Vision Transformer For Plant Health Detection

Ziyi Wang

1 Overview

Global food security is largely dependent on agricultural productivity. However, the plant diseases have been a long issue to impact the crop yields which can also lead to serious economic losses and food scarcity. Traditional methods of identifying plant health issues mostly rely on expert's knowledge and manual inspection which can be inefficient and impractical for large scale applications. With the recent research on machine learning and computer vision, we have potential tools to automating this classification process.

This project aims to leverage Vision Transformers to classify plant health conditions based on leaf images, which will address the need for accurate large-scale plant health condition detection in agriculture. This project combines my interest in cutting-edge machine learning techniques and their practical applications to address real-life problems. And it also provides me an opportunity to deepen my knowledge of Vision Transformers and their application to image classification tasks, which is the field I found most interesting in machine learning. The dataset come from the huggingface[1]. The expected outcomes include high classification accuracy, insights into key visual features influencing the model decisions, and a demonstration of AI's potential to revolutionize agricultural sustainability by efficient and scalable plant health monitoring.

2 Prior Work

Vision Transformer (ViTs) were introduced in 2020 by Dosovitskiy et al. as a groundbreaking architecture that adapts the transformer framework, initially designed for natural language processing, to computer vision tasks by dividing images into patched and processing them as sequences of tokens(Dosovitakiy et al., 2020). Unlike convolutional neural networks (CNNs), used widely in computer vision tasks, ViTs make use of self-attention mechanism to capture long-range dependencies and contextual information within images. Since their inception, ViTs have undergone significant advancements, including the development of hybrid architectures that integrate convolutional layers for local feature extraction, as well as training improvements such as self-supervised learning and knowledge distillation to address their data-intensive nature(Touvron et al., 2021). ViTs are now applied across a variety of real-world domains, such as medical imaging for precise organ and lesion segmentation, agricultural monitoring for crop disease detection, and autonomous driving for object detection and scene understanding. These advancements highlight the versatility and transformative potential of ViTs in addressing complex visual tasks across industries.

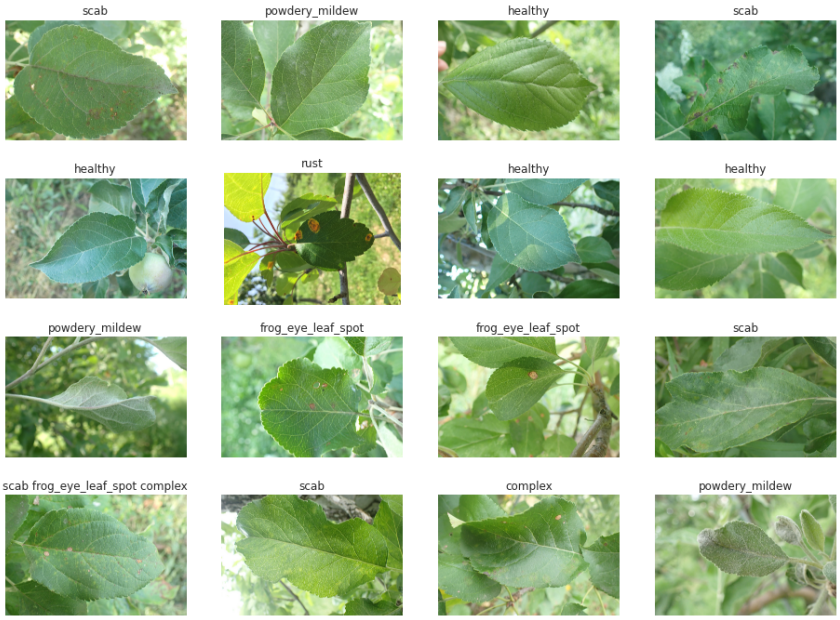
To classify plant health conditions using ViTs, firstly, it is essential to conduct data preprocessing to prepare the dataset, including resizing images to match the input size required by the ViT model (e.g., 224x224 pixels) and possible techniques to increase dataset diversity by applying data augmentation. Secondly, a pre-trained ViT model from the Hugging Face PyTorch Image Models library will be fine-tuned for this specific task by replacing the classification head with one tailored to the dataset's health categories.

[1] https://huggingface.co/datasets/timm/plant-pathology-2021

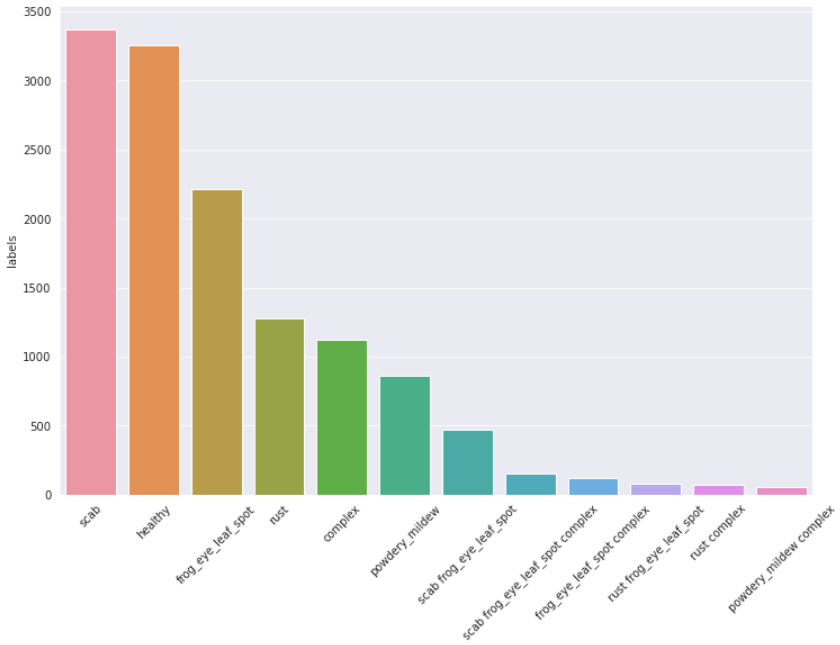
Preprint. Under review.

3 Preliminary Results

The Plant Pathology 2021 dataset contains labeled images of healthy and diseased plant leaves, specifically apple leaves. It has multiple classes corresponding to different conditions, such as healthy, rust, scab, and multiple diseases. Images vary in resolution but are uniform in format. Some sample images and a basic counting of the number of each label category are shown below.



(a) Some image samples of the plant health dataset.



(b) The number of images for various of 12 categories present.

Figure 1: Basic description of the dataset with some sample images.

Challenges such as class imbalance, noisy labels, and variations in environmental conditions (e.g. lighting and plant ages) can occur. Possible solutions include preprocessing and data augmentation techniques to enhance diversity and model robustness. Vision Transformers (ViTs) typically require

significant computational resources due to their attention mechanisms. GPUs with large memory, such as NVIDIA A100 or similar, will be essential for efficient training. However, since this project is based on the pre-trained model, it will have a lower computational demand.

Tools from the class such as NumPy, Pandas, and Matplotlib will be used for data analysis and visualization, PyTorch will be used for defining and training the model, Scikit-learn will be used for evaluation metrics and performance analysis. Some other tools I will explore during this project include how to use Hugging Face Transformers for accessing pre-trained ViT models and datasets and tools like TensorBoard or Weights & Biases for tracking experiments

4 Project Deliverables

If this project is successfully completed, a robust and accurate ViT model which is capable of classifying plant health conditions based on leaf images will be produced. The trained and fine-tuned ViT model should achieve high accuracy and balanced performance across all health categories. To achieve the project outcomes, some sub-goals will also be pursued including conducting a detailed analysis of model performance metrics and drawing attention maps highlighting the key image features used by the ViT (if possible).

5 Plan of Project (Timeline)

5.1 Week 1-2

Search on huggingface to select model for this project. Then, find the proper dataset for model training. Finish the proposal and conduct basic data analysis and preprocessing.

5.2 Week 3-4

Set up the environment and start to work on the fine-tuned model with the selected dataset.

5.3 Week 5

Finish the final report and upload the code and report to GitHub.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *preprint arXiv:2010.11929.*. https://arxiv.org/abs/2010.11929

[2] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877.*. https://arxiv.org/abs/2012.12877