

QWID: Quantized Weed Identification Deep neural network

Parikshit Singh Rathore

Maharana Pratap University of Agriculture and Technology

14.parikshitsingh@gmail.com

Abstract—In this paper, we present an efficient solution for weed classification in agriculture. We focus on optimizing model performance at inference while respecting the constraints of the agricultural domain. We propose a Quantized Deep Neural Network model that classifies a dataset of 9 weed classes using 8-bit integer (int8) quantization, a departure from standard 32-bit floating point (fp32) models. Recognizing the hardware resource limitations in agriculture, our model balances model size, inference time, and accuracy, aligning with practical requirements. We evaluate the approach on ResNet-50 and InceptionV3 architectures, comparing their performance against their int8 quantized versions. Transfer learning and fine-tuning are applied using the DeepWeeds dataset. The results show staggering model size and inference time reductions while maintaining accuracy in real-world production scenarios like Desktop, Mobile and Raspberry Pi. Our work sheds light on a promising direction for efficient AI in agriculture, holding potential for broader applications.¹

I. INTRODUCTION

Weeds are undesirable plants that compete with the agricultural crop plant for resources like soil nutrients, direct sunlight, water and to some extent space to grow. The weeding process plays a significant role in agriculture because weeds produce a major loss in crop yield. With as high as 50-71% yield reduction was seen in soybean, to around 40-71% in groundnut [1]. Weeds accounted for a total of 12.65 billion dollar losses in 2007 in India alone [2].

Weed identification plays an important role in weeding because deriving the weed type with appropriate hoeing depth, hoeing positions, and particular herbicides can be used [3]. Automating these processes minimizes human intervention which is high in cost and also labor-intensive.

Training the models on DeepWeeds dataset [4] consisting of 9 classes namely chinese apple, lantana, parkinsonia, parthenium, prickly acacia, rubber vine, siam weed, snake weed and negatives (other non-target plant life). The dataset was prepared in real-world conditions like dark shadows, canopy cover, high contrast, and variable distance between the camera and the plant. Sample of each class is represented in Figure 1.

Models used originally for this dataset were ResNet-50 [5] and Inceptionv3 [6], both transfer learned on the DeepWeeds dataset. We put forward the quantized versions of both ResNet-50 and Inceptionv3, transfer learned and fine-tuned to achieve



Fig. 1: Dataset samples [4]

almost the same accuracy but with significantly better inference time and model size. The int8 version is also better suited than the fp32 in production as it requires low computational power which is not readily available in farmlands.

The state-of-the-art (SOTA) object classification models aim to increase the model accuracy by building a denser neural network with precise calculations in terms of weights, biases, activation functions, matrix multiplications, etc., which leads to high accuracy. As a result, it increases the number of calculations which in turn is time-consuming even for inference (forward pass). These computationally heavy models require high-performance servers or workstations equipped with GPUs enabled for parallel computing. While on embedded devices, the computational resources present are very limited, an offline execution of the model takes place, i.e., the inference is performed on powerful servers instead of the target system. Another approach includes using lightweight deep neural network models like MobileNet [7] which uses depth-wise separable convolutions to reduce the model complexity, resulting in architectural changes which has limited capacity

¹GitHub: <https://github.com/parikshit14/QNN-for-weed>

for complex patterns. In this paper, we propose a quantized ResNet and quantized Inception weed classification model to address the limited computational resources on edge devices. Although training a quantized model leads to an accuracy drop from the SOTA models, the drop is very marginal, a mere 1-3% while being able to achieve more than 10 times gain in performance, in terms of inference time when compared to a non-quantized model.

The remainder of the paper is organized as follows. In Section II, we discuss the related research. In Section III, we present our approach, which includes the architecture, training methodology, and the associated issues with quantization. In Section IV, we present our findings on the DeepWeeds dataset and share model results based on relative accuracy, inference time, and complexity, followed by a conclusion in Section V.

II. RELATED WORK

A. Quantization

Although similar concepts had first appeared in the literature as early as 1898, the history of the theory and practice of quantization dates back to 1948. The early development of pulse code modulation systems led to the initial recognition of quantization in modulation and analog-to-digital conversion.

In neural networks, quantization is used to reduce the memory consumption of weight biases and activation by using low-precision datatypes like int8 instead of fp32. This reduces the model size by a factor of four. For context, the amount of multiplication and addition operations produced by operating a neural network on hardware can quickly reach many millions. High precision is typically not required during inference and could impede the use of AI in real-time or on devices with limited resources. Large computational gains and improved performance are obtained by combining lower-bit mathematical operations with quantized parameters for the intermediate calculations in a neural network.

Quantized neural networks improve power economy in addition to performance for two reasons, i.e., decreased memory access costs, and improved computation efficiency. By utilizing the lower-bit quantized data, less data must be moved both on and off-chip, reducing memory bandwidth and significantly reducing energy consumption. Mathematical operations with lower precision, such as an int8 multiplication as opposed to an fp32 multiplication, use less energy and have a higher compute efficiency, which results in less power being used. Sample comparison of power consumption in Figure 2.

B. CNN-Based Image classification

Several papers propose CNN-based models to improve the accuracy on the DeepWeeds dataset. The original paper on DeepWeeds [4] proposes a ResNet-50 and Inceptionv3 models trained using transfer learning with fine tuning on ImageNet weights, [9] presents 27 SOTA deep learning models through transfer learning evaluating accuracies and inference of each, [10] proposes a combination of predictions of a CNN and a secondary classifier for statistical features in weed images. [11] proposes a diffusion probabilistic model to generate

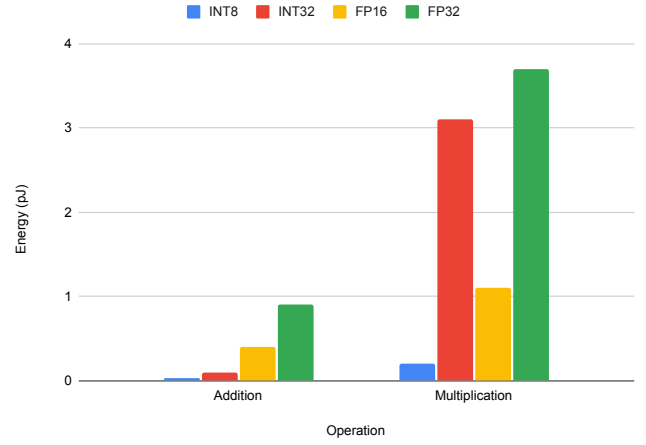


Fig. 2: Energy consumption on 45nm processor [8]

synthetic weed images of high quality to overcome the cost of data capture along with transfer learning, [12] proposes a combination of convolutional neural network and transformer structures for classification and feature extraction. However, none of the papers propose an improvement in terms of inference time and low computational power as present in actual agricultural environments.

C. CNN-Based Image classification for Embedded systems

The SOTA CNN models require high-end GPU not only for training but also for the purpose of inference, else it increases the inference time drastically. As a result, best-performing CNN models fail to outperform optimally on embedded systems on chips (SoC) [13]. For instance, a provider in the security camera market discovered that even after switching the YOLO back-end from GoogleNet to a more straightforward CNN like AlexNet, their embedded implementation only operates at a maximum frame rate of 5 frames per second on embedded GPUs [14].

In our proposed algorithm, a quantized convolutional neural network (Q-CNN), a form of compressed and accelerated CNN model, is implemented for the DeepWeeds dataset without significant accuracy loss. The proposed models use PyTorch [15], an open-source framework for model training and inference. It is heavily used in research and development tasks in industry and academia. The PyTorch quantization module currently provides support for x86 CPUs, ARM CPUs which are typically found in mobile/embedded devices, and early support for Nvidia GPU via TensorRT. On these systems, the suggested PyTorch-based model can benefit from similar architectural advantages and can be used for real-time object classification applications.

III. MODEL TRAINING

We used a combination of transfer learning and fine-tuning approaches to train the ResNet-50 and Inceptionv3 models. Transfer learning alone on a pre-quantized (int8) trained model with a custom-trained classifier head fails to give appropriate

accuracy. Also, transfer learning is not possible on a trained quantized model as it has no trainable parameters. To overcome these issues and use the advantages of a pre-trained model, we use the standard SOTA model (fp32) for transfer learning. It gives us the advantage of pretrained weights instead of random weight initialization which in turn would have required a lot of training. We replaced the SOTA 1,000 class classifier with a custom classifier head for 9 classes. Model parameters are kept unfrozen with a low learning rate of $1e-4$ for 30 epochs. The entire dataset is divided in a 60:20:20 ratio for training, validation, and testing, similar to that proposed in the original DeepWeeds paper. The models are trained on images with (224,224,3) shape. Most of the other parameters of the training are kept the same as those in the original paper. Adam optimizer [16] and cross-entropy loss function [17] are used in training.

A. Network Architecture

The overall architecture of the feature extractor remains the same for the most part as of a standard ImageNet model, except for the fact that in order to imitate the effects of int8, fake-quantization modules are inserted to model the effects of quantization via zero point shifting and scaling. To calculate zero point z and scale s .

Consider $x_q \in [\alpha_q, \beta_q]$ and $x \in [\alpha, \beta]$, where α and β are minimum and maximum values in their range.

$$s = \frac{\beta - \alpha}{\beta_q - \alpha_q} \quad (1)$$

$$z = \text{round} \left(\frac{\beta}{\alpha_q} - \frac{\alpha}{\beta_q} - \alpha \right)$$

where x_q = quantized value (int8), x = value (fp32), s = scale (fp32) and z = zero point (int8).

After the training process is completed, model conversion happens, i.e., the activations and weights are quantized to int8 from fp32, and the activations are fused into the preceding layer wherever possible. Since the transition from float to a lesser precision is a lossy process, we typically observe a large decline in accuracy. A quantization-aware training (QAT) [18] is used to assist in reducing this loss.

B. Training Methodology

In the proposed models, training is done using single-precision floating-point computation. There is no requirement to complete the training in fixed-point because it is done offline on a workstation. Fake-quant modules are inserted to replicate the effects of int8. This technique is termed as quantization-aware training.

$$v_{qc} = \text{clip}(v_q, \alpha_q, \beta_q)$$

$$\text{clip}(x, A, B) = \begin{cases} A & \text{if } x < A \\ x & \text{if } A \leq x \leq B \\ B & \text{if } x > B \end{cases} \quad (2)$$

QAT is frequently employed with training Q-CNNs and generates results with greater accuracy than static quantization.

Before the deep neural network is applied to the target, a conversion from a floating-point to a fixed-point representation must be made. This requires quantizing the deep neural network weights because the range of potential values for fixed-point and floating-point representations differs. Non-quantized values continue to be used in the backpropagation. The DNN can be pre-trained using a floating-point representation in order to initialize the parameters with reasonable values. This stabilizes the learning phase with the quantized version and yields better results. Although it is technically possible, there would be extra difficulties with the gradient calculation as discussed in section III-B6.

1) *Quantization Mapping*: The mathematical representation of mapping fp32 values to int8 values through quantization:

$$x_q = \text{round}(x/s + z) \quad (3)$$

and dequantization:

$$x = s(x_q - z) \quad (4)$$

2) *Weight Quantization*: CNN-based models are typically formed from convolutional layers and fully connected layers. These do require quantization-aware training for the parameters. The weights of the convolutional layer can be represented in a tensor as $(f_h, f_w, c_{in}, c_{out})$, and for a fully connected layer as (c_{in}, c_{out}) .² The output channel quantization bounds are calculated along each of them. There may be distinct and independent quantization boundaries for each output channel. This ensures smaller scaling factors and finer quantization ranges than the channel with a higher range in weight. Both Inceptionv3 and ResNet-50 have a large number of weight channels with notable magnitude fluctuations.

3) *Activation Quantization*: The activation functions are quantized by mapping their continuous output values to a discrete range of quantization levels. The numerical precision of activation is decreased during this process, enabling the use of low-precision hardware or memory-efficient deployment. The range of values is calculated similarly to that of convolutional and fully connected layers.

Standard ReLU [19]:

$$\text{ReLU}(x, 0, 0, 1) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \quad (5)$$

Quantized ReLU [20]:

$$\text{ReLU}_q(x_q, z_x, z_y, \frac{s_x}{s_y}) = \begin{cases} z_y & \text{if } x_q < z_x \\ z_y + \frac{s_x}{s_y}(x_q - z_x) & \text{if } x_q \geq z_x \end{cases} \quad (6)$$

4) *Layer Fusion*: For some combinations of neural network layers, such as Conv2D-ReLU and Conv2D-BatchNorm-ReLU, layer fusions are frequently used [21]. The idea of several layers is abstracted with layer fusion, which results in a single layer. When we do not use layer fusion, meaning we

² c_{in} = in channels, c_{out} = out channels, f_h = filter height, f_w = filter width

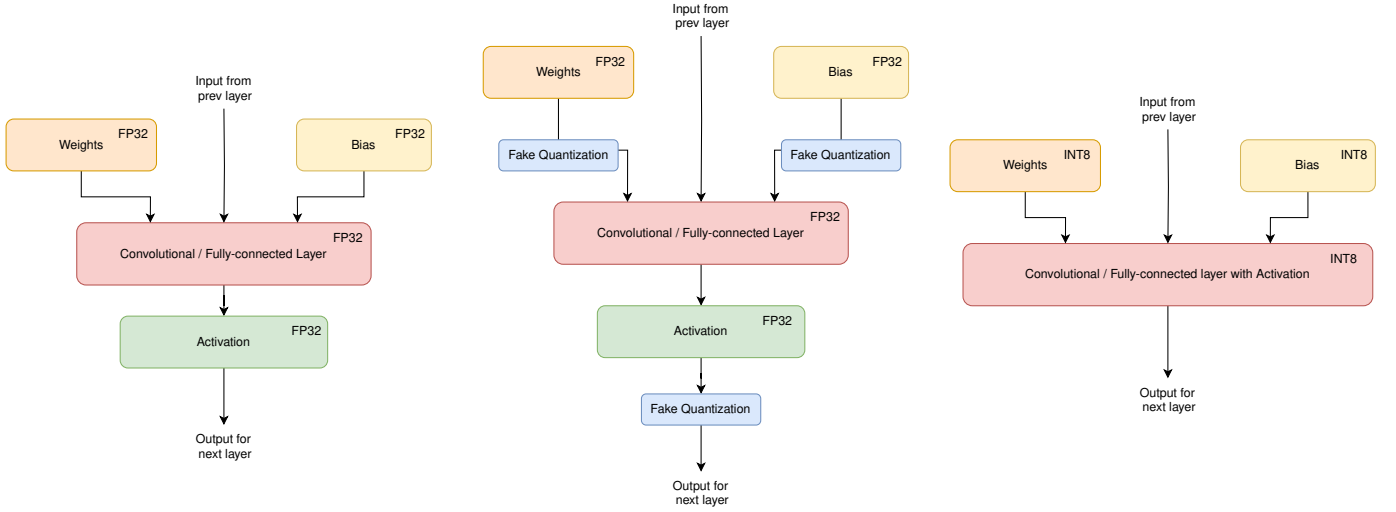


Fig. 3: Comparison between full-precision inference (left), model for QAT with simulated quantization (middle), and quantized model for integer-only inference (right)

have two separate layers a Conv2D layer followed by a ReLU activation layer we find that we must keep track of scales and zero points at multiple stages within the neural network. Firstly, we need scales and zero points for the inputs, denoted as x_0 , to the Conv2D layer. Next, we must monitor the scales and zero points for the outputs, x_1 , coming from the Conv2D layer, which also serves as inputs to the subsequent ReLU activation layer. Lastly, we need to consider the scales and zero points for the outputs, x_2 , produced by the ReLU activation layer. However, with the introduction of layer fusion, such as combining Conv2D and ReLU into a single Conv2D-ReLU layer, we can simplify this process. In this case, we only need to keep track of scales and zero points for the inputs x_0 to the Conv2D-ReLU layer and the outputs x_1 emerging from the same Conv2D-ReLU layer. This layer fusion technique streamlines the management of scales and zero points by integrating Conv2D and ReLU into a unified layer, reducing overall complexity. Figure 3 shows fake quantization modules and fused activations.

5) *Additional Layers*: The max-pooling layer only uses an element-wise maximum, and there is no need to quantize because the inputs from the preceding layer have already been quantized and the dynamic range cannot be increased. On the other hand, the element-wise addition layer needs quantization as adding two large values of parameters can exceed the dynamic range of the output. Therefore, the output scale factor is calculated using the same quantization method.

6) *Issues in QAT*: The inference accuracy from the quantized integer models is invariably worse than that from the floating point models due to information loss. The fact that the floating points are not perfectly recoverable after quantization and dequantization is the cause of this information loss.

$$x \neq f_{dq}(f_q(x, s_x, z_x), s_x, z_x) \quad (7)$$

where f_q is quantization function 3 and f_{dq} is dequantization function 4.

An error term Δ_x is introduced to consider the impact of such information loss during training:

$$x = f_{dq}(f_q(x, s_x, z_x), s_x, z_x) + \Delta_x \quad (8)$$

As a result, the model will have low inference accuracy loss.

Another issue with QAT is that the quantization and dequantization layers are not differentiable. The quantization/dequantization operation maps a continuous input to a discrete output, resulting in a step-like piecewise constant function. This discontinuity in the function makes it non-differentiable at the quantization points. However, there are strategies and procedures that can be used to approximate gradients and make it possible to train quantized networks, such as straight-through estimation [22] and Gumbel-softmax relaxation [23]. These strategies seek to differentiate the quantization layer so that gradient-based optimization is possible while still reaping the benefits of quantization. Figure 4 shows the close relation between the SOTA vs their quantized counterparts during training.

IV. RESULTS

The training of the proposed models was conducted on Nvidia Tesla P100 GPU and Intel Xeon 2.20 GHz CPU. The programming environment used for training was Python 3.9 and the deep learning framework was PyTorch 2.0.1. The inference was made on three different hardware, i.e., PC with Intel Core i5-8250U from the x86 family, and for the ARM architecture we are using a Mobile device with Tensor G2, and Raspberry Pi with Cortex-A72 for a complete sense of inference time on different architecture coverage. The hardware specifications are presented in Table III. The input image was run for 100 iterations (forward pass only) to get the average iterations per second shown in Table I. The inference time does not include the pre-processing of the input-image tensor nor the model loading time. In Table II, we compare

TABLE I: Model Comparison

Model Name	Accuracy %		Model Size (MB)	Inference Time (ms)		
	Top1	Top3		Core-i5	Tensor-G2	Cortex-A72
ResNet-50	95.95	99.43	94.45	112.71	173.62	1906.73
Quantized ResNet-50	94.77	99.49	23.72	42.61	90.99	218.42
Inceptionv3	95.09	99.63	87.59	90.17	120.92	838.79
Quantized Inceptionv3	94.57	99.40	22.04	34.46	71.15	187.56

Note: Accuracy values represent the Top1 and Top3 metrics. Model Size is in megabytes (MB), and Inference Time is measured in milliseconds (ms) on different hardware platforms.

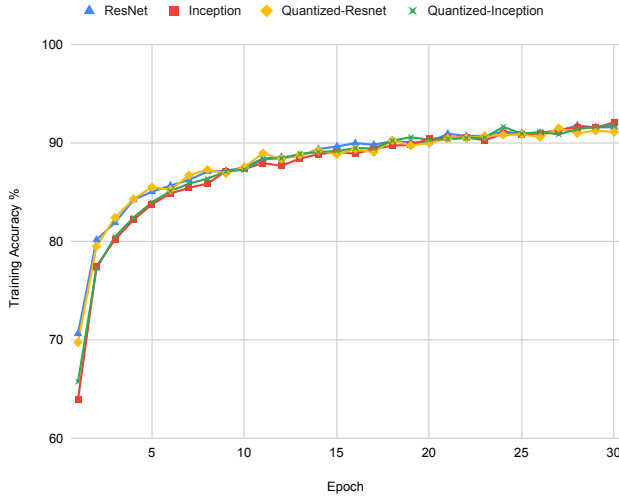


Fig. 4: Training Accuracy vs Epochs

the respective total operations, memory footprint represents the maximum RAM usage by any layer for a single image during inference.

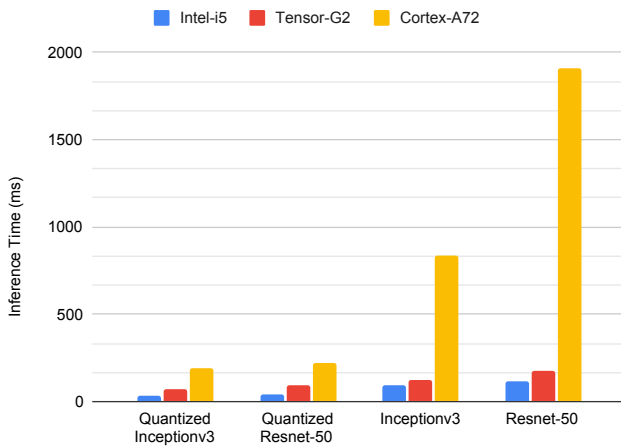


Fig. 5: Inference Time Comparison

Also, accuracy cannot be a single tool to evaluate the model results because of the presence of unbalanced classes. So, a

confusion matrix provides a much clearer picture of the trained model for each class. In Figure 6, we can see both SOTA models and their quantized counterparts perform equally well. Due to the uneven sample distribution, there are more instances where the eighth category or other categories are incorrectly identified. However, this inaccuracy rate can be tolerated because the negative weed category has approximately eight times as many images as the other categories. In Table IV we further represent our findings with model throughput on different hardware devices.

TABLE II: Performance Analysis

Model Name	GFLOPs/GOPs	Memory footprint (Mb)
ResNet-50	4.13	53.60
Quantized ResNet-50	4.13	6.70
Inceptionv3	2.85	34.64
Quantized Inceptionv3	2.85	4.58

Note: The complexity of the Non-Quantized models are represented in floating-point operations while quantized models use operations.

TABLE III: Hardware Specification

Hardware	Memory(GB)	Clock Speed(GHz)	CPU-Cores/Threads
Core-i5	8	3.40	4/8
Tensor-G2	12	2.85	8/8
Cortex-A72	8	1.80	4/4

Note: Different processors used are compared for clarity in performance capabilities.

V. CONCLUSION

In this paper, we proposed a Quantized ResNet-50 and Inceptionv3 model, a low complexity fully convolutional neural network for non-GPU enabled and embedded devices. These models perform comparably with respect to their SOTA counterparts in terms of accuracy. In terms of storage consumption, it takes almost $4\times$ less storage, achieving a $4\times$ speedup on CPU, $6\times$ speedup on Raspberry Pi, and $2\times$ speedup on a mobile processor. We believe that this strategy and the findings from our experimental study will make it easier to conduct future quantization research and develop industrial vision applications tailored to agriculture for devices with limited resources, enabling them to be more AI-enabled while using fewer resources.

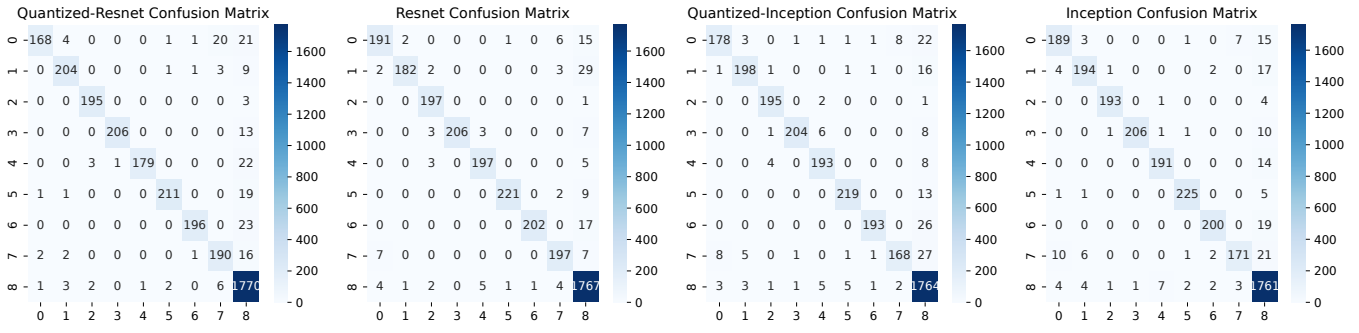


Fig. 6: Confusion matrix of respective models, X-axis represents the actual label, and Y-axis represents the predicted label. The respective labels are in the following sequence: 0: 'Chinese apple', 1: 'Lantana', 2: 'Parkinsonia', 3: 'Parthenium', 4: 'Prickly acacia', 5: 'Rubber vine', 6: 'Siam weed', 7: 'Snake weed', 8: 'Negative'.

TABLE IV: Throughput Comparison

Hardware	Model Throughput GFLOPs/GOPs per sec			
	ResNet-50	Quantized ResNet-50	Inceptionv3	Quantized Inceptionv3
Tensor-G2	23.79	45.40	23.60	40.12
Cortex-A72	2.16	18.91	3.40	15.21
Core-i5	36.65	96.96	31.65	82.83

Note: Comparing throughput of models with their quantized counterparts on the selected hardware.

REFERENCES

- [1] Y. Gharde, P. Singh, R. Dubey, and P. Gupta, "Assessment of yield and economic losses in agriculture due to weeds in india," *Crop Protection*, vol. 107, pp. 12–18, 2018.
- [2] J. G. Varshney and M. Babu, "Future scenario of weed management in india," *Indian Journal of Weed Science*, vol. 40, no. 1, pp. 1–9, 2008.
- [3] P. Wang, Y. Tang, F. Luo, L. Wang, C. Li, Q. Niu, and H. Li, "Weed25: A deep learning dataset for weed identification," *Frontiers in Plant Science*, vol. 13, p. 1053329, 2022.
- [4] A. Olsen, D. A. Konovalov, B. Philippa, P. Ridd, J. C. Wood, J. Johns, W. Banks, B. Girgenti, O. Kenny, J. Whinney, *et al.*, "Deepweeds: A multiclass weed species image dataset for deep learning," *Scientific reports*, vol. 9, no. 1, p. 2058, 2019.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.
- [6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- [7] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.
- [8] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 10–14, 2014.
- [9] D. Chen, Y. Lu, Z. Li, and S. Young, "Performance evaluation of deep transfer learning on multi-class identification of common weed species in cotton production systems," *Computers and Electronics in Agriculture*, vol. 198, p. 107091, 2022.
- [10] J. Huertas-Tato, A. Martín, J. Fierrez, and D. Camacho, "Fusing cnns and statistical indicators to improve image classification," *Information Fusion*, vol. 79, pp. 174–187, 2022.
- [11] D. Chen, X. Qi, Y. Zheng, Y. Lu, Y. Huang, and Z. Li, "Deep data augmentation for weed recognition enhancement: A diffusion probabilistic model and transfer learning based approach," in *2023 ASABE Annual International Meeting*, p. 1, American Society of Agricultural and Biological Engineers, 2023.
- [12] J. Zhang, "Weed recognition method based on hybrid cnn-transformer model," *Frontiers in Computing and Intelligent Systems*, vol. 4, p. 72–77, Jun. 2023.
- [13] J. Qiu, J. Wang, S. Yao, K. Guo, B. Li, E. Zhou, J. Yu, T. Tang, N. Xu, S. Song, Y. Wang, and H. Yang, "Going deeper with embedded fpga platform for convolutional neural network," (New York, NY, USA), Association for Computing Machinery, 2016.
- [14] S. Tripathi, G. Dane, B. Kang, V. Bhaskaran, and T. Nguyen, "Lcdet: Low-complexity fully-convolutional neural networks for object detection in embedded systems," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 94–103, 2017.
- [15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
- [16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [17] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.
- [18] E. Park, S. Yoo, and P. Vajda, "Value-aware quantization for training and inference of neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 580–595, 2018.
- [19] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.
- [20] L. Mao, "Quantization for Neural Networks — leimao.github.io," <https://leimao.github.io/article/Neural-Networks-Quantization/>. [Accessed 08-09-2023].
- [21] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. Van Baalen, and T. Blankevoort, "A white paper on neural network quantization," *arXiv preprint arXiv:2106.08295*, 2021.
- [22] P. Yin, J. Lyu, S. Zhang, S. Osher, Y. Qi, and J. Xin, "Understanding straight-through estimator in training activation quantized neural nets," *arXiv preprint arXiv:1903.05662*, 2019.
- [23] A. Potapczynski, G. Loaiza-Ganem, and J. P. Cunningham, "Invertible gaussian reparameterization: Revisiting the gumbel-softmax," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 12311–12321, Curran Associates, Inc., 2020.