An Energy-Efficient Quantized and Regularized Training Framework For Processing-In-Memory Accelerators

Hanbo Sun[1*], Zhenhua Zhu[1*], Yi Cai[1]
Xiaoming Chen[2], Yu Wang[1], Huazhong Yang[1]
[1]Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Abstract—Convolutional Neural Networks (CNNs) have made breakthroughs in various fields, while the energy consumption becomes enormous. Processing-In-Memory (PIM) architectures based on emerging non-volatile memory (e.g., Resistive Random Access Memory, RRAM) have demonstrated great potential in improving the energy efficiency of CNN computing. However, there is still much room for improvement in the energy efficiency of existing PIM architectures. On the one hand, current work shows that high resolution Analog-to-Digital Converters (ADCs) are required for maintaining computing accuracy, but they dominate more than 60% energy consumption of the entire system, damaging the energy efficiency benefits of PIM. On the other hand, the characteristic of computing in the analog domain in PIM accelerators leads to the computing energy consumption is influenced by the specific input and weight values. However, as far as we know, there is no energy efficiency optimization method based on this characteristic in existing work. To solve these problems, in this paper, we propose an energy-efficient quantized and regularized training framework for PIM accelerators, which consists of a PIM-based non-uniform activation quantization scheme and an energy-aware weight regularization method. The proposed framework can improve the energy efficiency of PIM architectures by reducing the ADC resolution requirements and training low energy consumption CNN models for PIM, with little accuracy loss. The experimental results show that the proposed training framework can reduce the resolution of ADCs by 2 bits and the computing energy consumption in the analog domain by 35%. The energy efficiency, therefore, can be enhanced by $3.4\times$ in our proposed training framework.

I. Introduction

Recently, Convolutional Neural Networks (CNNs) have made breakthroughs in various fields, such as image classification and object detection. However, CNN structures become more and more complex, with the amount of calculation and the energy consumption increase dramatically.

Previous work has demonstrated the great potential of Processing-In-Memory (PIM) architectures based on emerging non-volatile memory in improving energy efficiency in CNN computing. Because PIM architectures have the ability to complete the CNN computing in memory by converting convolution operations into analog-domain Matrix-Vector-Multiplications (MVMs), data movements are greatly reduced and energy efficiency can be enhanced by over $100\times$ compared with CMOS-based architectures [1].

Although PIM architectures can improve the energy efficiency of CNN computing, there is still much room for improvement. Firstly, in existing PIM accelerators, the resolution of ADCs has a crucial impact on accuracy and energy consumption of the entire system. On the one hand, PIM accelerators based on high-resolution ADCs can achieve high accuracy in large scale datasets and CNN models. However, the high-resolution ADCs severely damage the energy efficiency improvement brought by PIM, for the reason that ADCs occupy more than 60% energy consumption of the entire system [2] and high-resolution ADCs consumes more energy than low-resolution ADCs (e.g., the 8-bit ADC in [3] costs $8.66\times$ energy than the 4-bit ADC in [4] for one conversion). On the other hand, in order to reduce the overhead of ADCs, researchers have proposed several PIM accelerators with low precision interfaces. However, these work focuses on simple algorithm models (such as MLP based on MNIST datasets) [5] [6] or has high accuracy loss in large scale algorithms (e.g., Using 4-bit ADCs causes 8% accuracy loss on ResNet18 @ cifar10) [7]. Therefore, how to reduce the ADCs' energy consumption further, while ensuring the computing accuracy, is of great importance for improving the overall energy efficiency of the PIM accelerators.

Secondly, existing work does not consider the relationship between calculated data values and computing energy consumption. PIM accelerators perform MVMs in the analog domain, and the energy consumption is related to the specific data values. For example, in RRAM-based PIM accelerators, high resistance state (HRS) and low voltage level (LVL) are used to represent 0 in weights and input data, low resistance state (LRS) and high voltage level (HVL) represent 1. The LVL is usually set to 0V and the HVL is set to the NVM read voltage (e.g., $0.15 \sim 1$V [8] [9]), resulting in different energy consumption on different input values. Besides, in order to minimize the calculation error caused by the resistance deviations, the R ratio (i.e., HRS/LRS) needs to be large enough (e.g., $10\sim100$ [1]). In other words, when applying the same input voltage, the energy consumption gap between HRS and LRS can reach one to two orders of magnitude. Therefore, it is necessary and feasible to improve the energy efficiency of analog computing by adjusting the distributions of inputs and weights.

In this paper, we combine the characteristics of CNN computing and PIM accelerators, analyze and model the energy consumption of PIM computing, propose an energy-efficient quantized and regularized training framework for PIM accelerators. The proposed training framework is composed of a PIM-based non-uniform activation quantization scheme and an energy consumption aware weight regularization method. Our framework can improve the energy efficiency by reducing

*: Both authors contributed equally to this work.

978-1-7281-4123-7/20/$31.00

the ADC resolution requirements and training low energy consumption CNN models for PIM with little accuracy loss. The main contributions of this paper include:

i) We design a PIM-based non-uniform activation quantization scheme, including quantization range optimization, high-precision scale design, and non-uniform quantization method. As a result, the quantization resolution of ADCs can be reduced by 2 bits, with 70% energy consumption reduction and comparable accuracy to traditional activation quantization methods.
ii) We propose an energy consumption aware weight regularization method, consisting of an energy consumption model for PIM and a weight regularization method used in CNN training. The proposed weight regularization method can reduce 35% analog computing energy consumption by adjusting the distributions of inputs and weights.
iii) Experiments show that the training framework can improve $3.4\times$ energy efficiency of PIM accelerators with little accuracy loss. The equivalent energy efficiency is 9.02 TOPS/W, nearly $2.6 \sim 4.2\times$ compared with existing work.

II. Preliminaries and Related Work

A. CNN

CNNs mainly consist of convolutional (CONV) layers and fully-connected (FC) layers. CONV layers realize the convolution operation which is described as:

$$A_o(h,w,c) = \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{C_{in}} A_i(h+i, w+j, k) w_c(i,j,k) \quad (1)$$

where A_i and A_o represent the input feature map and output activations, respectively. w_c is a 3-dimensional convolutional kernel with the size of $K \times K \times C_{in}$. The computations in FC layers are similar to those in CONV layers.

B. PIM Architectures

In PIM architectures, crossbars based on emerging non-volatile memory (NVM, e.g., RRAM) construct the basic computing unit, as shown in Figure 1 (a). When applying the input voltage vector $\boldsymbol{V}$ to the word-lines (WLs) of crossbars and mapping the matrix values to the NVM conductance $\{g_{i,j}\}$, then we can get the MVM results, which are represented by the bit-line current $\boldsymbol{I}$, as shown in Equation 2:

$$i_{out,k} = \sum_{j=1}^{N} g_{k,j} v_{in,j} \quad (2)$$

where $v_{in,j}$ and $i_{out,k}$ are the component of $\boldsymbol{V}$ and $\boldsymbol{I}$, respectively. By performing the analog domain calculations, matrix data movements are eliminated, introducing energy efficiency improvements.

However, because of the analog computing pattern of PIM architectures, ADCs and DACs are used as interfaces between crossbars and peripheral digital circuits. Researchers have pointed out that the ADCs occupy more than 60% energy consumption of the overall PIM architectures, which damage the energy efficiency gains of PIM architectures [2]. To tackle this problem, in the hardware level, some work proposed low precision interface circuits design to substitute ADCs [5] [6]. But they only concentrate on small scale applications and algorithms (e.g., FFT and four layers CNN). In the software level, researchers design the low bit-width CNN for PIM architectures to reduce the resolution requirement of ADCs, which introduce additional accuracy loss overhead [7].

III. Framework Overview

The overall quantized and regularized training framework is shown in Figure 1 (a), which consists of a PIM-based non-uniform activation quantization scheme (Section IV) and an energy-aware weight regularization method (Section V). The PIM-based non-uniform activation quantization scheme includes quantization range optimization (Section IV.IV-B), high-precision scale implementation (Section IV.IV-C), and a non-uniform quantization method (Section IV.IV-D), and can reduce ADCs energy consumption from 2 bits ADC resolution reduction.

The energy-aware weight regularization contains crossbar computing energy modeling (Section V.V-A) and weight regularization (Section V.V-B), which can reduce the ratio of HVL on LRS (with high energy cost) by 41% and achieve 35% analog computing energy reduction.

The computational flow of the proposed training framework is shown in Figure 1 (b). Compared with traditional training methods, we optimize the uniform quantization with quantization range optimization and high-precision scale implementation. For each layer, in the forward phase, the input data and weights are quantized by the optimized uniform quantization. Then the quantized inputs and weights are used to calculate the output activation and the energy consumption indicator. The outputs are quantized by the PIM-based non-uniform activation quantization scheme and the energy consumption indicator is added to the final loss. Because all the above computational operators are derivable, the standard back-propagation algorithms can be used in the backward phase of the proposed training framework.

IV. PIM-based Non-uniform Activation Quantization Scheme

As mentioned in Section I, high resolution ADCs cause a heavy energy burden in PIM accelerators, while low resolution ADCs introduce larger quantization error and bring higher accuracy loss. To handle these problems, we propose the PIM-based non-uniform activation quantization scheme, which contains quantization range optimization, high-precision scale implementation, and non-uniform quantization, as shown in Figure 1. Based on this activation quantization scheme, we can reduce the ADC resolution requirements without any accuracy loss, and can further improve the computing energy efficiency.

A. Traditional Activation Quantization Method

Existing activation quantization methods used in CNN consist of the following three parts:

1) **Quantization scale determination:** Firstly, the maximum of the input vector absolute value $|x_{in}|$ is used as the quantization range. Then, the minimal value α which is an integer power of 2 and can cover the quantization range is found as a scaling parameter for next steps. The

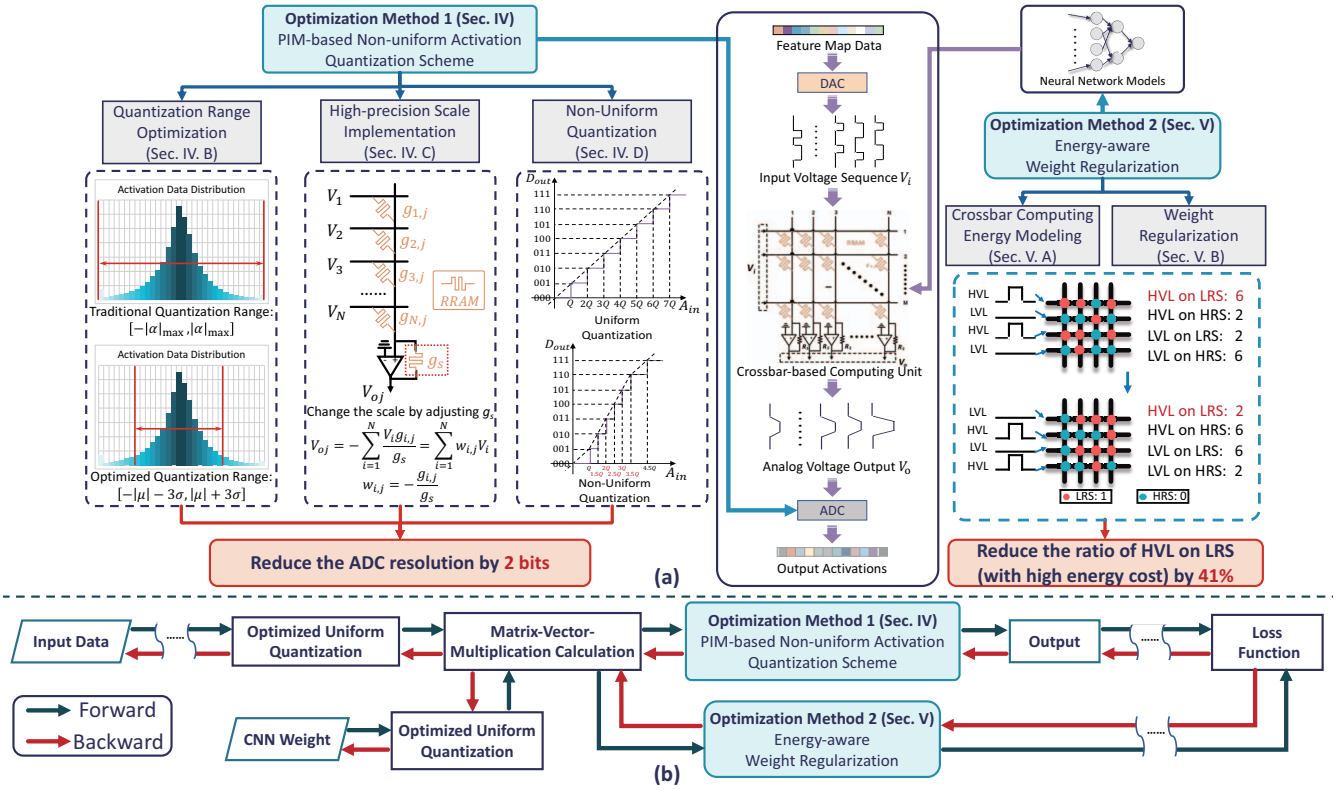

Fig. 1. (a) The overall quantized and regularized training framework. (b) Computational flow of the proposed training framework (only list one layer)

calculation formula of the scaling parameter α is shown in Equation 3.

$$\alpha = 2^{\lceil log_2(max(|x_{in}|)) \rceil} \tag{3}$$

2) **Linear uniform quantization:** Firstly, linear scaling is used to map the input vector x_{in} to the interval [-1, 1]. Then, a k-bit uniform quantization is performed as shown in Equation 4.

$$\begin{aligned} x_{q,out} &= Q_k\left(x_{in}, \alpha, k\right) \\ &= round\left(\left(2^{k-1}-1\right)\frac{x_{in}}{\alpha}\right)\frac{\alpha}{2^{k-1}-1} \end{aligned} \tag{4}$$

where $x_{q,out}$ is the quantization results.

3) **Gradient back-propagation:** Because the gradients of the function $round$ equal to zero at continuous points, straight-through estimator (STE) [7] is used to generate gradients to the input vector x_{in}.

B. Quantization Range Optimization

In the traditional activation quantization method, the quantization range is determined by the maximum of the input vector absolute values. However, the maximum of absolute values is susceptible to individual extreme data and can not reflect the overall data distribution. Figure 2 (a) implies less than 5% data appear in the interval $(max/4, max]$, which means using the maximum as the quantization boundary "wastes" 75% quantization bit width. Referring to [10], we can assume the activation data distribution is close to a Gaussian distribution. In Gaussian distribution, 3σ criterion is often used in anomaly detection on account of the probability of normal data appearing in $[\mu-3\sigma, \mu+3\sigma]$ is 99.73%. Inspired by this, we use $[-|\mu|-3\sigma, |\mu|+3\sigma]$ as the new quantization range instead of choosing the maximum value, which can cover $> 97\%$ data as shown in Figure 2 (a). Besides, Figure 2 (b) shows that the quantization range size is reduced to 25% after using the optimized quantization range. Therefore, the quantization range optimization method can reduce the requirements for ADCs resolution in PIM architectures.

Similar to the activation quantization method in [11], the proposed quantization range optimization method adopting a dynamic update method to the quantization range. However, our method determines the quantization range by the statistical characteristics of the data distribution, which means the proposed method can be independent of parameterized clipping and network training and can be applied to the weights.

C. High-precision Scale Implementation

Limited by the binary digital computing, the scaling parameter in traditional quantization methods must be an integer power of 2, which brings two problems. On the one hand, the integer power of 2 makes the scale calculated by Equation 3 larger than the actual scale (at most $2\times$), but the precision k in Equation 4 is unchanged, resulting in the increase of quantization error. On the other hand, in the training phase, different inputs require different scales. Traditional quantization methods choose the maximum scale as the scaling factor, causing high quantization error. To settle down these two problems, we propose PIM-based high-precision scale implementation. In the PIM accelerators, MVMs are carried out in the analog domain as shown in Figure 1 (a). Compared with existing

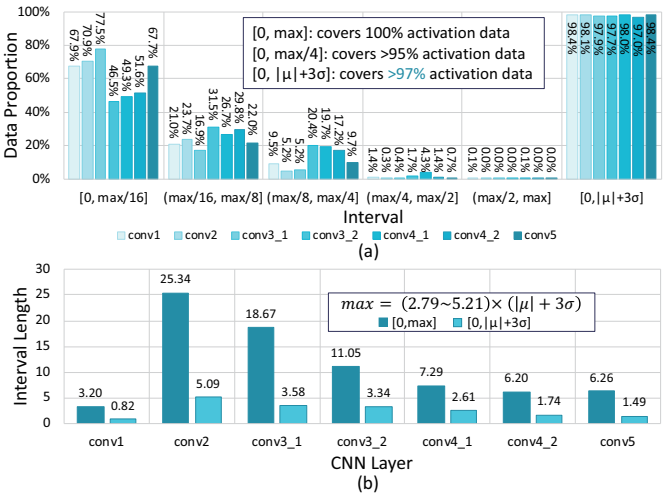

Fig. 2. (a) The percentage of activation data in different ranges w.r.t. different CNN layers in a well-trained VGG-8 model. (b) The quantization range sizes in different layers.

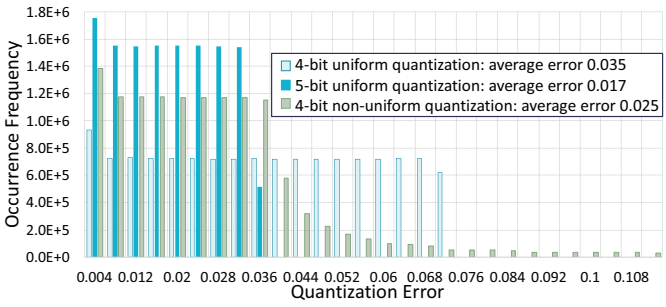

Fig. 3. Quantization error distributions under different quantization methods (i.e., 4-bit uniform quantization, 5-bit uniform quantization, and 4-bit non-uniform quantization)

work, we use an RRAM with adjustable resistance instead of a resistor as the load resistance. After that, we can transfer the bit-line current with different ranges to a fixed voltage range by adjusting the load resistance g_s, which is equivalent to achieve high scaling parameter other than an integer power of 2.

Besides, in the typical CNN training phase, one mini-batch data is trained at a time because of the storage limitation. Different mini-batch data bring different scales during training. The instability of the scaling parameter may cause severe jitter and non-convergence during training. In order to alleviate the jitter and non-convergence in the training phase, we introduce the momentum smoothing method. After a new quantization scaling parameter is generated, it is weighted summed with the previous scaling parameter. Then we can get a smoothed quantization scaling parameter, as shown in Equation 5(m denotes the momentum coefficient).

$$\alpha \leftarrow m\alpha + (1-m)\left(|mean(v_{in})| + 3std(v_{in})\right) \tag{5}$$

D. Non-uniform Quantization

Since the activation data in CNNs are close to the Gaussian distribution [10], uniform quantization causes large quantization error compared with non-uniform quantization, as shown in Figure 3. But non-uniform quantization brings non-uniform results, and the quantized results can not be directly used for computing. For example, in Figure 1 (a), the non-uniform quantization quantifies Q to 001 and $1.5Q$ to 010, but $(Q+Q=2Q) \neq (001+001=010=1.5Q)$. Thus, additional uniform mapping is required after non-uniform quantization for correct computing. In the CMOS-based computing architectures, the uniform mapping will cause extra area and energy overhead, and non-uniform quantization cannot improve hardware performance compared with uniform quantization. Therefore, non-uniform quantization implementation in CMOS-based computing architectures does not bring any benefits in energy consumption.

Different with CMOS-based computing architectures, in PIM architectures, the activation quantization is executed by ADCs. From Figure 3, non-uniform quantization can reduce 1 bit ADC resolution compared with uniform quantization without accuracy loss, and ADC overheads grow exponentially with resolution [12]. Therefore, adopting non-uniform quantization can reduce ADC energy overhead and improve overall energy efficiency. In our design, we realize the non-uniform quantization and uniform mapping by non-uniform ADCs and multiplexers (MUXes) in PIM accelerators.

In order to determine the specific non-uniform quantization method in non-uniform ADCs, we construct a transformation function which makes the transformed input data obey uniform distribution. Besides, it has been proved mathematically that the cumulative distribution function (CDF) can be used as such transformation function [13]. Moreover, for the activation data with approximate Gaussian distribution, we use Sigmoid function, i.e., $f(x) = 1/(1+e^{-x})$, to approximate the CDF of activation data. In the circuits design level, the non-uniform ADCs can be implemented by generating different reference voltage levels from adjusting the capacitance and the divider resistance value in SAR ADCs or Flash ADCs, with little area and energy overhead.

MUXes are leveraged to map low precision non-uniform quantization results to high precision uniform quantized data. In the training phase, we add a uniform mapping module to simulate the function of MUXes after the non-uniform quantization. After getting the well-trained CNN model, we can also determine the output of the uniform mapping module in each layer. In the deployment, the select signals of MUXes are connected with the output of the non-uniform ADCs, while the input signals are set as the output of the additional uniform mapping module.

Combining the above two parts, the proposed non-uniform quantization method implemented by non-uniform ADCs and MUXes can be expressed as Equation 6:

$$\begin{aligned}\hat{x}_{q,out} &= \hat{Q}_k\left(x_{in}, \alpha\right) \\ &= Q_{2k}\left(f^{-1}\left(\frac{clip\left(r\left(f\left(\eta\frac{x_{in}}{\alpha}\right)2^k\right), 1, 2^k-1\right)}{2^k}\right)\frac{\alpha}{\eta}, \alpha\right)\end{aligned} \tag{6}$$

where f, r and Q denote $sigmoid$, $round$ and uniform quantization. η is a parameter used to adjust the distribution.

V. Energy Aware Weight Regularization

Restricted by the immature manufacturing process, the computing frequency is ~ 100 MHz in the existing PIM accelerators, and it is hard to make further dramatic improvements. Besides, in order to ensure the PIM computing reliability, R ratio is large enough (e.g., $10 \sim 100$), causing a huge energy consumption gap among different conductance states and input voltage levels. Existing work focuses on mapping a well-trained CNN model on PIM architectures, ignoring the fact that different data values bring different energy consumption. To further enhance the energy efficiency of PIM architectures, it is necessary to model the relationship between computing data values and energy consumption, and design an energy-efficient CNN training method considering the relationship model.

A. Energy Consumption Model

As mentioned in Section I, PIM architectures perform MVMs in the analog domain. The input data, weights, and MVM results are represented by the input voltage on the word-line (V_i), cell conductance (g_{ij}), and current on the bit-line (I_j), respectively. According to the Trans-impedance amplifier (TIA) sensing model [14], the energy consumption of each MVM result in PIM can be calculated by Equation 7.

$$E_j = \sum_{i=1}^{N} V_i^2 g_{ij} t, \quad I_j = \sum_{i=1}^{N} V_i g_{ij} \tag{7}$$

where E_j denotes the energy consumption for generating I_j, and t represents the clock cycle of the analog computing.

Equation 7 gives the energy consumption of a single MVM in crossbars. Noticing that the differences between energy consumption expression and output current expression are Δt and the exponent of V_i. Therefore, we can derive the calculation formula of energy consumption by modifying the CNN computing formula. Besides, we also take the mapping relationship between algorithm parameters (i.e., weights and input data) and analog computing parameters (i.e., DAC full-scale voltage output V_{FS} and crossbar conductance parameter g determined by the RRAM resistance and the load resistor) into consideration. Because the mapping relationship is proportional, we can get the energy consumption model shown in Equation 8:

$$E = \phi\left(\left(\frac{x}{\alpha_x} V_{FS}\right)^2, \frac{w}{\alpha_w} g\right) t = V_{FS}^2 g t \phi\left(\left(\frac{x}{\alpha_x}\right)^2, \frac{w}{\alpha_w}\right) \tag{8}$$

where x, w, and ϕ denote the input data, weights, and the computing function, respectively. α_x and α_w are the quantization scale for x and w.

B. Weight Regularization

Considering the overall energy consumption in the loss function, we propose an energy consumption aware weight regularization method, which is expressed as:

$$L = L_s + \lambda V_{FS}^2 g t \sum_{l=1}^{L} \phi_l \left(\left(\frac{x_l}{\alpha_{x_l}}\right)^2, \frac{w_l}{\alpha_{w_l}}\right) \tag{9}$$

L_s is the softmax loss. λ is a coefficient. ϕ_l denotes the computing in layer l. x_l, w_l, α_{x_l}, and α_{w_l} are the input data, weights, and the corresponding quantization scale in layer l, respectively.

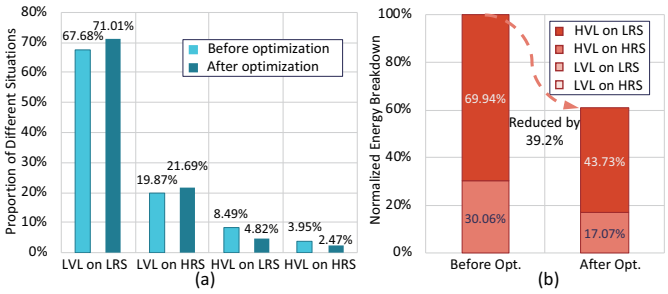

Fig. 4. (a) The proportion of different situations before/after weight regularization. (b) The normalized energy consumption of different situations before/after weight regularization

C. Optimization Results

In the PIM architecture, the input voltage level determines whether the energy consumption closes to zero or not, and the R ratio determines the difference of the energy consumption between HRS and LRS under the same voltage. HVL on LRS has the highest energy consumption and HVL on HRS consumes more energy than other two situations. The optimization results of the weight regularization are shown in Figure 4. The results show that the proposed method can reduce the proportion of HVL on LRS and HVL on HRS by $\sim 41\%$ and the total computing energy consumption by $\sim$40%.

VI. Experiment Results

A. Experiment Setup

We test the proposed training framework on three types of CNN models: LeNet [15], VGG-8 [16], and ResNet-18 [17]. All experiments are evaluated on the Cifar-10 dataset [18]. We change the network structure of ResNet-18 by reducing the number of the channel to a quarter of the original for faster training. The PIM architecture we used refers to [2], which is composed of crossbars with the size of 256×256. We model the crossbar computing energy according to the RRAM data from [9] (HRS, LRS, and read voltage are 150 KΩ, 30 KΩ, and 0.15V) and the system frequency is set to 100MHz. For the ADC part, we use the data from [3] (8-bit, 4mW @ 1.1GS/s), [19] (6-bit, 1.28mW @ 1GS/s), and [4] (4-bit, 0.7mW @ 1GS/s), and for the DAC part, we use the 1-bit DAC design mentioned in [12]. Besides, we synthesize the digital circuit modules at 45nm technology node with 500MHz using Cadence Encounter® RTL Compiler.

B. Energy Consumption and Accuracy Results

Table I shows the accuracy and energy consumption of PIM architectures compared with existing work [7]. On the one hand, under the premise that CNN parameters have the same precision, our framework improves $> 10\%$ classification accuracy with lower energy consumption as shown in the last two lines. On the other hand, the proposed training framework can achieve a comparable accuracy compared with the floating baseline, but brings $> 70\%$ energy conduction (equivalent to $3.4\times$ energy efficiency improvement).

TABLE I
ACCURACY AND ENERGY CONSUMPTION UNDER DIFFERENT SITUATIONS (THE NUMBERS FOLLOWING *A* AND *W* ARE THE PRECISION OF ACTIVATIONS AND WEIGHTS)

		LeNet		VGG-8		ResNet-18	
		Accuracy	Energy/nJ	Accuracy	Energy/nJ	Accuracy	Energy/nJ
float baseline		0.7448	-	0.9336	-	0.8887	-
[7]	A8W4	0.7499	1683	0.9308	2115889	0.8785	19767
	A6W4	0.7463	408	0.9243	458193	0.8756	4517
	A4W4	0.6375	154	0.1887	140889	0.7412	1563
Ours	A4W4	0.7467	**142**	0.9286	**138793**	0.8655	**1500**

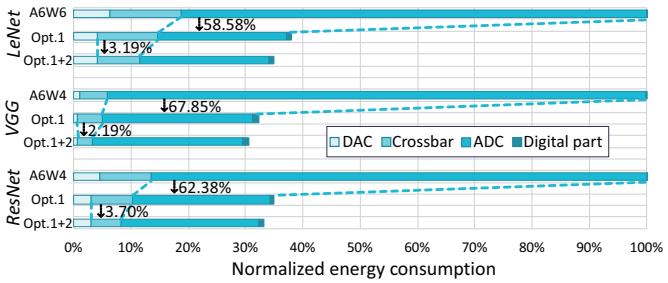

Fig. 5. The normalized energy consumption of each part (A6W4(6): existing work with traditional quantization method; Opt.1: optimized by non-uniform quantization scheme; Opt.1+2: optimized by non-uniform quantization scheme and energy-aware weight regularization)

C. Performance Analysis

Figure 5 shows the energy consumption of each part, which is normalized to the results of traditional quantization methods. According to the results, we know that the proposed non-uniform quantization scheme can reduce 65% energy consumption of the entire system. The reduction mainly comes from using low resolution ADCs (i.e., the power of 4-bit ADCs is 54.68% of 8-bit ADCs), which reduce 70% ADC energy consumption. It is worth mentioning that this improvement is not only due to the low power of ADCs, but also to the reduction of calculation cycles after the quantization. Besides, additional MUXes bring $\sim$ 3% extra overhead. The weight regularization method can reduce 35% crossbar computing energy consumption, which accounts for 3% of the total energy.

In a word, the proposed training framework can enhance the energy efficiency by $\sim$3.4$\times$. The equivalent energy efficiency of the computing units (e.g., RRAM computing banks) is 9.02TOPS/W, nearly 2.6 $\sim$ 4.2$\times$ compared with existing work [1] [2] [7].

VII. SUMMARY AND CONCLUSIONS

In this paper, we propose an energy-efficient quantized and regularized training framework, consisting of a PIM-based non-uniform activation quantization scheme and an energy-aware weight regularization. The proposed training framework can reduce the ADC resolution by 2 bits and the analog computing energy by 35%, and therefore improves the energy efficiency up to 3.4$\times$ compared with existing work. The equivalent energy efficiency of the computing units is 9.02TOPS/W.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China (No. 2017YFA0207600), National Natural Science Foundation of China (No. 61832007, 61622403, 61621091), Beijing National Research Center for Information Science and Technology (BNRist), and Beijing Innovation Center for Future Chips. Chen's work was supported by the Beijing Academy of Artificial Intelligence under Grant BAAI2019QN0402.

REFERENCES

[1] P. Chi *et al.*, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, June 2016, pp. 27–39.
[2] Z. Zhu *et al.*, "A configurable multi-precision cnn computing framework based on single bit rram," in *Proceedings of the 56th Annual Design Automation Conference 2019*. ACM, 2019, p. 56.
[3] H. Chen *et al.*, "A >3ghz erbw 1.1gs/s 8b two-sten sar adc with recursive-weight dac," in *2018 IEEE Symposium on VLSI Circuits*, June 2018, pp. 97–98.
[4] B. Nasri *et al.*, "A 700 μw 1gs/s 4-bit folding-flash adc in 65nm cmos for wideband wireless communications," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2017, pp. 1–4.
[5] B. Li *et al.*, "Merging the interface: Power, area and accuracy co-optimization for rram crossbar-based mixed-signal computing system," in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2015, pp. 1–6.
[6] L. Xia *et al.*, "Switched by input: Power efficient structure for rram-based convolutional neural network," in *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2016, pp. 1–6.
[7] Y. Cai *et al.*, "Low bit-width convolutional neural network on rram," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 1–1, 2019.
[8] M. Chang *et al.*, "19.4 embedded 1mb reram in 28nm cmos with 0.27-to-1v read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb 2014, pp. 332–333.
[9] Y. Lin *et al.*, "Demonstration of generative adversarial network by intrinsic random noises of analog rram devices," in *2018 IEEE International Electron Devices Meeting (IEDM)*, Dec 2018, pp. 3.4.1–3.4.4.
[10] D. Lin *et al.*, "Fixed point quantization of deep convolutional networks," in *International Conference on Machine Learning*, 2016, pp. 2849–2858.
[11] J. Choi, Z. Wang, S. Venkataramani, P. I.-J. Chuang, V. Srinivasan, and K. Gopalakrishnan, "Pact: Parameterized clipping activation for quantized neural networks," *arXiv preprint arXiv:1805.06085*, 2018.
[12] A. Shafiee *et al.*, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, June 2016, pp. 14–26.
[13] S. Han, H. Mao, and W. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.
[14] L. Xia *et al.*, "Stuck-at fault tolerance in rram computing systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 102–115, March 2018.
[15] Y. LeCun *et al.*, "Comparison of learning algorithms for handwritten digit recognition," in *International conference on artificial neural networks*, vol. 60. Perth, Australia, 1995, pp. 53–60.
[16] S. Wu *et al.*, "Training and inference with integers in deep neural networks," *arXiv preprint arXiv:1802.04680*, 2018.
[17] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[18] K. Alex *et al.*, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.
[19] K. D. Choo, J. Bell, and M. P. Flynn, "27.3 area-efficient 1gs/s 6b sar adc with charge-injection-cell-based dac," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, Jan 2016, pp. 460–461.