Leveraging Data-Free Pruning through Gradient Analysis towards Lightweight Maize Disease Detection

Carlos PADEIRO[y], Tse-Wei CHEN[yy], Takahiro KOMAMIZU[yyy], and Ichiro IDE[yyyy]

y Nagoya University, Graduate School of Informatics, Department of Intelligent Systems
Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601 Japan

E-mail: ypadeiroc@cs.is.i.nagoya-u.ac.jp, yytwchen@ieee.org, yyytaka-coma@acm.org, yyyyide@i.nagoya-u.ac.jp

Abstract The traditional method of diagnosing crop diseases by expert visual observations is expensive, time-consuming, and prone to errors. Convolutional Neural Network (CNN) models o□er promising alternatives, but are not easily accessible to farmers in rural areas due to high resource requirements and reliance on Internet connectivity. Lightweight models operating on resource-limited devices without Internet access are crucial to address this gap, where pruning is a promising approach. This report introduces a novel pruning method that uses gradient analysis. By examining the weight channel with a low magnitude value of gradient descent, we identify and eliminate less significant weights, without compromising its accuracy. Experiments show that the proposed method improves the accuracy, and reduces the number of parameters and inference time.

Key words Channel pruning, data-free pruning, crop diseasse, image classification

1. Introduction

Crop diseases pose a formidable challenge to global food security, causing substantial production and economic losses, estimated at $220 billion annually by the United Nations Food and Agriculture Organization (FAO) [12]. These diseases not only result in an insu□cient human food supply, but also disrupt farmers' income-generating activities. Maize, one of the dominant food crops, is particularly susceptible to various diseases, despite its global yield of $1.2 billion in 2020, with a productivity of 6.0 t/ha [13]. Conventional crop disease diagnosis methods rely on visual inspection by experts, utilizing their in-depth knowledge of crop diseases, and their symptoms. This process is not only time-consuming and expensive, but also prone to human error due to subjective perception.

The advent of Deep Neural Networks (DNNs), particularly in image processing techniques, has revolutionized precision agriculture promising to reduce labor costs, and high accuracy [43]. Previous studies, such as that by Zhang et al. [54], have proposed improved CNN (Convolutional Neural Network) models for crop disease detection. However, these models are computationally expensive and require significant memory due to over-parameterization, a common characteristic of DNNs [9]. This imposes high computational and memory demands for inference, making these solutions less accessible to farmers. Moreover, to address connectivity issues in remote cultivation areas, deploying these models on resource-limited devices is essential for wider adoption.

To address these challenges, we propose a lightweight object detection DNN model specifically designed to detect crop diseases. In this report, specifically, pruning [22] is the process of simplifying neural networks by removing unimportant and redundant weights, channels, filters, or neurons. It is usually divided into three types: unstructured [14], [45], semi-structured [17], [38], [36], and structured [35], [11], [48]. Unstructured pruning, a.k.a. non-structured pruning or weight-wise pruning, is the finest-grained case that can quickly achieve a higher compression ratio and accuracy on specialized hardware. Semi-structured pruning, a.k.a. pattern-based pruning [36], achieves high accuracy and structural regularity simultaneously. Structured pruning removes entire filters [53], channels [1], [32], Transformer attention heads [35], or even layers [37] after pruning and does not require specialized hardware. This e□ciency gain, however, may lead to a trade-o□ in accuracy compared to vanilla models. Recent research e□orts have focused on addressing this trade-o□ between e□ciency and accuracy in pruning.

For instance, Li et al. [29] assess channel importance by computing the sum of absolute weight values. Alternative channel pruning methods introduce channel-wise sparsity during training, subsequently pruning channels with low scaling factors and fine-tuning the model. Liu et al. [33] minimize the reconstruction error of feature maps between the pruned and original pre-trained models. He et al. [21] and Tang et al. [46] highlight that the significance of channels is strongly influenced by the input data and propose a method to distinct sub-networks for each input instance. However, assessing the importance of individual channels in neural networks is challenging due to the complexity of operations such as activation functions and batch normalization. One potential approach is to evaluate a channel's significance based on the increase of the final loss caused by its removal. However, this method becomes ine□cient in very deep networks. In such cases, the shallow layers often exhibit limited discriminative capacity, as the long propagation path diminishes their influence on the final output.

In contrast to previous studies, we propose BlockWIPG (Block Weight and Image Perturbation-induced Gradients) which estimates channel importance based on three factors; (a) E□cient analysis of intermediate layers or blocks, and a divided computation of intermediate gradient analysis. (b) Examines the contribution of intermediate channels in layers or blocks by observing image variations through gradients. (c) Assesses channel importance by measuring changes in intermediate layer or block outputs when small weight perturbations are introduced. In summary, the main contributions of this report are two fold:

(1) **BlockWIPG**: We propose a new pruning methodology that in-