electronics

MDPI

Article

High-Performance and Robust Binarized Neural Network Accelerator Based on Modified Content-Addressable Memory

Sureum Choi, Youngjun Jeon and Yeongkyo Seo *

Department of Information and Communication Engineering, Inha University, Incheon 22212, Korea
* Correspondence: yeongkyo@inha.ac.kr; Tel.: +82-32-860-7415

Abstract: The binarized neural network (BNN) is one of the most promising candidates for low-cost convolutional neural networks (CNNs). This is because of its significant reduction in memory and computational costs, and reasonable classification accuracy. Content-addressable memory (CAM) can perform binarized convolution operations efficiently since the bitwise comparison in CAM matches well with the binarized multiply operation in a BNN. However, a significant design issue in CAM-based BNN accelerators is that the operational reliability is severely degraded by process variations during match-line (ML) sensing operations. In this paper, we proposed a novel ML sensing scheme to reduce the hardware error probability. Most errors occur when the difference between the number of matches in the evaluation ML and the reference ML is small; thus, the proposed hardware identified cases that are vulnerable to process variations using dual references. The proposed dual-reference sensing structure has >49% less ML sensing errors than that of the conventional design, leading to a >1.0% accuracy improvement for Fashion MNIST image classification. In addition, owing to the parallel convolution operation of the CAM-based BNN accelerator, the proposed hardware achieved >34% processing-time improvement compared with that of the digital logic implementation.

Keywords: BNN accelerator; content-addressable memory; XNOR bit-counting operation; dual reference

Citation: Choi, S.; Jeon, Y.; Seo, Y. High-Performance and Robust Binarized Neural Network Accelerator Based on Modified Content-Addressable Memory. *Electronics* **2022**, *11*, 2780. https://doi.org/10.3390/electronics11172780

Academic Editor: Ping-Feng Pai

Received: 3 August 2022
Accepted: 31 August 2022
Published: 3 September 2022

1. Introduction

Convolutional neural networks (CNNs) have been widely used for a range of tasks, including image and speech recognition, and traffic prediction. This is because they achieve far higher accuracies than that of conventional techniques [1–5]. However, recently developed CNNs based on excessive floating-point parameters and operations require multiple power-hungry general-purpose processing units such as graphics processing units [6]. It is challenging to apply CNNs to low-power devices, and considerable research effort has been directed towards improving the operational efficiency of CNNs [7].

Binarized neural networks (BNNs) with binary weights and activations (such as –1 and +1) has gained significant attention as a potential candidate for satisfying desired computation and memory requirements with reasonable accuracy [8–12]. Many researchers have paid attention to improve the efficiency and/or classification accuracy of BNN by proposing various methods such as semi-binarized framework, semantic segmentation, and object recognition [13–15].

The computational benefit achieved by using BNNs in inference is that complex floating-point-based multiplying and accumulating (MAC) operations can be replaced with lightweight bitwise XNOR and bit-counting computations [9]. Such bitwise operations in a BNN exploit the processing-in-memory (PIM) architecture by performing XNOR bit-counting logic operations locally in the memory [16,17]. The PIM architecture reduces data movement between the host processor and memory, which improves the memory bandwidth and power consumption. In particular, PIM architectures developed using content-addressable memory (CAM) have attracted considerable attention because the binarized multiply operation in a BNN can be easily implemented into the CAM array.

Moreover, CAM has the distinct advantage of a fully parallel search operation, leading to high throughput and energy efficiency [17].

However, the main drawback of the CAM-based BNN accelerator [17] is that the reliability of the XNOR bit-counting operations is adversely affected by process variations, resulting in significant sensing errors. Thus, it is imperative to improve the operational reliability of XNOR bit-counting operations. In this paper, we propose a new design technique that reduces the number of operational errors. Considering that XNOR bit-counting operation errors in the PIM hardware mainly occur in cases of small match differences between the evaluated match-line (ML) and reference ML, the proposed hardware can find cases that were vulnerable to XNOR bit-counting errors using a dual-reference sensing scheme.

The remainder of this paper is organized as follows. Section 2 presents the basics of the CAM array based BNN accelerator design. In Section 3, we discuss the proposed CAM design for reducing XNOR bit-counting errors. Section 4 presents the simulation results, and Section 5 concludes the paper.

2. CAM-Based BNN Accelerator Design

2.1. XNOR Bit-Counting Operation

The BNN proposed in [9] restricts the inputs and weights to ± 1. Figure 1a shows the truth table of the multiply operations in the BNN, which are based on the binarized bits of ± 1. However, a Boolean expression could not represent -1 in a single digit; therefore, we assigned the logical value 0 instead of -1 for simplified BNN hardware implementations [9]. The multiplication operation based on +1 and 0 (Figure 1b) can be realized by an XNOR gate because the output of XNOR is 1 if both inputs matched and 0 otherwise [18]. In addition, the post-accumulation signum function in the BNN [9], as shown in Figure 2a, is mapped to a bit-counting operation with a comparator of $n/2$ (half the cumulative size) threshold (Figure 2b) [19]. If the number of matches is equal or greater than that of the threshold, the output activation is 1; otherwise, it is 0.

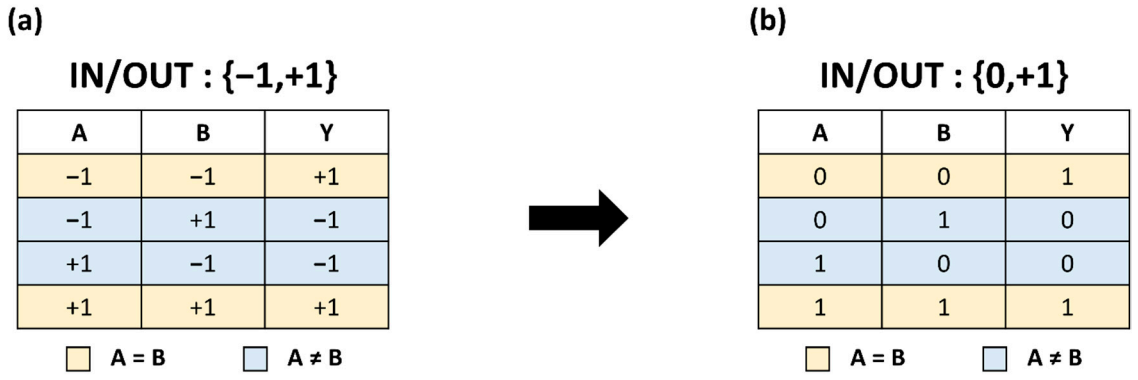


Figure 1. Truth tables for BNN multiply operations based on different in/out: (**a**) +1 and −1; (**b**) +1 and 0.

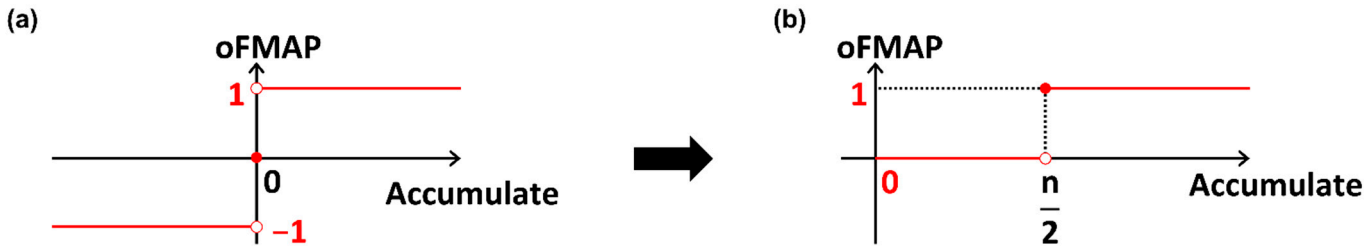
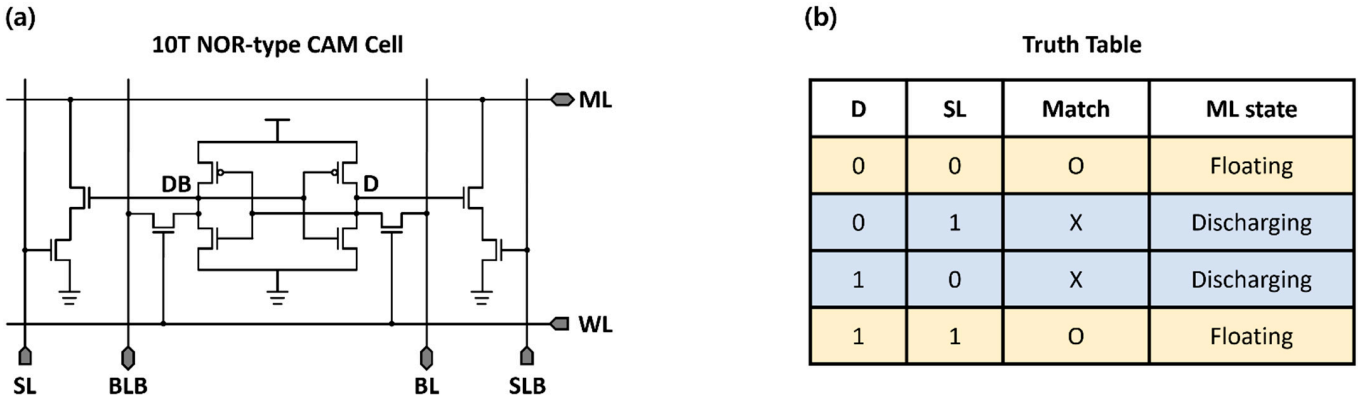


Figure 2. Graphs: (**a**) post-accumulate signum function; (**b**) bit-counting operation.

2.2. CAM-Based BNN Accelerator

CAM compared the input search data with stored data and returns the address of the matching location. Figure 3a shows a 10T-based CAM bit-cell that consists of an 6T SRAM and two stacked NMOS transistors on both sides [14,15]. The ML sensing operation of the CAM array is explained as follows (Figure 4a,b) [20,21]:

1. Prior to the ML pre-charge phase, the search-line (SL) pairs are deactivated to prevent unintentional ML discharges.
2. During the ML pre-charge phase, all MLs in the array are pre-charged to the supply voltage.
3. When SL pairs are set to search for data, the two stacked NMOS transistors compared the search and stored data.



D	SL	Match	ML state
0	0	O	Floating
0	1	X	Discharging
1	0	X	Discharging
1	1	O	Floating

Figure 3. (**a**) Schematic of cell; (**b**) truth table for 10 transistor-based CAM bit-cells.

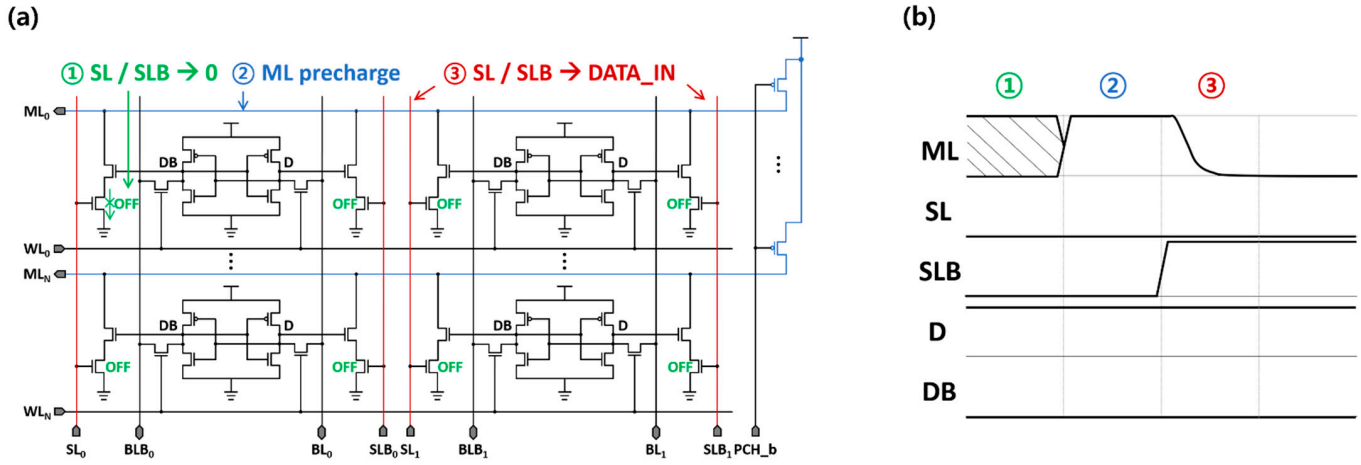


Figure 4. (**a**) The ML operation in a 2-by-2 CAM array; (**b**) timing diagram of the ML operation.

As shown in the truth table from Figure 3b, when the search data are not equal to the stored data, one of the two stacked NMOS transistors is turned on, and a pull-down path is formed such that the ML node is discharging. In contrast, when the search and stored data are matched, both stacked NMOS transistors are turned off, and there is no current path from the ML to the ground [20]. Hence, the XNOR operation for the BNN could be implemented as a search operation for CAM. To perform the convolution operations in the BNN (Figure 5a), the input (iFMAP) is mapped to the storage node of the CAM bit-cell, and the SL node is determined by the weight such that the XNOR operation can be performed by comparing the input and weight in the CAM bit-cell (Figure 5b) [17].

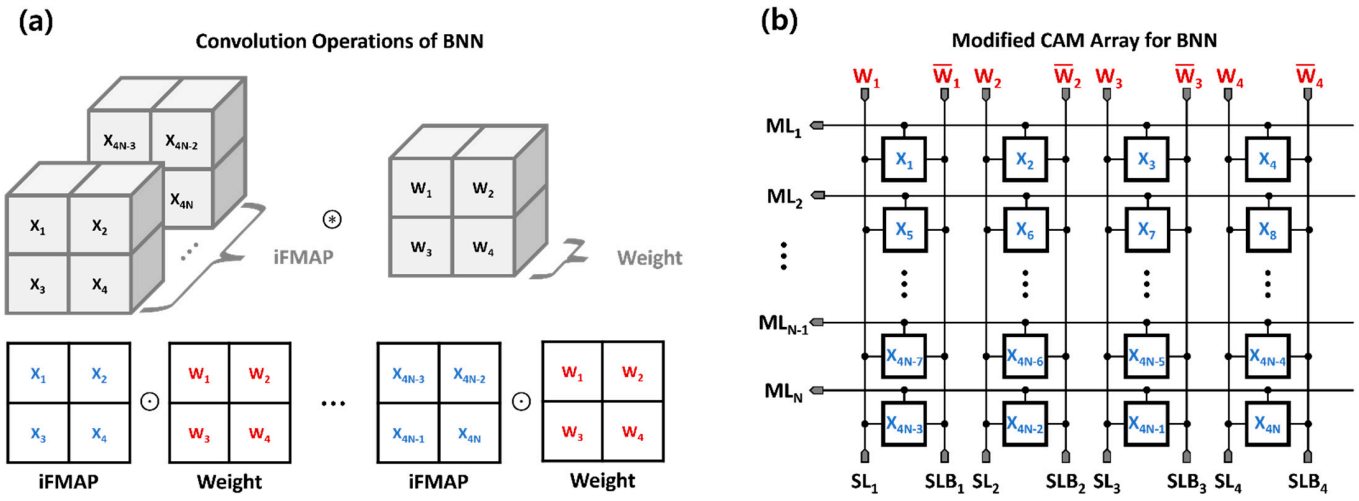


Figure 5. (**a**) Convolution operations of a BNN; (**b**) Modified CAM array architecture for a BNN.

The bit-counting operation for the BNN, which compares the number of matches with the threshold value, is conducted by comparing the voltage of the evaluated ML and reference ML [17]. As shown in Figure 6, the ML voltage depends on the number of matched CAM bit-cells (the higher the number of matched CAM bit-cells, the higher the ML voltage). Therefore, we can determine whether output activation is 1 or 0 by comparing the voltage of the evaluated ML with that of the reference ML using the sense amplifier. For example, if the voltage of the evaluated ML is higher than that of the reference ML, the number of matched CAM bit-cells in the evaluated ML is larger than that in the reference ML, such that the output activation became 1. However, the output activation became 0 when the voltage of the evaluated ML is lower than that of the reference ML.

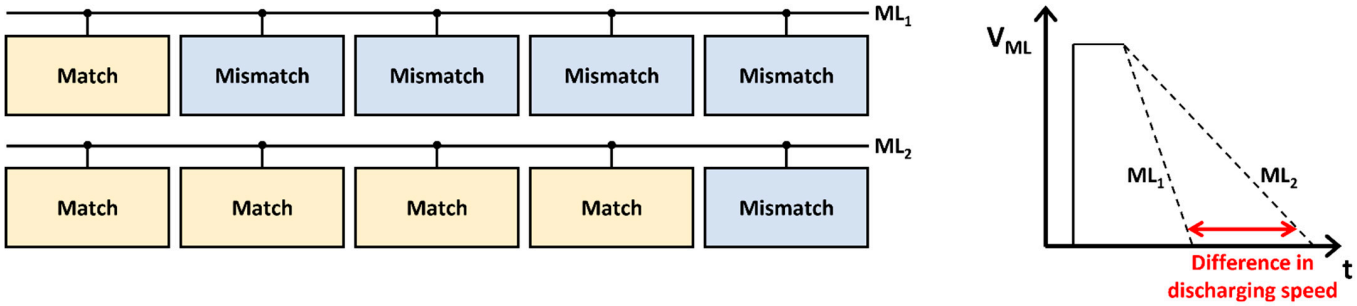


Figure 6. Example of difference in ML discharging speed with respect to the number of mismatched CAM bit-cells.

By performing the XNOR bit-counting operation in the modified CAM array, the computationally intensive convolution operations of the BNN are replaced with the XNOR bit-counting operation in the CAM array. Therefore, the power and performance overheads caused by the data movement between the digital processor and memory are mitigated. Moreover, a high-throughput XNOR bit-counting operation is enabled in CAM-based BNN accelerators because CAM arrays performed multiple ML operations in parallel.

3. Proposed BNN Accelerator Design

The main shortcoming of the CAM-based BNN accelerator [17] is that the reliability of the XNOR bit-counting operations is severely affected by process variations, resulting in significant sensing errors. Thus, it is imperative to improve the reliability of ML sensing operations.

In this paper, we propose a novel sensing technique that reduces the number of ML operation errors. XNOR bit-counting operation errors mainly occur when the difference in the number of matched CAM bit-cells between the evaluated ML and reference ML is small. This is due to the fact that the ML discharging speeds of the evaluated ML and reference ML are close to each other when there are small differences in the number of matches [17]. Thus, when considering local process variation, there is a high possibility that the output of the ML sense amplifier would be reversed. To address the issue of ML sensing reliability, we propose a new sensing scheme with dual references to identify cases prone to process variation. Two separate references are determined as follows. In the first reference ML (REF1), two additional CAM cells are mismatched; thus, the discharging speed of REF1 is higher than that of the conventional reference. The second reference ML (REF2) have two additional match cases; hence, the discharging speed of REF2 is lower than that of the conventional reference.

Figure 7 shows a detailed example of the proposed sensing scheme with a dual reference of a ± 2 bit match. REF1 and REF2 had 6 and 10 bit match bit-cells, respectively. For the 14 bit match input (Figure 7b), the ML discharges much slower than both dual references (REF1 and REF2); thus, the evaluated ML voltage is higher than that of the voltages of both references, and the outputs of the two sense amplifiers are identical to 1. The difference in number of matched CAM bit-cells between the reference ML and evaluated ML is significant; hence, the error probability of the XNOR bit-counting operation is close to 0, and the output of the XNOR bit-counting operation is determined as 1. For the case where two CAM bit-cells are matched (Figure 7c), the evaluated ML discharges significantly faster than those of REF1 and REF2, and the two sense amplifiers have the same outputs of 0; thus, the reliability of the XNOR-bit-counting operation is also high owing to the large difference in the matched bit-cell between the evaluated and reference ML. The result of XNOR bit-counting operation is reliable, and the output activation is considered 0.

In contrast, when the difference in the matched CAM bit-cells between the evaluated ML and reference ML is small, the discharging speed of the evaluated ML (8 bit match case) is similar to that of the two references. Hence, the reliability of the XNOR bit-counting operation is prone to process variations. In such a situation, the evaluated ML voltage is between the voltages of the two references, as shown in Figure 7d. Thus, we determined the case of a high sensing error probability when the sensing outputs from the two references are different. Because the ML sensing output is not reliable in this case, the proposed structure performs a digital logic-based XNOR bit-counting operation followed by an SRAM read operation (with an additional clock cycle) to reduce the XNOR bit-counting errors in the modified CAM array (caused by process variation) [22].

Figure 7a shows the proposed CAM array architecture. It should be noted that in our proposed memory structure, two reference MLs were located on top of the CAM array, and each evaluated ML is compared with the two reference MLs using two sense amplifiers. As aforementioned, if the results from the two sense amplifiers are the same, the result of the XNOR bit-counting operation is considered reliable, and the output is as is. Otherwise, the ML operation results are inaccurate, and additional digital logic-based operations are performed.

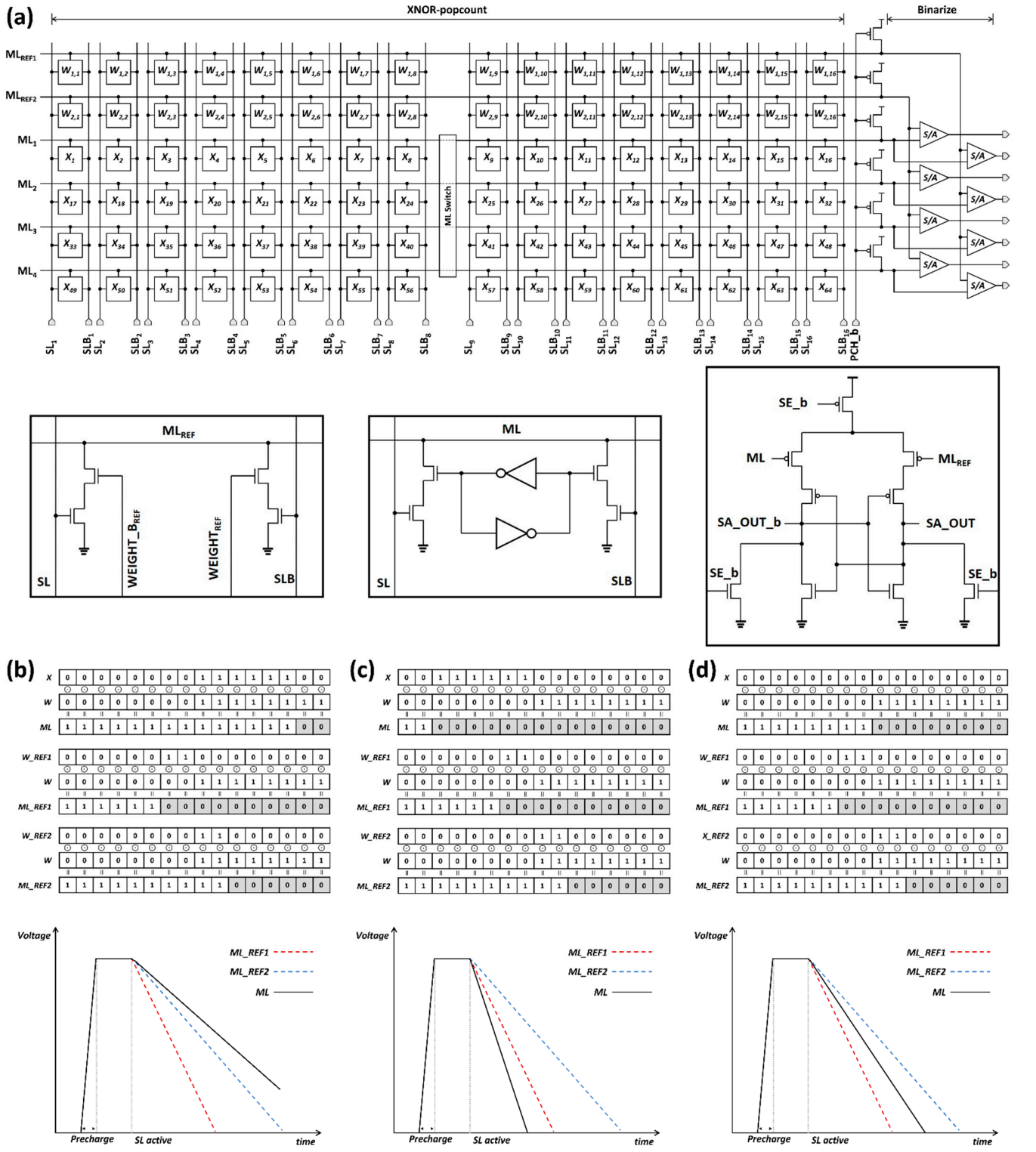


Figure 7. (**a**) Proposed CAM array design with dual-reference sensing scheme. The discharging speed comparison between reference and evaluation MLs for different match cases (**b**) 14 bit (**c**) 2 bit (**d**) 8 bit.

4. Results and Discussion

To evaluate the effectiveness of the proposed CAM-based BNN accelerator, the second convolutional layer of the LeNet-5 model [22] was implemented using commercial 28 nm CMOS technology. Five banks (150 cells per ML) separated by ML switches were imple-

mented in the modified CAM array architecture, with each bank consisting of a 30 × 11 array (nine rows for CAM cells and two rows for reference cells). SPICE circuit-level simulations are performed to evaluate the power, performance, and ML sensing reliability of our proposed CAM array with a conventional single-reference based CAM array [17] in TYPICAL 1 V and 25 °C corner (clock cycle: 250 MHz).

Figure 8a shows a comparison of the XNOR bit-counting failure rates between the conventional and the proposed designs (dual reference of ±2 and ±5 bit case) with respect to the difference in the number of matched CAM bit-cells between the evaluated ML and the reference ML. As previously discussed, the operation failure rates increase when the absolute values of the difference in matched CAM bit-cells between the evaluated ML and reference ML decreased [17]. The proposed dual-reference designs exhibited significantly lower operational failure rates than those of the conventional scheme. This is because the cases, which are vulnerable to ML operational reliability, are detected by the dual reference sensing scheme, and the digital logic-based operations are selectively performed for the reliable XNOR bit-counting results.

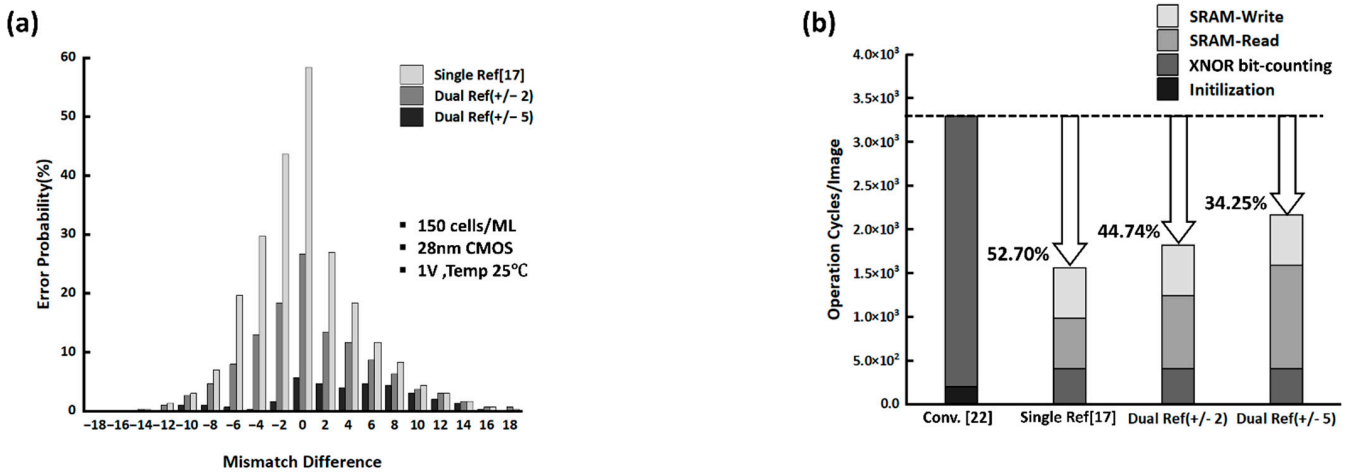


Figure 8. (**a**) Comparison of the error probabilities of the conventional single reference-based design and proposed design with the dual references; (**b**) operation cycles per image comparison of the BNN accelerators.

As listed in Table 1, when the ML sensing error probability of the conventional single reference design was applied to the XNOR bit-counting results, 8.83% of the total output activations for the Fashion MNIST test image were flipped, resulting in a top-1 accuracy degradation of 2.9% for the LeNet-5 [23,24]. However, the proposed memory architecture with dual references of ±2 bit mismatch achieve ~50% reduction in XNOR bit-counting operation errors (4.42% error), leading to an improvement in the classification accuracy (1.0%) as compared with that of the conventional design. The dual references of the ±5 bit mismatch design further improves the error probability reduction by 89% (1.00% error) because more errors are detected by a wider range of dual references. Thus, only a small classification accuracy reduction of 0.5% is observed in the Fashion MNIST dataset.

Table 1. Error probability and classification accuracy comparison of CAM-based BNN accelerators.

	XNOR-Bitcounting Error Probability	Fashion MNIST Classification Accuracy
TOP-1 Accuracy	-	84.4%
Single Ref. [17]	8.83%	81.5%
Dual Ref. (±2)	4.42%	82.5%
Dual Ref. (±5)	1.00%	83.9%

Moreover, as shown in Figure 8b, the proposed BNN accelerators with dual references of ±2 bit and ±5 bit mismatches achieve 44.74% and 34.25% reductions in the number of operation cycles for an input image, respectively, as compared with that of the digital logic-based XNOR-Net hardware in [22], owing to the parallel ML operations of the CAM array.

Figure 9 shows the area and power consumption comparison results between the proposed CAM and conventional CAM arrays shown in [17]. Our proposed CAM structure results in an increase of 5.97% in area as compared to that of the conventional design due to additional reference ML and sensing circuitries (Figure 7a). In addition, owing to the operation of additional reference ML and sensing amplifiers, our proposed memory array showed 7.49% higher power consumption than that of the existing scheme in [17].

Figure 9. CAM array comparison between conventional design and proposed dual reference design in terms of different aspects: (**a**) area; (**b**) power for XNOR-bitcounting operation.

5. Conclusions

We propose a reliability improvement sensing technique for fast and robust CAM-based BNN accelerators. The significant convolution operations of a BNN are replaced using a fully parallel search operation of a modified CAM array, leading to an improvement of >30% in operation performance. Moreover, the proposed sensing scheme with multiple references can reduce the XNOR bit-counting error probability by detecting the cases which are prone to the process variation. Our proposed BNN accelerator achieves >49% less XNOR bit-counting operation errors than that of the conventional design, resulting in a >1.0% accuracy improvement for Fashion MNIST image classification

Author Contributions: Conceptualization, Y.S.; Data curation, S.C. and Y.J.; Formal analysis, S.C.; Funding acquisition, Y.S.; Investigation, S.C., Y.J. and Y.S.; Methodology, S.C. and Y.S.; Project administration, Y.S.; Resources, S.C. and Y.J.; Supervision, Y.S.; Validation, S.C. and Y.S.; Visualization, S.C.; Writing—original draft, S.C. and Y.S.; Writing—review and editing, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the INHA UNIVERSITY Research Grant.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014. [CrossRef]
2. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv* **2013**. [CrossRef]
3. Ciresan, D.C.; Meier, U.; Schmidhuber, J. Multi-column deep neural networks for image classification. *arXiv* **2012**. [CrossRef]

4. Deng, L.; Hinton, G.; Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview. In Proceedings of the IEEE Int. Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013. [CrossRef]
5. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.-Y. Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 865–873. [CrossRef]
6. Dundar, A.; Jin, J.; Gokhale, V.; Martini, B.; Culurciello, E. Memory access optimized routing scheme for deep networks on a mobile coprocessor. In Proceedings of the IEEE High Performance Extreme Computing Conference, Waltham, MA, USA, 9–11 September 2014. [CrossRef]
7. Gong, Y.; Liu, L.; Yang, M.; Bourdev, L. Compressing deep convolutional networks using vector quantization. *arXiv* **2014**. [CrossRef]
8. Courbariaux, M.; Hubara, I.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or −1. *arXiv* **2016**. [CrossRef]
9. Rastegari, M.; Ordonez, V.; Redmon, J.; Farhadi, A. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016. [CrossRef]
10. Simons, T.; Lee, D.-J. A Review of Binarized Neural Networks. *Electronics* **2019**, *8*, 661. [CrossRef]
11. Gao, J.; Liu, Q.; Lai, J. An Approach of Binary Neural Network Energy-Efficient Implementation. *Electronics* **2021**, *10*, 1830. [CrossRef]
12. Choi, J.H.; Gong, Y.-H.; Chung, S.W. A System-Level Exploration of Binary Neural Network Accelerators with Monolithic 3D Based Compute-in-Memory SRAM. *Electronics* **2021**, *10*, 623. [CrossRef]
13. Chen, J.; Wang, J.; Shu, T.; de Silva, C.W. WSN optimization for sampling-based signal estimation using semi-binarized variational autoencoder. *Inf. Sci.* **2022**, *587*, 188–205. [CrossRef]
14. Austin, J. High speed image segmentation using a binary neural network. In *Neurocomputation in Remote Sensing Data Analysis*; Springer: Berlin/Heidelberg, Germany, 1997. [CrossRef]
15. Kung, J.; Zhang, D.; van der Wal, G.; Chai, S.; Mukhopadhyay, S. Efficient Object Detection Using Embedded Binarized Neural Networks. *J. Signal Processing* **2018**, *90*, 877–890. [CrossRef]
16. Liu, R.; Peng, X.; Sun, X.; Khwa, W.S.; Si, X.; Chen, J.J.; Li, J.F.; Chang, M.F.; Yu, S. Parallelizing SRAM Arrays with Customized Bit-Cell for Binary Neural Networks. In Proceedings of the ACM/ESDA/IEEE Design Automation Conference, San Francisco, CA, USA, 24–28 June 2018. [CrossRef]
17. Choi, W.; Jeong, K.; Choi, K.; Lee, K.; Park, J. Content Addressable Memory Based Binarized Neural Network Accelerator Using Time-Domain Signal Processing. In Proceedings of the ACM/ESDA/IEEE Design Automation Conference, San Francisco, CA, USA, 24–28 June 2018. [CrossRef]
18. Guo, P.; Ma, H.; Chen, R.; Li, P.; Xie, S.; Wang, D. FBNA: A Fully Binarized Neural Network Accelerator. In Proceedings of the International Conference on Field Programmable Logic and Applications, Dublin, Ireland, 27–31 August 2018. [CrossRef]
19. Kim, J.H.; Lee, J.; Anderson, J.H. FPGA Architecture Enhancements for Efficient BNN Implementation. In Proceedings of the International Conference on Field-Programmable Technology, Naha, Japan, 10–14 December 2018. [CrossRef]
20. Pagiamtzis, K.; Sheikholeslami, A. Content-addressable memory (CAM) circuits and architectures: A tutorial and survey. *IEEE J. Solid-State Circuits* **2006**, *41*, 712–727. [CrossRef]
21. Huang, P.-T.; Chang, W.-K.; Hwang, W. Low Power Pre-Comparison Scheme for NOR-Type 10T Content Addressable Memory. In Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems, Singapore, 4–7 December 2006. [CrossRef]
22. Yonekawa, H.; Nakahara, H. On-chip Memory Based Binarized Convolutional Deep Neural Network Applying Batch Normalization Free Technique on an FPGA. In Proceedings of the IEEE International Parallel and Distributed Processing Symposium Work, Lake Buena Vista, FL, USA, 29 May–2 June 2017. [CrossRef]
23. Kayed, M.; Anter, A.; Mohamed, H. Classification of Garments from Fashion MNIST Dataset Using CNN LeNet-5 Architecture. In Proceedings of the International Conference on Innovative Trends in Communication and Computer Engineering, Aswan, Egypt, 8–9 February 2020. [CrossRef]
24. Chen, Y.; Rouhsedaghat, M.; You, S.; Rao, R.; Kuo, C.-C.J. Pixelhop++: A Small Successive-Subspace-Learning-Based (Ssl-Based) Model For Image Classification. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020. [CrossRef]