Degradation analysis and optimization of temperature effect on MEMRISTOR-based Neural Network Accelerators by electro-thermal simulation

Mengjun Shang¹, Longxiang Yin²,* and Ning Xu¹,*

¹School of Information Engineering, Wuhan University of Technology, Wuhan, China
²State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

*Corresponding author: yinlongxiang@ict.ac.cn, Xuning@whut.edu.cn

Abstract. Nowadays, memristor-based neural network accelerators have been widely studied due to their outstanding performance in massive parallel vector matrix multiplication. However, the memristor is sensitive to temperature and its on/off state operation window can be seriously degraded by the increasing temperature, which may lead to computation failures in memristor-based NN accelerators. In this work, we establish an electro-thermal simulation platform to evaluate the temperature impact on memristor-based NN accelerators. With this platform, we first investigate the impact on computation accuracy with the temperature increase in different NN layers in the accelerators. We then apply a temperature-aware NN weight mapping scheme to the most temperature-sensitive layer and achieve 28.89% improvement in computation accuracy, which only has 0.06% difference with the improvement achieved by applied the mapping scheme to the whole NN model. This finding can help to simplify the temperature-aware hardware optimization design in memristor-based neural network accelerators and reduce the power consumption.

1. Introduction
In recent years, neural network has achieved great success in research and application. However, it is a memory-centric application, which is a great challenge to traditional von Neumann architecture and exacerbate the “memory wall” due to large-scale parameters frequently transferring between memory and the computing unit [1-3]. Recently, memristor, which can simultaneously realize memory and computing, has gained great attention from academia and industry. It is the most promising candidate to end the “memory wall” problem. Memristor-based crossbar structure can perform massive parallel vector matrix multiplication in O(1) time complexity based on Ohm’s Law and Kirchhoff’s Law and are widely used in novel memristor-based NN accelerator designs. In 2016, A. Shafiee et al. [4] proposed a memristor-based NN accelerator prototype “ISaac” with the energy efficiency up to 380.7GOPS /W with a 128x128 crossbar scale, which is 130% higher than DaDianNao [5]. In 2020, P. Yao et al. [6] reported an eight 2048-memristor array integration prototype, which showed an energy...
efficiency more than two orders of magnitude greater than state-of-art Tesla V100 [7].

However, memristor-based NN accelerator may experience severe thermal problem. In 2011, C. Walczyk et al. [8] found in experiment that the on/off state operation window of the memristor decreases to its half at 410K comparing with that at 300K, which is shown in Figure 1. It will lead to data failure and computation errors in the memristor-based NN accelerators.

![Figure 1](image.jpg)

**Figure 1.** The temperature increase results in the reduction of the memristor switching conductance window [8].

In this paper, we analyze and optimize the impact of temperature on memristor-based NN accelerators. The main contributions of the work are as follows:

- We established an electro-thermal simulation platform aimed at thermal-aware analysis of memristor-based circuit and architecture, consisting of a fast temperature distribution solver and the NeurSim simulator.
- We propose a simplified temperature-aware hardware optimization scheme for memristor-based NN accelerators, only by applying temperature-aware NN weight mapping scheme to the most temperature-sensitive layer can we get the same optimized results and effectively reduce the power consumption.

2. Related works

Temperature effect on memristor has begun for a long time. Early works focus on the microscopic mechanism of heating in a single memristor device. In 2009, B. Gao et al. [9] proposed a unified physical understanding for microscopic mechanism in memristor. Later, C. Walczyk et al. [8] confirmed in experiment that high temperature can lead to on-state conductance degradation in memristor.

Recent years, temperature-induced performance degradation in memristor-based NN accelerator gradually gets more attention. In 2018, P. Huang et al. [10] found that the accuracy of image recognition in memristor-based NN accelerator decreases to less than 30% after $10^4$s in 175°C. In the same year, M. V. Beigi et al. [11] found a max temperature 345K in memristor-based NN accelerator when performing online operation. In 2019, Y. Xiang et al. [12] proposed a new unit design of memristor crossbar array to reduce the impact of temperature effect, which leads to a higher hardware cost. M. V. Beigi et al. [11] proposed a temperature-aware mapping scheme which can selectively place the weights that have large impacts on neural network onto memristor crossbar cells with lower temperature and can successfully improve the accuracy in memristor-based NN accelerators. However, the temperature distributions utilized in above mentioned works are either simplified to a uniform
temperature hypothesis or achieved by a sophisticated numerical simulation. A fast and accurate way to achieve temperature distribution is badly needed.

3. Simulation methods and optimization scheme
In this section, we will discuss electro-thermal simulation platform that we proposed and the optimization scheme to get the optimized results.

3.1 The electro-thermal simulation platform
The proposed electro-thermal simulation platform consists of an in-house fast temperature distribution solver and the NeuorSim [13] simulator.

We establish a fast temperature distribution solver based on “Power Blurring” method [11]. The relationship of heat and temperature can be regarded as a linear signal system. The actual temperature distribution can be fast achieved by convoluting the actual heat density distribution and the ideal pulse temperature distribution according to the basic theory of linear signal systems, which is more convenient and almost the same accurate comparing with numerical computation.

The fast temperature distribution solver is shown below:

\[
T(x, y, t) = \frac{P(x,y,t)T_\delta(x,y,t)dt}{\int_0^\tau P_\delta(t)dt}
\]  

(1)

In equation (1), \(T_\delta(x, y, t)\) is the temperature at the position \((x, y)\) under ideal pulse heat density distribution in the memristor-based NN accelerator; \(P(x, y, t)\) represents the actual heat density at the position \((x, y)\) at the time \(t\); \(P_\delta(t)\) denotes the total heat value under ideal pulse heat density distribution at time \(t\); \(\tau\) is the relaxation time to reach the steady-state temperature distribution. We can obtain the final actual steady-state temperature distribution when \(t\) on the left is equal to \(\tau\).

![Figure 2. The test results of fast temperature distribution model.](image1)

![Figure 3. A snapshot of the thermal distribution of a part of memristor-based NN accelerator.](image2)

In order to verify the accuracy of the fast temperature distribution solver, we compared it with the numerical computation results generated by ANSYS [14]. We simulate a 5x5 memristor-based NN accelerator. In Figure 2 we compare the temperature distribution achieved by our solver and ANSYS. We can find that the results are almost the same and the maximum error is less than 0.2%, which verifies the correctness of our fast temperature distribution solver. We can get the temperature distribution [15] of large-scale memristor crossbar array fast. In Figure 3, we show a snapshot of the
temperature distribution in the 5x5 memristor-based NN accelerator.

NeuroSim [13] is an open source full-stack NVM-oriented simulator. This platform can perform the neural network simulation process of devices, circuits, algorithms and other aspects. We utilize this simulator to evaluate the degradation of performance in memristor-based NN accelerators caused by temperature effects.

3.2 Temperature-aware NN weight mapping scheme

![Temperature-aware NN weight mapping scheme](image)

**Figure 4.** The principle of temperature-aware NN weight mapping scheme.

Figure 4 shows the principle of temperature-aware NN weight mapping scheme. The weights from the same neuron are mapping onto the same row in the memristor-based NN accelerator. According to the characteristics of neural network, exchanging the positions of neurons in the same layer along with their inputs has no impact on the output of the neural network. We can utilize this feature to exchange the weights on hot cells with that on cold cells and then we can reduce the maximum temperature of a neural network layer, especially for consecutive rows of heat cells, since they are diffused rather than aggregated. So we can achieve the target of optimizing the temperature effect based on the memristor neural network accelerator.

4. Experimental results

In this section, we will study the temperature effect in a memristor-based NN accelerator mapping a two-layered MLP model and achieve some general disciplines related with temperature-induced degradation in memristor-based NN accelerators. We achieved improved accuracy in offline training which [11] is not considered and less hardware cost compared with [15].

4.1 Temperature-induced degradation difference in different layers

The weight value of the first layer in MLP was 40,000 and 1000 in second layer. Obviously, there were more data in the first layer and more content to be extracted for features, which will have an impact on the temperature-induced degradation. So we guess that in MLP, temperature had a greater impact on
the weight of the first layer than the second. We add a uniform temperature distribution onto different layers respectively, ranging from 310K to 390K, to observe the accuracy variation in different layers. The results are shown in Figure 5. We can see that the accuracy is not significantly decreased with increasing temperature only happened in the second layer. On the other hand, when temperature increasing results only in the first layer, show that the accuracy slowly decreases, and starts to drop sharply when the temperature rises to 340K, and the accuracy stabilizes at only 10% when the temperature reaches 360K. Therefore, we can conclude that the temperature variation in the first layer of the two-layered MLP has a great impact on the MLP output result.

4.2 Inference accuracy of the proposed scheme

There are four results we can get. As shown in the Figure 6, the computation accuracy of the two-layered MLP in memristor-based neural network accelerator without temperature-induced degradation (the ideal case) is 94.79%. However, when considering the temperature effect, the computation accuracy drops to only 44.82%. We then perform the temperature-aware NN weight mapping scheme in the memristor-based neural network accelerator to reduce the temperature effect. We compare the results of performing the mapping scheme only onto the first layer and onto the whole MLP model, which respectively improves the computation accuracy to 73.71% and 73.77%. It has the accuracy improvement of 28.89% and 28.95%, which effectively improves the calculation accuracy of the system. And there is almost no difference in results of the two situations. The mapping scheme onto the second layer has little improvement in the computation accuracy and unsatisfactorily increases the total power consumption. This finding can help us simplify the temperature-aware hardware optimization design in memristor-based neural network accelerators.

Figure 5. The influence of temperature on the accuracy of the two-layered MLP.

Figure 6. Comparison of the temperature-induced accuracy degradation in different situations.

5. Conclusion

In this work, we investigate the impact of temperature to computation accuracy of memristor-based accelerator. We propose the electro-thermal simulation platform to quickly get the real temperature distribution in the system. And find that accuracy degradation of the memristor-based neural network
accelerator is mainly influenced by the front layers in the MLP model. This finding can help to simplify the temperature-aware hardware optimization design in memristor-based neural network accelerators.

Reference
[1] S.H. Jo, T. Chang, I. Ebong, et al. "Nanoscale Memristor Device as Synapse in Neuromorphic Systems," Nano Letters, vol. 10, no. 4, pp. 1297-1301, 2010.
[2] S. Park, A. Sheri, J. Kim, et al. "Neuromorphic speech systems using advanced ReRAM-based synapse," In Proc. of 2013 IEEE IEDM, 2013, pp. 25.6.1-25.6.4.
[3] L. Gao, I-T. Wang, P-Y. Chen, et al. "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," Nanotechnology, vol. 26, no. 45, pp. 455204, 2015.
[4] A. Shafiee, A. Nag, N. Muralimanohar, et al. "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars," In Proc. of 2016 ISCA, 2016, pp. 14-26.
[5] Y. Chen, T. Luo, S. Liu, et al. "DaDianNao: A Neural Network Supercomputer," In Proc. of 2017 MICRO, 2017, pp. 609-622.
[6] P. Yao, H. Wu, B. Gao, et al. "Face classification using electronic synapses," Nature Communications, vol. 8, pp. 15199, 2017.
[7] S. Yu, "Neuro-Inspired Computing With Emerging Nonvolatile Memory," Proceedings of the IEEE, vol. 106, no. 2, pp. 260-285, 2018.
[8] C. Walczyk, D. Walczyk, T. Schroeder, et al. "Impact of Temperature on the Resistive Switching Behavior of Embedded-Based MEMRISTOR Devices," IEEE Transactions on Electron Devices, vol. 58, no. 9, pp. 3124-3131, 2011.
[9] B. Gao, B. Sun, H. Zhang, L. Liu, X. Liu, R. Han, J. Kang and B. Yu, "Unified Physical Model of Bipolar Oxide-Based Resistive Switching Memory," IEEE Electron Device Letters, vol. 30, no. 12, pp. 1326-1328, 2009
[10] P. Huang, Y.C. Xiang, Y.D. Zhao, et al. "Analytic Model for Statistical State Instability and Retention Behaviors of Filamentary Analog MEMRISTOR Array and Its Applications in Design of Neural Network," In Proc. of 2018 IEEE IEDM, 2018, pp. 40.4.1-40.4.4.
[11] M.V. Beigi and G. Memik, "Thermal-aware Optimizations of ReRAM-based Neuromorphic Computing Systems," In Proc. of 2018 DAC, 2018, pp. 1-6.
[12] Y. Xiang, P. Huang, Y. Zhao, et al. "Impacts of State Instability and Retention Failure of Filamentary Analog MEMRISTOR on the Performance of Deep Neural Network," IEEE Transactions on Electron Devices, vol. 66, no. 11, pp. 4517-4522, 2019.
[13] P.-Y. Chen et al., "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," in IEDM, 2017.
[14] [on line] ANSYS, https://www.ansys.com/
[15] X. Liu, M. Zhou, T.S. Rosing and J. Zhao, "HR3AM: A Heat Resilient Design for MEMRISTOR-based Neuromorphic Computing," In Proc. of 2019 ISLPED, 2019, pp. 1-6.