UIC ENGINEERING University of Illinois at Chicago UIC BUSINESS

Supported-BinaryNet: Bitcell Array-based Weight Supports for Dynamic Accuracy-Latency Trade-offs in SRAM-based Binarized Neural Network

Shamma Nasrin, Srikanth Ramakrishna, Theja Tulabandhula and Amit Ranjan Trivedi

Introduction

Neural Network Inference On Edge Devices

IMAGE DATA

INFERENCE RESULTS

ADVANTAGES	CHALLENGES
1. Reduction in Latency and Communicating costs	1. Power
2. Increased Security and Privacy	2. Memory and Compute
	3. Chip Area

- Efficient network Inference under these constraints is an active area of research today.
- We propose a novel approach to address these challenges without compromising on the performance.

Bottlenecks and Solutions

- For any ML algorithm, the fundamental computational operations are multiplication and accumulation, which create the energy bottleneck while designing a hardware accelerator.
- the number of weights in even a moderate size network for real-world applications can be hundreds to tens of thousands.
- In such networks, a typical von Neumann platform incurs high traffic to read weights from the memories and to write back neuron output and partial sums.

Challenges of existing solution

- A key challenge in the current designs is that the network weights must be binarized [1].
- weight binarization leads to higher inaccuracy by limiting the flexibility of weight space.
- Operation with multi-bit precision weights for in-SRAM neural networks requires considerably increased complexity and power- hungry implementation.

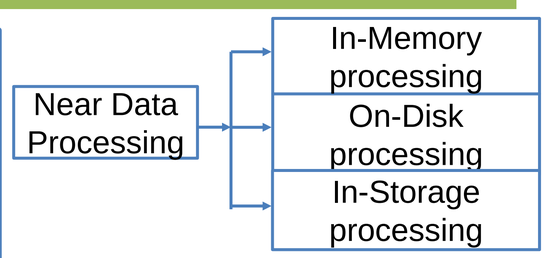


Fig.1. Existing solution. In-Memory and near memory computing.

Proposed Solution

- Support Optimization for In-Memory Computation using an SRAM Bank for increased power efficiency and solution to achieve better performance
- The proposed solution allows us to achieve higher accuracy than the binarized neural network with much lower hardware complexity.

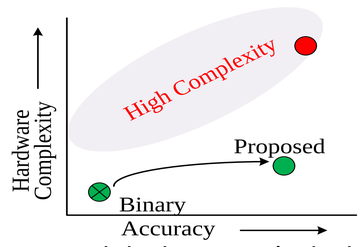


Fig.2. Proposed design complexity is much lower than the existing implementations.

Support Parameter Optimization Algorithm

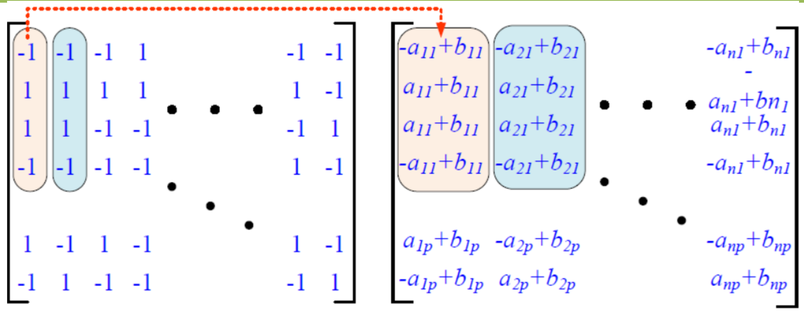
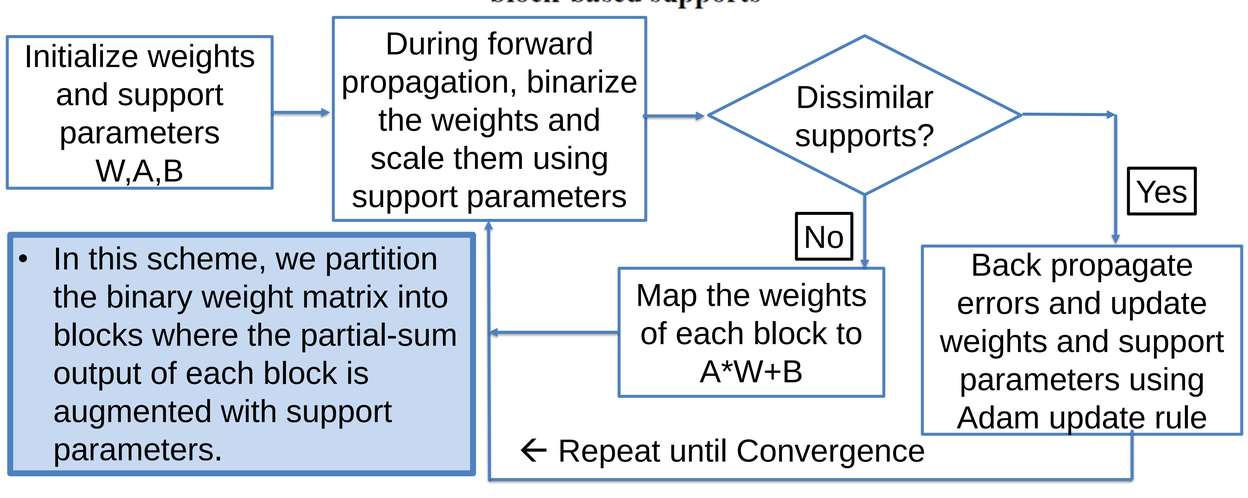
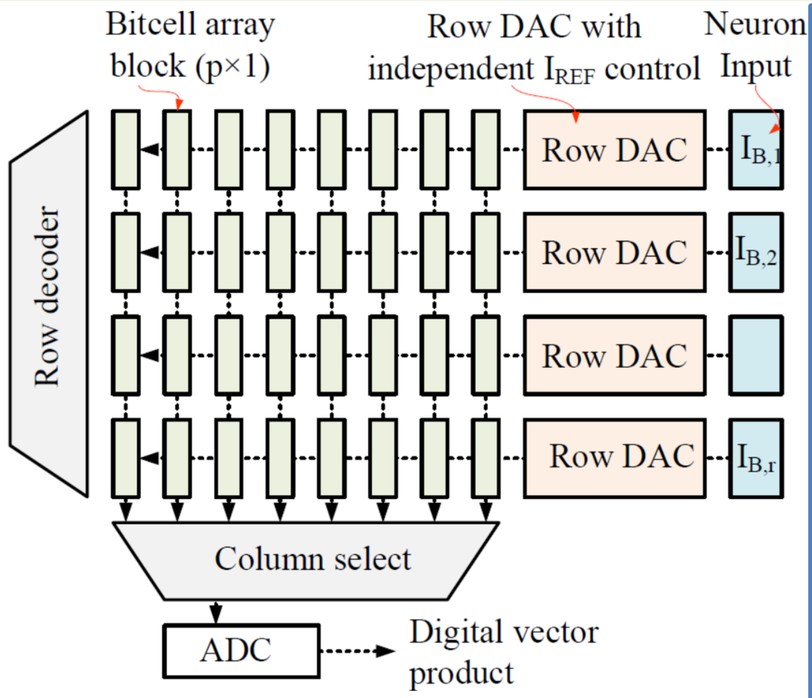


Fig.3. Introducing bitcell array-based support parameters in binarized weight matrix. Here, each block of binary weight matrices is mapped to Aw+B. A and B are learnable support parameters.



- In this scheme, we partition the binary weight matrix into blocks where the partial-sum output of each block is augmented with support parameters.

SRAM Bitcell Array-based Support Vectors



- Weight partitioning is considerate of the physical design and mapping to SRAM, thereby requires minimum interventions to introduce support parameters for each block while maximizing the flexibility of training weight-space.
- Bitcell array-based supports are implemented by modifying the design of digital-to-analog converter (DAC) at SRAM rows.

Fig.4. SRAM architecture for bitcell array-based support parameters. Support parameters to enhance weight space of binarized neural networks (BNNs) are stored in the buffer of row digital-to-analog converter (DAC).

Algorithmic Results

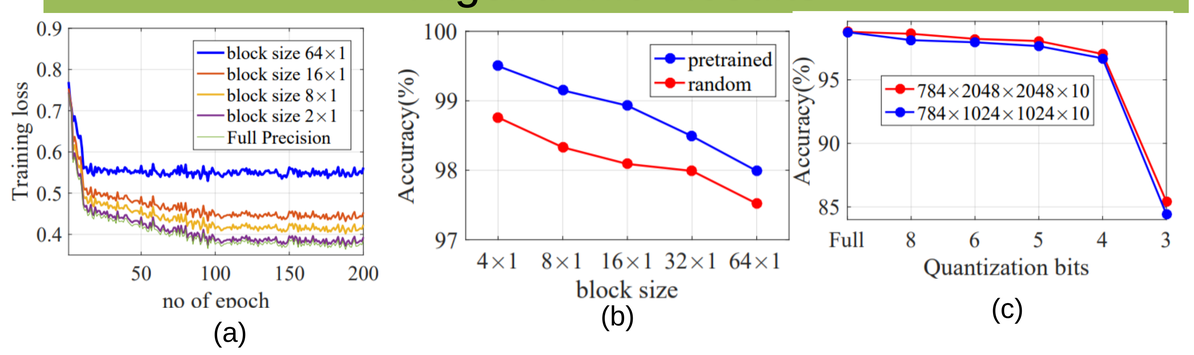


Fig.5. (a) Learning curve of fully connected neural network for MNIST training data set. (b) Accuracy on MNIST by performing support optimization on weights for various block sizes. We show results for two cases: (1) initialized with pre-trained network and (2) initialized randomly.(c) Accuracy with different bit precision of the support parameters.

Hardware Implementation Results

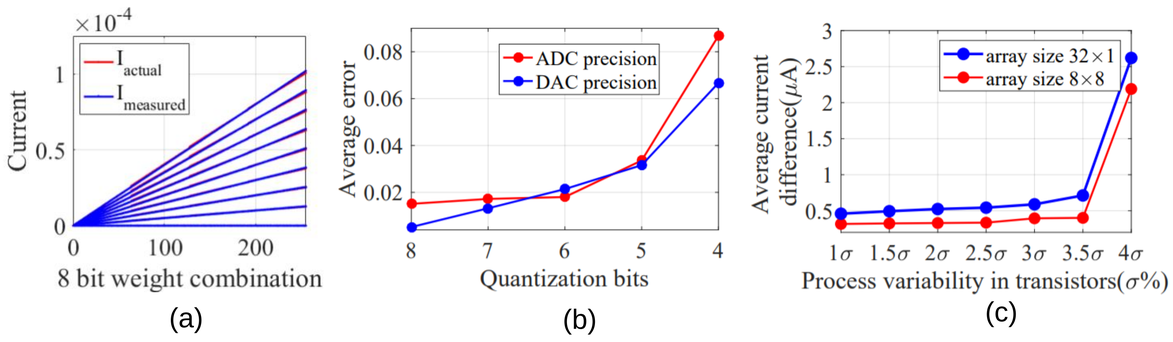


Fig.6. (a) Calculated and simulated current difference for 8×1 block without introducing any support parameters. (b) accuracy of a fully connected MLP for MNIST training data set with different bit precision of the DAC and ADC. Effect of V_TH variability in SRAM cell transistors to scalar product current. (c) Scalar product(as current) at different level of processes variation.

Performance of current work & Future Work

- Our approach reduces classification error in MNIST by 35.71% (error rate decreases from 1.4% to 0.91%).
- To reduce the power overheads, we propose a dynamic drop out a part of the support parameters. Our architecture can drop out 52% of the bitcell array-based support parameters without losing accuracy.
- However, currently we show the work only on fully connected layers of a convolutional neural network and as a future work we will be exploring how the convolutional layer matrices can be incorporated into the SRAM bank architecture for more efficient processing.
- Utilizing the sparsity of the weight matrices for further improvements in efficient processing is also considered as future work.

Refferences

[1] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," , 2016.