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

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.