Conference Paper Title

Oğulcan Aşık, Ji Wu & Kazuteru Namba
Graduate School of Science and Engineering
Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522 Japan
E-mail: ogasi100@uni-duesseldorf.de, cafa0748@chiba-u.jp & namba@ieee.org

***Abstract*—abstract**

***Index Terms*—SRAM, quantization, ECC, CNN**

I. Introduction

Recently computer-based classification has come very far. Convolutional neural networks (CNN) can perform various tasks similar to the way humans can. However, significant calculations and thus power requirements are required to achieve such goals. Previous studies proposed quantization [1] and low-voltage computation [2]. But lowering voltage comes with not insignificant bit error rates (BER) [3] that in turn lower the accuracy of a model in a significant way.

Quantization has been shown to be an effective method to use while mostly preserving good accuracy [4] [2]. In this paper we have expanded on the ideas put forth in [2] by exploring higher BERs in terms of classification accuracy coupled with a single bit error correcting code (ECC) using a computer simulation. In addition to using two different operating voltages, and hence a protected and unprotected bit range respectively, we have also shown the impact of using the ECC without such a split.

II. Prerequisites

A. ECC: Single-Bit Correcting Hamming Code

For the ECC the single-bit correcting Hamming Code as been chosen. The reason for this is its ease of use and minimal impact on performance and bit bandwidth usage while allowing to correct up to one error. In practice this means 4 data bits are protected by 3 parity bits, 8 bits are covered by 4 and 16 bits are covered by 5 parity bits.

B. CNN

The tested neural networks were trained on the MNIST dataset [5]. The CNNs were built with one input layer, three convolutionall layers, two pooling layers and two fully connected layers. SoftMax was used as the activation function in the hidden layer and the number of steps was 2000 using a batch size of 500. The model was trained for 32-bit weights.

C. Quantization in the CNN

The model has been quantized post-training to 4-bits, 8-bits and 16-bits using asymmetric uniform quantization. The following formula was used:

$$\text{Q} = \text{offset} + \text{round}(\frac{x}{\text{scale}})$$

This method has been shown to be lower energy costs with a low loss of accuracy [3].

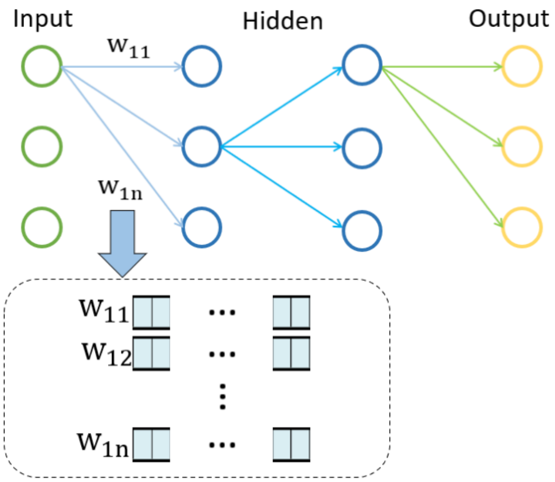

Fig. 1. Quantization of CNN weights, Image from [2]

III. Experimental Conditions

Expanding on [2], which proposed certain bits with a separate higher operating voltage while lowering the voltage on the other bits making them less protected, we attached ECC in order to lower the BER even further. In addition to considering two different BERs we also tested applying one BER to all bits instead. This was done in a simulation using Python and pytorch only.

The simulations with two different BERs tested with increasing size of protected range, starting with bit 0, only then 0-1, 0-2 etc. For each such combination the model was tested 10 times and the accuracy was averaged. The BERs in the protected range were set to the constant value of 10^{-4}. The simulations using only one BER do not have a protected range.

IV. Experimentation Results

Below we present the results of the simulations. We start with the idea of splitting the bits in protected and unprotected bits proposed by [2]. Afterwards, the results without such a split are presented. All simulations are compared to the baseline that does not use ECC to correct errors.

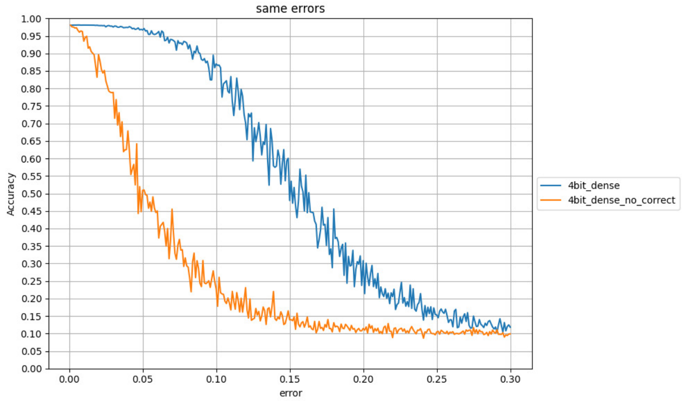

Fig. 2. 4-bit Quantization. The BER plotted against the overall prediction accuracy of the model.

A. *Different BER Simulations*

B. *Same BER Simulations*

1) 4-Bit Quantization: The results of the 4-bit quantization clearly show an improvement of accuracy at higher BERs allowing a reliable accuracy up to an error rate of even 0.05 after which we can see a sharp decline. This setup allows the highest BER in comparison to the other simulations of this category.

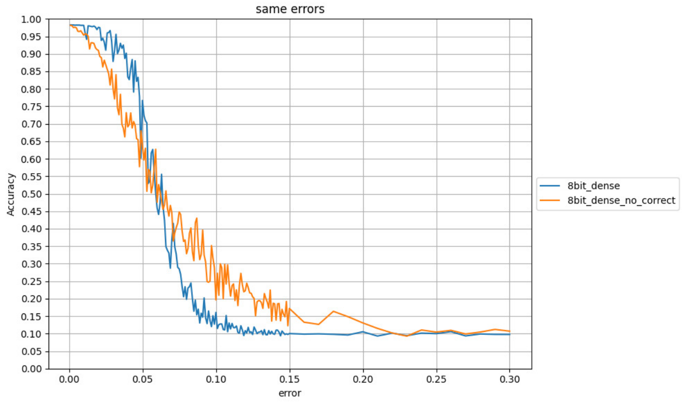

Fig. 3. 8-bit Quantization. The BER plotted against the overall prediction accuracy of the model.

2) 8-Bit Quantization:

REFERENCES

[1] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," 2017. [Online]. Available: https://arxiv.org/abs/1712.05877

[2] J. Wu and K. Namba, "Sram-based efficiency memory model for quantized convolutional neural networks," in *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 2023, pp. 499–500.

[3] L. Yang and B. Murmann, "Sram voltage scaling for energy-efficient convolutional neural networks," in *2017 18th International Symposium on Quality Electronic Design (ISQED)*, 2017, pp. 7–12.

[4] B. Rokh, A. Azarpeyvand, and A. Khanteymoori, "A comprehensive survey on model quantization for deep neural networks in image classification," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 6, Nov. 2023. [Online]. Available: https://doi.org/10.1145/3623402

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.