Utilizing Vision-Language Models for Detection of Leaf-Based Diseases in Tomtoes

James Blossom Eleojo
Bowen University
jamesblossom123@gmail.com

Abstract

Leaf based diseases in tomatoes such as early blight, late blight, and septoria leaf spot, pose a significant threat to global food security and have substantial economic impacts. Early detection of these diseases is crucial for improving crop yields. This paper explores the use of vision-language models(VLMs) for detecting tomato leaf diseases by fine-tuning a pre-trained model on a large dataset of tomato leaf images with corresponding disease annotations. This approach enhances disease detection accuracy and enables multi-modal learning, real-time monitoring, and automated diagnosis, offering promising applications in precision farming and food production.

Introduction

In this paper the significance of utilizing vision-language models for detecting leaf based diseases in tomatoes will be explored. According to the data collected and updated in December 2022 by the FAO, the total tomato production for both processing and fresh consumption in 2021 amounted to just over 189.1 million metric tonnes. This shows how much tomatoes are in demand and how tomato production is a critical component of the global food security. Manual disease detection in tomatoes rely on human expertise which is not accurate, time consuming, and tiring. By implementing automated plant disease diagnosis the yield of tomato production will face a possible increase.

Background

"Since 2021, we've seen an increased interest in models that combine vision and language modalities (also called joint vision-language models), such as OpenAI's CLIP. Joint vision-language models have shown particularly impressive capabilities in very challenging tasks such as image captioning, text-guided image generation and manipulation, and visual question-answering. This field continues to evolve, and so does its effectiveness in improving zero-shot generalization leading to various practical use cases." (A Dive Into Vision-Language Models, n.d.).

Previous studies have explored computer vision and machine learning approaches for plant disease detection. Notable works include:

Trivedi et al's (2021) "Early Detection and Classification of Tomato Leaf Diseases Using High-performance Deep Neural Network" makes use of Google colab to conduct experiments with a dataset containing 3000 images of tomato leaves affected by nine different diseases and a healthy one. The images were first preprocessed then further processed with varying hyper-parameters of the CNN model. Trivedi et al's (2021) findings showed that the proposed model predictions are 98.49% accurate.

Jung et al's. (2023) paper,"Construction of Deep Learning-Based Disease detection Model in Plants," presents a stepwise disease detection model using images of diseased-healthy plant pairs and a CNN algorithm. The model achieved high accuracy in classifying crops and disease types, potentially enhancing smart farming of Solanaceae crops. In this research low accuracy of non-model crops was improved by adding the non-model crops to the training dataset implicating expend-ability of the model.

Approach

By leveraging the strengths of both visual and semantic information, my approach enables accurate and robust disease detection, improving upon existing single-modal method.

This approach leverages a multi-modal learning framework, combining computer vision and natural language processing (NLP) technique to detect leaf-based diseases in tomatoes. By making use of a very large dataset of tomato leaf images, healthy and diseased.

For the Computer vision component,

1. **Image pre-processing**:Tomato leaf images are resized to a consistent dimension, normalized to stand-

ardize pixel values, and augmented to create variations in the dataset. This pre-processing enhances the diversity and robustness of the image data.

2. **Feature extraction**: A multi-stream convolutional neural network (MSCNN) extracts visual features from the preprocessed images (instead of a single CNN as used in previous times), in order to improve the accuracy of disease detection.
3. **Image Embedding** : The extracted features would be embedded into a vector space, in order to represent the visual characteristics of images. (Radford et al., 2021)

For the natural language component;

1. **Text pre-processing:** Disease labels and descriptions are pre-processed using NLP techniques such as tokenization, stopword removal.
2. **Text embedding**: The pre-processed text is embedded into a vector space, representing the semantic meaning of the disease labels.

For the multi-modal fusion;

1. **Joint embedding space:** The image and text embeddings are fused into a joint embedding space, enabling the model to learn cross-modal relationships.
2. **Vision-language model(VLM)**: A pre-trained VLM (e.g. CLIP, VisualBERT) is fine-tuned on the joint embedding space to learn disease-specific features.

For disease detection

1. **Image-text matching:** The fine-tuned VLM predicts disease labels for new, unseen images by matching image embeddings with text embeddings.(Kim et al., 2021)
2. **Disease Classification:** The predicted labels are used to classify the images into corresponding disease categories.

This approach is unique in the sense that it makes use of a multi-stream convolutional neural network and multi-modal fusion, which increases this models accuracy.

Evaluation

Assessing the effectiveness of the ability for a vision-language model to understand and generate text (give diagnosis of the tomato leaf) based on visual inputs such as images of the leafs, both healthy and diseased is very important. Here are some key aspects to assess:

1. **Disease Classification Accuracy**: The model's ability to correctly classify tomato leaves as healthy or diseased, and identify the specific disease (e.g., leaf spot, powdery mildew, early blight) would be evaluated.
2. **Disease Detection Sensitivity**: Asses the model's sensitivity in detecting disease at various stages(e.g.,early, moderate, severe) this helps in enabling users to detect whatever disease is affecting the tomato plant at an early stage, increasing crop yield.
3. **Leaf image Quality Robustness:** Test will be taken on the model's performance on images with varying quality, resolution, lighting conditions.
4. **Disease Severity Estimation:** Assess the model's ability to estimate disease severity level (e.g., mild, moderate, severe).
5. **Multi-class Disease Detection**: Evaluate the model's performance in detecting multiple diseases simultaneously.
6. **Performance Evaluation Metrics:** The accuracy, precision, recall, and F1-score measures are used to evaluate the model's performance.
7. **Real-world performance**: Test the model's performance on real-world images, including those with varying backgrounds, occlusions, and leaf orientations.

Discussion

VLMs offer a promising solution for detecting tomato leaf diseases by combining image recognition and text descriptions. However, they face challenges in real-world deployment. While VLMs improve accessibility with disease explanations, their performance is limited by the quality of training data. Miss-classifications occur when diseases have similar symptoms, and robust datasets are required for better accuracy.

Practical use is also hindered by image quality issues and the high computational demands of VLMs. To overcome this, lightweight models for mobile devices could enable real-time detection in farms, making these systems more accessible and usable in diverse areas.

Conclusion

This proposal demonstrates the potential of vision-language models for accurate and efficient detection of leaf-based diseases in tomatoes. By leveraging the combined power of computer vision and natural language processing, VLMs can provide a valuable tool for farmers and agricultural professionals in managing tomato crops and preventing significant economic losses. Future work may focus on further improving model performance, exploring new vision-language model (VLM) architectures and integrating disease detection systems into real-world scenarios.

References

A Dive into Vision-Language Models. (n.d.). https://huggingface.co/blog/vision_language_pretraining. Accessed: 2025-01-11

Jin, X., Tao, Z., & Kong, J. 2021. Multi-stream aggregation network for fine-grained crop pests and diseases image recognition. *International Journal of Cybernetics and Cyber-Physical Systems*, *1*(1), 52. https://doi.org/10.1504/ijccps.2021.113105

Jung, M., Song, J. S., Shin, A., Choi, B., Go, S., Kwon, S., Park, J., Park, S. G., & Kim, Y. 2023. Construction of deep learning-based disease detection model in plants. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-34549-2

Kim, W., Son, B., & Kim, I. 2021, February 5. VILT: Vision-and-Language Transformer without convolution or region Supervision. arXiv.org. https://arxiv.org/abs/2102.03334

Qiu, J., Lu, X., Wang, X., Chen, C., Chen, Y., & Yang, Y. 2024. Research on image recognition of tomato leaf diseases based on improved AlexNet model. Heliyon, 10(13), e33555. https://doi.org/10.1016/j.heliyon.2024.e33555

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. 2021, February 26. Learning transferable visual models from natural language supervision. arXiv.org. https://arxiv.org/abs/2103.00020

Trivedi, N. K., Gautam, V., Anand, A., Aljahdali, H. M., Villar, S. G., Anand, D., Goyal, N., & Kadry, S. 2021. Early detection and classification of tomato leaf disease using High-Performance Deep Neural Network. *Sensors*, *21*(23), 7987. https://doi.org/10.3390/s21237987

Zhang, J., Huang, J., Jin, S., & Lu, S. 2024. Vision-Language Models for Vision Tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5625–5644. https://doi.org/10.1109/tpami.2024.3369699

Zhang, J., Wang, G., Kalra, M. K., & Yan, P. 2024. Disease-Informed adaptation of Vision-Language models. In Lecture notes in computer science (pp.232–242). https://doi.org/10.1007/978-3-031-72120-5_22