Neural Network Weights Stored in One-Transistor (1T) Crossbar Arrays

TECHNOLOGY NUMBER: 2020-475

OVERVIEW

Improved calculations which optimize machine learning and artificial intelligence

- A neural network composed of a crossbar array with fixed weight nodes
- Performs vector matrix multiplication and generates inferences based off matrix inputs

BACKGROUND

Machine learning and artificial intelligence (AI) are steadily becoming more integrated into work produced across a wide spectrum of industries. The process of machine learning or (AI) involves the use of 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 refers to the process of presenting the network with new data for classification. The mainstay of many types of neural networks is vector-matrix multiplication, or computation between an input and weight matrix. Crossbar arrays perform analog vector-matrix multiplication naturally, with each row and column of the crossbar connected through a processing element (PE) that represents a weight in a weight matrix. Inputs can be applied to the rows as voltage pulses, and the resulting column currents are scaled, or multiplied, by the PEs according to physics. The total current in a column is defined by the summation of each PE current. Given the importance of these types of calculations to optimize machine learning and AI, a need exists to further scale measurement device dimensions to provide different on-resistance values.

INNOVATION

Researchers have invented a neural network composed of a crossbar array with fixed weight nodes to perform vector matrix multiplication and generate inferences based off inputs to the matrix. This innovation utilizes an analog neural network architecture and a crossbar array comprised of a row controller, column controller, and transistor gate voltage controller. The system is set up so that each row and column of the crossbar is connected through a processing element which represents a weight in a weight matrix. Inputs are applied to the rows as voltage pulses, and the resulting current columns are scaled by the processing elements, an action which is learned offline by tuning weight values using vast amounts of data. These weights do not vary, so the processing elements are made up of a single transistor where the weight is the transistor on-resistance. The on-resistance relates to the device dimensions and the voltage applied to the transistor gate. The transistor processor elements are therefore able to correlate the weight for each node of the array to subsequently perform matrix multiplication.

Technology ID
2020-475

Category
Hardware
Engineering & Physical Sciences

Author(s)
Justin Correll
Michael Flynn
Seungheun Song

Further information
Joohee Kim
jooheek@umich.edu

Learn more