PROJECT REPORT

TRANSFORMER HELP CNN SEE BETTER: A LIGHTWEIGHT HYBRID APPLE DISEASE IDENTIFICATION MODEL BASED ON TRANSFORMERS

Submitted to:-
Dr. Aparajita Ojha

Prepared by team: **Runtime_Terror**
Apurv Prakash Shrivastava - 22MCSA02
Ayushi Tiwari - 22MCSA03
Priya Gupta - 22MCSS05

02

INTRODUCTION

Apple is a major fruit species today and is very popular. However, apple leaf diseases seriously affect its yield and quality. Therefore, timely and accurate identification of apple leaf diseases is essential to improve apple yield and quality and promote the apple industry's healthy development. Farmers relied on planting experience and expert guidance to diagnose diseases in the early days. However, with the expansion of the planting scale, this method could no longer meet practical needs. The development of machine learning has provided a new approach to crop disease identification

CASE STUDIES

- Ref [1-4] started with working on machine learning models which require manual feature extraction and traditional machine learning models like Principal Component Analysis (PCA), genetic algorithm (GA), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), etc. but their dataset was small and recognition accuracy is low in noise data.
- Ref [5-7] started to work with CNN but the dataset they used contained a relatively homogenous background, which was ineffective in practical application scenarios.
- Ref [8-9] uses a single attention mechanism which caused overfitting of the model and has a large number of parameters and flops
- Ref [10] uses MobileNet which effectively reduces the parameters and FLOPs but the accuracy could not meet the practical needs.
- Vision Transformer Ref [12-17] inherits the Transformer's approach in NLP Ref [11], dividing the input image into small non-overlapping patches and flattening them into one-dimensional vectors for input into the cascaded Transformers. The multi-head attention mechanism in Transformer can establish long-distance dependence on the input image, providing different attention to different positions of the image. It lacks the inductive bias of CNN structure, so a large amount of data is required for training.

So, we combined the advantages of Transformer and CNN structures to propose a general-purpose, lightweight model for apple disease identification in complex environments—ConvViT.

03

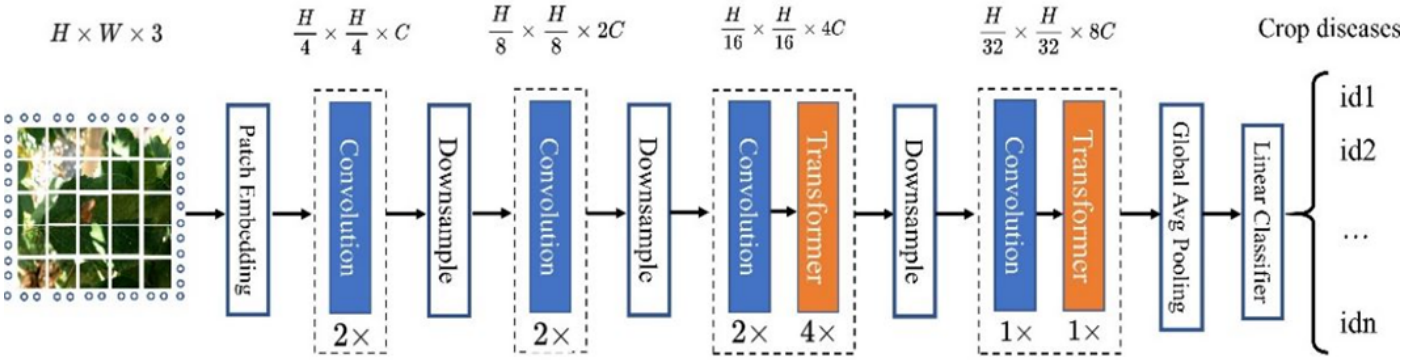
PROBLEM STATEMENT

The complex backgrounds of crop disease images and the small contrast between the disease area and the background can easily cause confusion, which seriously affects the robustness and accuracy of apple disease- identification models. There are several models that are used for identify apple leaf disease but they are not effective on complex backgrounds and works fine with homogenous background.

PROPOSED METHODOLOGY

- Our ConvViT includes convolutional structures and Transformer structures. The convolutional structure is used to extract the global features of the image, and the Transformer structure is used to obtain the local features of the disease region to help the CNN see better.
- We adopt lightweight designs and improve the patch embedding method of Transformer. The depthwise separable convolution reduces the computational complexity of the convolution structure, and the Global Average Pooling reduces the computational complexity of the Transformer structure to linear complexity before performing the attention operation. The improved overlapping patch embedding approach promotes the information exchange between adjacent patches, preserves the information of image edges, and ensures the continuity of image local information
- ConvViT fully combines the advantages of CNN and Transformer and obtains competitive results on the self-built apple disease dataset with much lower parameters and computational effort than other similar identification effect models.

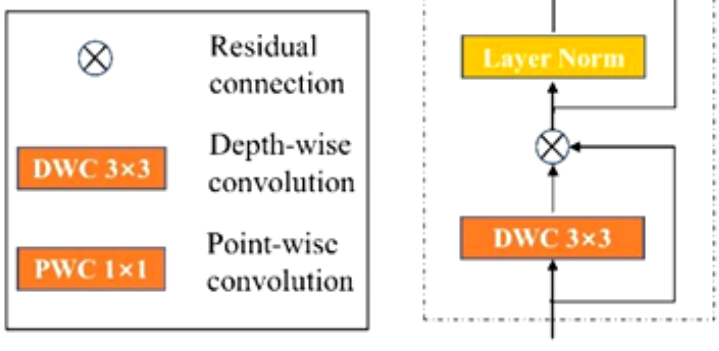
04



CONVVIT MODEL

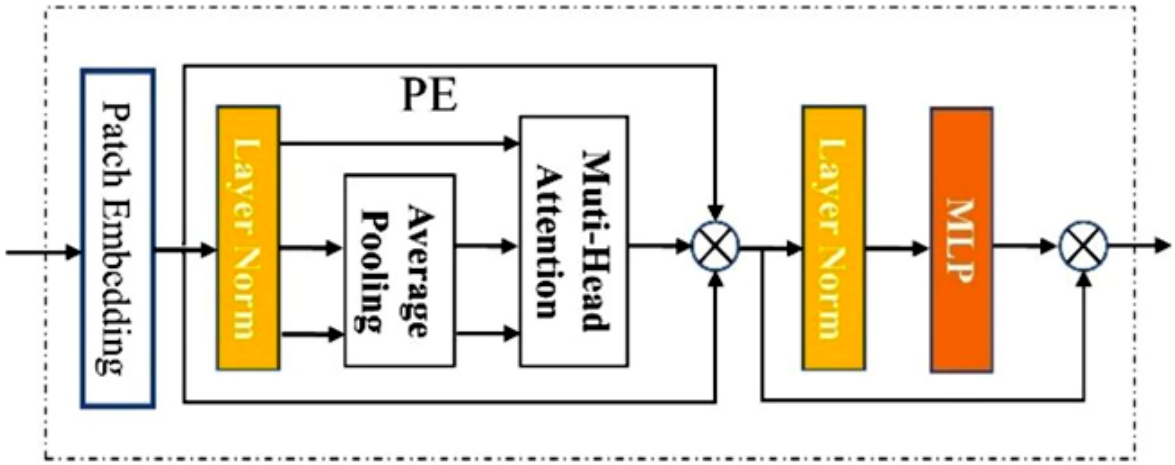
- The architecture contains four stages, each containing several convolutional and Transformer structures and a downsampling structure. It also contains patch embedding, Global Average Pooling, and a linear classifier auxiliary module.
- The model uses a multi-stage design, it reduces redundancy through downsampling, thus reducing the computational effort.
- Due to the extremely high computational cost of Transformer on the high-resolution feature maps, only the convolution structure is used in the first and second stages of the model. In the third and fourth stages of the model, the convolutional and Transformer structures are used alternately.
- Patch embedding is the first step of image processing by Transformer, where the input image is sliced into small patches and transformed into one-dimensional directions for subsequent operations. The Global Average Pooling and linear classifier are used to integrate the global spatial information of the feature map and output the disease classes.

DESIGN OF CONVOLUTION STRUCTURE



05

DESIGN OF TRANSFORMER STRUCTURE



DATASET USED:

WE HAVE USED THE GDPR DATASET OF APPLE LEAF WHICH CONTAINS:
622 - RUST
592 - SCAB
516 - HEALTHY
91 - IMAGES OF MULTIPLE DISEASES.

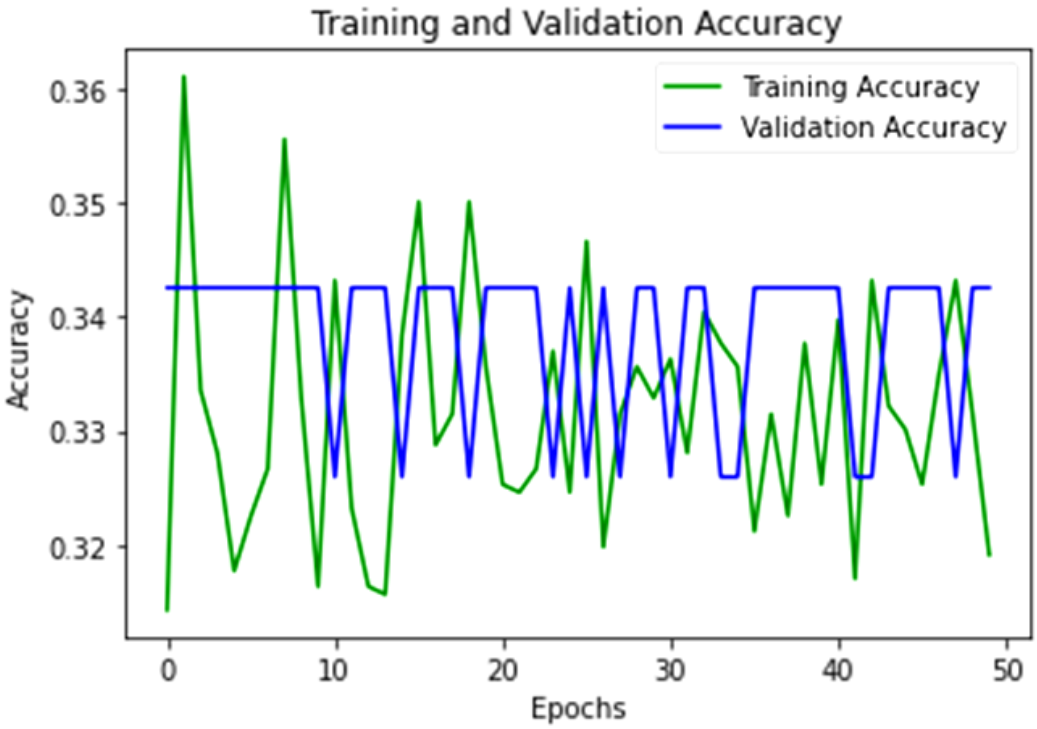
SPECIFIC EXPERIMENTAL METRICS FOR COMPARISON WITH STATE-OF-THE-ART METHODS

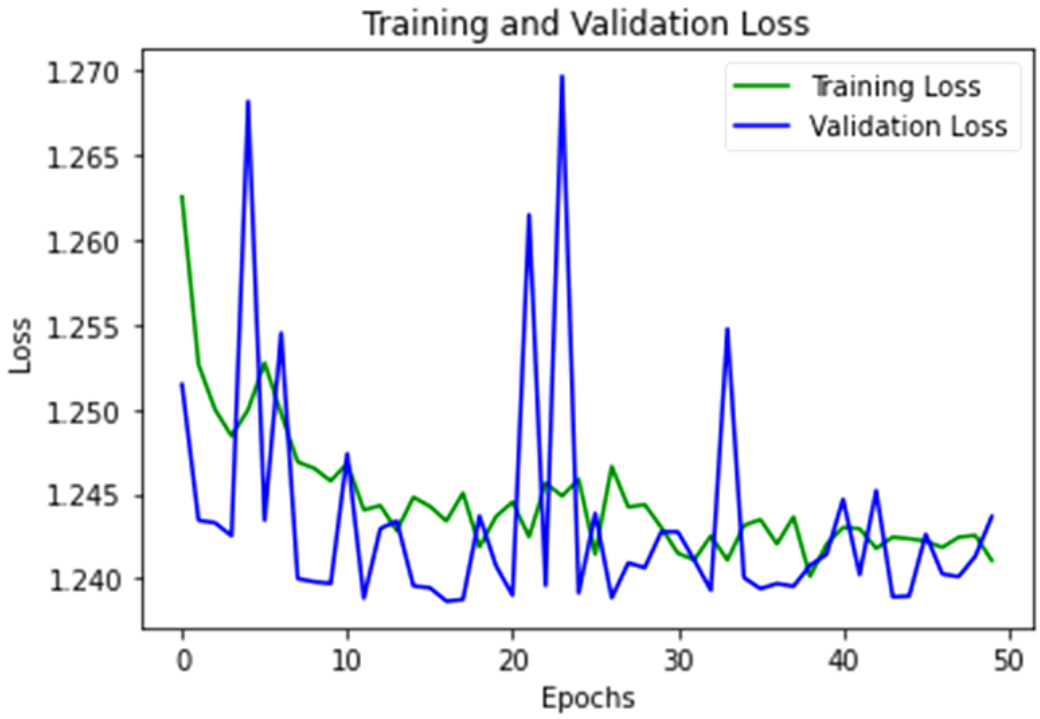
Model	Accuracy	Params	FLOPs	Recall	Precision	F1 Score
Vgg16	96.13%	138 M	15.5 G	96.2%	96.19%	96.20%
Resnet18	95.19%	11.5 M	1.71 G	95.19%	95.27%	95.23%
MobilenetV3	87.42%	5.4 M	0.22 G	87.11%	87.27%	87.19%
Efficientnet-b0	90.44%	5.3 M	0.41 G	90.19%	90.23%	90.21%
ViT-small	96.51%	22 M	4.24 G	96.27%	96.35%	96.31%
DeiT-small	95.56%	5.0 M	1.3 G	95.66%	95.65%	95.66%
Swin-tiny	96.94%	29 M	4.5 G	95.19%	95.27%	95.23%
ConvViT (ours)	96.85%	9.5 M	0.98 G	95.19%	95.21%	95.19%

06

RESULTS ACHIEVED

Training and validation accuracy





07

CONCLUSION

We found that the Transformer-based CNNs model can significantly improve the identification of apple leaf diseases in complex backgrounds.

- First, the input image is modeled locally using the convolutional structure to extract image features; the local nature of the convolutional operation makes it difficult to communicate between features that are far away, resulting in the extracted features containing parts with complex backgrounds, which affects the identification results. We solve this problem by connecting the Transformer structure after the convolutional structure. The Transformer optimizes the feature map by modeling the feature map at long distances, which is equivalent to guiding the CNN structure to focus on features useful for recognition, helping the CNN to see better.
- Secondly, the original patch embedding method is improved to retain more edge information of the input disease image and increase the continuity of the local information of the image.
- Finally, the computational complexity of the convolutional operation is reduced by using a depth-separable convolution method, and the computational complexity of the MHA algorithm is effectively reduced to linear complexity by using Global Average Pooling before performing the attention operation. The model's parameters and FLOPs are significantly reduced, enabling ConvViT to be applied to real-world scenarios. Compared with experimental results on other dominant network structures, the model achieves competitive recognition accuracy on a self-built apple dataset with much lower parameters and FLOPs than other models with the same performance.