

# 校准方法和存内训练相结合的忆阻器神经形态计算方法\*

杜湘瑜, 彭杰, 刘海军

(国防科技大学电子科学学院, 湖南长沙 410073)

**摘要:**基于忆阻器的神经形态计算架构在图像分类、语音识别等领域取得了较好的效果,但当忆阻器阵列存在低良率问题时,其性能会出现明显下降。提出一种基于忆阻器神经形态计算的校准方法和原位训练相结合的算法,利用校准方法提高乘累加计算的准确率,并利用原位训练方法降低训练误差。为了验证所提方法的性能,采用多层感知器架构进行仿真。从仿真结果来看,神经网络的精度有明显的提高(近40%)。实验结果表明,与单纯的校准方法相比,采用所提方法训练的网络精度提高了约30%,与其他主流的方法相比,所提方法训练的网络精度提高了0.29%。

**关键词:**忆阻器;神经形态计算;原位训练;校准方法

中图分类号:TP35 文献标志码:A 开放科学(资源服务)标识码(OSID):

文章编号:1001-2486(2023)05-202-05



听语音  
与作者  
互动  
聊科研

## Memristive neuromorphic computing approach combining calibration method and in-memory training

DU Xiangyu, PENG Jie, LIU Haijun

(College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China)

**Abstract:** Memristor based neuromorphic computing architecture has achieved good results in image classification, speech recognition and other fields, but when the memristor array has the problem of low yield, the performance declines significantly. A method combining memristive neuromorphic computing based calibration method with in-situ training was proposed, which increased the accuracy of multiplicative accumulation calculation by using the calibration method and reduced the training error by using the in-situ training method. In order to verify the performance of the proposed method, a multi-layer perceptron architecture was used for simulation. From the simulation results, the accuracy of the neural network is improved obviously (nearly 40%). Experimental results show that compared with the single calibration method, the precision of the network trained by the proposed method is improved by about 30%, and the precision of the network trained by the proposed method is improved by 0.29% when compared with other mainstream methods.

**Keywords:** memristor; neuromorphic computing; in-situ training; calibration method

近年来,深度学习技术发展迅速,尤其是在神经网络领域<sup>[1]</sup>。深度学习算法通常采用CPU或GPU进行训练<sup>[2]</sup>,这也带来了很大的功耗问题<sup>[3]</sup>。随着超大规模集成电路制造业的快速发展<sup>[4-5]</sup>,嵌入式系统逐渐成为研究的热点,它能够在功耗和占用面积非常小的情况下解决目标检测问题<sup>[6]</sup>。

在嵌入式神经形态处理系统的发展中<sup>[7]</sup>,基于忆阻器的神经形态计算体系架构引起了许多研究者的关注<sup>[8-9]</sup>。一些基于忆阻器的神经形态计算体系架构已被提出<sup>[10-11]</sup>。例如,Sun等<sup>[12]</sup>提出了一种基于忆阻器的神经形态计算架构,该架

构实现了三层全并行卷积神经网络(full-parallel convolutional neural networks, FP-CNN)。Sun等<sup>[13]</sup>在FP-CNN计算体系架构的基础上,以其为基本计算单元(basic computing units, BCUs)提出了一种级联神经网络体系架构。Yakopcic等<sup>[14]</sup>提出了一种基于忆阻器的卷积神经网络,该网络使用多个忆阻器交叉阵列。上述方法在网络架构方面取得了很好的效果,但当忆阻器阵列存在低良率问题时,其性能会出现明显下降<sup>[15-16]</sup>。

对此,本文提出了一种基于忆阻器的神经形态计算方法,将能够提高忆阻器阵列乘累加计算准确率的校准方法与可减少训练误差的原位训练

\* 收稿日期:2021-06-16

基金项目:国家自然科学基金资助项目(62074166,61974164,61804181,62004219,62004220)

作者简介:杜湘瑜(1976—),女,辽宁沈阳人,教授,博士,E-mail:xiangyu\_du@163.com;

刘海军(通信作者),男,山东德州人,副教授,博士,硕士生导师,E-mail:liuhaijun@nudt.edu.cn



低良率忆阻器阵列中实现的神经网络的识别率。算法流程如图 4 所示。

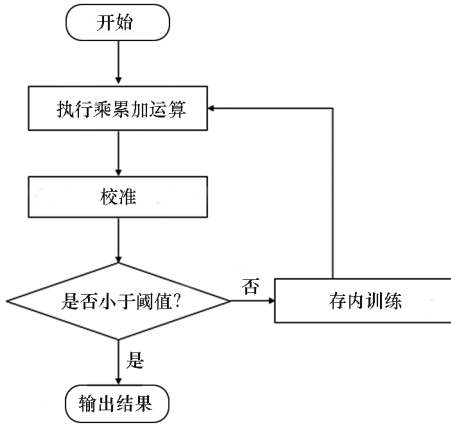


图 4 组合算法的执行流程

Fig.4 Execution flow chart of the combined algorithm

首先,使用忆阻器阵列进行乘累加计算,并使用 1.1 节校准方法进行校准,以获得校准输出。定义一个“阈值”(threshold)来确定使用校准算法后阵列的标准输出与原始输出之间的差值是否大于预期。阈值的确定与实际器件可支撑的擦写次数、网络精度有关。阈值越小,实现的网络精度会提高,对应的网络精度调整越细,训练时间会越来越长,网络训练过程中的权值更新越频繁,器件的实际擦写次数越多,对器件的要求就越高;反之,阈值越大,实现的网络精度会降低,对应的网络精度调整越粗,网络训练过程中的权值更新次数越低,对器件的要求也越低。在本文的仿真实验中,综合考虑器件可支撑的擦写次数、网络精度等因素,阈值被设为  $10^{-2}$ ,所采用的数据集经过 200 个 epoch 的训练,能够最终收敛。为了保证网络能够收敛,需要优选网络的初始化权值等参数,并综合考虑实际器件可支撑的擦写次数、网络精度等因素的影响合理设置阈值。如果校准输出与原始输出的差值大于“阈值”,则进行原位训练,训练后再进行校准验证,直到误差达到要求的精度为止。该算法的执行流程总结如下:

1) 根据公式  $V_{out} = V_{in} \times G \times R_s$ ,使用忆阻器阵列执行乘累加运算以获得模拟输出,其中,  $V_{out}$  为阵列输出电压向量,  $V_{in}$  为阵列输入电压向量,  $G$  为忆阻器阵列电导,  $R_s$  为运放电阻值。

2) 执行 1.1 节校准算法,根据图 1 得到校准后的输出  $\hat{V}_{out} = V_{out}^{STD} + \hat{V}'_{out} - \delta'$ ,即标准输出的校准值。

3) 确定校准输出和标准输出之间的差值是否大于阈值。如果差异大于“阈值”,进行原位训练,然后回到第一步。

4) 输出最终的计算结果。

需要说明的是,上述方法中第三步的阈值设定与实际器件可支撑的擦写次数、网络精度有关。阈值仅体现在训练过程中,仅影响训练过程的参数更新,对识别过程没有影响。本文方法将能够提高忆阻器阵列乘累加计算准确率的校准方法与可减少训练误差的原位训练方法相结合,从而提高网络在低良率阵列中的识别率。

## 2 仿真验证

### 2.1 仿真设置

使用 Tensorflow<sup>[18]</sup> 开源库进行实验,并使用 LTspice 进行电路仿真。为了更好地反映算法的实际应用性能,不使用任何数据增强技术<sup>[18-20]</sup>。选择 MNIST 数据集<sup>[21]</sup>进行验证,将每幅图像归一化为 0~1 的灰度值。本文采用 Ti/AlO<sub>x</sub>/TaO<sub>x</sub>/Pt<sup>[15]</sup> 忆阻器进行仿真实验。忆阻器阵列可以重复编程到从 1 kΩ 到 12 kΩ 的不同状态。在实验中,校准算法的误差阈值设置为  $10^{-2}$ 。

### 2.2 性能仿真

使用一个  $784 \times 100 \times 10$  的多层感知机 (multi-layer perceptron, MLP)<sup>[22]</sup> 来验证所提方法的性能。图 5 显示了 MLP 在两层忆阻器阵列上的实现。权重矩阵  $W_1$  被映射到一个  $784 \times 200$  的数组,矩阵  $W_2$  被映射到一个  $100 \times 20$  的数组。

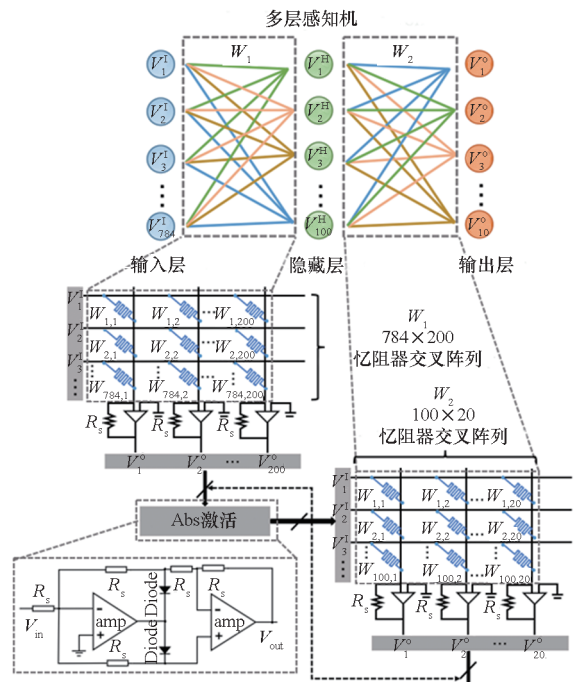


图 5 MLP 在两层忆阻器阵列上的实现

Fig.5 Implementation of MLP on a two-layer memristor array



需要说明的是,激活函数采用的是 Abs 激活函数,原因在于 Abs 激活函数在浅层神经网络中有助于提高识别率<sup>[12]</sup>。在 Abs 激活函数后,第一个数组的输出被视为第二个数组的输入。

在模拟实验中,本文考虑了器件的波动性和阵列的良率。在良率方面,本文主要考虑单比特错误 (single-bit failure, SBF)<sup>[23]</sup>。具体来说,当 SBF 发生时,忆阻器固定在高阻状态。

阻值变化遵循对数正态分布<sup>[24]</sup>。另一方面,器件的波动性可以表示为  $\tilde{G}_{ij} \leftarrow G_{ij} \cdot \exp(\theta_{ij})$ ;  $\theta_{ij} \sim N(0, \sigma^2)$ 。其中,  $\sigma$  表示实验中波动性变化的程度,  $\theta_{ij}$  表示阻值波动性的变化。在本次仿真实验中,  $\sigma$  的范围为 0.6 ~ 1.0。

图 6 展示了本文算法在 MNIST 数据集上的识别率。图中的识别率均为 100 次实验的平均识别率。在实际中,阵列同时受波动性与良率共同影响,因此本实验采用坏块比例的方法进行仿真。坏块比例 (defect percent) 考虑了良率与波动性两个因素。若当前的坏块比例为  $\alpha$ , 则受良率影响的阵列比例为  $\alpha \times \beta$ , 受器件波动性影响的阵列比例为  $\alpha \times (1 - \beta)$ 。在本文的仿真实验中,  $\beta$  被设为 0.5。受波动性和良率影响的器件的位置随机产生,其中波动性系数  $\sigma$  在 0.6 ~ 1.0 之间随机产生。

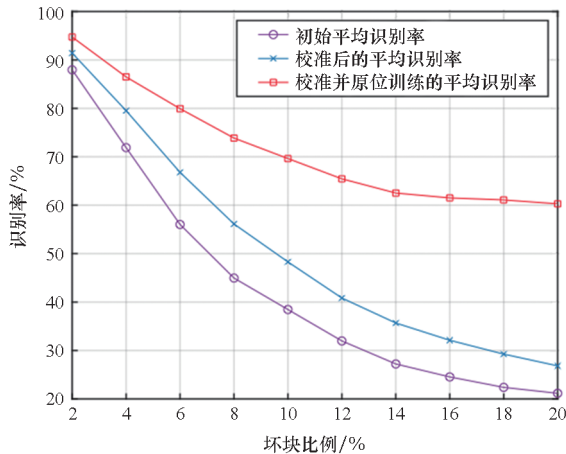


图 6 算法在 MNIST 数据集上的识别率

Fig. 6 Performance of the algorithms on the MNIST dataset

从图 6 中可以看出,采用原位训练和校准算法相结合的方法对网络进行训练后,识别率明显提高。例如,当坏块比例达到 10% 时,不进行任何校准的网络精度只能达到 38.3%,使用校准算法的网络精度可以达到 48.5%,使用校准算法与原位训练相结合的网络精度可以达到近 70%。在最坏的情况下,在坏块比例达到 20% 的情况

下,网络识别性能从 21.4% 提高到了 60%。

本文还将这项工作与文献[24]的方法进行了比较,后者对神经网络进行了再训练,以提高精度。表 1 展示了两种方法在 MLP 上的性能 ( $784 \times 256 \times 10$ )。假设器件损坏只存在于权矩阵  $W_1$  和  $W_2$  中,或者同时存在于  $W_1$  和  $W_2$  中。可以看出,文献[24]可以获得 92.93% 的识别率,本文工作比文献[24]的识别率高出 0.29%。

表 1 本文算法同主流方法的识别率对比

Tab.1 Recognition rate comparison of the proposed algorithm and the mainstream methods

方法	MLP( $W_1$ )	MLP( $W_2$ )	MLP( $W_1 \& W_2$ )
文献[24]	94.88	97.43	92.93
本文方法	94.93	97.66	93.22

### 3 结论

本文提出了一种基于忆阻器的神经形态计算方法,将校准方法与原位训练相结合。通过 MLP 架构的仿真验证了该方法的有效性。从以上的仿真结果来看,神经网络的精度有明显的提高(近 40%)。实验结果表明,与不进行校准的方法相比,该方法的网络精度提高了约 30%。与现有方法相比,该方法的精度比现有方法提高了 0.29%。

### 参考文献 (References)

- [1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.
- [2] DANG D, DASS J, MAHAPATRA R. ConvLight: a convolutional accelerator with memristor integrated photonic computing [C]//Proceedings of 2017 IEEE 24th International Conference on High Performance Computing (HiPC), 2017.
- [3] ANKIT A, SENGUPTA A, ROY K. TraNNsformer: neural network transformation for memristive crossbar based neuromorphic system design [C]//Proceedings of 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2017.
- [4] ANKIT A, SENGUPTA A, PANDA P, et al. RESPARC: a reconfigurable and energy-efficient architecture with memristive crossbars for deep spiking neural networks [C]//Proceedings of the 54th Annual Design Automation Conference (DAC), 2017.
- [5] 陈长林, 骆畅航, 刘森, 等. 忆阻器类脑计算芯片研究现状综述 [J]. 国防科技大学学报, 2023, 45(1): 1-14. CHEN C L, LUO C H, LIU S, et al. Review on the memristor based neuromorphic chips [J]. Journal of National University of Defense Technology, 2023, 45(1): 1-14. (in

- Chinese)
- [6] PARK S, NOH J, CHOO M L, et al. Nanoscale RRAM-based synaptic electronics: toward a neuromorphic computing device[J]. *Nanotechnology*, 2013, 24(38): 384009.
- [7] JANG B C, NAM Y Y, KOO B J, et al. Memristive logic-in-memory integrated circuits for energy-efficient flexible electronics [J]. *Advanced Functional Materials*, 2018, 28(2): 1704725.
- [8] JO S, SUN W, KIM B, et al. Memristor neural network training with clock synchronous neuromorphic system [J]. *Micromachines*, 2019, 10(6): 384.
- [9] CHEN C, LI L X, PENG H P, et al. Fixed-time synchronization of inertial memristor-based neural networks with discrete delay [J]. *Neural Networks*, 2019, 109: 81–89.
- [10] PARASTESH F, JAFARI S, AZARNOUSH H, et al. Chimera in a network of memristor-based Hopfield neural network[J]. *The European Physical Journal Special Topics*, 2019, 228: 2023–2033.
- [11] WEN S P, WEI H Q, YAN Z, et al. Memristor-based design of sparse compact convolutional neural network [J]. *IEEE Transactions on Network Science and Engineering*, 2020, 7(3): 1431–1440.
- [12] SUN S Y, LI Z W, LI J W, et al. A memristor-based convolutional neural network with full parallelization architecture [J]. *IEICE Electronics Express*, 2019, 16(3): 20181034.
- [13] SUN S Y, XU H, LI J W, et al. Cascaded architecture for memristor crossbar array based larger-scale neuromorphic computing[J]. *IEEE Access*, 2019, 7: 61679–61688.
- [14] YAKOPCIC C, ALOM M Z, TAHA T M, et al. Extremely parallel memristor crossbar architecture for convolutional neural network implementation [C]//*Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2017.
- [15] SUN Y, XU H, WANG C, et al. A Ti/AIO<sub>x</sub>/TaO<sub>x</sub>/Pt analog synapse for memristive neural network [J]. *IEEE Electron Device Letters*, 2018, 39(9): 1298–1301.
- [16] SUN S Y, XU H, LI J W, et al. Cases study of inputs split based calibration method for RRAM crossbar [J]. *IEEE Access*, 2019, 7: 141792–141800.
- [17] LI C, BELKIN D, LI Y N, et al. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks[J]. *Nature Communications*, 2018, 9(1): 2385.
- [18] ABADI M, BARHAM P, CHEN J, et al. Tensorflow: a system for large-scale machine learning [C]//*Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI)*, 2016: 265–283.
- [19] WANG J, PEREZ L. The effectiveness of data augmentation in image classification using deep learning [C/OL]//*Proceedings of Computer Vision and Pattern Recognition*, 2017.
- [20] TAKAHASHI R, MATSUBARA T, UEHARA K. Data augmentation using random image cropping and patching for deep CNNs[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, 30(9): 2917–2931.
- [21] WONG S C, GATT A, STAMATESCU V, et al. Understanding data augmentation for classification: when to Warp? [C]//*Proceedings of International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2016.
- [22] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278–2324.
- [23] YAKOPCIC C, TAHA M T. Memristor crossbar based implementation of a multilayer perceptron [C]//*Proceedings of 2017 IEEE National Aerospace and Electronics Conference (NAECON)*, 2017.
- [24] LIU C C, HU M, STRACHAN J P, et al. Rescuing memristor-based neuromorphic design with high defects [C]//*Proceedings of the 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2017.