Automatic Classification of Pollen Grain Images using Vision Transformer

Project Report for ECCE 633 Machine Vision and Image Understanding

Submitted By:

Adarsh Ghimire (100058927)

Selina Shrestha (100058926)

Submitted To:

Dr. Hasan Al Marzouqi

1. Introduction

The classification of pollen grains from their microscopic images has several applications in the field of medicine and biology as it allows the detection of allergens and prediction of potential allergy symptoms [1]. The performance of a pollen classifier depends on two tasks: distinguishing pollen grains correctly in images from non pollen grains, and classifying them. Traditionally, the pollen grain images are studied and classified by a qualified pathologist. However, this requires a considerable amount of processing time and is affected by subjectivity. Thus, to leverage the advantages of pollen grain classification in aerobiology, fast and automatic classification of the pollen images without human intervention is of great importance.

Owing to the popularity of machine vision algorithms in automatic image classification, one of the key areas of present-day research in pollen grain classification involves the use of machine vision [2]. In comparison to standard image processing techniques, machine learning based solutions have been found to perform the task of pollen grain classification better [3]. Earlier works for machine learning based approaches use hand-crafted features along with a machine learning classifier [4, 5, 6]. However, these hand-crafted features do not work so well because of the similar structural appearance of pollen grains. Thus, deep learning approaches, specifically convolutional neural networks (CNN), which do not require manual feature engineering have been found to be better suited for the task. Since the publicly available datasets for labelled pollen grain images are small in size, transfer learning approaches with pre-trained CNN models have been presented in the literature.

Vision transformers are getting highly popular in the computer vision domain for its ability to incorporate attention mechanisms and encode the contextual information of the image [7]. This has helped to bolster the research in the computer vision community since the attention mechanism can play a significant role in making the decision, resulting in substantial improvements in performance. The transformer models however have not been used for pollen grain classification yet. In this project, we explore the two forms of state-of-the art vision transformers for pollen classification tasks. We use the publicly available dataset in the ICPR 2020 Pollen challenge [3] for the exploration and testing of vision transformers for classification of pollen grains. The dataset contains labelled images of 3 classes of pollens and another class of

objects that could be mis-classified as pollens. Thus, the problem at hand is a 4 class classification task.

1.1 Contributions

1) This is the first work that applies vision transformers for microscopic pollen grain image classification
2) Exploration of different vision transformer architectures
3) Ensembling of multiple high performing transformer models to enhance the overall performance

2. Literature Review

Early works based on pollen grain classification based on classical machine learning use hand crafted features to train machine learning classifiers. [4] uses morphological features of the pollen grains with a support vector machine (SVM) classifier while [5] uses the texture features along with several known classifiers like SVM, KNN, multi-layer perceptrons. [6] on the other hand uses hybrid features. These works are however limited by challenges faced by the hand-crafted features because of similar morphological appearances of the pollen grains and other artifacts that might be present in the images.

To overcome the limitations posed by hand crafted features, several CNN based deep learning approaches that perform automatic feature engineering have been proposed. [3] has trained the AlexNet and SmallerVGGnet architectures on the proposed ICPR 2020 Pollen challenge dataset with data augmentation. The deep CNN architectures however have millions of trainable parameters and thus require very large datasets in order to perform well. Since the datasets available for pollen grain images are small in size, most of the recent works fine-tune CNN models that have been pre-trained on large image datasets. [8] uses a pre-trained AlexNet and fine tunes the last three layers on the Pollen23E dataset for pollen grain classification. Similarly, [9] fine-tunes the pre-trained AlexNet network on their own dataset obtained by capturing dark field microscopic images on the Classifynder system. However, the performances of these approaches are still not up to the mark for practical applications. An ensemble based approach that uses four fine-tuned CNNs: EfficientNetB0, EfficientNetB1, EfficientNetB2, and

SeResNeXt50 for classification of pollen images in the ICPR 2020 Pollen challenge dataset is proposed in [9].

Vision transformers [7] have been emerging as competitive alternatives to the CNNs in the field of computer vision due to their ability to perform attention mechanisms by learning important regions in images for better performance. While this has been applied in numerous computer vision related research areas, it has not been used in pollen grain image classifications yet. Thus, in this project, we explore the use of the state of the art vision transformer [7] for automatic pollen grain image classification using the ICPR 2020 Pollen challenge dataset.

3. Methodology

3.1 Vision Transformer

Transformers are deep learning models that have self-attention layers that assign different weights to different parts of the input data according to their significance. These transformers have been performing very well in NLP models and are now being applied to images for image recognition tasks, giving the name "Vision Transformer". The architecture of a vision transformer (ViT) is shown in the figure below.

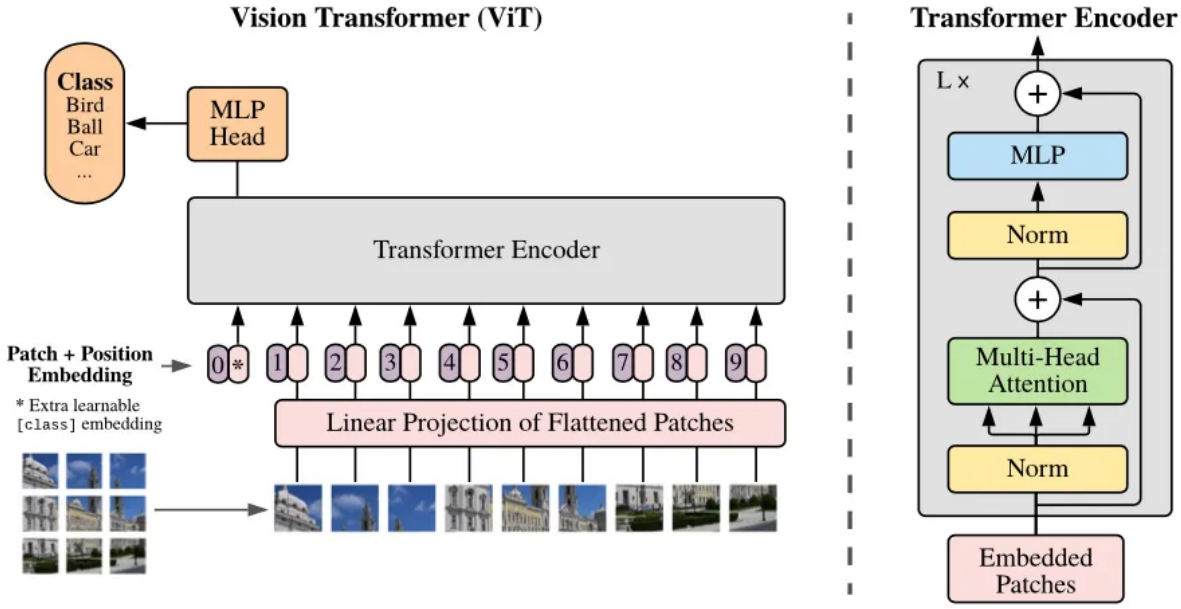
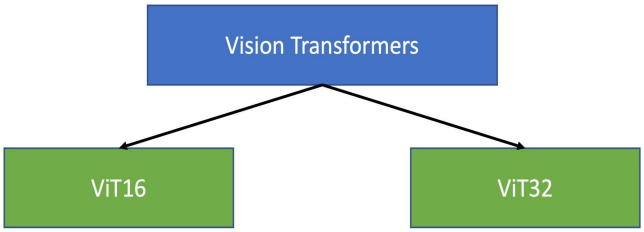


Figure: Vision Transformer Architecture [7]

The images are first split into fixed sized patches. Then, the patches are flattened and the linear projections of the flattened image patches are sent into the position encoder. Position embeddings are added to the patch embeddings to retain positional information. Next, the sequence of patches and position embeddings are fed into the transformer encoder. The transformer encoder consists of a series of blocks, each with a multi-headed self attention mechanism followed by a multi-layer perceptron. Each consecutive block encodes more refined contextual cues. In practice, ViT models pre-trained on a large image dataset are used via transfer learning for specific applications. The features generated by the ViT can be fed into MLP layers for specific classification tasks.

3.2 Model Architecture Selection



In this project, we consider two ViT models: ViT16 and ViT32. Architecturally, both the models are the same with 11 multi-headed self attention blocks. ViT16 divides the input image into 16 patches while ViT32 divides it into 32 patches. A higher number of patches means that the transformer encoder is forced to learn or encode the finer details between the patches. To explore the effects of higher numbers of patches versus fewer numbers of patches, we test for both the models in our work.

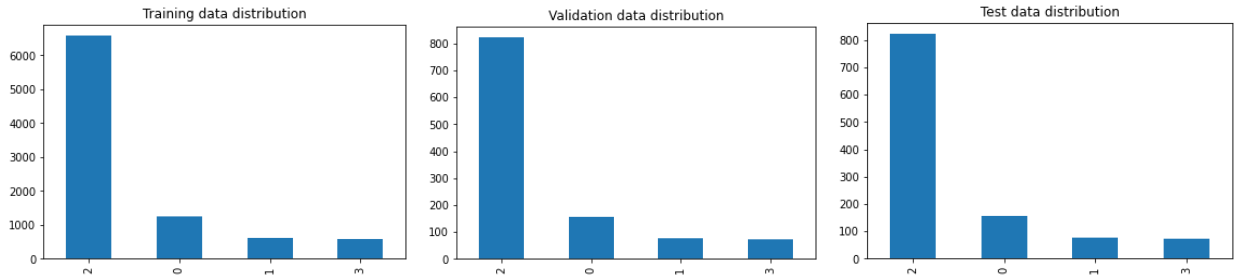
The transformer architecture has numerous trainable parameters. Thus, to train them, a large amount of data is required. However, in our case, the dataset available for pollen grain classification is not sufficient to train the transformer from scratch. So, we have used the weights of a transformer that was pre-trained on a large ImageNet dataset. The output of the pre-trained transformer is fed into a linear layer classification head that is trained to perform pollen grain classification.

3.3 Data Preparation

In this work, the ICPR 2020 Pollen challenge dataset has been used. The dataset contains 11,279 images of 4 categories:

1) Corylus avellana, well-developed pollen grains
2) Corylus avellana, anomalous pollen grains
3) Alnus, well-developed pollen grains
4) Debris (bubbles, dust and any non-pollen detected object)

The dataset is divided into training set (80%), validation set (10%), and test set (10%). And the dataset was divided into training, validation and test set was done such that distribution of the data remained the same in each. The inter-class distribution are maintained to be uniform over the training, validation and test sets as shown below:



To ease the preparation of the data for training, a lookup table for the image name and the corresponding label was generated. The labels assigned in the dataset are from 1 to 4. This has been updated to 0 to 3 to align with the class numbering convention used in the MLP output layer. Then, the following image preprocessing steps were performed:

1) Image resizing to 224x224
2) Image augmentation: vertical flip with probability of 0.3, horizontal flip with probability of 0.3, cropping
3) Normalization

3.4 Fine-Tuning Experiments

During training, for both ViT16 and ViT32 models, we add an MLP classifier with 4 prediction heads and unfreeze some of the last blocks of the pre-trained ViT. We then fine-tune the model on the pollen grain dataset. Several experiments have been performed by varying the number of unfreezed last blocks in the pre-trained transformer.

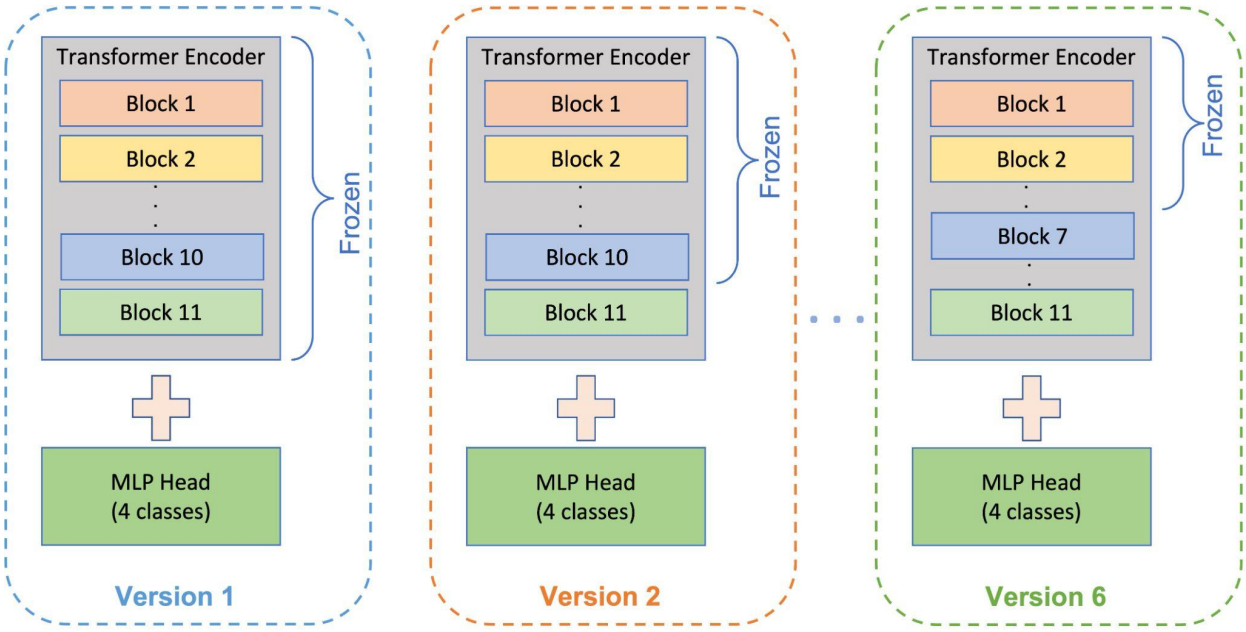


Figure: ViT fine-tuning [7]

The different experiments performed to fine-tune the models are shown in the table below.

ViT Model	**Version**	**Trainable Layer/s**
ViT16	1	MLP Head only
	2	MLP Head + Block 11
	3	MLP Head + Block 11 + Block 10
	4	MLP Head + Block 11 + Block 10 + Block 9
	5	MLP Head + Block 11 + Block 10 + Block 9 + Block 8
	6	MLP Head + Block 11 + Block 10 + Block 9 + Block 8 + Block 7
ViT32	1	MLP Head only

	2	MLP Head + Block 11
	3	MLP Head + Block 11 + Block 10
	4	MLP Head + Block 11 + Block 10 + Block 9
	5	MLP Head + Block 11 + Block 10 + Block 9 + Block 8
	6	MLP Head + Block 11 + Block 10 + Block 9 + Block 8 + Block 7

3.5 Training Process

For each case of fine-tuning described in the previous table, 2 different trainings were performed using the Adam and Adabelief optimizer respectively. The loss function used is the categorical cross-entropy and the performance metrics used are the commonly used ones for classification: precision, recall, F1 score, accuracy, weighted average, and macro average. Each model was trained for upto 10 epochs since models generally reached saturation in less than 10 epochs. The model that performs the best on the validation set within the 10 epochs is saved for each case. As well as, for later training if required, the latest checkpoints are also saved.

3.6 Model Ensembling

To further improve the performance, the best performing models with a validation accuracy greater than 95% were nominated for ensembling. Additionally, by analysing the confusion matrix of these best performing models, final models were selected. The predictions from these models were combined and weighted to get an enhanced performance.

4. Results

The results of all the models obtained by different fine-tuning settings using adam and adabelief optimizers for ViT16 and ViT32 models are shown in the table below.

ViT Model	Version	Trained Layers	Validation Set					
			Accuracy (%)		Macro Average		Weighted Average	
			Adam	Adabelief	Adam	Adabelief	Adam	Adabelief
ViT16	1	MLP Head only	92.4	92.7	0.86	0.86	0.92	0.92
	2	Up to Block 11	94.2	94.5	0.89	0.9	0.94	0.94
	3	Up to Block 10	94.5	**95**	0.9	0.91	0.94	0.95
	4	Up to Block 9	94.5	**95.1**	0.9	0.92	0.94	0.95
	5	Up to Block 8	94.6	94.7	0.91	0.91	0.95	0.95
	6	Up to Block 7	94.9	**95.2**	0.91	0.92	0.95	0.95
ViT32	1	MLP Head only	92.3	91.9	0.88	0.87	0.92	0.92
	2	Up to Block 11	94.5	94.5	0.91	0.92	0.94	0.94
	3	Up to Block 10	94.1	94.6	0.9	0.92	0.94	0.94
	4	Up to Block 9	94.5	**95.2**	0.91	0.91	0.94	0.95
	5	Up to Block 8	94.9	**95.1**	0.92	0.92	0.95	0.95
	6	Up to Block 7	94.5	**95.6**	0.91	0.94	0.94	0.96

In the table, the models with validation accuracy greater than 95% have been highlighted in yellow, with the best model highlighted in green. The top-6 models were selected for ensembling to enhance the model performance further. It can be observed that all the best performing models are the ones that were trained using the Adabelief optimizer. Additionally, it can be seen that the best performing model is the one that uses ViT32. However, there is a very slight difference in the performances of models using ViT16 and ViT32. Thus, it can be deduced that changing the number of patches from 16 to 32 does not affect the model performance much.

By analyzing the confusion matrix of the 6 selected models, it was observed that 4 models had comparatively lower False prediction. The other 2 models falsely predicted other classes as class 2. Thus, the 4 models given below with low false prediction were selected for further consideration.

Model	ViT	Version	Trained Layers
1	ViT16	4	Up to Block 9
3	ViT16	6	Up to Block 7
3	ViT32	5	Up to Block 8
4	ViT32	6	Up to Block 7

For ensembling, all possible combinations of these 4 models were generated and evaluated on the validation set. Each individual model was assigned an equal weight in the ensembled model. The results are shown in the table below.

Model Combination	Validation Set		
	Accuracy (%)	Macro Average	Weighted Average (%)
1,2	95.656	0.928	0.956
1,3	95.922	0.931	0.959
1,4	**96.365**	0.941	0.963
2,3	96.188	0.939	0.961
2,4	96.011	0.938	0.960
3,4	96.099	0.940	0.960
1,2,3	96.188	0.936	0.961
1,2,4	96.188	0.939	0.961
1,3,4	**96.454**	0.944	0.964
2,3,4	96.277	0.945	0.963
1,2,3,4	**96.454**	0.944	0.964

The top 3 models highlighted in yellow with the best validation accuracies were selected for final evaluation on the test set. The test accuracies reported for the final 3 ensembled models are given below.

Model Combination	Test Set		
	Accuracy(%)	Macro Average	Weighted Average
1,4	94.592	0.908	0.945
1,3,4	94.858	0.912	0.948
1,2,3,4	94.947	0.914	0.949

It can be seen that the ensemble model combining models 1,2,3, and 4 was able to achieve the best test accuracy of 94.947 %.

4.1 Performance Comparison with other works

The following table shows the test accuracy performance of our best ensemble model (combination of 1,2,3,4) in comparison to the test accuracies reported by other works in the literature.

Models	Accuracy(%)	Weighted Average
Alex-Net [3]	89.63	0.890
SmallVGGnet [3]	89.73	0.891
Ensembled networks [9]	94.48	0.945
Our Best Ensemble Model	**94.94**	**0.949**

It can be observed that our ensemble model outperforms the works presented in the literature. The works submitted in the ICPR 2020 Pollen challenge leaderboard however report a test accuracy of up to 97.5 %. But the results in the leaderboard are reported by testing on a different test set, which are not publicly disclosed. Thus, we cannot objectively compare our test scores to the ones reported in the leaderboard.

5. Conclusion

The state-of-the art multi-head self attention technique of vision transformers was applied for microscopic pollen grain image classification using the ICPR 2020 Pollen challenge dataset. Two pre-trained ViT models: ViT16 and ViT32 were fine-tuned on the dataset by varying the blocks unfreezed for training. The model obtained by unfreezing up to block 7 and trained using Adabelief optimizer was found to perform the best on the validation set. Furthermore, few best performing models were ensembled to enhance the model's performance. It is observed that our vision transformer based ensemble model is able to outperform CNN based models presented in the literature by a noticeable margin. The lower performance of CNN models might be attributed to the fact that they are prone to utilize local interaction between elements in the input domain. The transformer model, however is able to to utilize local interactions in a global context due to its self-attention mechanism, resulting in a higher performance.

Future developments could include trying other transformer architectures with different hyperparameter tuning and the ensembling of transformer based models with high performing CNN models.

References

[1] H. Ribeiro, S. Morales, C. Salmer´on, A. Cruz, L. Calado, M. I. Rodr´ıguez-Garc´ıa, J. D. Alch´e, and I. Abreu. Analysis of the pollen allergen content of twelve olive cultivars grown in Portugal. *Aerobiologia*, 29(4):513–521, 2013.

[2] Holt, Katherine & Bennett, Keith. (2014). Principles and methods for automated palynology. *New Phytologist*. 203. 10.1111/nph.12848.

[3] S. Battiato, A. Ortis, F. Trenta, L. Ascari, M. Politi and C. Siniscalco, "Detection and Classification of Pollen Grain Microscope Images," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 4220-4227, doi: 10.1109/CVPRW50498.2020.00498.

[4] C. M. Travieso, J. C. Briceño, J. R. Ticay-Rivas and J. B. Alonso, "Pollen classification based on contour features," *2011 15th IEEE International Conference on Intelligent Engineering Systems*, 2011, pp. 17-21, doi: 10.1109/INES.2011.5954712.

[5] Fernández-Delgado, M & Carrión, Pilar & Cernadas, E. & Galvez, J.F. (2003). Improved Classification of Pollen Texture Images Using SVM and MLP. *3rd IASTED International Conference on Visualization, Imaging and Image Processing (VIIP2003)*. 2.

[6] Chica, Manuel. (2012). Authentication of bee pollen grains in bright-field microscopy by combining one-class classification techniques and image processing. *Microscopy research and technique*. 75. 1475-85. 10.1002/jemt.22091.

[7] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[8] Sevillano, Victor & Aznarte, José. (2018). Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. PLOS ONE. 13. e0201807. 10.1371/journal.pone.0201807.

[9] Mahbod, Amirreza & Schaefer, Gerald & Ecker, Rupert & Ellinger, Isabella. (2021). Pollen Grain Microscopic Image Classification Using an Ensemble of Fine-Tuned Deep Convolutional Neural Networks. 10.1007/978-3-030-68763-2_26.

Appendix

Complete project can be found in the github repository below:

https://github.com/ghimireadarsh/Pollen-Data-Classification-using-Vision-Transformer

Best model checkpoints and latest checkpoints can be found in the link below:

https://kuacae-my.sharepoint.com/personal/100058927_ku_ac_ae/_layouts/15/onedrive.aspx?id=%2Fpersonal%2F100058927%5Fku%5Fac%5Fae%2FDocuments%2FPollen%20Classification%20Models