

赵法川, 徐晓辉, 宋涛, 等. 融合多头注意力的轻量级作物病虫害识别 [J]. 华南农业大学学报, 2023, 44(6): 986-994.
ZHAO Fachuan, XU Xiaohui, SONG Tao, et al. A lightweight crop pest identification method based on multi-head attention[J]. Journal of South China Agricultural University, 2023, 44(6): 986-994.

融合多头注意力的轻量级作物病虫害识别

赵法川[✉], 徐晓辉[✉], 宋涛, 郝淼淼, 汪曙, 朱伟龙
(河北工业大学 电子信息工程学院, 天津 300401)

摘要:【目的】解决当前病虫害识别方法参数多、计算量大、难以在边缘嵌入式设备部署的问题, 实现农作物病虫害精准识别, 提高农作物产量和品质。【方法】提出一种融合多头注意力的轻量级卷积网络 (Multi-head attention to convolutional neural network, M2CNet)。M2CNet 采用层级金字塔结构, 首先, 结合深度可分离残差和循环全连接残差构建局部捕获块, 用来捕捉短距离信息; 其次, 结合全局子采样注意力和轻量级前馈网络构建轻量级全局捕获块, 用来捕捉长距离信息。提出 M2CNet-S/B/L 3 个变体以满足不同的边缘部署需求。【结果】M2CNet-S/B/L 参数量分别为 1.8M、3.5M 和 5.8M, 计算量 (Floating point operations, FLOPs) 分别为 0.23G、0.39G 和 0.60G。M2CNet-S/B/L 对 PlantVillage 病害数据集取得了大于 99.7% 的 Top5 准确率和大于 95.9% 的 Top1 准确率, 对 IP102 虫害数据集取得了大于 88.4% 的 Top5 准确率和大于 67.0% 的 Top1 准确率, 且比同级别的模型表现优异。【结论】该方法能够对作物病虫害进行有效识别, 且可为边缘侧工程部署提供有益参考。

关键词: 病虫害识别; 轻量级; 多头注意力; 残差学习; 深度可分离

中图分类号: S435

文献标志码: A

文章编号: 1001-411X(2023)06-0986-09

A lightweight crop pest identification method based on multi-head attention

ZHAO Fachuan[✉], XU Xiaohui[✉], SONG Tao, HAO Miaomiao, WANG Shu, ZHU Weilong
(College of Electronic Information Engineering, Hebei University of Technology, Tianjin 300401, China)

Abstract: 【Objective】To solve the problems that the current pest identification method has many parameters, a large amount of calculation and is difficult to deploy embedded devices at the edge, so as to realize accurate identification of crop pests and diseases, and improve crop yield and quality. 【Method】A lightweight convolutional neural network called multi-head attention to convolutional neural network (M2CNet) was proposed. M2CNet adopted hierarchical pyramid structure. Firstly, a local capture block was constructed by combining depth separable residual and cyclic fully connected residual to capture short-range information. Secondly, a lightweight global capture block was constructed by combining global subsampling attention and lightweight feedforward network to capture long-distance information. Three variants, namely M2CNet-S, M2CNet-B, and M2CNet-L, were proposed by M2CNet to meet different edge deployment requirements. 【Result】M2CNet-S/B/L had parameter sizes of 1.8M, 3.5M and 5.8M, and floating point operations of 0.23G, 0.39G, and 0.60G, respectively. M2CNet-S/B/L achieved top5 accuracy greater than 99.7% and top1 accuracy greater than 95.9% in PlantVillage disease dataset, and top5 accuracy greater than 88.4% and top1 accuracy

收稿日期: 2022-08-31 网络首发时间: 2023-05-30 10:11:44

首发网址: <https://kns.cnki.net/kcms2/detail/44.1110.S.20230529.1406.002.html>

作者简介: 赵法川, 硕士研究生, 主要从事农业物联网研究, E-mail: 202031903037@stu.hebut.edu.cn; 通信作者: 徐晓辉, 研究员, 主要从事传感器及智能系统研究, E-mail: xxh@hebut.edu.cn

基金项目: 河北省重点研发计划 (20327201D)

greater than 67.0% in IP102 pest dataset, outperforming models of the same level in comparison. 【Conclusion】 Effective identification of crop diseases and pests can be achieved by this method, and it provides valuable references for edge engineering deployment.

Key words: Diseases and pests identification; Lightweight; Multi-head attention; Residual learning; Deeply separable

农作物病虫害种类多、密度大, 极易造成作物大量减产, 严重制约农业生产, 而快速高效地识别病虫害是防治的关键。随着智慧农业的兴起与发展, 利用深度学习技术对病虫害进行智能识别以辅助农业生产, 减少不必要的农药喷施, 对保护生态环境、提高农作物的品质, 有着十分重要的作用。

随着数据量的增长和算力的提升, 深度学习发展迅猛, 诞生出 CNN、Transformer 等特征提取器, 催生出一系列模型。比较经典的卷积网络如 VGGNets^[1]、ResNets^[2] 具有识别准确率高的优点, 但也存在参数量大、性能差、难以广泛应用于移动端的问题, 因此越来越多的学者将目光转向轻量级网络的研究。李静等^[3] 通过迁移学习对 GoogLeNet 的 Inception-v4 网络结构进行优化, 对玉米螟虫害识别任务达到 96.44% 的准确率。刘洋等^[4] 对轻量级网络 MobileNet 和 Inception V3 进行优化, 在 PlantVillage 数据集上分别达到 95.02% 和 95.62% 的识别准确率。陆健强等^[5] 提出一种基于 Mixup 算法和卷积神经网络的柑橘黄龙果实识别模型, 对柑橘黄龙病数据集的识别准确率达到 94.29%。邱文杰等^[6] 通过知识蒸馏得到压缩模型 Distilled-MobileNet, 该模型在 38 种常见病害中达到了 97.62% 的分类准确率, 且模型仅为 19.83M。轻量级卷积网络在作物病虫害识别中的应用研究已经颇有成效, 但其在模型参数和性能方面仍有继续提升的空间, 也许多头注意力机制将是一个突破点。

近两年, 在自然语言处理领域大火的 Transformer 也被成功应用到计算机视觉领域。Dosovitskiy 等^[7] 提出了直接应用于图像块序列的视觉 Transformer (Vision transformer, ViT), 在 ImageNet-1K 上取得了 88.55% 的准确率, 刷新了该榜单纪录。相较于 CNN(如 ResNet), ViT 依靠多头注意力机制捕获图像块之间的长距离依赖关系, 因此拥有更大的感受野, 能获取全局信息, 但长程的多头注意力也使得 ViT 很容易忽略图像的局部性质, 而刚好 CNN 能弥补这一点。相较于 ViT 网络, CNN 的卷积核大多尺寸较小, 具有局部特征提取能力, 且在现实的工业部署场景中, 执行 CNN 比大多数现有的 ViT 都要高效。由此可见, 将卷积和多头注意

力混合设计, 有效结合 CNN 和 ViT 的优点, 可进一步提升轻量级作物病虫害识别模型的性能。

为了将卷积与多头注意力有效结合, 设计出高效的作物病虫害识别方法, 本研究提出了一个全新的架构 M2CNet (Multi-head attention to convolutional neural network)。M2CNet 基于层级金字塔结构, 并引入深度可分离卷积和循环全连接层进行局部特征提取, 同时设计轻量级的全局特征捕捉块, 既提高了性能, 也节省了计算开销, 以期对病虫害精准识别提供新的思路, 为后续的边缘平台部署和作物病害检测系统的开发提供新的解决思路和方案。

1 材料与方法

1.1 数据集及处理

1.1.1 CIFAR100 数据集 CIFAR100 由 Krizhevsky 等^[8] 收集, 图片主要来自 Google 和各类搜索引擎。CIFAR100 数据集有 100 个类别, 每个类别有 600 张大小为 32 像素 × 32 像素的彩色图像, 其中 500 张作为训练集, 100 张作为测试集。这 100 类被分为 20 个超类, 每个图像都带有一个“精细”标签(它所属的类)和一个“粗略”标签(它所属的超类)。将 M2CNet 应用于 CIFAR100 数据集, 并与其他模型进行效果比较。

1.1.2 PlantVillage 数据集 PlantVillage 由 Hughes 等^[9] 创建, 在植物病理学专家的辅助下完成标注, 目的是帮助解决传染病导致的作物产量损失问题。该数据集包含 54 309 张图像, 涵盖南瓜白粉病、桃细菌性斑点病、樱桃白粉病、柑橘黄龙病、玉米枯叶病、玉米灰斑病、玉米锈病、番茄二斑叶螨病、番茄叶霉病、番茄斑枯病、番茄早疫病、番茄晚疫病、番茄细菌性斑点病、番茄花叶病、番茄轮斑病、番茄黄曲叶病、苹果疮痂病、苹果锈病、苹果黑腐病、苹果叶焦病、葡萄叶枯病、葡萄黑痘病、葡萄黑腐病、辣椒细菌性斑点病、马铃薯早疫病、马铃薯晚疫病共计 26 种作物疾病。试验中按数量 8:2 的比例划分训练集和测试集, PlantVillage 将用于检验 M2CNet 在作物病害识别任务中的表现。

1.1.3 IP102 数据集 IP102 是用于作物害虫识别

的野外构建的大规模数据集^[10], 共有 75 222 张图片, 涵盖了 102 种常见的害虫, 平均每种害虫 737 个样本, 这些图像呈现出自然的长尾分布。病虫害生命周期有不同阶段, 例如稻纵卷叶螟在幼虫时期呈现翠绿色的长条节状, 而在成虫时期呈现棕灰色的飞蛾形态, 与水稻二化螟类似, 因此 IP102 呈现出类间差异小和类内差异大的特点。试验中同样按数量 8:2 的比例来划分训练集和测试集, 各类具体害虫的训练集、测试集包含的图像数量汇总如表 1 所示。IP102 用于检验 M2CNet 在作物虫害识别任务中的表现。

表 1 IP102 数据集害虫分级分类体系
Table 1 Taxonomy of the IP102 dataset on different class levels

作物 Crop	害虫类别 Pest class	训练集 Training set	测试集 Test set
水稻 Rice	14	6 734	1 683
玉米 Corn	13	11 212	2 803
小麦 Wheat	9	2 734	684
甜菜 Sugarbeet	8	3 536	884
苜蓿 Alfalfa	13	8 312	2 078
葡萄 Grape	16	14 041	3 510
柑橘 Orange	19	5 818	1 455
芒果 Mango	10	7 790	1 948
总计 Total	102	60 177	15 045

1.2 构建病虫害识别网络

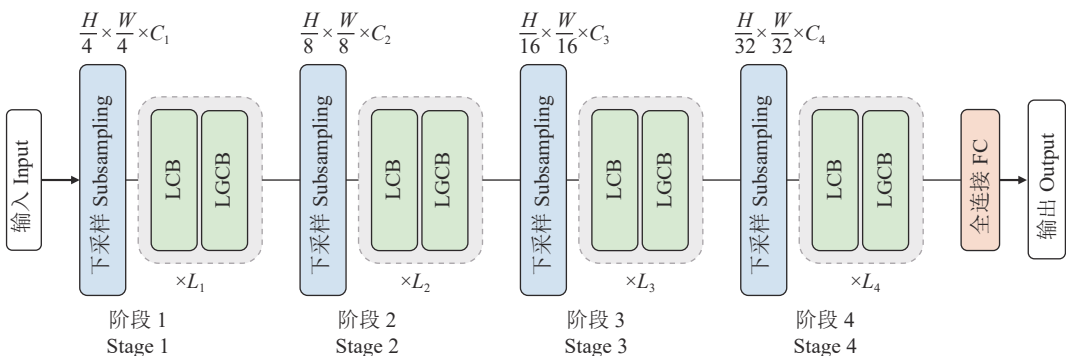
本文构建了一种识别作物病虫害的轻量模型——M2CNet, 该模型采用金字塔结构, 降低空间分辨率的同时能够在不同阶段扩展通道数。M2CNet 主要开发了 2 个重要组件, 首先构建了局部捕获块 (Local capture block, LCB), 该组件主要由深度可分离卷积和多层循环全连接构成, 用来捕捉

病虫害图片的短距离和细粒度信息; 其次构建了轻量级全局捕获块 (Lightweight global capture block, LGCB), 该组件由全局子采样注意力 (Global sub-sampling attention, GSA) 和轻量级前馈网络 (Lightweight feedforward network) 构成, 用来捕捉病虫害图片的长距离和高维信息。模型总体组成如图 1 所示, 下面将分别介绍局部捕获块和轻量级全局捕获块。

1.2.1 局部捕获块 局部捕捉块的结构如图 2 所示, 其中引入了残差学习^[11]的思想, 主要由 2 个连续的深度可分离卷积^[12]和 1 个多层循环全连接^[13]构成。深度可分离卷积由 1 个深度卷积和 1 个逐点卷积构成, 每层卷积后跟随一个批规范化^[14], 由于频繁地做非线性投影会有害于模型特征的信息传递^[15], 因此这里减少了激活层。深度可分离卷积先从空间维度获取局部信息, 再将获取的局部信息向通道维度传递, 最大程度地降低特征的损失; 多层循环全连接由 2 个伪核为 1×3 和 3×1 的循环全连接层构成, 其中也使用了残差学习以避免模型加深时出现的退化现象。多层循环全连接层通过阶梯状采样来增大其感受野以更好地集成上下文特征, 相比通道全连接有着一步操作就可以同时提取局部信息和融合通道信息的优势。

深度可分离卷积和多层循环全连接的感受野基本相当, 都可以关注局部信息, 但深度可分离卷积更侧重于空间维度, 多层循环全连接更侧重于通道维度。由于图片数据的纹理在空间维度表现更加明显, 因此在局部捕捉块中采取先空间后通道的思想, 深度可分离卷积在前, 多层循环全连接在后, 避免特征提取过程中图片纹理被过度压缩。

1.2.2 轻量级全局捕获块 轻量级全局捕获块由多个轻量结构组成, 旨在通过更少的参数来学习更



LCB: 局部捕获块; LGCB: 轻量级全局捕获块; H 和 W 分别代表输入图片的高度和宽度; C_i : 指用于阶段 i 的通道数; L_i : 阶段 i 的局部捕获块和轻量级全局捕获块数量

LCB: Local capture block; LGCB: Lightweight global capture block; H and W represent the height and width of the input image, respectively; C_i : Number of channels used for stage i ; L_i represents the number of local capture blocks and lightweight global capture blocks in stage i

图 1 M2CNet 网络总体组成

Fig. 1 Overall structure of the M2CNet network

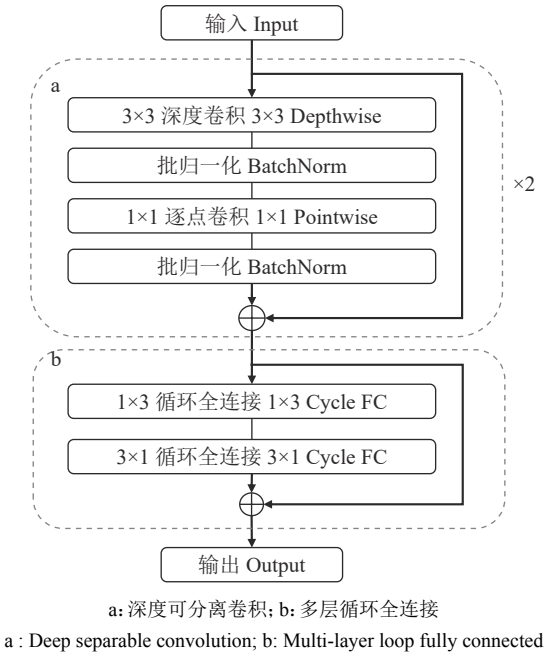


图 2 局部捕捉块结构图

Fig. 2 Structure diagram of a local snap block

鲁棒的表征。LGCB 最核心的部分是一种特殊的多头注意力: 全局子采样注意力^[16]。图 3 是标准多头注意力与全局子采样注意力的对比, 可以看到全局子采样注意力多出一个次采样 (Subsampling) 结构, 该结构把特征图分为多个不重叠的子窗口 ($s \times s$), 在子窗口上提取代表键 (K) 和值 (V), 但由于查询 (Q) 是全局的, 因此注意力仍可以恢复到全局, 这种做法显著减少了计算量。

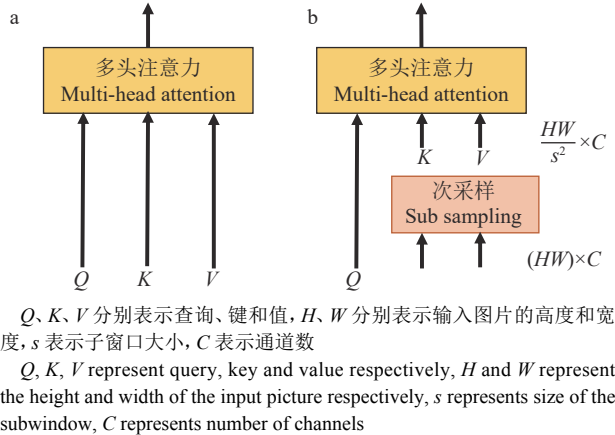


图 3 标准多头注意力 (a) 与全局子采样注意力 (b) 的对比
Fig. 3 Comparison of standard multi-head attention and global subsampling attention

轻量级全局捕捉块的整体结构如图 4 所示, LGCB 首先对输入特征图做条件位置编码^[17](Conditional position encoding, CPE), 将输入向量 $H \times W \times d_i$ 映射到高维空间, 然后在空间维度展平成向量 $HW \times d_i$, 过程中得到了输入特征图的位置信息。在全局子采样注意力阶段, 输入特征尺寸为

$HW \times d_i$, 次采样的输出尺寸为 $HW/s^2 \times d_i$, 其中 d_i 为通道维数, s 为子窗口的大小, 得到 $Q = HW \times d_i$, $K = V = HW/s^2 \times d_i/h$, h 为多头注意力头的数量, 将 QKV 共同送入多头注意力。最后经过轻量级前馈网络^[18]将输入从 d_i 降维到 d_i/r , 再从 d_i/r 升维到 d_i , 其中 r 为降维因子, 通常取 $r = 4$, 该操作用于提升模型容量。简单地, 轻量级全局捕捉块可以表述如下:

$$X = \text{CPE}[\text{Norm}(X_{\text{in}})], \tag{1}$$

$$Y = \text{GSA}[\text{Norm}(X)] + X, \tag{2}$$

$$X_{\text{out}} = \text{Lightweight FFN}[\text{Norm}(Y)] + Y, \tag{3}$$

式中, X_{in} 表示输入张量, Norm 是层归一化操作, CPE 是条件位置编码, GSA 是全局子采样注意力, Lightweight FFN 是轻量级前馈网络。所有这些操作都可以在标准深度学习平台通过常用和高度优化的操作来实现。

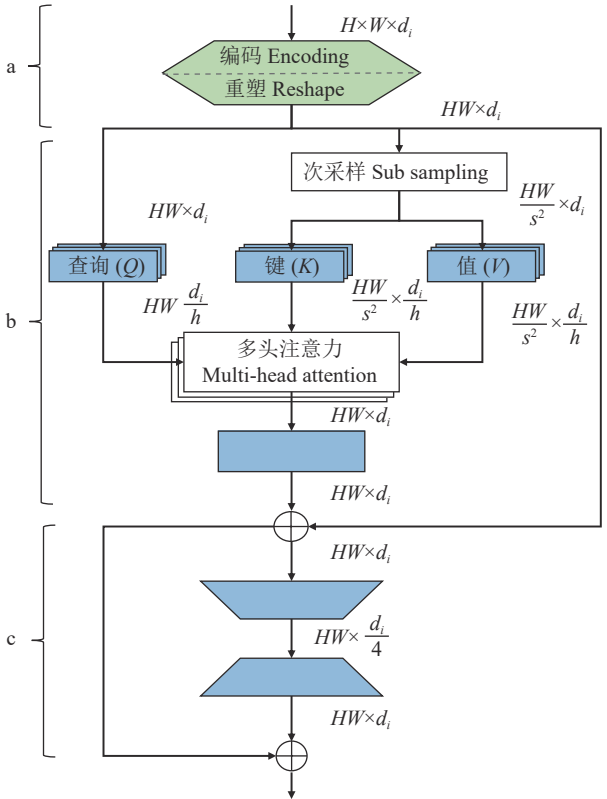


图 4 轻量级全局捕捉块

Fig. 4 Lightweight global capture block

1.2.3 M2CNet 模型架构 为满足不同的边缘部署需求, 本研究提出了 3 个典型的变体, 即 M2CNet-S/B/L。架构规范如表 2 所示, 对于归一化, 在局部

表 2 M2CNet-S/B/L 的网络架构¹⁾

Table 2 M2CNet-S/B/L network architecture

阶段 Stage	输出尺寸 Output size	层名称 Name of layer	M2CNet-S	M2CNet-B	M2CNet-L
1	56×56	Conv. 下采样	4×4, 36, stride 4	4×4, 48, stride 4	
	56×56	深度可分离卷积 多层循环全连接 全局子采样注意力 轻量级前馈网络	$\begin{bmatrix} 3\times 3, 1\times 1, 36 \\ 3\times 1, 1\times 3, 36 \\ H_1=1, s_1=4 \\ R_1=4 \end{bmatrix} \times 1$	$\begin{bmatrix} 3\times 3, 1\times 1, 48 \\ 3\times 1, 1\times 3, 48 \\ H_1=1, s_1=4 \\ R_1=4 \end{bmatrix} \times 1$	$\begin{bmatrix} 3\times 3, 1\times 1, 48 \\ 3\times 1, 1\times 3, 48 \\ H_1=1, s_1=4 \\ R_1=4 \end{bmatrix} \times 1$
2	28×28	Conv. 下采样	2×2, 72, stride 2	2×2, 96, stride 2	
	28×28	深度可分离卷积 多层循环全连接 全局子采样注意力 轻量级前馈网络	$\begin{bmatrix} 3\times 3, 1\times 1, 72 \\ 3\times 1, 1\times 3, 72 \\ H_1=2, s_1=2 \\ R_1=4 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times 3, 1\times 1, 96 \\ 3\times 1, 1\times 3, 96 \\ H_1=2, s_1=2 \\ R_1=4 \end{bmatrix} \times 1$	$\begin{bmatrix} 3\times 3, 1\times 1, 96 \\ 3\times 1, 1\times 3, 96 \\ H_1=2, s_1=2 \\ R_1=4 \end{bmatrix} \times 2$
3	14×14	Conv. 下采样	2×2, 144, stride 2	2×2, 192, stride 2	
	14×14	深度可分离卷积 多层循环全连接 全局子采样注意力 轻量级前馈网络	$\begin{bmatrix} 3\times 3, 1\times 1, 144 \\ 3\times 1, 1\times 3, 144 \\ H_1=4, s_1=2 \\ R_1=4 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times 3, 1\times 1, 192 \\ 3\times 1, 1\times 3, 192 \\ H_1=4, s_1=2 \\ R_1=4 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times 3, 1\times 1, 192 \\ 3\times 1, 1\times 3, 192 \\ H_1=4, s_1=2 \\ R_1=4 \end{bmatrix} \times 6$
4	7×7	Conv. 下采样	2×2, 288, stride 2	2×2, 384, stride 2	
	7×7	深度可分离卷积 多层循环全连接 全局子采样注意力 轻量级前馈网络	$\begin{bmatrix} 3\times 3, 1\times 1, 288 \\ 3\times 1, 1\times 3, 288 \\ H_1=8, s_1=1 \\ R_1=4 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times 3, 1\times 1, 384 \\ 3\times 1, 1\times 3, 384 \\ H_1=8, s_1=1 \\ R_1=4 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times 3, 1\times 1, 384 \\ 3\times 1, 1\times 3, 384 \\ H_1=8, s_1=1 \\ R_1=4 \end{bmatrix} \times 4$
输出 Output	1×1	全连接		100	
参数量(M) No. of parameters			1.83	3.52	5.76
计算量(G) Floating point operations			0.23	0.39	0.60

1)输入图像大小默认为224像素×224像素,Conv.代表卷积操作,stride表示卷积的步幅, H_i 和 S_i 是第*i*个全局子采样注意力的头数和次采样大小, R_i 是第*i*个轻量级前馈网络的特征尺寸缩放比

1) The input image size is 224×224 by default, Conv. stands for convolution operation, stride stands for convolution step, H_i and S_i are the number of heads and subsampling size of the i th global subsampling, and R_i is the scaling ratio of the feature size of the i th lightweight feedforward network

捕捉块中使用批归一化,在轻量级全局捕获块中使用层归一化,对于激活函数均使用 ReLU。

2 结果与分析

2.1 试验环境与评价指标

本研究在 Ubuntu 20.04 系统展开,该系统搭载 GeForce RTX 3 090 图形处理器并通过并行计算架构 CUDA 11.4 和 CUDNN 8.2.4 驱动,深度学习框架选择 PyTorch 1.10.1,编程语言为 Python 3.8.5。训练时 CIFAR100 和 IP102 的迭代次数设为 300,PlantVillage 的迭代次数设为 60,批次均为 64。学习率选择余弦衰减^[19]策略,PlantVillage 和 CIFAR-100 的初始学习率设为 0.000 5,IP102 的初始学习率设为 0.005,前 10 个迭代次数学习率均使用线性启动。优化器选择 Adamw^[20],并将权重衰减设置为 0.05,在训练中还使用了标签平滑^[21]和 Mixup^[22]数据增强来进一步探索模型性能。训练时

图像使用 224 像素×244 像素的随机裁剪,测试时使用 224 像素×244 像素的中心裁剪。

评价指标采用 Top1 准确率、Top5 准确率和损失值,Top1 准确率指预测概率排名第 1 的类别与实际结果相符的准确率,Top5 准确率是指预测概率排名前 5 的类别与实际结果相符的准确率。准确率 (Accuracy) 和损失值 (Loss) 的计算公式如下:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%, \tag{4}$$

$$Loss = -\sum_{i=1}^n p(x_i) \lg [q(x_i)], \tag{5}$$

式中,TP 为真正类,TN 为真负类,FP 假正类,FN 假负类; $p(x_i)$ 代表真实的标签, $q(x_i)$ 代表预测的概率。

2.2 模型性能对比

本研究将 M2CNet 应用于 CIFAR100,并与多种模型进行了比较,包括许多经典的计算量 (Floating point operations, FLOPs) 小于 1G 的轻量级卷积网络,例如 ShuffleNets^[23-24]、SqueezeNet^[25]、

MobileNetV2^[26]、MobileNetV3^[27]、MnasNet^[28]、EfficientNet^[29], 还包括 ViT 模型 MobileViT^[30] 和大型模型 VGG, M2CNet-S/B/L 的训练过程见图 5, 从

图 5 可以直观地看到随着 300 次迭代的收敛, M2CNet-S/B/L 在训练集和测试集的损失值逐渐降低, 直至趋于平稳。

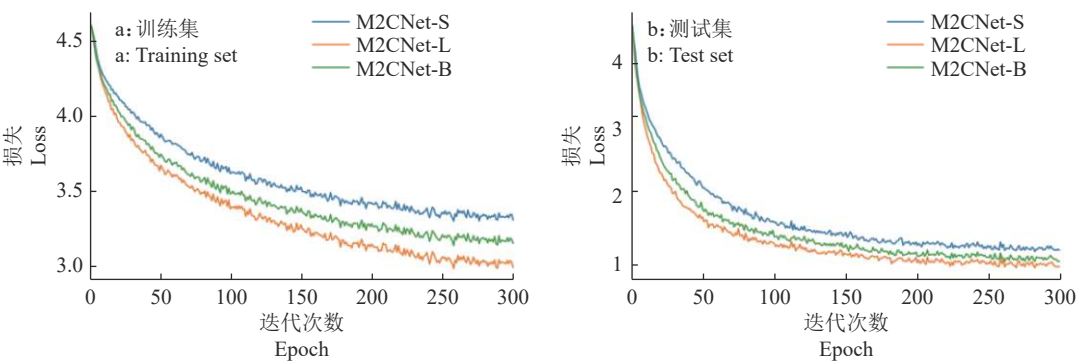


图 5 M2CNet-S/B/L 在 CIFAR100 数据集的训练过程
Fig. 5 M2CNet-S/B/L training process in the CIFAR100 dataset

表 3 是对比结果, 在参数量和计算量相似的情况下, M2CNet-S/B/L 占据一定优势, 且 M2CNet-L 参数量和准确率最优。与 ShuffleNet 系列相比, 本研究的 M2CNet-S/B 比 ShuffleNet-V2 1.5/2.0 分别在 Top1 准确率上实现了 4.53、2.53 个百分点的提升。与 MnasNet 系列相比, M2CNet-S/B/L 比 MnasNet 0.75/1.0/1.3 分别在 Top1 的准确率上实现了 1.89、2.62 和 1.75 个百分点的提升。与 MobileNet 系列相比, M2CNet-S/B/L 分别在 Top1 准确率上实现了 9.35、6.16 和 5.12 个百分点的提升。由此可见将多头注意力机制与卷积结合可以有效提升卷积模型的性能, 例如 M2CNet-S 与 MobileNet-V2、MobileNet-V3-Large 参数量和计算量相似, 但其识别精度却更优。在与 MobileViT 系列的对比中, M2CNet-S/B/L 同样在识别精度上展现出明显优势。本研究也将 M2CNet 与大型模型做对比, 可以看到 M2CNet-L 比 VGG 系列、ResNet 18 准确率更高, 而参数量仅为 ResNet 18 的一半, 是 VGG 系列的1/20。由此可见 M2CNet 可以在模型参数量和准确率之间保持平衡。

2.3 病虫害识别效果

为了更好地比较 M2CNet-S/B/L 的效果, 本研究针对每一种变体找到了在参数量和计算量上相似的对照, 即 M2CNet-S 对应 MobileViT-XS、MobileViT-XXS、MnasNet 0.75、MobileNet-V2; M2CNet-B 对应 MobileNet-V3-Large、EfficientNet B0、MnasNet 1.0; M2CNet-L 对应 EfficientNet B1、MobileViT-S、MnasNet 1.3。将以上网络分别在 PlantVillage 病害数据集和 IP102 虫害数据集上展开试验, 试验结果见图 6。

图 6a 是 PlantVillage 数据集识别结果, 可以看

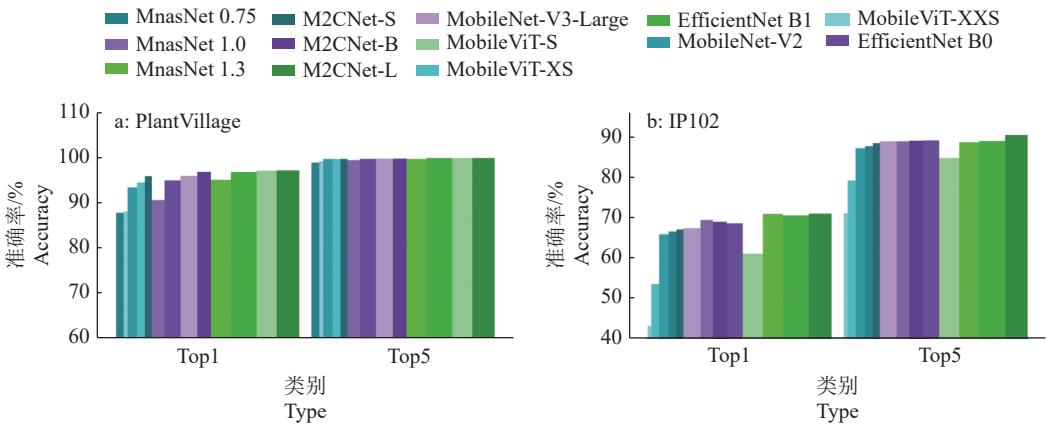
表 3 CIFAR100 数据集模型对比结果				
Table 3 Comparison results of CIFAR100 dataset model				
模型 Model	参数量 (M) No. of parameters	计算量 (G) Floating point operations	准确率/% Accuracy	
			Top5	Top1
ShuffleNet-V2 0.5	0.4	0.04	72.74	41.83
ShuffleNet-V2 1.0	1.4	0.15	86.21	59.65
ShuffleNet-V2 1.5	2.6	0.30	90.08	66.56
ShuffleNet-V2 2.0	5.6	0.56	93.06	72.79
SqueezeNet 1.0	0.8	0.75	78.48	49.68
SqueezeNet 1.1	0.8	0.30	78.12	50.14
MobileNet-V3-Small	1.6	0.06	87.90	61.74
MobileNet-V2	2.4	0.31	91.69	69.16
MobileNet-V3-Large	4.3	0.23	93.57	73.27
MnasNet 0.5	1.1	0.11	88.13	62.60
MnasNet 0.75	2.0	0.22	91.44	69.20
MnasNet 1.0	3.2	0.32	92.81	72.70
MnasNet 1.3	5.1	0.54	94.41	76.64
EfficientNet B0	4.1	0.40	94.63	76.00
EfficientNet B1	6.6	0.60	94.95	77.96
ResNet 18	11.2	1.80	94.66	76.85
VGG 11	129.2	7.60	94.25	75.82
VGG 13	129.4	11.30	94.38	76.46
VGG 16	134.7	15.50	94.63	78.19
VGG 19	140.0	19.60	95.25	78.19
MobileViT-XXS	1.0	0.33	84.98	55.96
MobileViT-XS	2.0	0.90	89.55	64.34
MobileViT-S	5.1	1.75	93.64	72.93
M2CNet-S	1.8	0.23	92.46	71.09
M2CNet-B	3.5	0.39	94.16	75.32
M2CNet-L	5.8	0.60	95.31	78.39

到在各组对照中 M2CNet-S/B/L 分别取得了 95.92%、96.82%、97.15% 的最大 Top1 识别准确率,在参数量相似的情况下取得了最优的结果。图 6b 是 IP102 数据集识别结果, M2CNet-S/L 依然延续了在 PlantVillage 上的表现,分别取得了 67.08%、71.0% 的最大 Top1 准确率和 88.49%、90.50% 的最大 Top5 准确率。在 M2CNet-B 对照中 MnasNet 1.0 取得了 69.46% 的最大 Top1 准确率,超出 M2CNet-B 0.47 个百分点,不过从整体来看, M2CNet 变体在作物病虫害识别任务中依然表现出色。M2CNet 变体能在对照试验中取得比其他轻量级网络更有竞争力的结果,分析原因在于融合多头

注意力的 M2CNet 不仅关注局部信息,也关注全局信息,因此能够灵活应对不同特征尺度的变化。

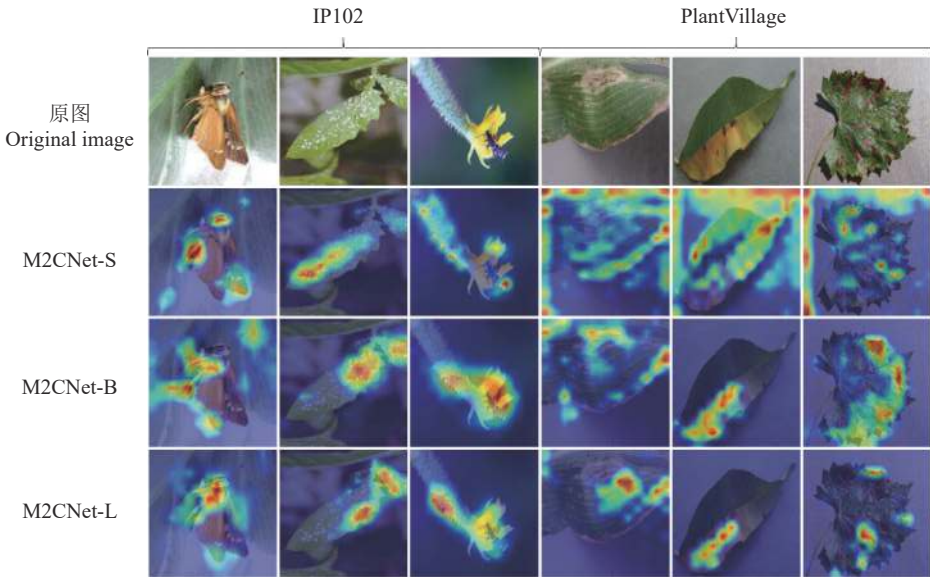
2.4 热力图可视化

为了进一步解释融合多头注意力后 M2CNet-S/B/L 关注的区域,这里使用 Grad-CAM^[31] 方法在病害和虫害的部分数据集上抽取特征图进行可视化,其可视化结果如图 7 所示。可以看到,由 M2CNet-S 到 M2CNet-B 再到 M2CNet-L,模型对于分类识别任务中更有判别性的特征区域给予了更高的关注,在一定程度上降低了背景特征的干扰,进而提升了模型识别精度。



柱状图的宽度与模型参数呈线性关系,参数量越大柱状图越宽;同一色系代表同一对照,同一色系中颜色最深的柱子对应 M2CNet 变体
The width of the bar chart is linearly related to the model parameters, the larger the number of parameters, the wider the bar chart; The same color system represents the same control, and the darkest column in the same color system corresponds to the M2CNet variant

图 6 病虫害数据集识别结果
Fig. 6 Identification results of pest data sets



红色高亮部分代表网络关注度高的区域,冷色发暗部分代表网络关注度低的区域
The highlighted areas in red represent areas with high network attention, while the dark areas in cool colors represent areas with low network attention

图 7 网络关注区域热力图
Fig. 7 Thermal map of the network focus area

3 结论

本研究为设计出轻量级作物病虫害识别方法, 将多头注意力机制捕捉长距离依赖关系的能力与卷积神经网络的局部特征提取能力相结合, 设计出满足不同边缘部署需求的 3 个变体: M2CNet-S/B/L。

为了验证 M2CNet-S/B/L 的特征提取能力, 在 CIFAR100 数据集上将其与其他轻量级网络展开对比, 在参数量和计算量相似的情况下, M2CNet 3 个变体均表现出良好的性能。在与经典的轻量级网络 MobileNet 系列比较中, M2CNet-S/B/L 在 Top1 准确率上分别实现了 9.35、6.16 和 5.12 个百分点的增益。

在作物病虫害数据集的试验中, M2CNet-S/B/L 在 PlantVillage 病害数据集上取得了大于 99.70% 的 Top5 准确率和大于 95.92% 的 Top1 准确率, 在 IP102 虫害数据集上取得了大于 88.4% 的 Top5 准确率和大于 67.0% 的 Top1 准确率, 且在同级别网络的对比中均占有优势, 证明 M2CNet 能够胜任作物病虫害识别任务。

M2CNet 网络有着参数量少的优点, 以 M2CNet-S 为例, 其参数内存仅占用 1.8M, 对硬件性能 (FLOPs) 要求仅为 0.23G, 这极大降低了对硬件平台的要求, 有利于后续的边缘平台部署和作物病害检测系统的开发和普及。

参考文献:

[1] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. arXiv: 1409.1556. <https://arxiv.org/abs/1409.1556>.

[2] HE K, ZHANG X, REN S, 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.

[3] 李静, 陈桂芬, 安宇. 基于优化卷积神经网络的玉米螟虫害图像识别[J]. 华南农业大学学报, 2020, 41(3): 110-116.

[4] 刘洋, 冯全, 王书志. 基于轻量级 CNN 的植物病害识别方法及移动端应用[J]. 农业工程学报, 2019, 35(17): 194-204.

[5] 陆健强, 林佳翰, 黄仲强, 等. 基于 Mixup 算法和卷积神经网络的柑橘黄龙病果实识别研究[J]. 华南农业大学学报, 2021, 42(3): 94-101.

[6] 邱文杰, 叶进, 胡亮青, 等. 面向植物病害识别的卷积神经网络精简结构 Distilled-MobileNet 模型[J]. 智慧农业 (中英文), 2021(1): 109-117.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. arXiv: 2010.11929. <https://arxiv.org/abs/2010.11929>.

<https://arxiv.org/abs/2010.11929>.

[8] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R/OL]. Technical report: University of Toronto, <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.

[9] HUGHES D P, SALATHE M. An open access repository of images on plant health to enable the development of mobile disease diagnostics[EB/OL]. arXiv: 1511.08060. <https://arxiv.org/abs/1511.08060>.

[10] WU X P, ZHAN C, LAI Y K, et al. IP102: A large-scale benchmark dataset for insect pest recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2020: 8779-8788.

[11] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[12] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. arXiv: 1704.04861. <https://arxiv.org/abs/1704.04861>.

[13] CHEN S, XIE E, GE C, et al. CycleMLP: A MLP-like architecture for dense prediction[C]//International Conference on Learning Representations. OpenReview. net, 2022: 1-21.

[14] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// Proceedings of the 32nd International Conference on Machine Learning. New York: ACM, 2015: 448-456.

[15] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA: IEEE, 2022: 11966-11976.

[16] CHU X, TIAN Z, WANG Y, et al. Twins: Revisiting the design of spatial attention in vision transformers[C]//Advances in Neural Information Processing Systems(NIPS). 2021, 34: 9355-9366.

[17] CHU X X, TIAN Z, ZHANG B, et al. Conditional positional encodings for vision transformers[EB/OL]. arXiv: 2102.10882. <https://arxiv.org/abs/2102.10882>.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM, 2017: 6000-6010.

[19] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with restarts[C]//International Conference on Learning Representations. Toulon: OpenReview. net, 2017: 1-16.

[20] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]//International Conference on Learning Representations. New Orleans: OpenReview. net, 2019: 1-19.

[21] MULLER R, KORNBLITH S, HINTON G E. When does label smoothing help? [EB/OL]. arXiv: 1906.02629. <https://arxiv.org/abs/1906.02629>.

[22] ZHANG H Y, CISSE M, DAUPHIN Y N, et al. Mixup: Beyond empirical risk minimization[EB/OL]. arXiv: 1710.09412. <https://arxiv.org/abs/1710.09412>.

[23] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 6848-6856.

[24] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV). New York: ACM, 2018: 122-138.

[25] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[EB/OL]//arXiv: 1602.07360. <https://arxiv.org/abs/1602.07360>.

[26] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted residuals and linear bottlenecks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 4510-4520.

[27] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2020: 1314-1324.

[28] TAN M X, CHEN B, PANG R M, et al. Mnasnet: Platform-aware neural architecture search for mobile[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2020: 2815-2823.

[29] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. Long Beach, CA, USA: LR, 2019: 6105-6114.

[30] MEHTA S, RASTEGARI M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer [EB/OL]. arXiv: 2110.02178. <https://arxiv.org/abs/2110.02178>.

[31] SELVARAJU R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[J]. *International Journal of Computer Vision*, 2020, 128(2): 336-359.

【责任编辑 霍 欢】