度特征提取不足及模型复杂度较高的问题。为此,本文以椰子幼苗病害为识别对象,提出一种基于改进ShuffleNetV2的椰子幼苗病害识别方法,通过引入多尺度特征提取与融合以及注意力增强,有效提升模型对不同尺度病斑与关键区域识别能力,最后基于L1范数结构化剪枝的模型压缩方法,实现精度下降可控与模型规模的显著压缩。本文提出的方法与当前具有代表性的8个基线方法对比,在各项评价指标上均取得了不同程度的提升,其中准确率、精确度、召回率和F1值的最大提升分别达到8.75%、8.21%、8.01%和8.09%,FLOPs 和参数量分别最高降低约80.00%和90.43%,最后压缩后的模型在各对比方法中取得了最优的综合表现,其准确率达到98.05%,仅较原始模型下降0.15%。同时模型参数量降低至1.10M,FLOPs降低至0.70G,压缩率和加速率分别达到 $2.44\times$ 和 $2.43\times$,说明所提出方法能够在模型压与性能保持之间取得更好的平衡。

[关键词] 椰子幼苗病害; 轻量化卷积神经网络; 多尺度特征; 注意力机制

中图分类号: S667.4 文献标识码: A

Research on Coconut Seedling Disease Identification Method Based on Improved ShuffleNetV2

Birong Liao Rui Li Xuening Li Jiahao Wang Minying Huang*

The College of Computer and Artificial Intelligence of Southwest Minzu University

[Abstract] Coconut seedlings are susceptible to leaf diseases. With the development of deep learning, convolutional neural networks have been widely applied to plant disease recognition. However, existing methods still suffer from insufficient multi-scale feature extraction and high model complexity. To address these issues, taking coconut seedling disease as the recognition target, this paper proposes a disease recognition method based on an improved ShuffleNetV2. By introducing a multi-scale feature extraction and fusion along with attention enhancement, the method effectively improves the recognition capability for disease spots of different scales and key regions. Finally, a model compression method based on L1-norm structured pruning is employed to achieve a significant reduction in model size while controlling the accuracy drop. Compared with eight representative baseline methods, the proposed method achieves varying degrees of improvement in various evaluation metrics, with maximum increases in accuracy, precision, recall, and F1-score reaching 8.75%, 8.21%, 8.01%, and 8.09%, respectively. The FLOPs and number of parameters are reduced by up to approximately 80.00% and 90.43%, respectively. The final compressed model achieves the best overall performance among all compared methods, with an accuracy of 98.05%, only 0.15% lower than that of the original model. Meanwhile, the number of parameters is reduced to 1.10M, FLOPs to 0.70G, with compression and acceleration ratios of $2.44\times$ and $2.43\times$, respectively, indicating that the proposed method achieves a better balance between model compression and performance preservation.

[Key words] coconut seedling diseases; lightweight convolutional neural network; multi-scale features; attention mechanism

前言

椰子作为亚热带重要经济作物,其苗木质量是产业可持续发展的关键。幼苗在高温高湿下易发生真菌性病害,严重时可导致死亡,影响产业效益^[1]。因此,开展病害精准识别具有重要意义。深度学习提取图像颜色、纹理特征,病害识别的精度和效率优势显著。然而,现有模型复杂度高、参数量大,难以满足边缘设备的实时与资源受限需求。现有方法聚焦于卷积神经网络(CNN),但传统CNN模型复杂度高,研究者提出了轻量级卷积神经网络(LWCNN)。LWCNN通过优化结构,在降低参数量与计算量同时,能够保持较高识别精度,在移动端实时检测中展现出良好应用潜力^{[2][3]}。尽管如此,受限于轻量化结构的参数约束以及多尺度特征利用不足,且复杂背景条件下难以提取充分病害特征。因此,在保证识别精度前提下提升多尺度特征表达能力并降低计算复杂度,成为亟需解决的问题。受卷积神经网络ShuffleNetV2方法成功的启发^[4],本文以椰子幼苗为研究对象,提出基于改进ShuffleNetV2与模型压缩的椰子幼苗病害识别方法,整体流程由三个阶段构成:首先,对椰子幼苗叶片病害图片进行规范化处理;其次,基于改进ShuffleNetV2构建识别模型MFEF-ShuffleNetV2,通过引入多尺度特征融合与注意力机制,实现对不同尺度病斑及关键区域特征的高效建模;最后,结合基于L1范数结构化剪枝的模型压缩对模型进行压缩与优化,在保持识别精度的同时显著降低模型复杂度。

1 相关工作

1.1 卷积神经网络的病害识别。近年来,随着深度学习的发展,卷积神经网络(CNN)已广泛用于农作物病害识别任务。但这类方法通常伴随着较大的参数规模与计算开销,从而限制了模型在移动设备和边缘设备上的部署。为缓解这一问题,研究者开始关注轻量化卷积神经网络结构的设计。MobileNet系列模型通过引入深度可分离卷积显著降低了计算复杂度^[5];在此基础上,MobileNetV2通过倒残差结构与线性瓶颈进一步提升了特征表达能力^[2]; MobileNetV3结合神经架构搜索与注意力机制,实现效率与性能的进一步优化^[6]。此外,EfficientNet通过复合缩放策略,在控制模型规模的同时获得了较高识别精度^[7]。这些方法为移动端实时病害识别提供重要技术支撑。由于轻量化网络在复杂环境下多尺度特征建模能力有限,相关研究开始引入注意力机制以增强关键区域特征表达能力。Hou等人提出的Coordinate Attention(CA)机制在通道注意力中融入空间位置信息,使模型能够同时关注重要特征和空间位置^[8];此外, Large Selective Kernel Network通过动态调整卷积核大小,实现对不同尺度特征的自适应建模,从而进一步提升识别性能^[9]。然而,现有轻量化模型在多尺度特征表达与复杂背景适应方面仍存在不足。不同于以往方法,本文基于改进ShuffleNetV2引入多尺度特征融合与注意力机制,实现对关键病害特征的高效建模与鲁棒识别。

1.2 模型压缩与结构化剪枝。随着卷积神经网络在图像识别中的广泛应用,模型计算与存储开销显著提升。为实现模型轻量

化,研究者提出多种压缩方法,包括量化、知识蒸馏、低秩分解和剪枝等方法^[10]。其中,剪枝具有显著优势。典型如基于L1范数的权重评估方法^[11]。然而,范数方法往往忽略滤波器间的冗余关系。为此,研究者从相似性角度进行建模,如基于几何中值的FPGM方法^[12]以及结合特征图相似性与信息熵的判别策略^{[13][14]}。此外,基于特征响应与信息量的方法也逐渐发展,包括基于特征图秩排序^[15]的剪枝方法。但大多仍依赖当前层的静态信息进行判断,缺乏对网络层与层之间信息传播关系的建模。忽略这种网络层与层依赖关系,容易导致关键通道被误剪,从而影响模型性能。在完成重要性评估之后,剪枝率合理分配也至关重要。传统方法采用固定比例或人工设定逐层剪枝率^[12],难以建模不同层之间的差异;基于全局阈值的统一筛选方法虽简化,但易导致剪枝不均^[16]。基于以上工作,本文基于L1范数结构化剪枝压缩方式,通过引入跨层通道传播分析与联合评估并实现通道合适的剪枝率分配,对于提升模型压缩效果具有重要意义。

2 基于MFEF-ShuffleNetV2模型与结构化剪枝病害识别

本文提出基于改进ShuffleNetV2与结构化剪枝的椰子幼苗病害轻量化识别方法,如图1所示,具体过程如下:(1)将数据集进行统一预处理,为后续病害识别MFEF-ShuffleNetV2模型训练提供规范化数据支撑。(2)将处理好的数据用于MFEF-ShuffleNetV2模型进行病害分类,并以该模型作为基于L1范数结构化剪枝的模型压缩的预训练模型。(3)将预训练的MFEF-ShuffleNetV2模型通过基于L1范数结构化剪枝的模型压缩,对模型中的冗余通道进行结构化压缩。

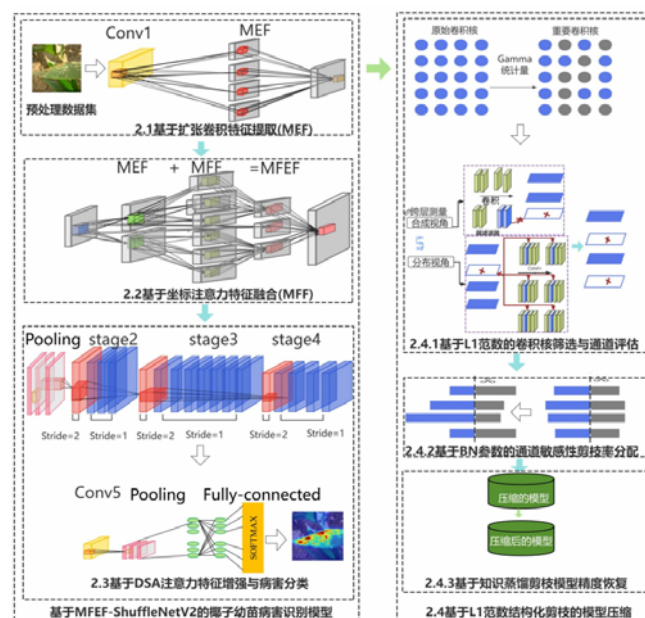


图1 基于MFEF-ShuffleNetV2模型与结构化剪枝病害识别方法

2.1 基于扩张卷积特征提取。现有轻量化网络虽在移动端部署具有优势,但受限于参数规模和计算能力,致其特征表达能力往往有限。为此,本文在MFEF-ShuffleNetV2网络的Conv1之后,随后将特征输入多尺度特征提取与融合模块(multi-scale

feature extraction and fusion module, MFEF) 中的多尺度特征提取模块 (multi-scale feature extraction and fusion module, MFE)。该模块以 ShuffleNetV2 的高效特征表达为基础^[4], 融合了多分支感受野建模^[17]与空洞卷积扩展机制^[18], 旨在不显著增加计算开销的前提下, 增强模型对细粒度纹理与大尺度语义结构的协同感知能力。预处理后的椰子幼苗叶片 RGB 图像经 Conv1 层及 BN 与 ReLU6 激活, 输出浅层特征图。随后进入 MFEF 模块的 MFE, 以四分支并行结构构建多尺度感受野, 捕获病斑局部特征与全局特征的过程, 经过四个分支的并行处理, MFE 模块最终输出四组尺寸一致的多尺度特征图并进行作为下一阶段 MFEF 模块中的多尺度特征融合模块 (multi-scale feature fusion, MFF) 的输入。

2.2 基于坐标注意力特征融合。为了充分利用不同尺度信息的优势并增强模型的表征能力。本研究采用特征融合, 其核心思想借鉴了 LSKNet^[9] 中的选择技术。该技术能够使每个神经元根据输入特征的内容自适应地调整感受野大小, 从而更有效地捕获多尺度信息。虽然 LSKNet 中的空间选择机制对空间信息进行全局编码, 但它将全局空间信息压缩到单个通道描述符中, 难以保留视觉任务中捕获空间结构所需的定位信息。因此, 本研究参考坐标注意力机制^[8], 并提出一种替代的空间选择机制, 其加权融合过程可统一表示为:

$$F_{out} = \sigma \left(\mathcal{A}_{CA} \left(\sum_{i=1}^N \frac{\alpha_i \cdot F_i}{\sum_{j=1}^N \alpha_j + \epsilon} \right) \right) \odot F \quad 2-1$$

其中, F_i 表示第 i 的多尺度特征, α_i 为可学习融合权重, ϵ 为稳定项, $\mathcal{A}_{CA}(\cdot)$ 表示坐标注意力编码过程, $\sigma(\cdot)$ 为 Sigmoid 激活函数, $\odot$ 表示逐元素加权操作, F 为融合后的特征表示。从整体流程来看, 首先对多尺度特征 F_i 进行归一化加权融合; 随后, 将融合特征输入坐标注意力模块, 通过沿水平方向与垂直方向的特征编码; 在此基础上, 经 Sigmoid 函数生成空间选择掩码, 并对融合特征进行逐元素重标定, 从而突出关键病害区域并抑制背景冗余信息; 最终得到兼具多尺度表达能力与空间感知能力的输出特征 F_{out} , 为后续 stage2、stage3 以及 stage4 阶段的输入, 用于进一步的深层特征提取。

2.3 基于 DSA 注意力特征增强与病害分类。为保证计算效率的前提下, 同时实现高效信息保留与关键特征增强。在融合后的多尺度特征基础上, 本研究通过构建 ShuffleNetV2 的 stage2、stage3 以及 stage4 阶段的内部组成单元, 其中该内部的组成单元由下采样单元以及基础单元构成。将其应用于深层病害特征提取并进行病害分类。该过程分为三个步骤: (1) 下采样单元构建 (Stride=2); (2) 基础单元构建 (Stride=1); (3) 病害分类。

(1) 下采样单元构建: 为提升模型的性能与计算效率, 本文在 ShuffleNetV2 模型的组成单元进行改进^[4]。在结构设计上, 借鉴 MobileNet^[5] 中深度可分离卷积的思想, 以进一步降低模型参数量与计算复杂度。具体而言, 将原有的 3×3 深度可分离卷积替换为 5×5 深度可分离卷积; 同时, 移除一层 1×1 逐点卷积层。在此

基础上, 为进一步增强特征表达能力并提升模型对关键病害区域的识别能力, 本文在分支特征拼接后引入双分支统计通道注意力机制 (Dual-Branch Shuffle Attention, DSA), 用于替代传统的通道混合操作。其中 DSA 结合分组卷积和通道混合操作取代传统通道混合操作, 并应用于步长为 2 的基本单元中。其整体计算过程可统一表示为:

$$\mu_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_c(i, j) \quad 2-2$$

$$\sigma_c = \sqrt{\frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W (X_c(i, j) - \mu_c)^2} \quad 2-3$$

$$A = \sigma \left(\text{Conv1D}_2 \left(\text{Conv1D}_1 (\text{Concat}(\mu, \sigma)) \right) \right) \quad 2-4$$

在上述过程中, 完成特征分组与拼接操作后, 得到输入特征矩阵 $X \in \mathbb{R}^{C \times H \times W}$, 其中每个通道 $X_c \in \mathbb{R}^{H \times W}$ 表示一个独立的特征响应单元。首先, 输入特征 X 通过平均池化得到通道统计量 μ_c ; 随后进一步计算标准差 σ_c 。接着, 将通道均值特征 μ 与标准差特征 σ 进行拼接得到融合特征 Z , 并通过两层一维卷积 $\text{Conv1D}_1(\cdot)$ 与 $\text{Conv1D}_2(\cdot)$ 实现跨通道信息交互, 经 Sigmoid 函数映射生成通道注意力权重 A 。在此基础上, 将注意力权重 A 作用于原始特征 X , 通过逐元素加权操作完成特征重标定, 从而增强关键通道响应并抑制冗余信息; 最后, 通过通道重排操作实现不同分组之间的信息交互, 得到最终增强后的深层特征, 以供基础单元构建的输入。

(2) 基础单元构建: 在完成步长为 2 的下采样单元构建后, 为进一步提升特征表达能力与信息利用效率。在保留通道分裂与通道重排机制的基础上, 引入 FasterNet Block 替代传统卷积分支, 其中部分卷积仅作用于部分通道, 其余通道采用恒等映射以减少计算冗余并保留特征信息^[19], 同时结合逐点卷积实现通道间信息融合。为增强通道判别能力, 在该结构中嵌入 ECA 注意力机制, 通过全局平均池化与一维卷积建模跨通道关系, 对关键特征进行自适应加权^[3]。整体上, 改进单元延续双分支结构, 一支用于保留原始信息, 另一支承担特征增强功能, 最后通过通道拼接与通道重排实现信息融合, 以增强特征的表达能力, 为后续的椰子幼苗叶片病害分类提供深层特征的输入。

(3) 椰子幼苗病害分类: 椰子幼苗叶片病害分类以 Stage2、Stage3 和 Stage4 输出的高层语义特征为输入, 通过卷积映射对特征进行通道升维; 随后利用全局平均池化对特征图进行空间压缩并聚合为一维全局特征向量; 在此基础上, 将得到的全局特征输入全连接分类器, 并通过 Softmax 函数映射为各类别的概率分布, 最终识别出对应的黄化病、叶枯病、炭疽病、灰斑病及健康的结果。

2.4 基于 L_1 范数结构化剪枝的模型压缩。

2.4.1 基于 L_1 范数的卷积核筛选与通道评估。为分析卷积核的重要性分布特征并进一步评估通道在网络中的实际贡献。首先, 在完成 MFEF-ShuffleNetV2 椰子幼苗病害识别预训练的模型后, 得到收敛的网络参数作为模型压缩阶段的输入。其次, 设第 l 层卷积核权重为 K^l , 其重要性度量统一表示为:

$$S_{n,m}^l = \|K_{n,m}^l\|_1 \cdot \mathbb{I}(\|K_{n,m}^l\|_1 \geq \mathbb{E}_l - c\sqrt{V_l}) \quad 2-5$$

其中, $\|K_{n,m}^l\|_1$ 表示卷积核的 L_1 范数, 用于刻画其对特征生成的贡献程度; $\mathbb{E}_l$ 与 V_l 分别为该层范数分布的均值与方差; c 为调节系数, $\mathbb{I}(\cdot)$ 为指示函数, $S_{n,m}^l$ 为卷积核重要性评分。在该过程中, 预训练的模型参数 K^1 通过 L_1 范数完成基础重要性建模, 并结合卷积核范数分布的统计特性(近似Gamma分布)^[11]构建自适应筛选阈值 $\mathbb{E}_l - c\sqrt{V_l}$, 以实现针对不同层卷积核重要性的差异化; 随后, 通过指示函数筛选高于阈值的关键卷积核, 从而完成从预训练的模型参数到卷积核重要性评分 $S_{n,m}^l$ 的映射, 为通道重要性评估提供可靠依据。

最后, 在完成卷积核重要性筛选后, 因此, 本文引入网络层之间的特征传播分析思想(Cross-layer Bidirectional Measurement, CLBM), 从生成视角与分布视角两个互补维度对通道重要性进行统一评估^[20]。通道重要性评估过程可表示为:

$$S_G^l(n) = \sum_{m=1}^{M_l} \|K_{n,m}^l\|_1 \cdot D_{n,m}^l \quad 2-6$$

$$S_R^l(n) = \sum_{i=1}^{N_{l+1}} \|K_{i,n}^{l+1}\|_1 \cdot D_{i,n}^{l+1} \quad 2-7$$

$$S_V^l(n) = S_G^l(n) \cdot S_R^l(n) \quad 2-8$$

其中, $S_G^l(n)$ 表示第 l 层第 n 个通道在当前层的特征生成重要性; $S_R^l(n)$ 表示该通道在下一层中的特征传播重要性; $S_V^l(n)$ 表示通道的综合重要性评分。 $K_{n,m}^l$ 与 $K_{i,n}^{l+1}$ 分别表示当前层与下一层卷积核权重。在该过程中, 预训练模型中的通道首先通过当前层卷积核权重 K^1 参与特征生成, 其重要性通过对卷积核范数加权求和得到 $S_G^l(n)$, 用于评估通道对当前层输出的贡献程度; 同时, 通过分析该通道在下一层卷积核 K^{l+1} 中的参与程度, 计算其在特征传播过程中的重要性 $S_R^l(n)$, 反映该通道对后续层特征构建的影响; 在此基础上, 将特征生成与特征传播两个维度的重要性进行乘性融合, 得到综合重要性评分 $S_V^l(n)$, 实现对通道在网络中的全局贡献评估, 从而为后续基于BN参数的通道敏感性剪枝率分配提供更加准确的依据。

2.4.2基于BN参数的通道敏感性剪枝率分配。在完成跨层通道重要性评估后, 需要根据各层通道对模型性能的影响程度自适应分配剪枝率。基于MFEF-ShuffleNetV2的预训练模型中Batch Normalization(BN)层参数^[16], 设第 l 层的剪枝率可统一表示为:

$$p_l = R \cdot \frac{\sum_i (|\gamma_i^l| + |\beta_i^l|)}{\sum_i \sum_n (|\gamma_i^l| + |\beta_i^l|)} \quad 2-9$$

在对模型性能的影响程度分配剪枝率中, 通过对预训练模型中的BN参数 γ_i^l 与 β_i^l 用于刻画各通道在特征分布调节中的作用强度, 当参数绝对值较小时, 对应通道对输出影响较弱, 反之则表示该通道对特征表达具有更强贡献; 在此基础上, 通过对各层BN参数进行统计聚合, 构建层级敏感性度量, 并将其作为剪枝

率分配依据, 实现不同层之间的差异化压缩; 进一步结合全局剪枝约束 R , 对各层剪枝比例进行分配。在确定各层剪枝率后, 依据通道重要性排序结果, 移除对应的低重要性通道, 以提供知识蒸馏模型精度恢复。

2.4.3基于知识蒸馏模型精度恢复。为了恢复剪枝模型的性能并提高其泛化能力, 在剪枝完成后引入知识蒸馏(Knowledge Distillation)^[10]对压缩后的模型进行微调训练。具体而言, 通过MFEF-ShuffleNetV2的预训练模型作为教师模型与剪枝后的学生模型输出经温度参数进行平滑处理得到软标签; 随后, 通过最小化两者之间的分布差异; 在此基础上, 将蒸馏约束与真实标签监督进行联合优化; 最终, 实现从剪枝模型到性能恢复模型的映射过程, 使压缩模型在降低参数数量与计算复杂度的同时, 尽可能维持原始模型的识别精度。

3 实验设置

3.1数据集准备和处理。本文主要研究椰子幼苗生长过程中较为频繁发生的黄化病、叶枯病、炭疽病、灰斑病。以下是椰子幼苗病害分类数据集的构建过程。该过程分为两个步骤: 数据集来源、数据预处理。

(1)数据集来源: Pacific Environment Data Portal是面向太平洋岛国及周边区域的环境与自然资源数据共享平台, 提供包含农业等领域在内的官方数据集。(2)数据预处理: 本研究对数据集进行了两项预处理。一是基于黄化病、叶枯病、炭疽病、灰斑病及健康叶片五类样本, 共计6939张图像, 其中各类别样本数量分别为1738、1764、1442、1109和886张。按照70%、15%、15%的比例将数据划分为训练集、验证集和测试集, 分别包含4855、1041和1043张图像。为提升模型泛化能力, 对训练集进行随机旋转、翻转、亮度调整等数据增强操作, 最终扩展至29130张训练样本; 二是将图像统一缩放至 224×224 像素, 并对RGB像素值归一化至 $[0, 1]$ 实现数据标准化, 从提高训练稳定性与收敛速度。

3.2评价指标。本文从分类性能与模型复杂度两个方面对所提出方法进行综合评估。

(1)分类性能指标: 用于评估模型在椰子幼苗病害识别任务中的判别能力, 主要包括准确率(Accuracy)、精确率(Precision)、召回率(Recall)以及F1-score。上述指标能够从整体预测正确性、样本识别精度以及类别平衡性等多个角度反映模型的分类性能。(2)模型复杂度指标: 用于衡量模型的轻量化程度与计算效率, 主要包括参数量(Parameters)与浮点运算量(FLOPs)。此外, 为更全面评估模型压缩效果, 引入压缩率(Compression Ratio, CR)与加速率(Acceleration Ratio, AR)作为辅助指标, 以量化剪枝前后模型在规模与计算开销上的变化。其计算方式如下:

$$CR = \frac{Params_{original}}{Params_{pruned}} \quad 3-1$$

$$AR = \frac{FLOPs_{original}}{FLOPs_{pruned}} \quad 3-2$$

其中, $Params_{original}$ 和 $FLOPs_{original}$ 分别表示原始模型的参数量和计算量, $Params_{pruned}$ 和 $FLOPs_{pruned}$ 表示剪枝后模型的参数量和计算量。

3.3 基线方法。本文从病害识别模型与模型剪枝方法两个层面构建基线, 分别评估模型的性能与压缩效果, 从而验证所提方法在识别精度与轻量化性能之间的综合优势。

(1) 病害识别模型基线方法: 本文对比八个模型按照模型结构复杂度与特征表征能力构建传统轻量卷积模型、结构优化型轻量模型、混合CNN-Transformer架构以及大规模Transformer上限模型作为基线方法进行系统对比: ①传统轻量级卷积神经网络: 选取经典轻量级卷积网络ShuffleNetV2^[4]和MobileNetV2^[2]作为对比模型。其中, ShuffleNetV2通过通道分离与重排机制提升信息流动效率; MobileNetV2采用倒残差结构与线性瓶颈设计。②结构优化型轻量CNN模型: 本文选取MobileNetV3^[6]、EfficientNetV2^[7]以及Mob-Res^[21]作为对比模型。其中, MobileNetV3在MobileNetV2基础上融合轻量级注意力机制与结构改进; EfficientNetV2基于复合缩放策略与训练优化; Mob-Res通过并行融合MobileNetV2与残差结构提升识别性能。③轻量级混合CNN模型: 本文选择MobilePlantViT^[22]和Tiny-LeViT^[23]代表性混合模型。此类模型在保持较低参数规模的同时, 引入自注意力机制以增强全局特征建模能力。④复杂模型Transformer: 本文选取Swin Transformer^[24]作为性能上限参考模型。该模型基于移动窗口自注意力机制。

(2) 模型剪枝基线方法: 为了验证本文提出的基于 L_1 范数结构化剪枝的模型压缩在压缩与性能保持方面的有效性, 本文选取了多种具有代表性的卷积神经网络剪枝方法作为对比基线方法, 包括FPGM^[12]、HRank^[15]、FSIM-E^[13]、EACP^[25]以及Graph-based Filter Pruning^[14]: ①FPGM方法(Filter Pruning via Geometric Median): 通过计算同一卷积层中滤波器的几何中值来识别具有冗余信息的滤波器, 并将距离几何中值较近的滤波器进行剪枝。②HRank方法(High-Rank Feature Map Pruning)方法: HRank通过计算每个滤波器输出特征图的平均秩, 并按照秩值大小进行排序并删除低秩特征图实现模型压缩。③FSIM-E方法: FSIM-E(Feature Similarity and Entropy Pruning)该方法首先利用特征相似性指标衡量不同特征图之间相似程度, 并通过二维信息熵评估特征图所包含的信息量。通过综合这两个指标识别冗余滤波器, 从而实现模型压缩。④EACP方法: 该方法基于滤波器特征相似性, 利用k-means++聚类对卷积核进行分组, 并删除冗余滤波器实现通道剪枝。⑤Graph-based Filter Pruning方法: 该方法利用图中节点的连接关系和边权信息判断特征图之间的相似程度, 并删除冗余特征图对应的滤波器, 从而实现一次性剪枝。

4 结果分析

本节将本文提出的方法从准确率、精确度、召回率、F1值、FLOPs、Param、CR以及AR等8个方面, 与基线方法在数据集上进行对比试验, 验证基于MFEF-ShuffleNetV2改进的椰子幼苗病害

识别的性能。

4.1 对比实验。表1给出了基于MFEF-ShuffleNetV2的椰子幼苗病害识别方法与8种主流对比模型的性能比较结果。由表中数据可以看出, 所提出模型在综合性能方面表现最优, 在Accuracy、Precision、Recall及F1-Score四项指标上分别达到98.20%、97.50%、97.43%和97.44%, 整体性能处于所有方法的领先水平。从整体对比结果来看, 该方法相较于轻量化模型(如MobileNetV2、MobileNetV3和ShuffleNetV2等), 在Accuracy上提升约2.6%至7.9%, Precision提升约2.4%至7.4%, Recall提升约2.2%至7.2%, F1-Score提升约2.3%至7.3%。同时, 相较于融合模型Tiny-LeViT与MobilePlantViT, 各项指标仍有约2.6%左右的提升, 说明该方法在兼顾轻量化的同时具备更强的判别能力。进一步与Mob-Res方法进行对比, 在Accuracy上提升0.35%, 但在Precision、Recall和F1-Score上分别略低0.20%、0.37%和0.31%。然而, Mob-Res的计算复杂度为3.12 GFLOPs, 参数量为3.51 M, 而本文模型仅为1.70 GFLOPs和2.68 M, 在计算量降低约45%、参数量减少约24%的情况下仍保持更高的准确率, 体现出更优的精度与复杂度权衡。

此外, 与高复杂度模型EfficientNetV2-S和Swin Transformer Tiny相比, 该方法在Accuracy上分别提升0.90%和0.65%, 同时计算量显著降低(由8.50 G和4.50 G降至1.70 G), 参数规模也大幅减少, 说明所提出方法在中等规模数据集条件下能够有效避免过拟合问题, 并具备更强的实际部署优势。

表1 MFEF-ShuffleNetV2与基线方法结果统计

模型	Acc (%)	Pre (%)	Re (%)	F1 (%)	FLOPs (G)	Param (M)
MobileNet V2 1.0×	90.3	90.1	90.2	90.15	0.3	3.5
MobileNet V3 Large	95.1	94.9	95	94.95	0.23	5.4
Tiny-LeViT	95.6	95.4	95.5	95.45	0.2	6.8
MobilePlantViT	95.62	95.55	95.65	95.6	1.12	3.69
ShuffleNetV2 ×2.0	95.63	95.5	95.7	95.6	0.59	5.4
EfficientNetV2-S	97.3	97.1	97.2	97.15	8.5	21
Swin Transformer Tiny	97.55	97.4	97.5	97.45	4.5	28
Mob-Res	97.85	97.7	97.8	97.75	3.12	3.51
MFEF-shuffleNetV2	98.2	97.5	97.43	97.44	1.7	2.68

为进一步验证基于 L_1 范数结构化剪枝的模型压缩的有效性, 如表2所示, 该模型压缩在各对比方法中取得最优综合表现, 其准确率达到98.05%, 仅较原始模型下降0.15%。同时模型参数量降低至1.10M, FLOPs降低至0.70G, 压缩率和加速率最高, 分别达到2.44×和2.43×, 说明该步骤能够在模型压缩与性能保持之间取得更好平衡。(1)FPGM方法: 通过几何中值删除相似卷积核实现参数压缩, 但缺乏对通道判别能力, 容易误剪关键特征通道, 导致准确率下降0.62%。相比之下, 本文方法能够更精准地区分冗余与关键通道, 从而显著降低性能损失。(2)Graphical方法^[14]: 基于图结构关系进行剪枝, 导致关键特征信息丢失最为明显, 准确率下降达到0.65%。而本文方法在特征表达层面具有更强适应性。(3)HRank方法: 利用特征图秩评估通道重要性, 准确率下降达到0.40%。相比之下, 本文方法通过更全面的重要性建

模机制,实现了更优的精度保持。(4)EACP方法:通过自适应策略优化剪枝过程,准确率下降0.48%,其复杂特征分布下仍存在误判风险,限制了剪枝效果的进一步提升。(5)FSIM-E方法:引入特征相似性度量以提升通道筛选质量,在精度下降0.45%,其仍未充分考虑通道在网络整体结构中的贡献程度,因此在压缩与性能平衡方面仍存在提升空间。

表2 基于 L_1 范数结构化剪枝的模型压缩与基线方法对比

方法	Acc (%)	Acc Drop (%)	Params (M)	FLOPs (G)	CR (×)	AR (×)
MFEF-shuffleNetV2	98.2	-	2.68	1.7	-	-
FPGM	97.58	0.62	1.25	0.74	2.14	2.3
Graphical	97.55	0.65	1.22	0.73	2.2	2.33
HRank	97.8	0.4	1.15	0.71	2.33	2.39
EACP	97.72	0.48	1.18	0.73	2.27	2.33
FSIM-E	97.75	0.45	1.2	0.74	2.23	2.3
CLBM-MFEF-shuffleNetV2	98.05	0.15	1.1	0.7	2.44	2.43

4.2 消融实验。为全面验证本文提出方法中各关键模块的有效性,本文从模型结构改进与模型压缩策略两个方面开展消融实验。

(1)模型结构改进模块消融实验:为验证各改进模块对模型的实际贡献,本文通过不同融合模块有效性比较,如表3所示。实验分别引入ECA注意力机制、对组成单元进行改进并引入DSA双分支注意力模块(unit-Dsa)、多尺度特征提取(MFE)以及多特征融合(MFF)的MFEF模块。

表3 MFEF-ShuffleNetV2模型消融实验结果

实验名称	Acc	Pre	Re	F1	FLOPs	Param
Baseline	95.63	95.5	95.7	95.6	0.59	5.4
(ECA+unit-Dsa)	96.33	96.2	96.4	96.3	0.57	2.67
MFEF	97.7	97.4	97.3	97.37	0.88	5.43
Full	98.2	98	97.9	97.95	1.7	2.68

①完整模型在Accuracy、Precision、Recall和F1-score上分别达到98.20%、98.0%、97.9%和97.95%,相较基线模型分别提升2.57%、2.5%、2.2%和2.35%,说明多模块联合能够显著提升模型整体判别能力。②与仅引入注意力机制的模型相比,完整模型各指标进一步提升约1.5%至1.8%,从模型复杂度角度来看,引入改进模块后模型FLOPs从0.59G增加至1.70G,说明多尺度特征提取与特征融合结构会带来一定计算开销。但与此同时,模型参数量由5.40M降低至2.68M,表明本文通过轻量化结构设计减少了冗余参数。结果表明MFEF的MFE模块能提取不同尺度的病害特征;MFEF的MFF模块则通过特征融合机制整合不同病害特征。③与仅采用多尺度模块的模型相比,完整模型仍有小幅提升约1%至2%,说明注意力机制能够进一步强化关键特征并抑制背景干扰。

(2)模型压缩消融实验:本文在统一剪枝率60%及相同模型规模(1.10M/0.70G)约束下创建了四类消融实验。如表4所示,具体来说,本文在基准方法(L_1 -norm剪枝+固定剪枝率+无蒸馏)的基础上,依次添加:① L_1 范数重要卷积核筛选(ks);②网络层之间的特征传播分析思想(CLBM);③BN参数敏感度动态分配(BN);④轻量化知识蒸馏(KD)。

表4 基于 L_1 范数结构化剪枝的模型压缩消融实验结果

方法	ks	CLBM	BN	KD	准确率 (%)	精度损失 (%)	参数量 (M)	FLOPs (G)
基线	✗	✗	✗	✗	96.82	1.38	1.1	0.7
+KS	✓	✗	✗	✗	97.05	1.15	1.1	0.7
+KS+CL	✓	✓	✗	✗	97.18	1.02	1.1	0.7
+KS+CL+BN	✓	✓	✓	✗	97.3	0.9	1.1	0.7
Full	✓	✓	✓	✓	98.05	0.15	1.1	0.7

消融实验结果表明,各改进模块对模型性能均具有逐步提升作用:首先,引入基于 L_1 范数的重要卷积核筛选后,模型准确率由96.82%提升至97.05%,有效减少冗余卷积核干扰;在此基础上,结合跨层传播机制,准确率进一步提升至97.18%,实现对通道重要性的更全面评估;随后,引入基于BN参数敏感度的剪枝率分配,准确率提升至97.30%,通过差异化剪枝提升整体压缩合理性;最后,结合知识蒸馏进行微调,模型准确率显著恢复至98.05%,有效弥补剪枝带来的信息损失。

4.3 不同剪枝率分析。为了进一步评估不同剪枝比例对模型性能的影响,本文对MFEF-ShuffleNetV2模型在不同剪枝比例条件下进行了实验分析,如表5所示。

表5 剪枝率敏感性对比

Rate	Acc (%)	模型大小 (MB)	参数量 (M)	内存占用 (MB)	FLOPs (G)	推理时间 (ms)
0%	98.2	10.72	2.68	356	1.7	55
10%	98.19	9.68	2.42	322	1.54	50
20%	98.18	8.64	2.16	290	1.38	45
30%	98.16	7.6	1.9	260	1.22	40
40%	98.12	6.6	1.65	232	1.05	35
50%	98.09	5.52	1.38	200	0.87	28
60%	98.05	4.4	1.1	162	0.7	21
65%	97.68	3.84	0.96	142	0.61	18
70%	97.4	3.32	0.83	124	0.53	16
80%	96.5	2.4	0.6	98	0.38	13

随着剪枝比例逐步增加,模型性能与压缩效果呈现明显的阶段性变化:在低剪枝阶段0%至30%,模型准确率最大下降仅0.04%,而参数量、FLOPs及推理时间均显著降低,说明模型存在一定冗余结构;然而,当剪枝比例超过65%后,模型精度开始明显下降70%为准确率97.40%,80%降至96.50%,说明过度剪枝会破坏关键特征表达能力。当剪枝比例提升至40%到60%时,模型压缩效果进一步增强,在60%剪枝下参数量降至1.10M、FLOPs降至0.70G、推理时间缩短至21ms,同时准确率仅下降0.15%,表明该方法60%剪枝取得精度与效率之间最优平衡。

5 总结

本文提出基于改进ShuffleNetV2的椰子幼苗病害识别方法,针对复杂背景下多尺度特征提取不足问题,引入多尺度特征提取与融合模块及双分支注意力机制,在椰子幼苗病害数据集上取得98.20%识别准确率;为进一步降低模型复杂度,基于 L_1 范数结构化剪枝的模型压缩,其通过跨层双向评估与BN参数敏感性分析实现自适应剪枝率分配,并利用知识蒸馏恢复精度。实验结果表明,在60%剪枝率下,模型参数量降至1.10M、计算量降至0.70G,识别精度仍保持在98.05%,实现高精度与轻量化有效平衡。

[参考文献]

- [1]Gurr G M,Johnson A C,Ash G J,et al.Coconut lethal yellowing diseases: a phytoplasma threat to palms of global economic and social significance[J].Frontiers in plant science,2016,7:1521.
- [2]Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510–4520.
- [3]Wang Q,Wu B,Zhu P,et al.ECA-Net:Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.2020:11534–11542.
- [4]Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018:116–131.
- [5]Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [J].arXiv preprint arXiv:1704.04861,2017.
- [6]Howard A,Sandler M,Chu G,et al.Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF international conference on computer vision.2019:1314–1324.
- [7]Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning.PMLR,2019:6105–6114.
- [8]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.
- [9]Li Y,Hou Q,Zheng Z,et al. Large selective kernel network for remote sensing object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 16794–16805.
- [10]Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.
- [11]Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient convnets[J]. arXiv preprint arXiv:1608.08710,2016.
- [12]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.
- [13]Liu Y,Fan K, Wu D, et al. Filter pruning by quantifying feature similarity and entropy of feature maps[J].Neurocomputing,2023,544:126297.
- [14]Li J,Shao H,Zhai S,et al.A graphical approach for filter pruning by exploring the similarity relation between feature maps[J].Pattern Recognition Letters,2023,166:69–75.
- [15]Lin M, Ji R, Wang Y, et al. Hrank: Filter pruning using high-rank feature map[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1529–1538.
- [16]Liu Z,Li J,Shen Z,et al.Learning efficient convolutional networks through network slimming[C]//Proceedings of the IEEE international conference on computer vision.2017:2736–2744.
- [17]Liu S, Huang D. Receptive field block net for accurate and fast object detection[C]//Proceedings of the European conference on computer vision(ECCV).2018:385–400.
- [18]Yu F,Koltun V.Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.
- [19]Chen J,Kao S,He H,et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023:12021–12031.
- [20]Lian S,Cai J, Zhao Y, et al. A cross-layer bidirectional measurement pruning for model compression guided by feature propagation[J].Information Sciences,2026:123216.
- [21]Pal C,Karmakar S, Mukherjee I, et al. A lightweight and explainable CNN model for empowering plant disease diagnosis [J].Scientific Reports,2025,15(1):30720.
- [22]Tonmoy M R, Hossain M M,Dey N,et al.MobilePlantViT:a Mobile-friendly hybrid ViT for generalized plant disease image classification[J].arXiv preprint arXiv:2503.16628,2025.
- [23]Thai H T,Le K H,Nguyen N L T.Towards sustainable agriculture:A lightweight hybrid model and cloud-based collection of datasets for efficient leaf disease detection[J].Future Generation Computer Systems,2023,148:488–500.
- [24]Liu Z,Lin Y,Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision.2021: 10012–10022.
- [25]Liu Y, Wu D, Zhou W, et al. EACP: An effective automatic channel pruning for neural networks[J]. Neurocomputing, 2023, 526:131–142.

作者简介:

廖碧蓉(1995--),女,汉族,四川成都人,在读研究生,研究方向或从事工作: 农业信息化。

*通讯作者:

黄闽英(1975--),女,汉族,江西南昌人,副教授,研究方向: 农业信息化。