一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IMQ2023-90,IE2023-145,MVE2023-119(2024-03)

Towards Similarity-Preserving Post-Training Quantization for Lightweight Maize Disease Detection

Carlos VICTORINO PADEIRO[y], Tse-Wei CHEN[y], Takahiro KOMAMIZU[y], and Ichiro IDE[y]

y Nagoya University, Nagoya, Aichi 464-8601 Japan

E-mail: ypadeiroc@cs.is.i.nagoya-u.ac.jp ; twchen@ieee.org ; taka-coma@acm.org ; ide@i.nagoya-u.ac.jp

Abstract The traditional crop disease diagnosis process, relying on expert visual observation, is costly, time-consuming, and prone to human error. Though Convolutional Neural Network (CNN) o□ers promising alternatives, it su□ers from high resource demands, making them inaccessible to farmers. Lightweight models that can work on resource-restricted devices without network access are needed to address this issue. In this perspective, we propose a solution that leverages Post-Training Quantization (PTQ) to convert high-precision models into lower-precision ones while preserving similar features. While quantization is a promising method for building lightweight CNN models for crop disease detection, the quantized models' quality needs to be improved. To overcome this problem, we propose a PTQ method called Similarity-Preserving Quantization (SPQ), which ensures equivalent activation patterns for similar crop images in both original and quantized models. Experimental evaluation applying SPQ to MobileNetV2 showed that the proposed method improved about 30% of throughput while keeping detection performance.

Key words Quantization, post-training quantization, similarity-preserving, crop Disease.

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) [5]. These diseases contribute to insu□cient human food supply and 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 [6]. 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 time-consuming, expensive, and prone to human error due to subjective perception. The advent of Convolutional Neural Networks (CNN), particularly in image processing techniques, has revolutionized precision agriculture, labor costs, and high accuracy [21].

Previous studies, such as Zhang et al. [33], have proposed improved deep CNN for crop disease detection. Besides, these models are computationally expensive and require significant memory due to over-parameterization, a common characteristic of deep neural networks [3]. This imposes high computational and memory demands for inference, making these solutions less accessible to farmers. Moreover, to address the network connectivity issues in remote cultivation areas, deploying these models on resource-constrained devices is essential for broader adoption. To address these challenges, we develop a lightweight object detection neural network model designed explicitly to detect crop diseases.

This applies Post-Training Quantization (PTQ) [8] to reduce a neural network model's memory and computational requirements. PTQ is widely regarded as one of the most e□cient compression methods practically, benefitting from its data privacy and low computational costs. Emerging as a promising solution to resource limitations, PTQ enables deploying resource-e□cient CNN for crop disease detection. Unlike traditional quantization requiring extensive calibration data and retraining, PTQ minimizes computational overhead by bypassing iterative fine-tuning. This e□ciency gain, however, may lead to a minor trade-o□ in accuracy compared to full-precision models.

Recent research e□orts have addressed this trade-o□ between accuracy and e□ciency in PTQ. For instance, Nagel et al. [15] introduced soft quantization with learnable parameters by constructing new optimization functions based on second-order Taylor expansions of the loss functions before and after quantization. This approach e□ectively balances model accuracy and e□ciency. Li et al. [12] proposed a block-by-block reconstruction method instead of the traditional layer-by-layer approach and utilized diagonal Fisher matrices to approximate the Hessian matrix, conserving more information during quantization. This strategy further improved quantization accuracy without compromising e□ciency. Wei et al. [31] discovered that randomly disabling a subset of activation quantization elements can smooth the loss surface of the quantization weights, leading to improved accuracy. Though PTQ has been studied and improved, these existing methods have yet to consider pairwise activation similarities between the full-precision and quantized models.

This presents a quantization reconstruction method called SPQ (Similarity-Preserving Quantization) that preserves pairwise activation similarities between input pairs in the quantized model rather than directly mimicking the representations of the full-precision models. In summary, the main contributions of this method are three-fold:

(1) **SPQ**: We propose a new method focusing on preserving pairwise activation similarities between input pairs in the quantized model and full-precision representation.
(2) **Loss Function**: We apply a reconstruction error for preserving layer/block similarity between full-precision and quantized models to object detection neural network architectures.
(3) **Validation**: We validate that SPQ enhances quantized network calibration outcomes and o□ers a valuable adjunct to established PTQ techniques.

2. Related Works

This section analyzes combinations of quantization techniques for enhancing resource-constrained disease detection applications.

2.1 Crop Disease Detection

Current techniques in precision agriculture are commonly preceded by analyzing the images captured by devices and sensors. These images are then used to detect crop diseases. The problem of crop disease detection has been addressed in numerous studies. However, most of them focused only on the problem of classifying foliar diseases, such as the method by k-means clustering and deep learning to detect orange diseases and predict their names from image [9]. Their architecture is designed based on GoogLeNet's [20] inception networks and AlexNet [11] for identifying and recognizing apple leaf diseases as proposed by Liu et al [13]. A related work proposed a method based on an improved VGG-16 [24] network to identify apple leaf diseases [32]. Unlike these works, we apply object detection networks to detect crop diseases on plant leaves.

2.2 Quantization

Quantization of neural networks has been studied for a while, and there are numerous methods [8], [12], [16], [31]; all of them are based on following equation;

$$w_q = \&D0=\complement^1|^{\circ}- \tag{1}$$

$$\&D0=\complement^1|^{\circ} = 2;0<? \quad \frac{|^{\mathbb{m}}}{(}\ ,\ /;0-2^n \quad 1\ - \tag{2}$$

$$(= \frac{|<0G \quad |<8=}{2^n \quad 1}- \tag{3}$$

where S denotes the scaling factor to convert the range of w to n bit-width, and b•e is the round-to-nearest. The clamp() function restricts a given value between an upper and lower bound. While the w_q is quantized weights as outputs of a quantization function Quant() similarity from [17]. The / is used to decide which quantized value zero is mapped to.

In contrast, the dequantization integer value represents w^0 2 $\mathbb{R}$ obtained by.

$$w^0 = w_q{}^1(\quad /^{\circ} \tag{4}$$

The activation values quantization procedure is analogous to the weight values quantization, except that the minimum and maximum values are determined by analyzing activations from a limited calibration data set and employing a moving average.

2.3 Post-Training Quantization

Quantization is a powerful technique for compressing neural networks, enabling their deployment on resource-constrained devices, by applying Eq. 2. Two primary quantization methodologies exist Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ). QAT [4] [8] [23] incorporates quantization into the network training phase, while PTQ [17] applies quantization after training completion. PTQ o□ers significant computational advantages, making it the preferred choice for network deployment. The primary objective of PTQ is to determine the quantization parameters for weights and activations in each layer. Despite incorporating fine-tuning during quantization, these PTQ methods remain distinct from QAT. QAT employs the entire labeled training dataset to adjust the model's weights, while PTQ solely optimizes the quantization parameters using a subset of unlabeled data, making it e□cient. Post-Training Quantization (PTQ) has emerged as a promising solution to address these challenges, enabling the deployment of CNN with significantly reduced memory footprint and computational complexity. Traditional quantization methods, such as full-precision training followed by quantization, often require large amounts of calibration data to fine-tune the quantized model, resulting in substantial computational overhead. In contrast, PTQ eliminates the need for iterative quantization training, significantly reducing computational costs and enabling e□cient model deployment. However, This efficiency often comes at the partial sacrifice of accuracy due to the reduction in precision, which can lead to information loss and a diminished ability to represent fine-grained details in the model.

2.4 Quantization Reconstruction Error

This sacrifice of accuracy is associated with the error that occurs when a neural network model is quantized, typically to lower-precision numerical representations. This quantization error can be controlled by quantization reconstruction error, which acts as a regularizer that reduces generalization error by aligning corresponding components of the quantized and full-precision models. AdaRound [15] analyzes that it is not advisable to round full precision weight to its nearest fixed-point value and proposes a novel rounding mechanism that assigns a continuous variable to each weight value, determining whether it should be rounded up or down rather than employing the traditional nearest rounding method. BRECQ (Block REConstruction Quantization) [12] establishes block-wise reconstruction between the full-precision and quantized network outputs, balancing cross-layer dependency and generalization error. Additionally, BRECQ incorporates trainable clipping for activations. Similar methods have been explored in earlier works [2] [7]. AQuant [27] enhances activation quantization strategy and overall quantization performance, although at the cost of increased inference overhead. PD-Quant [14] addresses the discrepancy between the distribution of calibration activations and their corresponding real activations by proposing a technique for adjusting the calibration activations accordingly. Previous research has investigated the influence of the calibration dataset on the performance of quantized models [7]. Besides, the study [1] has explored the reconstruction of features by calculating the feature output distance between quantized and full-precision models, making the quantized model mimic the full-precision model. In contrast to the previous method, in SPQ, the quantized model is not required to mimic the representation space of the full-precision but rather to preserve the pairwise similarities in its own representation space.

3. Methodology

The central concept of the proposed method is exploring important information in the activation map of the full-precision model and transferring this vital information into the quantized model. Moreover, we set / equal to zero to eliminate the zero-point o□set in Eq. 2 and Eq. 4; this method simplifies the accumulation operation and reduces computational overhead. However, this simplification comes at the cost of a restricted mapping between the integer and floating-point domains. While suitable for one-tailed distributions like ReLU activations. As claimed by Nagel et al. [17].

This uses a di□erent method to reconstruct each quantized model by layer or block named Block Similarity-Preserving for Post-Training Quantization (SPQ). This method is inspired from [12], [25]. Tung et al. [25] apply similarity preservation at the batch level, focusing primarily on the similarity of the last convolution layers, while SPQ focuses on intermediate blocks. While Li et al. [12] focusd on applying Fisher information, SPQ focuses on similarity-preservation instead.

3.1 Overview of SPQ

Figure 1 shows the overall procedure of the proposed method. SPQ is a technique for reconstructing a quantized neural network model using knowledge from a full-precision model. Unlike the BRECQ [12] and QDrop [29], reconstruction methods match output values or class probabilities, eliminating the di□erence in activation output. SPQ aims to preserve the similarity in the relationships

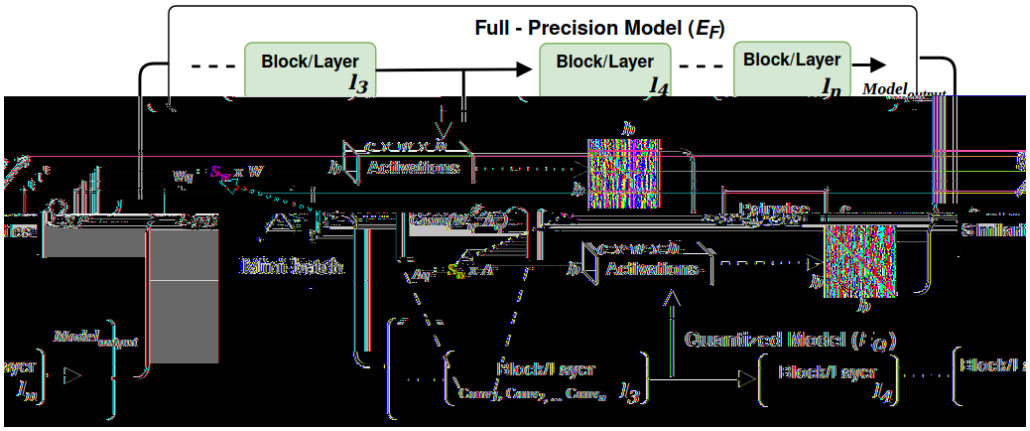

Fig. 1: Overview of the proposed SPQ. A pipeline overview of the proposed method is applied to find activation scaling factors ($(_O$) to quantize the activation. Yellow and green rectangles denote the quantized and full-precision layer/block, respectively. $L_{(\%}$ is similarity-preserving quantization reconstruction loss that can be combined with di□erent loss functions.

between activations in the full-precision and quantized. Such that input pairs' samples and their corresponding relations across the feature space that produce similar or dissimilar activations in the full-precision and the quantized model are maintained. In the context of quantization reconstruction error, we hypothesize that aligning the activation patterns produced by the quantized model with those generated by the full-precision model for highly similar input pairs, as measured by a chosen similarity metric, can yield benefits for quantization reconstruction error. Moreover, we focus on integrating similar information into constructing a new data representation in a quantized model, significantly improving quantization reconstruction error tasks. More importantly, the proposed idea can be readily applied to other post-training quantization methods such as BRECQ [12] and PD-Quant [14]. Furthermore, we aim to preserve the original data learning similarity information from the full precision model. To this end, we use the widely used spatial and channel similarity activation map.

The following gives the formal explanation of SPQ. For a given mini-batch of input pair samples, let the activation map generated by the full-precision model (resp. the quantized model $\&$) at a specific layer or block $;$ as $^{1;o}$- $_{\&}^{1;o}$ 2 $\mathbb{R}^{11\ 2}$ $|^{o}$, where 1 represents the batch size, 2 signifies the number of output channels, and and | denote the spatial dimensions. We define quantization reconstruction loss that penalizes di□erences in the $!_2$-normalized inner products of $^{1;o}$ and $_{\&}^{1;0o}$. Suppose - is either or $\&$,

$$/_{-}^{1;o} = \frac{A\ {}_{-}^{1;o}\ \ A\ {}_{-}^{1;o}\ >}{\Gamma}- \tag{5}$$

where $A^{1\ o}$ is a reshape function of $_{-}^{1;o}$ (detail will be in the subsequence section), and Γ is the normalization factor. Intuitively, entry $^{18-\ 9o}$ in $/_{-}^{1;o}$ encodes the similarity of the activations at $;$ elicited by the 8-th and 9-th images in the mini-batch. We define the similarity-preserving quantization reconstruction loss as follows:

$$L_{(\%^{1}\ -\&^{o}} = \frac{1}{1^2} \tilde{O}_{1;-;^{0o}2:} /^{1;o} \quad /_{\&}^{1;0o>\ 2}{}_{5}- \tag{6}$$

where $:$ collects the corresponding layer pairs (e.g., layer pair at the end of the same block) and $k\ k_5$ is the Frobenius norm. Eq. 6 represents a summation, across all paired layers $^{1;-;^{0o}}$ 2 $:$, of the mean squared element-wise di□erence between the Gramian matrices $/^{1;o}$ and $/_{\&}^{1;0o}$ of the full-precision and quantized model, respectively. Finally, the total loss for training the student network is defined as:

$$L = L \quad {}^{1}G\text{-}W^{1\sim oo}\ ,\ VL_{(\%^{1}\ -\&^{o}} \tag{7}$$

where V represents the regularization loss imposed on quantization reconstruction loss, and L is any Knowledge Distillation loss to regularize the output probabilities of the quantized model.

3.2 Similarity-Preservation Strategy

SPQ reconstructs quantized features through pairwise similarity across spatial, channel, and batch dimensions. This multi-level similarity leverages fine-grained information for e□ective reconstruction. Further details on this method are discussed in this section.

- **Batch Similarity-Preservation**: Semantically similar images exhibit high pairwise similarity of activation maps, while dissimilar images exhibit low pairwise similarity. This property can be exploited during calibrating by measuring pairwise similarities within the activation map of a batch of images obtained from the full-precision model. These relationship similarities among image batches can then be used to guide the calibration of the quantized model. Batch similarity-preserving is computed by re-shape function $A_{\text{batch}} : \mathbb{R}^{11\ 2\ |^{o}} ! \mathbb{R}^{11\ 12\ \ |^{oo}}$. Therefore, $/^{1;o}\text{-}/_{\&}^{1;o}$ 2 $\mathbb{R}^{11\ 1^{o}}$.
- **Spatial Similarity-Preservation**: Unlike batch similarity, spatial pairwise similarity measures the proximity between individual pixels within an image based on pixel-wise correlation. It computes at the image-level by $A_{\text{spatial}} : \mathbb{R}^{11\ 2\ \ |^{o}} ! \mathbb{R}^{11\ 1\ |^{o}\ 2^{o}}$. Therefore, $/^{1;o}\text{-}/_{\&}^{1;o}$ 2 $\mathbb{R}^{11\ 1\ |^{o}\ 1\ |^{oo}}$.
- **Channel Similarity-Preservation**: Unlike spatial similarity, it reshapes features and calculates similarity across channels, resulting in a di□erent output size. A 1×1 convolution ensures compatible channel dimensions before reshaping by $A_{\text{channel}} : \mathbb{R}^{11\ 2\ \ |^{o}} ! \mathbb{R}^{11\ 2\ 1\ |^{oo}}$. Therefore, $/^{1;o}\text{-}/_{\&}^{1;o}$ 2 $\mathbb{R}^{11\ 2\ 2^{o}}$.
- **Spatial and Channel Similarity-Preservation**: It is achieved by fusing spatial and channel pairwise similarities. We compute the quantization reconstruction loss ($L_{(\%}$) by Eq. 7 using transformed activation maps from both types of similarities and linearly combine their individual losses.

3.3 Computational E□ciency

We use post-quantization Bit OPerations (BOPs) to evaluate accuracy-power trade-o□s at di□erent bit-width. Unlike prior

Table 1: Comparison of Faster-RCNN [19] between Full-Precision (FP-32) and Quantized models (W8A8 and W6A6). Based on MobileNetV2 [22] backbone on the validation dataset. Best value in each row is bold-faced and the numbers in braces show the improvement from the full-precision model. Note that the Inference and Throughput were measured using an image with shape [3, 4032, 3024] on the CPU.

	FP-32	W8A8	W6A6
Inference* [ms]	1,373.52	1,029.34 (" 1•33)	**1,023.23** (" 1•34)
Throughput*	0.73	0.97 (" 1•33)	**0.98** (" 1•34)
Model Memory [MiB]	346.94	111.26 (# 3•12)	**91.63** (# 3•78)
Top-1 mAP@50 [%]	37.10	**37.11** (" ¸0•01)	35.08 (# 2•02)
Top-1 F1-Score [%]	97.46	**97.47** (" ¸0•01)	95.43 (# 2•03)
Bit Operations [T]	26.80	1.67 (" 16•05)	**0.94**(" 28•51)

works [26], BOPs do not guide our quantization, but measure its e□ciency by Eq. 8, following [28].

$$\text{BOPs} = |_1 0_1 \text{MAC-} \quad (8)$$

$$\text{MAC} = 2_8 1 \ _> |_> : \ : |\, 2_> - \quad (9)$$

where $|_1$ and 0_1 are the bit-width of weights and activations, MAC (Multiply and ACcumulate) operations (Eq. 9), 2_8 and $2_>$ are the input and output channel size, $_>$ and $|_>$ are the output height and width, $:$ and $: |$ are the kernel height and width, and 1 is the batch size, respectively.

4. Experimental Evaluation

This section presents a comprehensive evaluation of the performance of the proposed algorithm. We begin by outlining the experimental setup and implementation details. Subsequently, we compare our quantized method, evaluated across various low-bit-width configurations, against the current state-of-the-art in our proposed dataset of crop disease. Finally, we conduct systematic ablation studies to gain deeper insights into the key properties and contributions of our method.

4.1 Implementation Setup

Here, we analyzed inference time for an image with a dimension of 4032 3024 3 on NVIDIA RTX A6000 GPU and CPU 2.3GHz 8-Core Intel Core i9 to demonstrate that models quantized by the proposed method can reduce the model memory and accelerate inference with negligible accuracy drop. We tested this technique on the backbone of Faster R-CNN [19] under MobileNetV2 [22]. The experiments were implemented using the PyTorch [18] framework. After quantizing the model, it was reconstructed block by block with the following setting to recover the accuracy. The weight rounding scheme adopted in our work adhered to the method specified in [15]. For other hyperparameters related to the reconstruction process, such as the number of iterations and loss ratios, we maintained consistency with those reported in QDrop [29] and BRECQ [12]. Notably, we deviated by employing an 8-bit representation for the output of the first and last layers in all experiments, which positively impacted accuracy. Batch sizes 16 and 10 epochs, Adam optimizer [10], and the initial learning rate 0.003 were used.

4.2 Settings

Dataset: We selected Maize leaves from three disease classes; namely, *Northern corn Leaf Blight* (NLB), *Fall ArmyWorm* (FAW) and *Maize Streak Virus* (MSV). The dataset comprises four di□erent datasets; more than 18,222 images annotated with 105,735 NLB lesions were collected in the USA [30]. Image datasets were collected across three di□erent Sub-Saharan African countries (Ghana, Uganda, and Namibia) in the field with two di□erent classes: FAW and MSV. The dataset was split into four sets: train, calibration, validation, and test in the proportion of 70 : 20 : 5 : 5. The train, validation, and test sets were used to train and test the full precision model. The calibration and test sets were applied to recalibrate and test the quantized model.

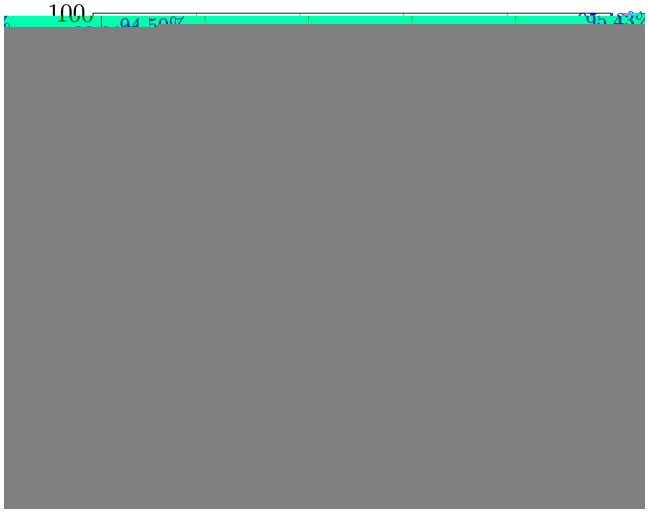

Fig. 2: Sensitivity of V hyper-parameter on Faster-RCNN [19] with MobileNetV2 [22] backbone on validation set

Metrics: We measured the network e□ciency in four dimensions; *Memory*, *Inference*, *Bit Operations*, and *Accuracy*. In terms of Accuracy, F1-Score and mean Average Precision (mAP) were used. F1-Score is the harmonic mean of precision and recall for the optimized confidence score threshold. In contrast, mAP summarizes the global trade-o□ between precision and recall. We applied mAP@50 considering the nature of some crop diseases being detected, which exhibit distinct and well-defined symptoms, making them easier to detect. However, certain diseases have more di□use symptoms and vary in appearance, making detecting them challenging. The Inference is the time needed to make a prediction, and the short time indicates that the model runs faster. The Memory measures the memory footprint needed to run the model, and the smaller memory size consumption is better, which leads to a better memory footprint. To quantify the computational e□ciency, we measured Bit OPerations (BOPs) for a single forward pass of the model using the equation Eq. 8.

4.3 E□ects of Hyper-Parameters

Note that V is a weight parameter for balancing the regularization loss imposed on quantization reconstruction loss on similarity-preserving loss in Eq. 7. From Fig. 2, we find the quantized model achieves the best performance when V is 1,000. In other words, the larger V is, the more quantized model will learn from the similar activation in the full-precision model. That is to say, the greater V is, the greater the e□ect of similarity preservation loss is. We set V to 0, 10, 100, and 1,000. We found that the similarity-preservation brought about an improvement of 0•93% on Top-1(F1-Score). When the V was equal to 1,000, the quantized model did not improve on Top-1 (mAP@50). When V gradually decreased from 10 to 0 (V = 0, without similarity-preserving), the improvement of F1-Score and mAP@50 results became smaller and smaller as 33•04% and 93•01% and 29•04%, 41•07%, respectively for Top-1 (mAP@50) and Top-1 (F1-Score). This proves that the introduction of similarity-preserving can learn the useful feature information between full-precision and quantized model to improve the quantization reconstruction error results. The overall results are in Fig. 2. It shows that the increase of V can improve the disease detection results, reflecting the quantized model's good generalization ability.

Table 2: E□ect of Quantization Faster-RCNN [19] with MobileNetV2 [22] backbone on validation dataset. The best scores in each row are bold-faced.

	Top-1 mAP@50 in [%]			Top-1 F1-Score in [%]		
	FP-32	W8A8	W6A6	FP-32	W8A8	W6A6
BRECQ [12]	37.10	37.10	34.79	97.46	**97.47**	95.29
QDrop [29]	37.10	37.08	34.11	97.46	97.44	95.11
SPQ (Ours)	37.10	**37.11**	**35.08**	97.46	**97.47**	**95.43**

4.4 Results

Table 1 shows the results of the detection model with the backbone on di□erent quantizer bit-width. Before being quantized, the model achieved 37•10% for Top-1 (mAP@50), and 97•46% for Top-1 (F1-Score), on MobileNetV2 [22] backbone. The results show that full precision (FP-32) required significant inference time and memory footprint. When the quantizer bit-width was set to W8A8 (Weight bit = 8 and Activation bit = 8), the model outperformed the full-precision by 0•01% in Top-1 (mAP@50) and Top-1 (F1-Score) respectively. However, we continued seeing significant improvement in model e□ciency with a good balance between F1-Score and detector. We reduced the inference and throughput by 1•33 times faster and Bit Operations to 1•67 value. The model memory footprint improved by 3•78 times smaller, and the Bit Operations improved more than 16•05 times compared to FP-32, which is an outstanding achievement even though increasing the accuracy in the MobileNetV2 backbone. The results show that the models e□ectively balanced accuracy, speed, and memory footprint even after quantizer bit-widths reached 8-bit.

When the quantizer bit-width reached W6A6, the model's efficiency was significantly increased and lightweight, which can be confirmed by the Bit Operations reaching the lowest value of 0•94, making the model 28•51 times e□ciency. However, the accuracy significantly dropped, achieving 35•08% Top-1(mAP@50) and 95•43% for Top-1(F1-Score), respectively, making the model slightly lose balance between accuracy and e□ciency. Nevertheless, the results show that the models achieved an e□ective improvement of speed and memory footprint, leading the model on both quantizer bit-width suitable for disease detection despite a decrease of accuracy by 2•02% and 2•03% Top-1(mAP@50) and Top-1(F1-Score) respectively.

4.5 Performance Analysis

We extensively compared SPQ with various PTQ algorithms across various bit-width configurations without misleading QAT comparisons due to their inherent training di□erences. Notably, SPQ consistently outperformed other methods, particularly at low bit-widths. Encouraging only the quantization to mimic di□erent aspects of the full precision representation space to optimize activation scaling factors proved to be insu□cient for low-bit scenarios. Therefore, SPQ focuses on integrating similar information into constructing a new data representation in a quantized model, significantly improving the optimization of rounding values and activation scaling factors. We benchmark against QDrop [29] and BRECQ [12], the strongest-performing PTQ method. To ensure consistency, we applied Spatial and Channel Similarity-Preservation as described in the methodology section. All experiments performed include this method.

Table 2 summarizes the results respectively for Top-1 (mAP@50) and Top-1 (F1-Score). The table results demonstrate substantial improvements achieved by SPQ compared to strong PTQ baselines. While gains at W8A8 were modest, they became more pronounced at lower bit-widths. For instance, SPQ outperformed QDrop and BRECQ at W8A8 settings quantization, by improving MobileNetV2 [22] by 0•01% and 0•03% for Top-1 (mAP@50), respectively. The same performance is noticed in W6A6 where SPQ also outperformed QDrop and BRECQ by 0•29% and 0•97% for Top-1 (mAP@50), respectively. The same performance trend does not change the Top-1 (F1-Score) metric. The SPQ surpasses QDrop and BRECQ at W8A8 settings quantization, by enhancing MobileNetV2 by 0•01% and 0•03%, respectively. Furthermore, we noted that performance in W6A6 where SPQ outperforms QDrop and BRECQ by 0•29% and 0•97% , respectively. This underlines the significance of SPQ's optimization strategy. Moreover, SPQ requires no additional computation for inference after optimization, ensuring e□ciency. Additionally, SPQ focuses on being hardware-friendly by allowing bit homogeneity through bit-width hyper-parameters. Like earlier PTQ work, we maintain the first and last layers at 8 bits. While some previous works reach higher accuracy with an additional 8-bit first-layer output, we experiment with these settings to validate SPQ's e□cacy.

Table 3: Loss function comparison. L_{KD} loss only, L_{SP} only, and both L , $L_{(\%}$ are compared.

	Top-1 mAP@50 in [%]		Top-1 F1-Score in [%]	
	W8A8	W6A6	W8A8	W6A6
L	29.04	29.04	39.05	41.07
$L_{(\%}$	32.01	30.03	42.01	44.00
L , $L_{(\%}$	37.11	35.08	97.47	95.43

4.6 Ablation Study

We conducted experiments of quantization reconstruction error based only on similarity-preserving loss in Eq. 7 using MobileNetV2 [22]. We applied W8A8 and W6A6 bit quantization for all layers except for the first and the last layer. From Tab 3, we can observe that Similarity-Preserving loss ($L_{(\%}$) has stable accuracy improvement at all bits. This result implies that the generalization of quantization error in $L_{(\%}$ consistently outperformed the L in all settings.

5. Conclusion

This revealed that encouraging only the quantized model to mimic di□erent aspects of the full precision representation space to optimize activation scaling factors proved insu□cient for low-bit scenarios and complex datasets. Meanwhile, we observed that capturing global structure in activation map information and preserving the original pairwise similarities between the activation map points in the full model and the quantized model in the embedding space is apromising method . The proposed technique is particularly suitable for problems sensitive to sample similarity, such as classification, detection, anddrug similarity in recommender systems, and disease detection in healthcare informatics. The proposed method can improve the performance quantization significantly. This is because our method is based on similarity, while other methods are based on Euclidean distance, which is unsuitable for complex tasks such as detection. Furthermore, our hardware-friendly method allows bit homogeneity through bit-width hyper-parameters.

Future work could explore how to improve our method, such as combining our method with self-supervised learning. We will also consider how to apply our method to model pruning.

Acknowledgement. This work was partly supported by JSPS KAKENHI JP21H0355.

References

[1] R. Banner, Y. Nahshan, E. Ho□er, and D. Soudry. ACIQ: Analytical Clipping for Integer Quantization of neural networks. *Computing*

Resaerch Repository arXiv Preprints, arXiv:1810.05723, 2018.

[2] R. Banner, Y. Nahshan, and D. Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems 32*, pages 7948–7956, 2019.

[3] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus. Exploiting linear structure within convolutional networks for e□cient evaluation. In *Advances in Neural Information Processing Systems*, volume 27, 1269–1277, 2014.

[4] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha. Learned step size quantization. *Computing Resaerch Repository arXiv Preprints*, arXiv:1902.08153, 2019.

[5] Food and Agriculture Organization of the United Nations. New standards to curb the global spread of plant pests and diseases, 2021. `https://www.fao.org/news/story/en/item/1187738/icode/` (Visited: 04-July-2023).

[6] Food and Agriculture Organization of the United Nations. Agricultural production statistics. 2000–2020. FAOSTAT Analytical Brief Series No. 41, 2022.

[7] I. Hubara, Y. Nahshan, Y. Hanani, R. Banner, and D. Soudry. Accurate post training quantization with small calibration sets. In *Proceedings of the 38th International Conference on Machine Learning*, 4466-4475, 2021.

[8] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for e□cient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2704-2713, 2018.

[9] K. Kaur and M. Kaur. Prediction of plant disease from weather forecasting using data mining. *International Journal on Future Revolution in Computer Science Communication Engineering*, 4(4):685–688, 2018.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computing Resaerch Repository arXiv Preprints*, arXiv:1412.6980, 2017.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 25:1097–1105, 2012.

[12] Y. Li, R. Gong, X. Tan, Y. Yang, P. Hu, Q. Zhang, F. Yu, W. Wang, and S. Gu. {BRECQ}: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.

[13] B. Liu, Y. Zhang, D. He, and Y. Li. Identification of apple leaf diseases based on deep convolutional neural networks. *Symmetry*, 2018.

[14] J. Liu, L. Niu, Z. Yuan, D. Yang, X. Wang, and W. Liu. PD-Quant: Post-Training Quantization Based on Prediction Di□erence Metric. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24427–24437, 2023.

[15] M. Nagel, R. A. Amjad, M. Van Baalen, C. Louizos, and T. Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7197–7206, 2020.

[16] M. Nagel, M. Baalen, T. Blankevoort, and M. Welling. Data-free quantization through weight equalization and bias correction. In *17th IEEE/CVF International Conference on Computer Vision*, pages 1325–1334, 2019.

[17] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. Baalen, and T. Blankevoort. A white paper on neural network quantization. *Computing Resaerch Repository arXiv Preprints*, arXiv:2106.08295, 2021.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[21] M. H. Saleem, J. Potgieter, and K. M. Arif. Plant disease detection and classification by deep learning. *Plants*, 8(11):468–479, 2019.

[22] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *Computing Resaerch Repository arXiv Preprints*, arXiv:1801.04381, 2018.

[23] Y. Shang, D. Xu, B. Duan, Z. Zong, L. Nie, and Y. Yan. Lipschitz continuity retained binary neural network. In *European Computer Vision Association*, 2022.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Resaerch Repository arXiv Preprints*, arXiv:1409.1556, 2016.

[25] F. Tung and G. Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[26] M. van Baalen, B. Kahne, E. Mahurin, A. Kuzmin, A. Skliar, M. Nagel, and T. Blankevoort. Simulated quantization, real power savings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2756–2760, 2022.

[27] C. Wang, D. Zheng, Y. Liu, and L. Li. Leveraging inter-layer dependency for post -training quantization. In *Advances in Neural Information Processing Systems*, 2022.

[28] Y. Wang, Y. Lu, and T. Blankevoort. Di□erentiable joint pruning and quantization for hardware e□ciency. *Computing Resaerch Repository arXiv Preprints*, arXiv:2007.10463, 2020.

[29] X. Wei, R. Gong, Y. Li, X. Liu, and F. Yu. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. *Computing Resaerch Repository arXiv Preprints*, arXiv:2203.05740, 2022.

[30] T. Wiesner-Hanks and M. Brahimi. Image set for deep learning: Field images of maize annotated with disease symptoms, 2019. `https://osf.io/p67rz/` (Visited: 04-July-2023).

[31] X. Xia, X. Xiao, X. Wang, and M. Zheng. Progressive automatic design of search space for one-shot neural architecture search. In *In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, page 2455–2464, 2022.

[32] Q. Yan, B. Yang, W. Wang, B. Wang, P. Chen, and J. Zhang. Apple leaf diseases recognition based on an improved convolutional neural network. *Sensors*, pages 3535–3547, 2020.

[33] X. Zhang, Y. Qiao, F. Meng, C. Fan, and M. Zhang. Identification of maize leaf diseases using improved deep convolutional neural networks. *IEEE Access*, 6:30370–30377, 2018.