Tolerance of intrinsic device variation in fuzzy restricted Boltzmann machine network based on memristive nano-synapses

To cite this article: Teng Zhang et al 2017 Nano Futures 1 015003

View the article online for updates and enhancements.

Related content
- High-speed true random number generation based on paired memristors for security electronics
  Teng Zhang, Minghui Yin, Changmin Xu et al.
- Metal oxide resistive random access memory based synaptic devices for brain-inspired computing
  Bin Gao, Jinfeng Kang, Zheng Zhou et al.
- 3D Ta/TaOx/TiO2/Ti synaptic array and linearity tuning of weight update for hardware neural network applications
  I-Ting Wang, Chih-Cheng Chang, Li-Wen Chiu et al.
Tolerance of intrinsic device variation in fuzzy restricted Boltzmann machine network based on memristive nano-synapses

Teng Zhang¹, Minghui Yin¹, Xiayan Lu¹,², Yimao Cai¹,³, Yuchao Yang¹,³,⁴ and Ru Huang¹,³,⁴

¹ Key Laboratory of Microelectronic Devices and Circuits (MOE), Institute of Microelectronics, Peking University, Beijing 100871, People’s Republic of China
² Academy of Advanced Interdisciplinary Studies, Peking University, Beijing 100871, People’s Republic of China
³ Innovation Center for Micro/NanoElectronics and Integrated System, Beijing 100871, People’s Republic of China
⁴ Authors to whom any correspondence should be addressed.

E-mail: yuchaoyang@pku.edu.cn and ruhuang@pku.edu.cn

Keywords: neuromorphic computing, device variation, memristor, fuzzy restricted Boltzmann machine, electronic synapses

Supplementary material for this article is available online

Abstract

Inspired by the architecture and principle of the human brain, neuromorphic computing has attracted enormous research interest due to its potential for massively parallel and energy-efficient computing, where nanoscale memristors are considered as perfect building blocks for hardware neural networks, serving as compact, analog synapses. However, the inherent variation in memristors has been regarded as a major obstacle to their practical application in neuromorphic computing. Here, for the first time, we demonstrate that this long-standing issue can be addressed by introducing fuzziness into the neural networks. We found that the cycle-to-cycle and device-to-device conductance variations in the on and off states of Pt/TaOx/Ta memristors statistically follow Gaussian distributions, and using an experimentally verified compact synapse model based on the electrical characteristics of Pt/TaOx/Ta devices, a fuzzy restricted Boltzmann machine (FRBM) network was constructed where all the weight states were fuzzified to accommodate device stochasticity. The FRBM network has shown significantly improved tolerance to device variation, as confirmed by increased accuracy in the benchmark test of MNIST handwritten digit classifications. This study thus provides a new route towards highly robust neuromorphic computing, even if the computing elements can be stochastic and inhomogeneous.

1. Introduction

Brain-inspired neuromorphic computing has provided new impetus to overcome the fundamental limits in conventional von Neumann computers in terms of intelligence, scalability, and power efficiency and leads to a massively parallel and energy-efficient computing paradigm [1]. As the fourth passive circuit element proposed in 1971 [2] and experimentally confirmed in 2008 [3], memristors have been considered as perfect candidates for synaptic devices in neuromorphic systems [4–7]. On the one hand, brain-inspired algorithms rely heavily on vector-matrix multiplications for the calculation of neuronal outputs, while high-density memristor crossbar arrays naturally offer the ability to perform such vector-matrix multiplication in parallel in the analog domain with extremely low power. On the other hand, the conductance of the memristor physically acts as the connection weight of the synapse that can be facilely tuned by applying voltage pulses, therefore offering great potential for adaptive systems with online learning capability [4, 8]. These desirable properties make memristors well-suited for electronic synapses in neuromorphic hardware and thus are widely exploited nowadays [4, 9–13]. However, it has been realized that nanoscale memristors suffer from large device variability that originates from intrinsic stochasticity during the ion migration as well as filament formation/dissolution processes [14–20]. As a matter of fact, the emergence of variation is a common phenomenon as the dimension of electronic/ionic devices approaches the nanoscale regime [21, 22], and therefore it is a generalized concern not only for memristive synapses but for any other inhomogeneous and stochastic neuronal or synaptic elements such as...
CMOS and phase-change devices. Although it is believed that neural network algorithms themselves are naturally capable of tolerating certain levels of device stochasticity [9], the existence of large variations will still inevitably deteriorate the network performance [9, 10, 23]. This has become a serious bottleneck that must be addressed before the practical application of memristor-based neuromorphic hardware can become a reality.

In this study, we demonstrate a new strategy to tackle such device stochasticity via introducing fuzziness into neural networks, that is, by fuzzifying the device parameters governing the learning processes. The cycle-to-cycle (C2C) and device-to-device (D2D) variations in on- and off-state conductance of Pt/TaOx/Ta devices fabricated in this work were systematically examined and found to follow Gaussian distributions. In consideration of such variability, we constructed a fuzzy restricted Boltzmann machine (FRBM) network based on an experimentally verified memristor model, using fuzzified weights to accommodate the variation in synaptic connections. The performances of the FRBM and traditional restricted Boltzmann machine (RBM) network were compared, and the results unambiguously demonstrated the ability of the FRBM network to tolerate variations in memristive synapses. Such fuzzified network could thus serve as a generalized approach to building neuromorphic systems that are immune to intrinsic device stochasticity, and therefore greatly facilitate the development of intelligent and energy-efficient computing paradigms.

2. Experimental details

The Pt/TaOx/Ta memristive devices in this work were fabricated on SiO2 (300 nm)/Si substrates, where a recessed bottom electrode structure was employed, as schematically shown in the inset of figure 1(a). Such cell geometry is favorable for enhancing the uniformity of electric field distribution inside the device structure and thus improving the device performance [24]. The fabrication process starts from the preparation of SiO2 trenches with a depth of ~50 nm, which were achieved by photolithography and reactive ion etching (RIE). Using the same photoresist pattern as the mask, Ta bottom electrodes were then prepared by magnetron sputtering, followed by a lift-off process. Subsequently, the Ta electrodes were subjected to an oxidation step in oxygen atmosphere at ~400 °C, hence creating a TaOx layer on top of the electrodes with a TaOx thickness of ~12 nm. Afterwards, Pt top electrodes placed perpendicularly to the Ta bottom electrodes were prepared by photolithography and lift-off processes. In the end, photolithography and RIE were performed to etch TaOx and hence expose the contact pads of the Ta electrodes. The size of the Pt/TaOx/Ta devices in this work was 2 × 2 μm².

In order to characterize the switching behavior and device variations of the Pt/TaOx/Ta memristors, electrical characterizations including direct current (DC) and pulse tests were performed using an Agilent B1500A semiconductor parameter analyzer together with a Cascade probe station. The electrical stimuli were applied on the Ta electrode, with the Pt electrode grounded in all the DC and pulse measurements.

Figure 1. Resistance switching characteristics of Pt/TaOx/Ta memristors and an experimentally verified memristor model. (a) DC I–V characteristics of the Pt/TaOx/Ta device showing bipolar switching. The gray curves show the I–V characteristics in 20 cycles and the black line represents the averaged curve. The red curve shows the simulation result with the memristor model. Inset: schematic of the Pt/TaOx/Ta device used in this study. The TaOx layer has a built-in gradient of oxygen vacancy concentration due to the thermal oxidation process. (b) Potentiation/depression measurements on the Pt/TaOx/Ta devices by applying 100 identical 0.8 V 100 ns potentiation pulses and 100 identical −0.93 V 100 ns depression pulses, showing incrementally tunable resistance in repetitive measurements. The device state was read at 0.1 V. The red curve shows the simulation result with the memristor model.
3. Results and discussion

Figure 1(a) shows the current–voltage (I–V) characteristics of the Pt/TaOx/Ta devices, showing typical bipolar resistive switching with a set voltage of around 0.7 V and reset voltage of around ~1.5 V. Such bipolar switching behavior can be attributed to the formation and dissolution of oxygen vacancy-based conducting filaments as disclosed by direct microscopic observations [25, 26], and the switching polarity is in agreement with the existence of built-in oxygen vacancy (VO) concentration gradient inside the TaOx layer, that is, the VO concentration gradually decreases from the Ta electrode to the Pt electrode due to the thermal oxidation process. Moreover, the switching polarity also coincides with the asymmetric electrode configuration in the Pt/TaOx/Ta cell, where Ta can be considered as a reservoir of VO\[27\] compared with the inert nature of Pt. Figure 1(b) shows incremental conductance changes of the device upon application of 100 identical potentiation pulses (0.8 V 100 ns), followed by 100 identical depression pulses (~0.93 V 100 ns). One can see that a large number of incremental conductance states can be accessed, making the Pt/TaOx/Ta devices potential candidates for electronic synapses. Figure 1 also unambiguously shows that device variation is indeed present in both the DC and pulse measurements (gray lines). Note that the above DC and pulse characteristics in Pt/TaOx/Ta devices can all be self-consistently explained and fitted by a compact memristor model, as described below:

\[
I = \alpha (1 - e^{-\beta V}) + w \gamma \sinh(\delta V)
\]

\[
\frac{dw}{dt} = \lambda \sinh(\eta V) \begin{cases} \frac{1 - w}{\tau_0} & (V > 0) \\ \frac{w}{\tau_0} & (V < 0) \end{cases}
\]

Equation (1) describes the I–V relationship of the memristor, where the overall device conduction is composed of two components: Schottky emission in the non-filament region and tunneling conduction in the filament region [28, 29], as described by the first and second terms, respectively. Here w is a state variable representing the overall area fraction of the filament, formed due to the increased concentration of oxygen vacancies upon electrical programming. Equation (2) in turn describes the dynamics of the state variable under the applied signals, that is, the relationship between the change rate of w and the instantaneous applied inputs, where \(\alpha\), \(\beta\), \(\gamma\), \(\delta\), \(\eta\), \(\lambda\), and \(\tau_0\) are all material-dependent parameters. It can be clearly seen from figures 1(a) and (b) that both the DC and the pulse characteristics of the Pt/TaOx/Ta devices can be consistently explained by the same model (red lines), showing that our model successfully captures the essential features of the Pt/TaOx/Ta devices. This therefore allows subsequent construction and simulation of large-scale neural networks based on memristive nano-synapses.

In order to understand the variation of memristive devices, we have also experimentally examined the variations existing in the conductance states of Pt/TaOx/Ta devices. To do this, the device was repeatedly switched for >10⁴ cycles, using relatively strong programming (1.2 V) and erasing (~1.8 V) pulses. In this way, the devices were switched to on and off states during the cyclic switching, so that the conductance distributions around these two target states could be extracted and statistically analyzed. Otherwise, it is difficult to directly perform statistical examination on arbitrary analog states, because the device variation has been found to cause significant overlaps between different analog states. Figures 2(a) and (b) exhibit the probabilistic distributions of the on- and off-state conductance of the Pt/TaOx/Ta device, respectively. It was found that both the on- and off-state conductance generally follows Gaussian distribution, although the fitting in the on state seemed better compared with that in the off state, which may be ascribed to the fact that some intermediate conductance states (with conductance up to about 0.2 mS) were still included on the high–conductance (right) side of figure 2(b). It is therefore reasonable to deduce that similar variations can be expected for arbitrary conductance states, when the memristors operate in the analog domain as electronic synapses. Besides the C2C variation, we have also examined the D2D variation in the Pt/TaOx/Ta devices by testing ~300 cells, and the experimental results show that it also seems to follow Gaussian distribution in both the on and off states (see supplementary figure S1 available at stacks.iop.org/NANOFL/1/015003/mmedia). However, as we are tackling the uncertainty in weight adjustments that could deviate from the training algorithm, C2C variation is more relevant than D2D variation. The impact of the latter is mainly limited to the initialization of the network. We have thus adopted similar Gaussian distributions in the fuzzified weight states in the FRBM network in order to tolerate such device variations (to be discussed in detail afterwards), in contrast to the deterministic weight states that are always considered in conventional neural networks. It is also worthwhile pointing out that the variation in the off state seems much more significant compared with that in the on state. This may be attributed to the fact that the off-state resistance strongly depends on the location of the filament breakage during reset as well as the size of the gap that is formed after the filament dissolution, which are both stochastic and unpredictable processes. In contrast, the overall low resistance caused by filament formation has largely restricted the extent of variation in the on-state resistance, leading to the much smaller spread in figure 2(a).
In order to disclose the influence of device variation on neural network performances and evaluate its role in neuromorphic computing, a regular RBM network was constructed first using the above experimentally verified memristor model. Classical MNIST handwritten digit classification was employed as a benchmark test for the assessment of the network performance, as shown in figure 3, where different levels of variation were introduced into the synaptic weights, and corresponding learning processes as well as recognition rates were simulated and plotted. One can clearly see that the recognition rate constantly decreases as the device variation increases (figure 3), thus showing a detrimental effect on the network training. A large number of previous studies have also revealed similar impacts of device variations on network performances \[9, 10, 23\]. Although it is believed that neural network itself is capable of tolerating certain levels of variation \[9\], it is obvious from figure 3 that large variations will still inevitably deteriorate the network performance. This variation issue has to be addressed in order to facilitate further studies on neuromorphic computing when using memristive nanodevices as synapses.

Here, we put forth a new strategy to tackle such device inhomogeneities via introducing fuzziness into neural networks, that is, by fuzzifying the synaptic weights governing the learning processes. We demonstrate this idea using an FRBM network herein as an example, but the principle is widely applicable to other neural networks. Figure 4 sketches the principle of the FRBM network, where each previously deterministic weight state in conventional neural networks is now replaced by a fuzzy set of weights following Gaussian distributions, as has been experimentally confirmed in figure 2. Using the MNIST data set of handwritten digits, we have trained the FRBM network by a contrastive divergence approach \[30\] for 50 epochs, and each epoch includes 5000 samples divided into 100 minibatches. The network was tested after each training epoch, using new samples that were not

---

**Figure 2.** C2C conductance variation in the on and off states of the Pt/TaOx/Ta devices. (a) Distribution of on-state conductance (shaded area) and fitting results with Gaussian distribution (line). (b) Distribution of off-state conductance (shaded area) and fitting results with Gaussian distribution (line). Some intermediate conductance states (with conductance up to about 0.2 mS) might be included on the high-conductance (right) side, thus affecting the overall fitting in (b).

**Figure 3.** Recognition rate of MNIST data set as a function of variation in device conductance (σ/μ) using the RBM network. The accuracy decreases as device variation increases.
included during the training process. Figure 5 shows the evolution in the error rate of the FRBM network during training (blue triangles), compared with a normal RBM network with perfectly uniform elements where no device variation is considered (black squares) and stochastic elements with variation of $\sigma/\mu = 1$ (red circles). Once again, one can see that the existence of device variation (red circles) deteriorates the network performance via augmenting the error rate by $\sim 10\%$, compared with idealized networks with homogeneous devices (black squares, figure 5). However, when the synaptic weights are fuzzified (blue triangles), even if the same level of device variation exists, the fuzzified network persistently learns during training and the error rate in recognition reaches a comparable level with that of idealized networks (black squares, figure 5) after the FRBM network is trained by $>40$ epochs. This strongly suggests that the device variation has been effectively tolerated by introducing fuzziness into the neural network, therefore showing great prospect of the present approach achieving highly robust computing.

Figure 6(a) further shows the fuzzy weight maps of the network after learning. Notably, each weight state is now accompanied by an error component, which is the most important feature of the fuzzified networks. In conventional unfuzzified networks, the network state in any stage during training is deterministic, and thus any departure from the network state that is targeted by the training algorithms is likely to negatively impact the network performance, which is inevitable when the weight tuning is performed on memristive synapses that are intrinsically stochastic determined by the microscopic mechanism $[14, 15]$. However, when the weight states themselves are fuzzified as we have done here, the network state in any stage is no longer deterministic in the training algorithm but effectively expanded to a collection of network states around the optimal state as targeted.
by the algorithm. As a result, the representation capability of the network model is greatly enhanced, and the uncertainty in the network state during the network training and electrical programming as a result of device variations is likely to have been represented by the network collection [30], which is therefore the origin of the improved tolerance to variations, as demonstrated in figure 6(a). Figure 6(b) further shows that upon inputting different digits into the network, the images can be correctly reconstructed from the trained FRBM network. Likewise, the reconstructed images also include error components due to the errors in the weights. We believe such a fuzzied network proposed here is not limited to specific networks but offers a generalized strategy for tolerating device imperfections, and therefore is significant for neuromorphic computing.

To further confirm the robustness of the FRBM network constructed above, figure 7 shows the recognition rate of the network under different levels of device variation. As can be clearly seen, the accuracy in MNIST handwritten digit classification only slightly decreases as device variation increases, in stark contrast to the pronounced drop of recognition accuracy in figure 3. This once again unambiguously demonstrates the advantage of the fuzzied network proposed above and highlights the prospect of the present approach in achieving robust learning, even if the computing elements themselves can be inhomogeneous and stochastic, like nanoscale memristors.

Figure 6. (a) Weight maps of the FRBM network after training. In this case, each weight state is no longer a constant, but includes an error component. (b) Reconstructed images upon inputs of different handwritten digits after the FRBM network is trained. The reconstructed images also contain error components due to the fuzzy nature of the weight states.

Figure 7. Recognition accuracy of the FRBM network as a function of relative standard dispersion. The accuracy only decreases slightly as the device variation increases.
4. Conclusions

In summary, we have shown a novel fuzzified approach to addressing the long-standing issue of device variations limiting the construction of highly robust neuromorphic computing systems. The C2C and D2D variations in the on- and off-state conductance of Pt/TaOx/Ta devices were found to follow Gaussian distributions, and a compact device model was developed to self-consistently explain the experimental switching characteristics. Using MNIST handwritten digit classification as a benchmark test, the performance of the FRBM network with fuzzified weight states demonstrates significant improvements compared with traditional unfuzzified RBM networks with deterministic weight states, hence unambiguously showing the potential of the fuzzified network in tolerating device variations and achieving highly robust neuromorphic computing. We believe the fuzzified network proposed here could serve as a generalized approach to building neuromorphic systems with great immunity to intrinsic device stochasticity, and will thus contribute to the development of next-generation intelligent, energy-efficient, and fault-tolerant computing paradigms.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61674006, 61421005, 61376087, 61574007) and the Beijing Municipal Science & Technology Commission Program (Z161100000216148). YY acknowledges supports from the ‘1000 Youth Talents Program’ of China.

References