LICENSING OPPORTUNITY: QUASI-SYSTOLIC PROCESSOR AND QUASI-SYSTOLIC ARRAY

DESCRIPTION

Problem

During training of typical AI models, in order to update the matrix of weights in a particular network layer, a developer must calculate an additional update matrix to transfer into the main weight array. This update matrix is just as big as the main array, which is already very large. This large matrix calculation dramatically reduces the advantages of using certain emerging classes of vector matrix multipliers, such as crossbar arrays, since this large matrix requires lots of traditional computing resources. Also, it's well known, even during traditional training, transmission of weight updates and network gradients consumes a lot of computing resources.

Invention

The invention is a computer architecture which efficiently calculates a weight matrix update for an AI model, especially ones designed to be implemented in hardware neural networks. Our architecture uses an approximation algorithm to do this calculation with less memory overhead. The architecture uses special matrix decomposition methods, such as streaming principal component analysis, to calculate an approximation of this update matrix using far fewer parameters which need to be stored in memory.

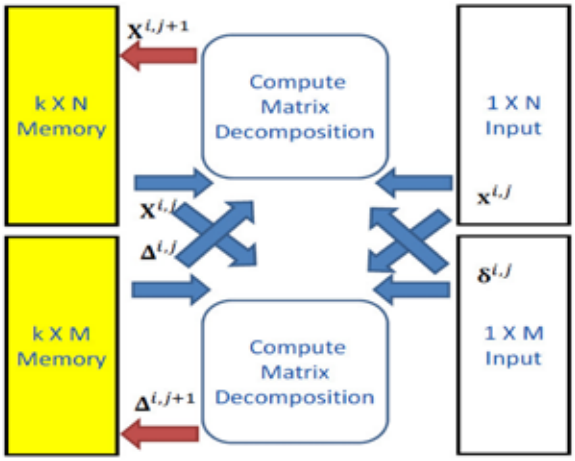
BENEFITS

Commercial Application

The primary use of this technology is for training of AI models in the cloud or at the edge. This approach potentially yields many of the critical benefits of batch update, and requires substantially less overhead, and has a significantly lower computational cost.

Competitive Advantage

The fewer the number of memory locations and calculations needed to train the network, then the less time, area, and energy needed to operate a AI hardware system.



The matrix decomposition system block.

Contact: licensing@nist.gov

NIST TECHNOLOGY PARTNERSHIPS OFFICE

NIST Technology Partnerships Office
National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899-2200