

面向主粮作物病害诊断的跨模态注意力模型研究

左春婷* 王帅 张淑艳 于大海 于晓飞 姜峰

长春光华学院电气信息学院, 中国·吉林 长春 130033

摘要: 针对单模态深度学习方法在复杂田间环境下作物病害特征表征能力不足的问题, 提出一种基于跨模态注意力机制的多模态融合诊断模型。模型采用图像-文本双流编码, 通过交叉注意力实现视觉与语义特征对齐, 并引入门控融合进行动态自适应选择。在 Soybean Disease 与 PlantVillage 数据集上分别取得 98.74% 与 99.08% 的识别准确率, 显著优于单模态模型。消融实验验证了各模块有效性。该方法可有效挖掘图像与文本间的互补信息, 为田间复杂场景下的病害快速诊断提供技术方案。

关键词: 跨模态注意力; 多模态融合; 病虫害诊断; 交叉注意力

Research on Cross-Modal Attention Model for Major Grain Crop Disease Diagnosis

Zuo Chunting*, Wang Shuai, Zhang Shuyan, Yu Dahai, Yu Xiaofei, Jiang Feng

School of Electrical and Information Engineering, Changchun Guanghua University, China Jilin Changchun 130033

Abstract: To address the insufficient feature representation of single-modal methods for crop disease diagnosis in complex field environments, this paper proposes a multimodal fusion model based on cross-modal attention. The model employs image-text dual-stream encoding, aligns features via cross-attention, and introduces gated fusion for adaptive selection. It achieves 98.74% and 99.08% accuracy on Soybean Disease and PlantVillage datasets, significantly outperforming single-modal baselines. Ablation studies validate the effectiveness of each module. The method effectively exploits complementary information between image and text, providing a technical solution for rapid disease diagnosis in complex field scenarios.

Keywords: Cross-modal attention; Multimodal fusion; Crop disease diagnosis; Cross-attention

0 引言

近年来深度学习模型在植物病害图像识别中取得显著进展, 但单一图像模态在田间复杂环境下面临光照、背景、病征变异等挑战, 性能受限。与图像互补的文本描述, 如农民对症状的口述, 包含丰富语义线索, 可为病害识别提供额外判别信息。视觉-语言模型与跨模态注意力机制的发展为模态融合提供了技术可能, 但现有方法在模态对齐效率、自适应融合及轻量化部署方面仍有不足。为此, 本文提出一种基于交叉注意力与门控融合的多模态病害诊断模型, 主要贡献包括: (1) 设计图像-文本双流编码框架, 采用 ShuffleNet-v2 和 BERT 分别编码图像与文本; (2) 提出跨模态交叉注意力模块实现特征深度交互; (3) 引入门控融合机制实现自适应权重分配。

1 研究方法

1.1 模型整体架构

本文提出的基于跨模态注意力机制的农作物病虫害诊断模型整体架构如图 1 所示。模型由图像编码器、文本编码器、交叉注意力融合模块、门控融合模块以及分类预测模块构成。

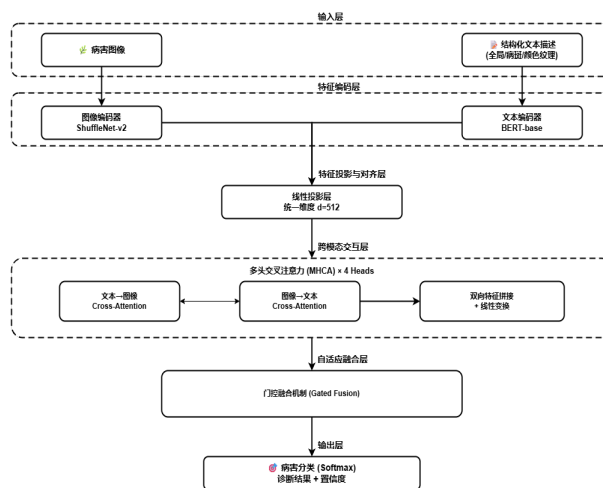


图1 模型整体架构图

1.2 双流特征编码

图像编码器: 本文采用 ShuffleNet-v2 作为图像特征提取主干网络。输入图像经过预处理后尺寸调整为 $224 \times 224 \times 3$, 通过 ShuffleNet-v2 提取全局视觉特征。设输入图像为 I , 图像编码器 ϵ_v 输出视觉特征向量 $F_v \in R^{d_v}$, 其中 $d_v=1024$ 。

文本编码器: 文本描述的生成与编码是多模态融合

的关键环节。对于每张病害图像，模型生成三类描述：（1）全局描述——整体叶片外观与病害概况；（2）局部病斑描述——病斑位置、形状、颜色等细节；（3）颜色纹理描述——叶片颜色变化与表面纹理特征。生成的文本描述通过 BERT 模型编码为语义向量 $F_i \in R^{d_i}$ ，其中 $d_i=768$ 。文本特征提取完成后，模型将视觉特征与语义特征投影至统一的对齐空间。

1.3 跨模态交叉注意力模块

为实现图像与文本模态的深度交互，本文引入交叉注意力机制。具体而言，设视觉特征序列为 $X_v \in R^{n \times d}$ ，文本特征序列为 $X_t \in R^{m \times d}$ ，其中 n 为视觉 patch 数量， m 为文本 token 数量， d 为特征维度。从文本到图像的交叉注意力定义为：

$$CrossAttn(Q_v, K_v, V_v) = softmax(Q_v K_v^T / \sqrt{d}) V_v$$

其中 $Q_t = X_t W_Q$ 为文本模态的查询矩阵， $K_v = X_v W_K$ 和 $V_v = X_v W_V$ 分别为图像模态的键矩阵和值矩阵。相似地，从图像到文本的交叉注意力可对称定义。本文采用双向交叉注意力机制，将图像到文本的注意力输出与文本到图像的注意力输出沿通道维度拼接后，经线性变换投影回原始维度，实现双向信息交流。

1.4 门控融合机制

交叉注意力模块实现了模态间的深度交互，但不同病害样本中图像与文本的贡献强度存在差异。为此，本文引入门控融合模块，实现模态特征的自适应动态选择。

门控融合模块设计为：

$$g = \sigma(W_g [F_v'; F_t'] + b_g)$$

$$F_{fused} = g \odot F_v' + (1-g) \odot F_t'$$

其中 F_v' 和 F_t' 分别为经由交叉注意力模块后的视觉与文本特征， W_g 和 b_g 为可学习参数， $\sigma(\cdot)$ 为 sigmoid 激活函数， $\odot$ 表示逐元素相乘。门控向量 $g \in [0, 1]^d$ 的每个维度指示模型在该特征维度上应更依赖视觉信息（ $g \rightarrow 1$ ）还是文本信息（ $g \rightarrow 0$ ），从而实现细粒度的自适应模态融合。

1.5 分类预测

融合特征 F_{fused} 经过两层全连接网络与 Dropout 正则

化（丢弃率 0.5）后，输入 softmax 层输出病害类别的概率分布：

$$\hat{y} = softmax(W_c \cdot Dropout(ReLU(W_f F_{fused} + b_f)) + b_c)$$

模型采用交叉熵损失作为优化目标：

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c})$$

其中 N 为样本数， C 为病害类别数， $y_{i,c}$ 为真实标签的 one-hot 编码。

2 实验与分析

2.1 实验设置

本研究采用两个公开数据集进行模型验证，两个数据集均按 60%/20%/20% 的比例划分为训练集、验证集和测试集。Soybean Disease 数据集包含大豆叶片病害图像，涵盖霜霉病、灰斑病、锈病、花叶病毒病等多种常见病害类型。PlantVillage 数据集：包含 54,303 张健康与患病叶片图像，涵盖 14 种作物品种和 38 个疾病类别，在受控条件下采集以降低背景噪声与光照多样性。为适配多模态实验，本文为两个数据集的每张图像均配以 Zhipu.ai 多模态模型自动生成的结构化文本描述（全局描述、局部病斑描述、颜色纹理描述）。

选取以下模型作为对比基线：ResNet50、ViT-B/16、ShuffleNet-v2、仅 BERT 编码的文本描述、图像与文本特征简单拼接、本文交叉注意力 + 门控融合模型。

本研究采用准确率、精确率、召回率、F1 分数和模型参数量作为评价指标。优化器采用 AdamW，初始学习率设为 $1e-4$ ，权重衰减 $1e-5$ ，batch size 为 64，训练轮数为 50 轮。所有实验在 NVIDIA GeForce RTX 3090 GPU 上进行。文本编码器选用 bert-base-uncased（参数量 110M），图像编码器选用 ShuffleNet-v2（参数量约 1.5M），冻结预训练权重，仅微调下游分类层。

2.2 实验结果

2.2.1 主实验

本文模型及对比方法在 Soybean Disease 与 PlantVillage 数据集上的性能如表 1 所示。

表1 各模型在两个数据集上的性能对比

模型	Soybean Disease Acc (%)	PlantVillage Acc (%)	参数量(M)
ResNet50	85.32 ± 1.2	94.56 ± 0.8	25.6
ViT-B/16	87.64 ± 1.0	96.23 ± 0.6	86.0
ShuffleNet-v2（仅图像）	83.95 ± 1.3	96.59 ± 0.5	1.5
仅文本	76.48 ± 1.5	72.31 ± 1.4	110
简单拼接	88.21 ± 1.1	97.04 ± 0.5	111.5
本文模型	98.74 ± 0.3	99.08 ± 0.2	112.0

实验结果表明：本文模型在 Soybean Disease 数据集上取得了 98.74% 的识别准确率，在 PlantVillage 数据集上达到 99.08%，分别超越 ShuffleNet-v2 基线模型 14.79 和 2.49 个百分点。相比简单拼接方法，本文模型分别提升了 10.53 和 2.04 个百分点，充分证明了交叉注意力与门控融合机制在模态对齐与深度交互方面的有效性。

2.2.2.2 各病害类别诊断准确率

为检验模型对不同病害类型的诊断能力，本文在 PlantVillage 数据集上对各类病害分别进行了统计，结果如表 2 所示。

从表 2 可以看出，番茄晚疫病的识别准确率为 98.80%，略低于番茄早疫病（99.20%），两者病斑形态相似是造成混淆的主要原因，但文本描述中关于病斑边缘特征与扩展速度的差异为模型提供了关键的区分线索。类似地，马铃薯早疫病（98.50%）与番茄早疫病症状相近，但玉米锈病（99.43%）与苹果黑星病（99.38%）等特征鲜明的病害类别取得了更高准确率。这表明，在症状相似的病害之间，文本模态提供的语义信息能够有效弥补视觉特征区分度不足的问题。

2.2.3 消融实验

为评估模型中各模块的贡献，本文在 Soybean Disease 数据集上设计了消融实验，逐步移除交叉注意力模块和门控融合模块，结果如表 3 所示。

消融实验结果表明移除交叉注意力后模型准确率下降 9.10 个百分点，门控融合机制移除导致准确率下降 2.46 个百分点。移除两者后，模型退化为简单特征拼接，准确率进一步降至 88.21%，与仅图像模型（83.95%）的差距主要源自文本模态的补充信息。

2.2.4 模型参数量与推理效率

本文统计了各模型在不同硬件条件下的推理时间和参数量，如表 4 所示。

3 讨论与展望

本研究验证了交叉注意力与门控融合机制在多模态农作物病害诊断中的有效性。实验结果表明，完整的跨模态融合框架相比仅图像基线提升 14.79 个百分点，相比简单拼接提升 10.53 个百分点。交叉注意力实现了图像病斑区域与文本症状描述之间的细粒度对齐，使模型能够“看到”病斑并同时“理解”农民对病斑的口头描述。这一机制尤

表2 本文模型在PlantVillage数据集各病害类别上的识别准确率

作物	病害类别	样本数	准确率(%)	易混淆情况
番茄	早疫病	500	99.20	—
番茄	晚疫病	500	98.80	与早疫病病斑形态相似
番茄	叶霉病	450	99.11	—
玉米	大斑病	380	98.95	—
玉米	锈病	350	99.43	—
马铃薯	早疫病	400	98.50	与番茄早疫病症状相近
马铃薯	晚疫病	400	98.75	—
苹果	黑星病	320	99.38	—
苹果	锈病	300	99.00	—
葡萄	黑腐病	350	99.14	—
平均	—	3,950	99.04	—

表3 消融实验结果（Soybean Disease数据集）

模型配置	准确率(%)	精确率(%)	召回率(%)	F1分数
完整模型	98.74	98.52	98.36	98.44
移除门控（仅交叉注意力）	96.28	95.97	95.82	95.89
移除交叉注意力（仅门控）	89.64	89.21	88.93	89.07
移除两者（简单拼接）	88.21	87.85	87.63	87.74
仅图像（无文本）	83.95	83.61	83.27	83.44

表4 模型参数量与推理效率对比

模型	参数量(M)	GPU推理时间(ms/样本)	CPU推理时间(ms/样本)
ResNet50	25.6	5.2	32.4
ViT-B/16	86.0	8.7	58.2
本文模型（完整）	112.0	10.3	67.5

其适用于田间边缘场景，提升模型在真实环境下的鲁棒性。

参考文献:

[1] WANG H Y, ZHOU G X, CHEN G Y. CEA-Net: A multi-modal model for corn disease classification with dynamic fusion and cross-layer connection mechanism[J]. Pattern Recognition, 2026, 173: 112788.

[2] LI H N, CHEN B Y, CHEN J J. ITIMCA: Image-text information and cross-attention for multi-modal cassava leaf disease classification based on a novel multi-modal dataset in natural environments[J]. Crop Protection, 2025, 189: 106981.

[3] 马晓慧等. 结合注意力机制与双向特征融合的叶片病害检测方法[J]. 中国农机化学报, 2024(10).

[4] 张家瑜等. 基于选择性注意力神经网络的木薯叶病害检测算法[J]. 农业机械学报, 2024(05).

[5] 李坤等. 基于分层特征交叉注意力的小样本马铃薯

病害叶片识别[J]. 江苏农业科学, 2024, 52(10): 210-216.

[6] 薛悦平, 胡彦蓉, 刘洪久等. 基于多模态预训练模型的水稻病虫害图像描述生成研究[J]. 南京农业大学学报, 2024, 47(4): 782-791.

[7] 杨巨成, 沈杰, 刘建征等. 基于图文模型和对比学习的水稻病害识别方法[J/OL]. 天津科技大学学报.

[8] 张净, 邵文文, 刘晓梅等. 基于超图的双模态特征融合的作物病害识别算法[J]. 江苏农业科学, 2023, 51(15): 164-173.

基金项目: 吉林省教育厅科学研究项目“基于多模态大模型的吉林省主粮作物病害智能诊断系统”(JJKH20261965KJ)。

作者简介: * 通讯作者: 左春婷(1990-), 女, 汉族, 湖北省松滋市人, 副教授, 硕士, 研究方向: 人工智能、计算机视觉。