Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M17-204P
Published: 5/31/2022

Inventors

Jae-Sun Seo
Shihui Yin
Mingoo Seok
Zhewei Jiang

Contact

Shen Yan
shen.yan@skysonginnovations.com

SRAM Design with Embedded XNOR Functionality for Binary and Ternary Neural Networks

Background Deep neural networks, and in particular convolutional neural networks, are being used with increasing frequency for a number of tasks such as image classification, image clustering, and object recognition. In a forward propagation of a conventional convolutional neural network, a kernel is passed over one or more tensors to produce one or more feature maps. Recent work in the field has focused on reducing the necessary computing power for implementing convolutional neural networks. One approach, referred to as a "binary neural network," uses binary weight values in the kernel. By converting the weight values in the kernel to binary values, a forward propagation of the binary neural network can be computed using only addition and subtraction. Another approach, referred to as an "XNOR neural network," uses binary input values in the tensors and binary weight values in the kernel. By converting the tensor input values and the kernel weight values to binary, a forward propagation of the XNOR neural network can be computed using only an exclusive nor (XNOR) operation and a bit count operation, where a bit count operation is simply a count of the number of high bits in a given stream of binary values. XNOR neural networks have applications in mobile and other low-power devices. However, conventional computing systems are not well suited for efficient XNOR neural network implementation. Accordingly, there is a need for computing systems, and in particular memory architectures, that are capable of efficiently supporting the operation of XNOR neural networks for improvements in speed and efficiency. Invention Description Researchers at Arizona State University and Columbia University have developed a novel SRAM bitcell and array design that embeds XNOR functionality, which is suitable for efficiently computing binarized neural networks (e.g., XNOR-Net) in hardware. The proposed SRAM array performs bitwise XNOR and bitcount operation in parallel, enabling direct mapping of neural networks with binary (+1, -1) weights and binary (+1, -1) / ternary (+1, 0, -1) activations onto the SRAM array for in-memory computing. This innovation is covered by U.S. Pat. No. 11,170,292. Potential Applications • XNOR neural networks • Deep neural networks • In-memory computing Benefits and Advantages • Performs parallel in-memory XNOR computing of binarized neural networks Related Publication: XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks (PDF)Research Homepage of Professor Jae-sun Seo