基于文本提示增强 CLIP 的苹果叶片病害诊断模型

张瑞程 22354189

2024 年 12 月，广东 · 深圳

摘 要: 本项目提出了一种新颖的基于文本提示增强的对比语言-图像预训练（CLIP）模型，用于苹果叶片病害诊断。我们首先利用大型语言模型（LLM）生成与苹果病害相关的病理描述，并将其作为文本提示集成到 CLIP 模型中，以增强模型对病害特征的理解能力。通过引入 LoRA（Low-Rank Adaptation）技术，我们对 CLIP 模型的关键层进行了微调，以适应特定的疾病分类任务。实验结果表明，经过文本增强和 LoRA 微调的 CLIP 模型在多标签分类任务中表现出了优异的性能，精确率和召回率均有显著提升。此外，我们还采用了 t-SNE、CAM 等多样的可视化分析方法，对预测结果、中间层特征、可解释性进行了全面的评估，进一步验证了模型的可解释性和准确性。本研究为农业病害诊断领域提供了一种新的技术途径，并为预训练多模态模型在专业领域的应用提供了有价值的参考。

关键词: CLIP，细粒度图像分类，多标签分类，植物疾病诊断，多模态。

1 引言

苹果是世界上最重要的温带水果作物之一。叶面病害对苹果园的整体生产力和质量构成重大威胁。目前苹果园的病害诊断过程是基于人工侦察，这既费时又昂贵。近年来，深度学习彻底改变了计算机视觉领域，现在正在成为许多应用程序的标准工具。已经开发了许多基于深度学习的植物病害自动诊断技术，旨在支持农民并减少植物生产力方面的损失 [1]。

然而，苹果叶片病害诊断是一项具有挑战性的任务。不同苹果品种的单一疾病视觉症状的巨大差异。这些变化源于自然环境和图像捕获环境的差异，例如，叶子的颜色、形态、感染程度、不均匀的图像背景以及成像过程中不同的光线照明等。同时，苹果叶片病害具有高度细粒度性，要求模型必须能够有效聚焦占整张图片面积很小的病灶区域，这对深度学习模型造成了很大的挑战。

近些年来，该领域的进展主要集中在设计更为先进的网络架构（如 YOLOv8[2]）、训练策略（如知识蒸馏）、损失函数（如 Arcfaceloss[3]）、数据增强方法（如 LeafGAN[4]）等。这些方法和策略已经取得了很大的成功，但仍存在几个关键问题。首先，作为一个典型的细粒度分类问题，准确识别到微小病灶并有效区分是十分困难的，在模型没有相关背景知识的情况下训练损失不易收敛。其次，实际的苹果叶片病害数据集通常在每个类别上的数量不平衡。虽然已经提出了几种技术来解决这个数据不平衡问题，但疾病分类模型通常偏向于具有更多样本和更高变化的类别。第三，过拟合问题在植物诊断任务中特别严重，因为提供诊断线索（即分类证据）的图像特征通常比一般物体识别问题中的特征小得多。我们发现，特别是在早期阶段，诊断线索可能仅由图像中的一个小点或微弱的皱纹组成。

对比维度	纯视觉模型	CLIP 模型
预训练知识	仅在大规模图像数据上进行视觉特征学习，知识局限于图像领域通用特征。仅依赖于单一的视觉模态信息，无法利用其他可能与疾病诊断相关的文本描述、环境因素（如种植地区的气候、土壤条件等文字信息）等额外信息来辅助诊断。	基于海量的文本 - 图像对进行预训练，不仅学习到丰富的视觉特征，还掌握了大量与文本相关的语义知识，能够理解和关联各种疾病名称等文本信息与视觉特征之间的关系，例如知晓“苹果锈病”“苹果黑斑病”等文本描述对应的叶片视觉特征模式，为疾病诊断提供更丰富的知识储备。
特征表示	主要提取图像的视觉特征，如颜色、纹理、形状等低层次特征和通过卷积神经网络学习到的一些抽象特征，但这些特征对于复杂的苹果叶片疾病情况，可能缺乏足够的语义信息来准确区分不同疾病类别，尤其是在疾病症状相似的情况下容易混淆。	能够将图像特征与文本特征进行联合学习和映射，生成具有更强语义表达能力的特征表示。通过对比图像与文本之间的相似性，可以更好地捕捉到叶片疾病的细微特征差异，并利用文本语义信息来辅助区分相似病症，例如准确区分“锈病早期”和“轻微的真菌感染”等易混淆的疾病状态，提高诊断的准确性。
零样本学习能力	缺乏零样本学习能力，需要在特定的苹果叶片疾病数据集上进行大量标注数据的训练才能进行诊断任务，对于新出现的未在训练集中的疾病类别难以进行准确判断。	具有强大的零样本学习能力，即使未在特定的苹果叶片疾病数据集上进行训练，凭借其预训练阶段学到的通用文本 - 图像关联知识，通过合理设计的文本提示（如“苹果叶片患有{疾病名称}病”），可以直接对从未见过的疾病类型进行初步诊断和分类，极大地拓展了模型的应用范围和适应性，能够快速应对新出现的病害情况。
应对数据不平衡问题	在面对苹果叶片疾病数据集中不同疾病类别样本数量不平衡的情况时，容易倾向于学习样本数量较多的疾病类别特征，而忽视样本较少的类别，导致在诊断少见疾病时准确率较低，模型的性能受到数据分布不均衡的严重影响。	通过文本提示和多模态信息的关联，可以在一定程度上缓解数据不平衡带来的问题。即使某些疾病类别在图像数据中较为少见，但通过丰富的文本描述和与其他常见疾病的对比学习，CLIP 模型能够更好地捕捉到这些少见疾病的独特特征，从而在诊断时给予它们合理的关注和准确的判断，提高对各类疾病的综合诊断能力，保障在实际农业生产中对各种可能出现的疾病情况都能进行有效的监测和诊断。

图 1: CLIP 模型与纯视觉模型在苹果叶片疾病诊断任务上的优势对比。对于本问题，CLIP 模型在预训练知识、特征表示、零样本学习能力、应对数据不均衡问题等方面具有巨大优势，这是我们选择使用 CLIP 模型和多模态方法的动机。

对比语言图像预训练模型（CLIP）[5] 和大语言模型（LLM）的出现为解决上述问题提供了全新的思路。首先，LLM 可以轻松地生成疾病相关生物学知识和病症描述，无需人类专家费时费力的标注。这些上下文信息在 CLIP 中与图像特征进行深度比对，能够为疾病监测提供更为精细的的指导。其次，CLIP 的对比学习策略本质上进行的是特征相似度对齐，天生对类别不均具有鲁棒性，甚至能够有效预测训练集中不存在的样本（Zero Shot 能力）。此外，这种通过文本建立监督的方式迫使模型的诊断线索更加聚焦细粒度病症，而非图像上的其他干扰特征。尽管 CLIP 模型在苹果叶片疾病诊断任务上展现出巨大的潜力，目前几乎没有工作在此方面进行探究。

受启发于多模态模型在分类任务上的出色表现，我们提出了一种新颖的文本提示增强的对比语言图像预训练（CLIP）模型，用于苹果叶片病害诊断。我们的模型使用 OpenAI 提出的经典 CLIP 框架，包括一个图像编码器和一个文本编码器。这两个编码器共同学习将图像和文本映射到同一个嵌入空间，使得相关的图像和文本在该空间中彼此接近，从而实现通过相似度匹配进行病害分类。在此基础上，我们进一步跟据病理学任务特点对模型进行了文本提示增强，使用大语言模型为不同病害类型建立相应的“病例”。“病例” 对每种病害的发病原因、病灶现象以及病理学原理进行了详细的分析，这种语义知识不仅能够帮助模型精确捕获病灶的特点，其病理性的解释内在编码了不同疾病之间的相关互关系，对多病症诊断提供了额外的指导。

文本提示中的语义信息，天然地蕴含了特征向量间的距离关系。以类别名称为例，”scab”（疮痂病）和”rust”（锈病）这两个词向量在语义空间中的距离，理论上应比”scab”与”healthy”（健康）之间的距离更近。这种语义亲近性为 CLIP 模型提供了丰富的类间关系信息，使其能够深刻理解不同病害之间的微妙差异。得益于这种深度的类间关系感知能力，CLIP 模型能够有效地区分和避免不恰当的类别组合，例如同时将一个样本错误地预测为”scab” 和”healthy”。同时，不同疾病之间也可能存在相关性，比如我们发现，在数据集中具有数量较多的同时感染 Scab (疮痂病) 和 Frog Eye Leaf Spot (蛙眼叶斑病) 的叶片，经查阅资料我们发现，这两种病害都是由真菌引起的，并且病原体在相似的环境条件下繁殖，如高湿度和适宜的温度。在我们的文本提示增强 CLIP 中，文本提示包含这类病理性的重要信息，能够有效帮助模型理解不同病症之间的相互关联。这种能力是传统独热编码方法所不具备的，后者在处理类别关系时往往缺乏这种细腻的语义理解。

```
# Map each category to its description for text prompts
category_descriptions = {
    "scab": "Caused by the fungus Venturia inaequalis, the initial symptoms on the leaves inclu
    "healthy": "Healthy apple leaves typically exhibit a vibrant green color, free from spots
    "frog_eye_leaf_spot": "This disease is caused by the fungus Cercospora. It manifests as rou
    "rust": "Rust disease is caused by the fungus Puccinia. It is characterized by the appearan
    "complex": "This term may refer to diseases caused by a combination of multiple pathogens o
    "powdery_mildew": "Caused by the fungus Erysiphe, this disease is identified by the presenc
    "scab,frog_eye_leaf_spot":"Simultaneously having Scab and Frog Eye Leaf Spot.",
    "scab,frog_eye_leaf_spot,complex": "Simultaneously having scab, complex disease and frog ey
    "frog_eye_leaf_spot,complex": "simultaneously having scab and complex disease.",
    "rust,frog_eye_leaf_spot": "simultaneously having rust and frog eye leaf spot.",
    "rust,complex": "simultaneously having rust and complex disease.",
    "powdery_mildew,complex": "simultaneously having powdery_mildew and complex disease.",
}
```

图 2: 病理提示文本展示。提示文本中涵盖各类苹果叶病害的发病原理分析、发病现象描述和不同阶段病情特点等信息。

在此项任务中，我们首先使用预训练 CLIP 对苹果叶片进行病害诊断，探究了 CLIP 模型强大的 Zero Shot 能力。然后，我们通过 LoRA 技术 [6] 通过在模型的特定层中引入低秩矩阵来分别微调模型的语言编码器和视觉编码器，使模型适应新的任务。具体实验细节详见下文分析。

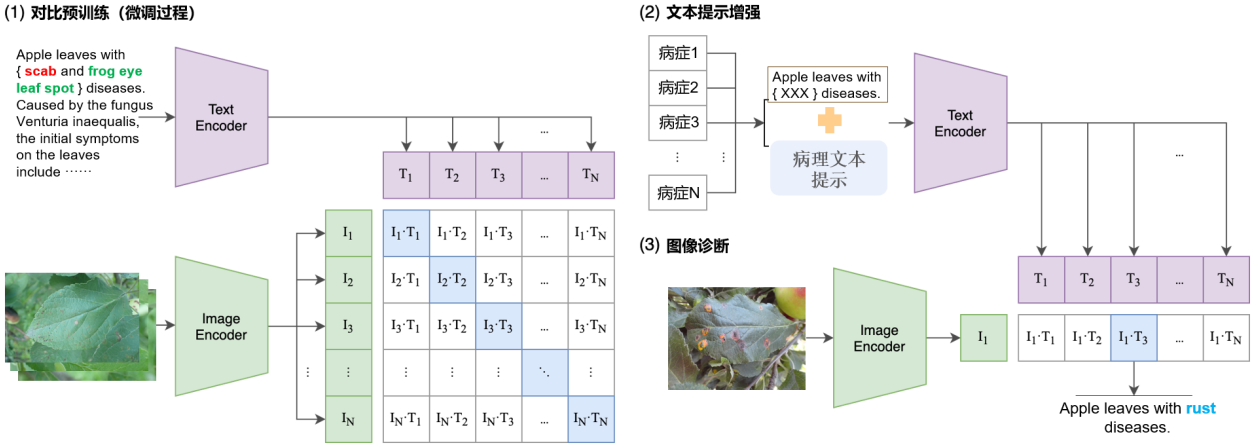

图 3: 模型架构。

2 相关工作

2.1 对比语言-图像预训练模型（CLIP）

受到对比预训练的启发，Radford 等人 [5] 提出了对比语言图像预训练（Contrastive Language Image Pre-training，简称 CLIP），这是一种从文本监督中学习可解释视觉表示的方法。与大多数仅依赖视觉信息的对比预训练方法 [7, 8] 不同，CLIP 综合利用了视觉和语言信息。具体而言，CLIP 将文本描述视为图像的语言维度，并假定这种描述与图像内容具有内在的一致性。基于此，CLIP 致力于在潜在空间中将配对的图像和文本表示尽可能地拉近，以实现图像-文本对的紧密对齐。这种对齐是通过 CLIP 的视觉编码器和文本编码器共同完成的，使得视觉编码器能够编码广泛的知识。得益

于图像与文本之间的精确对齐，CLIP 能够从文本监督中吸收丰富的知识，并已在多个领域显示出其效用，包括图像生成 [9, 10]、分割 [11, 12]、检测 [13, 14] 以及分类 [15, 16]。

近年来，CLIP 在疾病诊断领域也得到了越来越多的关注（本次植物疾病诊断可以大致归结为此类问题）。这一趋势可以归因于 CLIP 的独特能力，它能够将神经网络的表征与人类认知进行对齐，满足人工智能在诊断领域对可解释性的需求 [17]。相较于其他仅依赖视觉信息的模型，疾病诊断任务的特殊性在于它需要更深层次的专业知识，这对于纯视觉模型来说是一个难以逾越的障碍 [18]。传统的解决方案试图通过细粒度的注释，如边界框 [19, 20] 和分割掩码 [21, 22]，来提高模型的专业性，但这种方法不仅耗时而且费力，难以实现规模化扩展。与此相对，大型语言模型（Large Language Model，简称 LLM）通过整合网络知识和视觉理解，能够以较低的成本为诊断任务提供精确而细致的上下文知识。这些上下文提示不仅包含了丰富的语义信息，而且相较于细粒度标注，它们在成本和效率上具有明显优势。CLIP 模型通过大规模的图文对比学习，将视觉与语言信号嵌入到同一特征空间中，从而实现了图像与文本之间的高度对齐。CLIP 编码的人类级知识和人类级知识对应可以作为一些需要注释的任务（如植物疾病特征）的外部监督信号，从而提高视觉模型的性能。

2.2 植物病害诊断

植物病害的自动诊断是农业领域最活跃的研究领域之一。及时准确地检测植物病害对于确保全球粮食安全和农业生态系统的可持续性至关重要 [1]。得益于深度学习模型在图像识别和分类任务中的卓越表现，植物病害诊断领域中深度学习技术的应用正日益增多。与传统的人工诊断方法相比，基于计算机视觉的方法能够处理大量数据，并从中学习复杂的模式。具体来说，这些模型能够识别植物叶片上的病害症状，并将这些症状与已知的病害类型相匹配。这使得基于计算机视觉的模型在植物病害鉴定方面显示出前景。

Kawasaki 等人 [23] 训练了一个三层卷积神经网络（CNN），对来自真实农场的图像中的三种黄瓜病害类别进行诊断，其中目标对象在复杂背景下出现。Fuentes 等人 [24] 结合了三种目标检测方法，在广泛的西红柿数据集上同时进行疾病检测和诊断，实现了 86.0% 的平均精度。DeChant 等人 [25] 提出了一种自动系统，用于识别玉米植株田间采集图像上的北方叶枯病病变，并在测试集上实现了 96.7% 的准确率。Lu 等人 [26] 使用自然水稻图像数据集训练他们的系统以识别十种常见的水稻病害。在十折交叉验证策略下，所提出的 CNN 模型实现了 95.48% 的准确率。

3 问题分析

3.1 数据集介绍

Plant Pathology 2021 数据集是植物病理学领域一个具有重要价值的公开数据集，其在推动植物病害识别技术的发展中发挥着关键作用。该数据集主要聚焦于苹果树叶的病害情况，共包含 18632 张高质量的图像数据，全面涵盖了复杂多样的叶面病害类型及健康状态，具体涵盖了诸如 complex、frog_eye_leaf_spot、powdery_mildew、rust、scab 等 12 种苹果叶面病害及多种复合病害类型，能够精准地反映出实际农业生产中所面临的复杂病害情形。在本次实验中，我们使用划分好的 3000 张图片作为训练集，600 张图片作为测试集，600 张图片作为验证集。

图 4: 数据集展示。

3.2 任务挑战

基于图像的植物诊断是一项具有挑战的任务，因为需要细粒度的物体识别。一般来说，像 CNN 或 transformer 这样的深度分类器倾向于捕获大面积的图像特征（亮度和颜色），而不是可能指示疾病的微弱特征。此外，当使用分为训练集、验证集和测试集的数据（其中应用交叉验证）评估分类器时，数据集中的“潜在相似性”（如背景、亮度和/或目标与相机之间的距离）起到积极偏差的作用，通常仅提高表面的诊断准确率，而在其他未知环境下评估的准确率变得非常低 [4]。例如，在从广角图像进行黄瓜病害诊断中，在同一农场的诊断性能在 F1 分数上显示为 86.0%，但在不同农场下降到 20.7% [4]。

之前许多工作尝试使用卷积神经网络、变换器解决这些挑战，它们在预测准确率上取得了优异的表现。随着网络架构的不断优化、训练策略的不断提升、数据增强策略的广泛普及，本细粒度分类问题已得到初步的解决，3 年前的 SOTA 方案（如 SwinTransformer [27]）就已经可以达到超过 95% 的准确率。这些工作的代码均可以在网络上找到，因此，单纯准求预测准确率的提升已没有太多意义和提升空间，我希望通过本次项目进行一种全新的尝试。

4 方法

4.1 模型原理

对比语言-图像预训练（CLIP）是一种由 OpenAI 开发的预训练方法，旨在弥合图像与文本之间的鸿沟。它联合优化视觉编码器和文本编码器，确保图像-文本对在一个共享的潜在空间中紧密对齐。与其他依赖广泛人工监督或复杂结构的方法不同，CLIP 遵循奥卡姆剃刀原则[1] [28]。

训练：CLIP 模型通过对比预训练（Contrastive pre-training）优化，其负责对图像与文本对进行精确对齐。具体来说，在给定批量大小（Batch size）为 N 的情况下，可以构造出 N^2 个图像-文本对，其中包括 N 个匹配的图像-文本对（正样本

[1]奥卡姆剃刀原则（Occam's Razor）是一种解决问题和理论构建的方法论原则，它建议在所有可能的解释中，最简单的解释（即假设最少的解释）往往是最正确的。在机器学习和人工智能领域，奥卡姆剃刀原则有着重要指示意义，在选择模型时，如果一个简单的模型和一个复杂的模型具有相似的性能，那么通常会倾向于选择更简单的模型，因为它更易于理解和维护，也更不容易过拟合。

对，如图 2 所示以蓝色突出显示），以及 $N^2 - N$ 个不匹配的图像-文本对（负样本对）。图像编码器的预训练目标函数定义如下：

$$L_{\text{img}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\Phi\left(I_i, T_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\Phi\left(I_i, T_j\right)/\tau\right)}, \tag{1}$$

其中，$\Phi(\cdot,\cdot)$ 表示余弦相似度，τ 是一个可学习的温控参数，I_i 和 T_i 分别代表第 i 个图像嵌入向量和文本嵌入向量。文本编码器的目标函数定义与之对称：

$$L_{\text{txt}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\Phi\left(T_i, I_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\Phi\left(T_i, I_j\right)/\tau\right)}. \tag{2}$$

CLIP 模型的总体优化目标是通过对图像编码器和文本编码器的目标函数取平均值来实现的：

$$\mathcal{L}_{\text{total}} = \frac{\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}}}{2}. \tag{3}$$

这种优化策略确保了模型在苹果病害诊断任务中，能够有效地学习到叶片病灶和病例文本之间的关联性，从而提高病害识别的准确性和鲁棒性。通过对比学习，CLIP 模型能够从大量的图像特征中提取出稀疏而关键的病症信息，这对精细化分类提供了有益的指导。

推理：在预训练阶段，CLIP 经过专门训练，旨在精准预测图像与文本描述是否匹配，这一特性使其在具有文本提示的分类任务中展现出天然的优势。其核心原理是通过细致对比图像嵌入和文本嵌入来达成分类目的。具体而言，设 I 为给定苹果叶片图像 x 经图像编码器提取得到的图像特征，而 $\{W_i\}_{i=1}^{K}$ 则是由文本编码器生成的类别嵌入集合，其中 K 代表类别总数。每一个 W_i 均源自于特定的文本提示，例如 “一张具有 [CLASS] 病的苹果叶片照片 + 病理信息”，在此处，类别标记会被具体的类别名称所替换，并且附加不同类别的病例信息作为本文增强，从而生成具有针对性的类别嵌入。预测概率的计算通过如下公式实现：

$$p(y = i|I) = \frac{\exp\left(\Phi\left(I, W_i\right)/\tau\right)}{\sum_{j=1}^{K}\exp\left(\Phi\left(I, W_j\right)/\tau\right)}, \tag{4}$$

在上述公式中，τ 是在预训练期间通过学习确定的温度参数，它对于概率计算的准确性和稳定性起着重要作用。而 $\Phi(\cdot,\cdot)$ 代表的是余弦相似性，通过这一指标来衡量图像特征与类别嵌入之间的相似程度，进而确定图像属于各个类别的概率分布。

与传统分类器的学习方式相比，传统方法通常是从零开始学习封闭集的视觉概念，这种方式往往局限于预先定义好的有限类别集合，缺乏对新类别和未知概念的拓展能力。而 CLIP 的预训练机制则别具一格，它允许通过文本编码器对开放集的视觉概念进行探索。这意味着 CLIP 能够突破传统分类器的局限，接触到更为广泛的语义空间，涵盖了大量在预训练过程中接触到的各种文本描述所对应的视觉概念。这种广泛的语义覆盖使得 CLIP 学到的表示具有更强的通用性和灵活性，从而在迁移到下游多标签分类任务时更加自然和高效。

4.2 网络架构

CLIP 无缝地将视觉模型与语言模型集成在一起。视觉编码器可以基于 ResNet [29] 或视觉变换器（ViT）[30]，而语言编码器则基于类似 BERT [31] 的基于变换器的模型。它在每次迭代中接收一批苹果叶片图像及其相应的病理文本描述作为输入。经过编码过程后，嵌入被标准化并映射到联合图像-文本潜在空间。也就是说，输入的图像和文本分别被编码为 $I \in R(N\times D)$ 和 $T \in R(N\times D)$，其中 N 表示批量大小，D 代表嵌入维度。

在本次实验中，我们尝试了三种不同的视觉编码器配置，分别为 RN50×4，ViT-B/16 和 ViT-B/32。其中，RN50×4 基于 ResNet-50 架构，并通过 EfficientNet[32] 的扩展规则进行了扩展，将标准的 ResNet-50 模型的规模增大了 4 倍。ViT-B/16 和 ViT-B/32 是 Vision Transformer（ViT）模型的常用变体，其中”B” 代表基本大小（Base），而数字指的是输入图像被分割成 16×16 和 32×32 像素大小的 patches，具体细节请详见 [30]。对于语言编码器，我们选择使用了 chatgpt-2 架构，其封装在 OpenAI 的 CLIP 库中，词汇量为 49408，最大上下文长度为 77。

4.3 LoRA 迁移学习

OpenAI 的 CLIP 模型在多个大规模数据集上进行了预训练，这些数据集包括但不限于 LAION-400M [33]、LAION-5B [34] 等。这些数据集为 CLIP 提供了丰富的图像和文本对，使其能够在对比学习框架下学习到图像和文本之间的关联性。在实验中我们发现，预训练好的 CLIP 模型已经具备一定的零镜头诊断能力，能够大致区分一些病害类型，这得益于我们有效的文本增强方法，具体效果详见实验部分。

为了使模型能够更好的适应任务特点，我们使用 LoRA（Low-Rank Adaptation）[6] 技术在 Plant Pathology 2021 数据集上对预训练权重进行了微调。LoRA 技术可以无缝集成到现有的神经网络架构中，通过将模型的大权重矩阵分解为两个低秩矩阵，显著减少了需要调整的参数数量，从而实现了对预训练模型的轻量级微调。这种方法不仅可以降低计算复杂度和内存使用，还能帮助防止过拟合，特别是在训练数据有限的情况下。

在代码中，LoRA 技术被应用于 CLIP 模型的特定层，以实现对模型的低秩调整。具体来说，在保持预训练模型权重不变的情况下，通过向卷积层和 Transformer 层添加低秩矩阵来实现参数的高效更新。

LoRALayer_vit 类：该类是对 Vision Transformer（ViT）中的自注意力层进行低秩调整的实现。它通过在原始的自注意力层上添加两个低秩矩阵（rank_down 和 rank_up），并在前向传播中计算这两个矩阵的乘积来更新自注意力层的权重。这种设计允许模型在保持大部分预训练权重不变的同时，通过调整低秩矩阵来适应新的任务。初始化时，矩阵 A 通过高斯函数初始化，矩阵 B 为零初始化，使得训练开始之前旁路对原模型不造成影响，即参数变化量为 0。对于权重的输入 x，输出计算为

$$h = W_0 x + BAx, \tag{5}$$

其中 W_0 是原始权重，BA 是低秩矩阵的乘积。

```
self.lora_layers = nn.ModuleList([LoRALayer_vit(layer) for layer in layers_to_replace])
for i in range(len(layers_to_replace)):
    self.clip_model.transformer.resblocks[i].attn = self.lora_layers[i]
```

图 5: LoRALayer_vit 的应用。首先，为 layers_to_replace 中的每个层创建一个 LoRALayer_vit 实例，并将这些实例收集到一个 nn.ModuleList 容器中。然后遍历 layers_to_replace 中的每个索引 i，并将 CLIP 模型的 Transformer 部分中的第 i 个残差块（resblocks）的自注意力层（attn）替换为 lora_layers 列表中对应的 LoRALayer_vit 实例。

LoRAConv2d 类：该类是对卷积层进行低秩调整的实现。它通过在原始卷积层的权重上添加一个低秩矩阵（通过两个参数 A 和 B 计算得到）来更新卷积层的权重。这种方法同样允许模型在保持预训练权重的基础上，通过调整低秩矩阵来适应新的任务。在 LoRAConv2d 中，低秩矩阵的计算和应用方式与 LoRALayer_vit 类似，但针对的是卷积操作。

这两种 LoRA 实现方式都是通过在模型的关键层中引入低秩矩阵来完成微调，而不是直接修改预训练模型的权重。这种方法在保持模型预训练知识的同时，能够更高效地适应叶片诊断的特异化任务需求。

```
self.lora_layers = []
for layer_name, block_idx in layers_to_replace:
    layer = getattr(self.clip_model, layer_name)[block_idx]

    layer.conv1 = LoRAConv2d(layer.conv1)
    layer.conv2 = LoRAConv2d(layer.conv2)
    layer.conv3 = LoRAConv2d(layer.conv3)

    getattr(self.clip_model, layer_name)[block_idx] = layer

    self.lora_layers.extend([layer.conv1, layer.conv2, layer.conv3])
```

图 6: LoRALayer_conv 的应用。首先，使用 getattr 函数和 block_idx 索引来获取 CLIP 模型中指定的原始层。然后，将 LoRAConv2d 类应用于该层的三个卷积层（conv1, conv2, conv3）。这意味着每个卷积层都会被一个 LoRA 层包裹，从而在不改变原始卷积层权重的情况下，通过添加低秩矩阵来调整其行为。最后，将经过 LoRA 调整的层重新赋值给原始模型的相应位置。这样，当模型进行前向传播时，这些层将使用 LoRA 层来进行计算。

在添加 LORA 后，我们统计了可训练参数的占比，如表1所示。可以看出，我们仅需更新约 0.61% 的参数，即可实现对大模型的微调，这为我们在 8 张 NVIDIA RTX 3090 GPU 上训练模型提供了可能。

表 1: 可训练参数统计

Trainable params	All params	Trainable rate
531462	86724102	0.6128192598638842%

4.4 实施细节

文件说明 本项目目录下存在 3 个 python 文件：generate_text_feature_npys.py，fine_tune.py 和 reference.py。其中，generate_text_feature_npys.py 用于将大模型生成的病理描述编码为词向量，并以 np 数组的形式进行保存。fine_tune.py 用于在本任务数据集上对 CLIP 进行微调。它首先会从 OpenAI 提供的 CLIP 库中加载预训练好的 CLIP 模型，并使用 LORA 技术对其中的关键层附加低秩矩阵。然后建立本数据集的 Dataset 类，使用成对的图片和文本进行训练。reference.py 用于推理与测试过程，给定一个图片，它会使用所有类别构建出多个不同的文本提示，CLIP 通过计算图片与每种文本提示间的余弦相似度判定图片中病害的类别。

运行流程 微调过程（训练过程）：使用大模型生成病理提示——> 建立病理提示的词向量编码——> 建立输入图像的视觉编码——> 计算文本向量与视觉向量余弦的余弦相似度——> 使用 LORA 更新损失。推理过程：建立输入图像的视觉编码——> 建立所有类别的文本提示词向量编码——> 计算计算文本向量与视觉向量余弦的余弦相似度——> 找到最匹配的类别。

5 实验

5.1 数据预处理

5.1.1 文本提示

在实验过程中，我们率先借助 ChatGPT O1 大模型针对 6 种基础类型展开了病理描述性分析，具体内容如图 7 所示：

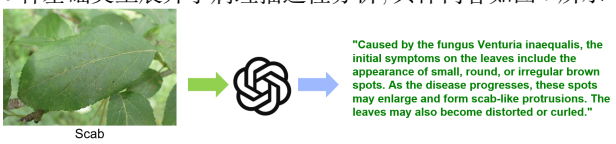

图 7: 文本提示展示。

针对多种疾病的组合情形，我们在文本提示中采用了对各类疾病进行简单组合描述的方式。

5.1.2 独热编码

Plant Pathology 2021 数据集有以下 12 个类别，不同的类别之间有重复，例如 [scab] 存在于多个类别中，这可能会影响模型的训练精度。因此，我们选择将本问题转换成多标签分类问题，对标签进行 one-hot 编码（如下图所示）再进行训练。

```
1  path,scab,healthy,frog_eye_leaf_spot,rust,complex,powdery_mildew
2  8a4195529f6b95e3.jpg,0,0,0,0,1
```

图 8: 独热编码。

5.1.3 数据增强

数据增强是一种用于提升机器学习模型泛化能力的技术。在训练过程中，它通过对输入数据进行随机变换，以此模拟出更丰富的数据多样性。

在我们的代码中，采用了多种数据增强方法：

1. **随机垂直翻转（RandomVerticalFlip）**：以 0.3 的概率对图像进行垂直翻转操作。这种操作有助于模型学习到图像在垂直方向上的上下文对称性。

2. **随机水平翻转（RandomHorizontalFlip）**：同样以 0.3 的概率对图像进行水平翻转。这一操作能让模型具备对图像左右翻转的不变性特征。

3. **随机旋转（RandomRotation）**：对图像进行随机角度的旋转，旋转角度设定为 15 度。这有助于模型适应图像的旋转变化，从而具备对图像旋转的不变性。

4. **颜色抖动（ColorJitter）**：对图像的亮度、对比度和饱和度进行随机调整。通过这种方式，模型能够在面对不同光照条件和颜色变化时，具备更强的鲁棒性。

上述这些数据增强技术被应用于训练数据，进而提升模型的泛化能力。在 create_train_transform 函数中，这些增强操作被整合进一个基于验证集（val_transform）的转换序

列中。随后，利用这个经过增强处理的转换序列来生成训练集（train_transform）。

如图9所示，在训练过程中，这些数据增强技术会随机应用于每个训练批次的图像上。这样一来，无需扩大实际数据集的规模，就能增加数据的多样性，进而有效防止模型出现过拟合现象，并提升模型对新数据和未见过的数据的适应能力。

```
1 usage
def create_train_transform(val_transform: tsfm.Compose) -> tsfm.Compose:
    train_transform_list = copy.deepcopy(val_transform.transforms)
    train_transform_list.insert(2, tsfm.RandomVerticalFlip(p=0.3))
    train_transform_list.insert(3, tsfm.RandomHorizontalFlip(p=0.3))
    train_transform_list.insert(4, tsfm.RandomRotation(degrees=15))
    train_transform_list.insert(5, tsfm.ColorJitter(brightness=0.2, contrast=0.2, saturation=0.2))
    train_transform = tsfm.Compose(train_transform_list)
    return train_transform
```

图 9: 数据增强。

5.2 实验设置

训练设置 我们按照老师提供的数据集划分方式，train 文件夹下的 3000 张图片作为训练集，text 文件夹下的 600 张图片作为测试集。训练过程使用 AdamW 优化器和余弦退火学习率调度器，学习率为 $5e^{-5}$，训练轮数为 50 轮，所有实验在 8 张 NVIDIA RTX 3090 GPU 上进行。

评估指标 我们选取精确率（Precision）和召回率（Recall）作为评估指标，并通过 AUROC（Area Under the Receiver Operating Characteristic Curve）指标衡量模型在所有分类阈值上的性能。

精确率（Precision）用于衡量模型预测为正类别中实际为正类别的比例，计算方法为：$\text{Precision} = \frac{TP}{TP+FP}$，即正确预测的正例数（TP）除以模型预测的正例总数（TP+FP）。

召回率（Recall）用于衡量所有实际为正类别中被模型正确预测的比例，计算方法为：$\text{Recall} = \frac{TP}{TP+FN}$，即正确预测的正例数（TP）除以实际的正例总数（TP + FN）。

ROC 曲线是真正例率（TPR）与假正例率（FPR）之间关系的图形表示。由于本问题为多分类问题，对每个类别，分别绘制 TPR 与 FPR 的曲线，然后计算曲线下的面积。AUROC 值范围从 0 到 1，值越接近 1 表示模型性能越好。

5.3 zero shot 实验

鉴于 CLIP 本身具备极为强大的零镜头泛化能力，我们首先选择直接运用预训练模型的参数来开展叶片病害诊断工作。在这一过程中，我们采用了与文献 [5] 所描述类似的提示集合策略，具体而言，就是使用形如 “Apple leaves with { · · · } diseases.” 这样的提示语句，并且采用相同的提示模板集。针对每一个类别名称，我们对所有模板对应的文本嵌入进行平均计算，进而得到平均文本嵌入值。后续，我们借助这些平均嵌入值来衡量每一张测试图像与各类别嵌入值之间的相似度。不同视觉编码器的测试结果表2所示。

表 2: zero shot 实验结果

视觉编码器架构	$RN50 \times 4$	$ViT-B/16$	$ViT-B/32$
精确率	21.32%	44.75%	51.02%

令人感到意外的是，尽管该模型从未在当前所使用的数据集中进行过训练，但仅仅依靠我们所采用的文本增强策略，竟然最高能够达到 51.02% 的精准率。这归功于 CLIP 卓越的文本 - 图像匹配机制和其在预训练阶段所积累的广泛而通用的知识表示。在预训练过程中，CLIP 接触了海量来自不同领域的文本 - 图像对，使其能够学习到丰富且具有高度概括性的语义特征。通过我们精心设计的文本增强策略，CLIP 能够更好地将叶片病害类别与预训练中习得的知识联系起来。例如，它能够依据病灶现象在文本空间中的语言描述表示（如：“· · · the appearance of yellow or orange, powdery, rust-like spots · · · (*Rust disease*)”），合理地推断出对应在图像空间中的特征，进而对测试图像进行较为准确的分类。这种能力不仅展示了 CLIP 在零 - 样本学习场景下的巨大潜力，也为在农业病害诊断领域利用预训练模型提供了新的思路和方法。

5.4 Fine-tune 模型实验

5.4.1 训练过程

我们进一步针对 Plant Pathology 2021 数据集，采用文本增强策略对 CLIP 模型展开了长达 50 轮的微调操作。在微调过程中，我们借助 LORA 技术来实现对原模型权重的调整。这种技术仅对原模型中的少量关键层权重进行修改，其优势在于能够确保预训练模型在已有的知识基础上，快速地实现损失收敛。考虑到算力资源的限制，我们仅仅选择在性能最为优越的 ViT - B/32 架构上进行这一微调操作。

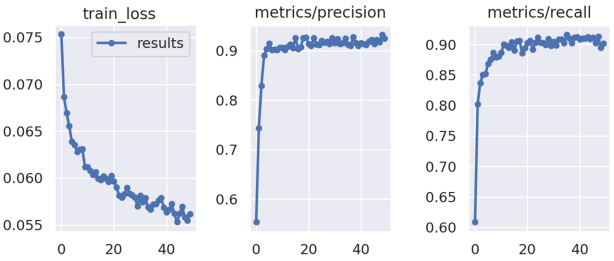

图 10: Fine-tune 训练过程展示。

图10清晰地呈现了在训练过程中，损失、精准率以及召回率随着训练轮数变化的趋势。从图中可以直观地看到，随着训练轮数的不断增加，训练损失呈现出平稳下降的态势。在这一过程中，精准率最高能够达到令人瞩目的 93.15%，与之相对应的召回率为 90.21%。值得注意的是，由于 CLIP 本身具备强大的零镜头能力，这使得在训练初期，精确率和召回率并非是从 0 开始逐步上升的。

5.4.2 结果分析

在追求模型性能最优的过程中，我们按照 AUROC 最大原则来对精准率和召回率进行均衡。通过这种方式，我们得到了最优的性能组合：精准率为 92.46%，召回率为 92.27%，此时能够实现最大的 AUROC 值，达到 85.31%。这一结果充分表明，通过 LORA 技术对 CLIP 模型进行微调，结合文本增强策略，在植物病理学数据集上能够取得非常优异的分类效果。同时，这也为利用预训练模型在特定领域进行高效微调提供了有力的实践依据，为后续在类似领域的模型优化和应用提供了有价值的参考方向。

5.4.3 可视化分析

混淆矩阵 针对微调后模型的诊断结果，我们绘制了相应的混淆矩阵。在该矩阵中，横坐标表示预测类别，纵坐标表示真实类别。通过观察图 11 可以发现，本模型能够对各类别进行有效预测。

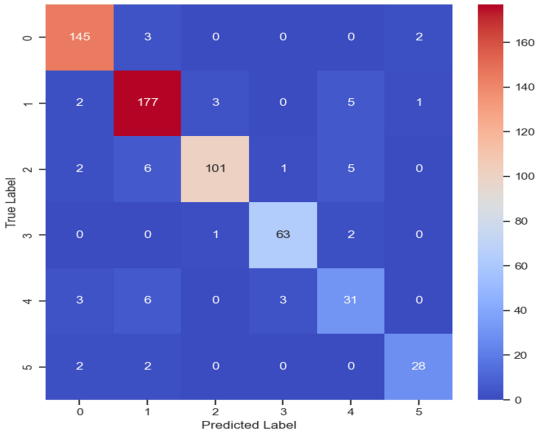

图 11: 诊断结果混淆矩阵。

t - SNE 可视化 我们同时对从 CLIP 微调前后所学习到的图像特征进行了 t - SNE（t-Distributed Stochastic Neighbor Embedding）可视化[2]对比。如图12所示，12 种不同的颜色分别代表了 12 种可能出现的叶片状态（包括各种病害状态及其组合）。通过对比可以清晰地发现，在对 CLIP 模型进行微调之后，图像特征在 t - SNE 空间中的分布呈现出更具优势的特性，具体表现为类边界更加清晰，簇质心更加明确。这意味着经过微调后，模型能够更好地区分不同类型的叶片状态，各类别之间的区分度得到了提高。

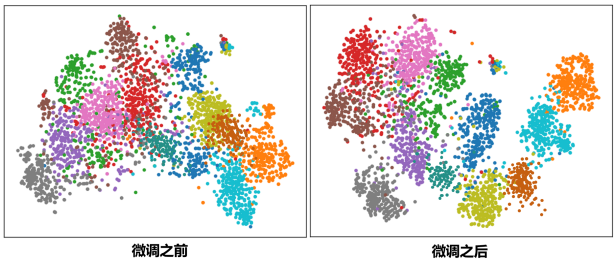

图 12: 微调前后 CLIP 学习的图像特征的 t-SNE 可视化对比。

特征图可视化 为了深入探究和验证模型在苹果叶片病害诊断任务中是否真正捕捉到了关键的病理特征，我们输出了 CLIP 视觉编码器最后一层的特征图（16 维）。通过观察图13中的特征图，我们可以清晰地看到每个通道都在识别图像的不同方面发挥着作用。例如，某些通道可能专注于边缘检测，从而帮助模型识别叶片轮廓和病灶的边界；另一些通道可能对颜色变化更为敏感，这使得模型能够区分健康叶片与病变区域之间的色差；还有一些通道可能专注于识别特定的斑点或纹理，这对于识别病害类型至关重要。

这些特征图的可视化结果不仅证实了 CLIP 视觉编码器能够有效地从图像中提取出丰富的视觉信息，而且还揭示了模型在识别苹果叶片病害时所依赖的关键视觉线索。特别是，一些通道对病灶斑点的显著关注，如第二行第二列的子图所示，这些通道的特征很可能是模型进行病害分类决策的重要依据。

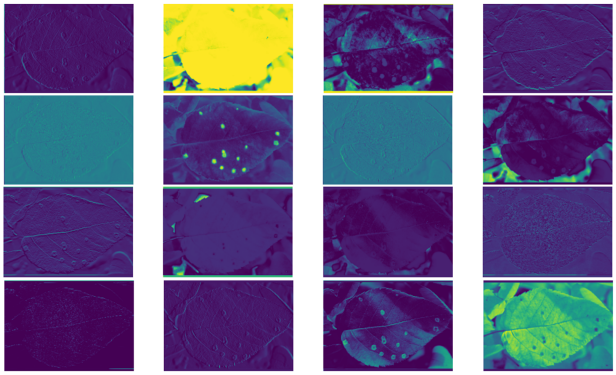

图 13: 视觉编码器特征图展示。

可解释性分析 在探究模型决策机制的过程中，为了深入验证模型是否能够精准地聚焦于叶片的病灶区域，从而以此为依据进行准确分类，我们采用了 CAM [35]（Class Activation Mapping，类激活映射）和 Grad - CAM [36]（Gradient-weighted Class Activation Mapping，梯度加权类激活映射）这两种极具价值的后验神经网络可解释性方法，对模型的焦点进行了全面且细致的可视化分析。

CAM 方法基于全局平均池化层的输出特征图与最终的分类权重之间的关联，通过计算每个特征图通道对于特定类别的重要性得分，生成相应的类激活映射图，以此直观地展示模型在进行分类决策时所关注的图像区域。而 Grad - CAM 则是通过计算目标类别相对于特征图的梯度，将这些梯度信息与特征图进行加权组合，从而生成可视化的显著图，突出显示模型在做出决策时所依赖的关键像素区域。

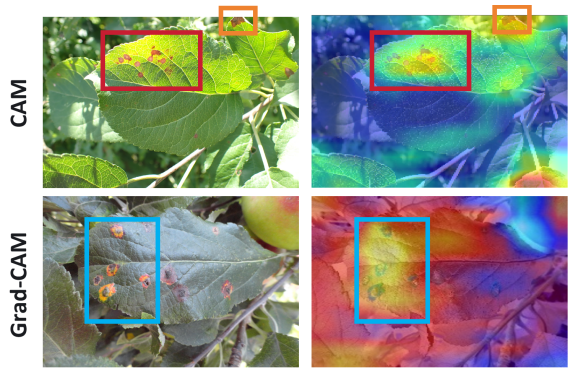

图 14: 模型可解释性可视化。

通过对图14的深入分析可知，模型在诊断过程中能够高效且精准地聚焦于叶片的病灶区域，这一现象有力地证明了模型所采用的诊断判据具有高度的合理性和正确性。模型并非是基于随机或无关的图像区域进行判断，而是能够准确地识别出与疾病相关的关键特征区域。

[2]t-SNE 是一种非线性降维算法，它通过特征间的概率分布将高维数据映射到低维空间（通常是二维或三维空间），以便能够进行可视化展示，同时尽可能保留数据在原始高维空间中的局部结构和相对关系。

5.5 消融实验

为验证我们所提出的病理文本增强策略的实际有效性，我们精心设计并开展了一系列消融实验，分别将该策略应用于预训练模型以及微调模型之上。我们设置了两组对比条件，即分别为 CLIP 模型增加病理文本提示以及不增加任何病理文本提示，以此来清晰地观察和衡量文本提示所带来的影响。所有参与实验的视觉编码器均统一采用了 ViT - B/32 架构，以确保实验条件的一致性和可比性。

通过对实验结果的细致分析，我们惊喜地发现，无论是对于预训练模型还是经过微调的模型而言，引入文本提示这一操作均能够有效地提升模型的预测精准率。进一步观察数据可以发现，这种提升效果在预训练模型上表现得更为显著。这一现象背后可能蕴含着多种深层次的原因：

一方面，预训练模型在其初始训练阶段虽然已经接触到了海量的通用图像和文本数据，但对于特定领域（如植物病理学）的专业知识和语义理解仍相对有限。当我们为其引入病理文本提示时，这些具有针对性的文本信息能够迅速填补其在专业领域语义理解上的空白，使其能够更加精准地捕捉到图像中与病理相关的特征信息，从而在预测过程中做出更为准确的判断，进而显著提升了预测精准率。

另一方面，预训练模型的参数在初始阶段尚未针对特定任务进行过多的调整和优化，具有较高的灵活性和可塑性。因此，当受到病理文本提示的引导时，模型能够相对容易地调整其内部的特征表示方式，以更好地适应病理分类这一特定任务的需求，从而使得文本提示的作用能够得到更加充分的发挥，最终导致预测精准率的大幅提升。相比之下，微调模型在经过一定程度的特定任务训练后，其参数已经在一定程度上朝着特定任务的最优解方向进行了收敛，尽管文本提示仍然能够带来一定程度的提升，但这种提升的幅度相对预训练模型而言就没有那么明显了。

表 3: 消融实验

文本提示增强	预训练 CLIP	微调 CLIP
有	21.34%	89.03%
无	51.02%	92.46%

从表格中的数据可以直观地看出，在预训练模型中，有无文本提示的预测精准率差异较大，而微调模型中虽然也存在差异，但相对较小，这与我们上述的分析结果高度吻合，进一步证实了文本提示对于不同阶段模型的影响特点和程度。

6 未来工作

在本项目中，我们创新性地运用文本提示增强 CLIP 方法来处理苹果叶片病害细粒度分类问题。通过引入病理相关语义信息，我们成功地证明了这种方法能够显著提升模型对疾病的理解能力和诊断准确性，这一成果为其他类似任务提供了极具价值的启发和借鉴。一系列实验也充分验证了本方法的有效性。然而，受人力、算力以及时间等多方面因素的限制，当前方法还存在一些有待改进之处。

首先，在图像编码器与语言信息交互方面存在局限性。在目前的情况，图像编码器从语言侧所接收到的监督信号较为有限，因为相同的图像始终仅与固定的单词进行配对。这种固定的配对模式导致语言侧能够为图像侧所提供的指导信息相对匮乏，进而限制了图像编码器从语言信息中获取更丰富、更具多样性知识的可能性。

其次，文本编码器在训练过程中存在过度拟合风险。在每个训练周期（Epoch）中，文本编码器反复处理完全相同的文本。这种重复性操作增加了文本编码器对特定文本过度拟合的可能性，进而影响到最终训练得到的图片编码器的泛化性能。

更为理想的解决方案是为每一张图像生成一段具有针对性的文本描述，并且在训练过程中的不同 Epoch 阶段，对这些文本描述不断进行重写和增强。这样做能够为图像编码器提供更加丰富、多样化且有针对性的语言指导信息，从而更好地挖掘图像中的特征信息，提升模型的性能。然而，这一设想的实现面临着巨大的挑战。它需要投入大量的人力进行文本创作和优化，同时对算力提出了极高的要求。此外，考虑到生成高质量、针对性强的文本描述可能需要借助像 ChatGPT 这样的语言模型，这又涉及到对相关工具使用权限的获取和管理问题。目前来看，这无疑是一项极具挑战性且工作量庞大的任务，需要我们在未来进一步探索和解决。

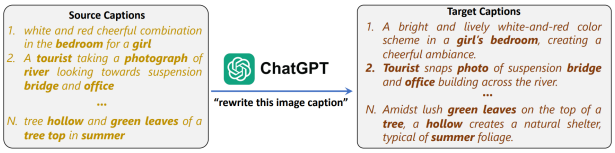

图 15: 本图展示了使用 ChatGPT 重写增强文本提示的示例，从中可以直观地看到针对性文本描述对于图像信息补充的潜力，但同时也反映出实现这一操作的复杂性和难度。

7 结论

通过本项目的研究，我们成功地将文本提示增强的 CLIP 模型应用于苹果叶片病害的细粒度分类问题。实验结果表明，我们的 CLIP 模型仅在少量轮数的微调情况下即可有效解决叶片诊断的细粒度分类问题，并表现出强大的泛化能力和优秀的零样本预测能力。文本提示增强策略为 CLIP 提供了有效的病理信息，显著提升了模型对病害的识别能力。LoRA 技术的引入进一步优化了模型的参数调整，使得模型在有限的训练资源下能够快速收敛并提高性能。此外，我们还通过 t-SEN、CAM 等多样的可视化分析方法，对预测结果、中间层特征、可解释性进行了全面的评估，这增强了模型决策过程的透明度和可信度，也帮助我们更好地理解模型原理。但是本研究也存在一些局限性，如语言信息交互的局限性和文本编码器的过度拟合风险。未来的工作将集中在解决这些挑战上，包括生成更有针对性的文本描述和优化文本编码器的训练策略。总体而言，本项目不仅为农业病害诊断提供了创新的解决方案，也为多模态预训练模型在专业领域的应用开辟了新的可能性。

8 致谢

感谢老师和助教提供的数据集。感谢中山大学华为智能基座协会提供的 8 卡 NVIDIA RTX 3090 GPU 服务器。

参考文献

[1] Serge Savary, Laetitia Willocquet, Sarah Jane Pethybridge, Paul Esker, Neil McRoberts, and Andy Nelson. The global burden of pathogens and pests on major food crops. *Nature Ecology amp; Evolution*, page 430–439, Feb 2019.

[2] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024.

[3] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022.

[4] Quan Huu Cap, Hiroyuki Uga, Satoshi Kagiwada, and Hitoshi Iyatomi. Leafgan: An effective data augmentation method for practical plant disease diagnosis. *IEEE Transactions on Automation Science and Engineering*, page 1258–1267, Apr 2022.

[5] Alec Radford, JongWook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Askell Amanda, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Cornell University - arXiv,Cornell University - arXiv*, Feb 2021.

[6] HuEdward J., Yulong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv: Computation and Language,arXiv: Computation and Language*, Jun 2021.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

[8] Jure Zbontar, Jing Li, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Mar 2021.

[9] Yael Vinker, Ehsan Pajouheshgar, JessicaY. Bo, RomanChristian Bachmann, AmitHaim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching.

[10] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific Reports,Scientific Reports*, Jul 2017.

[11] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. Nov 2022.

[12] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022.

[13] Hanoona Rasheed, Muhammad Maaz, MuhammadUzair Khattak, Salman Khan, Shahbaz Fahad, and Khan Khan. Bridging the gap between object and image-level representations for open-vocabulary detection.

[14] Jiayi Lin and Shaogang Gong. Gridclip: One-stage object detection by grid-level clip representation learning.

[15] Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. Improving zero-shot generalization for clip with synthesized prompts.

[16] Kaiyang Zhou, Jingkang Yang, Chen Loy, Ziwei Liu, and JLDZMSDZZLM D. Conditional prompt learning for vision-language models.

[17] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, page 101985, Apr 2021.

[18] Zihao Zhao, Yuxiao Liu, Han Wu, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Xiang Li, Zhiming Cui, Qian Wang, and Dinggang Shen. Clip in medical imaging: A comprehensive survey. Dec 2023.

[19] Philip Müller, Felix Meissen, Johannes Brandt, Georgios Kaissis, and Daniel Rueckert. Anatomy-driven pathology detection on chest x-rays. Sep 2023.

[20] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. Apr 2023.

[21] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G. Elmore, and Linda Shapiro. *Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images*, page 893–901. Jan 2018.

[22] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[23] Yusuke Kawasaki, Hiroyuki Uga, Satoshi Kagiwada, and Hitoshi Iyatomi. *Basic Study of Automated Diagnosis of Viral Plant Diseases Using Convolutional Neural Networks*, page 638–645. Jan 2015.

[24] Alvaro Fuentes, Sook Yoon, Sang Kim, and Dong Park. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, page 2022, Sep 2017.

[25] Chad DeChant, Tyr Wiesner-Hanks, Siyuan Chen, Ethan L. Stewart, Jason Yosinski, Michael A. Gore, Rebecca J. Nelson, and Hod Lipson. Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning. *Phytopathology®*, page 1426–1432, Nov 2017.

[26] Yang Lu, Shujuan Yi, Nianyin Zeng, Yurong Liu, and Yong Zhang. Identification of rice diseases using deep convolutional neural networks. *Neurocomputing*, page 378–384, Dec 2017.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[28] B.A. Perdue. Occam’ s razor. Jul 2020.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Oct 2020.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019.

[32] Mingxing Tan and QuocV. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. May 2019.

[33] C Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, C.T. Mullis, Katta Aarush, Coombes Theo, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *Cornell University - arXiv,Cornell University - arXiv*, Nov 2021.

[34] Christoph Schuhmann, §§°°romain Beaumont, Vencu Vencu, Ade Gordon, Wightman Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, °°jenia Jitsev, UC Berkeley, and Gentec Data. Laion-5b: An open large-scale dataset for training next generation image-text models.

[35] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, page 336–359, Feb 2020.