doi:10.6041/j.issn.1000-1298.2025.12.045

基于轻量化多模态 Blend-CNN 模型的小麦病虫害识别方法

郝　霞　周子钰　宋　扬　李广烨　王志军　郭旭超
（山东农业大学信息科学与工程学院，泰安 271017）

摘要：小麦作为全球重要粮食作物之一，对其病虫害进行及时诊断与防治可有效减少粮食损失。然而，深度学习方法通常依赖于大量的训练数据和高性能计算资源，在小样本学习数据和资源有限情况下存在限制。为了解决上述问题，提出一种基于知识增强的轻量化多模态小麦病虫害识别模型 Blend-CNN（Multi-modal blend convolutional neural network）。该模型以双分支卷积神经网络为主体框架，在采用 EfficientNet 骨干网络提取病虫害图像特征基础上，引入多路 TextCNN 骨干网络提取病虫害文本描述特征，以得到更多的病虫害特征信息，从而提高模型识别准确率；提出基于卷积网络的多模态学习方法，使模型能够从全局上更优地融合 2 种模态信息。为了减少传统多模态方法的准确率损失，构建了梯度拼配损失函数；为了验证所提模型的有效性，在包含 880 个样本的自建数据集上进行了对比实验。结果表明，所提模型不仅在该数据集上取得最高识别准确率（96.95%），并且与其他对比模型相比，参数量更少、复杂度更低，更加轻量化，可以用于边缘设备，可为小样本和资源有限场景下的小麦病虫害识别提供一定理论支撑。

关键词：小麦；病虫害识别；深度学习；多模态融合；多模态学习；小样本学习
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1000-1298(2025)12-0490-09　　**OSID**：

Wheat Disease and Pest Recognition Method Based on Lightweight Multimodal Blend-CNN Model

HAO Xia　ZHOU Ziyu　SONG Yang　LI Guangye　WANG Zhijun　GUO Xuchao
(*College of Information Science and Engineering*, *Shandong Agricultural University*, *Taian* 271017, *China*)

Abstract: Wheat, as one of the world's essential food crops, requires timely diagnosis and control of diseases and pests to significantly reduce yield losses, which is crucial for global food security. However, deep learning methods typically rely on large amounts of training data and high-performance computing resources, which pose limitations in scenarios with small sample sizes and limited resources. To address these issues, a knowledge-enhanced lightweight multi-modal wheat diseases and pests identification model named multi-modal blend convolutional neural network (Blend-CNN) was proposed. The model was structured around a dual-branch convolutional neural network framework. It utilized an EfficientNet backbone network to extract image features of diseases and pests and incorporated a multi-branch TextCNN backbone network to extract textual features of diseases and pests descriptions, thereby obtaining more feature information and improving identification accuracy. Additionally, an innovative convolutional network-based multi-modal fusion method was introduced, allowing the model to optimally integrate information from both modalities globally. Furthermore, to mitigate the accuracy loss common in traditional multi-modal methods, the gradient blend loss function was developed. Finally, to verify the model's effectiveness, comparative experiments were conducted on a constructed dataset containing 880 samples. The results demonstrated that the proposed model achieved the highest identification accuracy of 96.95% on this dataset, compared with other models, it had fewer parameters, lower complexity, and it was more lightweight which was applicable for edge devices, offering theoretical support for wheat diseases and pests identification in scenarios with limited sample sizes and resources.

Key words: wheat; diseases and pests identification; deep learning; multi-modal fusion; multi-modal learning; few-shot learning

收稿日期：2024-07-26　修回日期：2024-10-08
基金项目：山东省自然科学基金项目(ZR2023QF016)和国家自然科学基金项目(32401702)
作者简介：郝霞(1990—)，女，副教授，主要从事计算机视觉和植物表型研究，E-mail：haoxia@sdau.edu.cn
通信作者：郭旭超(1992—)，男，副教授，主要从事知识图谱构建与应用和多模态学习研究，E-mail：guoxc@sdau.edu.cn

0　引言

根据国家统计局数据，2022 年我国小麦播种面积和产量分别占全国粮食作物种植总面积和农作物总产量的 19.88% 和 20.06%，在粮食生产中占据重要地位。然而，小麦生产过程中面临诸多挑战，其中病虫害是最为普遍且破坏力强的因素之一，严重威胁着粮食产量和品质。有效的病虫害防治对于减少小麦粮食损失、保障粮食供应安全至关重要。

传统的小麦病虫害诊断主要依赖人工观察与经验判断。由于病害症状复杂且易混淆，诊断过程要求极高的专业知识与实践经验，人工识别不仅效率低、主观性强，且难以实现大规模监测。随着人工智能与计算机视觉的快速发展，深度学习凭借其强大的特征表示与端到端优化能力，已经被广泛应用于农业领域[1-3]。目前，众多学者通过不断优化和改进诸如 AlexNet[4]、ResNet[5] 和 MobileNet[6]、Transformer[7] 以及 YOLO[8] 系列等多种深度学习模型，已经将其应用于多种农作物的各种病虫害分类检测与识别[9-11]，且研究对象不限玉米、小麦、苹果、杂草等。BRAHIMI 等[12] 将 Inception[13]、ResNet、SqueezeNet 等[14]深度学习模型应用于 14 种农作物病害检测中；XU 等[15] 提出 RFE－CNN 模型，不仅可以检测小麦病虫害，而且还可以预测小麦长势。TURKOGLU 等[16]基于多组卷积神经网络与长短期记忆网络构建了多模型的神经网络系统 MLP-CNNs，并成功用于苹果病虫害检测。RÖSSLE 等[17]使用 EfficientNet 对小麦赤霉病的 RGB 图像进行分类检测。SLADOJEVIC 等[18] 及 FERENTINOS[19] 将包括 CaffeNet[20]、AlexNet、GoogLeNet[21] 在内的多种模型应用总计 38 种植物的病虫害识别上。尽管上述方法在特征提取与识别准确率方面取得了良好效果，但这类深度学习模型通常需要大规模标注数据进行训练[5]，在数据有限的农业场景表现不稳定。现有研究表明病虫害文本描述往往包含丰富的病虫害症状信息，理论上对提高病虫害识别准确率具有一定辅助作用[22]，但在现有的小样本学习方法中尚未充分利用[23-24]。

多模态学习通过融合图像与文本等异质信息，为小样本场景下的特征补偿与泛化提升提供了新思路[25-27]。多模态深度学习融合方式主要包括早期融合、混合融合以及晚期融合[28]。其中，晚期融合被广泛应用于视觉语言任务，如 BAE 等[29] 提出基于 VGG[30] 与 TextCNN[31] 的多模态分类模型 Modified m-CNN，假借类似 m-CNN[32] 模型的三组 CNN 结构，得到更全面的特征，实现了细粒度花卉分类。ZHOU 等[33] 提出 ITK-Net 模型，通过注入文本描述，提高了作物病害检测准确率。HE 等[34] 提出了双分支视觉-语言模型 CVL，尝试解决多模态分类问题。随后，文献[35－37]相继提出了多种双分支多模态分类模型，均在晚期使用多层感知机实现融合特征的学习。然而，现有研究存在如下问题：由于多模态模型的学习能力更强，在数据有限情况下可能会导致过拟合；多个分支骨干网络各自学习速率不一，易造成不同模态网络之间泛化程度与过拟合程度不平衡问题，从而造成性能损失[25]；大部分多模态模型为了提高准确率，体量通常较大，难以部署在计算资源有限的嵌入式设备上，限制了以边缘设备为载体在实际小麦生产中的应用。

为此，本文提出一种适用于小样本的轻量化多模态小麦病虫害识别模型 Blend-CNN。该模型以双路 CNN 作为骨干网络，引入 EfficientNet[38-39] 提取图像特征，同时引入多路 TextCNN 模型从描述文本中提取文本上下文信息；提出基于卷积网络的多模态学习方法，使模型能够从全局上最优地融合 2 种模态信息；构建梯度拼配损失函数（Gradient blend loss function），以解决不同模态分支网络之间训练速度不同导致准确率下降的问题。

1　Blend-CNN 模型

如图 1 所示，Blend-CNN 模型主要由 3 部分组成：基于 EfficientNet-b3[38] 的视觉模态骨干网络、基于多路 TextCNN[31] 的文本模态骨干网络和进行模态融合特征学习的融合 CNN 网络。

图像模态主要从视觉角度捕捉病虫害的外观信息，文本模态则能传达更细致的语义信息，如病虫害的症状、成因和发展情况。这种互补性使得模型在小样本情况下，即使图像信息不足，也可以从文本中获得额外的判别线索，增强整体特征表示能力。通过引入文本模态，模型输入数据维度增加，相当于对输入数据进行了扩展，丰富了数据多样性。同时，多模态信息可以帮助模型更好地捕捉病虫害本质特征，减少因样本数量不足导致的过拟合。

1.1　视觉特征学习

该模块旨在通过卷积神经网络提取图像特征信息，由 EfficientNet-b3 构成骨干网络，包含多个 MBConv（Mobile Inverted Bottleneck）[40-41] 模块，其中每个 MBConv 块内建压缩和激励网络[42] 模块，可以通过显式建模通道间的相互依赖，自适应地重新校准通道特征响应。其详细结构如表 1 所示，其中每一行表示第 i 阶段，有 $\hat{L}_i$ 层网络，$\langle \hat{H}_i \times \hat{W}_i \rangle$ 表示输入分辨率，$\hat{C}_i$ 表示输出通道。“MBConv”字符串

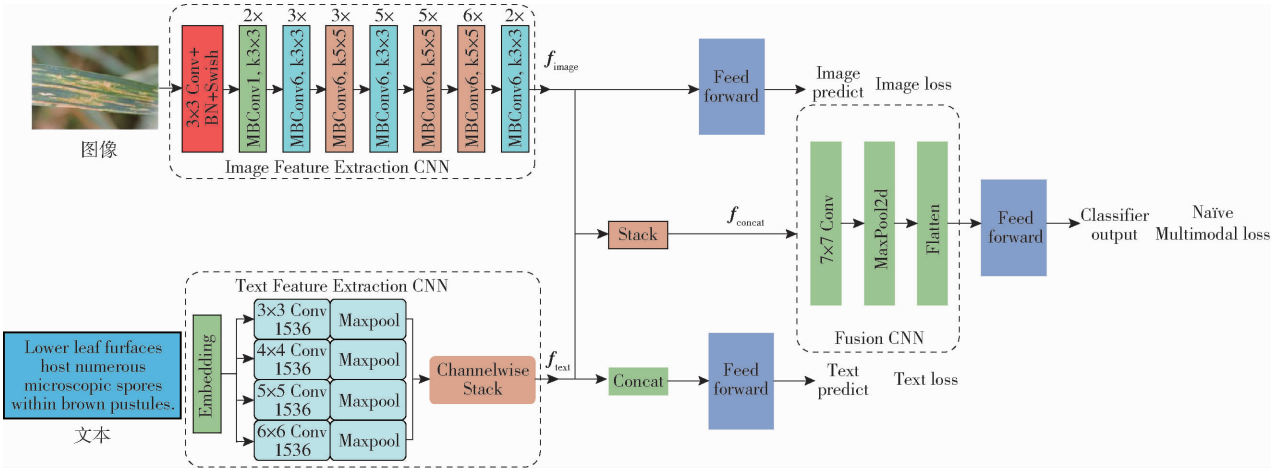


图 1 Blend-CNN 模型结构图

Fig. 1 Illustration of Blend-CNN architecture

表 1 视觉 CNN 模块结构

Tab. 1 Construction of visual CNN module

阶段	运算	分辨率/（像素×像素）	通道数	层数
1	3×3 Conv, BN, Swish	300×300	40	1
2	MBConv1, k3×3	150×150	24	2
3	MBConv6, k3×3	150×150	32	3
4	MBConv6, k5×5	75×75	48	3
5	MBConv6, k3×3	38×38	96	5
6	MBConv6, k5×5	19×19	136	5
7	MBConv6, k5×5	10×10	232	6
8	MBConv6, k3×3	10×10	384	2

后数字表示缩放因子，k3×3 与 k5×5 表示卷积核尺寸。

如图 2 所示，MBConv 模块由卷积核尺寸为 1×1 的卷积层、深度可分离卷积层、归一化层（Batch normalization, BN）[43]，以及 Swish 非线性激活函数、Squeeze-and-Excitation 模块构成。其中，最初 1×1 卷积层使用输入通道数 6 倍的 1×1 卷积核进行升维。SE 模块由一个平均池化层与 2 个全连接层组成，第 1 个全连接层维度是输入特征矩阵通道数量的 1/4，使用 Swish 激活函数；第 2 个全连接层维度等于深度可分卷积层输出特征的通道数，使用 Sigmoid 激活函数。当 MBConv 模块输入特征矩阵维度与输出相等时，通过引入残差链接，缓解梯度爆炸/梯度消失问题[5]。

通常情况下，模型性能越高，意味着参数量越多计算复杂度越高。例如，Vision Transformer（ViT）[44] 以及其他基于 Transformer 或 Swin Transformer[45-46] 的模型虽然能够取得较高准确率，但它们通常有更多的参数量以及更高的计算复杂度，意味着在诸如嵌入式设备等计算资源有限情况下难以应用。因此，在不降低识别准确率情况下，采用轻量化的视觉骨干网络（参数量仅为 1.0×10^7）很有必要。

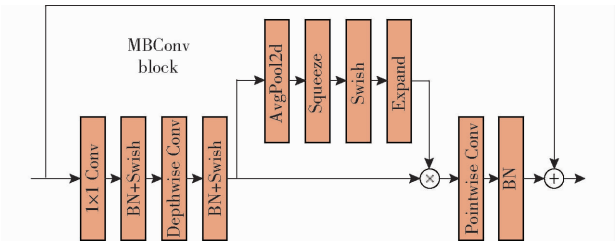


图 2 MBConv 模块结构图

Fig. 2 Illustration of MBConv block

1.2 文本特征学习

选用多分支文本卷积神经网络作为文本特征提取器，每个分支使用不同尺寸卷积核分别学习输入文本 x_i，得到不同语义空间的特征图公式为

$$x_{1:n}=\mathrm{concat}(x_1,x_2,\cdots,x_n) \tag{1}$$

采用 4 个分支，卷积核尺寸分别为 3、4、5 和 6。卷积步骤是使用尺寸为 h 的卷积核 $w\in\mathbf{R}^{h\times k}$，应用到 h 个词中产生特征 c_i，即

$$c_i=\mathrm{ReLU}(wx_{i:i+h-1}+b) \tag{2}$$

式中 $b\in\mathbf{R}$ 是可学习偏置项。最终一段文本特征表示 $c\in\mathbf{R}^{n-h+1}$ 为

$$c=\mathrm{concat}(c_1,c_2,\cdots,c_i,\cdots,c_{n-h+1}) \tag{3}$$

针对每一个卷积核卷积产生的特征图进行最大池化操作，对于每一个文本 CNN 分支，使用 1 536 个相同尺寸的卷积核产生相同数量不同的特征图。设向量 $\hat{\boldsymbol{c}}$ 表示每一个分支输出得到的特征向量，公式为

$$\hat{\boldsymbol{c}}=\mathrm{concat}(\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_{384}) \tag{4}$$

其中 $$\hat{c}_i=\max(c_i)$$

实验结果表明，卷积核尺寸越大，模型准确率基本不变（详见 2.4 节）。因此，在 Blend-CNN 中，4 个文本 CNN 分支的卷积核尺寸分别设置为 3、4、5 和 6。通过 4 个分支文本 CNN 后，得到 4 个不同空间

的文本特征向量。

1.3　多模态特征融合

经过上述步骤，分别得到视觉特征向量为 $\boldsymbol{f}_{\text{image}}$，4 个维度相同的文本特征向量为 $\boldsymbol{f}_{\text{text}_i}(i=0,1,2,3)$，将视觉特征向量分别与文本特征向量两两堆叠得到混合特征 $\boldsymbol{f}_{\text{stack}}\in\mathbf{R}^{1\times8\times1536}$，公式为

$$\boldsymbol{f}_{\text{stack}}=(\boldsymbol{f}_{\text{image}},\boldsymbol{f}_{\text{text}_0},\cdots,\boldsymbol{f}_{\text{image}},\boldsymbol{f}_{\text{text}_3}) \tag{5}$$

为了使 2 种模态特征充分融合，同时为输出得到 1 维特征张量，采用尺寸为 7×7 卷积核 $w\in\mathbf{R}^{7\times7}$ 对级联后特征进行卷积操作，然后进行最大池化操作得到融合特征向量 $\hat{\boldsymbol{f}}$，公式为

$$\hat{\boldsymbol{f}}=[\hat{\boldsymbol{f}}(i,j)]\quad(i=0,j=0,1,\cdots,765) \tag{6}$$

其中 $\hat{\boldsymbol{f}}(i,j)=\max_{m=\{0,1\}}\max_{n=\{0,1\}}\boldsymbol{f}(2i+m,2j+n)$

$$\boldsymbol{f}=w\boldsymbol{f}_{\text{stack}}+b$$

此时，模态融合特征向量 $\hat{\boldsymbol{f}}\in\mathbf{R}^{765}$ 即是对于视觉与文本 2 种模态学习后综合得到的混合特征。最后使用线性前馈神经网络进行分类，得到最终输出 y，公式为

$$y=\text{FFN}(\hat{\boldsymbol{f}})=w\hat{\boldsymbol{f}}+b \tag{7}$$

1.4　多模态的梯度拼配损失函数

理论上，多模态模型相较于单模态理应拥有相当或更高的性能，因为多模态网络能够获取更多信息。而 WANG 等[47]研究表明，目前性能最高的单模态模型在不同任务、不同模态组合下有时会比多模态模型展现出更出色的性能。原因为多模态模型能力过强会导致其更易过拟合；不同模态在训练过程中收敛速度不同，泛化程度与过拟合程度不同，使用单一的优化策略可能会影响优化的效果。

过拟合意味着模型从训练集上学习到的某种模式无法泛化到目标分布上。设第 N 轮（Epoch）训练过程中，L_N^T 表示模型在训练集 T 上平均损失值（Loss），L_N^* 表示模型在目标分布下平均损失值，L_N^* 通常可以近似使用模型在测试集 V 上平均损失值 L_N^V 表示。则模型在第 N 轮训练后过拟合程度 O_N 为 L_N^* 与 L_N^T 之差，即

$$O_N\equiv L_N^*-L_N^T\approx L_N^V-L_N^T \tag{8}$$

假设模型继续训练 n 轮，此时模型在 2 个时刻之间的训练变化程度可以使用过拟合程度与泛化性程度的变化（$\Delta O_{N,n}$ 与 $\Delta G_{N,n}$）描述

$$\begin{aligned}\Delta O_{N,n}&=O_{N+n}-O_N\\ \Delta G_{N,n}&=L_{N+n}^*-L_N^*\end{aligned} \tag{9}$$

为了量化模型训练过程学习到的信息，定义过拟合程度与泛化程度比率 OGR，即泛化信息量与未泛化信息量之比为

$$OGR\equiv\left|\frac{\Delta O_{N,n}}{\Delta G_{N,n}}\right|=\left|\frac{O_{N+n}-O_N}{L_{N+n}^*-L_N^*}\right| \tag{10}$$

在模型训练过程中两个时刻之间的 OGR 可用于衡量模型所学习到的信息量，通常使用交叉熵损失函数。为了便于最小化 OGR，在多模态情况下，结合各个模态梯度估计并最小化 OGR^2 以保证每一步得到的增益不会小于单个性能最好的模态。

晚期融合多模态模型的训练过程可以看作最小化损失函数 L_{multi} 的过程，公式为

$$L_{\text{multi}}=L(C(\varphi_{m_1}+\varphi_{m_2}+\cdots+\varphi_{m_k}),y) \tag{11}$$

其中 $\{m_1,m_2,\cdots,m_k\}$ 分别表示 k 个模态，$\{\varphi_{m_1},\varphi_{m_2},\cdots,\varphi_{m_k}\}$ 表示学习每个模态的特征提取网络分支。“+”表示某种混合操作，比如说拼接或堆叠等，混合后的特征输入到分类器 C 中（通常是前馈神经网络）进行分类得到分类结果，最后与正确答案 y 通过损失函数 L（通常为交叉熵损失）计算得到多模态损失 L_{multi}。

为了平衡每个模态的收敛过程，定义拼配损失 L_{blend} 为

$$L_{\text{blend}}=w_{\text{multi}}L_{\text{multi}}+\sum_{i=1}^{k}w_iL_i \tag{12}$$

拼配损失为每个模态损失（包括朴素的多模态损失 L_{multi}）的加权和，拼配权重 w_i 可由最小化 OGR 过程得到，以平衡各个模态的收敛过程。单个模态损失可以通过在该模态的网络分支后添加临时分类器得到（图 1）。选择恰当的拼配权重 w_i 可以调整每个模态训练节奏，平衡各模态泛化程度与过拟合程度，以最小化 OGR，达到最佳训练效果。最终得到计算每个模态拼配权重的算法，即

输入：φ^N：第 N 轮训练后的模型

　　n：训练轮数

输出：$k+1$ 个模态（包含朴素的多模态）拼配权重的集合

1 for i in (all $k+1$ modalities) do

2　$\varphi_{m_i}^N\leftarrow load(\varphi^N)$；/* 初始化单模态子模型，包含该模态专用的分类器 $C_{m_i}\subset\varphi_{m_i}^N$ */

3　$L_N^{T'},L_N^V=validate(\text{model}=\varphi^N)$；/* 在训练集子集 T' 与验证集 V 上验证子模型 */

4　$train(\text{model}=\varphi^N,\text{dataset}=T,\text{epochs}=n)$；/* 训练 n 轮 */

5　$L_{N+n}^{T'},L_{N+n}^V=validate(\text{model}=\varphi^N)$；/* 再次验证得到训练后的损失 */

6　$O^i=\Delta O_{N,n},G^i=\Delta G_{N,n}$；/* 计算子模型过拟合程度与泛化程度 */

7 end

8 $\{w_i\}_{i=1}^{k+1}=\dfrac{1}{Z}\dfrac{G^i}{2O^{i2}}$；/* 计算各模态的拼配权重

其中 Z 是归一化常数 */

2 实验结果与分析

2.1 数据集构建

本文构建的小样本多模态小麦病虫害数据集包含图像和文本 2 种模态。其中图像模态数据来自公开数据集 LWDCD 2020[48]，共有小麦根冠腐烂、小麦叶锈病、小麦散斑、小麦白粉病、小麦孢囊线虫病、小麦赤霉病、小麦红蜘蛛、小麦茎基腐、小麦全蚀病、小麦纹枯病、小麦蚜虫共 11 种病虫害类别，从其中随机选择 880 幅图像构建成为图像模态，平均每个类别 80 幅，数据样本量较小，可用于客观验证模型的小样本学习能力，部分图像样本如图 3 所示，图像样本量分布情况如图 4 所示。

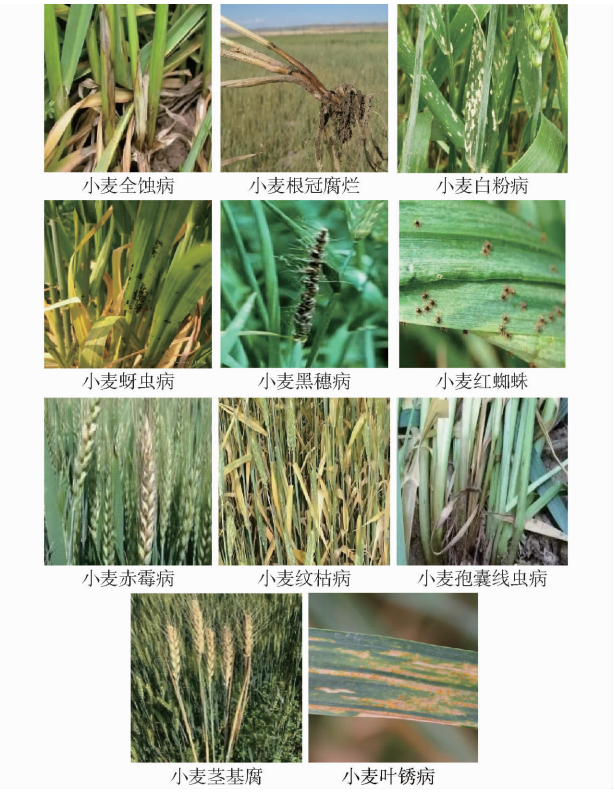


图 3 图像数据集样本示例

Fig. 3 Example of image samples

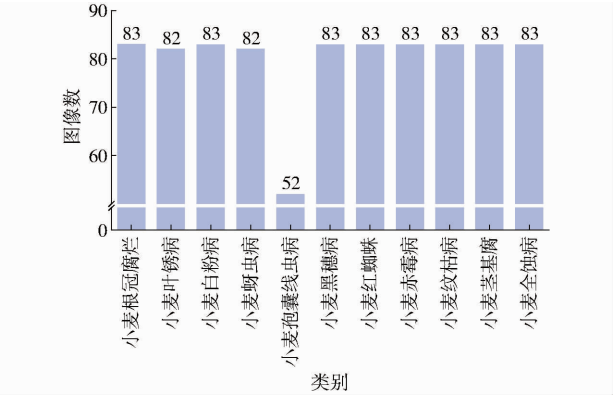


图 4 图像数据样本分布

Fig. 4 Distribution of samples in all categories

文本数据主要来源于华盛顿州立大学、内布拉斯加大学、加利福尼亚大学、西澳大利亚州第一产业和区域发展部农业和食品部数据库、维基百科等数据库及百科网站，数据量不足者通过手工撰写补足。图 5 为以小麦叶锈病为例的图像-文本样本描述。数据集按比例 8∶2 随机分为训练集与验证集，由于原始数据集中图像分辨率不同，而视觉 CNN 所接受的分辨率为 300 像素×300 像素，因此，将数据输入网络模型前需进行数据预处理。

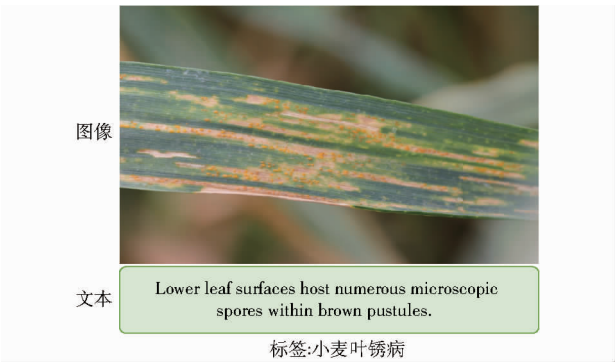


图 5 图像-文本样本描述示例

Fig. 5 Example of image-text sample descriptions

首先将原始图像使用双三次插值法变换为 320 像素×320 像素，再中心裁剪为 300 像素×300 像素。随后，使用随机水平、垂直翻转变换进行数据增强，以提高模型的泛化能力与鲁棒性，缓解过拟合，提升模型小样本学习的总体性能。最后，图像将会按照均值(0.49，0.51，0.37)，方差(0.21，0.20，0.20)进行归一化，样本均值与方差按照数据集中总体数据计算得到。归一化处理可以消除数据中不同特征之间量纲的差异，以保证不同特征对模型训练产生的影响大小类似。

2.2 实验环境与评价指标

实验环境为 x86_64 Arch Linux 系统，GPU 为 Nvidia RTX 4070 Ti，CUDA 版本为 12.1。模型使用 PyTorch 2.2.1 框架实现，模型优化器使用 Nesterov 动量的随机梯度下降 SGD 算法，初始学习率为 10^{-2}，权重衰减为 10^{-4}，动量为 0.9。使用拼配权重估计算法得到朴素多模态、图像模态、文本模态的梯度拼配权重分别为 0.18、0.73 和 0.09。

采用准确率衡量模型分类准确程度，并采用模型参数量和浮点运算量评估模型性能。

2.3 对比实验

为了验证本文模型性能，选用 ResNet、Vision Transformer 等 5 种传统计算机视觉模型、BiLSTM、BERT 等 3 种自然语言处理模型，以及 Modified m-CNN、OpenAI CLIP 等 4 种近年来提出的图文多模态分类模型进行对比实验，分别对比了各模型参数量、

浮点运算量、准确率以及在边缘设备上的适用性。其中，边缘设备上的适用性是通过在 Raspberry Pi 4b 设备上使用 CPU 进行推理测试，按照各模型能否正常推理进行判断。由表 2 可知，BiLSTM 模型表现最差，准确率仅为 9.76%，如图 6 所示，原因是 BiLSTM 模型参数为随机初始化，未经过预训练，在小样本任务中由于数据量太少而导致严重欠拟合；在对比的 CV 模型中，Vision Transformer 与 Swin Transformer 准确率低于其他 CNN 模型，由图 6 及表 2 可知，Vision Transformer 虽然在大数据图像分类任务上表现出色，但在样本量极少的小样本学习任务中会出现过拟合，相比之下 CNN 模型更能胜任这类小样本任务。在对比的 NLP 模型中，BERT 表现最佳，TextCNN 次之；虽然 BERT 在训练后期也出现了过拟合现象，其结果准确率仍比较出色，但其参数量为 1.10×10^8，是 TextCNN 的 11 倍，因此在 Blend-CNN 模型具体实现中，选择更轻量的骨干网络 TextCNN。在多模态模型中，OpenAI CLIP 模型在准确率上表现最差。这一结果可能源于 OpenAI CLIP 模型的核心机制：使用对比学习将图像和文本映射到一个共同的潜在空间，并优化模型使得相关联的图像和文本在该空间中尽可能靠近。然而，本研究构建的多模态小麦病虫害数据集主要目的是通过文本提供分类相关的语义知识，而不是实现严格的图文对齐。因此，OpenAI CLIP 模型在本数据集上无法充分发挥其设计的优势，导致其性能不如其他多模态模型。

表 2　不同深度学习算法对比

Tab. 2　Comparison between different algorithms

类型	算法	参数量	浮点运算量	准确率/%	适用性
Visual	ResNet-152[5]	6.00×10^7	1.151×10^{10}	72.04	✓
Visual	ViT-B[44]	8.80×10^7	4.410×10^9	62.68	✓
Visual	Swin-v2-b[46]	8.80×10^7	2.032×10^{10}	69.51	✓
Visual	EfficientNet-b3[38]	1.10×10^7	1.830×10^9	86.46	✓
Visual	Inception-v3[13]	2.50×10^7	5.710×10^9	70.79	✓
Textual	BiLSTM[49]	3.00×10^6	6.760×10^6	9.76	✓
Textual	TextCNN[31]	1.00×10^7	1.220×10^9	82.31	✓
Textual	BERT[50]	1.10×10^8	5.097×10^{10}	85.63	×
Multi	Modified m-CNN[29]	1.38×10^8	3.217×10^{10}	76.71	×
Multi	文献[36]	1.39×10^8	1.339×10^{11}	66.30	×
Multi	文献[37]	1.35×10^8	6.242×10^{10}	86.56	×
Multi	OpenAI CLIP[51]	1.51×10^8	1.478×10^{10}	30.68	×
Multi	Blend-CNN	2.10×10^7	3.190×10^9	96.95	✓

注：类型列中“Visual”表示 CV 算法，“Textual”表示 NLP 算法，“Multi”表示多模态算法。

此外，NAWAZ 等[36]构建的基于 NTS-NET 与 BERT 多模态模型准确率稍低，推测可能是由于其图像骨干网络 NTS-NET 是为 CUB-200-2011 数据集定制的，泛化性偏弱，在其他数据集上表现不佳；GALLO 等[37]提出的多模态模型准确率仅次于 Blend-CNN，为 86.56%，比 Modified m-CNN[29]准确率更高，原因是文献[37]使用 Inception v3 与 BERT 作为图像与文本模态的骨干网络，而 Modified m-CNN 则是使用 VGG 与 TextCNN，骨干网络性能稍逊于前者。而本文模型在自构建数据集上取得最优准确率为 96.95%，原因是图像与文本模态的骨干网络选择了该数据集上准确率表现较高者，同时使用了 CNN 进行模态融合，而非仅仅使用多层感知机进行融合特征分类，并且构建了梯度拼配损失函数，使得多模态训练效果更佳。此外，Blend-CNN 模型参数量仅为 2.10×10^7，比 ResNet-152 少 3.90×10^7。边缘设备的适用性测试显示，仅有 Blend-CNN 及其他一些参数量较少的单模态模型能够在边缘设备上正常推理。因此，上述实验结果不仅表明本文模型在小样本条件下病虫害识别的有效性，而且更轻量化，可应用于边缘设备上。此外，本文还以 Blend-CNN 模型为例，可视化了其结果对应的混淆矩阵，如图 7 所示。整体来看，模型对大多数类别的分类

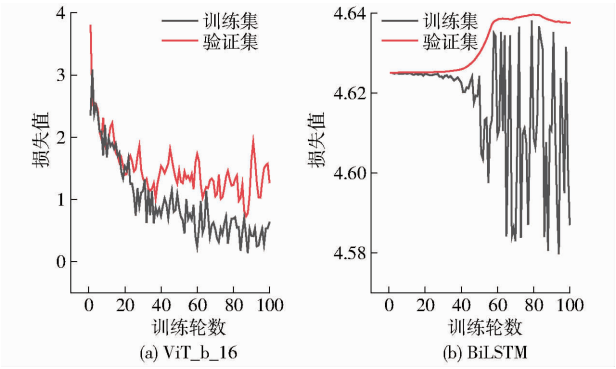


图 6　ViT 与 BiLSTM 损失值曲线

Fig. 6　Loss curves for ViT and BiLSTM

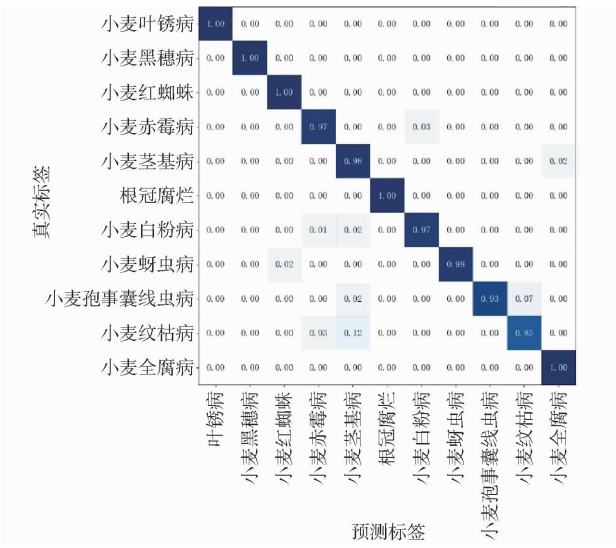


图 7　Blend-CNN 混淆矩阵

Fig. 7　Confusion matrix of Blend-CNN

准确率较高,特别是“小麦叶锈病”“小麦黑穗病”“小麦红蜘蛛”“小麦根冠腐烂”和“小麦全蚀病”预测精度基本达到 100%。然而,部分类别间存在一定混淆,例如“小麦纹枯病”与“小麦茎基腐”之间的错误分类比例最高,其原因是这 2 种病害之间差异较小、分类难度较高,在许多场景下难以区分。尽管如此,从混淆矩阵也可以看出模型整体平均准确率较高,表明其具备较为优秀的分类性能。

2.4 消融实验

为了说明本文模型的骨干网络、融合网络与损失函数等各改进模块有效性,进行了 5 组消融实验。如表 3 所示,图像骨干列表示视觉模态骨干网络,文本骨干列表示文本模态骨干网络,融合列表示图像与文本模态融合方法,CNN 表示基于卷积神经网络融合,FFN 表示基于前馈神经网络融合(若为多模态,则多个特征拼接;若为单模态,则直接使用 FFN)。由表 3 可知,单模态图像模型(EfficientNet)与单模态文本模型(TextCNN)准确率仅为 86.46% 和 82.31%,比 Blend-CNN 低约 10 个百分点;而朴素的多模态混合模型(直接拼接隐藏层特征并使用 FFN 分类)准确率为 93.12%,高于上述模型,但是仍然低于基于 CNN 融合的 Blend-CNN。另外,使用交叉熵损失函数(组 5)准确率仅有 90.85%,甚至比模态融合的实验组准确率低,这是由于文本模态的训练节奏远高于图像模态(由 1.4 节得出的拼配权重可看出),在图像模态收敛泛化时文本模态早已出现过拟合现象,导致总体准确率下降。而 Blend-CNN 使用拼配损失函数,能够平衡不同模态之间的训练过程,缓解多模态准确率损失。

表 3 不同模型结构准确率对比

Tab.3 Comparison between different model architectures

组别	图像骨干	文本骨干	融合	损失函数	准确率/%	描述
1	EfficientNet	TextCNN	CNN	Blend Loss	96.95	Blend-CNN
2	EfficientNet	Absent	FFN	Cross Entropy Loss	86.46	Visual Only
3	Absent	TextCNN	FFN	Cross Entropy Loss	82.31	Textual Only
4	EfficientNet	TextCNN	FFN	Blend Loss	93.12	Naïve Fusion
5	EfficientNet	TextCNN	CNN	Cross Entropy Loss	90.85	Naïve Loss

2.5 可解释性讨论

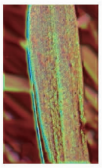
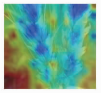
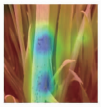
为了进一步说明本文模型的有效性,开展特征图可视化实验以及网格搜索实验。

对于特征图可视化,采用 Grad-CAM,即通过反向传播,获取模型最后隐藏层激活值对输入图像的梯度,将梯度转换成图像,实现特征图可视化。采用相同方法标注文本模态,使用红色标注权重最大的单词。同时将训练得到的 VGG 16 模型输出特征图可视化,用于对比,结果如表 4 所示。由表 4 可看出,图中高亮部位以及文本高亮字段实现了与病虫害实际特征的对应,表明本文模型 Blend-CNN 能够学习到与目标最相关的特征信息,在图像与文本 2 种模态上有一定学习与泛化能力。

为了探究文本模态骨干网络 TextCNN 中卷积核尺寸选择的合理性,进行网格实验。在其他变量不变情况下,卷积核分别从[3, 4, 5, 6]到[42, 44, 46, 48]依次递增,得到准确率变化如图 8 所示。由图 8 可知,卷积核尺寸不同,结果也不同。其中,采用[3, 4, 5, 6]、[12, 14, 16, 18]、[22, 24, 26, 28]等尺寸的卷积核得到的结果差距较小,说明此类卷积核对模型性能影响较小。卷积核尺寸过大,其模型性能反而更差,这是因为数据集中描述性文本相对较短,此类文本捕捉局部的关键点相对于长期记忆更重要。而[3, 4, 5, 6]取得最优值,因此,在 Blend-CNN 具体实现时,分别将卷积核尺寸设置为 3、4、5 和 6。

表 4 特征可视化

Tab.4 Feature visualization

类别	原图	可视化图(Blend-CNN)	可视化图(VGG16)	文本
小麦叶锈病				Lower leaf surfaces host numerous microscopic spores within brownpustules.
小麦赤霉病				Brown rot in wheat plant basal tissues, with small light brownspots on spikelets and glumes, turningthem yellow.
小麦黑穗病				A fine gray membrane enveloping the outer surface of wheat ears, which disintegrates upon maturity, dispersing black spores.
小麦红蜘蛛				Wheat round spiders or wheat long-legged spiders, characterized by black dots and redlegs, infest the leaves.

其模型性能反而更差,这是因为数据集中描述性文本相对较短,此类文本捕捉局部的关键点相对于长期记忆更重要。而[3, 4, 5, 6]取得最优值,因此,在 Blend-CNN 具体实现时,分别将卷积核尺寸设置为 3、4、5 和 6。

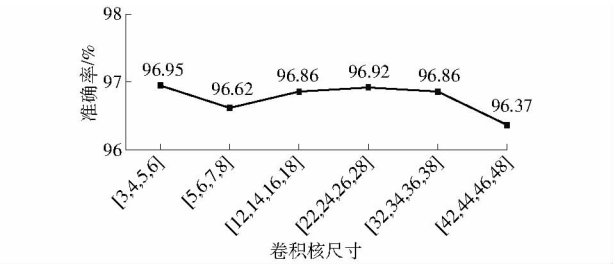


图 8　准确率随卷积核尺寸的变化

Fig. 8　Accuracy variation with filter size

3　结束语

为解决小样本条件下小麦病虫害识别问题，提出了一种新型轻量化的小样本多模态小麦病虫害识别模型 Blend-CNN，使用双路卷积神经网络架构，对图像与文本进行特征提取，并使用融合 CNN 对多模态的特征进行融合学习，以捕获到全局的综合信息。同时，使用梯度拼配方法构建损失函数并训练模型，使各个模态的分支网络学习状态达到相对平衡，以缓解传统多模态模型中各模态学习速率差异所引起的泛化与过拟合不平衡问题，从而提升模型的整体准确率。在自构建数据集上的实验结果表明本文模型性能最优，且参数量更少，为在计算资源有限的嵌入式设备部署提供可能，并为实际应用于作物病虫害防治提供了方法支撑。

参　考　文　献

[1] 慕君林，马博，王云飞，等. 基于深度学习的农作物病虫害检测算法综述[J]. 农业机械学报，2023，54(增刊 2)：301－313.
MU Junlin，MA Bo，WANG Yunfei，et al. Review of crop disease and pest detection algorithms based on deep learning [J]. Transactions of the Chinese Society for Agricultural Machinery，2023，54(Supp. 2)：301－313. (in Chinese)

[2] 翟肇裕，曹益飞，徐焕良，等. 农作物病虫害识别关键技术研究综述[J]. 农业机械学报，2021，52(7)：1－18.
ZHAI Zhaoyu，CAO Yifei，XU Huanliang，et al. Review of key techniques for crop disease and pest detection[J]. Transactions of the Chinese Society for Agricultural Machinery，2021，52(7)：1－18. (in Chinese)

[3] HAO X，CAO Y，ZHANG Z，et al. CountShoots：automatic detection and counting of slash pine new shoots using UAV imagery [J]. Plant Phenomics，2023，5：65.

[4] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems：25. Curran Associates，Inc.，2012.

[5] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，2016.

[6] HOWARD A，SANDLER M，CHU G，et al. Searching for MobileNetV3[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)，2019.

[7] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Advances in Neural Information Processing Systems：30. Curran Associates，Inc.，2017.

[8] WANG A，CHEN H，LIU L，et al. YOLOv10：real-time end-to-end object detection[C]// Advances in Neural Information Processing Systems：37. Curran Associates，Inc.，2024：107984－108011.

[9] SUNIL C K，JAIDHAR C D，PATIL N. Systematic study on deep learning-based plant disease detection or classification[J]. Artificial Intelligence Review，2023，56(12)：14955－15052.

[10] PACAL I. Enhancing crop productivity and sustainability through disease identification in maize leaves：exploiting a large dataset with an advanced vision transformer model[J]. Expert Systems with Applications，2024，238：122099.

[11] 徐雪环，贾岚，李红丹，等. 深度学习在小麦害虫分类研究中的应用[J]. 四川农业大学学报，2023，41(6)：1079－1089.
XU Xuehuan，JIA Lan，LI Hongdan，et al. Application of deep learning to wheat pest classification study[J]. Journal of Sichuan Agricultural University，2023，41(6)：1079－1089. (in Chinese)

[12] BRAHIMI M，ARSENOVIC M，LARABA S，et al. Deep learning for plant diseases：detection and saliency map visualisation [M]// Cham：Springer International Publishing，2018：93－117.

[13] SZEGEDY C，VANHOUCKE V，IOFFE S，et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，2016.

[14] IANDOLA F N，HAN S，MOSKEWICZ M W，et al. SqueezeNet：AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[J]. arXiv：1602.07360，2016.

[15] XU L，CAO B，ZHAO F，et al. Wheat leaf disease identification based on deep learning algorithms[J]. Physiological and Molecular Plant Pathology，2023，123：101940.

[16] TURKOGLU M，HANBAY D，SENGUR A. Multi-model LSTM-based convolutional neural networks for detection of apple diseases and pests[J]. Journal of Ambient Intelligence and Humanized Computing，2022，13(7)：3335－3345.

[17] RÖSSLE D，PREY L，RAMGRABER L，et al. Efficient noninvasive FHB estimation using RGB images from a novel multiyear，multirater dataset[J]. Plant Phenomics，2023，5：0068.

[18] SLADOJEVIC S, ARSENOVIC M, ANDERLA A, et al. Deep neural networks based recognition of plant diseases by leaf image classification[J]. Computational Intelligence and Neuroscience, 2016, 2016(1): 3289801.
[19] FERENTINOS K P. Deep learning models for plant disease detection and diagnosis[J]. Computers and Electronics in Agriculture, 2018, 145: 311 - 318.
[20] REYES A K, CAICEDO J C, CAMARGO J E. Fine-tuning deep convolutional networks for plant recognition[J]. CLEF (Working Notes), 2015: 1391.
[21] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[22] DAI G, FAN J, DEWI C. ITF - WPI: image and text based cross-modal feature fusion model for wolfberry pest recognition [J]. Computers and Electronics in Agriculture, 2023, 212: 108129.
[23] TIAN L, FENG J, CHAI X, et al. Prototypes-oriented transductive few-shot learning with conditional transport[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 16317 - 16326.
[24] LEE D, SONG S, SUH J, et al. Read-only prompt optimization for vision-language few-shot learning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 1401 - 1411.
[25] XU P, ZHU X, CLIFTON D A. Multimodal learning with transformers: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113 - 12132.
[26] 张友为, 王鑫鑫, 范晓飞. 基于深度学习的玉米和番茄病虫害检测技术研究进展[J]. 江苏农业科学, 2023, 52(10): 10 - 20.
ZHANG Youwei, WANG Xinxin, FAN Xiaofei. Research progress on corn and tomato diseases and pests detection technology based on deep learning[J]. Jiangsu Agricultural Sciences, 2023, 52(10): 10 - 20. (in Chinese)
[27] 邵明月, 张建华, 冯全, 等. 深度学习在植物叶部病害检测与识别的研究进展[J]. 智慧农业(中英文), 2022, 4(1): 29 - 46.
SHAO Mingyue, ZHANG Jianhua, FENG Quan, et al. Research progress of deep learning in detection and recognition of plant leaf diseases[J]. Smart Agriculture, 2022, 4(1): 29 - 46. (in Chinese)
[28] ZHANG Y, SIDIBÉ D, MOREL O, et al. Deep multimodal fusion for semantic image segmentation: a survey[J]. Image and Vision Computing, 2021, 105: 104042.
[29] BAE K I, PARK J, LEE J, et al. Flower classification with modified multimodal convolutional neural networks[J]. Expert Systems with Applications, 2020, 159: 113455.
[30] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2015.
[31] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv: 1408.5882, 2014.
[32] MA L, LU Z, SHANG L, et al. Multimodal convolutional neural networks for matching image and sentence[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV),2015.
[33] ZHOU J, LI J, WANG C, et al. Crop disease identification and interpretation method based on multimodal deep learning[J]. Computers and Electronics in Agriculture, 2021, 189: 106408.
[34] HE X, PENG Y. Fine-grained image classification via combining vision and language[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[35] KIELA D, GRAVE E, JOULIN A, et al. Efficient large-scale multi-modal classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32: 1.
[36] NAWAZ S, CALEFATI A, CARAFFINI M, et al. Are these birds similar: learning branched networks for fine-grained representations[C]//2019 International Conference on Image and Vision Computing New Zealand (IVCNZ), 2019: 1 - 5.
[37] GALLO I, RIA G, LANDRO N, et al. Image and text fusion for UPMC Food - 101 using BERT and CNNs[C]//2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), 2020: 1 - 6.
[38] TAN M, LE Q. EfficientNet: rethinking model scaling for convolutional neural networks[C]// Proceedings of the 36th International Conference on Machine Learning, 2019: 6105 - 6114.
[39] TAN M, LE Q. EfficientNetV2: smaller models and faster training[C]// Proceedings of the 38th International Conference on Machine Learning, 2021: 10096 - 10106.
[40] 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 (CVPR), 2018.
[41] TAN M, CHEN B, PANG R, et al. MnasNet: platform-aware neural architecture search for mobile[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[42] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[43] 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, 2015: 448 - 456.

(下转第 559 页)

[31] 郭文龙,罗熊,姜惠娟,等. 基于 TF－IDF 算法的方剂构成相似度可视化研究[J]. 中国中医药信息杂志,2019,26(7):104－108.
GUO Wenlong, LUO Xiong, JIANG Huijuan, et al. Research on visualization of formula composition similarity based on TF－IDF algorithm[J]. Chinese Journal of Traditional Chinese Medicine Information, 2019, 26(7): 104－108. (in Chinese)

[32] 赵磊,张涛,郑书恒,等. 运行植物诊所破解绿色防控技术推广难题[J]. 中国植保导刊,2018,38(2):84－85.
ZHAO Lei, ZHANG Tao, ZHENG Shuheng, et al. Running a plant clinic to solve the problem of promoting green prevention and control technology[J]. China Plant Protection Guide, 2018, 38(2): 84－85. (in Chinese)

[33] CHAPPELKA A H, GRULKE N E. Disruption of the 'disease triangle' by chemical and physical environmental change[J]. Plant Biology, 2016, 18: 5－12.

[34] MUNIR M. Plant disease epidemiology: disease triangle and forecasting mechanisms in highlights[J]. Hosts Virus, 2018, 5(1): 7－11.

[35] 潘红丽,李迈和,蔡小虎,等. 海拔梯度上的植物生长与生理生态特性[J]. 生态环境学报,2009,18(2):722.
PAN Hongli, LI Maihe, CAI Xiaohu, et al. Plant growth and physiological ecological characteristics on altitude gradient[J]. Journal of Ecological Environment, 2009, 18(2): 722. (in Chinese)

[36] 刘秉儒. 生物多样性的海拔分布格局研究及进展[J]. 生态环境学报,2021,30(2): 438.
LIU Bingru. Research and progress on the altitudinal distribution pattern of biodiversity [J]. Journal of Ecological Environment, 2021, 30(2): 438. (in Chinese)

[37] ALEXANDER H M. Disease in natural plant populations, communities, and ecosystems: insights into ecological and evolutionary processes[J]. Plant Disease, 2010, 94(5): 492－503.

(上接第 498 页)

[44] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv: 2010.11929, 2021.

[45] 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 (ICCV), 2021: 10012－10022.

[46] LIU Z, HU H, LIN Y, et al. Swin Transformer V2: scaling up capacity and resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 12009－12019.

[47] WANG W, TRAN D, FEISZLI M. What makes training multi-modal classification networks hard? [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[48] GOYAL L, SHARMA C M, SINGH A, et al. Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture[J]. Informatics in Medicine Unlocked, 2021, 25: 100642.

[49] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5): 602－610.

[50] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding [J]. arXiv: 1810.04805, 2019.

[51] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning, 2021: 8748－8763.