Cassava Leaf Diseases Classification using DeiT and SWin Transformer

Pio Lauren T. Mendoza
Institute of Electrical and
Electronics Engineering
University of the Philippines - Diliman

***Abstract*—Cassava plants heavily suffer from leaf diseases. Early detection of these diseases can improve the plants' yield. These diseases can be identified through the aid of computer vision techniques. In this project the use of transformer based classifier networks for cassava leaf disease detection were explored. Pretrained DeiT-Small and SWin-Tiny models were fine tuned on cassava leaf diseases dataset. DeiT-Small model was able to achieve 85.73% accuracy while SWin-Tiny achieved an accuracy of 87.8%.**

I. Introduction

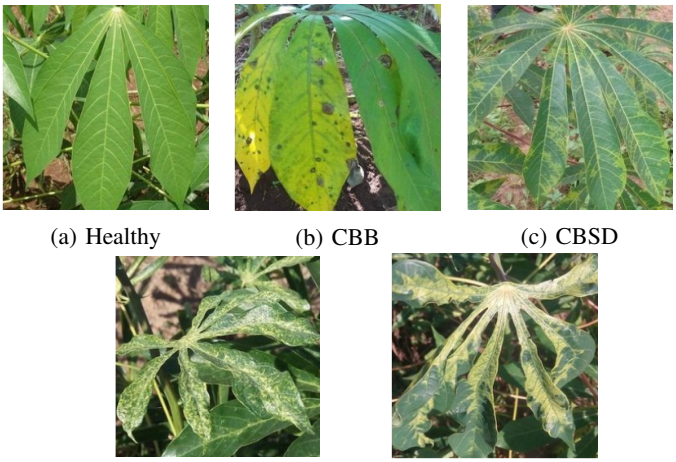
Cassava is one of the most carbohydrate-rich food. It is usually processed into food products such as puddings, cakes, and dried chips. It is massively produced in tropical countries for their consumption and exports. In the Philippines alone 722k metric tonnes of cassava were produced during the 3rd quarter of year 2021 [1].

Cassava plants are resistant to various whether conditions. It can grow in either the rainy or sunny seasons. Unfortunately, it is highly susceptible to various diseases. These diseases are seen on the cassava leaves. Even though these diseases have varying level and effects to the cassava plants, their appearances may look the same on an untrained eye. Identifying the diseases manually by the experts is costly and tedious. Given this, automated classification through the use of computer vision methods are being studied. One of them is through the use of machine learning. These methods are much efficient, low-cost and are scalable [2], [3].

II. Related Work

A. Cassva Dataset

Building cassava leaf disease dataset is quite costly. Fortunately, Artificial Intelligence lab in Makerere University and the National Crops Resources Research Institute, (NaCRRI) released their crowdsource dataset [2]. It is composed of 5 classes namely Cassava mosaic disease (CMD), Cassava brown streak disease (CBSD), Cassava bacterial blight (CBB), Cassava green mite (CGM), and the healthy class. Their samples images are shown in figure 1. The dataset contains 9,436 labeled images. However, the dataset is unbalanced 72% of the images are under the disease classes CMD and CSBD. Other dataset is also found in [3]. The dataset has more types of disease classes to offer. In their study [3] they have used two types of dataset a main cassava dataset and leaflet cassava dataset. The later are cropped images of the leaflet of the cassava plants. Regrettably, it is not available to the public.



(a) Healthy (b) CBB (c) CBSD

(d) CGM (e) CMD

Fig. 1: Cassava Diseases seen on Leaves [2]

B. Vision Transformer

The transformer architecture revolutionized the field of Natural Lanuage Processing (NLP) in the year of 2017. It utilizes an encoder-decoder network. The decoder network uses attention which was first seen in [5] while the encoder network uses self-attention which was first introduced in [6]. Currently, transformer architecture is the de facto standard for NLP problems. From then on various researches of possible usage of transformer networks in computer vision were done. In 2020, the landmark paper [4] were published on arXiv. It introduced the Vision Transformer (ViT) which is used for image classification. The Vision Transformer architecture, shown in figure 2, is not novel. It is the exactly the encoder part of the transformer [7]. The network divides an image into smaller sub-images called patches. Patches are analogous to the token input of the transformer. These patches are then converted into vector. Then, positional embeddings are added to the patches. But, unlike the Transformer, ViT has learnable positional embeddings. Aside from the patches, another token is utilize called the CLS token. These tokens are then fed into stack of Multi-head Self Attention (MSA) layers. Finally, a dense layer is added on top of the CLS token to compute for the classification output vector. ViT beats ResNet [8], a CNN based architecture, on publicly available dataset by a small

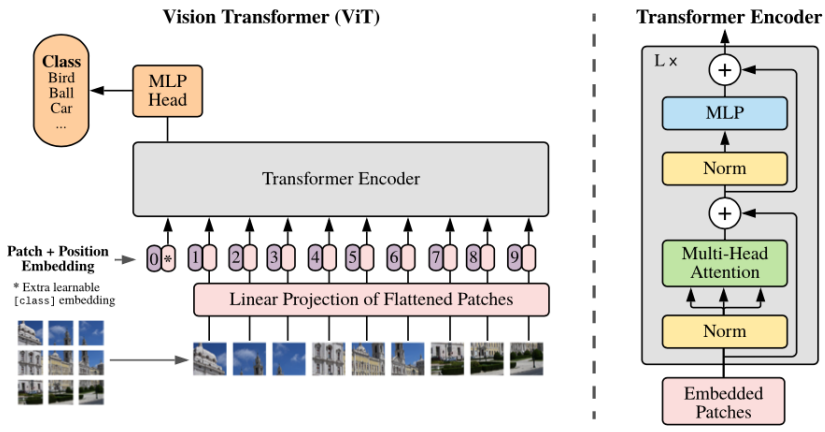


Fig. 2: ViT Architecture [4]

margin (87.54% vs 87.76%). This is due to the fact that ViT were trained on Google private dataset, the JFT, which has 300 million images. ViT is data hungry. ResNet still reigns supreme against ViT that are trained with dataset that has less than a 100 million images.

C. Data-Efficient Image Transfomers

Three months after ViT was released, Facebook introduced their Data Efficient-Image Transformer (DeiT) [9]. They were able to achieve better performance and image throughput as compared to ViT if both models are trained with ImageNet (77.9% vs 83.4%; 85.9image/s vs 290.9image/s). They mitigated the ViT's data hunger through the use of distillation, regularization, and augmentation techniques. The distillation technique caused the drastic reduction of the needed dataset size and training time. Distillation is a way of transferring the learnt knowledge of one network, the teacher network, to another, the student network. DeiT introduced two distillatin strategies, the hard and soft distillation. For the soft distillation, the error between the output distribution of the teacher and student networks are added to the loss function. On the other hand, hard distillation treats the classification of the teacher as a true label. Its error with respect to the output of the student is added to the loss function. This is done through the use of distillation token as seen in figure 3. Unlike the class token whose goal is to learn the true label, distillation token aims to learn the label produced by the teacher network. DeiT uses the state-of-the-art CNNs as the teacher network for its distillation process.

D. Shifted Windows Transformer

Shifted Windows Transformer (SWin Transformer) [10] is a general-purpose computer vision backbone. It is also based on ViT but it is hierarchical and uses the concept of shifted windows. Windows are groups of image patches. Unlike ViT and DeiT where attention of one patch is computed for all the patches, in SWin transformer, attention of a single patch is only computed for the patches within the same window. This drastically diminishes the time complexity from quadratic to

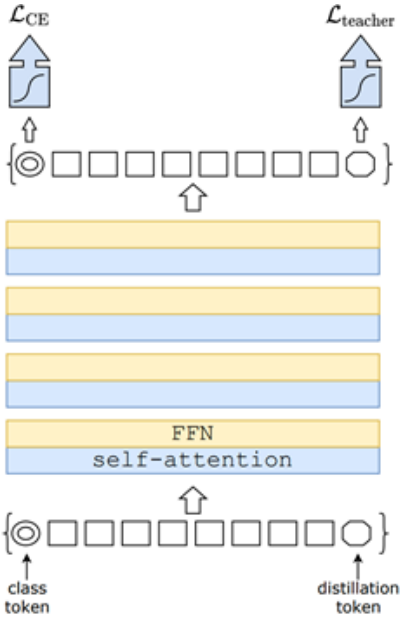


Fig. 3: DeiT Architecture [9]

a linear time complexity, with respect to the image resolution. The shifting is done to promote cross-window connection. SWin transformer blocks always come in pair like shown in figure 4. For each stage, neighboring patches are merge. The SWin transformer performs better as compared to ViT with similar number of parameters (77.9% vs 83.5%).

III. Method

In this project DeiT and SWin transformer were used for classifying cassava leaf diseases. To be more specific, DeiT-small and SWin-Tiny were used for the classification task. DeiT-small has 22M parameters while SWin-Tiny has 28M parameters. These models were created using the Timm: Pytorch Image Models library [11]. The pretrained weights of the networks were used while replacing their heads with the appropriate linear layers. The pretrained transformer weights

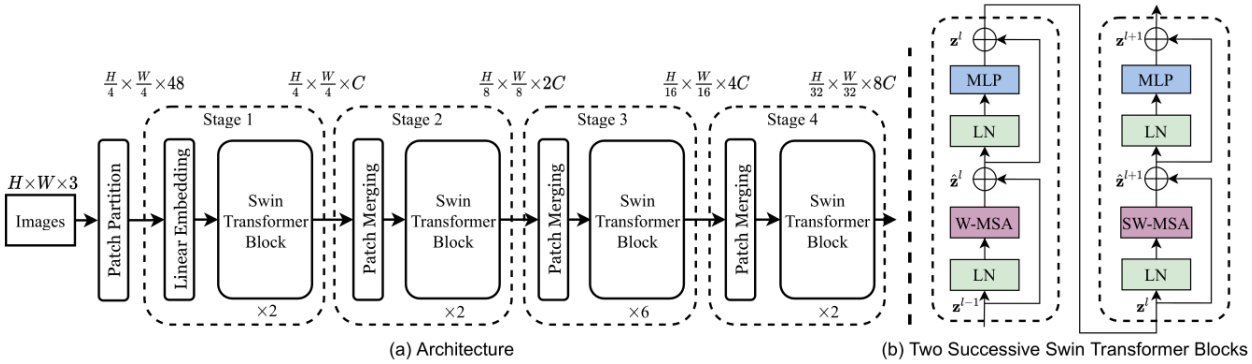


Fig. 4: SWin Transformer Architecture [10]

were leveraged since they have great feature extraction performance and also to lessen the needed training time.

The networks were trained using [2] dataset. The train dataset was transformed using various augmentation as suggested by [9]. The following are the augmentation done on the training images:

- Resize and crop
- Horizontal flip
- Vertical flip
- Random brightness-contrast adjustment
- Shift-scale-rotate
- Normalization using ImageNet mean and variance
- Coarse dropout

The dataset balancer of Pytorch were also setup and used since the cassava dataset is unbalanced. The validation split were also used during training to verify the training status. It was transformed by resizing the images to 224×224 pixels and normalizing it by using Imagenet mean and variance. Transfer learning was done. Various learning rates and batch sizes were experimented in finding the optimal network. Their performances are measured in terms of the networks accuracy.

IV. Results and Analysis

The training specs of the best models are shown in table I. The networks body were first frozen while training the head. Afterwards, the whole network is unfrozen and were trained. The models experienced some overfitting after 7 epochs. The final validation loss and validation accuracy of the DeiT is 0.511 and 84.4%, respectively. On the other hand, the SWin Transformer achieve a final validation loss of 0.39 and final validation accuracy of 87.2%. DeiT was trained for 27.5 minutes while the Swin transformer was trained for 42 minutes.

Parameter	DeiT Value	SWin T Value
Batch Size	16	8
Optimizer	AdamW	Adam
Learning Rate	1e-4	1.5e-5
Epoch	10	10

TABLE I: Training Hyperparameters

After training, the networks were tested against the never been seen test split. The final test accuracy for the DeiT and SWin Trasnformer models are 85.73% and 87.8%, respectively. The SWin Transformer model has more than $\sim 2\%$ accuracy bump as compared to DeiT model. The confusion matrix of the DeiT is shown in figure 5 while for the SWin Transformer model is shown in figure 6. The DeiT model was able to differentiate disease class CBB, 70%, even though it has very little amount of samples as compared to disease class CMD. However for the SWin Transformer, it has a very low capacity to identify the CBB class, 58%, even though it had a better performance as compared to DeiT model.

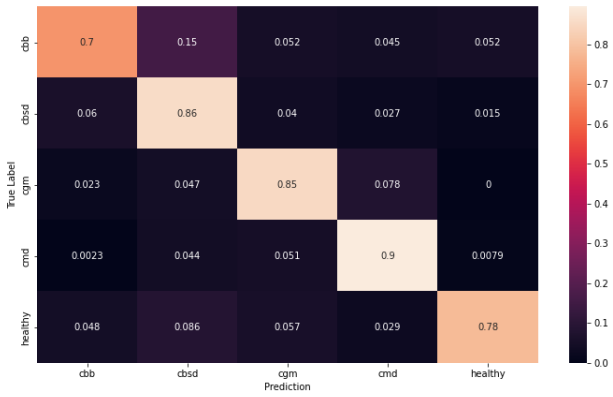


Fig. 5: DeiT Confusion Matrix

In [2] they have tallied the results of their Kaggle competition [12]. The best model was able to classify 93.86% of the test dataset. Unfortunately, the solution is private and no details about its implementation was discussed. But probably is not a transformer based solution since it was released in 2019. This project models are still far from the best but is also able to perform well its task of classifying the cassava leaf diseases.

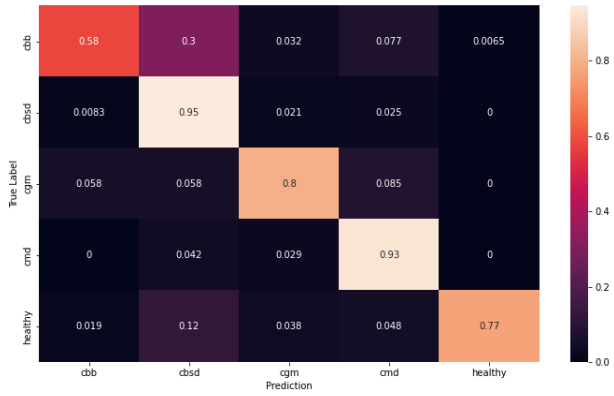


Fig. 6: SWin Transformer Confusion Matrix

V. Conclusion

In this project it was shown that DeiT and SWin Transformer models can perform the task of cassava leaf diseases detection. Fine tuning of the pretrained networks were done to be able to achieve detection accuracy of 85.73% and 87.8% for the DeiT and SWin Transformer models, respectively. These results are still from perfect. Further optimization must be done. The dataset [2] has another unlabeled split. This can be utilized for unsupervised learning for vision transformer models [13], for the further studies.

References

[1] D. S. Mapa, "Major vegetables and root crops quarterly bulletin." [Online]. Available: https://psa.gov.ph/sites/default/files/Major%20Vegetables%20and%20Rootcrops%20Quarterly%20Bulletin_July-September%202021_0.pdf

[2] E. Mwebaze, T. Gebru, A. Frome, S. Nsumba, and J. Tusubira, "icassava 2019 fine-grained visual categorization challenge," *arXiv preprint arXiv:1908.02900*, 2019.

[3] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep learning for image-based cassava disease detection," *Frontiers in plant science*, vol. 8, p. 1852, 2017.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[6] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[11] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019.

[12] "Cassava disease classification." [Online]. Available: https://www.kaggle.com/c/cassava-disease/overview

[13] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," *arXiv preprint arXiv:2105.04553*, 2021.