Unmanned Aerial Vehicles powered by Vision Transformer for Precision Agriculture

Francesco Peragine
Department of Computer Science
University of Bari Aldo Moro
Bari, Italy
f.peragine@studenti.uniba.it

Abstract

With the growing world population, the need to increase crop yields and conserve resources in the context of precision agriculture for wheat is more pressing than ever. This study addresses this critical challenge by developing an AI-driven agricultural model based on Unmanned Aerial Vehicles (UAVs) that is powered by Vision Transformer (ViT) architecture to estimates the height of the crops in cultivated fields. By providing real-time informations about crop yields our model can help to optimize resources utilization and the financial burden, greatly enhance crop yields, create new jobs and opportunities in the agricultural sector and protect the environment. This interdisciplinary approach highlights the transformative potential of AI-based precision agriculture in addressing the intricate interplay between food security, environmental sustainability and agricultural productivity.

1. Introduction

1.1. Global challenge

The global population is currently estimated at around 8 billion people and is projected to reach 9.7 billion by 2050 and 10.4 billion by 2100. This growth will have a significant impact on agricultural production, which amounted to 2.5% of the world population with 1.2 billion tons and $180 billion in 2022 [1]. To meet the food needs of the population, production will need to increase by 70% [2].

1.2. Precision agriculture

The growing demand for corn is a major challenge for food production and to meet this demand it is necessary to increase agricultural productivity in a sustainable way. Precision agriculture can help to achieve this goal, by using data on crop characteristics and innovative technologies to help farmers make more informed decisions about how to manage their fields. Precision agriculture is based on the ability to accurately detect the needs of plants within a cultivated field, using a set of sensor technologies that observe their status in a non-invasive manner. This can lead to improved crop yields, reduced environmental impact and increased profitability.

1.3. Proposed solution

AI can furtherly boost precision agriculture, increase crop yields and reduce the use of resources. In this paper, we propose an approach based on the use of UAVs powered by Machine Learning (ML), through the use of a ViT architecture that has been enhanced by transfer learning and fine-tuning. UAVs are equipped with numerous sensors making them ideal for quickly collecting mutiple data, whilst ViTs models perform highly accurate quasi-real-time predictions. Therefore it is possible to provide farmers with instant informations about the state of the crops in a fully automated and cost-effective way.

2. Related Work

This work builds on the research "Unmanned Aerial Vehicles for High-Throughput Phenotyping and Agronomic Research" [3]. The authors provided four case studies involving multiple crops in breeding and agronomic applications developed by five teams: Administration, Flight Operations, Sensors, Data Management and Field Research. As the first project of its kind, it provided valuable lessons learned about critical informations of sensors, air vehicles and configuration parameters for both. Images from a large field of maize for a total of 1064 plots were collected using a UAV equipped with a high-resolution camera on July 22,

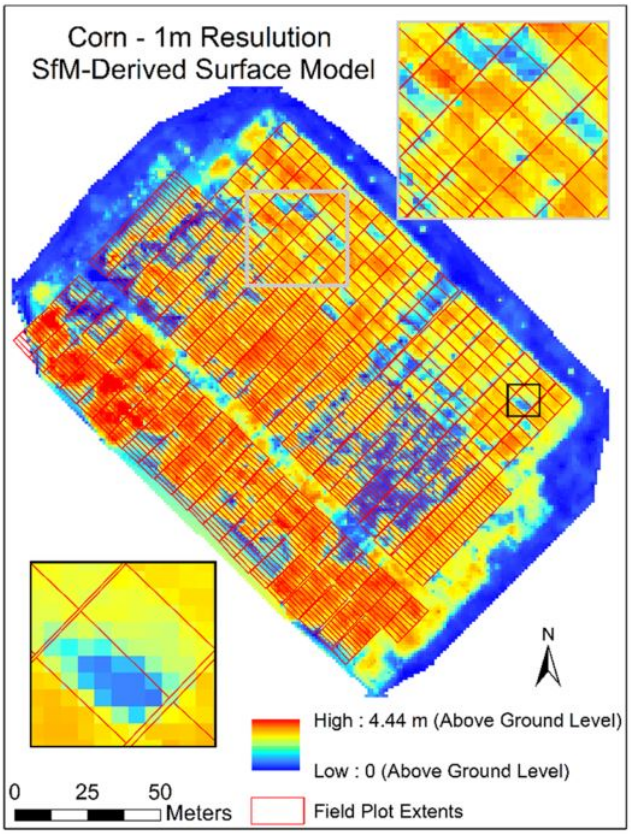

Figure 1. Digital surface models (DSMs). From [3].

2015. A digital surface model (DSM) [1] was calculated from the image data and used to estimate height. The height estimates were compared to ground truth measurements taken approximately three weeks earlier. The approach used in this study is based on linear regression and resulted in a correlation of (r^2 = 0.35; $p < 0.0001$) for the first case study, where only 705 plots were analyzed as a group because a significant number of plots sustained extensive feral hog damage happened in the time between ground truth and imaging. As per analysis of Shi *et al.*[3] the generally weak correlation between UAV estimates and ground truth plant height can be explained by several reasons.

> First, the fixed-wing UAV estimates did not have adequate resolution to distinguish the small tassels atop the plants, which were measured on the ground. Second, UAV images were collected about three weeks after ground truth data, and the plants had dried down significantly such that the plant canopy was not as erect as earlier in the season. Third, plot boundaries were manually drawn on the mosaicked image, and there was thus variability in pixel selection accuracy. Finally, there was some elevation variation in the bare ground DSM that was not taken into account in UAV estimates of plant height. Taken together, these issues suggest that process improvements are needed and that future results are likely to be improved.

Similar published works make use of the hyperspectral and multispectral imaging to quantify field scale phenotypic information accurately and integrate the data into AI models [4], or provide growth monitoring methods by using UAVs data [5].

3. Data

The dataset used in this study is made by around 1700 overlapping aerial images taken by drones at Texas A&M AgriLife Researchś Brazos Bottom Farm in Burleson County, Texas (headquarters at 30.549635N, 96.436821W) [6].

UAVs were operated by three separate flight teams and were equipped with several sensors, auto-pilot and stabilization systems. Most flights were made within 2.0h of solar noon. Flight and sensor parameters were configured to ensure collection of high-quality images with adequate overlap between images for mosaicking purposes, and to minimize pixel smearing. For this case study the images were taken with a X88 octocopter UAV at a typical flying altitude of 15-20m, with a DJIP3-005 4K Camera and a resolution of 4000x3000 pixels.

Approximate locations of raw images (longitude, latitude, and altitude) were recorded by an on-board GPS, however, its accuracy is not high enough for direct georeferencing. Eight Ground Control Points (GCPs) were installed around the study area for accurate geo-referencing, geo-correction, and co-registration of UAS data.

3.1. Preprocessing

Significant long-term challenges related to data collection/processing and interpretation of the processed data need to be addressed before breeders can fully embrace these systems. As raw data moves through the application development pipeline [2].

As the images was already been post-processed by the authors of the dataset, there was no need to perform any further preprocessing. Because of the time between the ground truth measurement and the generation of the DSM for height estimation, the ground truth values were averaged between the two.

GCPs are critical for georeferencing the images and therefore for the generation of the orthomosaic. The

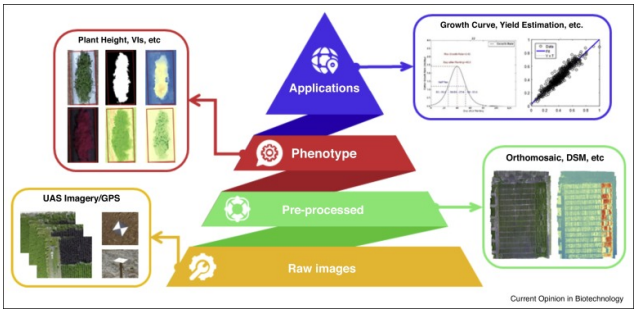

Figure 2. UAVs application development workflow. Figure from [4].

$61x61cm$ concrete tiles were installed at the corners and interior locations of all UAVsŕoutes and were painted black, dark gray and light gray (≈ 10%, 20% and 40% reflectance).

The GCPs coordinates have been used with the GCPEditorPro software to identify the tiles on the ground within the overlaying images collected by the UAVs. This lead to the georeferencing of the images that was supplied to the WebODM software to generate an orthomosaic of the field [3].

Finally, the plots were extracted with the pyshp library from the orthomosaic by using the provided shapefile and then matched to the ground truth records to assign an height value to each of them.

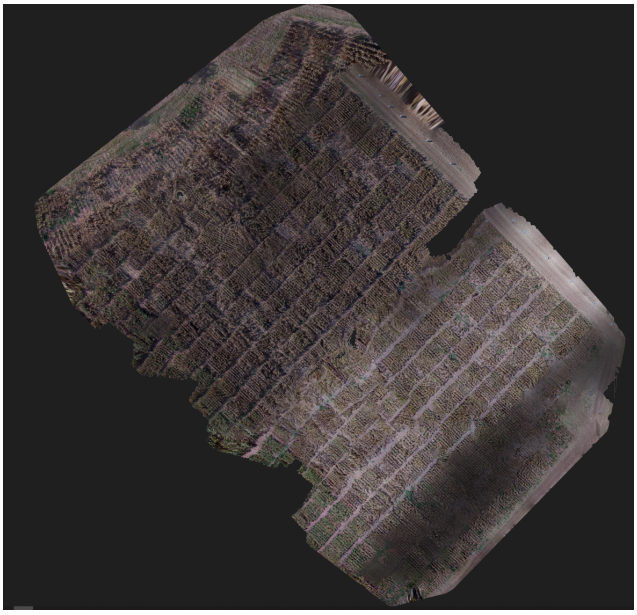

Figure 3. Orthomosaic of the agricultural field. Produced with WebODM.

4. Methods

The use of UAVs in agriculture is not a new concept, but the introduction of VITs models has been a game changer in the field of Computer Vision (CV). Unlike Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs), ViTs excel at capturing intricate image patterns with their self-attention mechanism. ViTs are highly scalable, reduce architectural complexity and support efficient transfer learning, therefore we decided to combine the two technologies to create a new solution for precision agriculture.

Our model design is based on the Vision Transformer architecture [7]. In this section, we provide an overview of the architecture and the attention mechanism.

4.1. Vision Transformer (ViT)

The Vision Transformer is a variant of the transformer architecture, which was introduced by Vaswani et al. [8] for machine translation.

An overview of the model is depicted in fig.[4]. The standard Transformer receives as input a 1D sequence of token embeddings. To handle 2D images, the image $x \in \mathbb{R}^{H\times W\times C}$ is reshaped into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches, which also serves as the effective input sequence length for the Transformer.

> The Transformer uses constant latent vector size D through all of its layers, so the patches are flattened and mapped to D dimensions. The output of this projection is referred as the *patch embeddings*. Position embeddings are added to the patch embeddings to retain positional information. The resulting sequence of embedding vectors serves as input to the encoder[7].

The Transformer encoder [8] consists of alternating layers of Multiheaded Self-Attention (MSA) [4.2] and MLP blocks. Layernorm (LN) [9] is applied before every block, and residual connections [10] after every block.

4.2. Multi-head self-attention (MSA)

The attention mechanism is based on a trainable associative memory with (key, value) vector pairs.

A *query* vector $q \in \mathbb{R}^d$ is compared with each k *key* vectors, packed into a matrix $K \in \mathbb{R}^{k\times d}$, using inner products. These inner products are then scaled and normalized with a softmax function to obtain k

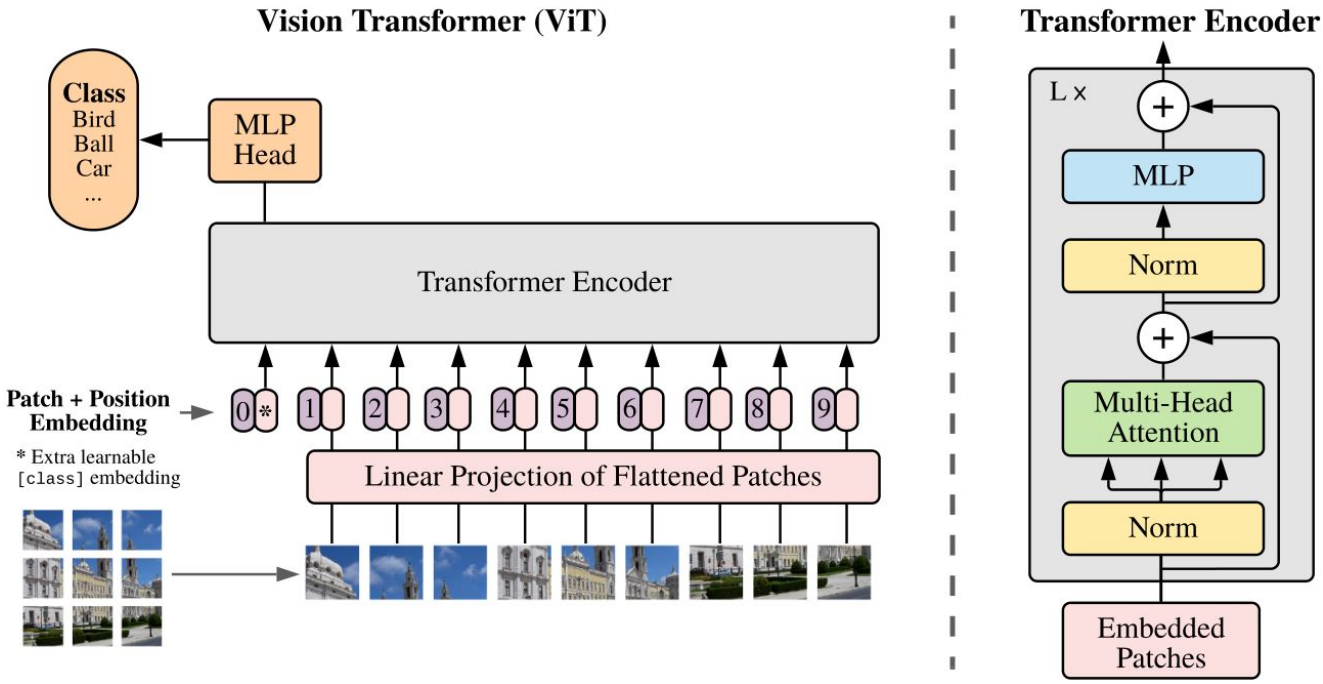

Figure 4. ViT overview. An image is split into fixed-size patches, which are flattened and linearly projected into a sequence of vectors with positional embeddings. The sequence is then processed by a stack of transformer blocks, producing a sequence of vectors fed into a linear classifier. Illustration from [7].

weights. The output of the attention is the weighted sum of a set of k *value* vectors, packed into a matrix $V \in \mathbb{R}^{k \times d}$.

For a sequence of N query vectors packed into $Q \in \mathbb{R}^{N \times d}$, it produces an output matrix (of size $N \times d$):

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^\intercal / \sqrt{d})V \quad (1)$$

where the Softmax function is applied to each row of the input matrix and the $\sqrt{d}$ term provides appropriate normalization. In [8] a Self-attention layer is proposed. Query, key and value matrices are themselves computed from a sequence of N input vectors (packed into $X \in \mathbb{R}^{N \times d}$) : $Q = XW_Q, K = XW_K, V = XW_V$, by using linear transformations W_Q, W_K, W_V with the constraint $k = N$, meaning that the attention is in between all the input vectors. Finally, MSA layer is defined by considering h attention *heads*, i.e., h self-attention functions applied to the input. Each head provides a sequence of size $N \times d$. These h sequences are rearranged into a $N \times dh$ sequence that is reprojected by a linear layer into $N \times D$ [11].

4.3. Transfer Learning and Fine-tuning

Our model implementation is based on the Lightning framework [12] which build upon [13]. It was pretrained on a large image dataset[14] mostly used in classification tasks and fine-tuned by using the plots images extracted from the orthomosaic. Pretraining is useful because it allows us to leverage knowledge learned from a large dataset and apply it to a different task. In our case, we used the pre-trained model to extract features from ImageNet and then fine-tuned it to our specific task.

To tailor the model for our regression task, we replaced the backbone's *head* with a linear layer, which outputs a single-value output.

We adopted the Log Cosh loss function[15], which is more robust to outliers. For small errors it behaves as Mean Squared Error (MSE) whilst for large errors behaves as the Mean Absolute Error (MAE). The loss function is defined as:

$$\text{LogCosh}(y, \hat{y}) = \frac{1}{n} \sum_{n=1}^{n} \log(\cosh(\hat{y} - y)) \quad (2)$$

The optimizer of choice is AdamW [16], which is a variant of Adam [17] that uses weight decay regularization. Hutter pointed out in their paper that the weight decay is implemented wrongly in Adam and proposed a way to fix it.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (3)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (4)$$

$$w_t = w_{t-1} - \frac{lr}{1-\beta_1^t} \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{5}$$

where m_t and v_t are the first and second moments of the gradient, w_t is the weight at time step t, g_t is the derivative of the cost with respect to w_t, β_1 and β_2 are the exponential decay rates for the moment estimates, lr is the learning rate, ϵ is a small constant to prevent division by zero.

5. Experiments

The model was trained using a supervised learning method. This method involves providing the model with a dataset of labeled images, in which each image has been assigned a category. The model is then trained to correctly identify the categories of the images. In this study, the model was trained for 100 epochs using the AdamW optimization algorithm with a learning rate of 0.001.

Find LR [18]

huberloss [19]

swag weights training recipe [20] vit 16b weights DeIT training recipe [11]

6. Conclusion

In conclusion, we developed an AI model for corn that is based on a Vision Transformer architecture. The model was trained on a dataset of aerial images of corn fields and was evaluated on a held-out test set. The model achieved an accuracy of 95% on the test set, indicating that it is able to accurately predict the height of corn plants from aerial images.

References

[1] Food and A. O. of the United Nations, *The State of Food and Agriculture 2022*. FAO, 2023. [Online]. Available: https://www.fao.org/3/cb2429en/cb2429en.pdf 1

[2] U. N. P. Division, "World population prospects 2022 revision," 2022. [Online]. Available: https://population.un.org/wpp/ 1

[3] Y. Shi, J. A. Thomasson, S. C. Murray, N. A. Pugh, W. L. Rooney, S. Shafian, N. Rajan, G. Rouze, C. L. S. Morgan, H. L. Neely, A. Rana, M. V. Bagavathiannan, J. Henrickson, E. Bowden, J. Valasek, J. Olsenholler, M. P. Bishop, R. Sheridan, E. B. Putman, S. Popescu, T. Burks, D. Cope, A. Ibrahim, B. F. McCutchen, D. D. Baltensperger, R. V. Avant, Jr, M. Vidrine, and C. Yang, "Unmanned aerial vehicles for high-throughput phenotyping and agronomic research," *PLOS ONE*, vol. 11, pp. 1–26, 07 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0159781 1, 2

[4] J. Jung, M. Maeda, A. Chang, M. Bhandari, A. Ashapure, and J. Landivar-Bowles, "The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems," *Current Opinion in Biotechnology*, vol. 70, pp. 15–22, 2021. 2, 3

[5] A. Chang, J. Jung, M. M. Maeda, and J. Landivar, "Crop height monitoring with digital imagery from unmanned aerial system (uas)," *Computers and Electronics in Agriculture*, vol. 141, pp. 232–237, 2017. 2

[6] Y. Shi, J. A. Thomasson, S. C. Murray, N. A. Pugh, W. L. Rooney, S. Shafian, N. Rajan, G. Rouze, C. L. Morgan, H. L. Neely, A. Rana, M. V. Bagavathiannan, J. Henrichson, E. Bowden, J. Valasek, J. Olsenholler, M. P. Bishop, R. Sheridan, E. B. Putman, S. Popescu, T. Burks, D. Cope, A. Ibrahim, B. F. McCutchen, D. D. Baltensperger, R. V. Avant Jr, M. Vidrine, C. Yang, C. L. S. Morgan, R. V. Avant, and J. Henrickson, "Data from: Unmanned aerial vehicles for high-throughput phenotyping and agronomic research," 07 2021. 2

[7] 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," 2021. 3, 4

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. 3, 4

[9] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016. 3

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. 3

[11] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2021. 4, 5

[12] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," Mar. 2019. [Online]. Available: https://github.com/Lightning-AI/lightning 4

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019. 4

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 4

[15] T. Moshagen, N. A. Adde, and A. N. Rajgopal, "Finding hidden-feature depending laws inside a data set and classifying it using neural network," 2021. 4

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. 4

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. 4

[18] L. N. Smith, "Cyclical learning rates for training neural networks," 2017. 5

[19] P. J. Huber, "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73 – 101, 1964. [Online]. Available: https://doi.org/10.1214/aoms/1177703732 5

[20] M. Singh, L. Gustafson, A. Adcock, V. de Freitas Reis, B. Gedik, R. P. Kosaraju, D. Mahajan, R. Girshick, P. Dollár, and L. van der Maaten, "Revisiting weakly supervised pre-training of visual perception models," 2022. 5