Image Classification of Nutrient Deficiencies Technical Report

Xinyao Liu *
Xi'an Jiaotong University
Xi'an
2205124667@stu.xjtu.edu.cn

Guang Liang †*
Xi'an Jiaotong University
Xi'an
2204313319@stu.xjtu.edu.cn

Abstract

This technical report provides an overview of our solution submitted to the 8th Workshop on Computer Vision in Plant Phenotyping and Agriculture (CVPPA) at the IEEE/CVF International Conference of Computer Vision (ICCV) 2023 for image classification. Effective plant phenotyping is of utmost importance to support the sustainability of our planet and its inhabitants. The involvement of strong community structures and computer vision scientists in this field is now more critical than ever. We utilized the mmpretrain framework for image training and experimented with various models, including ResNet, ViT, Swin Transformer v1, and Swin Transformer v2. Ultimately, we fine-tuned the Swin v2 model. Locally, we divided the provided dataset (trainval) into training and validation sets to obtain output results for different models. We selected the model that performed the best on the validation set, then retrained it on all training set images and conducted inference for submission. In the end, our model achieved an average top-1 accuracy of 93.3% on two datasets.

1. Introduction

In recent years, significant progress has been made in the fields of computer vision and machine learning, fundamentally transforming various domains, including agriculture and plant phenotyping. The capability to automatically analyze and classify visual data holds the promise of greatly enhancing our understanding of crop health and yield. One crucial aspect is the detection of nutrient deficiencies in crops, which can have far-reaching implications for agricultural output and food security.

This technical report aims to provide an overview of our image classification solutions submitted to the 8th Workshop on Computer Vision in Plant Phenotyping and Agriculture (CVPPA) at the IEEE/CVF International Conference on Computer Vision (ICCV) 2023. The primary objective of this challenge is to employ computer vision techniques to classify nutrient deficiencies in winter wheat and winter rye, utilizing a carefully curated dataset known as DND-Diko-WWWR. This dataset comprises 3,600 RGB images captured by unmanned aerial vehicles (UAVs) and offers a unique opportunity to explore the impact of nutrient variations on crop health.

Our approach is based on the mmpretrain framework and is aimed at addressing nutrient deficiencies in winter wheat and winter rye through the training of image classification models. In the process, we explored several different model architectures, including ResNet, ViT, Swin Transformer V1, and Swin Transformer V2, among others. Ultimately, we conducted detailed parameter tuning on the Swin Transformer V2 model and locally partitioned the provided dataset (trainval) into training and validation sets to evaluate the performance of each model.

During the model selection phase, we chose the model that performed best on the validation set and further fine-tuned it using the entire training dataset for inference and generating the final classification results. Our ultimate model achieved impressive results, averaging a top-1 accuracy of 93.3% across both datasets, providing a robust solution for addressing nutrient deficiencies in winter wheat and winter rye.

2. Dataset

DND-Diko-WWWR[6] is a UAV-based RGB dataset specifically curated for the classification of nutrient deficiencies in winter wheat and winter rye within the context of a long-term fertilizer experiment. This dataset is characterized by its precision and relevance, as it offers image-level labels that facilitate the accurate identification of nutrient deficiencies in these crops.

One notable feature of the DND-Diko-WWWR dataset is its exceptional class balance, ensuring an equal distribution of samples across different nutrient deficiency categories. This balance enhances the dataset's suitability for training machine learning models, as it mitigates the risk of

*equal contribution
†Corresponding author

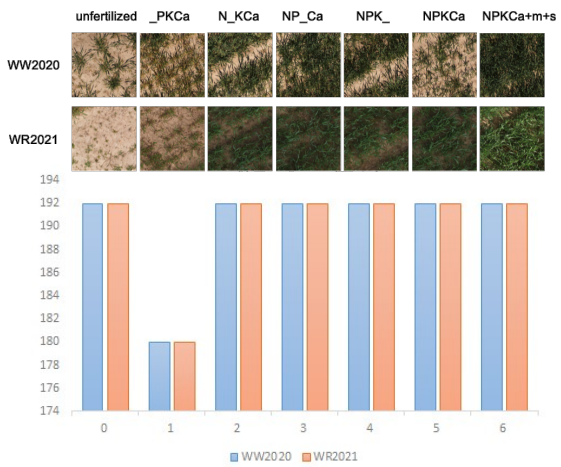

Figure 1.

biases that can arise from imbalanced class distributions.

Additionally, the dataset exhibits a remarkable absence of discernible noise or spurious images, further contributing to its quality and reliability. This absence of noise ensures that the dataset maintains a high signal-to-noise ratio, allowing us in the competition to focus on the core task of nutrient deficiency classification without the interference of irrelevant or erroneous data points.

3. Method

3.1. Mainstream Method

VGG: VGG[5] is a classic convolutional neural network architecture primarily used for image classification tasks. Its hallmark is its depth, consisting of multiple convolutional and pooling layers. VGG's simplicity and ease of reproducibility make it an attractive choice, and its deep structure aids in learning intricate features. However, its large model size, substantial parameter count, and slower training and inference speeds make it less competitive compared to newer models in terms of performance.

ResNet (Residual Networks): ResNet[2], short for Residual Networks, stands as a seminal deep convolutional neural network architecture widely employed in tasks like image classification, object detection, and semantic segmentation. It uniquely addresses the vanishing gradient problem by introducing residual connections, enabling the construction of exceptionally deep networks. Its advantages lie in its ability to capture complex features and its ease of training.

ViT (Vision Transformer): Vision Transformer, or ViT[1], is a novel image classification model based on self-attention mechanisms that partition images into smaller patches and employ self-attention to capture both global and local information. ViT has demonstrated exceptional performance in various visual tasks, including image classification and object detection. Its advantages include a straightforward model architecture, adaptability to different image resolutions, and excellent performance when trained on large-scale datasets. Nevertheless, ViT may not perform as well on smaller datasets and incurs higher computational and memory costs.

Swin Transformer: Swin Transformer[4] is an emerging self-attention model designed for image classification and object detection tasks. It introduces windowed attention mechanisms to reduce computational complexity while maintaining strong performance across multiple visual tasks. Swin Transformer's strengths include versatility across different vision tasks, efficient windowed attention for reduced computational demands, and overall strong performance.

Swinv2 (Swin Transformer Version 2): Swinv2[3], an improved version of Swin Transformer, continues to excel in image classification and object detection tasks. It inherits the strengths of its predecessor, offering even better performance and generalization capabilities.

Our Method Our image classification task is conducted using the MMPretrain framework. We have explored various relevant models and methods, and ultimately chose the Swin Transformer v2 model as the approach for submission, as shown in the figure 2. During training, we utilized the model's associated pre-trained model with 1k images.

3.2. Data Augmentation

Data augmentation is a widely employed technique in deep learning, primarily utilized to enhance model performance by expanding the training dataset and enhancing the model's generalization capabilities. The data augmentation techniques we employ are described as follows.

Mixup:[8] For each iteration, we randomly select two examples, denoted as (x_i, y_i) and (x_j, y_j). We then construct a new example through a weighted linear interpolation of these two instances:

$$\begin{aligned} \hat{x} &= \lambda x_i + (1-\lambda) x_j \\ \hat{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{1}$$

where λ is a random number chosen from the interval $[0, 1]$.

CutMix:[7] CutMix is primarily employed to increase the diversity of the training dataset, thereby enhancing model performance and generalization capability.

CutMix operates as follows: During each training iteration, two random samples, denoted as (x_i, y_i) and (x_j, y_j), are first selected. Subsequently, a new sample is constructed through a weighted linear interpolation of these two samples:

$$\begin{aligned} \hat{x} &= M \odot x_i + (1-M) \odot x_j \\ \hat{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{2}$$

Here, λ is a random number chosen from the interval $[0, 1]$,

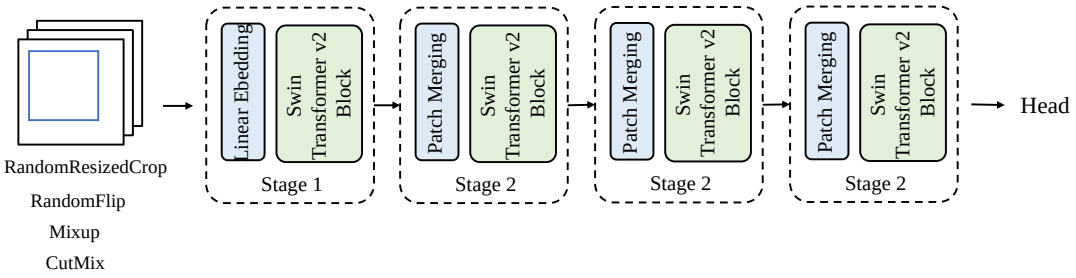

Figure 2. Network Structure

Table 1. Training Details for Image Classification Networks Based on MMretrain

Model	Model Size	Neck	Window_size	Pretrain	Img_size	Top-1 acc[1](WW)	Top-1 acc(WR)
resnet	resnet50	LinearClsHead	-	in1k	224×224	68.91	90.26
resnet	resnet50	LinearClsHead	-	in1k	896×896	77.53	88.39
swinv1	small	LinearClsHead	7	in1k	224×224	52.43	61.80
swinv1	small	LinearClsHead	7	in1k	896×896	85.39	94.76
swinv1	small	LinearClsHead	28	in1k	896×896	87.64	94.38
swinv1	base	LinearClsHead	7	in1k	896×896	89.51	94.76
swinv1	small	ArcFaceClsHead	7	in1k	896×896	80.90	91.01
swinv2	small	LinearClsHead	8	in1k(w8)	224×224	34.83	41.57
swinv2	small	LinearClsHead	4	in1k(w8)	896×896	88.02	95.51
swinv2	small	LinearClsHead	8	in1k(w8)	896×896	92.88	97.38
swinv2	small	LinearClsHead	28	in1k(w8)	896×896	93.26	97.75
swinv2	base	LinearClsHead	8	in1k(w8)	896×896	93.26	98.13

[1] Select the result that performs best on the validation set in the last 5 epochs of training

and M is a binary mask used to determine which parts come from x_i and which come from x_j.

The core idea behind CutMix is to blend two samples together, increasing the diversity of the data and helping the model better understand relationships between different regions. This reduces the model's reliance on local features, improves its generalization performance, and effectively mitigates overfitting issues. By introducing CutMix, we encourage the model to adapt better to varying data distributions during training, resulting in improved performance across various applications.

3.3. Loss

We use **LabelSmooth Loss** to compute the loss. LabelSmooth Loss is a loss function used in deep learning, typically used in conjunction with Label Smoothing to improve model training and generalization performance.

The definition of LabelSmooth Loss is as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\left[\alpha\cdot\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)+(1-\alpha)\cdot\frac{1}{K}\sum_k e^{f_k}\right] \tag{3}$$

where N is the number of samples in a batch, K represents the total number of categories in a classification problem, and α is a hyperparameter ranging between 0 and 1. f_i denotes the model's output for the i-th sample.

The key idea behind LabelSmooth Loss is to incorporate the concept of label smoothing into the loss function to reduce the model's confidence in its predictions for each sample. This encourages the model to be more cautious about each category during training, avoiding excessive confidence in one class and, thus, improving the model's generalization performance.

By introducing LabelSmooth Loss, we better constrain the model's outputs to adapt to various uncertainties, enhancing the model's robustness in various applications. This approach is commonly used for classification problems and can be combined with other regularization techniques as needed to achieve better results.

4. Experiment

The DND-Diko-WWWR dataset contains 7 categories both for WW2020 and WR2021. Following the guidelines of the challenge, we use top-1 accuracy to evaluate the classification results on each dataset(WW2020 or WR2021).

In our local experiments, we split the given training set

(trainval) into a training set and a validation set using an 80-20 ratio. Training details are provided in the table 1. The reported top-1 accuracy here represents the best performance achieved by the model on the validation set during the last 5 epochs of training. For the final submission, we trained on all the images within the given training set and then submitted the inference results.

During inference, we employed the TTA (Test Time Augmentation) technique. In addition to conventional data augmentation techniques such as VerticalFlip, HorizontalFlip, and Fivecrop, we also found that refraining from applying data augmentation on lighting and contrast (ImageEnhance.Brightness, ImageEnhance.Contrast) yielded improvements in inference performance.

5. Conclusion

Our approach involved training image models using the mmpretrain framework, where we explored different models such as ResNet, ViT, Swin Transformer v1, and Swin Transformer v2. Fine-tuning was performed on the Swin v2 model. Locally, we divided the provided dataset (trainval) into training and validation sets to obtain output results for various models. We selected the model that exhibited the best performance on the validation set and then conducted retraining on the entire training dataset, followed by inference for submission.

Our model achieved a top-1 accuracy of 94.9% on the WW2020 dataset and a top-1 accuracy of 91.7% on the WR2021 dataset. This outcome underscores the potential of computer vision in addressing the critical challenges of plant phenotyping and agriculture, and we remain committed to exploring further advancements.

6. Acknowledgement

I would like to express my heartfelt gratitude to my mentors, family, friends, and the National Training Program of Innovation and Entrepreneurship for Undergraduates for their invaluable support during the completion of this dissertation. Your guidance, encouragement, and financial support have played a crucial role in my research achievements. Thank you very much for your support and assistance.

References

[1] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[2] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[3] Ze Liu et al. “Swin transformer v2: Scaling up capacity and resolution”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 12009–12019.

[4] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[5] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[6] Jinhui Yi et al. “Non-Invasive Diagnosis of Nutrient Deficiencies in Winter Wheat and Winter Rye Using Uav-Based Rgb Images”. In: *Available at SSRN 4549653* ().

[7] Sangdoo Yun et al. “Cutmix: Regularization strategy to train strong classifiers with localizable features”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6023–6032.

[8] Hongyi Zhang et al. “mixup: Beyond empirical risk minimization”. In: *arXiv preprint arXiv:1710.09412* (2017).