Vision Transformer For Plant Health Detection

Ziyi Wang
Dept. of Statistics
Univercity of Michigan-Ann Arbor
Ann Arbor, US
ziyiwan@umich.edu

I. Introduction

Global food security heavily depends on agricultural productivity, yet plant diseases have long posed significant challenges to crop yields, leading to severe economic losses and exacerbating food scarcity. Traditional methods for identifying plant health issues typically rely on expert knowledge and manual inspection, which are time-consuming, inefficient, and impractical for large-scale applications. Recent advancements in machine learning and computer vision offer promising solutions to automate this process, paving the way for scalable and precise plant health monitoring. Vision Transformers (ViTs) have emerged as a groundbreaking tool in computer vision, leveraging self-attention mechanisms to address image classification tasks with unprecedented accuracy and scalability. This project aims to leverage ViTs to classify plant health conditions using leaf images, providing an efficient, accurate, and scalable solution for detecting plant diseases. By implementing ViTs, this project not only aligns with cutting-edge developments in machine learning but also contributes to sustainable agriculture and global food security. The expected outcomes include achieving high classification accuracy, gaining insights into the visual features that influence the model's decisions, and demonstrating AI's potential to revolutionize agricultural sustainability by enabling efficient and scalable plant health monitoring.

The Vision Transformer (ViT), introduced in 2020 by Dosovitskiy et al. (1), revolutionized computer vision by adapting the transformer architecture from natural language processing to image-based tasks. ViTs divide images into patches, embed each patch as a token, and process the sequence of tokens using self-attention mechanisms. Unlike convolutional neural networks (CNNs), which rely on local feature extraction, ViTs excel at capturing long-range dependencies and global contextual information. Since their introduction, advancements such as hybrid architectures combining convolutional layers, and training techniques like self-supervised learning and knowledge distillation, have enhanced ViTs' performance and efficiency (2). These innovations have made ViTs highly effective in diverse applications, including medical imaging, agricultural monitoring, and autonomous driving, demonstrating their versatility and transformative potential.

ViTs' ability to tokenize images, analogous to tokenizing sentences in natural language processing, underscores their versatility. The process involves splitting an image into patches, embedding each patch through a linear projection, and passing the resulting sequence of embedded patches as input tokens to the transformer model. This innovative design enables ViTs to excel in capturing long-range dependencies and contextual information within images, a critical factor in tasks requiring a global understanding of visual data. By leveraging this architecture, the project aims to achieve high classification accuracy, understand the visual features influencing the model's decisions, and demonstrate the potential of ViTs to revolutionize agricultural monitoring.

II. Method

To classify plant health conditions using Vision Transformers (ViTs), this project formulated the problem as a supervised learning task where the input consists of leaf images, and the output is a categorical label representing the plant's health condition. The dataset, sourced from the Plant Pathology 2021 competition[1], contains images labeled into six categories: powdery mildew, frog eye leaf spot, complex, healthy, scab, and rust. To address this problem, data preprocessing was conducted, including resizing the images to the required input dimensions (224x224 pixels) and applying normalization to ensure compatibility with the ViT model.

A pre-trained ViT model [2] was loaded from the Hugging Face PyTorch Image Models library. The classifier head of the model was replaced with one tailored to the six target health categories in the dataset. The weights of the backbone were initialized using the pre-trained ImageNet model, leveraging transfer learning to extract robust features for plant health classification. Training was conducted on a filtered subset of the dataset to include only single-labeled images, ensuring consistency in class labeling. The AdamW optimizer was used to update model weights, and the cross-entropy loss function was employed to minimize prediction error. Throughout the training process, validation loss and accuracy were monitored to identify signs of overfitting and guide model optimization. At the end of each epoch, the model was evaluated on a held-out test set to compute the test loss and accuracy.

III. Result

The data pipeline begins with the preprocessing of the Plant Pathology 2021 dataset, where images are resized, normalized, and transformed into tensors using Hugging Face's

[1]https://huggingface.co/datasets/timm/plant-pathology-2021
[2]https://huggingface.co/google/vit-base-patch16-224

AutoImageProcessor. The dataset is then split into training, validation, and test sets, with a filtered subset of single-labeled images comprising 1% of the original data for fine-tuning. The fine-tuned Vision Transformer (ViT) model was set up with a classifier head matching the six target categories. During training, the model was optimized using the AdamW optimizer and cross-entropy loss, and its performance was validated after each epoch.

The fine-tuned Vision Transformer (ViT) model was adapted to classify plant health conditions into six categories. The pre-trained ViT backbone was combined with a modified classification head to output six target classes. The training process used the AdamW optimizer with a learning rate of 5×10^{-5} and cross-entropy loss to minimize prediction error. The model was trained over 10 epochs, with the training, validation, and test losses monitored to ensure convergence and mitigate overfitting. The training loss reached 0.0042, indicating that the model fit the training data effectively. Validation and test losses stabilized at 0.4160 and 0.4118, respectively, demonstrating strong generalization. The final test accuracy of 86.67% underscored the model's capability to classify plant health conditions accurately on unseen data.

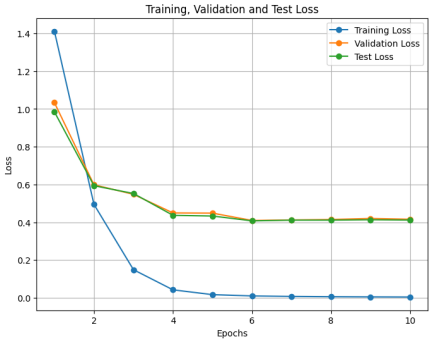

Fig. 1. Loss Trends

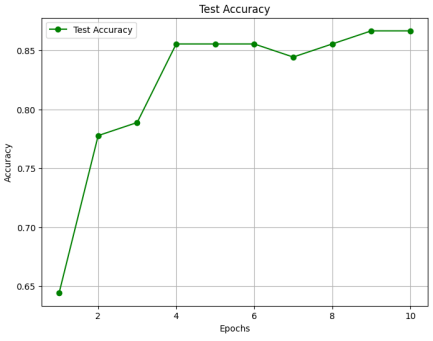

Fig. 2. Test Accuracy

The figures presented provide additional insights into the model's performance. The training and validation loss curve show a steady decline in training loss, stabilizing near-zero, while validation and test losses converge at a similar level. This indicates that the model generalizes well without signs of overfitting. The test accuracy curve shows a steady increase across

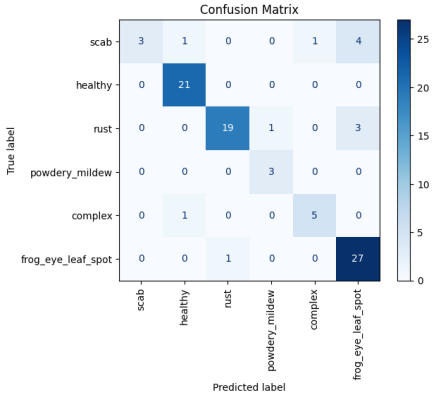

Fig. 3. Confusion Matrix

epochs, plateauing at 86.67%, which reflects the model's robustness in making predictions on unseen data. The confusion matrix highlights the class-wise performance, with strong results for categories like "healthy" and "frog_eye_leaf_spot." However, minor misclassifications were observed, such as slight confusion between "scab" and "frog_eye_leaf_spot."

The results validate the effectiveness of transfer learning using Vision Transformers for plant health classification tasks. The pre-trained ViT model effectively leveraged its knowledge from ImageNet while adapting to the domain of plant disease detection. The high accuracy and stable loss metrics highlight the model's ability to generalize well to unseen data. However, minor misclassifications in visually similar categories suggest potential areas for improvement, such as using data augmentation or increasing the dataset's diversity. Overall, the results demonstrate that the model is well-suited for automated agricultural monitoring and decision-making.

IV. Conclusion

In this project, a pre-trained Vision Transformer (ViT) model was fine-tuned to classify plant health conditions based on leaf images. By leveraging the robust feature extraction capabilities of the pre-trained ViT model and tailoring its classification head to the specific task, the model achieved high accuracy and demonstrated strong generalization on unseen test data. The training process highlighted the efficiency of transfer learning, where a model pre-trained on a large-scale dataset such as ImageNet could be effectively adapted to a domain-specific problem with a relatively small dataset. The results, including loss trends and the confusion matrix, show the model's ability to correctly classify most categories, with minimal misclassifications. Future work could explore data augmentation and additional fine-tuning techniques to further improve performance, particularly for underrepresented classes. This project underscores the viability of using state-of-the-art transformer-based models for agricultural monitoring and disease detection, paving the way for scalable and automated plant health diagnostics.

References

[1] [Dosovitakiy et al., 2020]Dosovitskiy2022 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 et al., 2021]Touvron2021 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