Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5641 - 5654
Received 15 May 2021; Accepted 20 May 2021.

Wheat Heads Detection using Deep Learning Algorithms

Movva Nitin Datta[1*], Yash Rathi[2], M. Eliazer[3]
[1,2,3]SRM Institute of Science and Technology, Kattankulathur
*mm1178@srmist.edu.in

ABSTRACT

This paper aims to analyze the use of object detection in the field of wheat phenotyping and the components used inside it. Recently computer vision has started to play a pivotal role in agriculture, so the special emphasis has been given to the state-of-the-art object detection models to compare them against one another. The comparison of each algorithm is performed based on performance on a dataset known as GWHD and its Box AP and APS values on the benchmark dataset. The most efficient networks for object detection, comprising of a single-stage detector known as YOLO and a two-stage detector known as Faster R-CNN, have been studied, in general, and compared on many fronts to get an overall and comprehensive comparison.

Keywords:
Faster R-CNN, YOLO v4, YOLO v5, GWHD, wheat awns.

Introduction

The field of agriculture [1] has been gaining attention lately as it can be useful in genome mapping [2], another upcoming field where Artificial Intelligence [3] is playing a very crucial role. Plant phenotyping, although a subfield of agriculture, itself is an umbrella under which many sub-fields exist, like Postharvest [4], Development [5], Physiology [6], and Morphology [7]. All of the before mentioned fields cover all the domains in plant phenotyping, ranging from its health detection to counting organs to fruits. With the recent advancement in computer vision and neural networks, more improved solutions have been possible. Since it is a field that requires extensive research and is upcoming hence not much has been established and is still under research. New benchmarks are being set daily with the coming of new advancements, and this paper aims to use many such advanced techniques that were released recently to create robust models that can detect wheat heads accurately.

One of the main reasons to detect wheat-heads is that it is one of the important crops that plays a pivotal role in feeding humans across the globe. Globally, wheat production was around 758.3 million tonnes in 2020.

Literature Review

Since the paper is a combination of object detection and field of agriculture, many papers have been published by many renounced researchers.

Yu Jiang et al. [8] in which there is abundant information that tells how to use CNN’s for understanding stress evaluations, plant development, and post-harvest quality assessment. There are different architectures presented that explain image segmentation, object detection. They even provide some SOTA solutions for certain phenotyping applications.

Wu Wei et al. [9] proposed detection and enumeration of wheat grains based on a deep learning method under various scenarios and scales which gives us information about wheat grains. It states that the number of grains plays a pivotal role in determining yield. The authors use Faster

R- CNN to detect wheat grains with loss less than 0.5 and mAP:0.9. The detection time is quick, under 2 seconds. It is also robust to different backgrounds and different levels of grain crowding.

Zhong-Qiu Zhao et al. [10] wrote about object detection with deep learning and it is discussed about the most popular object detection architectures along with their working. This paper also explains about application domains of object detection, two of which are face detection, pedestrian detection. Some algorithms that are discussed are R-CNN, Fast R-CNN, Faster-RCNN along with different feature extractors such as FPN, Single Feature Map etc.

Ajeet Ram Pathak et al. [11] in the paper proposed an explanation about deep learning techniques that are used for object detection. Some of the SOTA algorithms are discussed in this paper. It also discusses some of the benchmark datasets used for object detection. The paper even explains different types of object detection methods: Deep Saliency Network, Adversarial Learning, Fine-grained object detection.

Cong Tang et al. [12] analyses object detection based on deep learning in the paper. It explains in-depth about real-time object detection. It also challenges present in the current circumstances and proposes solutions on how to improve techniques.

Steven C.H. Hoi et al.[13] elaborates about recent advances in deep learning for object detection about a general introduction given to the object detection. Then the authors dwell into 3 major parts which are detection components, learning strategies, applications with benchmarks. Some of the components are feature learning, proposal generation, sampling strategies etc. At the end, the authors provide future directions on what more can be improved.

Roman Solovyev et al. [14] proposed weighted box fusion for object detection models. The paper introduces a novel technique discussed known as weighted box fusion. Weighted box fusion has a better performance than NMS (non-max suppression) and NMS-soft. It works better when using ensemble methods.

Barret Zoph et al. [15] discusses about data augmentation. Augmentations play a crucial in object detection. There were many experiments with and without augmentations to show accuracies of different detectors. It even explains why the model regularizes and ends with explanations of many different augmentations.

Guotai Wang et al. [16]explains why using augmentations at test time helps in improving the robustness and accuracies of the models. There was an experiment conducted with MRI scans of fetal brains and brain tumors which provided better model-based uncertainty.

Yukang Chen et al. [17] explains mosaic data augmentation. It is explained that mosaic augmentation works very well with small objects. It also helps in exploiting the loss in statistics and helps in scale balancing large and small objects.

Alexey Bochkovsky et al. [18] dwells deep on YOLOv4. The authors discuss how it is improved over YOLO v3 significantly over speed, accuracy and newer technology used. Many architectures were used in this paper which had different combinations of backbone, neck, and detector. Few of the backbones that were considered in this paper were CSPResNeXt50, CSPDarknet53, EfficientNet B3. Upon experimentation of all these backbones, CSPDarknet53 showed the best results. Upon this the SPP block (Spatial Pyramid Pooling) was used to separate out the most significant context features while maintaining the same inference speed. Instead of FPN, which is extremely popular the authors prefer PANet (Path Aggregation Network) [19] and the head used was YOLO v3.

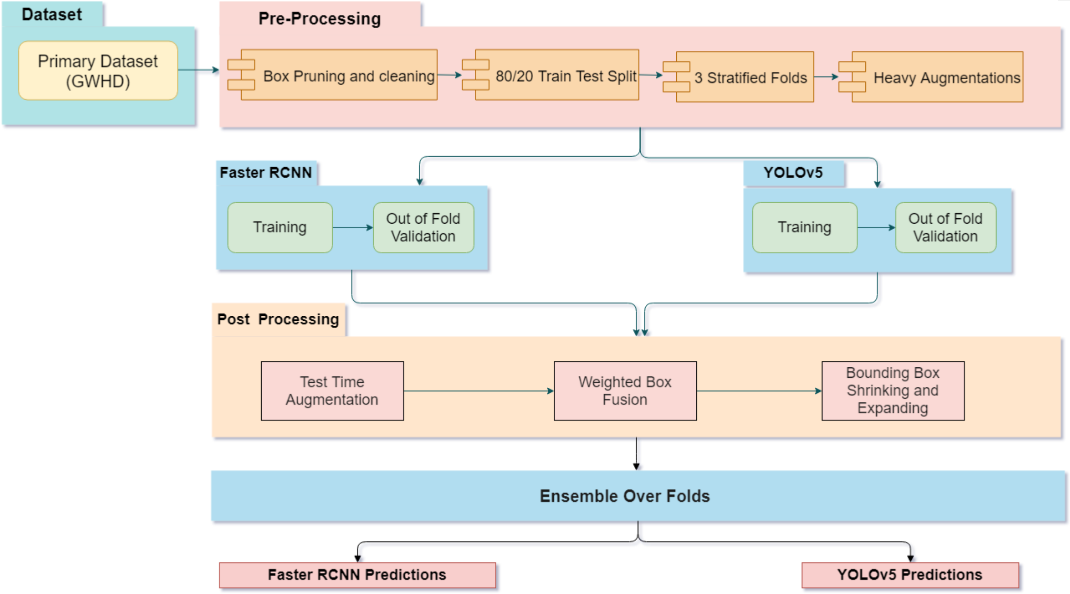


Figure 1. Architectural diagram of the proposed work

Dataset

Wheat heads are the main source of information for characteristics such as the presence of awns, health, maturity stage, and size. The heads are the main source of information for almost all the plants in nature. Over the years many models and methods have been developed for the detection and study of the high-resolution RGB wheat head images, but the main hindrance in further development has been the scarce and missing presence of a diverse dataset, as shown in Figure 2, that tests the methods on various fronts. High variability in genotypic differences, observational conditions, development stages, and head orientation are challenges that increase the complexity of the problem. Overlapping of the wheat heads in a concentrated environment and motion blur aggravate the problems.

GWHD[20] dataset helps to overcome such problems by providing diversity as well as quantity of well-labeled images. The collection of 190,000 labelled wheat heads in 4,700 high-resolution RGB images is assembled from several countries spanning across different continents around the world to encompass a wide range of genotypes.

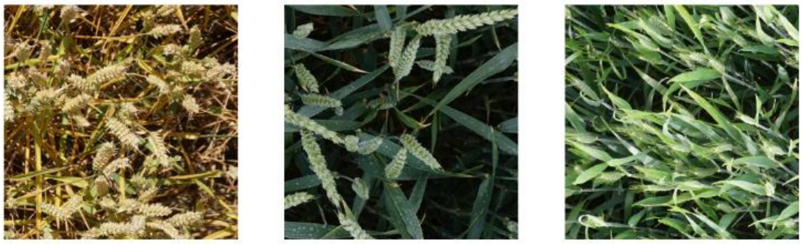


Figure 2. Diversified dataset

Proposed Methods

Pre-Processing phase

The first phase is the pre-processing of the images and the data are given from the dataset, as shown in Figure 1 contains four main methods or components. Since it is important that the images are of good resolution and there is no inherent skewness in the dataset hence detailed and rigorous processes are undertaken to make sure the images of utmost quality are used. Apart from images we even need to focus on the bounding boxes given since the data is manually labelled and prone to errors. We will use few techniques, that are listed below, to eliminate those outliers.

Box Pruning and Cleaning

This step is done to remove images that are very blurry or have a very a smaller number of bounding boxes, as very little information can be extracted from them. From the selected images the x and y coordinates along with width (w) and height(h) of the bounding boxes are stored. Then x-center and y-center coordinates are calculated, which are the center coordinates of the box on the image, and stored which are used by the YOLO model for training. Bounding boxes whose area is between 40 and 130000 pixels are only taken into consideration because boxes outside these are either too small to interpret for the result or contain too many extra and unnecessary parts of the image except the awns. We even plot bounding boxes over images and manually try to find anomalies that have escaped the area filters. Five such bounding boxes were anomalous and were removed.

Train-Test split

As we have 3373 images that make up the GWHD dataset we split the data in such a way so that the test set contains 400 images, and the rest of the images go into the training set. The training set is further divided into 3 stratified folds which contain 991 images each.

Stratified folds

As the dataset inherently contains data from different source countries, using stratification helps in the splitting of data in such a way so that all folds are equally balanced with diversified data. Hence consistency wise all the folds are very similar in nature and any inherent skewness in the dataset is also taken care of by making stratified folds based on the country (source) from which the image is taken. The dataset can be broken down into 3 folds containing 991 images each to accommodate the diversity of the dataset. During training, we will use the Out of Fold validation technique to have 3 different models.

Augmentation

To achieve higher accuracy and better detection, using images with different contrast levels, different brightness levels, rotations, flipping, etc. is done. These are the basic types of augmentation techniques and more advanced and effective techniques like CutOut[21], MixUp[22], Mosaic[23] are used to further improve the robustness of the model. Each augmentation is used with a certain probability so that all the images are different, and dataset is diverse.

Basic Techniques

These are simple techniques in which the image is rotated across planes, color and brightness are changed so as to consider corner cases in which the images might not be that clear and hence makes our models even more robust to real-time scenarios.

Random Size Cropping

This augmentation essentially is related to cropping. The algorithm randomly selects one part of the image depending on the size given as input and crops the rest of the image. This is useful as it crops the images, and a very concentrated part of the image is taken into consideration.This augmentation essentially is related to cropping. The algorithm randomly selects one part of the image depending on the size given as input and crops the rest of the image. This is useful as it crops the images and a very concentrated part of the image is taken into consideration.

Hue Saturation

This augmentation technique randomly changes the hue, saturation, and value of the input image. Ranges, that are between 0 and 255, are given as input for all the 3 categories and it randomly changes within the limit.

Random Brightness Contrast

Brightness and contrast of the images are taken care of in this technique. Like the previous algorithm, brightness and contrast can be randomly changed within the specified limits.

Converting images to Gray

The input image is in the RGB format and hence contains a lot of noise throughout the 3 channels. Making it monochrome helps in reducing the unwanted noise by reducing the channel count by 2 to 1. This is a very common technique as it is one of the most effective ways to reduce noise without losing the characteristics and the features of the image.

Horizontal Flip

This is a simple technique of increasing the size of the dataset by flipping the image horizontally by 180° to generate a mirror image of the input image in which the y-axis acts as a mirror.

Vertical Flip

This technique is a variation of the flipping algorithm in which the image is rotated vertically to generate a mirror image in which the x-axis acts like a mirror.

Advanced Techniques

These are newer and more effective sort of augmentation techniques in which different images are mixed or a part of the image is cut out. All of these makes the models even more robust as it makes the model learn to detect by only looking at a part of the whole and not the entire image.

CutOut

This augmentation technique is used to make the model robust to the problem of object occlusion. Often the object that is to be detected is covered by some other object hence not revealing the entire object and making it difficult for the model to detect it. In the technique, it randomly masks out square regions of input during training. This helps the model in regularizing and helps in improved learning of the objects.

Mixup:

It is a data augmentation technique that generates weighted combinations of random image pairs from the training data. Given two images and their ground truth labels: (xi, yi),(xj, yj), a synthetic training example $(\hat{x},\hat{y})$ is generated as:

$$\hat{x}=\lambda\ x_i +(1-\lambda)\ x_j$$
$$\hat{y}=\lambda\ y_i +(1-\lambda)\ y_j$$

Received 15 May 2021; Accepted 20 May 2021.

where λ~Beta(α=0.2) is independently sampled for each augmented example. Advantages of using Mixup is that it is extremely good at regularizing models and is really fast as well.

Mosaic

This technique essentially combines 4 images into 1 based on predefined ratios. Hence it allows the models to learn to detect and identify objects at a smaller scale than usual. Figure 3 gives an illustration of this type of augmentation.

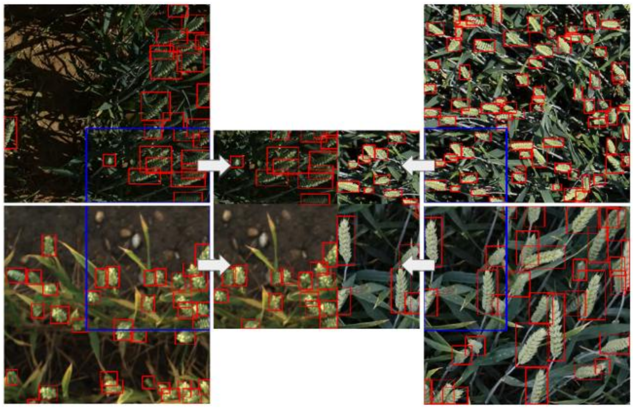


Figure 4. Shows Mosaic augmentation where 4 image corners are taken and merged to form a new image.

Training phase

Faster RCNN

The Faster R-CNN model, which is a 2-stage detector, is trained for 80 epochs for each fold using a custom LR scheduler which is built upon the LR scheduler known as Cosine Annealing Warm Restarts. Upon a lot of experimentation, it was found that this specific LR scheduler along with some tweaks produces accurate results with a lower number of epochs. After comparison of results using different backbones such as ResNet-50, ResNet-101[24], ResNeSt-101, ResNeXt-101, it was found that ResNeSt-101 had the highest accuracy and took the least amount of time to train. As claimed by the authors ResNeSt outperforms Efficientnet[25] in terms of accuracy in image classification problems. It has also achieved good results on the benchmark datasets while serving as a backbone. The below formula shows RPN loss function.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*)\,\lambda\,\frac{1}{N_{reg}}\sum_i p_i^*\,L_{reg}(t_i,t_i^*)\ (3)$$

Moving on to some important parameters, the base learning rate used was 0.005, the optimizer used was SGD (Stochastic Gradient Descent) with a momentum of 0.9 and weight decay of 0.0005. During the training stage bounding box regression loss (add formula), RPN box regression loss, best threshold, etc were tracked. These metrics provided a clear idea of when to stop the training. The training was done on 3 folds and all the 3 models were tested on the 400-image test dataset. The accuracy of each model was around 0.69. On using Weighted Box Fusion for ensemble by giving higher weightage to the fold with maximum accuracy best score achieved on the ensemble was 0.71.

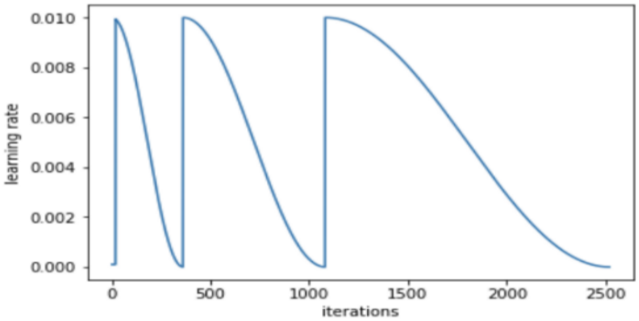


Figure 4. Cosine Annealing learning rate scheduler. Further minute modifications have been adopted for each algorithm.

YOLO

Although there is no official paper on YOLO v5 it has few improvements over the YOLO v4 version. The YOLO v4 algorithm had many new components in it which were described as Bag of Freebies and Bag of Specials by its authors. For the backbone, that is CSPDarknet-53[26], data augmentation was applied along with some advanced augmentations such as Mosaic and CutMix. In addition to this dropblock regularization[27] and class label smoothing was used in YOLO v4 which were not present in the previous versions. Instead of using FPN (Feature Pyramid Network), which is used in many of the State-of-the-Art algorithms such as YOLO v3 and R-CNN variants, the authors used PANet (Path Aggregation Network)because it can preserve spatial information accurately along with few more advantages. Figure 5 illustrates PANet architecture that is further subdivided into 4 different components.

Few more components that were new in the YOLO v4 detector were CIOU loss (Complete Intersection over Union)[28], CmBN (Cross Mini-Batch Normalization)[29], self-adversarial training[30], cosine annealing scheduler as shown in Figure 4, etc. Apart from these the authors also proposed Bag of Specials which contains Mish activations, CSP (Cross Stage Partial)[31] connections, etc.

However, the YOLO v4's performance on custom datasets is not optimal. This might be because YOLO v4 uses anchors that are based on the Microsoft COCO dataset[32]. The YOLO v5 algorithm tries to solve this issue by using a genetic algorithm that generates new anchors. In case if the pre-loaded anchors do not fit well with the custom data and fall below a certain matching threshold, then YOLO v5 automatically starts training new anchors which fit well with the custom data at hand.

As the GWHD dataset has bounding boxes that are small and crowded, YOLO v5 might have performed better by generating new anchors. Similar to Faster R-CNN the same folds using the same split were used to train 3 different models of YOLO v5. To improve the speed and to increase the batch size Nvidia Apex was used which incorporates mixed precision and 16-bit training. Some of the important hyperparameters were learning rate, which was 0.001, optimizer used was SGD (Stochastic Gradient Descent) along with a momentum of 0.937 and the weight decay was 0.0005. During the training stage, some important metrics were tracked which were GPU memory usage, Precision, Recall, mAP@.5, etc. The accuracy of each of the three models on the 400-image test dataset was around 0.70. Similar to Faster R-CNN the Fold-2 model produces lesser accuracy and the ensemble of all 3 folds produced the best accuracy of 0.719. These results were achieved by using TTA (Test Time Augmentation)[33] and hyperparameter tuning.

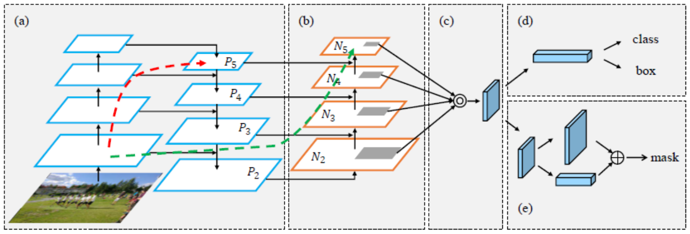


Figure 5.PANet architecture over FPN is used in YOLO v4.

Post Processing Techniques

Post Processing Techniques play a very important role in analysing the formed bounding boxes and how can these be made more accurate. Techniques used are both manual and automatic in nature to achieve the highest possible degree of precision.

Test Time Augmentation

The purpose of Test Time Augmentation (TTA) is to perform random augmentations on the test images or performing pre-defined set of augmentations. Thus, instead of showing the regular, "clean" images, only once to the trained model, it will show the augmented images several times. This will help the model to create several sets of bounding boxes for one single image. All these sets of bounding boxes will be passed on to a technique called weighted box fusion, which ensembles all the predictions to give a final prediction.

Weighted Box Fusion

After training is done, multiple boxes are generated. All the generated boxes have a confidence value attached to them. In Weighted Box Fusion, all the predicted boxes are averaged using specific weights or ensembled to generate one single prediction box which is the most suitable and has the highest IoU (Intersection over Union) for the object. The threshold is set to consider only boxes that are predicted highly accurately by the model. It is an advanced method over the Non-Maximum Suppression (NMS)[34] method.

Bounding Box Shrinking and Expanding

This is a manual post-processing technique in which the predicted boxes are taken into consideration and then manually the boxes are either shrunk or expanded depending on what is the necessity in the majority of the cases. This method allows the user to manually assess results as well as help in increasing the accuracy of the model minutely.

Results

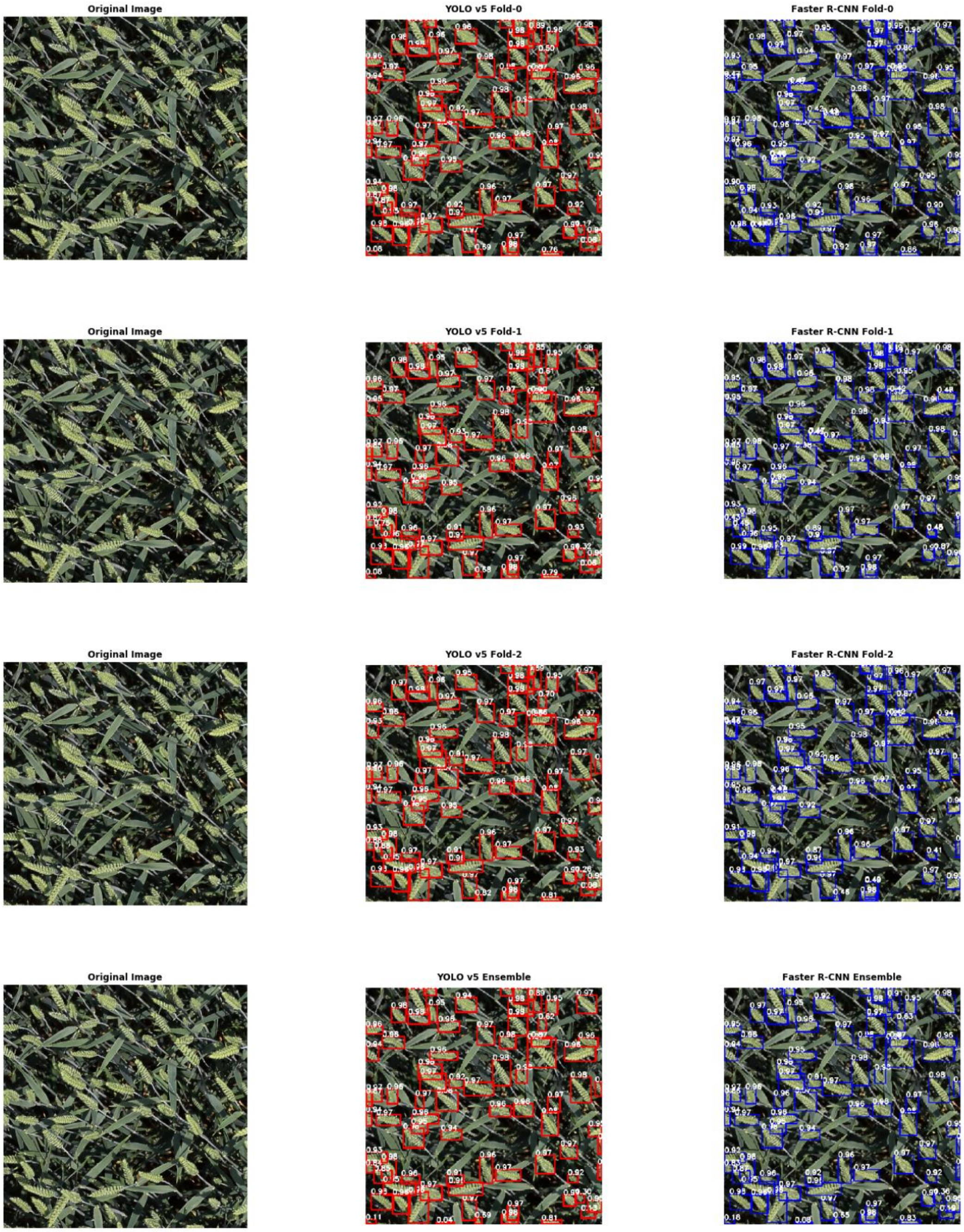
After thorough hyperparameter tuning on both the algorithms which are Faster R-CNN and YOLO v5 respectively, it was found that YOLO v5 has higher accuracy. One of the reasons why YOLO outperforms Faster R-CNNmight be because the GWHD dataset was labelled using a variant of YOLO. The results are evident for each fold from the Figure 6. In addition, from the Table 1 it is evident that Fold-2 had least accuracy for both the algorithms and an ensemble over 3 folds produced the best results in each of the algorithm. The ensemble which is performed by Weighted Box Fusion [14] gives higher weightage to Fold 0 and Fold 1 as these both single models have better performance than Fold 2 single model.

Table 1. Accuracy of different models on different folds.

Algorithm	Fold-0	Fold-1	Fold-2	Ensemble
Faster R-CNN	0.6986	0.698	0.6931	0.7085
YOLO v5	0.7106	0.7098	0.7057	0.7192

Conclusion and Future Work

Further modification of these models can be done to achieve higher accuracy. Apart from that, one can use these models to detect the amount of yield from a given image. One can even try to detect the continent from which the images come from. There are around 15 stages during the wheat growth process by compiling a dataset for each stage one can integrate it into this project to detect its stages which can help the farmers immensely by providing them in-depth information about how to grow the wheat efficiently at each stage. By using a drone, the farmers can even lookout for pests and diseases if any on the wheat heads. This can be relatively simple because the model will not detect heads that are infested with pests as it is not trained on that kind of data. So the number of heads detected will fall significantly indicating a threat to heads using this the farmers can delve deeper to find the root cause of the wheat heads.



Formatted: Font: 12 pt

Figure 6. Output of images for different folds.

References (APA 6th edition)

[1] Li, Lei, Qin Zhang, and Danfeng Huang. "A review of imaging techniques for plant phenotyping." Sensors 14.11 (2014): 20078-20111.https://doi.org/10.3390/s141120078

[2] Mohan, M., Nair, S., Bhagwat, A., Krishna, T.G., Yano, M., Bhatia, C.R. and Sasaki, T., 1997. Genome mapping, molecular markers and marker-assisted selection in crop plants. Molecular breeding, 3(2), pp.87-103.https://link.springer.com/article/10.1023/A:1009651919792

[3] Russell, Stuart, and Peter Norvig. "Artificial intelligence: a modern approach." (2002).https://storage.googleapis.com/pub-tools-public-publication-data/pdf/27702.pdf

[4] Magan, N. and Aldred, D., 2007. Post-harvest control strategies: minimizing mycotoxins in the food chain. International journal of food microbiology, 119(1-2), pp.131-139.https://dspace.lib.cranfield.ac.uk/bitstream/handle/1826/2390/Post-Harvest control strategies-2007.pdf?sequence=1

[5] Inzé, D. and De Veylder, L., 2006. Cell cycle regulation in plant development. Annu. Rev. Genet., 40, pp.77-105.https://www.annualreviews.org/doi/abs/10.1146/annurev.genet.40.110405.090431

[6] Krause, G.H. and Weis, E., 1984. Chlorophyll fluorescence as a tool in plant physiology. Photosynthesis research, 5(2), pp.139-157.https://link.springer.com/article/10.1007/BF00028527

[7] Kaplan, D.R., 2001. The science of plant morphology: definition, history, and role in modern biology. American Journal of Botany, 88(10), pp.1711-1741.https://bsapubs.onlinelibrary.wiley.com/doi/full/10.2307/3558347

[8] Jiang, Y. and Li, C., 2020. Convolutional neural networks for image-based high-throughput plant phenotyping: a review. Plant Phenomics, 2020.https://www.sciencedirect.com/science/article/pii/S2095311919628030

[9] Wei, W.U., YANG, T.L., Rui, L.I., Chen, C.H.E.N., Tao, L.I.U., Kai, Z.H.O.U., SUN, C.M., LI, C.Y., ZHU, X.K. and GUO, W.S., 2020. Detection and enumeration of wheat grains based on a deep learning method under various scenarios and scales. Journal of Integrative Agriculture, 19(8), pp.1998-2008.https://www.sciencedirect.com/science/article/pii/S2095311919628030

[10] Zhao, Z.Q., Zheng, P., Xu, S.T. and Wu, X., 2019. Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), pp.3212-3232.https://arxiv.org/pdf/1807.05511.pdf&usg=ALkJrhhpApwNJOmg83O8p2Ua76PNh6tR8A

[11] Pathak, A.R., Pandey, M. and Rautaray, S., 2018. Application of deep learning for object detection. Procedia computer science, 132, pp.1706-1717.https://www.sciencedirect.com/science/article/pii/S1877050918308767

[12] Tang, C., Feng, Y., Yang, X., Zheng, C. and Zhou, Y., 2017, July. The object detection based on deep learning. In 2017 4th International Conference on Information Science and Control Engineering (ICISCE) (pp. 723-728). IEEE.https://ieeexplore.ieee.org/abstract/document/8110383/

[13] Wu, X., Sahoo, D. and Hoi, S.C., 2020. Recent advances in deep learning for object detection. Neurocomputing, 396, pp.39-64.https://arxiv.org/pdf/1908.03673

[14] Solovyev, R., Wang, W. and Gabruseva, T., 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing, p.104117.https://doi.org/10.1016/j.imavis.2021.104117

[15] Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J. and Le, Q.V., 2020, August. Learning data augmentation strategies for object detection. In European Conference on Computer Vision (pp. 566-583). Springer, Cham.https://arxiv.org/pdf/1906.11172

[16] Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S. and Vercauteren, T., 2018. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation.https://openreview.net/pdf?id=Byxv9aioz

[17] Chen, Y., Zhang, P., Li, Z., Li, Y., Zhang, X., Meng, G., Xiang, S., Sun, J. and Jia, J., 2020. Stitcher: feedback-driven data provider for object detection. arXiv preprint arXiv:2004.12432.https://arxiv.org/pdf/2004.12432

[18] Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.https://arxiv.org/pdf/2004.10934

[19] Liu, S., Qi, L., Qin, H., Shi, J. and Jia, J., 2018. Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8759-8768).https://arxiv.org/abs/1803.01534

[20] David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgessner, N., Ishikawa, G., Nagasawa, K., Badhon, M.A. and Pozniak, C., 2020. 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. arXiv preprint arXiv:2005.02162.https://arxiv.org/abs/2005.02162

[21] DeVries, T. and Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.https://arxiv.org/abs/1708.04552

[22] Zhang, H., Cisse, M., Dauphin, Y.N. and Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.https://arxiv.org/abs/1710.09412

[23] Wei, Z. and Duan, C., 2020. AMRNet: Chips Augmentation in Areial Images Object Detection. arXiv preprint arXiv:2009.07168.https://arxiv.org/abs/2009.07168

[24] Ghosal, P., Nandanwar, L., Kanchan, S., Bhadra, A., Chakraborty, J. and Nandi, D., 2019, February. Brain tumor classification using ResNet-101 based squeeze and excitation deep neural network. In 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP) (pp. 1-6). IEEE.https://ieeexplore.ieee.org/document/8882973

[25] Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114). PMLR.https://arxiv.org/abs/1905.11946

[26] Deng, J., Xuan, X., Wang, W., Li, Z., Yao, H. and Wang, Z., 2020, November. A review of research on object detection based on deep learning. In Journal of Physics: Conference Series (Vol. 1684, No. 1, p. 012028). IOP Publishing.https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012028

[27] Ghiasi, G., Lin, T.Y. and Le, Q.V., 2018. Dropblock: A regularization method for convolutional networks. arXiv preprint arXiv:1810.12890.https://arxiv.org/abs/1810.12890

[28] Jörgensen, E., Zach, C. and Kahl, F., 2019. Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss. arXiv preprint arXiv:1906.08070.https://arxiv.org/abs/1906.08070

[29] Yao, Z., Cao, Y., Zheng, S., Huang, G. and Lin, S., 2020. Cross-iteration batch normalization. arXiv preprint arXiv:2002.05712.https://arxiv.org/abs/2002.05712

[30] Chou, C.J., Chien, J.T. and Chen, H.T., 2018, November. Self adversarial training for human pose estimation. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 17-30). IEEE.https://arxiv.org/abs/1707.02439

[31] Wang, C.Y., Bochkovskiy, A. and Liao, H.Y.M., 2020. Scaled-YOLOv4: Scaling Cross Stage Partial Network. arXiv preprint arXiv:2011.08036.https://arxiv.org/abs/2011.08036

[32] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.https://arxiv.org/abs/1405.0312

[33] Wang, G., Li, W., Ourselin, S. and Vercauteren, T., 2018, September. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In International MICCAI Brainlesion Workshop (pp. 61-72). Springer, Cham.https://arxiv.org/abs/1810.07884

[34] Neubeck, A. and Van Gool, L., 2006, August. Efficient non-maximum suppression. In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 3, pp. 850-855). IEEE.https://ieeexplore.ieee.org/document/1699659