Proceedings of the 5th International Conference on Inventive Research in Computing Applications (ICIRCA 2023)
IEEE Xplore Part Number: CFP23N67-ART; ISBN: 979-8-3503-2142-5

RetinaNet and Vision Transformer-based Model for Wheat Head Detection

Dr. Gurram Sunitha*
Department of AI & ML, School of Computing, Mohan Babu University
Tirupati, India.
gurramsunitha@gmail.com

Adluru Sudeepthi
UG Scholar, Department of CSE, Sree Vidyanikethan Engineering College,
Tirupati, India.
adlurusudeepthi@gmail.com

Balija Sreedhar
UG Scholar, Department of CSE, Sree Vidyanikethan Engineering College,
Tirupati, India.
sreedharbalija12345@gmail.com

Abdul Bari Shaik
UG Scholar, Department of CSE, Sree Vidyanikethan Engineering College,
Tirupati, India.
abdul.sheikh456@gmail.com

C. Farooq
UG Scholar, Department of CSE, Sree Vidyanikethan Engineering College,
Tirupati, India.
cherukurufarooq@gmail.com

2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA) | 979-8-3503-2142-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIRCA57980.2023.10220614

***Abstract*—Vision transformers have achieved cutting-edge results on numerous object detection benchmarks, showcasing its potency as a robust object detection framework. Our aim is to design and develop a vision transformer based deep learning model towards the goal of smart agriculture. Wheat head detection is an important task in precision agriculture for estimating crop yields and monitoring plant health. This research study proposes a two-stage detector system by combining the RetinaNet object detection architecture with the vision transformer to improve the wheat head detection performance. RetinaNet is used as a proposal generator to generate candidate bounding boxes, and ViT is used as a backbone network for object classification/localization. The model was evaluated using Focal loss and Smooth L1 loss functions to jointly optimize classification and bounding box regression performance. The combination of RetinaNet and ViT took advantage of the strengths of both the approaches. RetinaNet generated candidate bounding boxes with high recall, while ViT processed these candidates efficiently and accurately, potentially reducing the number of false positives and improving overall detection accuracy. Experimental results on a wheat head detection dataset demonstrated that the proposed RetinaNet+ViT model is a promising and potentially efficient approach for wheat head detection, and has shown promising results in object detection task.**

Keywords— Object detection, Vision transformer, RetinaNet, Smart agriculture, Wheat farming.

I. INTRODUCTION

Traditional deep learning networks, have been widely successful in various computer vision tasks. [1-5]. However, traditional deep learning networks have considerable limitations. Traditional deep learning networks provide limited global context. They typically operate locally and lack a direct mechanism to capture long-range dependencies in the input data. They may be less effective in capturing global contextual information because they are built to extract features hierarchically based on local spatial patterns. Traditional deep learning networks require fixed-size inputs, and any variations in input size or aspect ratio often require preprocessing steps like resizing or cropping. This limitation can be problematic when dealing with images of different resolutions, scales, or aspect ratios.

The main purpose of conventional deep learning networks is to handle data, where spatial relationships are important. They are not efficient for the classification tasks on datasets where temporal dependencies play a significant role. Traditional deep learning networks often have a large number of parameters, making them computationally expensive and memory-intensive. Training and inference of these models can be resource-demanding, requiring powerful hardware and extensive computational resources.

The transformer architectures are a newly developed neural network that utilizes self-attention to critically acclaim most relevant global dependencies from the images. They are capable of processing variable-size inputs and can handle both spatial and sequential information effectively. Additionally, vision transformers have shown promising results with significantly fewer parameters, making them more memory-efficient compared to traditional deep learning networks.

Drawing inspiration from the significant success of transformer-based approaches, researchers recently are into transformer architectures to perform computer vision tasks. Transformer architectures are currently demonstrating its potential as a replacement for convolutional neural networks, despite the fact that CNNs are still thought of as the core component in vision applications. When Vision Transformer (ViT) uses a pure transformer to classify the entire image directly from sequences of image patches. Since it was first proposed, it has performed at the cutting edge on a number of image classification tasks [6, 7]. Transformers have been utilised to tackle a multitude of computer vision tasks other than classification.

Wheat is an essential crop for agriculture, as it contributes significantly to food security, livelihoods, and economic

growth worldwide. Computer vision has emerged as a critical technology in modern agriculture, including wheat farming. By leveraging machine learning and artificial intelligence algorithms, computer vision can monitor and analyze plant growth patterns, detect and diagnose diseases and pests, and optimize irrigation and fertilizer usage. By providing farmers with real-time information about their crops, computer vision can help them make more informed decisions, improve yield and quality, and reduce waste and costs.

To get detailed and accurate information about wheat fields around the world, plant scientists use image detection of "wheat heads"—spikes on the plant's top that hold grain. These photos are used to evaluate the size and form of various types of wheat heads. However, it could be visually difficult to distinguish wheat heads in pictures of spacious fields. Other object detection models exist, such as the neural network (RCNN) family, YOLO method (You Look Only Once), and SSD (Single Shot Detection), but using transformer-based models is a promising objective. In this paper, we propose a deep learning model that combines the RetinaNet object detection architecture with the Vision Transformer to improve wheat head detection performance. RetinaNet is used as a proposal generator to generate candidate bounding boxes, and ViT is used as a backbone network to extract features from the candidate boxes and classify them as wheat heads or background.

II. Related Works

Vision transformers are becoming a common option for image classification, object detection, and semantic segmentation tasks due to the emergence of large-scale pre-training datasets and the accessibility of sophisticated hardware [6, 7]. It is anticipated that vision transformers will continue to play an important role in the future of computer vision with continued research and development [8-11].

The creation of a new object detection model using Transformers architecture commonly used in Natural Language Processing solutions was revealed. In 2020, the Facebook research team published their research proposing an End-to-End Item Object Detection with Transformers [12]. Researchers presented the Detection Transformer (DETR), a significant new method for panoptic segmentation and object identification. When compared to earlier object detection systems, DETR radically alters the design. Transformers have been effectively incorporated as a key building piece in the detection pipeline for the first time by this object detection framework.

A wheat head detection system was proposed that uses Faster R-CNN, a popular object detection framework, to detect wheat heads in field images [9]. An automated system that can accurately detect wheat heads in images was designed. The authors used the Faster R-CNN framework to integrate the feature extractor and RPN into a single end-to-end system.

A new approach is proposed for detecting actions in videos using transformers [13]. They introduced a temporal transformer architecture that processes a stream of frames into a stream of feature vectors that represent temporal information.

A deformable transformer has been introduced that improves the detection accuracy on small objects and highly occluded objects [14]. The notion of self-supervised learning, in which a model learns to detect objects without the requirement for explicit supervision, was established in response to the constraints of supervised learning in object detection. The few-shot object detection is the technique where a model is trained to detect objects with very limited annotated data [15].

We acknowledge that the Vision Transformer is a potent instrument for detecting objects [16-18]. With this initiative, we propose a two-stage detector based on RetinaNet and vision transformer approaches to perform object detection in wheat farm images. We have specifically designed the model for wheat head detection. Our model works efficiently towards wheat head detection in agricultural images, where the wheat heads can be present in varying sizes, orientations, and locations in the image.

III. RetinaNet+ViT Model for Wheat Head Detection

Our proposed deep model is based on vision transformer approach for object detection. We have customized and designed a model for building an efficient learning model for wheat head detection. ViT has shown great potential for improving the performance of object detection tasks, including wheat head detection. The attention mechanism in ViT allows it to handle object occlusion and cluttered backgrounds effectively, which is a common issue in wheat head detection, where the wheat heads are often close to each other and can be partially blocked by debris like leaves. Overall, the promising results of ViT in multitude of computer vision tasks and its ability to capture global contextual information make it a potential game changer for wheat head detection in agricultural images.

Figures 1 and 2 shows the proposed learning pipeline and the proposed RetinaNet+ViT model respectively for efficient wheat head detection. RetinaNet is a popular object detection algorithm that can be used to detect objects in images, including wheat heads in agricultural images [19]. We have designed a two-stage detector model with the RetinaNet acting as the proposal generator and ViT being the backbone network typically followed by a set of detection heads. The object class and bounding box coordinates are predicted by the detection heads using the extracted features. RetinaNet followed by ViT are used as a part of a pipeline for object detection. In this pipeline, to generate a set of candidate object proposals, RetinaNet is used as the proposal generator. These proposals would be passed through the ViT backbone network to extract features, which would be used to classify the objects and refine their bounding boxes.

The feature maps are extracted by the RetinaNet from the wheat head farm images. Information about the two-dimensional structure is retained in this way. The data is then flattened in the subsequent stages, giving a one-dimensional structure. It is then passed to the encoder-decoder mechanism after positional encoding. Finally, the Feed Forward Network receives each output. In the Positional Encoding section, the vector of each image patch is recreated according to its place in the array. Thus, the same image patch can have different vectors at different positions in the array. The output embedding is created from N object queries. The feed forward network is used to perform final estimate processes in the next phase.

1. Generate object proposals using RetinaNet: Pass the input image through the RetinaNet architecture to generate a set of candidate object proposals. Each candidate proposal would consist of a candidate bounding box and an objectness score.
2. Extract features using ViT: ViT being the backbone network would process each candidate box to extract features from the candidate boxes and classify them as wheat heads or background to produce final classification and bounding box regression output. These outputs would be used as input to the subsequent classification heads.
3. Apply classification heads: Apply separate classification and regression heads to the cropped features for each proposal to predict the class probabilities and refine the bounding box coordinates.
4. Non-maximum suppression: Apply NMS to the proposals to remove duplicates and keep the proposals with the highest objectness scores.
5. Output: Return the final set of object detections, each represented by a bounding box and associated class label.

Fig. 1. The Proposed RetinaNet + ViT Pipeline for Wheat Seed Detection

For the RetinaNet+ViT model proposed for wheat head detection, the Smooth L1 loss and Focal loss are used for calculating the bounding box and classification loss respectively. These functions were chosen as they are predominantly used with RetinaNet architectures.

Smooth L1 loss is a common choice for calculating the bounding box loss in object detection tasks, including those involving wheat head detection. The Smooth L1 loss function provides a smooth gradient near zero and reduces the sensitivity to outliers, making it suitable for bounding box regression. It strikes a balance between the L1 loss and the L2 loss, allowing for robust and stable training of the object detection model.

Focal loss is a popular choice for calculating the classification loss in object detection tasks, including wheat head detection. Focal loss addresses the issue of class imbalance commonly found in these tasks by assigning higher weights to hard, misclassified examples and lower weights to easy, well-classified examples. This helps in focusing the model's attention on challenging instances and improving its ability to distinguish between different classes. Focal loss effectively tackles the problem of foreground-background class imbalance and has shown promising results in boosting the performance of object detection models, particularly when dealing with datasets with imbalanced class distributions.

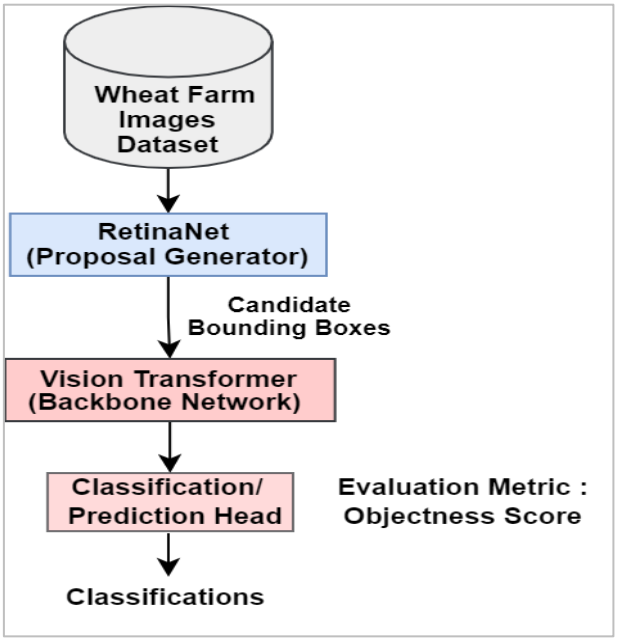

Fig. 2. Proposed RetinaNet + ViT Model for Wheat Seed Detection

The bounding box loss is then combined with the classification loss to obtain the final detection loss to obtain the objectness score. Let there be *K* object categories, and for each category *k*, we have M_k proposed regions with their corresponding proposal scores $P_{k,i}$ and classification scores $C_{k,i}$ where *i* represents index of the i^{th} proposed region for an object category *k*. Let $O_{i,j}$ represent the j^{th} highest objectness score among all the proposed regions of an object category *k*, Then, the objectness score $O_{i,j}$ can be calculated as shown in Equation (1).

$$O_{i,j} = P_{i,j} \times C_{i,j} \quad \ldots\ldots\ldots \quad (1)$$

Objectness scores are used to select the regions with the highest scores as the final detections for each object category. The usage of transformer topologies offers a significant benefit in some types of object detection issues not simply in terms of performance. The use of transformer architectures provides a great advantage not only in terms of speed but also in some specific type of problems for object detection problems. The prediction is made according to the image content of the object detection algorithm via this architecture. Thus, higher success is achieved with this approach where the context is important in the images. When recurrent neural networks are used for object detection projects, it has been seen that the accuracy is lower and the model runs slower. Because the operations are sequential. Since these operations are carried out in parallel from transformer architectures, we get a much faster model.

IV. Experimentation Results and Discussions

The design of the proposed model is presented in Table 1. We have experimented our model on the wheat dataset, which comprises of the images of wheat farms [20]. Images having wheat heads are pre-annotated with bounding boxes surrounding each wheat head. Figure 3 shows sample training images of wheat farm with wheat heads and without wheat heads, with bounding boxes, without bounding boxes. Our proposed model is expected to predict the presence of bounding boxes accurately if present in the image. The size of the input wheat images dataset of model training is referred as the dataset size. The sizes that we used are 1K, 2K and 3.432K as the original wheat image dataset size is 3.432K. We have experimented with different batch sizes, including 32 and 64, that were employed in the studies. The training of the model is repeated for 20 and 50 epochs.

Table 1. The Design of the RetinaNet+ViT Model

Model Type	Proposal Generator	Backbone Network
Two-Stage Object Detector	RetinaNet	ViT-L/16

The experimental evaluation of the proposed RetinaNet+ViT model is presented in Table 2. In Table 3, we have tabulated the comparative evaluation of the proposed RetinaNet + ViT model with other deep learning models. The mAP@0.5 (mean average precision at intersection over union threshold of 0.5) for different models have been tabulated, interpreted and conclusions were drawn. From the results, it is observed that the RetinaNet+ViT model with ViT as the backbone network achieved the highest mAP of 0.9186, followed by EfficientNet-B4 with a mAP of 0.8710. ResNet-50 and ResNet-101 have significantly lower mAP scores of 0.8136 and 0.716, respectively. This indicates that the ViT backbone network is better suited for the wheat head detection task than the ResNet backbone networks. It is observed that using the combination of RetinaNet and ViT as a two-stage detector performs better than using either RetinaNet or ViT alone. Overall, this suggests that the RetinaNet+ViT model is an effective and efficient approach for wheat head detection. The model's performance is particularly impressive given the challenging nature of the wheat head detection task, with small and closely spaced wheat heads present in the images.

(a) Wheat Farm Image with Wheat Heads and with Bounding Boxes

(b) Wheat Farm Image with Wheat Heads and without Bounding Boxes

(c) Wheat Farm Image without Wheat Heads

Fig. 3. Sample Training Images of Wheat Dataset

Table 2. Performance of the Proposed RetinaNet+ViT Model

Dataset Size	Batch Size	Epochs	Accuracy (%)	F1-Score (%)
1K	32	20	90.35	91.55
		50	91.23	93.18
	64	20	91.50	90.87
		50	91.86	90.46
2K	32	20	92.33	91.64
		50	91.66	91.56
	64	20	91.23	90.59
		50	89.64	91.82
3.432K	32	20	91.44	91.63
		50	90.87	90.70
	64	20	89.85	90.06
		50	90.02	89.34

Table 3. Comparative Evaluation of the proposed Model

Model	mAP@0.5
ResNet-50	0.8429
ResNet-101	0.7160
EfficientNet-B4	0.8710
RetinaNet	0.8767
ViT	0.8543
RetinaNet + ViT	0.9186

RetinaNet performed well detecting small objects such as wheat heads. Its focal loss and anchor-free design are particularly well-suited for detecting objects of varying sizes. The combination of RetinaNet and ViT took advantage of the strengths of both approaches. RetinaNet generated candidate bounding boxes with high recall, while ViT processed these candidates efficiently and accurately, potentially reducing the number of false positives and improving overall detection accuracy. The model could be trained end-to-end using a single loss function that combined both the classification and regression losses, which simplified the training process and lead to better overall performance. Overall, the RetinaNet+ViT model appears to perform well on the given task, achieving accuracy scores above 90% in most experiments which ranges from 89.34% to 91.86%. The model can also be fine-tuned using transfer learning, leveraging pre-trained weights for both RetinaNet and ViT on large-scale datasets. This can help reduce the amount of labeled data needed for training and improve the model's performance on smaller datasets.

V. CONCLUSION

This study has proposed a novel two-stage detector for wheat head detection that combines the RetinaNet object detection architecture with the Vision Transformer. Our model leverages the strengths of both approaches, using RetinaNet as a proposal generator and ViT as a backbone network to extract and classify features. The RetinaNet architecture is used to generate object proposals from the wheat farm images. The extracted object proposals were fed to the transformer layer as the encoder-decoder mechanism followed with the computation of the loss value using the Focal loss and Smooth L1 loss functions for calculating the overall detection loss of the model. The combination of RetinaNet and ViT took advantage of the strengths of both approaches. RetinaNet generated candidate bounding boxes with high recall, while ViT processed these candidates efficiently and accurately, potentially reducing the number of false positives and improving overall detection accuracy, suggesting that it has great potential for practical applications in precision agriculture. In conclusion, the results suggest that the RetinaNet+ViT model with ViT as the backbone network is a highly effective two-stage detector for wheat head

detection. The results also highlight the importance of choosing an appropriate backbone network for the task at hand, with the ViT network outperforming the ResNet networks in this case. However, it is worth noting that vision transformers may require larger amounts of training data and longer training times to achieve optimal performance compared to CNNs, especially in scenarios with limited annotated data. Furthermore, vision transformers may face challenges when applied to tasks that require precise localization or handling fine-grained details due to their inherent downsampling nature. Further research is imminent and warranted to address and overcome the limitations of vision transformers, enabling their broader applicability and improved performance across various computer vision tasks.

REFERENCES

[1] Rajasekhar Nennuri, et. al., "A Multi-Stage Deep Model for Crop Variety and Disease Prediction," 14th International Conference on Soft Computing and Pattern Recognition, vol. 48, pp. 52-59, Springer, 2023.

[2] N. S. Charan, et. al., "Solid Waste Management using Deep Learning," 14th International Conference on Soft Computing and Pattern, vol. 648 pp. 44-51, Springer, 2023.

[3] M. Shereesha, et. al., "Precision Mango Farming: Using Compact Convolutional Transformer for Disease Detection," 13th International Conference on Innovations in Bio-Inspired Computing and Applications, Springer, vol. 649, pp. 458-465, 2023.

[4] N. Balakrishna, et. al., "Tomato Leaf Disease Detection Using Deep Learning: A CNN Approach," International Conference on Data Science, Agents & Artificial Intelligence, IEEE, 2022.

[5] D. Sudarsana Murthy, et. al., "An Investigative Study of Shallow, Deep and Dense Learning Models for Breast Cancer Detection based on Microcalcifications," In 2022 International Conference on Data Science, Agents & Artificial Intelligence, vol. 1, pp. 1-6, IEEE, 2022.

[6] A. Dosovitskiy, et. al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, arXiv preprint arXiv:2010.11929.

[7] K. Han, et. al., "A survey on vision transformer," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 1, pp. 87-110, 2022.

[8] X. Li, W. Fan, Y. Wang, L. Zhang, Z. Liu, C. Xia, "Detecting Plant Leaves Based on Vision Transformer Enhanced YOLOv5," In 2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML), pp. 32-37, IEEE, 2022.

[9] S. Khaki, N. Safaei, H. Pham, L. Wang, "Wheatnet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting," Neurocomputing, vol. 489, pp. 78-89, 2022.

[10] S. Wu, Y. Sun, H. Huang, "Multi-granularity feature extraction based on vision transformer for tomato leaf disease recognition," In 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), pp. 387-390, IEEE, 2021.

[11] R. Castro, I. Pineda, W. Lim, M.E. Morocho-Cayamcela, "Deep learning approaches based on transformer architectures for image captioning tasks," IEEE Access, vol. 10, pp. 33679-33694, 2022.

[12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers,". In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pp. 213-229, Springer International Publishing, 2020.

[13] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, X. Bai, End-to-end temporal action detection with transformer," IEEE Transactions on Image Processing, vol. 31, pp. 5427-5441, 2022.

[14] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," 2020, arXiv preprint arXiv:2010.04159.

[15] G. Huang, I. Laradji, D. Vazquez, S. Lacoste-Julien, P. Rodriguez, "A survey of self-supervised and few-shot object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[16] T. Ma, M. Mao, H. Zheng, P. Gao, X. Wang, S. Han, E. Ding, B. Zhang, D. Doermann, "Oriented object detection with transformer," 2021, arXiv preprint arXiv:2106.03146.

[17] A. Gupta, S. Narayan, K.J. Joseph, S. Khan, F.S. Khan, M. Shah, "Ow-detr: Open-world detection transformer," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 9235-9244, 2022.

[18] Z. Dai, B. Cai, Y. Lin, J. Chen, "Unsupervised Pre-Training for Detection Transformers," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[19] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection," In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988, 2017.

[20] E. David, S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng, S. Liu, N. Kirchgessner, G. Ishikawa, K. Nagasawa, M.A. Badhon, C. Pozniak, "Global Wheat Head Detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods," Plant Phenomics, 2020.