Journal of Semiconductor Engineering Vol. 1, Issue 1, June, 2020
https://doi.org/10.22895/jse.2020.0004
Received 15 Jun 2020, Accepted 16 Jun 2020 ISSN(On-line) 2733-6719

An Approximate DRAM Architecture for Energy-efficient Deep Learning

Duy Thanh Nguyen[1] and Ik-Joon Chang[1]

[1] Electronics, Kyunghee University, Yongin, 17104, Korea
Corresponding Author: Ik-Joon Chang (ichang@khu.ac.kr)

Funding Information: This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT1602-03.

ABSTRACT

We present an approximate DRAM architecture for energy-efficient deep learning. Our key premise is that by bounding memory errors to non-critical information, we can significantly reduce DRAM refresh energy without compromising recognition accuracy of deep neural networks. To validate the key premise, we make extensive Monte-Carlo simulations for several well-known convolutional neural networks such as LeNet, ConvNet and AlexNet with the input of MINIST, CIFAR-10, and ImageNet, respectively. We assume that the highest-order 8-bits (in single precision) and 4-bits (in half precision) are protected from retention errors under the proposed architecture and then, randomly inject bit-errors to unprotected bits with various bit-error-rates. Here, recognition accuracies of the above convolutional neural networks are successfully maintained up to the 10^{-5}-order bit-error-rate. We simulate DRAM energy during inference of the above convolutional neural networks, where the proposed architecture shows the possibility of considerable energy saving up to 10 ~ 37.5% of total DRAM energy.

KEY WORDS

DRAM, deep learning, energy efficiency, approximate computing.

1. INTRODUCTION

Deep neural network (DNN) succeeds in making a quantum leap in some areas such as image or speech recognition accuracy. Recently, even in the areas such as medicine and musical composition, where people have conventional belief that machine cannot deliver human-competitiveness, machine learning based on DNN, so called deep learning, demonstrates successful performance [1]. These successes make people consider that deep learning may lead to paradigm shift in human life, making strong momentum for the development of DNN to provide better recognition accuracy. Hence, researchers have continued to develop a new convolutional neural network (CNN), an effective DNN structure for image recognition, such as AlexNet, VGGNet, GoogleNet and ResNet [2]-[5]. Many researchers have shown that the above CNN algorithms achieve recognition accuracy as good as human [4].

It should be noted that for the inference and the training of these CNN algorithms, myriads of weight and activation parameters need to be referred or generated [6]. Hence, large capacity of main memories, implemented as DRAM, is required to efficiently operate deep learning systems based on these CNN algorithms. Under such circumstance, it is highly probable that the energy dissipation of DRAM becomes significant. Hence, researchers have explored many techniques to reduce the energy dissipation of DRAM in deep learning systems. For instance, the authors of [6] propose a novel framework for the compression of deep neural networks, which dramatically reduces the number of deep learning parameters. The research groups of [18] also develop ASICs dedicated for deep learning, where they invent techniques to reuse data stored in on-chip memories. The aim of these data-reusing techniques is to mitigate main memory footprints.

The above researches make significant technical progress in deep learning since they simultaneously improve system energy and throughput by reducing the number of DRAM accessing. However, these works have a limitation to be applied only for the inference of deep learning. Moreover, previous

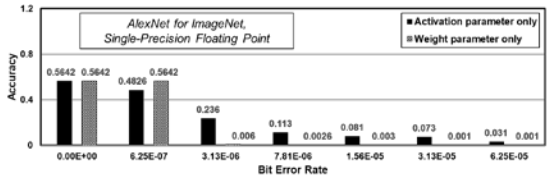


Figure 1. Inference Accuracy with respect to data errors

literature [8] has shown that in DRAM, refresh energy occupies considerable portion of total energy dissipation. The authors of [8] show that as the density of DRAM increases, refresh energy becomes more critical. These make strong motivation to improve DRAM refresh energy in both inference and training of deep learning.

In this work, we present an approximate DRAM architecture to reduce DRAM refresh energy in deep learning systems. In our approximate DRAM architecture, we allow retention errors to mitigate the constraint of auto-refresh time, 64ms in JEDEC specifications [19], or self-refresh time. However, we make these retention errors are bounded to non-critical information, which is least significant bits of weight and activation parameters and regulate bit-error-rate (BER) due to expanded refresh time below a certain level. In our approximate DRAM architecture, the range of critical and non-critical information and the regulated BER of non-critical information are reconfigurable and hence, can be varied by users depending on their target applications. These schemes can be implemented with very minor hardware modification of DRAM control parts. Consequently, corresponding overhead with respect to throughput and latency is negligible.

The remaining parts of this paper are organized as follows. In section 2, we briefly review several representative works regarding DRAM refresh energy reduction and analyze their challenge. In section 3, we discuss the major contribution of this paper and the detailed architecture of the proposed DRAM. In section 4, we make some analysis to support the possibility of our proposed DRAM architecture. In section 5, we show our simulation results, validating the proposed architecture. In section 6, our energy estimation results are discussed. Lastly, section 7 concludes this paper.

2. RELATE WORKS AND THEIR CHALLENGE

As the density of DRAM becomes higher, refresh energy of DRAM is predicted to become more significant [8], as mentioned in section 1. Hence, many researchers have made much effort to alleviate such a challenge. For instance, in RAIDR [8], retention times of all DRAM rows are fully profiled and then, a corresponding memory controller categorizes these DRAM rows to several bins with respect to retention time. After referring to the categorized bins, the memory controller adaptively issues auto-refresh command with the optimal refresh period of each row. This approach leads to almost 75% refresh energy improvement.

The authors of [14] note that the row-level refresh control of RAIDR may accompany considerable performance overhead due to large number of ACT/PRE commands. They efficiently address this challenge by using a dummy-refresh technique, referred as REFLEX. The REFLEX technique uses the same mechanism as normal auto-refresh operations and hence, does not degrade performance. Both RAIDR and REFLEX can be applied under the assumption that retention times of all memory rows can be fully profiled. However, all chips should be individually characterized with respect to retention time for RAIDR and REFLEX, impractical due to large verification overhead.

The authors of [16] develop a technique named as Flikker, which exploits significance-driven approximate computing to improve DRAM refresh energy efficiency. They categorize pages to critical and non-critical ones. Memory rows to store critical pages are refreshed with the normal period while non-critical pages are placed to rows with low-rate refreshes, potentially exposed to retention errors. This provides good energy-quality scalability since only non-critical pages experience retention errors. To further improve energy-quality scalability, the authors of [17] characterize retention times of whole physical pages. Then, they sequentially sort physical pages according to the characterized retention times. Logical pages are also sorted by their significance and then, sequentially mapped to the sorted physical pages. Depending on their applications, critical and non-critical pages are varied. System administrators adaptively control refresh time not to affect data integrity of critical pages. These techniques commonly assume that pages can be categorized according to the importance of data. However, in most cases it is difficult to relatively evaluate the importance of data, making the page categorization challenging. In addition, deep learning systems have a hazard that the classification accuracy is significantly degraded even under extremely low BER situation of DRAM.

Figure 1 shows our simulation results. Even small BER of 10^{-7}~ 10^{-6}, which possibly occur at high temperature with the slight increment of refresh time [15], results in significant accuracy degradation for both the first and second scenarios. The simulation results of Figure 1 imply that when the techniques of [16], [17] are applied for AlexNet, we may suffer from considerable classification accuracy degradation. Furthermore, the approach of [17] is valid under the assumption that retention times of physical pages can be fully profiled. However, this is challenging due to the large verification effort, as mentioned above.

3. OUR CONTRIBUTION

Main memories of sever machines are implemented in a dual in-line memory module (DIMM). In the DIMM, the function of error checking and correction (ECC) can be optionally added. In this work, we explain our approach with un-buffered non-ECC DRAM. However, our proposed concepts can be easily extended to DRAMs with ECC.

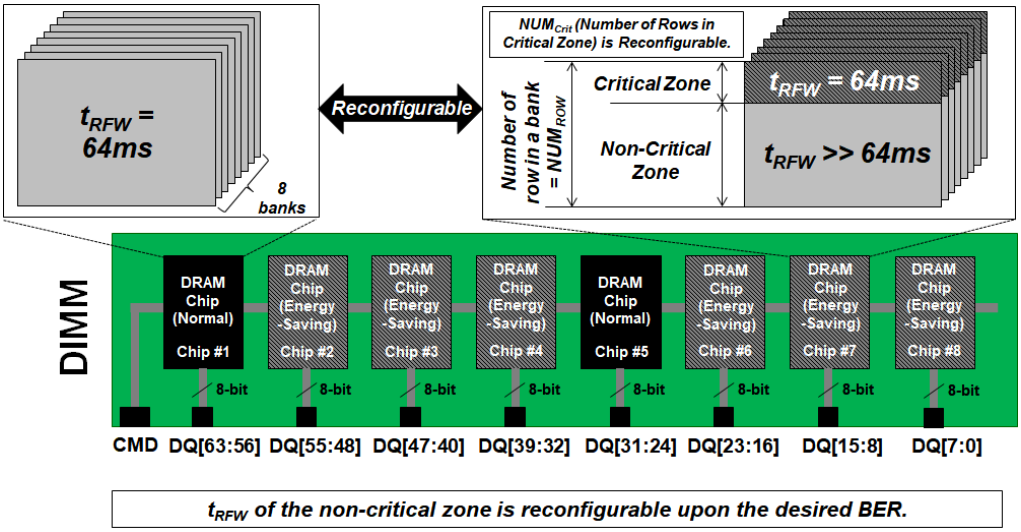


Figure 2. Our approximate DRAM architecture

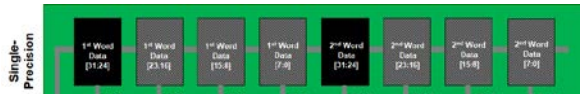
The concept of our approximate DRAM architecture can be summarized to Figure 2. Here, t_{RFW} expresses a refresh time of DRAM. According to JEDEC specification [19], the normal value of t_{RFW} is 64ms below 85°C for auto-refresh operation. Mostly, a DIMM has two ranks, which respectively consist of eight DRAM chips under the scenario of non-ECC. In this work, we assume the typical case that each DRAM chip has 8-bit data lines, which can be changed dependently on DRAM specifications. In the proposed architecture, we categorize eight DRAM chips of a DRAM rank to normal and energy-saving ones. Memory rows of the energy-saving chip are classified to two zones, which are critical and non-critical ones, while the normal chip has the same refresh behavior as conventional DRAM chips, where total rows are auto-refreshed with the t_{RFW} of 64ms below 85°C. In the energy-saving chip, t_{RFW} of the critical zone is same as that of the normal chip, whereas that of the non-critical zone is significantly larger than the normal refresh time.

Unlike Flikker [16], all deep learning parameters are stored to the non-critical zone without page categorization according to their significance and hence, all parameters are exposed to retention errors. The critical zone is necessary only to store program codes or control information, or to support other system programs under multi-tasking environment. We ensure that for the energy-saving chip, BER of its non-critical zone can be regulated below a level desired by system administrators. We assume that the system administrators decide a suitable BER level not to affect classification accuracy of their deep learning system, which can be obtained from system-level analysis, and then, send a BER regulation command to DRAM chips. Then, the DRAM chips adjust t_{RFW} of their non-critical zone to satisfy this.

A. OUR ARCHITECTURE

Let us further explain the architecture of Figure 2. We assume that critical MSBs of deep learning parameters are stored in Chip #1 or Chip #5, and other remaining bits are mapped to the other remaining chips. Throughout this work, we consider two most representative data formats in hybrid CPU-GPU platforms, single-precision and half-precision floating points. Our simulations show that up to the BER of 10^{-5}, 8-bit MSB (for single-precision) and 4-bit MSB protections (for half-precision), which is obtained by storing these bits to normal chips, deliver the same classification accuracies for various CNN algorithms as their corresponding data integrity cases, whose detail discussion is shown in section 4. The above two protection cases can be simply achieved under the architecture of Figure 2 discussed in section 3.B.

The optimal range of protected critical bits can be dependent on the currently used CNN algorithm or data format. Then, system administrators may want to change this setting to optimize their systems. Considering this, we make that it is reconfigurable whether a DRAM chip is normal or energy-saving, which enables us to change the range of protected bits in a DRAM rank. The number of rows placed in the critical zone of energy-saving chips, which is described as NUM_{Crit} in Figure 1 is also reconfigurable. In DDR4 DRAMs, the base granularity of refresh operation is eight rows [19] and hence, NUM_{Crit} should be controlled with the same granularity, which can be easily implemented in



half-precision scenario with only two normal chips, expressed in black color, as shown in Figure 3(b).

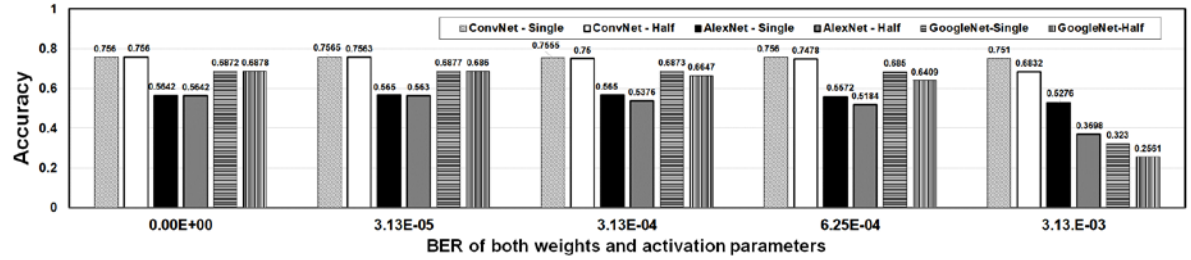


Figure 4. Inference simulation results for various BER scenarios

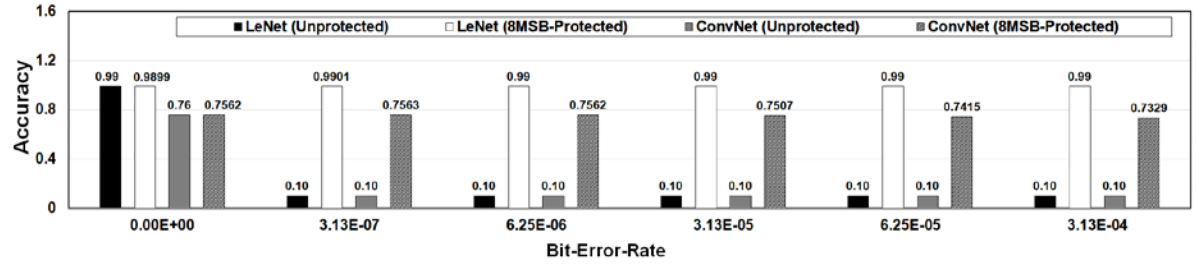


Figure 5. Training simulation results for various BER scenarios

DRAM control part. In Figure 2, the number of banks is expressed to be eight, which is borrowed from the specification of DDR3. However, this work can be easily applied to DDR4, where the number of bank can be further larger.

B. DATA MAPPING RULE

Firstly, we consider two representative data format cases, single- and half-precision floating points, as mentioned above. Under these data formats, our studies show that 8-bit and 4-bit MSB protections are sufficient, respectively. This conclusion is derived from our hypothetical experiments where we inject bit-errors to a special bit-position with a certain BER. The results gathered by various and numerous experiments show that classification accuracies are sensitive to only 8-bit MSBs (for single-precision) and 4-bit MSB (for half-precision).

For the single-precision, the 8-bit MSB protection is simply supported under the architecture of Figure 2, as shown in Figure 3(a). However, to protect 4-bit MSB in the half-precision scenario, the number of normal chips, expressed in black color, needs to be incremented as shown in the case 1 of Figure 3(b) When we assume that due to large t_{RFW}, the refresh energy of energy-saving chips are much smaller than that of normal chips, the scheme of Figure 3(a) leads to 75% refresh energy saving while that of the case 1 in Figure 3(b) obtains only 50% reduction. To further improve the energy efficiency of the half-precision scenario, we present a data mapping such as the case 2 of Figure 3(b) in DIMM, which can be simply implemented with the support of memory controllers. This mapping enables the 4-bit MSB protection of the

4. VALIDATION

As discussed in section 3, in our architecture data integrity is guaranteed for only several critical MSBs. Under the 8-MSB (for single-precision) and the 4-MSB (for half-precion) protection, our hypothetical experiments show that deep learning systems provide sufficiently good classification accuracies even in the existence of run-time retention errors of LSBs, which is the key premise of the proposed architecture. We validate the key premise by running simulation in inference and training of several CNNs.

A. INFERENCE

In the similar way to Figure 1, we run simulations. Under the architecture of Figure 2, the data integrity of 8-MSB (for single-precision) and 4-MSB (for half-precision) is guaranteed. Hence, we generate bit-errors only for the other LSB bits. Let us call the 8-MSB protection of single-precision as the first scenario and the other one as the second one in this section. We perform the simulation for a hybrid CPU-GPU platform. We observe classification accuracies of four CNNs, which are ConvNet (for CIFAR-10 [10, 12]), AlexNet (for ImageNet [13]), and GoogleNet (for ImageNet).
Our simulation results are shown in Figure 4. At the first scenario, our proposed architecture delivers the same classification accuracy as the case with data integrity, with up to 10-4-order BER for all data-sets of CIFAR-10, and ImageNet. At the second scenario, small accuracy drops of ImageNet (less than 3%) are observed from the BER of 10^{-4}-order. Up to 10^{-5}-order BER, classification accuracies are same as the data

integrity case for all data-sets. The classification accuracies of CIFAR-10 are conserved up to the BERs of 10^{-4}-order.

Under the 4-MSB protection of half-precision floating point data, we can assure the data integrity of only sign bit and three higher-order exponent bits, and hence, two remaining exponent bits are still exposed to retention errors. On the other hand, at the first scenario, only single exponent bit experiences retention errors. This makes that the scheme of the second scenario is more sensitive to retention errors compared to that of the first one. In the case that the retention errors can occur in any of all bits without the protection of MSBs, classification accuracies are significantly degraded even at the BER of 10^{-7}-order, shown in Figure 1. This clearly validates the efficiency of our proposed protection schemes in inference operations of deep learning systems.

B. TRAINING

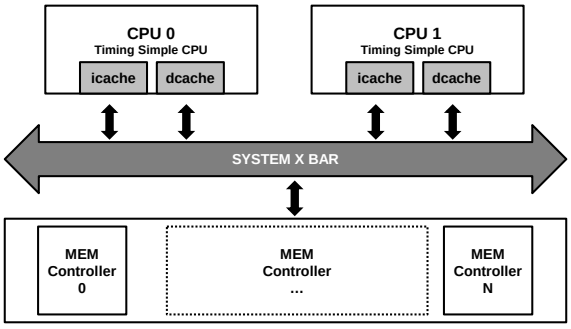
We also make simulations for training, whose results are shown in Figure 5. Here, we generate bit-flipping for unprotected bits, where the bit-flipped positions are randomly chosen. By using the similar method to the above inference simulations, we mitigate the problem that training performance is affected by the positions of bit-flipping. For the training, we only consider the single-precision floating-point data format.

Our simulation results show that the proposed scheme is so effective to training as well. Without the MSB protection, considerable training performance drop is observed for both MNIST [12] and CIFAR-10 at the small BER of 10^{-7}-order. However, the proposed 8-MSB protection makes the training performance to be conserved up to the BER of 10^{-5}-order. At the BER of 10^{-4}-order, only small classification accuracy drop of 2.3% is observed in CIFAR-10.

5. ENERGY SIMULATION

In this section, we simulate DRAM energy dissipation of deep learning systems by employing Gem5 [20] and DRAMPower [21] simulators. According to [15], both auto- and self-refresh energies are proportional to the number of refreshed rows. Hence, under the assumption that besides refresh energies, other energy components are same, we can suitably derive energy dissipation of our architecture. Due to computational overhead, we estimate the energy under CPU platforms instead of hybrid CPU-GPU platforms.

Figure 6 shows our simulation configurations. We imitate a general CPU platform, which have two CPU cores with their own instruction and data caches to operate at 4GHz clock. These CPU cores are connected to memory controllers through 1GHz crossbar switch bus. The number of memory controllers is varied, where three scenarios of 1, 2, and 4 (the number of memory controllers) are considered. We assume that for all cases, the number of ranks is



Simulator	Gem5 and DRAMPower
Framework	Tiny-DNN
Simulation configuration (Based on normal PC configuration)	2 CPUs – 4GHz System Bus – 1GHz # of Memory –channel and # of rank (NC2R) +DDR3_2133_8x8 (2133 MHz, 8x8) +DDR4_2400_8x8 (2400 MHz, 8x8) +DDR4_2400_16x4 (2400 MHz, 16x4)

Figure 6. Our simulation setup

fixed to two since most DDR3 and DDR4 DRAMs have two ranks. The numbers of memory controllers (=N) and rank (=M) are expressed as NCMR in Figure 6 and Figure 7. We employ three kinds of DRAM in the simulation, which are DDR3_2133_8X8 (2133 MHz, 8 chips X 8 DQ lines per chip in a rank = 64 DQ lines), DDR4_2400_8X8 (2400 MHz, 8 chips X 8 DQ lines per chip in a rank = 64 DQ lines) and DDR4_2400_16X4 (2400 MHz, 16 chips X DQ lines per chip in a rank = 64 DQ lines). We utilize the framework of Tiny-DNN [22]. This provides light-weight C++ DNN codes and hence, is suitable for the power simulation.

Figure 7 show the energy simulation and estimation results for single image input of AlexNet (with ImageNet). Due to extremely large computational overhead of training, we can simulate DRAM energy for only inference operation. Here, we define a parameter R_{Crit}, which is the value of 'NUM_{Crit}/NUM_{row}' in Figure 2. When R_{Crit} is one, which implies that all memory rows of energy-saving chips are included in critical region, the energy dissipation of the proposed DRAM architecture is the same as that of the conventional ones. As R_{Crit} is reduced, the number of rows placed in non-critical region becomes larger. For various cases of R_{Crit}, we suitably estimate energy dissipation by using the method mentioned above.

In our DRAM simulation, total refresh energy, the summation of auto- and self-refresh energies, occupies a substantial portion of total DRAM energy, roughly 14.7 ~ 50% for AlexNet. The DDR4 DRAMs used in our simulation have sixteen memory banks while in DDR3 DRAM, the number of memory banks is eight, making that refresh energy becomes more critical in DDR4 DRAM compared to DDR3 DRAM. Such trend clearly appears in our simulation results, which also show that with the increment of memory channel number, refresh energy tends to become larger. This is highly correlated to the fact that as the number of memory channel increases, the number of refreshed rows increases. It should be noted that our

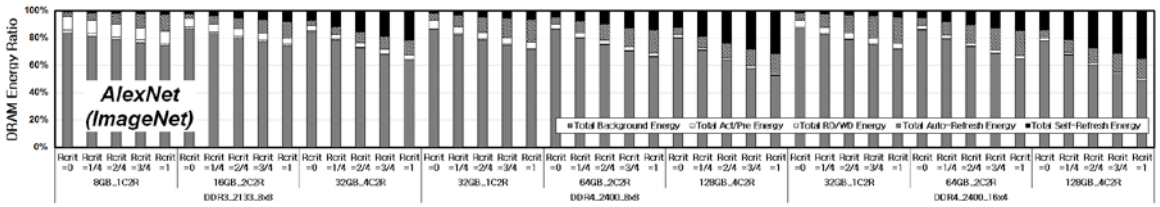


Figure 7. Energy Simulation Results

approximate DRAM architecture significantly reduces the refresh energy, up to 75% when R_{Crit} is zero. Such refresh energy saving corresponds to 11~ 37.5% of the total DRAM energy for AlexNet.

6. CONCLUSION

We present an approximate DRAM architecture to reduce refresh energy of main memories. In the proposed architecture, critical MSBs are stored to normal chips, which are normally refreshed, while energy-saving chips, where other reminding bits regarded relatively non-critical are stored, have expanded refresh times. Here, only non-critical bits are exposed to retention errors. By suitably regulating bit-error-rate of energy-saving chips, we can significantly improve DRAM energy without compromising classification accuracy of deep learning inference. This concept is validated through extensive simulations. Also, we show that the proposed architecture can be applicable to training of deep learning.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proceedings of ICLR*, 2015.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan,V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions,"in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[6] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "Eie: efficient inference engine on compressed deep neural network," in *Proceedings of the 43rd International Symposium on Computer Architecture*, IEEE Press, 2016.

[7] Y.-D. Kim, E. Park and D. Shin, "Compression of deep convolutional neural networks for fast and low power mobile applications," *ICLR*, 2016

[8] J. Liu, B. Jaiyen, R. Veras, and O. Mutlu, "Raidr: Retention-aware intelligent dram refresh," in *Proceedings of the 39th Annual International Symposium on Computer Architecture*, ACM, 2012.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, 1998.

[10] K. Alex, "Cifar-10 cuda-convnet model," https://code.google.com/archive/p/cuda-convnet/.

[11] Y. LeCun, "The mnist database of handwritten digits," http://yann. lecun.com/exdb/mnist/.

[12] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," http://www. cs. toronto. edu/kriz/cifar. html, 2014.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2009.

[14] I. Bhati, Z. Chishti, S.-L. Lu, and B. Jacob, "Flexible auto-refresh: Enabling scalable and energy-efficient dram refresh reductions," in *Proceedings of the 42th Annual International Symposium on Computer Architecture*, 2015.

[15] J. Liu, B. Jaiyen, Y. Kim, C. Wilkerson, and O. Mutlu, "An experimental study of data retention behavior in modern dram devices: Implications for retention time profiling mechanisms," in *Proceedings of the 40th Annual International Symposium on Computer Architecture*, 2013.

[16] S. Liu, K. Pattabiraman, T. Moscibroda, and B. G. Zorn, "Flikker:Saving dram refresh-power through critical data partitioning," in *Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser, 2011.

[17] A. Raha, S. Sutar, H. Jayakumar, and V. Raghunathan, "Quality config-urable approximate dram," *IEEE Transactions on Computers*, vol. 66, 2017.

[18] V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, "Efficient process-ing of deep neural networks: A tutorial and survey," *arXiv preprintarXiv*:1703.09039, 2017

[19] "Jedec, ddr3 sdram specification," 2010

[20] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashtiet et al., "The gem5 simulator," *ACM SIGARCH Computer Architecture News*, vol. 39, 2011.

[21] K. Chandrasekar, C. Weis, Y. Li, B. Akesson, N. Wehn, and K. Goossens, "Drampower: Open-source dram power & energy estimation tool," http://www. drampower. info, vol. 22, 2012.

[22] Tiny-DNN, "header only, dependency-free deep learning framework inc++," https://github.com/tiny-dnn.

AUTHOR BIOGRAPHIES

Duy Thanh Nguyen is currently pursing Ph.D. degree under the supervision of Prof. Ik-Joon Chang at Kyung Hee University (KHU), Republic of Korea. Before joining in KHU, he obtained bachelor degree in Computer Engineering from Ho Chi Minh City, University of Technology (HCMUT). His research interests are approximate computing, energy-efficient architecture and memory systems.

Ik-Joon Chang (M'12) received the B.S. degree, with summa cum laude, in electrical engineering from Seoul National University, Seoul, Korea. And he acquired his M.S. and Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, in 2005 and 2009, respectively. After his graduation, he worked in Samsung flash design team for two years. Now, he is an associate professor of Kyunghee University, Korea.