DAT: Leveraging Device-Specific Noise for Efficient and Robust AI Training in ReRAM-based Systems

Chanwoo Park[1], Jongwook Jeon[1,2], Hyunbo Cho[1]
[1]Research & Development Center, Alsemy Inc., Seoul, Korea
[2]School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, Korea

***Abstract*—The increasing interest in artificial intelligence (AI) and the limitations of general-purpose graphics processing units (GPUs) have prompted the exploration of neuromorphic devices, such as resistive random-access memory (ReRAM), for AI computation. However, ReRAM devices exhibit various sources of variability that impact their performance and reliability. In this paper, we propose Device-Aware Training (DAT), a robust training method that accounts for device-specific noise and resilience against inherent variability in ReRAM devices. To address the significant computational costs of noise-robust training, DAT employs sharpness-aware minimization and a low-rank approximation of the device-specific noise covariance matrix. This leads to efficient computation and reduced training time while maintaining versatility across various model architectures and tasks. We evaluate our method on CIFAR-10 and CIFAR-100 datasets, achieving a 38.2% increase in test accuracy in the presence of analog noise and a 5.9x faster training time compared to using a full-rank covariance matrix. From a loss landscape perspective, we provide insights into addressing noise-induced challenges in the weight space. DAT contributes to the development of reliable and high-performing neuromorphic AI systems based on ReRAM technology.**

***Index Terms*—Resistive random-access memory (ReRAM), Robust training, Neuromorphic AI systems**

I. Introduction

The demand for efficient computational platforms for AI models, driven by the rapid growth of these models and the limitations in power efficiency and inference speed of traditional GPUs, has led to the investigation of neuromorphic devices, such as Resistive Random Access Memory (ReRAM), as viable alternatives. These devices, while promising, exhibit considerable variability due to intrinsic factors, extrinsic factors, device aging, and programming noise. This variability has direct implications on their performance and reliability when employed for AI applications. To ensure robustness and performance, it is imperative that AI models trained on ReRAM devices take into account these multiple sources of device-specific variability.

ReRAM devices display significant variability originating from various sources, which directly impacts their operational stability and effectiveness. This variability can arise from manufacturing irregularities causing differences between devices and from programming noise, where identical set/reset pulses can lead to different resistive states. Environmental factors, such as thermal instability, can further influence device properties, inducing unregulated changes in resistance states.

An additional source of complexity is resistance drift - the gradual change of resistance under constant voltage or after programming. Notably, within a crossbar array (CBA), neighboring ReRAM cells might show more similar noise characteristics due to shared local conditions and process-induced variations, establishing a spatial noise correlation. Successfully addressing these sources of variability is critical for ensuring the robust and reliable operation of AI models on ReRAM-based neuromorphic systems.

Email: chanwoo.park@alsemy.com

In this paper, we present a robust training method, namely Device-Aware Training (DAT), designed to address the complex nature of device-specific noise and inherent variability in ReRAM devices. Our approach encourages learning robust representations capable of adapting to weight perturbations based on each device's unique noise characteristics. One key aspect of DAT is its utilization of a comprehensive noise profile, represented as a multivariate normal distribution with a full-rank covariance matrix, which enables it to model the possible correlations among the noise affecting each parameter.

Training AI models with high-dimensional noise representations poses significant computational challenges, especially considering the trend towards larger foundation models in recent AI research. The cost of training with the full-rank covariance matrix scales as $O(n^3)$, which can be prohibitive as the number of parameters increases. To address this, our method employs optimization techniques such as sharpness-aware minimization and efficient approximations. Specifically, we utilize the eigenstructure of the covariance matrix to construct a low-rank approximation, reducing the computational overhead typically associated with full-rank covariance matrix noise profiles. As a result, DAT offers a scalable solution while preserving key aspects of the noise distribution. AI models trained using DAT are thereby more robust, versatile, and suitable for a broad range of model architectures and tasks. Consequently, DAT provides an effective approach for training AI models on ReRAM devices under realistic noise conditions.

II. Related Work

A. Resilience to Noise in Training

The study by [1] explores the link between the flatness of the weight loss landscape and robust generalization under adversarial training, suggesting a dual-perturbation mechanism for improved model performance. Meanwhile, [2] investigates the benefits of noise injection as a regularization method, proposing an approach that enhances robustness against adversarial attacks. The theoretical examination by [3] suggests that injecting artificial noise into training data introduces a form of weighted ridge regularization, providing a deeper understanding of this commonly used random perturbation technique. These contributions highlight the importance and complexity of developing robust training strategies in noisy environments.

B. Loss Landscapes and Generalization

The behavior of loss landscapes and its relationship with model generalization has been investigated in several works. [4] emphasizes the crucial role of flat minima, noting their significance in keeping loss low even when there are slight shifts in the test environment. The study also suggests that using smaller batch sizes can induce beneficial noise that aids in avoiding sharp minima during training. Conversely, [5] proposes that the generalization gap is more related to the limited number of updates rather than batch size, describing the process of minimizing loss as a random walk influenced by mini-batch noise. To help visualize these phenomena, [6] offers a novel method to illustrate n-dimensional loss functions, emphasizing the need to consider weight scale. Meanwhile, [7] highlights the potential advantages of employing higher learning rates to achieve flatter minima and enhance model generalization, recommending the use of the highest tolerable learning rates to prevent training loss from diverging.

C. Sharpness-Aware Minimization

Several advancements have been made to enhance optimization algorithms with the focus on sharpness of the loss landscape. This includes the Sharpness-Aware Minimization (SAM) by [8] and a similar approach by [9], both aiming to converge towards minima that ensure low loss and sharper landscapes, consequently improving generalization performance. Variants of SAM like the one proposed by [10], not only maintain comparable accuracy gains but also significantly reduce computational overhead, thereby enabling efficient training of large-scale models. Additionally, [11] has demonstrated the effectiveness of sharpness-aware strategies in improving the performance and accuracy of model-agnostic meta-learning in few-shot learning tasks. Lastly, [12] introduced adaptive sharpness-aware minimization, which addresses sensitivity issues of sharpness measures and improves model generalization performance, further validating these techniques' relevance in different contexts.

D. Robust Training in Neuromorphic Devices

Several studies have addressed the challenges of enhancing robustness in deep neural network (DNN) training for ReRAM-based systems. [13] offers a characterization of both deterministic and stochastic noise in ReRAM crossbars. In parallel, [14] examines noise-tolerant strategies across different levels, from circuits and algorithms to entire systems. A distinct approach is taken by [15], which focuses on the development of an analytical noise model that correlates device variability with parameter noise. The investigation by [16] centers on the optimization of binary weight mapping onto ReRAM crossbars, demonstrating increased robustness against adversarial attacks. Lastly, [17] proposes a solution to the issue of bitcell conductance variations in crossbar-based in-memory architectures, achieving higher storage density with minimal loss in DNN accuracy.

III. Device-aware Training

A. Noise Profile

The provision of a comprehensive noise profile is critical for reliable and robust device operation, especially when these devices are used for AI model deployments. In our experimental setting, we assume this noise profile as a multivariate normal distribution with a full-rank covariance matrix, symbolized as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This assumption enables us to model potential correlations among noises in ReRAM devices that affect each parameter due to shared manufacturing processes, common environmental conditions, and interdependent device properties.

This modeling approach, however, poses significant computational challenges. Training AI models with noise sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ becomes computationally prohibitive as the number of parameters increases, with the complexity of covariance matrix decomposition scaling cubically with the number of parameters ($O(n^3)$). To address these computational challenges, we propose Device-Aware Training (DAT), a novel technique that efficiently integrates device-specific noise profiles into the AI model training process. By employing cost-effective approximations, DAT significantly reduces the computational overhead associated with incorporating full-rank covariance matrix noise profiles, thus facilitating scalable and robust AI model training for ReRAM devices.

B. Method

The DAT approach, as outlined in Algorithm 1, is designed to maximize efficiency and robustness in training AI models on ReRAM devices. At the core of DAT lies a unique method that extracts the principal noise components from a full-rank covariance matrix. This reduces the computational complexity from a high-dimensional n to a lower rank k, allowing for efficient noise sampling during the training process.

Each training iteration in the DAT process involves sampling a noise vector from these principal noise components and adding it to the model parameters. The choice of the noise vector is critical: the vector is selected such that the loss is maximized in the direction of these components. This methodology forms the foundation of DAT's optimization objective, which is to identify and minimize the loss where it is most perturbed along the major direction of the noise.

By customizing the training process to match the specific noise characteristics of each device, DAT enhances robustness against device-specific noise. Further, this approach helps reduce loss sharpness, thereby contributing to the overall robustness of the trained models. The utilization of sharpness-aware minimization in this step ensures the model converges to flatter, more robust minima, significantly improving the model's performance.

DAT's adaptability is another key advantage. It can be molded to fit a diverse range of model architectures and tasks, making it versatile in the face of varying requirements. This feature, along with the computational benefits offered by DAT, presents an effective solution to the challenges of training AI models on ReRAM devices under realistic noise conditions.

Algorithm 1 Device-Aware Training

```
Input: Device noise profile $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, loss function $L(\boldsymbol{w})$;
Compute eigenvectors $\boldsymbol{V}_k$ and diagonal matrix $\boldsymbol{D}_k$ corresponding to the $k$ largest eigenvalues of $\boldsymbol{\Sigma}$, where $k \ll n$;
Compute $\boldsymbol{A}_k = \boldsymbol{V}_k \boldsymbol{D}_k^{\frac{1}{2}}$;
while not converged do
    Sample minibatch $\mathcal{B} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{b}$;
    Compute gradient $\boldsymbol{g}_t = \nabla_{\boldsymbol{w}} L_{\mathcal{B}}(\boldsymbol{w_t})$;
    for $j = 1, \ldots, p$ do
        Sample $\boldsymbol{z_j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$; transform $\boldsymbol{\eta}_j = \boldsymbol{\mu} + \boldsymbol{A}_k \boldsymbol{z_j}$;
        Compute projection $\boldsymbol{\epsilon}_j = \frac{\boldsymbol{g}_t \cdot \boldsymbol{\eta}_j}{\|\boldsymbol{\eta}_j\|^2} \boldsymbol{\eta}_j$;
        Perturb weights $\boldsymbol{w}_t^j = \boldsymbol{w}_t + \alpha \boldsymbol{\epsilon}_j$;
        Compute gradient at perturbed weight $\nabla L(\boldsymbol{w}_t^j)$;
    end for
    Aggregate the gradients: $\boldsymbol{g}_{avg} = \frac{1}{p} \sum_{j=1}^{p} \nabla L(\boldsymbol{w}_t^j)$;
    Update the weights: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \beta \boldsymbol{g}_{avg}$;
    $t = t + 1$.
end while
```

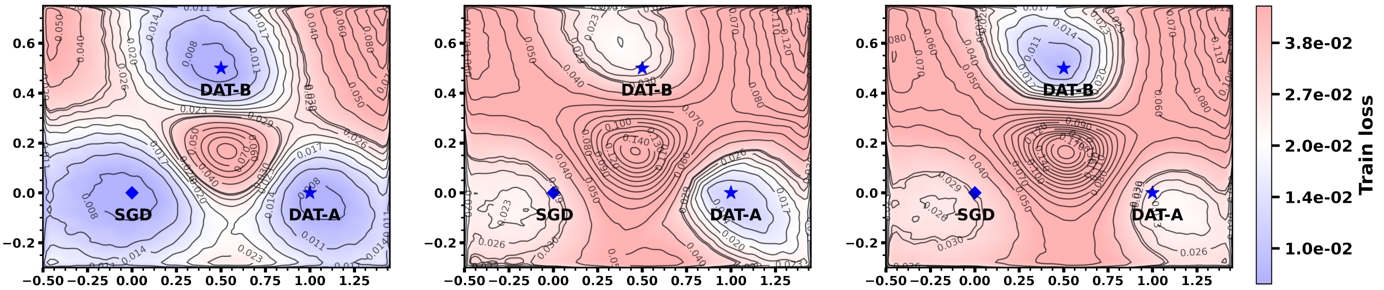
C. Loss Landscape Visualization

We provide a visual representation of the loss landscape by illustrating its behavior in weight space, as detailed below:

- **1D Plot:**

$$g(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \ell \left(f_{\mathbf{w}+\alpha \mathbf{d}} \left(\mathbf{x}_i \right), y_i \right),$$

where α denotes the weight perturbation magnitude, and $\mathbf{d}$ is sampled from a multivariate normal distribution



(a) without noise

(b) with noise profile of device A.

(c) with noise profile of device B.

Fig. 1. **2D Loss Surfaces: Comparing Model Performance with Device-Specific Noise Profiles** Training loss surfaces demonstrate model robustness (a) without noise, (b) with the noise profile of device A, and (c) with the noise profile of device B. The model trained with DAT maintains low loss values, indicating enhanced robustness under specific noise conditions.

TABLE I
IMPACT OF NOISE ON TRAIN/TEST ACCURACY FOR CIFAR-10 AND CIFAR-100 DATASETS

Model	**Method**	**CIFAR-10**			**CIFAR-100**		
		Train Acc. (wo/w noise)	Test Acc. (wo/w noise)	Train time (min.)	Train Acc. (wo/w noise)	Test Acc. (wo/w noise)	Train time (min.)
CNN (6-layer)	**SGD**	99.9/ 76.2	87.6/ 74.9	23.4	99.8/ 70.3	60.8/ 42.5	23.6
	SGD-BN	99.9/ 92.5	90.2/ 83.7	23.2	99.8/ 92.7	66.4/ 61.3	23.3
	DAT-Fast	99.9/ 94.3	90.5/ **86.4**	**24.0**	99.8/ 95.9	66.5/ **65.6**	**25.7**
	DAT-Full	99.9/ 95.7	90.3/ **87.9**	**25.6**	99.8/ 96.6	66.6/ **66.1**	**195.6**
ResNet18	**SGD**	100.0/ 95.1	91.8/ 87.3	47.8	99.9/ 58.2	72.1/ 43.4	48.0
	SGD-BN	100.0/ 95.9	95.3/ 89.6	47.3	99.9/ 94.4	77.6/ 70.8	47.7
	DAT-Fast	100.0/ 98.3	95.2/ **93.0**	**53.0**	99.9/ 97.1	77.7/ **75.6**	**54.6**
	DAT-Full	100.0/ 98.6	95.5/ **93.4**	**54.8**	99.9/ 97.3	77.5/ **76.1**	**232.3**

wo/w noise indicates accuracy measured without and with analog noise, respectively.

characterized by a specific device, and then normalized by the Frobenius norm $\|\mathbf{d}\|_F$.

- **2D Surface:**

$$g(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} \ell \left(f_{\mathbf{w}+\alpha\mathbf{u}+\beta\mathbf{v}} \left(\mathbf{x}_i \right), y_i \right),$$

with $\mathbf{w} = \mathbf{w}_{\text{SGD}}$, $\mathbf{u} = \mathbf{w}_{\text{DAT-A}} - \mathbf{w}_{\text{SGD}}$, and $\mathbf{v} = 2\mathbf{w}_{\text{DAT-B}} - \mathbf{w}_{\text{DAT-A}} - \mathbf{w}_{\text{SGD}}$. $\mathbf{w}_{\text{SGD}}$ denotes the SGD-trained model, and $\mathbf{w}_{\text{DAT-A}}$ and $\mathbf{w}_{\text{DAT-B}}$ correspond to models trained with DAT using the noise profiles of device A and B, respectively. We plot $g(\alpha, \beta)$ over a grid with $\alpha \in [-0.5, 1.5]$ and $\beta \in [-0.3, 0.8]$ to visualize the weight loss landscape.

IV. Experiments and Results

A. Experimental Setup

The experimental setup involved training on the CIFAR-10 and CIFAR-100 datasets using both a 6-layer CNN and ResNet18 architecture. We evaluated four different optimization methods: SGD, SGD with batch normalization (SGD-BN), our proposed DAT-Fast method, and the DAT-Full method. Each model was trained for 200 epochs with a batch size of 128. The learning rates were set at 0.015 for the CNN and 0.075 for ResNet, both with a weight decay of 2×10^{-4}. We used a learning rate scheduler that decayed the rate by a factor of 0.8 every 10 epochs. The train and test loss, along with accuracy, were measured using five different random seeds and then averaged.

For noise injection, the DAT-Full method sampled noise from $N(\mu, \sigma)$ for every iteration. In the DAT-Fast method, we used a low-rank approximation with a rank of $k = 10$. For simplicity, we performed one weight perturbation per iteration for each gradient update step. It is important to note, however, that the model's robustness to noise could be further enhanced by increasing the number of weight perturbations per iteration, at the expense of increased computational time.

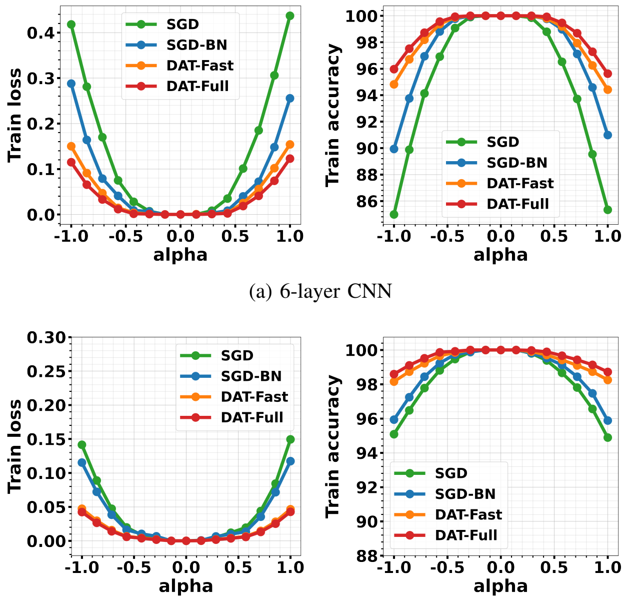
B. Performance Analysis

In experiments simulating the deployment of models on ReRAM devices with distinct noise profiles, the DAT method demonstrated enhanced robustness, as depicted in Fig.2. Evaluating 2D loss surfaces (Fig. 1) showed that models trained with DAT maintained robust performance under different noise profiles, emphasizing the importance of device-specific training.

Batch normalization's effect on the loss landscape was significant, which could be attributed to its role in normalizing layer inputs, thereby maintaining the stability of the network when noise is introduced. Consistent with previous research, the implementation of skip connections in ResNet promoted smoother minimizers, effectively suppressing instability and enhancing noise resistance [6]. While noise robust training employing full covariance yielded a marginally flatter loss landscape, the DAT method demonstrated comparable robustness.

The comparison of noise impacts on different datasets and networks is detailed in Table I. Noise was injected into the last fully connected layer for simplicity. While DAT-full exhibited minor enhancements in noise robustness, its exponential rise in computational complexity with an increase in parameters made it less practical for larger models. Alternatively, DAT-Fast, our proposed algorithm, maintained robustness and generalization performance on par with DAT-Full, but with a training time comparable to that of SGD.

Overall, DAT demonstrated superior performance over SGD, enhancing the average test accuracy by 10.9% and 65.4% (for CNN and ResNet18 respectively) on the CIFAR-10 and CIFAR-100 datasets. The effectiveness of DAT was particularly noticeable on CIFAR-100, a more complex task requiring a larger number of parameters in the final fully connected layer, thereby increasing sensitivity to noise. Importantly, DAT-Fast managed to match the robustness and generalization performance of DAT-Full but with significantly reduced computational demands. This is achieved by coupling a low-rank approximation of the covariance matrix with minimizing loss landscape sharpness, positioning DAT-Fast as an efficient solution for training larger models.



(a) 6-layer CNN

(b) ResNet18

Fig. 2. **Noise Robustness Evaluation with Diverse Architectures on CIFAR10.** The DAT method identifies flatter minima in noise directions for both 6-layer CNN and ResNet18 architectures. Loss curves for ResNet exhibit a broader shape due to the effect of skip-connections.

V. Conclusion

In this paper, we introduced DAT, a novel method designed to address device-specific noise in AI models deployed on ReRAM and other analog devices. Our experiments demonstrated the effectiveness of DAT in noisy conditions, particularly for classification tasks, where it facilitated significant accuracy improvements despite the presence of noise. DAT leverages a low-rank approximation of the covariance matrix and optimizes the loss landscape, achieving a balance between computational efficiency and robustness against variability. This resilience is crucial considering the inherent variability characteristic of neuromorphic devices. DAT's scalability makes it particularly suitable for training large AI models, addressing the computational challenges associated with these models. The results of this study contribute to the advancement of reliable and high-performing neuromorphic AI systems based on ReRAM technology, and its implications extend to other analog devices.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) under grant NRF-2022M3F3A2A01076569.

References

[1] D. Wu, S.-T. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2958–2969, 2020.

[2] Z. He, A. S. Rakin, and D. Fan, "Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 588–597.

[3] O. Dhifallah and Y. Lu, "On the inherent regularization effects of noise injection during training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2665–2675.

[4] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[5] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[6] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," *Advances in neural information processing systems*, vol. 31, 2018.

[7] S. Seong, Y. Lee, Y. Kee, D. Han, and J. Kim, "Towards flatter loss surface via nonmonotonic learning rate scheduling." in *UAI*, 2018, pp. 1020–1030.

[8] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*.

[9] J. Du, D. Zhou, J. Feng, V. Tan, and J. T. Zhou, "Sharpness-aware training for free," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 439–23 451, 2022.

[10] Y. Liu, S. Mai, X. Chen, C.-J. Hsieh, and Y. You, "Towards efficient and scalable sharpness-aware minimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 360–12 370.

[11] M. Abbas, Q. Xiao, L. Chen, P.-Y. Chen, and T. Chen, "Sharp-maml: Sharpness-aware model-agnostic meta learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10–32.

[12] J. Kwon, J. Kim, H. Park, and I. K. Choi, "Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5905–5914.

[13] Z. He, J. Lin, R. Ewetz, J.-S. Yuan, and D. Fan, "Noise injection adaption: End-to-end reram crossbar non-ideal effect adaption for neural network mapping," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[14] X. Yang, C. Wu, M. Li, and Y. Chen, "Tolerating noise effects in processing-in-memory systems for neural networks: A hardware–software codesign perspective," *Advanced Intelligent Systems*, vol. 4, no. 8, p. 2200029, 2022.

[15] Y. Long, X. She, and S. Mukhopadhyay, "Design of reliable dnn accelerator with un-reliable reram," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 1769–1774.

[16] A. Bhattacharjee and P. Panda, "Switchx: Gmin-gmax switching for energy-efficient and robust implementation of binarized neural networks on reram xbars," *ACM Transactions on Design Automation of Electronic Systems*, vol. 28, no. 4, pp. 1–21, 2023.

[17] S. K. Gonugondla, A. D. Patil, and N. R. Shanbhag, "Swipe: Enhancing robustness of reram crossbars for in-memory computing," in *Proceedings of the 39th International Conference on Computer-Aided Design*, 2020, pp. 1–9.