Peter deNoyelles

CSCI 5593

February 5th, 2019

**Report on “Accelerator for Sparse Machine Learning”**

Sparse matrix by vector multiplication is a bottleneck in today’s machine learning and data mining algorithms. This paper proposes an accelerator they have created to be integrated with the CPU for address Space Matrix by Sparse Vector multiplication (SpMSpV). Their accelerator algorithm reduces memory access while also gaining power efficiency. There have been many accelerators proposed for space matrix multiplication including but not limited to; parallel architecture comprising nnz (number of nonzero matrix elements) processing elements, and space matrix multiplication on the Distributed Array Processor (a parallel SIMD architecture).

The paper looks at two ways of multiplying a matrix by a vector. The first on they look at is implementing an inner dot product of each row of the matrix and the vector (CSR). The second is multiplying each column of the matrix by an appropriate vector element and accumulating the product vector (CSC). However, with the sparsity the researchers were testing, CSR’s performance of SpMSpV was expected to be extremely low, so the implementation of that algorithm was not done for this study.

While the CSC algorithm could be implemented via software, it can be sped up significantly via a designed hardware unit. The researchers used an architecture comprised of a Sparse Matrix Fetch Engine, a FMAC, and a Product Cache. Results proved to be significant. Compared to nnz and sps, researchers found 70 times increase in performance, 8 times better power efficiency, and 29 times overall energy reduction for SpMSpV based applications.

Through research of machine learning articles, I learned that this is essentially a new field for computer architecture. There are many bottlenecks researchers and daily users are finding in their ML, DL, and data mining operations that have potential fixes in new architecture designs. Many of the current designs don’t account for repeated use of matrix multiplication and more research into these designs could make for large movements in efficiency for the new ways we process big data.

**References**

[1] L. Yavits and R. Ginosar, "Accelerator for Sparse Machine Learning", *IEEE Computer Architecture Letters*, vol. 17, no. 1, pp. 21-24, 1 Jan.-June 2018. URL: <http://ieeexplore.ieee.org.aurarialibrary.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=7946089&isnumber=8316264> [Accessed February 1, 2019]