Plant disease identification in leaf images using multi-scale high-resolution visual transformers

Rakhat Yskak
School of Engineering and Digital Sciences
Nazarbayev University
Astana, Kazakhstan

***Abstract*—In this project, I developed a framework for automated plant disease identification using Vision Transformers and High-Resolution Networks. The model was trained on a custom subset of the PlantVillage dataset and evaluated using micro and macro-averaged metrics, achieving a micro-averaged F1-score of 0.98 and a Hamming loss of 0.025. The results highlight the model's effectiveness and potential for agricultural applications.**

I. Introduction

The accurate and early identification of plant diseases is crucial for ensuring food security and reducing agricultural losses. Traditional methods of disease detection often rely on manual inspections, which are time-consuming and prone to human error. Deep learning techniques, particularly convolutional neural networks (CNNs), have improved image-based disease detection but struggle to simultaneously capture global features and high-resolution details.

To address these limitations, I propose an approach that integrates Vision Transformers (ViTs) with High-Resolution Networks (HRNet) for efficient and accurate plant disease classification. This approach leverages ViTs for global feature extraction and HRNet for multi-scale feature representation, aiming to enhance classification performance.

II. Methodology

A. Model Architecture

The core of my project is a model, which combines the principles of ViT and HRNets to effectively capture and process high-resolution features. This design leverages recent techniques from [1] and [2].

The model divides each input image $x \in \mathbb{R}^{H\times W\times C}$ into non-overlapping patches of size $p \times p$. The total number of patches N is computed as:

$$N = \frac{H \cdot W}{p^2} \tag{1}$$

Each patch is embedded and combined with positional embeddings:

$$z = E \cdot x_{\text{patch}} + P \tag{2}$$

where E is the patch embedding matrix and P denotes the positional embeddings, ensuring the model preserves the spatial structure as described in [3].

Multi-scale architecture inspired by HRNet was apoted to maintain high-resolution feature maps throughout the model [1]. The fusion mechanism across multiple resolutions is defined as:

$$x_i^l = \sum_{j=1}^{s} F_{ij}(x_j^{l-1}) \tag{3}$$

where x_i^l represents the feature map at resolution i in layer l, s is the number of resolutions, and F_{ij} is the transformation function, which combines convolutional and attention operations for effective feature fusion.

B. Dataset and Preprocessing

For this project, I constructed a custom subset inspired by the PlantVillage dataset, similar to the one described in ViTaL [4]. My dataset contains 11 classes with 275 images per class to ensure class balance. Each image was resized to a fixed resolution of 256×256 pixels, and I applied a series of preprocessing and augmentation techniques to enhance the model's performance.

The preprocessing pipeline includes:

1) Resizing with bilinear interpolation.
2) Min-max normalization:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{4}$$

where $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ [3].

C. Data Augmentation

To reduce overfitting and improve the model's generalization capabilities, I applied the following augmentations:

1) Random horizontal flip with a probability of 0.5 to introduce variability.
2) Random rotation from $[-15°, 15°]$ to account for different orientations.
3) Color jitter to simulate varying lighting conditions.

These augmentations were essential in ensuring the model performed well under different conditions, as highlighted in related work [4].

D. Training

Model was trained using the cross-entropy loss function and the AdamW optimizer over 50 epochs with a batch size of 32, a learning rate of 1×10^{-4}, and a weight decay of 1×10^{-5}. Moreover, implemented early stopping if the validation loss did not improve over 10 consecutive epochs [4].

III. Results

The model's performance was evaluated using micro and macro-averaged precision, recall, and F1-score, along with the Hamming loss as in [4]. The results are summarized in Table I.

TABLE I
Performance Metrics

Metric	Micro Average	Macro Average
Precision	0.98	0.975
Recall	0.98	0.975
F1-Score	0.98	0.975
Hamming Loss	0.025	

The Hamming Loss of 0.025 indicates a low rate of incorrect predictions, demonstrating the model's strong performance across all classes.

The confusion matrix in Figure 1 shows only a few misclassifications, primarily in visually similar classes.

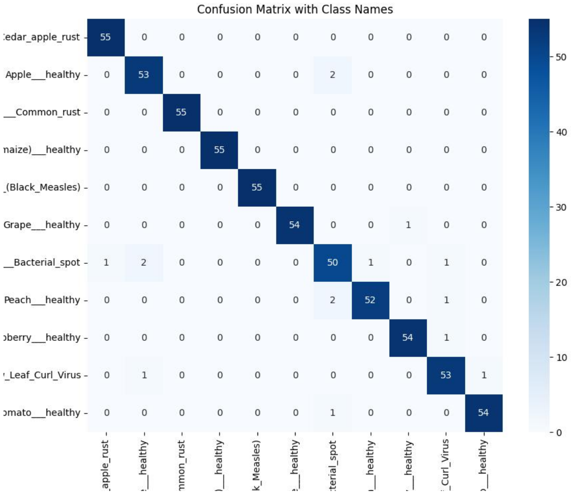

Fig. 1. Confusion Matrix with Class Names

The training and validation loss curves in Figure 2 show a steady decrease, indicating good generalization despite the small dataset.

IV. Discussion

Proposed model demonstrates high precision, recall, and F1-scores, with an overall strong performance. Despite the small dataset size, the model generalizes well, as evidenced by the consistent training and validation loss curves. However, minor misclassifications in similar disease classes suggest that a larger and more diverse dataset could further improve performance. Future work could explore additional augmentation techniques to enhance class separation.

V. Conclusion

In this project, I developed the model for automated plant disease identification using Vision Transformers and High-Resolution Networks. The model demonstrated strong performance, suggesting its potential for practical agricultural applications. Future improvements could include training on a larger, more diverse dataset to enhance model robustness.

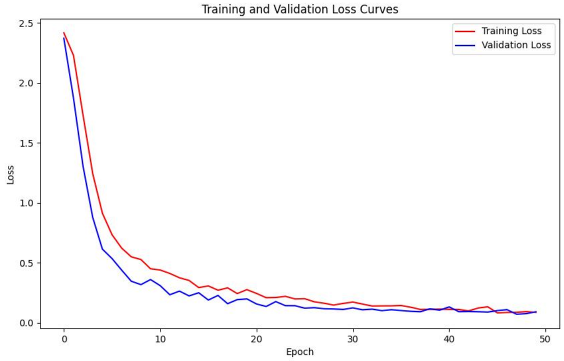

Fig. 2. Training and Validation Loss Curves

References

[1] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," 2020.

[2] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," 2021.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[4] A. Sebastian, A. A. Fathima, R. Pragna, S. MadhanKumar, G. Y. Kannan, and V. Murali, *ViTaL: An Advanced Framework for Automated Plant Disease Identification in Leaf Images Using Vision Transformers and Linear Projection for Feature Reduction*, p. 31–45. Springer Nature Singapore, 2024.