Work-in-Progress: Enabling NVM-Based Deep Learning Acceleration Using Nonuniform Data Quantization

Hao Yan
Department of Electrical and Computer Engineering
University of Texas at San Antonio
San Antonio, TX 78249
hao.yan@utsa.edu

Ethan C. Ahn
Department of Electrical and Computer Engineering
University of Texas at San Antonio
San Antonio, TX 78249
chiyui.ahn@utsa.edu

Lide Duan
Department of Electrical and Computer Engineering
University of Texas at San Antonio
San Antonio, TX 78249
lide.duan@utsa.edu

CCS CONCEPTS

• **Computer systems organization** → **Neural networks**; **Processors and memory architectures**; • **Hardware** → **Memory and dense storage**;

KEYWORDS

Non-volatile memory (NVM), deep learning acceleration, nonuniform data quantization

ACM Reference format:
Hao Yan, Ethan C. Ahn, and Lide Duan. 2017. Work-in-Progress: Enabling NVM-Based Deep Learning Acceleration Using Nonuniform Data Quantization. In *Proceedings of CASES '17 Companion, Seoul, Republic of Korea, October 15–20, 2017,* 2 pages.
DOI: 10.1145/3125501.3125516

1 ABSTRACT (INTRODUCTION)

Apart from employing a co-processor (e.g., GPU) for neural network (NN) computation, utilizing the unique characteristics of non-volatile memories (NVM), including RRAM, phase change memory (PCM), and STT-MRAM, to accelerate NN algorithms has been extensively studied. In such approaches, input data and synaptic weights are represented using word line voltages and cell resistance, with the resulting bit line current indicating the calculation result. However, the limited number of resistance levels in a NVM cell largely reduces the algorithm data precision, thus significantly lowering the model inference accuracy. Motivated by the observation that the conventional, uniformly generated data quantization points are not equally important to the model, **we propose a nonuniform data quantization scheme** to better represent the model in NVM cells and minimize the inference accuracy loss. Our experimental results show that the proposed scheme can achieve highly accurate deep learning model inference using as low as only 4 bits for synaptic weight representation. This effectively enables a NVM with few cell resistance levels (e.g., STT-MRAM) to perform NN calculation, and also results in additional benefits in performance, energy, and memory storage.

CASES '17 Companion, Seoul, Republic of Korea

DOI: 10.1145/3125501.3125516

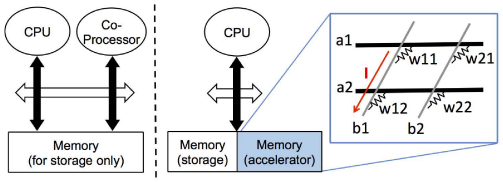

Figure 1: The co-processor approach (left) vs. the NVM approach (right) in accelerating NN algorithms.

2 BACKGROUND

Accelerating NN algorithms in hardware is typically performed via a co-processor [1], such as a GPU, FPGA, or ASIC device, or via a part of the NVM [2]. Figure 1 demonstrates these two approaches. NVM-based NN acceleration relies on the unique crossbar structure of the memory chip and the multiple resistance levels that can be configured in a NVM cell. Figure 1 (right part) performs a typical NN operation $b_j = \sum_{i=1}^{2} a_i \cdot w_{ji}$, where j ranges from 1 to 2. The input data a_i is applied as analog input voltages on the horizontal word lines; the synaptic weights w_{ji} are programmed into the NVM cell conductance (i.e., 1 / cell resistance). The resulting current I flowing out of the vertical bit line indicates the calculation result.

3 MOTIVATION

The number of resistance levels in a NVM cell determines the bit width (bw) of the fixed-point number it represents. Since a NVM cell only has a limited number of resistance levels, programming a synaptic weight into the cell conductance significantly reduces the weight precision. For example, if only 8 resistance levels exist in the NVM cell, the original 32-bit fixed-point weight values have to be reduced to only 3 bits. This is shown in Figure 2, where g(x) is the distribution of all the synaptic weights. A weight value needs to be quantized to one of the 8 quantization points that correspond to the 8 cell resistance levels. The static quantization [1] uses a fixed fractional length (fl) for the whole model independent of its weight value range, whereas the dynamic quantization [2, 3] allows tuning the fl to achieve the precision with lowest errors. The difference can be seen in the figure as the weight value range (on the x-axis) being static or dynamic. Nevertheless, **both the static and dynamic schemes have uniform quantization points.**

These conventional quantization schemes are not suitable to be used for highly complex deep learning models. Figure 3 shows how the inference accuracy of VGG19 (a complicated CNN model) varies with the fixed-point number bit width. As can be seen, the

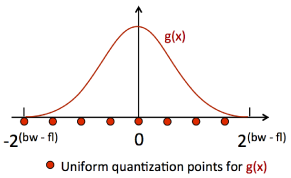

Figure 2: An example of uniform quantization (static and dynamic).

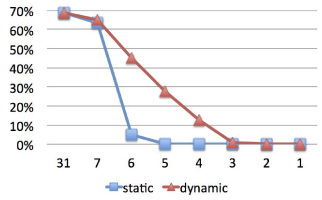

Figure 3: The inference accuracy of VGG19 varies with the bit width using uniform quantization.

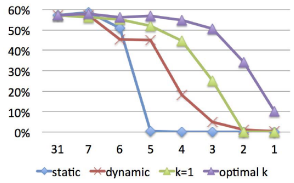

Figure 6: The inference accuracy of AlexNet varies with the bit width.

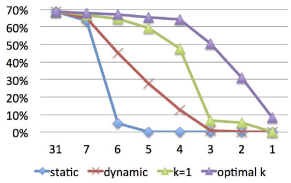

Figure 7: The inference accuracy of VGG19 varies with the bit width.

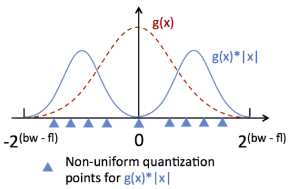

Figure 4: An example of nonuniform quantization.

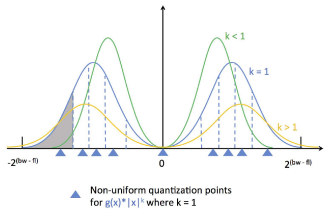

Figure 5: An illustration of the proposed nonuniform quantization scheme.

model accuracy starts to decrease significantly when the bit width goes below 7. A NVM cell has a typical precision of 2 to 7 bits [2]. The fundamental reason of the seen low accuracy is because **the quantization points are not equally important**. When it is too close to zero, it has minimal impact on the model inference despite the large number of weights being quantized to it; when it is close to the range boundary, it also shows limited impact due to the extremely low weight quantity. In other words, the most important quantization points are not uniformly distributed.As depicted in Figure 4, a nonuniform quantization scheme is needed to better represent the model and minimize the accuracy loss.

4 DESIGN

In this work, we propose a nonuniform data quantization scheme to achieve better deep learning model accuracy using fewer weight bits. Specifically, we construct an importance function $g(x) \cdot |x|$ to approximate the importance of different quantization values to the model accuracy. This is shown as the blue curve in Figure 4 and Figure 5. This function takes into account both the weight value and amount, indicating that the most important quantization points are around its peaks. To quantify these points, we evenly partition the area between the function curve and the x-axis into $2^{bw+1} + 1$ regions. For a bit width of 2 (not counting the sign bit), the area is partitioned into 9 regions. For the center region around 0, its quantization point is forced to be 0; for the other regions, the quantization point is the value that divides the region into halves (analogous to the center of mass).

Second, we generalize the importance function to be $g(x) \cdot |x|^k$ to prioritize the weight value differently by adjusting k. We test different values of k and pick the one that gives the highest model accuracy. Figure 5 shows the nonuniform quantization points for k=1 and two other function curves for k<1 and k>1.

5 EXPERIMENTS

Due to space, we show the model accuracy variations with reduced fixed-point number bit width for two workloads: AlexNet (Figure 6) and VGG19 (Figure 7). As can be seen, at a bit width of 4 or larger, our proposed scheme ("optimal k") demonstrates negligible model accuracy loss (< 2-3%) compared to the original model with 32-bit numbers. Therefore, the benefits of our scheme are: (1). a memory storage compression rate of 2 than conventional static/dynamic schemes with almost no accuracy loss; (2). reduced computation due to a quantization point at 0 and fewer bits to represent a number; and (3). more importantly, this enables STT-MRAM, which is found to have fewer cell resistance levels than ReRAM and PCM, to have the capability of deep learning acceleration.

6 CONCLUSIONS AND FUTURE WORK

This paper proposes a nonuniform data quantization scheme to identify the quantization points most important to deep learning model inference accuracy. Our future work include: (1). more quantitatively evaluating the benefits of our scheme in performance, energy, memory storage, etc., using a wider variety of workloads; (2). implementing the nonuniform quantization on the input data; and (3). developing an efficient NVM cell resistance reconfiguration scheme to accommodate different models.

ACKNOWLEDGMENTS

The work is supported by the National Science Foundation under Grant No. CCF-1566158. The authors would also like to thank the anonymous reviewers for their invaluable comments and helpful suggestions.

REFERENCES

[1] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. 2014. DianNao: A Small-footprint High-throughput Accelerator for Ubiquitous Machine-learning. In *ASPLOS*.

[2] Ping Chi, Shuangchen Li, Cong Xu, Tao Zhang, Jishen Zhao, Yongpan Liu, Yu Wang, and Yuan Xie. 2016. PRIME: A Novel Processing-in-memory Architecture for Neural Network Computation in ReRAM-based Main Memory. In *ISCA*.

[3] Jiantao Qiu, Jie Wang, Song Yao, Kaiyuan Guo, Boxun Li, Erjin Zhou, Jincheng Yu, Tianqi Tang, Ningyi Xu, Sen Song, Yu Wang, and Huazhong Yang. 2016. Going Deeper with Embedded FPGA Platform for Convolutional Neural Network. In *International Symposium on Field-Programmable Gate Arrays (FPGA)*.