Effect of Device Variation on Mapping Binary Neural Network to Memristor Crossbar Array

Wooseok Yi, Yulhwa Kim, and Jae-Joon Kim
Creative IT Engineering, *Pohang University of Science and Technology (POSTECH)*, Pohang, Republic of Korea
Email: wooseok.yi@postech.ac.kr, yulhwa.kim@postech.ac.kr, jaejoon@postech.ac.kr

***Abstract*—In memristor crossbar array (MCA)-based neural network hardware, it is generally assumed that entire word-lines (WLs) are simultaneously enabled for parallel matrix-vector multiplication (MxV) operation. However, the error probability of MxV in a memristor crossbar array (MCA) increases as the resistance ratio (R-ratio) of a memristor decreases and the resistance variation and the number of simultaneously activated WLs increase. In this paper, we analyze the effect of R-ratio and variation of memristor devices on read sense margin and inference accuracy of MCA-based Binary Neural Network (BNN) hardware. We first show that only a limited number of WLs should be enabled to ensure correct MxV output when the R-ratio is small. On the other hand, we also show that, if the resistance variation becomes higher than a certain level, simultaneous activation of large number of WLs produces the higher accuracy even when R-ratio is small. Based on the analysis, we propose the Accuracy Estimation (AE) factor to find the optimal number of word lines that are simultaneously activated.**

I. Introduction

As Deep Neural Networks (DNNs) show remarkable performance in various artificial intelligence fields, research on acceleration of DNN computation is getting more attention. Because the intensive data movement between on-chip processors and off-chip memory is one of the important bottlenecks in the existing DNN acceleration, Binary Neural Network (BNN) is receiving attention for its extremely small amount of memory requirement.

In addition, interest in acceleration of DNN using Memristor Crossbar Array (MCA) is also increasing. Because MCA can realize theoretically the smallest area ($4F^2$), it is suitable for highly integrated DNN acceleration hardware. One of the major advantage of MCA is the 1-step Vector-Matrix Multiply-and-Accumulation (1-step VM MAC) operation.

Thus far, memristors such as Magnetic RAM (MRAM), Resistive RAM (RRAM), and Phase Change RAM (PRAM) have been actively studied for neural network hardware implementation. Among them, the MRAM using Magnetic Tunnel Junction (MTJ) has advantages such as fast read/write speed, high endurance and low writing energy consumption compared to PRAM and RRAM [1], [2], but it has relatively low R-ratio (R_{HRS}/R_{LRS}) between high resistance state (HRS) and low resistance state (LRS). Generally, the R-ratio of PRAM and RRAM is 5∼1000, whereas MRAM has the R-ratio of 1∼3 [3]. This property greatly restricts the effectiveness of MRAM MCA for DNN acceleration.

Along with the small on/off resistance ratio, MTJ resistance variation also increases the possibility that the weighted sum converted from the analog BL current deviates from the target digital value. Therefore, in the case of MRAM crossbar array with small R-ratio, if all WLs are activated and entire rows are read at the same time, the loss of inference accuracy becomes large.

In order to solve this problem, it is necessary to give up some advantages of using MCA. To be more specific, for more accurate results, only a limited number of WLs should be activated at the same time (N_{OTR}). Then the BL current is converted into digital data by a Sense Amplifier (SA) or Analog-to-Digital Converter (ADC) and the partial sum data needs to be stored in the accumulator. Similar to conventional MRAMs which uses WL decoders, reading one WL at a time will yield the highest accuracy, but it produces the lowest throughput. On the other hand, if multiple WLs are read at the same time, the accuracy may decrease while the throughput increases.

In this paper, considering above characteristics of the MRAM crossbar array, we investigate how much the accuracy of BNN inference varies depending on device variation. We will derive an accuracy estimation factor as a function of the relative standard deviation (RSD), R-ratio, and N_{OTR} to predict how the inference accuracy will change when mapping BNN to MCA. Using the estimation factor, we can decide the optimal N_{OTR} to achieve the target accuracy drop at the given device R-ratio and RSD conditions.

II. Preliminary

A. Memristor Crossbar Array (MCA) with MTJ Cells

Magnetic Tunnel Junction (MTJ) is a kind of memristors which consists of two ferromagnetic layers and an insulating layer. The MTJ typically stores 1-bit binary data [4]. When the magnetization directions of two ferromagnetic layers are same (Parallel), the MTJ is in Low Resistance State (LRS), and when the magnetization directions are opposite (Anti-Parallel), the MTJ is in High Resistance State (HRS). MTJ-based MRAM switching by Spin-Transfer Torque (STT) or Spin-Orbit Torque (SOT) method is known to have faster write time, higher endurance and smaller write energy compared to PRAM and RRAM. On the other hand, the disadvantage is that MTJ-based cell size is larger and its R-ratio is smaller [1], [2]. As a result, the MTJ-cell based memory is getting attention for embedded system which requires fast read/write time for real-time applications [5], [6].

978-3-9819263-2-3/DATE19/©2019 EDAA

B. MCA Binary Neural Network (BNN) Hardware

In BNN, binary precision value (-1, 1) is used for both weights and activation values at the inference phase [7] [8]. BNN has merits in terms of hardware resource because the n-bit multiply operation in multi-bit neural network is replaced by the binary XNOR operation, which requires 58 times fewer operations [7]. In addition, the accumulation of multiplication results in multi-bit neural network is replaced by the simpler bit-counting operation in BNN. Therefore, the hardware resource can be further reduced in BNN. Mapping BNN on memory crossbar array structure has even greater potential because it can eliminate the use of multi-bit ADC and DAC which incur significant area and power overhead by using 1-bit word line driver and sense-amp for input and output.

III. Proposed AE factor & Design Flow

In this section, we analyze the device variations using Monte Carlo (MC) simulation in a 65nm CMOS technology and a PMAMTJ model from [9]. For both learning and inference, we experimented with binarized MLP and CNN network structure in [8] using Torch.

A. MCA-BNN with device variation & small R-ratio

The I_{BL} value tends to follow a normal distribution with non-negligible standard deviation (σ) value because of various factors. Here, the 3-σ value of the read sense margin (RSM), which distinguishes the difference between two levels, becomes smaller as the R-ratio of the device becomes smaller and the RSD becomes larger. Also, even with the same RSD, the sense margin becomes smaller as the number of WLs read simultaneously (N_{OTR}) increases (Fig. 1). For example, if 6 or more cells in the LRS state are read under the condition of R-ratio=2.5, RSD=5%, and N_{OTR}=8, 3-σ RSM becomes negative value because the distribution with adjacent levels is overlapped. Assuming that the RSD for R_{LRS} and R_{HRS} are similar, I_H and I_L will have similar RSD values, so the variation of I_H will become larger.

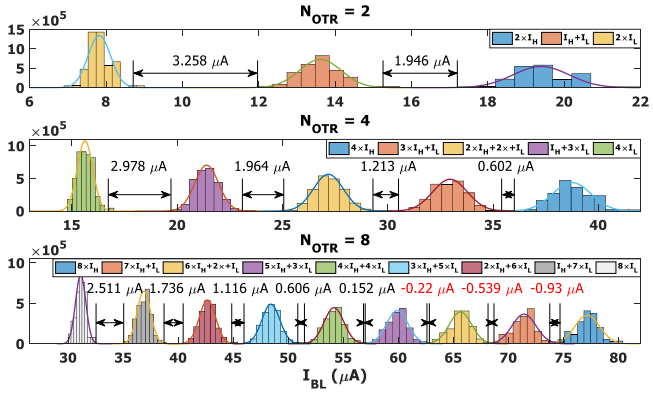

Fig. 1. Current level (I_{BL}) distribution according to the number of WL (N_{OTR}) (RSD=5%, R-ratio=2.5).

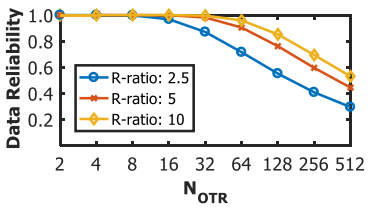

Fig. 2. The probability that the final read value will be equal to the ideal original value as the N_{OTR} is changed ($RSD = 5\%$, $N_{LRS} = N_{HRS}$).

Compared to RRAM and PRAM, MRAM is more likely to be affected by the variations because the R-ratio is much smaller (as small as ~ 2.5) in MRAM. In previous RRAM or PRAM based neural network hardware designs, it is generally assumed that all WLs are turned-on simultaneously for vector-matrix multiplication. However, if all WLs are turn-on simultaneously for an MCA which consists of MRAM cells with small R-ratio, the digitized weighted sum value can be significantly different from the correct value (Fig. 2). Therefore, unlike the RRAM/PRAM-BNN hardware, part of WLs, not all the WLs, may need to be turned-on simultaneously to produce the correct output.

B. Proposed Accuracy Estimation (AE) factors

As can be seen in Fig. 1, if the resistance is affected by the same RSD at each level, the variation is increased as the average value of I_{BL} increases. Therefore, it is most reasonable to consider the largest value case ($\mu = N_{OTR} \times I_H$) in judging how much each level of the I_{BL} value is overlapped with adjacent levels when reading N_{OTR} WLs simultaneously. As can be seen in Fig. 3, the probability that μ_N is misread as μ_{N-1} is equal to the area of pdf of the I_N from $-\infty$ to $(\mu_N - d_\mu/2)$, which can also be indirectly estimated by $\sigma_N/(d_\mu/2)$. If the R-ratio is γ and $(RSD) = \sigma_H/I_H$, then $I_H = \gamma I_L$, $d_\mu = \mu_N - \mu_{N-1} = I_H - I_L = (\gamma - 1)I_L$ and $\sigma_N = \sigma_H \times \sqrt{N_{OTR}} = (RSD)I_H \times \sqrt{N_{OTR}}$. Then,

$$\frac{\sigma_N}{d_\mu/2} = \frac{(RSD)I_H \times 2\sqrt{N_{OTR}}}{(\gamma - 1)I_L} = \frac{2(RSD)\gamma\sqrt{N_{OTR}}}{(\gamma - 1)}. \quad (1)$$

We name the $\sigma_N/(d_\mu/2)$ as the Accuracy Estimation (AE) factor. From our experiments, if the value of the AE factor is less than 0.328, the read result guarantees 3-σ accuracy and if the AE factor is less than 0.192, the 6-σ accuracy is guaranteed.

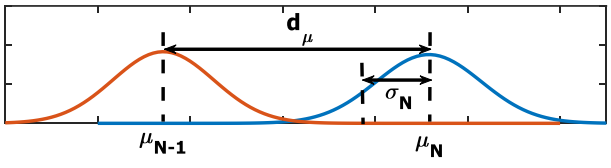

Fig. 3. The Probability Density Function (PDF) for current levels when reading N memristors simultaneously.

In the MCA-BNN hardware, one or more partial sums must be added depending on N_{OTR} in order to obtain one neuron activation value. Since small number of N_{OTR} requires a large number of partial sums to obtain the final neuron activation value, more errors are generated and integrated in the digitizing process of the partial sum. Therefore, in order to obtain a more accurate estimation factor, the effect of N_{OTR} in the AE factor should be decreased. Therefore, we define the adjusted-AE (adj-AE) at a design point as following equation

$$AE_{adj,dp} = \frac{2(RSD)\gamma(N_{OTR})^{0.5-k}}{(\gamma-1)} \quad (2)$$

where k is a value related to the structure of the network, which can vary depending on the depth of the network and the size of layers. Conversely, we can obtain the target N_{OTR} (N_{dp}) through a given $AE_{adj,dp}$, RSD, and γ as following equation

$$N_{dp} = \left(\frac{(AE_{adj,dp})(\gamma-1)}{2\gamma(RSD)}\right)^{1/(0.5-k)}. \quad (3)$$

Using Eq. (3) with given adj-AE factors and k values, we can find the maximum value of N_{OTR} that minimizes the accuracy drop for a MCA-BNN with certain RSD and R-ratio (γ) values (N_{bdp}). We call the design point which satisfies such conditions as the Best Design Point (BDP). We can also find the value of N_{OTR} for the Secondary Design Point (SDP) using the adj-AE factor value of the worst design point ($AE_{adj,wdp}$), where the worst design point is derived from the Lowest Accuracy Point (LAP) in a MCA-BNN with given RSD and γ. The $AE_{adj,wdp}$ helps to find the reasonable N_{OTR} when it is not possible to find the BDP at the given device conditions.

To find optimal N_{OTR} using the $AE_{adj,bdp}$ and $AE_{adj,wdp}$, a designer needs to follow two steps:

1) If the N_{bdp} value of Eq. (3) for the given R-ratio, RSD, k and $AE_{adj,bdp}$ is larger than 1, the optimal N_{OTR} becomes the largest possible integer value less than N_{bdp}. In this case, the device is in the Low RSD region.

2) If the N_{bdp} value is less than 1, it means that BDP does not exist. In this case, calculate the value of N_{wdp} as in the Eq. (3) to use it for finding a design point which has the smallest accuracy loss for given process conditions. If the N_{wdp} value is larger than 4, then the device is in the Middle RSD region, and the optimal $N_{OTR} = 1$ (=SDP). On the other hand, if the N_{wdp} value is smaller than 4, the device is in the High RSD region, and the optimal $N_{OTR} = 512$ (or maximum value in given condition).

We can see that there is a correlation between adj-AE factor value and the inference accuracy of the BNN, which will be discussed in detail in the section IV.

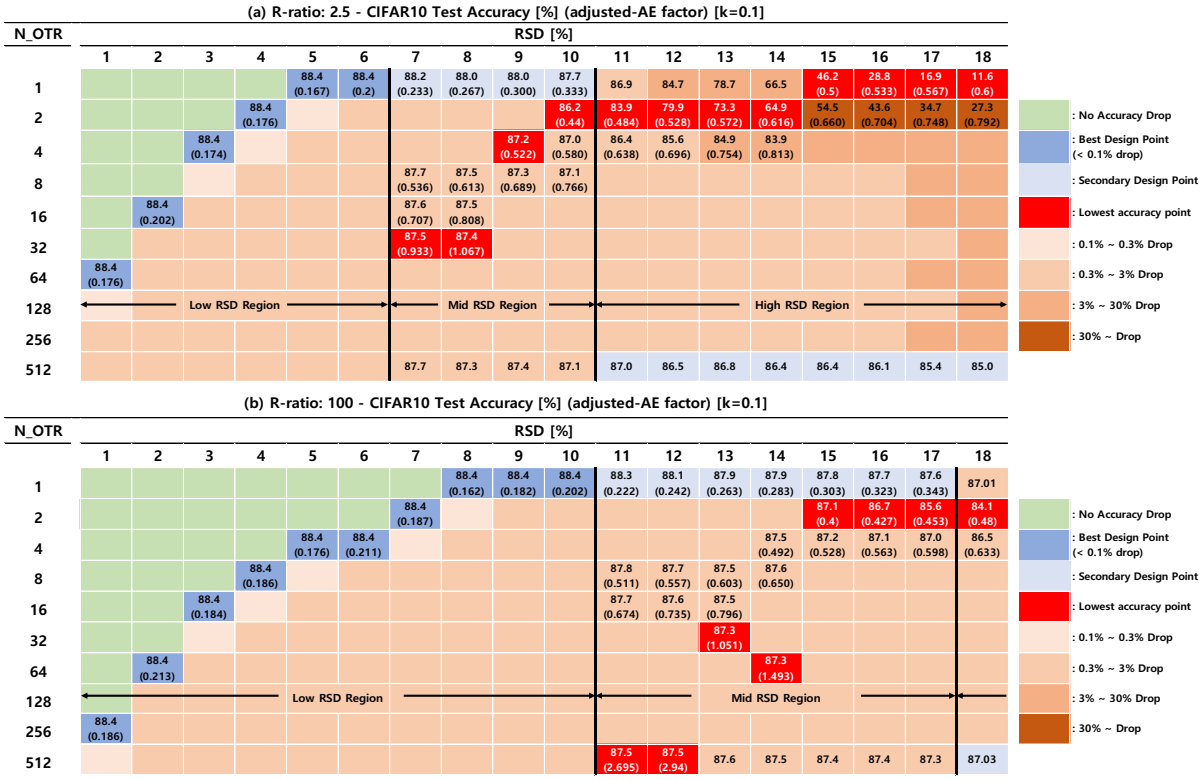

Fig. 4. Adj-AE factor table for CIFAR-10 with (a) R-ratio=2.5 and (b) R-ratio=100

TABLE I
K-VALUES AND THE DESIGN POINT VALUE OF ADJUSTED-AE FACTOR

Dataset	k-value	$AE_{adj,bdp}$	$AE_{adj,wdp}$
CIFAR-10	0.1	0.215	0.61
SVHN	0.26	0.32	0.5
MNIST	0.19	0.42	0.627

IV. RESULTS

We swept RSD and N_{OTR} for all cases of R-ratio=2.5 (MRAM), 5, 25, and 100 (PRAM, RRAM) to examine the variations in inference accuracy of CIFAR-10, SVHN (BCNN [8]) and MNIST (BMLP). For each RSD value, the point with the highest N_{OTR} with negligible drop of accuracy ($\leq 0.1\%$ for CIFAR-10, $\leq 0.03\%$ for MNIST and SVHN) was set as the Best Design Point (BDP, The blue entries in Fig. 4) and the point where the greatest accuracy drop occurred was set as the Lowest Accuracy Point (LAP, The red entries in Fig. 4). Based on the LAP values, we can choose the $AE_{adj,wdp}$ to define the boundary between Middle-RSD region and High-RSD region.

A. CIFAR-10 Dataset

The summary of the mean inference accuracy values are shown in Fig. 4. The cases with R-ratios = 2.5, 5, 25, and 100 were tested and only the cases with R-ratio=2.5 and 100 are shown in Fig. 4 due to space limit. The k-value and the adj-AE factors at BDP and WDP are shown in Table. I.

The adj-AE factor table (Fig. 4) consists of three parts; Low, Middle, and High RSD region. First, the Low RSD region is where the BNN hardware can be designed without accuracy loss. Therefore, BDP points can be found in the region only. The results confirm our expectation that large number of WLs can be turned on for RRAM (PRAM) BNN hardware but limited number of WLs should be turned on for MRAM BNN hardware even when RSD is well controlled.

In the Middle and High RSD regions, BNN hardware cannot be implemented without accuracy loss due to variations. However, we can still find design points (SDP) which show the best accuracy given the device conditions as suggested in section III. B. The major difference between Middle and High RSD regions was the N_{OTR} value of the SDP. To decide the boundary between Middle and High RSD region, we compared the inference accuracies of $N_{OTR} = 1$ and $N_{OTR} = 512$. If the accuracy for $N_{OTR} = 1$ is higher(lower) than the accuracy for $N_{OTR} = 512$, the points belong to Middle (High) RSD region. In the Middle RSD region, the $N_{OTR,SDP}$ becomes 1 and in the High RSD region, $N_{OTR,SDP}$ becomes 512. The $AE_{adj,wdp}$ in Eq. (3) was derived using the LAP values.

B. Comparisons Among Different Datasets

For SVHN Dataset, similar to the network for CIFAR-10, BCNN network consisting of 6 convolution layers and 2 FC layers was used. However, the network has half the channel size for each convolution layer compared to the network for CIFAR-10. For MNIST dataset, Binary-MLP network with the size of 784x2048x2048 was used.

From Table I, it is observed that the complex dataset tends to have the smaller $AE_{adj,bdp}$ values and the dataset with smaller $AE_{adj,bdp}$ has more stringent requirement for $N_{OTR,BDP}$ and RSD values.

We can also see that the k value of SVHN is larger than that of the other cases. We think that if the network depth and width are larger than the optimal depth and width the dataset requires, the k value becomes high. This is because, as the network size increases, the number of partial sums that are needed to calculate the final activation value increases. Also, the large k value means that the effect of N_{OTR} is reduced. In other words, the error which occurs when the large number of WLs are turned on becomes smaller than the error that is accumulated with addition of digitized partial sums.

V. CONCLUSION

Based on the analysis of the effect of various memristor device characteristics on MCA-BNN hardware, we conclude that conventional parallel vector-matrix multiplication for MCA, in which all the WLs are turned-on simultaneously, does not work when R-ratio is small. A direct implication from the observation is that only part of WLs should be activated for MRAM BNN hardware which has relatively small R-ratio. We also proposed an adjusted Accuracy Estimation (adj-AE) factor that predicts the accuracy drop in MCA-BNN that may occur due to the variation of real devices. The factor was derived only with the physical and structural characteristics of the memristor, such as R-ratio, RSD, and N_{OTR}. In the proposed scheme, the design point can be expressed with a single adj-AE value when the MNIST, SVHN, and CIFAR-10 are inferred on a MCA-BNN hardware.

ACKNOWLEDGMENT

This research was supported by the "Nano-Material Technology Development Program" through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (NRF-2016M3A7B4910249) and the MSIT(Ministry of Science and ICT), Korea, under the "ICT Consilience Creative program" (IITP-2018-2011-1-00783) supervised by the IITP(Institute for Information communications Technology Promotion), and Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-04.

REFERENCES

[1] S. Yu and P.-Y. Chen, "Emerging memory technologies: recent trends and prospects," *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43–56, 2016.

[2] A. Chen, "A review of emerging non-volatile memory (nvm) technologies and applications," *Solid-State Electronics*, vol. 125, pp. 25–38, 2016.

[3] C.-M. Dou, W.-H. Chen, C.-X. Xue, W.-Y. Lin, W.-E. Lin, J.-Y. Li, H.-T. Lin, and M.-f. Chang, "Nonvolatile circuits-devices interaction for memory, logic and artificial intelligence," in *VLSI Circuits, 2018 Symposium on*. IEEE, 2018, pp. T162–T163.

[4] Z. Sun, H. Li, Y. Chen, and X. Wang, "Voltage driven nondestructive self-reference sensing scheme of spin-transfer torque memory," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 11, pp. 2020–2030, 2012.

[5] S. Salehi, D. Fan, and R. F. Demara, "Survey of stt-mram cell design strategies: Taxonomy and sense amplifier tradeoffs for resiliency," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 3, p. 48, 2017.

[6] Z. He, Y. Zhang, S. Angizi, B. Gong, and D. Fan, "Exploring a sot-mram based in-memory computing for data processing," *IEEE Transactions on Multi-Scale Computing Systems*, 2018.

[7] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 525–542.

[8] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," *arXiv preprint arXiv:1602.02830*, 2016.

[9] Y. Zhang, W. Zhao, Y. Lakys, J.-O. Klein, J.-V. Kim, D. Ravelosona, and C. Chappert, "Compact modeling of perpendicular-anisotropy cofeb/mgo magnetic tunnel junctions," *IEEE Transactions on Electron Devices*, vol. 59, no. 3, pp. 819–826, 2012.