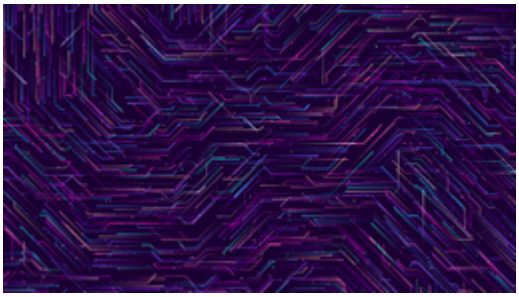
INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

Crossbar Mapping of DNN Weights

TECHNOLOGY NUMBER: 2020-479



Technology ID
2020-479

Category
Hardware
Engineering & Physical Sciences

Inventor
Wei Tang
Zhengya Zhang

Further information
Joohee Kim
jooheek@umich.edu

Learn more

OVERVIEW

Method to map weights for kernels onto a crossbar array comprised of non-volatile memory cells

- Storage of multiple bits of information across memory cell groups for deep neural network information weighting
- May be implemented by one or more computer programs that are run by a single or multiple processors

BACKGROUND

Machine learning or artificial intelligence (AI) tasks use neural networks to learn and then to infer. Learning refers to the process of tuning the weight values by training the network on vast amounts of data, while inference describes the process of presenting the network with new data for classification. The workhorse of many types of neural networks is vector-matrix multiplication, a procedure that produces computations between an input and weight matrix. Crossbar mapping is a method of implementing deep neural network (DNN) weights that can achieve high energy efficiency and low latency in hardware implementations. This technique involves mapping the weight values of a DNN onto the programmable resistive crossbar structure of a memristive (i.e., resistive random-access memory) device. The crossbar structure is composed of a set of parallel conductive wires that are connected to programmable resistive elements, which can store the weight values of the DNN. Still, a need exists to map the weights of one or more kernels of a neural network onto a crossbar array in order to improve computational efficiency.

INNOVATION

Researchers have developed a method that maps weights for kernels onto a crossbar array comprised of non-volatile memory cells that serves as an aid for machine learning and

inference by deep neural networks (DNN). This technology involves the arrangement of memory cells into columns and rows such that each row of the array is interconnected by a respective drive line and each column of the array is interconnected by a respective bit line. As the system receives two or more kernels of a neural network that are represented as values in a matrix that are converted into column vectors and stored. The computing system either employs an analog approach where an analog value is stored in the memristor of each memory cell or a digital approach which stores binary values in the memory cells. For a binary number comprised of multiple bits, the memory cells are collected into groups of memory cells, such that the value of each bit in the binary number is stored in a different memory cell within the group of memory cells.

For example, a value for each bit in a five-bit binary number is stored in a group of five adjacent columns of the array, where the value for the most significant bit is stored in memory cell on the leftmost column of a group and the value for the least significant bit is stored in memory cell in the rightmost column of a group. In this way, a multiplicand of a multiply-accumulate operation is a binary number comprised of multiple bits and stored across a one group of memory cells in the array, facilitating crossbar weighting of the DNN values.

The hardware arrangement of the system includes a data bus that is interfaced with or connected to the data bus in addition to a plurality of crossbar modules that are also connected to the data bus. Peripheral hardware includes a drive line circuit, a wordline circuit, and a bitline circuit that are designed to minimize the number of switches and level-shifters needed for mixing high-voltage and low-voltage operation. This peripheral hardware supports read and write operations in relation to the memory cells which comprise the crossbar array. If a circumstance arises in which the kernel dimensions or number of kernels in one layer are too large to fit into a single crossbar array, multiple crossbar arrays may be utilized. As described, this system may be implemented by one or more computer programs that are run by a single or multiple processors.