(SD2020-340) Algorithm-Hardware Co-Optimization For Efficient High-Dimensional Computing

Tech ID: 32309 / UC Case 2020-340-0

BACKGROUND

With the emergence of the Internet of Things (IoT), many applications run machine learning algorithms to perform cognitive tasks. The learning algorithms have been shown effectiveness for many tasks, e.g., object tracking, speech recognition, image classification, etc. However, since sensory and embedded devices are generating massive data streams, it poses huge technical challenges due to limited device resources. For example, although Deep Neural Networks (DNNs) such as AlexNet and GoogleNet have provided high classification accuracy for complex image classification tasks, their high computational complexity and memory requirement hinder usability to a broad variety of real-life (embedded) applications where the device resources and power budget is limited. Furthermore, in IoT systems, sending all the data to the powerful computing environment, e.g., cloud, cannot guarantee scalability and real-time response. It is also often undesirable due to privacy and security concerns.

Thus, we need alternative computing methods that can run the large amount of data at least partly on the less-powerful IoT devices. Brain-inspired Hyperdimensional (HD) computing has been proposed as the alternative computing method that processes the cognitive tasks in a more light-weight way. The HD computing is developed based on the fact that brains compute with patterns of neural activity which are not readily associated with numerical numbers. Recent research instead have utilized high dimension vectors (e.g., more than a thousand dimension), called hypervectors, to represent the neural activities, and showed successful progress for many cognitive tasks such as activity recognition, object recognition, language recognition, and bio-signal classification.

TECHNOLOGY DESCRIPTION

Researchers from UC San Diego have developed the first sparse hyperdimensional (HD) computing method that enables sparsity on the trained HD model. They name the proposed method "SparseHD". SparseHD improves the efficiency of hyperdimensional computing through reducing the number of elements in trained classes (which are in the form of hypervectors). This element reduction (i.e., sparsification) is done in two different ways.

In the first approach, which they call dimension-wise sparsification, SparseHD finds if elements of a specific index of all class/model hypervectors have an equal or close-to-equal value (e.g., dimension 2 of all class hypervectors lies in a limited range). This dimension is then removed from all class hypervectors because during the similarity checking (of an incoming query hypervector with class hypervectors) it adds the same score to all comparisons, so is ineffectual in finding out the maximum similarity score.

In the second approach, which they call class-wise sparsification, SparseHD finds the ineffectual dimensions 'within a class'. Essentially it removes the dimensions with values close to zero, because when such dimension is multiplied to the corresponding dimension of an incoming query hypervector, the score will also be close to zero. So such dimensions can be removed without significant impact on the score of a class.

The inventors also propose to use an automated technique which iteratively retrains HD models to compensate the potential quality loss that might be incurred by above-mentioned model sparsification. Basically retraining calibrates the values of remaining dimensions to enhance the

CONTACT
University of California, San Diego
Office of Innovation and Commercialization
innovation@ucsd.edu
tel: 858.534.5815.

OTHER INFORMATION

KEYWORDS
Robust Machine Learning, Efficient Machine Learning, Low-Power Internet of Things (IoT)

CATEGORIZED AS
- **Computer**
  - Security
  - Software

RELATED CASES
2020-340-0

accuracy.

They also propose an FPGA implementation of sparse HD computation that supports both the dimension-wise and class-wise sparse models. Their FPGA accelerator is hand-crafted in a pipelined structure to effectively utilize the FPGA resources to maximize performance. Specifically, they implement a Compress-Sparse-Column (CSC) architecture to enable class-wise sparsification as it is more complicated than dimension-wise sparsification that removes a specific dimension of all classes.

APPLICATIONS

This design enables low-power and efficient machine learning for IoT and embedded devices.

INTELLECTUAL PROPERTY INFO

US Patent Rights pending.

UC San Diego is seeking companies interested in commercializing this technology. Non-exclusive licenses are currently available to interested companies.

RELATED MATERIALS

- M. Imani, S. Salamat, B. Khaleghi, M. Samragh, F. Koushanfar and T. Rosing, "SparseHD: Algorithm-Hardware Co-optimization for Efficient High-Dimensional Computing," 2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), San Diego, CA, USA, 2019, pp. 190-198, doi: 10.1109/FCCM.2019.00034. - 04/28/2019

PATENT STATUS

Patent Pending

University of California, San Diego
Office of Innovation and Commercialization
9500 Gilman Drive, MC 0910, ,
La Jolla,CA 92093-0910

Tel: 858.534.5815
innovation@ucsd.edu
https://innovation.ucsd.edu
Fax: 858.534.7345