Plant disease identification in leaf images using multi-scale high-resolution visual transformers with linear projection for feature reduction

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

***Abstract*—For this project, I used Vision Transformers and High-Resolution Networks to create a framework for automatically identifying plant diseases. The model achieved a micro-averaged F1-score of 0.98 and a Hamming loss of 0.025 after being trained on a customised subset of the PlantVillage dataset and assessed using micro and macro-averaged metrics. The outcomes demonstrate the efficacy of the model and its potential for use in agriculture.**

I. Introduction

Reducing agricultural losses and guaranteeing food security depend on the precise and timely detection of plant diseases. Conventional disease detection techniques frequently rely on labour-intensive, human-error-prone manual checks. The significance of precision in disease detection becomes particularly evident in regions like Kazakhstan, where agriculture constitutes a substantial portion of the economy. According to recent updates, Kazakhstan has faced challenges such as low yields and grain quality in key crops like wheat and barley, largely attributed to climate conditions and inadequate disease management practices.

Deep learning approaches, such as convolutional neural networks (CNNs), have made strides in image-based disease identification but struggle to simultaneously capture high-resolution details and global patterns. In order to overcome these constraints, I propose a method that combines High-Resolution Networks (HRNet) [1] with Vision Transformers (ViTs) for the effective and precise classification of plant diseases. In order to improve classification performance, this method makes use of HRNet for multi-scale feature representation and ViTs for global feature extraction.

II. Methodology

A. Model Architecture

My project's central component is a model that efficiently captures and processes high-resolution information by fusing the ideas of ViT and HRNets [1]. Inspired by HRViT, this design adopts a multi-branch structure where features are extracted and processed in parallel streams of varying resolutions [2].

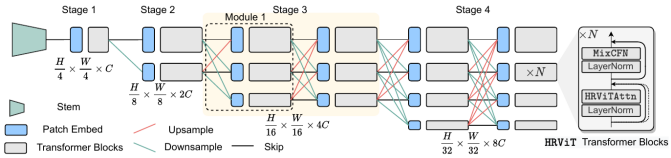

Fig. 1. The overall architecture of HRViT.

First, each input image $x \in \mathbb{R}^{H \times W \times C}$ is divided into non-overlapping N patches of size $p \times p$:

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

Then, to each embedded patch we add their positional embeddings:

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

where E is the patch embedding matrix and P denotes the positional embeddings. This ensures that the model preserves the spatial structure [3].

My inspiration to combine HRNet with ViT, came from the architecture of multi-scale high-resolution ViTs by [2], containing multiple stages from left to right and multiple increasing resolution streams from top to bottom. Each stage contains a patch embed followed by a transformer block, as shown in Figure 1. Each stream contains details of the image in various levels.

To combine and share these features between one stream to other the following fusion mechanism used:

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

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, combining convolutional and attention operations for effective feature fusion.

B. Dataset and Preprocessing

For the dataset, I have decided to use the subset of the commonly used PlantVillage dataset, which includes 54,305 images across 38 distinct classes, captured in RGB format,

with grey-scale and background-removed versions. For this project, a subset of 3,025 images across 11 balanced classes was curated, ensuring sufficient diversity while focusing on computational efficiency. The selected images included both healthy and diseased samples, with diseases such as Cedar Rust, Common Rust, Black Measles, Bacterial Spot, and Yellow Leaf Curl Virus. This selection of classes is due to [3] for a fair comparison between models.

C. Preprocessing

Preprocessing is essential for preparing the dataset for effective training. I applied a series of preprocessing techniques to enhance the model's performance, which are:

1) Resizing with bilinear interpolation. This process adhered to the original aspect ratio wherever possible:

$$\text{New Height} = \text{Target Width} \times \text{Aspect Ratio} \tag{4}$$

where the aspect ratio is defined as the ratio of height to width of the original image.

2) Each image was normalized to bring pixel values into the range [0, 1], reducing input variance and expediting model convergence. This was achieved using min-max normalization:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{5}$$

where $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ as recommended in [4].

D. 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 [3].

E. Training

The proposed model's training procedure was created to overcome issues like overfitting while striking a balance between computational viability and efficient learning. In order to avoid running out of resources, the model design and optimisation techniques have to be carefully adjusted to the constraints of an NVIDIA RTX 3060 GPU mobile with 6GB VRAM.

As seen in Figure 2, the model's design was divided into three stages and tailored for GPU memory and computational limitations. While preserving competitive model performance, this lightweight architecture made sure that GPU memory was used effectively.

Despite the model's reduced size, the training process faced challenges with overfitting due to the small dataset. Dropout layers were added to lessen this, and during training,

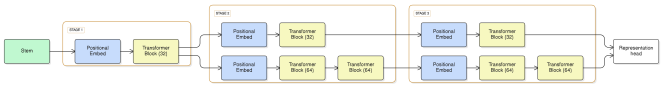

Fig. 2. Used Architecture

the dropout rates progressively increased from 0.1 to 0.3. Likewise, transformer blocks with the same range of rates were subjected to DropPath regularisation, which added diversity to the forward routes and pushed the model to pick up a wider variety of representations. An early stopping mechanism with patience of 10 was included to the training procedure to better combat overfitting.

The final parameters 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 [3].

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 [3]. 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 3 shows only a few misclassifications, primarily in visually similar classes.

The training and validation loss curves in Figure 4 show a steady decrease, indicating good generalization despite the small dataset.

IV. Discussion

ViTs and HRNet worked well together to capture intricate patterns linked to plant illnesses, like minute texture or colour changes in leaves. Although the outcomes are encouraging, the model's applicability to real-world situations where images might show more variation in background, illumination, and resolution is limited by its dependence on consistent backgrounds and controlled settings in a carefully selected portion of the PlantVillage dataset. When used in real-world situations, this can result in decreased performance.

The computing requirements of the model are another drawback. Even though the transformer-based architecture is optimised for an NVIDIA RTX 3060 GPU, it still necessitates

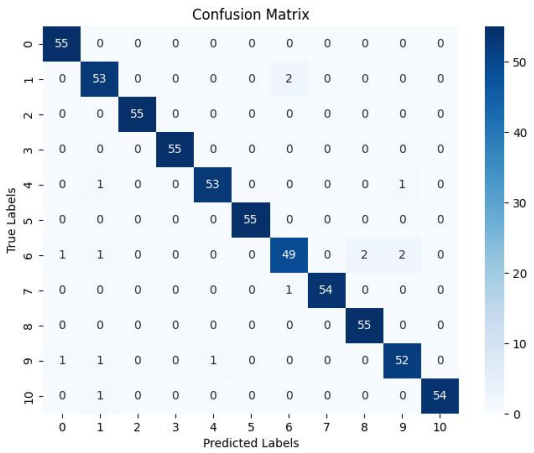

Fig. 3. Confusion Matrix

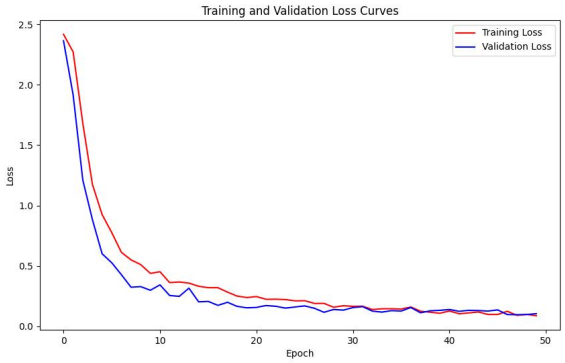

Fig. 4. Training and Validation Loss Curves

a substantial amount of processing power. This limits its use on devices with limited resources, including edge devices or mobile platforms, which are frequently employed in agricultural environments.

A. Future Work

Future work could address these limitations in several ways. Expanding the dataset to include images from diverse environments would enhance the model's robustness and applicability. Additionally, lightweight variants of the architecture, such as MobileViT or Lite-HRNet, could be explored to reduce computational overhead, making the model more suitable for deployment on low-power devices. Furthermore, employing attention mechanisms specifically tuned for high-resolution details, such as local-global attention mechanisms, might help mitigate the challenges posed by visually similar classes.

B. Practical Implications

Despite its limitations, the model has significant potential for practical applications in agriculture. Its ability to achieve near-perfect precision and recall suggests its utility in early disease detection, enabling timely interventions that could prevent large-scale crop losses. Moreover, by automating disease diagnosis, the model could reduce reliance on human expertise, making it particularly valuable in regions with limited access to agricultural specialists.

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.

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. 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.

[4] 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.