融合 CNN 与 Transformer 的轻量化农作物叶片病害分类模型

吉一甲[1],唐秀芳[1],王梦蝶[1],李赛赛[1],李书松[2]

(1.安阳职业技术学院,河南 安阳 455000;
2.河南耀皓电子股份有限公司,河南 南阳 473400)

摘　要： 针对视觉 Transformer 在农作物叶片病害识别中的性能限制,提出一种融合卷积神经网络(Convolutional Neural Network,CNN)与 Transformer 的轻量化分类模型。构建局部-全局特征融合编码器,通过并行提取病斑局部纹理与区域间关联信息,增强复杂背景下病害区域的特征表征能力;结合层级式知识蒸馏策略与线性注意力机制,在降低计算开销的同时提升分类性能。实验结果表明,该模型能够保证识别精度,同时降低参数规模与推理延迟,具有较好的轻量化特征和应用潜力。

关键词： 农作物叶片病害;视觉 Transformer;知识蒸馏;轻量化模型

中图分类号： TP391　　**文献标识码：** A

Lightweight crop leaf disease classification model fused with CNN and Transformer

JI Yijia[1],TANG Xiufang[1],WANG Mengdie[1],LI Saisai[1],LI Shusong[2]

(1.Anyang Vocational and Technical College,Anyang,Henan 455000,China;
2.Henan Yaohao Electronics Co.,Ltd.,Nanyang,Henan 473400,China)

Abstract: To address the performance limitations of vision Transformer in crop leaf disease recognition, a lightweight classification model integrating Convolutional Neural Network (CNN) and Transformer is proposed. A local-global feature fusion encoder is constructed to enhance disease feature representation in complex backgrounds by parallelly extracting local lesion textures and inter-region correlation information. A hierarchical knowledge distillation strategy and linear attention mechanism are introduced to improve classification performance while reducing computational overhead. Experimental results demonstrate that the proposed model maintains recognition accuracy while reducing the number of parameters and inference latency, showing favorable lightweight characteristics and application potential.

Key words: crop leaf disease,vision Transformer,knowledge distillation,lightweight model

1　引言

农作物叶片病害的自动识别是精准农业中的重要技术环节。传统 CNN 能够有效提取病斑纹理和边缘等局部特征,在病害识别任务中得到广泛应用。然而,受限于卷积核感受野,CNN 难以建模病斑区域与叶片整体结构之间的长程依赖关系[1]。

视觉 Transformer 通过自注意力机制实现全局特征建模,能够捕获图像区域间的关联信息,但其对局部细节特征的感知能力相对不足。同时,田间病害图像通常存在样本规模有限、背景复杂和病斑尺度变化明显等问题,容易导致模型过拟合。此外,较大的参数规模和计算开销限制了视觉 Transformer 在移动端设备中的应用。

针对上述问题,本文提出一种融合 CNN 与 Transformer 的轻量化农作物叶片病害分类模型。通过设计局部-全局特征融合编码器,增强局部纹理与全局语义信息的联合表征;引入层级式知识蒸馏策略,缓解小样本训练过程中的过拟合问题;结合线性注意力机制和结构压缩方法以降低模型计算开销,实现病害分类精度与模型轻量化之间的平衡。

2　融合 CNN 与 Transformer 的模型设计

2.1　基础模型框架

为实现局部纹理信息与全局语义信息的联合表征,

基金项目: AIGC 助力安阳市文旅融合高质量发展研究(ASKG250722)

首先构建融合CNN与视觉Transformer的基础分类网络。该网络由卷积特征提取模块、Transformer编码模块和分类模块组成。

卷积特征提取模块采用3层卷积块提取病斑边缘、纹理等局部特征，将输入尺寸为224×224×3的图像逐步降采样至28×28×256的特征图，各卷积层均采用3×3卷积核，步长设置为2。卷积输出特征按照2×2区域划分为196个图块后输入Transformer编码模块，通过多头自注意力机制建立不同区域特征之间的关联关系。该模块包含6层Transformer编码器，隐藏层维度设置为384，注意力头数设置为6，用于学习病斑区域与叶片整体结构之间的全局语义信息。编码后的特征经过全局平均池化处理，并由全连接层输出病害类别概率。基础模型参数量约为12.8×10^6，作为后续特征融合、知识蒸馏和轻量化优化的基础。

2.2　模型优化

2.2.1　局部-全局特征融合编码器

基础模型中的Transformer编码器通过自注意力机制实现全局特征关联建模，但基于图块序列的特征表示对病斑边缘细节和局部纹理信息的提取能力有限。为增强局部特征与全局语义信息的联合表达，在Transformer编码层中并行引入深度可分离卷积分支[2-3]，构建局部-全局特征融合编码器。

深度可分离卷积分支采用5×5卷积核对输入特征进行空间特征提取，以保留病斑轮廓和细粒度纹理信息；自注意力分支用于建立不同区域之间的语义关联。两路特征通过可学习权重进行融合，融合过程表示为：

$$F_{out}=\alpha\cdot F_{att}+(1-\alpha)\cdot F_{conv},\ \alpha\in[0,1] \tag{1}$$

式中，F_{att}表示自注意力分支输出特征；F_{conv}表示卷积分支输出特征；α表示融合权重参数，初始化为0.5，并在训练过程中动态调整。通过局部卷积特征与全局注意力特征的联合建模，编码器能够同时利用病斑局部细节和区域间关联信息，提高病害特征表达能力。

2.2.2　层级式知识蒸馏策略

农作物病害数据集受田间采集条件限制，样本规模通常较小，Transformer学生网络在有限标注数据下训练时容易出现过拟合。为提高小样本条件下的泛化能力，引入层级式知识蒸馏策略，选取ImageNet预训练的MobileNetV3-Small作为教师网络，其参数量约为2.5×10^6。

蒸馏过程包括特征层蒸馏和输出层蒸馏两个部分。特征层蒸馏将教师网络倒数第二层输出的576维特征与学生网络全局平均池化后的384维特征分别映射至256维空间，并计算均方误差（Mean Squared Error，MSE）作为特征约束[4]；输出层蒸馏则根据教师网络与学生网络的类别概率分布计算库尔巴克-莱布勒散度（Kullback-Leibler Divergence，KLD），并通过温度系数τ平滑概率分布，以传递类别间关联信息。总损失函数表示为：

$$L_{total}=L_{CE}+\lambda_1 L_{MSE}+\lambda_2 L_{KLD} \tag{2}$$

式中，L_{CE}表示学生网络预测结果与真实标签之间的交叉熵损失；λ_1，λ_2分别表示特征蒸馏与输出蒸馏的权重系数。

训练初期将λ_1设为1.0，λ_2设为0.5，使学生网络优先学习教师网络的中间特征表示；当训练轮次达到总轮次的60%后，λ_1线性衰减至0.2，λ_2提升至1.0，使训练重点逐步由特征约束转向分类结果约束。软标签概率采用下式计算：

$$p_i=\frac{\exp(z_i/\tau)}{\sum_{j=1}^{C}\exp(z_j/\tau)} \tag{3}$$

式中，z_i为第i类的输出logit值；C为病害类别总数；τ表示温度系数，本文设为4，用于平滑概率分布。

2.3　模型轻量化实现

完成局部-全局特征融合和层级式知识蒸馏优化后，对网络结构进行轻量化处理，如图1所示。

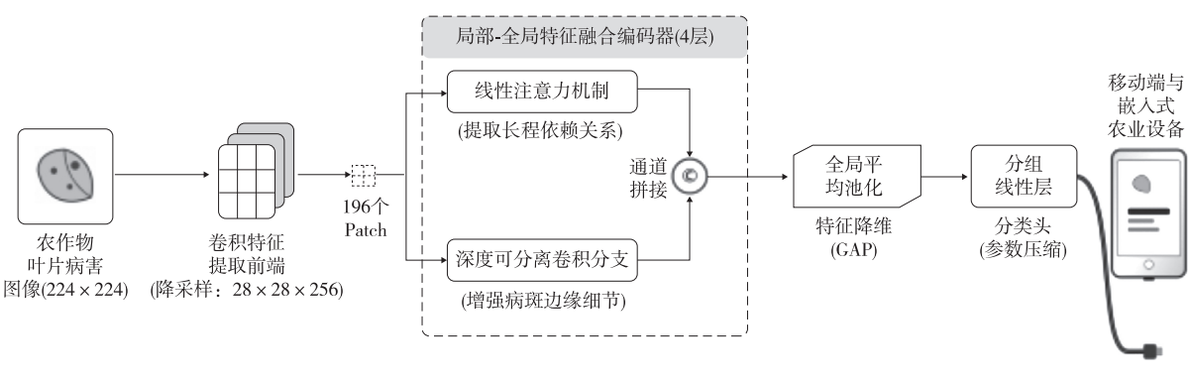

图1　轻量化Transformer架构

首先，在自注意力计算中引入线性注意力机制，对查询向量和键向量进行特征映射后完成关联计算，以避免显式构建图块数量平方规模的注意力矩阵，将计算复杂度由平方级降至线性级[5]。其次，压缩Transformer编码结构，将编码层数由6层减少至4层，隐藏层维度由384调整为256，注意力头数由6调整为4，单头维度保持为64。

在分类阶段，采用分组线性层替代标准全连接层，将

256维特征划分为4组分别进行映射后输出分类结果，从而减少分类层参数量。经上述轻量化处理，模型参数量约为 4.1×10^6，单张 224×224 图像的浮点运算量约为 0.86 GFLOPs，内存占用低于 50 MB，单帧推理延迟低于 30 ms。

3 模型应用效果分析

3.1 案例选取与实验设计

选取 PlantVillage 公开数据集中的番茄叶片病害子集开展实验，该子集包括早疫病、晚疫病、叶霉病和健康叶片 4 类图像，每类约 1 200 张。按照 8∶2 的比例划分训练集和测试集。实验平台采用 NVIDIA RTX 3080 GPU，显存为 10 GB，深度学习框架为 PyTorch 1.12；批量大小设置为 32，初始学习率设置为 0.000 1，采用 AdamW 优化器训练 120 轮。

评价指标包括分类准确率、单帧推理延迟和参数量。其中，分类准确率用于衡量病害识别性能，推理延迟和参数量用于评价模型轻量化程度。同时，每 5 轮记录一次验证集损失，用于观察不同模型的收敛过程。对比模型采用 ResNet-18 和 ViT-Tiny，参数量分别约为 11.7×10^6 和 5.7×10^6。3 种模型采用相同的数据划分和训练参数，并通过配对样本 t 检验分析性能差异，显著性水平设置为 0.05。

3.2 结果讨论

按照上述实验设置完成训练与测试，结果如表 1 所列。

表 1 3种模型性能对比

模型	准确率/%	推理延迟/ms	参数量
ResNet-18	91.2	18	11.7×10^6
ViT-Tiny	89.6	42	5.7×10^6
本文模型	95.3	26	4.1×10^6

由表 1 可知，本文模型分类准确率为 95.3%，较 ResNet-18 与 ViT-Tiny 分别提高 4.1 和 5.7 个百分点，且与 ResNet-18 的比较结果具有统计学意义（$p<0.01$）；参数量降至 4.1×10^6，低于两种对比模型；单帧推理延迟为 26 ms，较 ViT-Tiny 缩短 16 ms，但仍高于 ResNet-18。总体来看，本文方法在提高分类准确率的同时压缩了参数规模和 Transformer 结构的推理开销。不同模型的验证集损失变化如图 2 所示。

由图 2 可知，ResNet-18 在约第 60 轮后趋于稳定；ViT-Tiny 自第 40 轮后出现验证损失回升；本文模型则持续下降，并在约第 100 轮后趋于稳定，最终验证损失约为 0.18。与两种对比模型相比，本文模型在训练后期未出现明显的损失反弹，收敛过程相对稳定。结合前述蒸馏策略可以看出，层级式知识蒸馏对有限样本条件下的训练稳定性具有一定改善作用。

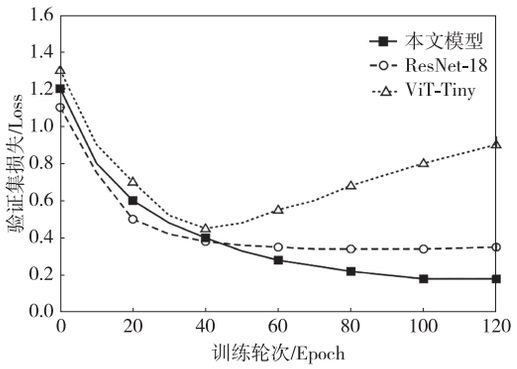

图 2 损失收敛曲线

4 结束语

围绕农作物叶片病害识别中的局部特征提取、过拟合和模型轻量化问题，本文对 CNN 与 Transformer 的融合方式进行了改进。通过局部-全局特征融合、层级式知识蒸馏和结构压缩，在降低参数量与推理延迟的同时提高了分类准确率。实验结果表明，本文模型准确率达到 95.3%，参数量降至 4.1×10^6，单帧推理延迟为 26 ms，可兼顾分类性能与计算效率。

然而，受实验数据范围限制，模型在复杂田间环境下的适应性仍有待进一步检验。后续可结合自然光照、遮挡及多尺度病斑等实际场景扩充样本，并进一步测试不同边缘设备上的运行性能，从而提高模型在实际农业场景中的适用性。

参考文献：

[1] 柏月鸿，张昕. 基于特征增强的 Transformer-RegNet-Y 网络的车道线检测研究[J]. 通信与信息技术，2026(2)：53-58.

[2] 杨丽霞，鲍雅君，张瑞，等. 面向跨域高光谱图像分类的嵌入 Transformer 层的轻量型 Res-3D-CNN[J]. 中国图象图形学报，2026，31(3)：927-943.

[3] 万勇，李骏杰，孙伟峰，等. 基于 CNN-Transformer 架构的电磁传播损耗预测算法[J]. 现代电子技术，2026，49(6)：43-48.

[4] 喻真妍，李少梅，马京振，等. RGAN：混合 Transformer 与生成对抗网络结合的地貌写景图风格迁移方法[J]. 地球信息科学学报，2026，28(3)：751-771.

[5] 连远锋，王鑫，崔超，等. 基于关键点感知和窗口-时序注意力 Transformer 的模糊手势重建方法[J]. 浙江大学学报(理学版)，2026，53(2)：172-180.

作者简介：

吉一甲(1999—)，硕士，研究方向：计算机技术、农业图像识别、软件开发、计算机网络。