Extended Abstracts of the 2020 International Conference on Solid State Devices and Materials, VIRTUAL conference, 2020, pp97-98

Impact and Solution of Nonlinear Characteristics of Resistive Memory in Analog Computing-in-memory Architecture for Deep Neural Network

Yu-Hsuan Lin, Po-Kai Hsu, Dai-Ying Lee, Ming-Hsiu Lee, Hsiang-Lan Lung, Kuang-Yeu Hsieh, Keh-Chung Wang, and Chih-Yuan Lu

Emerging Central Lab. Macronix International Co., Ltd.

16 Li-Hsin Rd. Hsinchu Science Park, Hsinchu, Taiwan, R.O.C.

TEL: +886-3-5786688 ext. 78024, FAX: +886-3-5789087, Email: sharonlin@mxic.com.tw

Abstract

Computing-in-memory (CIM) with nonvolatile memory (NVM) draws lots of attentions due to its potential of overcoming the von Neumann bottleneck. The non-ideal nonlinear voltage-current characteristic of resistive memory causes severe inference accuracy loss in the analog CIM architecture for deep neural networks. This work demonstrates that using differential ReRAM pairs helps to maintain the inference accuracy of image recognition in CIFAR-10. A novel input limiting mapping technique is proposed to obtain a near ideal output distribution. Applying this method together with BN calibration can increase the inference accuracy by more than 60%.

1. Introduction

Computing-in-memory (CIM) is a potential architecture which is able to overcome the von Neumann bottleneck between the processor and memory. Nonvolatile memory (NVM) such as resistive random access memory (ReRAM), phase change memory (PCM), and NOR/NAND flash are candidates for CIM. Many ReRAM characteristics, including noise [1], retention [1], and gradual switching [2], have been studied for multiplication-and-accumulation (MAC) operation. But the impact of nonlinear current-voltage (I-V) characteristic has not been carefully evaluated. In this work, we discuss the ReRAM nonlinear effect in analog NVM neural network for inference.

2. Memory Device and Neural Network

A 1Mb 1T1R (one-transistor-one-ReRAM) array with WO_x/TiO_x ReRAM [3] was fabricated and characterized with read voltages from 0.1V to 0.7V. The ReRAM device shows nonlinear I-V characteristic that the conductance (G) increases as the V_{read} increases (Fig. 1). ReRAM in high resistance state (HRS) has more serious conductance change than that in low resistance state (LRS). The read current of the WO_x/TiO_x ReRAM device follows the Poole-Frenkel emission (Fig. 2). The voltage related term, a, and the barrier related term, b, as functions of ReRAM conductance state are fitted and simulated (Fig. 3). The predicted read current for different ReRAM conductance states match well with the experimental data (Fig.4).

The nonlinear ReRAM devices were simulated as analog weights in a convolution neural network (CNN) classifier with 6 convolution layers and 3 fully-connected layers for CIFAR-10 image recognition (Fig. 5 and Table I). The inference accuracy of this network with ideal analog weights is 0.904.

3. Results and Discussions

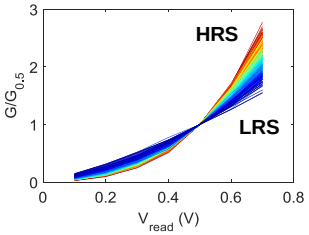
When we replace weights of the first convolution layer (Conv 1) by nonlinear conductances of ReRAM devices, in which a single ReRAM device represents a single weight, the inference accuracy drops dramatically from 0.904 to 0.138. Fig. 6 shows that the output distribution of the Conv 1 MAC with single ReRAM weight carries shallow low bound and wide high bound. The asymmetric distribution is due to ReRAM's nonlinear I-V that the low and high input voltages cause reduced and increased MAC results, respectively. Using two ReRAM devices in differential configuration ($G=G^+-G^-$) [5] for one weight helps to eliminate the asymmetric output distribution (Fig. 6) thus shows a better inference accuracy 0.792, since the positive and negative weights are mirrored (Table II). Fig. 7 shows the differential ReRAM array for MAC operation.

However, applying such differential nonlinear ReRAM pairs to other *single* convolution layer does not provide as significant accuracy improvement as applying to Conv 1 (Table III). This is because the input value of the Conv 1 has a confined and more uniform distribution than those of other Conv layers (Fig. 8). When a broad input range, e.g., Conv 2 input distribution (Fig. 9), maps to input voltage 0 to 1V linearly, most of the differential ReRAM pairs receive low input bias. According to nonlinear ReRAM characteristics, the MAC output distribution is narrower than the ideal output (Fig.10). An input limiting method to confine input distribution is thus proposed to relieve the nonlinear impact, by which the tail of input distribution is set to a boundary value (Fig. 9). Fig. 10 shows that the MAC output distribution changes as we change the value of input limit. Fig. 11 is the flow with input limiting method that the limiting function is applied before mapping the input value to voltage. Fig. 12 shows that an optimized input limit value (related to the weight combination after training) for each *single* ReRAM convolution layer can be obtained to get high inference accuracy. The improvement may be degraded with high input limit because of the insufficient resolution on the low input values. On the other hand, applying too-low input limit would result in reduced inference accuracy because it truncates too much high-valued input information. While implementing ***all*** the 6 convolution layers with nonlinear ReRAM differential weights receives an unacceptable 0.103 inference accuracy, applying the optimized input limits for each layer can effectively recover the inference accuracy back to 0.590 (Table IV). Further combining the proposed techniques with the batch normalization (BN) calibration method [4] on the output distribution in every convolution layer, the inference accuracy can achieve 0.760 for the all-ReRAM convolution CIM array.

4. Conclusions

The simulated nonlinear I-V curves of WO_x/TiO_x ReRAM based on Poole-Frenkel emission match well with the experimental data. Analog differential ReRAM pairs for weights in the CNN classifier with 6 convolution layers and 3 fully-connected layers has balanced MAC results. The proposed input limiting mapping method relieves the influence of the nonlinear ReRAM. By using this methodology combining with the previously reported BN calibration for every convolution layer, the inference accuracy can be substantially improved.

References [1] Y.-H. Lin *et al.*, *IEEE Trans. Electron Devices* **66** (2019) 1289. [2] C.-C. Chang *et al.*, *IEDM* (2017). [3] C.-H. Wang *et al.*, *SSDM* (2018). [4] D.-Y. Lee *et al.*, *SSDM* (2019). [5] G. W. Burr *et al.*, *IEEE Trans. Electron Devices* **62** (2015) 3498.



◂ Fig. 1 Conductance change with read voltage (V_{read}) of 1024 ReRAM devices. Taking conductance at 0.5V ($G_{0.5}$) as reference level for every device.

Poole –Frenkel Emission:

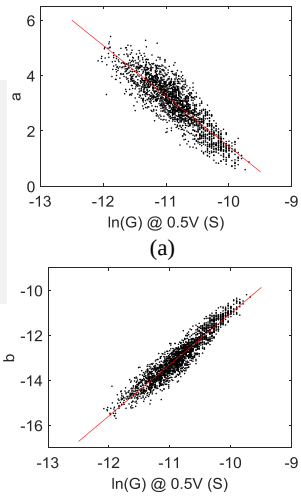
$$J = E \cdot exp\left[\frac{-q(\phi_B - \sqrt{qE/\pi\varepsilon})}{kT}\right]$$

$$\sim V \cdot exp\left(\frac{q\sqrt{q/\pi\varepsilon d}}{kT}\sqrt{V} - \frac{q\phi_B}{kT}\right)$$

$$\rightarrow I = V \cdot exp(a \cdot \sqrt{V} + b)$$

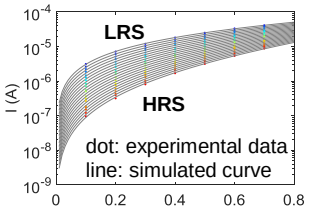
$$\begin{cases} a = -1.82 \cdot \ln G_{0.5} - 16.81 \\ b = 2.28 \cdot \ln G_{0.5} + 11.83 \end{cases}$$

Fig. 2 Poole-Frenkel Emission and the fitted conductance-dependent parameters. J is current density. E is electric field. T is temperature. φ_B is barrier. ε is the permittivity. d is dielectric thickness.



▴ Fig. 3 The Poole-Frenkel parameter (a) a and (b) b are linear functions of logarithm of ReRAM conductance at 0.5V.

◂ Fig. 4 The simulated current of different ReRAM states based on Poole-Frenkel emission.



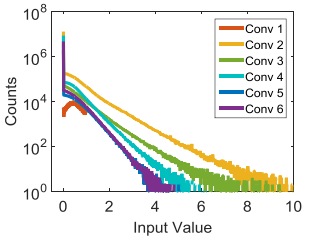


Fig. 8 Ideal input neuron distributions of six convolution layers.

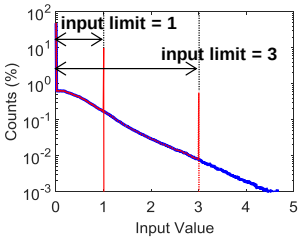


Fig. 9 Conv 2 input neuron distribution. In the input limiting method, the input values outside the limiting range will be set to the boundary.

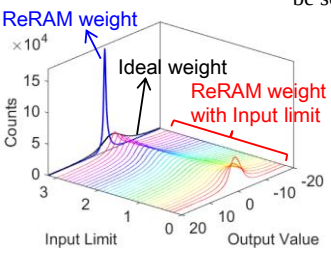


Fig. 10 Conv 2 output distribution with differential ReRAM weights. The output distribution can be modified to the near ideal output distribution by input limiting method.

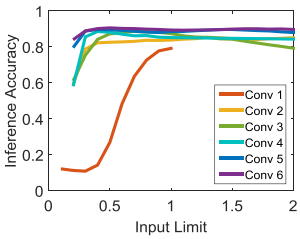


Fig. 12 Inference accuracy change with input limit for replacing single layer to differential ReRAM weights.

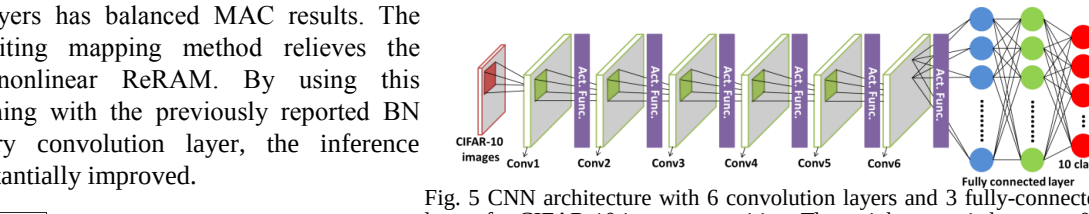


Fig. 5 CNN architecture with 6 convolution layers and 3 fully-connected layers for CIFAR-10 image recognition. The weight range is between -0.2 and 0.2 [4]. For ReRAM-based CIM, the minimum and maximum ReRAM conductance (G_{min} and G_{max}) weights in this work are $2.8x10^{-6}$S and $2.8x10^{-5}$S, respectively. The input voltage is from 0 to 1V.

	Conv 1	Act. Func.	Conv 2	Act. Func.	Conv 3	Act. Func.	Conv 4	Act. Func.	Conv 5	Act. Func.	Conv 6	Act. Func.	FC 1	Act. Func.	FC 2	Act. Func.	FC 3
Filter	128	BN	128	BN	256	BN	256	BN	512	BN	512	BN		BN		BN	
Neuron		ReLU		max pooling		ReLU		max pooling		ReLU		max pooling	1024	dropout	1024	dropout	10
				dropout				dropout				dropout		ReLU		ReLU	
				ReLU				ReLU				ReLU					

Table I Detailed CNN model including filter numbers for convolution layers, neuron numbers for fully-connect layers and activation functions.

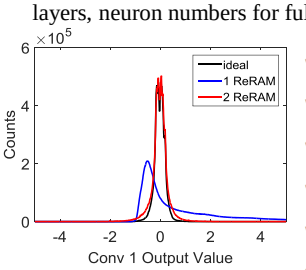


Fig. 6 Output distributions after Conv 1 MAC with ideal weights, ReRAM weights, and differential ReRAM weights in Conv 1.

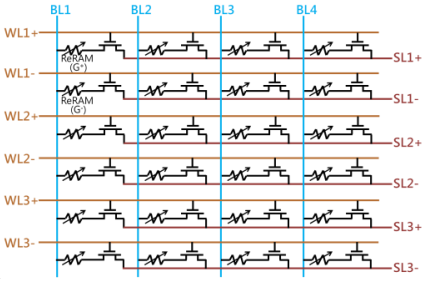


Fig. 7 The schematic diagram of a NOR-type differential ReRAM CIM array.

	W_{min}	→	0	→	W_{max}
G^+	G_{min}	G_{min}	G_{min}	→	G_{max}
G^-	G_{max}	←	G_{min}	G_{min}	G_{min}

Table II Modification of differential ReRAM weight. G^+ is modified from G_{min} to G_{max} and G^- is fixed at G_{min} for positive weight, while the conductance values of G^+ and G^- exchange for negative weight.

Conv	Accuracy
1	0.792
2	0.215
3	0.257
4	0.224
5	0.336
6	0.724

Table III Inference accuracy with differential ReRAM weights in single convolution layer.

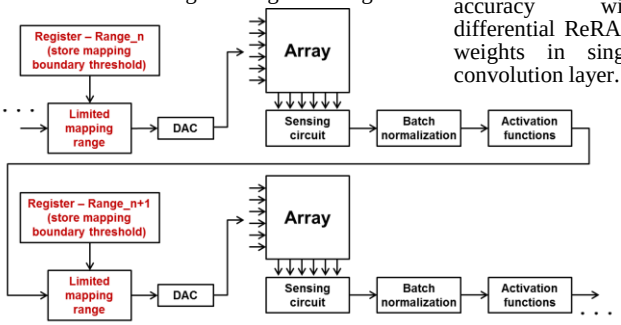


Fig. 11 An input limiting mapping function is added before the DAC of input signal. Registers are required to store the range boundary values

Conv	Input limit	Input limit	+ BN calibration					
1	X	1	V	V	V	V	V	V
2		1.3		V	V	V	V	V
3		0.7			V	V	V	V
4		0.4				V	V	V
5		0.4					V	V
6		0.5						V
Accuracy	0.103	0.590	0.636	0.762	0.759	0.729	0.769	0.760

Table IV The inference accuracy with input limiting method and BN calibration in the ReRAM-based neural network with differential ReRAM weights in six convolution layers. The optimum values of input limit in every layer is chosen from Fig. 12. The inference accuracy achieves 0.760.