SCIENCE CHINA
Information Sciences

• LETTER •

October 2025, Vol. 68, Iss. 10, 209402:1–209402:2
https://doi.org/10.1007/s11432-024-4430-6

Model compression and quantization optimization methods for memristive neural networks

Yu WANG[1], Xiaobing YAN[2*], Tuo SHI[3], Bo LYU[3], Yincheng QI[1],
Ying LIU[2], Jikang XU[2], Biao YANG[2] & Pengfei LI[2]

[1]*Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China*
[2]*Institute of Life Science and Green Development, Key Laboratory of Brain-like Neuromorphic Devices and Systems of Hebei Province, College of Electronic and Information Engineering, Hebei University, Baoding 071002, China*
[3]*Zhejiang Laboratory, Hangzhou 311122, China*

Received 31 October 2024/Revised 26 December 2024/Accepted 3 May 2025/Published online 2 September 2025

Citation Wang Y, Yan X B, Shi T, et al. Model compression and quantization optimization methods for memristive neural networks. Sci China Inf Sci, 2025, 68(10): 209402, https://doi.org/10.1007/s11432-024-4430-6

Memristors exhibit significant potential for neural network inference acceleration through computing-in-memory (CIM) architectures [1]. However, practical implementations face two fundamental limitations: non-ideal fabrication processes and insufficient hardware resource utilization. This study proposes a ferroelectric memristor-based algorithm-hardware co-optimization strategy (Figure 1(a)) that achieves efficient deployment through structured pruning, weight quantization, and hardware-aware training. The proposed strategy first employs a statistics-based adaptive threshold algorithm to perform row-column bidirectional structured pruning on convolutional kernels, dynamically adjusting layer-wise sparsity thresholds to reduce parameter scale while maintaining model capacity. Subsequently, it leverages the 8-state (3-bit) conductance characteristics of ferroelectric memristors to implement weight quantization, with a dequantization mechanism designed to ensure training stability. Finally, a joint optimization of cross-entropy loss and bit-distribution entropy regularization explicitly constrains weight distribution to enhance robustness against memristor non-ideal characteristics.

Methods. Emerging ferroelectric memristors based on hafnium oxide ((HfO_2)) materials have attracted significant research interest due to their excellent CMOS compatibility and high scalability. In this study, we fabricated Si-doped Si:HfO_2-based ferroelectric memristors with preferred (111) orientation on Si substrates using pulsed laser deposition (PLD) and magnetron sputtering techniques. As shown in Figure 1(b), the devices demonstrate excellent 8-state conductance retention characteristics, with all states maintaining stability for over 8000 s. Device characterization and image processing applications are provided in Appendixes A and B.

Network pruning serves as a fundamental technique for model compression. This study proposes an adaptive pruning method based on statistical characteristics of parameters. As detailed in Appendix C, our approach first computes the L1-norm of convolutional sub-kernel weights as importance metrics $\|W_i\|_1 = \sum_{j=1}^{n} |W_{ij}|$, then dynamically generates pruning thresholds according to layer-wise parameter distributions $T = \mu + \alpha\sigma$, where μ and σ represent the mean and standard deviation of layer weights, respectively, and α is a learnable scaling factor. This row-column bidirectional structured pruning strategy effectively balances model compression ratio with accuracy preservation.

The memristor-based CIM architecture implements weight mapping for neural network hardware through a bit-slicing scheme [2]. This study proposes a weight quantization method based on 8-state memristors, where the weight $B(w_l^i)$ is represented as

$$B(w_l^i) = \sum_{j=0}^{7} b_j \cdot 2^j, \tag{1}$$

where b_7 denotes the most significant bit (MSB), and b_0 represents the least significant bit (LSB). To accommodate the eight stable conductance states of Si:HfO_2-based memristor, the 8-bit weight is partitioned into three slices: $\{b_7, b_6\}$, $\{b_5, b_4, b_3\}$, and $\{b_2, b_1, b_0\}$. Each slice resides in separate memristor cells, enabling efficient CIM by exploiting multi-conductance states. The weights in the l-th layer are defined as

$$\mathrm{S}(W_l) = \left\lceil \log_2 \left(\max_{w_l^i \in W_l} \left(|w_l^i| \right) \right) \right\rceil, \tag{2}$$

where w_l^i represents the weight element indexed by i in the l-th layer, and W_l denotes the set of all weights in that layer. The quantization step size Q_{step} is defined as

$$Q_{\text{step}} = 2^{S(W_l)-8}. \tag{3}$$

The weight element w_l^i is quantized into an 8-bit integer $B(w_l^i)$, which is calculated as

$$B\left(w_l^i\right) = \text{round}\left(\frac{w_l^i}{Q_{\text{step}}}\right), \tag{4}$$

[*] Corresponding author (email: yanxiaobing@ime.ac.cn)

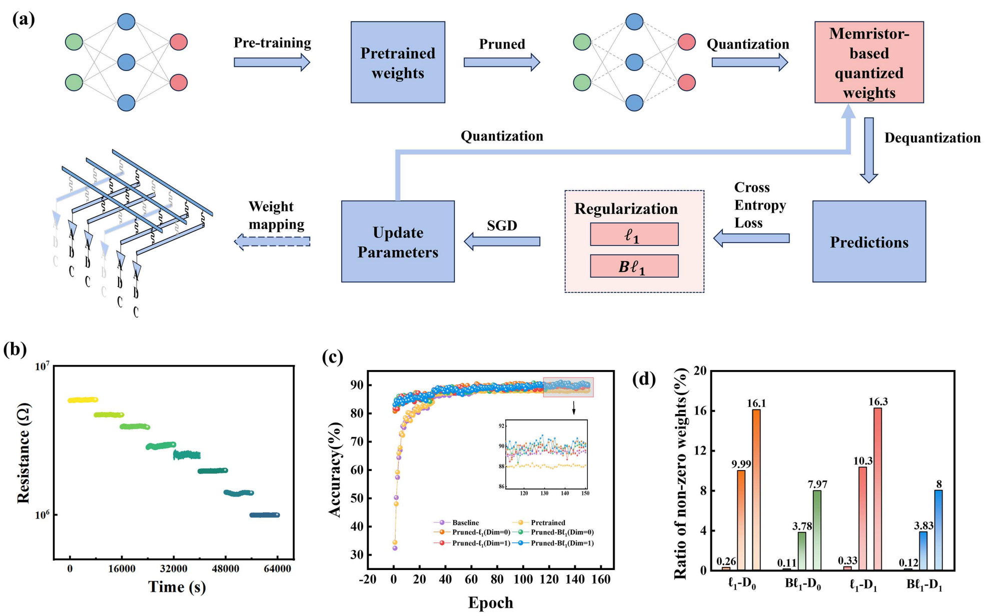

Figure 1 (Color online) (a) Flowchart of quantization-aware training; (b) eight conductance states; (c) accuracy comparison; (d) sparsity comparison.

where $B(w_l^i)$ ranges from [0, 255]. After the computation is completed in each ReRAM crossbar array, the dynamic range of the original weights is restored by applying a simple shift operation:

$$Q\left(w_l^i\right) = B\left(w_l^i\right) \cdot Q_{\text{step}}. \tag{5}$$

The quantized weight $B(w_l^i)$ is mapped to three separate crossbars, and the bit-level sparsity is optimized through ℓ_1 regularization to suppress the impact of extreme values on computation. The $B\ell_1$ regularization is defined as

$$B\ell_1\left(W_l\right) = \sum_{i,k} B_l^{i,k}, \tag{6}$$

where its gradient is backpropagated to the full-precision weights W_l using the straight-through estimator (STE), enabling bit-level sparsity while maintaining training stability. For further details on the quantization procedure, see Appendix D.

During training iterations, the weight w_l^i is first quantized to $B\left(w_l^i\right)$ and then dequantized to $Q\left(w_l^i\right)$, which is used to compute the cross-entropy loss $\mathcal{L}_{\text{CE}}$ and the penalty imposed by the $B\ell_1$ regularization term. Gradients are calculated in full precision, and the weights are updated according to the following rule:

$$q^{(t)} = Q\left(w_l^{(t)}\right), \tag{7}$$

$$w_l^{(t+1)} = q^{(t)} - h \times \left(\nabla_q \mathcal{L}_{\text{CE}}\left(q^{(t)}\right) + \alpha \nabla_q B\ell_1\left(q^{(t)}\right)\right), \tag{8}$$

where $w_l^{(t+1)}$ represents the updated weight at the $(t+1)$-th iteration, $q^{(t)}$ denotes the quantized weight at the t-th iteration, h is the learning rate, $\nabla_q \mathcal{L}_{\text{CE}}(q^{(t)})$ is the gradient of the $\mathcal{L}_{\text{CE}}$ with respect to the quantized weights, $\nabla_q B\ell_1\left(q^{(t)}\right)$ is the gradient of the $B\ell_1$ regularization with respect to the quantized weights, and α is the regularization parameter.

Experiments. Experiments were conducted on an NVIDIA GeForce RTX 3090 GPU (24 GB) using the VGG11-Lite architecture. The model was trained on the CIFAR-10 dataset for 150 epochs with a batch size of 128. Optimization was performed using stochastic gradient descent (SGD) with a momentum coefficient of 0.9. Additional ablation studies and architecture comparisons are provided in Appendix D for further analysis. Figure 1(c) demonstrates that our method achieves 90.2% accuracy with row-wise pruning and 90.8% under column-wise pruning. As shown in Figure 1(d), the proposed $B\ell_1$ regularization improves network sparsity by 2× compared to standard ℓ_1 approaches.

Conclusion. This study proposes an efficient deployment solution for memristor-based neural networks from three dimensions: device, algorithm, and hardware. By employing ferroelectric memristors to achieve efficient quantization, adopting statistics-driven structured pruning to reduce computational complexity, and utilizing bitwise regularization to enhance network sparsity, the proposed approach significantly improves computational efficiency while maintaining model accuracy, providing an effective technical pathway for hardware implementation.

Acknowledgements This work was supported by National Key R&D Plan "Nano Frontier" Key Special Project (Grant No. 2021YFA1200502), Cultivation Projects of National Major R&D Project (Grant No. 92164109), National Natural Science Foundation of China (Grant Nos. 61874158, 62004056, 62104058).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhao Z, Abdelsamie A, Guo R, et al. Flexible artificial synapse based on single-crystalline BiFeO$_3$ thin film. Nano Res, 2022, 15: 2682–2688

2 Zhang J, Yang H, Chen F, et al. Exploring bit-slice sparsity in deep neural networks for efficient ReRAM-based deployment. In: Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition, 2019. 1–5