Aphid Recognition using Multi-Scale Feature Representations with Vision Transformers

Fredrik-Otto Lautenbag
Supervisors: Henry Maathuis, Maya Aghaei Gavari

Winter 2023

Introduction

- Aphids are mainly responsible for infecting seed potatoes with viruses during their growth [1].
- To find out if pest control is more effective when applied at the right time, aphids must be accurately detected on crops.
- Currently, the detection of aphids is done by placing yellow sticky plates next to the agricultural crops and letting personnel check them.
- This research specifically investigates the classification performance of separating aphids from other insects.
- *Vision Transformers* have proven to be a well-performing alternative to CNNs [2].
- Single-scale architectures, based on CNN and ViT are used as baselines.
- The Cross-Attention Multi-Scale Vision Transformer (Cross-ViT) is used to classify the wide range of image resolutions and assess the impact of multi-scale feature representations [3].

Materials and Methods

- We gathered 6508 images. There are 682 images of aphids and 5826 that are not. Due to this class imbalance, the aphid images are oversampled.
- The labels for the images are determined in cooperation with a domain expert in which aphids are separated from non-aphids. Figures 1 and 2 depict examples of aphids and non-aphids.
- Multiple data cleaning iterations are performed to improve the quality of the annotations.

Abstract

We evaluated Multi-Scale and Single-Scale architectures to improve aphid classification on images. As a result, we used three different deep-learning models (ResNet, ViT, and Cross-ViT) to classify aphids using our own dataset. The model with the highest F1-score (84.88%) was Cross-ViT. Images are tokenized into various sizes because Cross-ViT is based on ViT but has been expanded to a Multi-Scale architecture. The Multi-Scale approach shows promising performance.

Figure 1. Examples of non-aphids stuck on the yellow sticky plate.

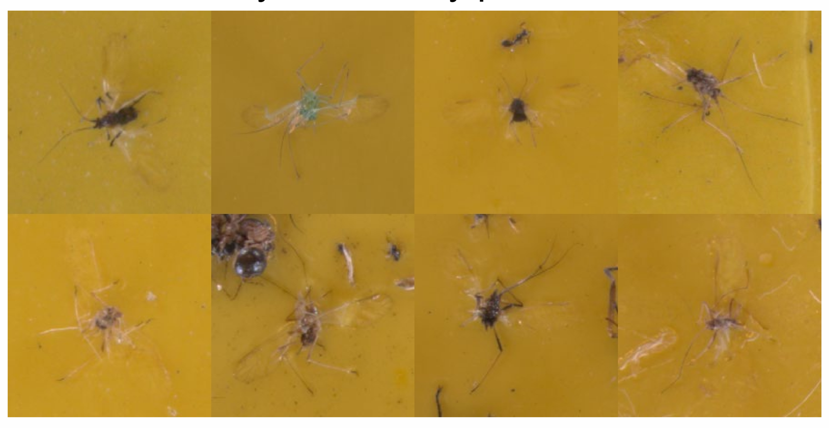

Figure 2. Example of aphids stuck on the yellow sticky plate.

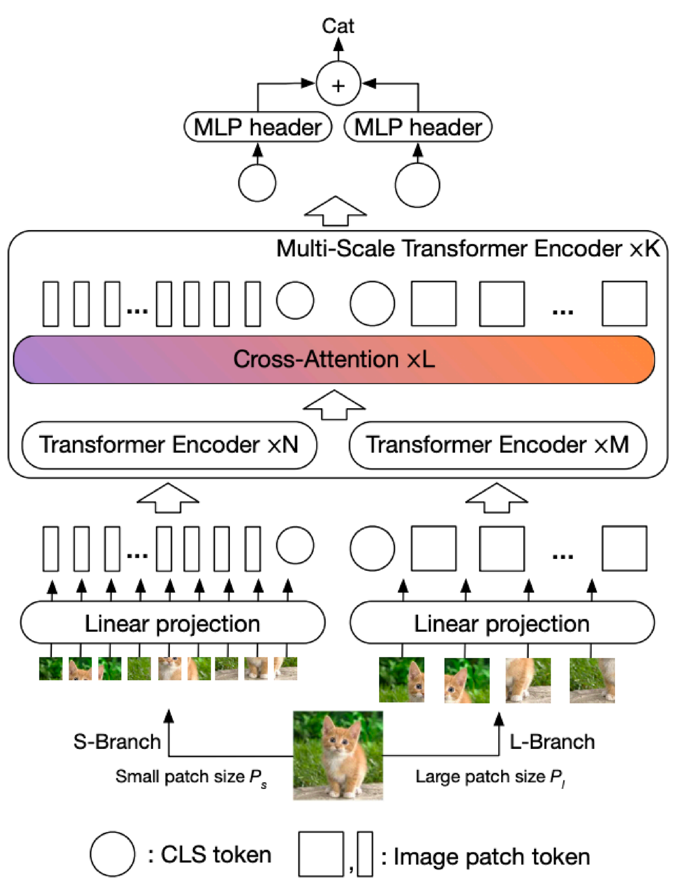

Figure 3. Cross-Attention transformer architecture illustration

- Three architectures were evaluated:
  - ResNet
  - Vision Transformers (ViT)
  - Cross-Attention Multi-Scale ViT (Cross-ViT) [3], as illustrated in Figure 3
- ResNet and ViT, being single-scale approaches, while Cross-ViT is a multi-scale approach.
- Model selection is based on pre-training and the number of parameters.
- Augementations are applied to vary the orientation, colour and noise during training.
- The F1-score is used to compare the performance of the model.

Experiments and Results

- A grid search is used to find the best configuration for each of the three architectures.
- Each model was evaluated with 108 configurations.
- Each configuration was performed four times to collect average scores.
- Table 1 depicts the best-performing configurations sorted by the F1-score.

Model	Start LR	Patience	Batch size	Augmentations	F1-Score
ResNet-50	0,05	5	16	Enabled	84.95% ±2.09%
ViT16	0,0001	20	8	Disabled	86.51% ±2.15%
ViT32	0,0001	10	8	Disabled	85.33% ±0.77%
Cross-ViT 12/16	0,0005	10	8	Enabled	85.33% ±1.37%

Table 1. Results of the grid search performed on the validation set.

- ViT16 is selected as a baseline since it outperformed ViT32.
  The numbers following the abbreviation, 16 and 32, are patch sizes.
- Augmentations improve ResNet-50 and Cross-ViT performance but not plain ViT's.

- The final experiments were done with the parameters found in the grid search (see Table 1).
- The final classification performance is depicted in Table 2, and more metrics are included to make our model selection clear.
- The ResNet-50 baseline model is 95.11 % accurate in classifying aphids.
- ResNet-50 was outperformed by ViT16 in terms of a higher F1-score of 1.94 % and a lower standard deviation of 3.66 %.
- Cross-ViT outperforms ViT16 by 3.51 % on the F1-score. Besides this, the standard deviation is lower.
- To understand the shortcomings of the classifier and dataset, some misclassified images are depicted in Figure 4.

Model	Accuracy	Precision	Recall	F1-score
ResNet-50	95.11% ±1.69%	90.41% ±2.56%	71.12% ±8.26%	79.43% ±5.70%
ViT16	95.81% ±0.61%	88.72% ±1.84%	75.29% ±4.45%	81.37% ±2.04%
Cross-ViT 12/16	96.54% ±0.28%	94.36% ±2.34%	77.18% ±2.20%	**84.88% ±1.06%**

Table 2. Final results from experiments on the test set.

Figure 4a. Example of multiple insects **Figure 4b.** Example of misclassified insect **Figure 4c.** Example of misclassified insect **Figure 4d.** Example of misclassified insect

Conclusions

- We can conclude that Multi-Scale representations of features with Vision Transformers improve classification for this problem.
- Model selection based on precision and recall was not needed since the Cross-ViT model outperformed ResNet-50 and ViT16 on all calculated metrics.
- The current classification performance provides a stable foundation for supplying the necessary data for the pesticide feasibility study.

Acknowledgements

This project is financially supported by ELFPO and performed within the POP3+ Fryslân project Innovatie luizendetectie.

NHL STENDEN computer vision & data science

provinsje fryslân provincie fryslân

GEORGE PARS GRAANHANDEL B.V.

Europees Landbouwfonds voor Plattelandsontwikkeling: Europa investeert in zijn platteland

GeJo Grading Services B.V.

Elyt. CONSULTANCY

Ana Vita

Fokkema Farming

References

- [1] J.A.L.M. Kamp, P.M. Blok, G. Polder, J.M. van der Wolf, and Henk Jalink. Smart Ziekzoeker 2015: Detectie van virus- en bacteriezieke pootaardappelen met behulp van vision- en sensortechnologie, volume 703 of PPO/PRI rapport.
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. CoRR, abs/2010.11929, 2020.
- [3] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. CoRR, abs/2103.14899, 2021.