南京农业大学学报 2024,47(4):782-791
Journal of Nanjing Agricultural University
http://nauxb.njau.edu.cn
DOI:10.7685/jnau.202308009

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.
XUE Yueping,HU Yanrong,LIU Hongjiu,et al. Research on image description generation of rice diseases and pests based on multimodal pre-training model[J]. Journal of Nanjing Agricultural University,2024,47(4):782-791.

基于多模态预训练模型的水稻病虫害图像描述生成研究

薛悦平,胡彦蓉*,刘洪久,童莉珍,葛万钊

(浙江农林大学数学与计算机科学学院/浙江省林业智能监测与信息技术研究重点实验室/
林业感知技术与智能装备国家林业和草原局重点实验室,浙江 杭州 311300)

摘要:[**目的**]针对水稻病虫害图像分类技术缺少对病症描述的问题,本文提出一种轻量化的水稻病虫害图像描述模型,对水稻病虫害图像进行更为具体的描述。[**方法**]以白叶枯病、细菌性条斑病、恶苗病、三化螟虫、稻瘟病、稻曲病、纹枯病、飞虱、稻蓟马、胡麻斑病这十类常见的水稻病虫害开展研究,构建了水稻病虫害图像中文描述数据集。首先采用多模态预训练模型 CLIP 生成图像向量,其中包含基本的图像信息以及丰富的语义信息,采用映射网络将图像向量映射到文本空间里生成文本提示向量,语言模型 GPT-2 根据文本提示向量生成图像描述。[**结果**]在水稻病虫害图像描述数据集上,本文模型的指标总体明显优于其他模型,本文算法的 BLEU-1、BLEU-2、BLEU-3、BLEU-4、ROUGE、METEOR 指标较传统的 CNN_LSTM 模型分别提升 0.26、0.27、0.24、0.22、0.22、0.14。生成的图像描述具有准确、详细、语义丰富等优点。另外使用实际稻田图片对模型进行测试,实际田间的场景更为复杂多样,生成的图像描述指标与数据集指标对比总体仅有轻微下降,仍高于其他对比模型。本文模型对水稻病虫害的总体识别准确率达 97.28%。[**结论**]基于多模态预训练模型的水稻病虫害图像描述方法能够准确识别水稻病虫害病症并形成相应的病症描述,为水稻病虫害检测提供一种新思路。

关键词:多模态预训练模型;水稻病虫害;图像描述生成;诊断
中图分类号:TP391 **文献标志码**:A **文章编号**:1000-2030(2024)04-0782-10

Research on image description generation of rice diseases and pests based on multimodal pre-training model

XUE Yueping,HU Yanrong*,LIU Hongjiu,TONG Lizhen,GE Wanzhao

(College of Mathematics and Computer Science/Zhejiang Key Laboratory of Forestry Intelligence Monitoring and Information Technology Research/Key Laboratory of Forestry Sensing Technology and Intelligent Equipment, National Forestry and Grassland Administration,Zhejiang A&F University,Hangzhou 311300,China)

Abstract:[**Objectives**]Aiming at the lack of disease description in rice diseases and pests image classification technology,a lightweight rice diseases and pests image description model was proposed in this paper to describe rice diseases and pests image more specifically. [**Methods**]Ten common rice pests and diseases,such as rice bacterial blight,rice bacterial streak disease,rice bakanae disease,rice three chemical borers,rice blast,rice false smut,rice sheath blight,rice planthopper,rice thrip and rice brown spot were studied,and Chinese description data set of rice pests and diseases image was constructed. Firstly,the multimodal pre-training model CLIP was used to generate image vectors,which contained basic image information and rich semantic information. The mapping network was used to map the image vectors into the text space to generate text prompt vectors. Finally,the language model GPT-2 generates image descriptions according to the prompt vectors. [**Results**]The test results showed that the indexes of the model in this paper were significantly superior to other models in the image description data set of rice pests and diseases. Compared with the traditional CNN_LSTM model,the indexes of BLEU-1,BLEU-2,BLEU-3,BLEU-4,ROUGE and METEOR improved 0.26,0.27,0.24,0.22,0.22 and 0.14,respectively. And the generated image description had the advantages of accurate,detailed and rich semantics. The model was tested by using actual rice field pictures. The actual field scenes were more complex and diverse,and the generated image description index only slightly decreased compared with the data set index,which was still higher than other comparison models. The overall recognition accuracy of the model was 97.28%. [**Conclusions**]The image description method of rice diseases and pests based on multimodal pre-training model can accurately describe the rice diseases and pests,and provide a new idea for the detection of rice diseases and pests.

Keywords:multimodal pre-training model;rice diseases and pests;image description generation;diagnosis

收稿日期:2023-08-17
基金项目:教育部人文社会科学研究规划基金项目(18YJA630037,21YJA630054);浙江省自然科学基金资助项目(LY18G010005)
*通信作者:胡彦蓉,博士,副教授,研究方向为自然语言处理,rosehyr2004@aliyun.com。

水稻病虫害是制约水稻丰收的重要因素之一[1],专家预计 2023 年水稻病虫害发生 7 733.3 万 hm^2 次,同比增加 25.1%[2]。我国农业生产仍以分散经营为主,大多数农民对病虫害的认识不全面,不能准确诊断水稻的病变种类,普遍存在盲目施药情况,错误施药或过度施药都极易造成环境污染、农产品质量下降等问题。准确识别水稻病虫害种类有利于针对性采取防治措施,对中国粮食安全问题具有重要的现实意义。

随着计算机技术的不断发展,深度学习被广泛应用于图像分类[3-4]、目标检测[5-6]、自然语言处理[7-9]中,水稻病虫害也逐渐从人工识别转向智能识别[10-12],大大减少了人力、物力。以往的病虫害图像检测研究大多仅预测图像的标签,缺乏对图像内容的解释。参考实际生产场景中专家观察病虫害的特征并得出诊断结果的过程,本文将图像描述技术引入了水稻病虫害诊断工作,该技术根据水稻病虫害图像形成病症的详细描述,从而更具针对性地进行水稻病虫害防治工作。

图像描述是一种计算机视觉和自然语言处理相结合的跨模态任务。根据输入的一张或一系列图像,系统自动生成一段文字描述,这段描述要求和图片内容高度相似。早期的图像描述主要有两类方法:基于模板法和基于检索法。基于模板法[13]首先检测出目标图像中的对象、属性、行为和场景信息,然后将这些信息填入到预先设定的句子模板中。虽然可以生成语法正确的描述,但模板是预定义的,无法生成可变长度的描述,这类方法的主要问题为生成的文本缺乏多样性。基于检索法[14]从现有的图像数据库中检索出与给定图像契合的图像及其描述,这些描述语句被用作候选描述。检索法生成的句子符合语法规范,但若目标图片与训练集差别较大,则生成的语句会不准确,同时也不能生成特定于图像的正确描述。近年来,神经网络在计算机视觉和自然语言处理领域有了广泛应用,同时也促进了图像描述的发展,深度学习可以直接从大量数据中学习图像到描述语句之间的映射关系,实现图像到生成文字的“翻译”过程,利用深度学习方法生成的图像描述[15]更加准确多样。

图像描述技术在农业领域的研究较少。关于水稻病虫害图像描述生成任务存在以下难点:一是当前有关图像描述的研究主要基于 COCO[16]、Flickr30k[17]等公开数据集,但此类数据集的图像主要为生活场景,并不涉及农业生产场景。二是为达到较好的水稻病虫害图像描述生成效果,往往需要大量的计算资源,而农业生产信息化程度较低,能获取到的训练数据十分有限,需考虑如何利用小样本数据集完成模型训练。三是近年来图像描述生成的研究大多基于英文,而中国作为水稻生产大国,英文描述在实际生产场景的使用中具有局限性。

综上,本文提出一种水稻病虫害图像描述生成方法,以十类常见的水稻病虫害为研究对象,创建水稻病虫害图像描述数据集,基于中文语言环境,实现水稻病虫害图像描述生成。首先采用多模态预训练模型 CLIP(contrastive language-image pre-training)对输入图像进行编码,提取图像中的视觉信息;其次利用映射网络将视觉信息转换为语言提示向量序列[18-19];最后将其输入到语言模型,并逐步生成图像描述语句。CLIP 模型在大量图像和文本描述中使用对比损失函数进行训练,它的视觉和文本表现具有很好的相关性,弥合了图像底层特征和高层语意信息不一致的语义鸿沟问题[20]。

1　材料与方法

1.1　试验材料

1.1.1　图像采集及预处理　当前图像描述的研究主要基于公共图像描述数据集,例如 COCO、Flickr30k 等,此类公共数据集含有上万张图片,每张图片有对应的 5 句描述语言,图片内容以普通生活场景为主,尚未发现农业生产场景相关的图像描述数据集,因此需自制水稻病虫害图像描述数据集。因为水稻病虫害种类繁多,本文选取典型的十类水稻病虫害开展研究[21],分别为白叶枯病、细菌性条斑病、恶苗病、三化螟虫、稻瘟病、稻曲病、纹枯病、飞虱、稻蓟马、胡麻斑病。水稻病虫害种类繁多且发病时期大不相同,各类病虫害在不同时期有不同表现,同一病虫害在不同株上的症状也各有不同,加之农业生产的信息化程度较低,采集数据难度较大,因此采用各类水稻病虫害典型时期的图片进行研究。例如飞虱选取了幼虫、成虫以及虫害泛滥时的图片,稻瘟病选取了节瘟、叶瘟、谷粒瘟图片。

图片主要通过网络爬虫从网络上获取,相较于依次手动下载图片,网络爬虫技术可以有效减少人工获取数据的时间和精力。爬虫的具体流程:创建存储图片的文件夹,使用正则表达式获取图片链接,通过循环遍历完成图片爬取并保存到本地。爬虫方式虽然可以快速获取广泛数据,但总体数据质量不高,其中不

乏错误数据、重复数据等，需要对爬取到的数据进行一系列预处理工作。整体预处理流程见图 1。首先是数据的规范化处理，爬虫获取到的图片名称和图片格式多样，利用 Python 修改图片名和格式，采用固定格式的连续命名并统一为 JPG 格式，便于后续批量处理，然后使用感知哈希算法去除重复图片，应用图像平滑技术去除图片噪声。网络数据来源不可靠，仅通过代码自动处理无法保证数据质量，还需人工手动检查并剔除低质量图片，例如过于模糊的图片、内容范围过大的图片、无病变特征的图片，进一步提高图片质量。另外，还实地采集了水稻田间病虫害图片，由于拍摄于开放的环境，光照条件多变，导致有些图片的部分区域过亮或过暗，采用图像自适应亮度调整方法，调整图片中过亮或过暗区域。

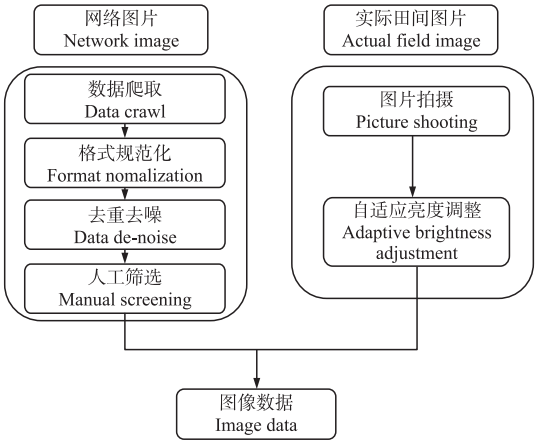


图 1 图片预处理流程

Fig. 1 Image preprocessing process

为减少模型过拟合风险，增强泛化能力和鲁棒性，通过样本增强的方式增加样本量。采用随机旋转、亮度调整、翻转和添加噪声的方式进行数据增强，增加样本的多样性，为后续的图像描述工作奠定了基础。采用的图片共 2 238 张，训练集、验证集、测试集划分的比例为 8∶1∶1。为进一步验证算法的优越性，另外使用 405 张实际稻田图片进行模型测试。

1.1.2 图像描述 图像描述数据集由图像和相应的描述语句构成，参考 Flick_30k_CN 数据集，为数据集中每个图片制作参考描述，描述图片中所有重要部分。图像描述尚无开源工具可以使用，因此结合人工标注和 Python 自动化的方式为图片制作描述语句。首先读取每个图片对应的名称，然后为每个图片都标注 5 句描述语言，且每句话的长度不少于 8 个字，主要内容为水稻病虫害的类型诊断以及特征说明，如表 1 所示，创建的描述语句共有 11 190 条，最终将描述语句存入 txt 文件。

表 1 水稻病虫害图像描述示例

Table 1 Example of rice pests and diseases image description

图片 Image	人工标注的参考语句 Manually annotated reference sentences
	患稻曲病的谷粒长有黑色稻曲球 The grain affected by rice false smut disease has black rice stalk balls 患水稻稻曲病的谷粒呈黑色 The grain affected by rice flase smut is black 患有稻曲病的谷粒常有黑菌核 Black sclerotium is common in grains with rice false smut 水稻患有稻曲病，稻穗上有黑色稻曲球 Rice suffers from rice false smut disease，which is a black stalk ball on the ear of rice 水稻受稻曲病影响，谷粒形成黑色菌核 Black sclerotium was formed in rice grains due to rice flase smut disease

1.2 试验方法

1.2.1 模型总体结构 本文采用传统的编码器-解码器架构实现水稻病虫害图像描述，算法模型主要由 3 部分构成：图像编码器、映射网络、语言模型。本文采用 CLIP 作为图像编码器提取图像特征，CLIP 编码器提取到的特征不仅包含基本的视觉信息还包含了丰富的语义信息，将提取到的信息通过映射网络转换为文本提示向量序列，并将其输入语言模型 GPT-2(generative pre-trainning)以生成图像描述。

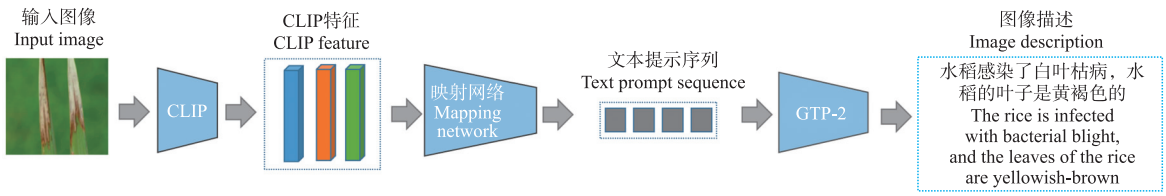


图 2 算法模型结构

Fig. 2 Algorithm model structure

图像描述的任务是为一张图像生成准确详细的描述语句。图像为 x,描述语句文本为 s,将已有的图像文本对记为 $\{x^i,s^i\}_{i=1}^N$,每句文本信息的长度扩充为 l,文本信息表示为 $s^i=s_1^i,\cdots,s_l^i$,θ 为模型可学习的训练参数,训练目标如式(1)所示:

$$\max_{\theta}\sum_{i=1}^{N}\log p_\theta(s_1^i,\cdots,s_l^i\mid x^i) \tag{1}$$

使用预训练的 CLIP 模型提取图像 x^i 的视觉信息后,采用一个轻量的映射网络 F 将 CLIP 图片向量映射为 k 个图像向量,即 $c_1^i,\cdots,c_k^i$,如式(2)所示:

$$c_1^i,\cdots,c_k^i=F(CLIP(x^i)) \tag{2}$$

将获得的图像向量与文本向量连接起来,如式(3)所示:

$$Z^i=c_1^i,\cdots,c_k^i,s_1^i,\cdots,s_l^i \tag{3}$$

在训练过程中,向语言模型提供 $\{Z^i\}_{i=1}^N$,使用交叉熵损失来训练映射网络,$p_\theta(s_j^i\mid c_1^i,\cdots,c_k^i,s_1^i,\cdots,s_{j-1}^i)$ 为生成字向量的概率,如式(4)所示:

$$L_X=-\left(\sum_{i=1}^{N}\sum_{j=1}^{l}\log p_\theta(s_j^i\mid c_1^i,\cdots,c_k^i,s_1^i,\cdots,s_{j-1}^i)\right) \tag{4}$$

1.2.2　图像特征编码器　通常图像描述模型首先将输入图像编码为特征向量,然后用于生成最终的单词序列。早期的工作利用预训练分类网络从中提出特征,例如 VGG、GooleNet、ResNet 等。而后的工作利用对象检测网络中更具表现力的特征,例如 YOLO 等。为了进一步挖掘图像中包含的信息,通常使用注意力机制来关注特定的特征。

CLIP[22] 是一种从文本中学习图像的方法,CLIP 打破了固定种类标签的范式,通过对互联网上获取的 4 亿个图像-文本对进行预训练,产生了视觉和文本数据共享的丰富语义潜在空间,CLIP 模型的泛化能力和迁移能力极强,因此在大量下游任务中都有应用,例如图像分类、目标检测、图像分割等。CLIP 模型主要由文本编码器和图像编码器 2 部分构成,其中文本编码器用来提取文本的特征,图像编码器用来提取图像的特征,对 2 个特征进行线性投射,并进行 L2 归一化,通过计算文本向量和图像向量的相似度以预测其是否为一个数据对。CLIP 模型结构如图 3 所示。

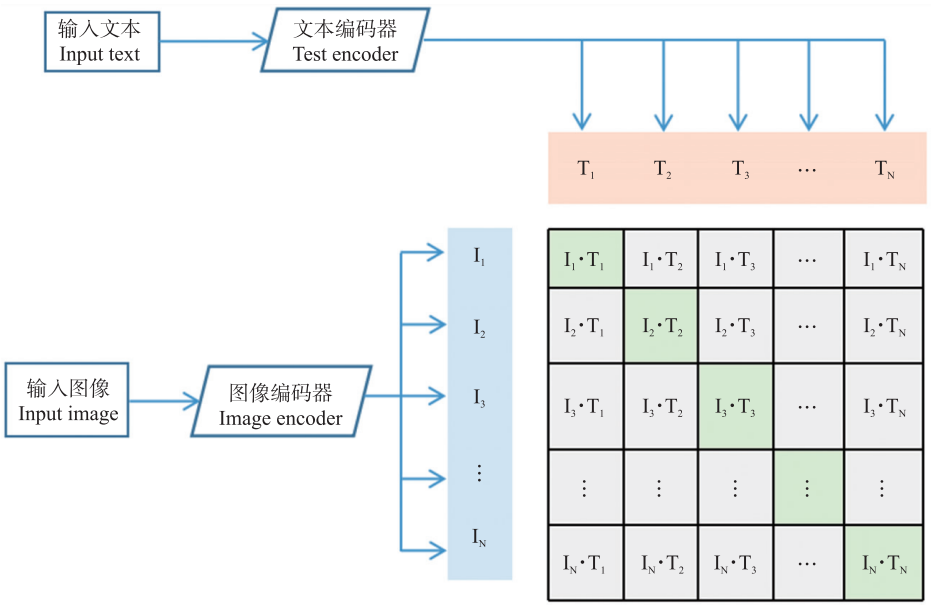


图 3　CLIP 模型结构

Fig. 3　CLIP model structure

本文采用经过预训练的 CLIP 模型作为图像特征编码器,CLIP 模型具有丰富的视觉-文本潜在相关性,可有效减少训练时间和数据需求,弥补图像描述任务的语义鸿沟问题。

1.2.3　映射网络　映射网络是整个图像描述结构中的重要组件,它是视觉空间和语言空间的桥梁。尽管 CLIP 模型和 GPT-2 模型都在大量的数据上进行预先训练,但两者之间是相对独立的。此外,水稻图像描述数据集较为特殊,与预训练使用的数据集差距较大,因此本文考虑微调语言模型,以获得更优秀的图像

描述结果。在微调语言模型的情况下,映射网络所面对的任务较轻,因此映射网络只需采用简单的全连接网络就能实现图像描述生成,其整体结构如图 2 所示。微调语言模型的训练方式可以使生成的描述语句更具表现力,但缺点是大幅增加了训练参数。为减少模型的训练参数,本文还比较了冻结语言模型的方法。在只训练映射网络的情况下,需要采用更具表现力的 Transformer[23] 作为映射网络,如图 4 所示,Transformer 在处理序列数据方面有更好的建模能力和计算效率,具有较强的特征抽取能力,并且可以实现并行计算。在使用 Transformer 作为映射网络时,将 CLIP 抽取的图像特征以及可学习常数作为 Transformer 的输入,常数可以从 CLIP 抽取的图像特征中检索有意义的信息。

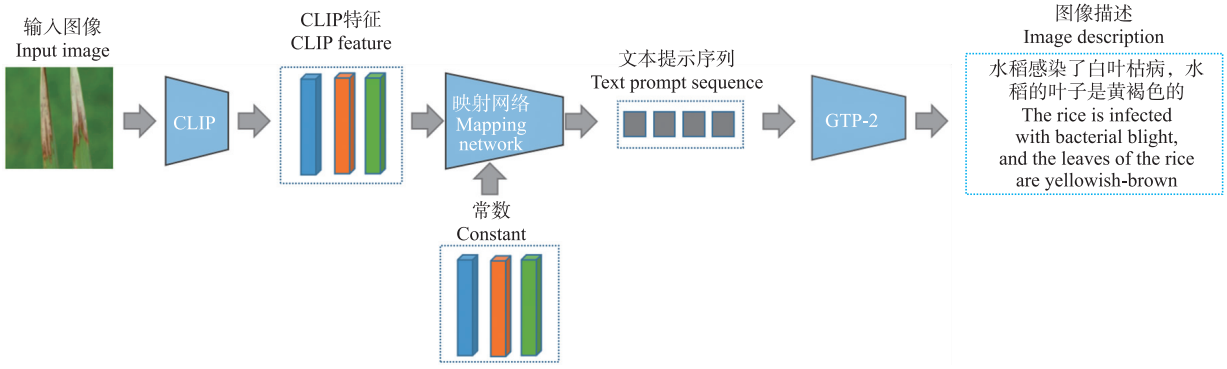


图 4 冻结语言模型模式结构

Fig. 4 Freeze the language model schema structure

1.2.4 语言模型 GPT-2 是通用的预训练语言模型,在大量高质量数据预训练过程中学习各类任务的相关信息,能够直接应用于文本翻译、文本摘要、文本生成等下游任务中[24]。GPT-2 是一个自回归模型(auto-regression),基于 Transformer 的解码器部分构成。GPT-2 每次产生新单词后,将新单词加到原输入句后面,作为新输入句,损失函数计算预测值与实际值之间的偏差。解码器模块加入了带掩码的自注意力机制(masked self-attention),只考虑在待预测值左侧的词对待预测词的影响。GPT-2 如图 5 所示。

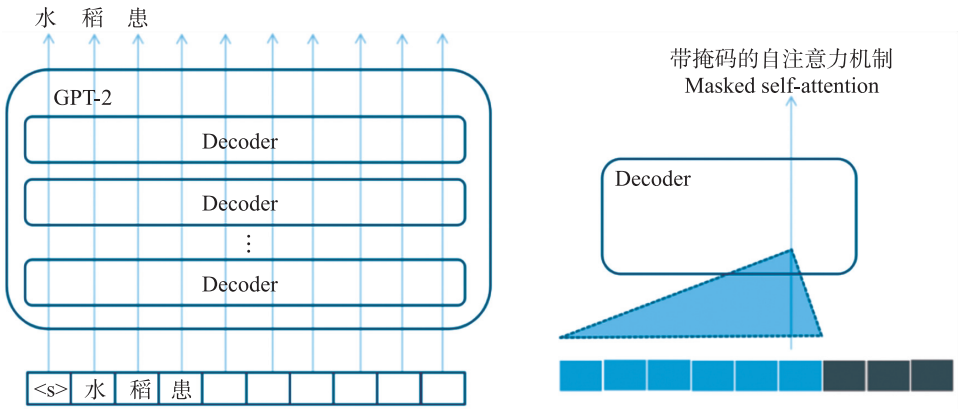


图 5 GPT-2 语言模型结构

Fig. 5 GPT-2 language model structure

本文使用 GPT-2 作为语言模型,GPT-2 在海量数据上完成大量参数训练,具备强大的语言理解能力,同时也可以生成丰富多样的文本。在生成结果的过程中,将映射网络生成的文本提示向量序列输入语言模型,逐步预测下一个词。

1.3 试验平台

使用的软件为 Python3.8 和 PyTorch1.11.0 深度学习框架,运行内存为 32 GB,搭载 AMD Ryzen 7 3800X 8-Core Processor 3.90 GHz 处理器,配备 NVIDIA GeForce RTX2080 SUPER,操作系统为 Windows 10,CUDA 版本为 11.7.102。

采用微调语言模型与冻结语言模型 2 种模式进行模型训练,2 种训练模式在训练时都冻结 CLIP 模型的权重,CLIP 模型权重的调整并不能带来性能的提升。在微调语言模型的情况下,映射网络采用 2 层全连接网络;在冻结语言模型的情况下,映射网络采用 Transformer 对其权重进行微调,且不加载预训练权重

为随机初始化。训练时的 batch size 为 20,epoch 为 50,学习率为 2e-5。

1.4　评价指标

采用性能对比分析与可视化结果分析 2 种评价方法对图像描述质量的评价。性能对比分析采用 BLEU[25](bilingual evaluation understudy)、METEOR[26] (metric for evaluation of translation with explicit ordering)、ROUGE[27] (recall-oriented understudy for gisting evaluation)、CIDEr[28] (consensus-based image description evaluation)4 个指标评价。

BLEU 指标是机器翻译常用的评价方法,应用于图像描述领域,分析候选描述中有多少 *n-gram* 出现在参考描述中,*n-gram* 为 n 个相邻字,可根据 *n-gram* 的选用形成多种评价指标,常用的有 BLEU-1、BLEU-2、BLEU-3、BLEU-4。BLEU 的计算公式如式(5)所示:

$$\begin{cases} CP_n(C,S)=\dfrac{\sum\limits_i\sum\limits_k \min(h_k(c_i),\max_{j\in m}h_k(s_{ij}))}{\sum\limits_i\sum\limits_k h_k(c_i)} \\ b(C,S)=\begin{cases}1 & l_c>l_s \\ e^{1-l_s/l_c} & l_c\leqslant l_s\end{cases} \\ BLEU_N(C,S)=b(C,S)\exp\left(\sum\limits_{n=1}^{N}w_n\lg\ CP_n(C,S)\right)\end{cases} \tag{5}$$

式中:$CPn(C,S)$为 *n-gram* 的匹配准确度,c_i 为候选描述;s_{ij}为参考描述,一个或多个连续单词在候选描述语句 c_i 和参考描述 s_{ij}中出现的次数记作 $h_k(c_i)$和 $h_k(s_{ij})$;l_c 是候选描述的长度,l_s 是参考描述语句的总长度;$b(C,S)$是一个简洁性惩罚机制,使用乘以简洁性惩罚参数的方法以防止很短的句子获得很高的分数;w_n 为权重。

METEOR 常应用于评价机器翻译,也适用于图像描述生成的评价。在 BLEU 的基础上利用 WordNet 等知识扩充了同义词集,主要评价候选描述与参考描述同义词的匹配情况,更接近于人工评价。

ROUGE 是一个评价召回率的指标,主要考察的是生成描述的忠实性和充分性,与 BLEU 的计算方式非常相近。

CIDEr 是一个主要应用于图像描述的指标,通过计算每个 n 元组的词频-逆文本频率(term frequency-inverse document frequency,TF-IDF)权重得到待评测语句与人工描述语句之间的余弦相似度,据此评价生成描述的效果。

图像描述生成由以上评价指标衡量,上述几项指标的得分越高,表示模型生成的描述越好,生成句子的质量越好。鉴于客观量化评价部分存在重复、重要信息缺失、缺乏多样性等问题,还采用了一种可视化结果分析方法,从描述的准确性、流畅性等多个维度对不同模型生成的描述语句进行了人工主观评估。

2　结果与分析

2.1　性能对比分析

相比传统的图像描述模型,本文的轻量化图像描述模型性能表现更好,更适用于水稻病虫害的图像描述。为了充分体现该方法针对水稻病虫害检测的适用性,在水稻病虫害图像描述数据集上结合相应的评价指标进行了评估试验,并与其他算法进行对比,其中 CNN_LSTM 模型是最基础的基于深度学习的图像描述模型,CNN_LSTM_ATT 是基于其施加注意力的模型。表 2 展示了本文算法与上述 2 种算法在自制水稻病虫害图像描述数据集上的结果。

由表 2 可知,本文算法的 BLEU-1、BLEU-2、BLEU-3、BLEU-4、ROUGE、METEOR 指标较传统的 CNN_LSTM 模型分别提升 0.26、0.27、0.24、0.22、0.22、0.14。本文算法生成的图像描述语句质量整体优于其他算法。另外还对比了微调语言模型与冻结语言模型的情况,从试验结果可知微调语言模型可以得到更好的结果,但需要增加更多训练参数、训练资源。

以上结果证明本文算法的有效性,其原因在于 CLIP 和 GPT-2 已有丰富的语义表示信息,可在少量训练资源的情况下,有效实现水稻病虫害的诊断与描述。

表 2 本文算法与其他算法的指标对比

Table 2 The index comparison between the algorithm in this paper and other algorithms

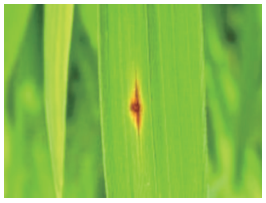
算法 Algorithm	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR
CNN_LSTM	0.29	0.13	0.07	0.04	0.26	0.31	0.19
CNN_LSTM_ATT	0.47	0.36	0.30	0.25	1.62	0.46	0.31
冻结语言模型(本文)Frozen language model(this paper)	0.51	0.37	0.29	0.24	1.50	0.52	0.32
微调语言模型(本文)Fine-tuning language model(this paper)	0.55	0.40	0.31	0.26	1.58	0.53	0.33

2.2 可视化结果及分析

为了更加直观地验证本文方法的有效性,在评价指标基础上加入人工主观评价的方法分析本文算法的性能,将本文方法与其他算法对于同一张图像生成的描述语句进行对比,对比结果如表 3 所示。

表 3 本文算法与其他算法的生成结果对比

Table 3 The results of this algorithm are compared with those of other algorithms

图像 Image	图像描述 Image captioning
	1、水稻受稻瘟病影响,水稻叶片出现灰霉病斑 Rice was affected by blast and gray mold spots appeared on rice leaves 2、白叶枯病叶缘呈深褐色 The leaf margin of bacterial blight is dark brown 3、水稻受白叶枯病的影响,水稻叶片边缘出现褐色条纹 Rice was affected by bacterial blight,and brown stripes appeared on the edge of rice leaves 4、水稻受到白叶枯病的影响,叶子边缘有褐色病斑 Rice is affected by bacterial blight,with brown spots on the edge of the leaves
	1、水稻受稻瘟病影响,病株叶片有淡黄色晕斑 Affected by rice blast,the leaves of infected rice plants have light yellow halo spots 2、水稻患有细菌性条斑病 Rice suffers from bacterial streak disease 3、水稻感染细菌性条斑病,出现黄色水渍状斑点 Rice infected with bacterial streak disease,yellow water-stained spots appear 4、水稻感染细菌性条斑病,水稻叶片发黄枯萎 Rice infected with bacterial streak disease,rice leaves yellow wilt
	1、飞虱有油性光泽 Rice planthoppers have an oily sheen 2、蓟马的头上有 2 个触角 Thrip has two antennaes on the head 3、蓟马的头上有触角 Thrip has tentacles on the head 4、蓟马的头上有一对线性触角 Thrip has a pair of linear antennae on the head
	1、水稻受胡麻斑病影响,病株出现大的不规则褐斑 Rice was affected by brown spot disease,with large irregular brown spots on infected plants 2、水稻患稻瘟病,叶脉间有褐色斑点 Rice suffers from rice blast,there are brown spots between the veins 3、水稻患稻瘟病,叶脉间有针头大小褐斑 Rice suffers from rice blast,and there are pinhead sized brown spots between the veins 4、水稻患稻瘟病,植株有针头大小的褐色斑点 Rice suffers from blast,and the plant has brown spots about the size of a pinhead
	1、水稻受稻瘟病影响,病株叶片有淡黄色晕斑 Affected by rice blast,the leaves of infected rice plants have light yellow halo spots 2、水稻有稻曲病,穗部有黑色菌核 Rice has rice false smut disease,ear has black sclerotium 3、水稻受稻曲病的影响,导致谷物中有黑色菌核 Rice is affected by rice false smut disease,resulting in black sclerotium in the grain 4、水稻稻曲病产生黑菌核 Black sclerotium occurs in rice false smut
	1、水稻受稻瘟病影响,水稻叶片有淡黄色晕灰色霉层 Affected by blast disease,rice leaves have pale yellow halo gray mold layer 2、水稻受病害,又细又高 The rice is thin and tall because of disease 3、水稻染上恶苗病,植株细高 The rice was infected with bakanae disease and the plants were slender and tall 4、水稻染上恶苗病,使病水稻又长又细 The rice is infected with bakanae disease,which makes the sick rice long and thin

注:1、2、3、4 分别对应 CNN_LSTM 模型、CNN_LSTM_ATT 模型、本文模型冻结语言模型模式和本文模型微调语言模型的自动描述结果。

Note:1,2,3,4 correspond to the automatic description results of the CNN_LSTM model,the CNN_LSTM_ATT model,the language model freezing mode of the proposed model,and the fine-tuning language model of the proposed model,respectively.

本文模型生成的描述语句更为丰富。例如第 1 张水稻感染了白叶枯病的图像,CNN_LSTM_ATT 模型生成的语句仅是对叶片病症的简单描述:“白叶枯病叶缘呈深褐色”,而本文模型生成了更详细完整的描述语句:“水稻受白叶枯病的影响,水稻叶片边缘出现褐色条纹;水稻受到白叶枯病的影响,叶子边缘有褐色病斑”。第 6 张图像 CNN_LSTM_ATT 模型仅描述了病症外观:“水稻受病害,又细又高”,未判断病症类型,本文模型生成的描述语句既有病症类型判断也病症外观描述:“水稻染上恶苗病,植株细高”。

本文模型能提取更多的图像特征。例如第 2 张图像 CNN_LSTM_ATT 仅描述了水稻患有细菌性条纹病的病症,本文模型补充了叶片发黄枯萎的具体描述。针对稻瘟病和白叶枯病都存在褐色病斑现象,模型能准确识别相应病症并分别进行了具体描述,表明本文模型对不同病害的相似症状具备区分能力。在部分图像中 CNN_LSTM_ATT 模型与本文模型一致或近似,例如第 3 张图像输入这 2 个模型后都得到了关于蓟马头上有触角的描述结果。

为进一步验证本文模型性能,使用 405 张实际稻田图片对模型进行测试,结果如表 4 所示。由于实际田间的场景更为复杂多样,生成的图像描述指标与数据集指标对比总体仅有轻微下降,但仍高于其他对比模型,证明本文模型具有较强的泛化能力,可对实际种植环境中的水稻病虫害实现有效描述。

表 4　实际稻田图片评估结果

Table 4　Results of actual rice field image evaluation

算法 Algorithm	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR
微调语言模型(本文)Fine-tuning language model(this paper)	0.53	0.37	0.29	0.24	1.54	0.51	0.31

在生成的图像描述中是否准确描述病虫害类别也是模型的重要考量。本文提取生成描述中的病症类别信息,统计分析了各个模型对病虫害类别的识别准确率,从图 6 结果可见,CNN_LSTM 模型准确率仅 37.28%,带注意力机制的 CNN_LSTM_ATT 准确率为 87.41%,本文模型的识别准确率达 97.28%,高于其他模型,表明本文模型具有良好的水稻病虫害识别能力。

本文模型对各类病虫害的识别准确率如表 5,对于白叶枯病和稻曲病的识别准确率最高,从单一类别来看,模型对各个病症的识别效果也均很好。

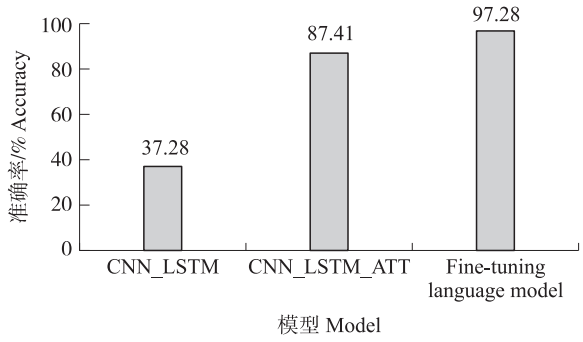


图 6　水稻病虫害总体识别准确率

Fig. 6　Overall accuracy of rice disease and pest identification

表 5　各类水稻病虫害识别准确率

Table 5　Identification accuracy of various rice pests and diseases

类别 Category	准确率/% Accuracy
白叶枯病 Rice bacterial blight	100.00
稻蓟马 Rice thrip	94.12
细菌性条斑病 Rice bacterial streak disease	97.96
恶苗病 Rice bakanae disease	96.30
三化螟虫 Rice three chemical borers	95.00
胡麻斑病 Rice brown spot	97.37
飞虱 Rice planthopper	97.29
稻瘟病 Rice blast	98.00
稻曲病 Rice false smut	100.00
纹枯病 Rice sheath blight	94.87

3　讨论与结论

当前图像描述研究主要基于英文开展,基于中文的研究较少,英文文本与中文文本的语言习惯有很大不同,例如分词方式、语法逻辑等,如果从英文文本直接翻译为中文文本,所对应的语言习惯仍为英文。为了使生成的中文图像描述语句更准确流畅,本文构建一个中文图像描述的模型,并使用中文标注的数据集进行训练。进一步提升了模型在实际场景中的应用。

为弥合图像和文本间的语义鸿沟,本文采用一种轻量化的图像描述方法,基于传统的编码器-解码器框架,编码端使用多模态预训练模型 CLIP,CLIP 模型抽取的图像特征含有丰富的语义信息,解码端使用预训练语言模型 GPT-2,通过训练简单的映射网络,使语言模型生成图像描述语句。根据试验结果,本文方法的总体性能都高于基础的 CNN_LSTM 模型以及加入注意力的模型。生成的图像描述不仅能实现水稻病虫害的类别诊断,更能加深对水稻具体病害症状的认识。

常用图像描述数据集的数据量庞大且内容丰富,包含多样的生活场景以及详细的语言描述。本文仅探究了常见的7种水稻病害和3种水稻虫害,且描述文本相对简单,因此相较于公共数据集,本文的研究对象较为局限,泛化性和多样性欠佳,将考虑扩大数据集规模,引入更加丰富的农业生产场景,提升模型在不同场景中的描述性能。此外,近年随着越来越多超大规模预训练模型的兴起以及大规模算力的投入,预训练模型的性能不断提升,本文采用的预训练模型较为单一,未来将引入更先进的预训练模型,进一步提升目标任务的效率,提升文字描述的精准性、多样性。本文模型主要是在电脑平台上运用,后续计划将该模型部署到移动设备上,应用于实际生产,使得图像描述在实际农业生产领域发挥更大价值。

综上,本文研究探索了图像描述技术在水稻病虫害识别领域的应用,水稻病虫害图像描述技术可为水稻病虫害的诊断提供有效依据。

1)构建了一个水稻病虫害图像中文描述模型,使用中文标注的水稻病虫害数据集进行训练,该模型生成的文本符合中文语法与语言习惯,并能合理表达描述图像内容。

2)预训练模型已在大量数据上进行训练,从中提取出尽可能多的共性特征,利用预训练模型可以让特定的学习负担变轻;利用多模态预训练模型抽取的图像特征不仅包含基础的图像信息还有丰富的语义信息,可以有效弥合图像描述任务的语义鸿沟问题。

3)本文基于多模态预训练模型的图像描述算法总体性能指标优于其他模型,能在少量训练资源的情况下,完成对水稻病虫害图像的病害类别判断以及生成详细的病症描述语句。

参考文献 References:

[1] 王艳青. 近年来中国水稻病虫害发生及趋势分析[J]. 中国农学通报,2006,22(2):343-347.
Wang Y Q. Incidence and trend analysis of rice pests and diseases in China in recent years[J]. Chinese Agricultural Bulletin,2006,22(2):343-347(in Chinese with English abstract).

[2] 刘杰,曾娟,杨清坡,等. 2023年农作物重大病虫害发生趋势预报[J]. 中国植保导刊,2023,43(1):32-35.
Liu J,Zeng J,Yang Q P,et al. Forecast of occurrence trend of major crop pests and diseases in 2023[J]. Chinese Plant Protection,2023,43(1):32-35(in Chinese).

[3] 杨堃,范习健,薄维昊,等. 基于视觉加强注意力模型的植物病虫害检测[J]. 南京林业大学学报(自然科学版),2023,47(3):11-18.
Yang K,Fan X J,Bo W H,et al. Plant disease and pest detection based on visiual attention enhancement[J]. Journal of Nanjing Forestry University(Natural Sciences Edition),2023,47(3):11-18(in Chinese with English abstract).

[4] 陆舟,沈明霞,刘龙申,等. 基于轻量化网络与注意力机制的育肥猪采食行为识别方法研究[J]. 南京农业大学学报,2023,46(4):802-812. DOI:10.7685/jnau.202208023.
Lu Z,Shen M X,Liu L S,et al. Study on feeding behavior recognition method of fattening pigs based on lightweight networks and attention mechanism[J]. Journal of Nanjing Agricultural University,2023,46(4):802-812(in Chinese with English abstract).

[5] Ren S Q,He K M,Girshick R,et al. Faster R-CNN:towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(6):1137-1149.

[6] 徐会杰,黄仪龙,刘曼. 基于改进YOLOv3模型的玉米叶片病虫害检测与识别研究[J]. 南京农业大学学报,2022,45(6):1276-1285. DOI:10.7685/jnau.202110039.
Xu H J,Huang Y L,Liu M.Research on pest detection and identification of corn leaf based on improved YOLOv3 model[J]. Journal of Nanjing Agricultural University,2022,45(6):1276-1285(in Chinese with English abstract).

[7] Young T,Hazarika D,Poria S,et al. Recent trends in deep learning based natural language processing[J]. IEEE Computational Intelligence Magazine,2018,13(3):55-75.

[8] Otter D W,Medina J R,Kalita J K. A survey of the usages of deep learning for natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems,2021,32(2):604-624.

[9] Li H. Deep learning for natural language processing:advantages and challenges[J]. National Science Review,2018,5(1):24-26.

[10] 高雨亮,徐向英,章永龙,等. 融合分组注意力机制的水稻病虫害图像识别算法[J]. 扬州大学学报(自然科学版),2021,24(6):53-57.
Gao Y l,Xu X Y,Zhang Y l,et al. Image recognition algorithm of rice diseases and insect pests based on Shuffle attention mechanism[J]. Journal of Yangzhou University(Natural Science Edition),2021,24(6):53-57(in Chinese with English abstract).

[11] 周维,牛永真,王亚炜,等. 基于改进的YOLOv4-GhostNet水稻病虫害识别方法[J]. 江苏农业学报,2022,38(3):685-695.
Zhou W,Niu Y Z,Wang Y W,et al. Rice pests and diseases identification method based on improved YOLOv4-GhostNet[J]. Journal of Jiangsu Agricultural Sciences,2022,38(3):685-695(in Chinese with English abstract).

[12] 许童羽,赵冬雪,周云成,等. 基于word2vec和Attention-Seq2Seq的水稻病虫害智能问答方法研究[J]. 沈阳农业大学学报,2019,50(3):378-384.
Xu T Y,Zhao D X,Zhou Y C,et al. Research on method of intelligent Q & A for rice pests and diseases based on word2vec and Attention-

Seq2Seq[J]. Journal of Shenyang Agricultural University,2019,50(3):378-384(in Chinese with English abstract).

[13] Kulkarni G,Premraj V,Ordonez V,et al. BabyTalk:understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(12):2891-2903.

[14] Kuznetsova P,Ordonez V,Berg T L,et al. Treetalk:composition and compression of trees for image descriptions[J]. Transactions of the Association for Computational Linguistics,2014,2(9):351-362.

[15] Xu K,Ba J L,Kiros R,et al. Show,attend and tell:neural image caption generation with visual attention[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. July 6-11,2015,Lille,France. ACM,2015:2048-2057.

[16] Chen X L,Fang H,Lin T Y,et al. Microsoft COCO captions:data collection and evaluation server[J/OL]. arXiv Preprint:1504.00325,2015.

[17] Plummer B A,Wang L,Cervantes C M,et al. Flickr30k entities:collecting region-to-phrase correspondences for richer imageto-sentence models[J]. International Journal of Computer Vision,2017,123(1):74-93.

[18] Li X L,Liang P. Prefix-tuning:optimizing continuous prompts for generation[J/OL]. arXiv Preprint:2101.00190,2021.

[19] Mokady R,Hertz A,Bermano A H. Clipcap:clip prefix for image captioning[J/OL]. arXiv Preprint:2111.09734,2021.

[20] 温超,耿国华. 基于内容图像检索中的“语义鸿沟”问题[J]. 西北大学学报(自然科学版),2005,35(5):536-540.
Wen C,Geng G H. Review and research on “semantic gap” problem in the content based image retrieval[J]. Journal of Northwest University (Natural Science Edition),2005,35(5):536-540(in Chinese with English abstract).

[21] 章烈辉,刘占山,肖启明,等. 我国水稻病虫害综合防治技术研究现状及发展建议[J]. 中国稻米,2009(1):6-9.
Zhang L H,Liu Z S,Xiao Q M,et al. Research status and development suggestions on integrated control technology of rice diseases and pests in China[J]. Chinese Rice,2009(1):6-9(in Chinese with English abstract).

[22] Radford A,Kim J W,Hallacy C,et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. New York:PMLR,2021:8748-8763.

[23] Vaswani A,Shazeer N,Parmar N,et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach:NIPS,2017:5998-6008.

[24] Radford A,Wu J,Child R,et al. Language models are unsupervised multitask learners[J]. OpenAI blog,2019,1(8):9.

[25] Papineni K,Roukos S,Ward T,et al. BLEU:a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphi,US:Association for Computational Linguistics,2002:311-318.

[26] Banerjee S,Lavie A. METEOR:an automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Michigan,USA:Association for Computational Linguistics,2005,65-72.

[27] Lin C Y. ROUGE:a package for automatic evaluation of summaries[C]//Proceedings of the Workshop on Text Summarization Branches Out. Barcelona,Spain:Association for Computational Linguistics Press,2004:74-81.

[28] Vedantam R,Zitnick C L,Parikh D. CIDEr:consensus-based image description evaluation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Boston,MA,USA. IEEE,2015:4566-4575.

责任编辑:沈　波