LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-SR-747311

Exploring Deep Learning and Sparse Matrix Format Selection

Y. Zhao, C. Liao, X. Shen

March 6, 2018

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Exploring Deep Learning and Sparse Matrix Format Selection

Student: Yue Zhao
Supervisors: Chunhua (Leo) Liao (LLNL technical contact), Xipeng Shen (Professor at NC State)

Recent several years have witnessed a rapid progress of deep learning, which has led to some significant advancements in image processing, speech recognition, and many other tasks. The potential of deep learning has however remained largely unexploited for High Performance Computing (HPC). In this work, we proposed to leverage deep learning to help remove some long-standing barriers for high performance sparse matrix computations.

This work focuses on Sparse Matrix Vector Multiplications (SpMV). As many studies have observed, a key factor for SpMV performance is the use of a proper storage format to represent sparse matrices in memory. More than 50 different storage formats (e.g., CSR, COO, DIA, ELL, etc.) have been proposed to match with various kinds of matrices and computer architectures. The formats significantly affect the data locality, cache performance, and ultimately the end-to-end performance of SpMV (for as much as several folds of difference).

We proposed to explore the use of Deep Neural Networks (DNN) for addressing the long-standing barriers. The recent rapid progress of DNN technology has created a large impact in many fields, which has significantly improved the prediction accuracy over traditional machine learning techniques in image classifications, speech recognitions, machine translations, and so on. To some degree, these tasks resemble the decision makings in many HPC tasks, including the aforementioned format selection for SpMV and linear solver selection. For instance, sparse matrix format selection is akin to image classification—such as, to tell whether an image contains a dog or a cat; in both problems, the right decisions are primarily determined by the spatial patterns of the elements in an input. For image classification, the patterns are of pixels, and for sparse matrix format selection, they are of non-zero elements. DNN could be naturally applied if we regard a sparse matrix as an image and the format selection or solver selection as classification problems.

To exert the potential, this work has developed several novel techniques to bridge the gap between SpMV format selection and DNN. First, it presents a set of fixed-size representations of sparse matrices that we have designed to suit the needs of DNN, including a binary scaled image, a density image, and a histogram-based representation. We have systematically explored the impact of the formats.

Second, this work empirically reveals the influence of CNN structures on sparse matrix format selection, and identifies a *late merging* structure as a CNN structure that suites the needs of Sparse Matrix format selection. The structure delays the integration of the information from different parts of the input representation to a late stage of the CNN processing, making it better match the input representations of sparse matrices.

Third, it introduces a concept in machine learning, transfer learning, into HPC, and reveals its potential for alleviating the cross-architecture migration difficulties for CNN-based models to serve for matrix format selection. It proposes two ways to materialize transfer learning in this new context, empirically compares their effectiveness, and demonstrates the large savings of the model migration overhead the technique brings.

Fourth, using an expanded set of sparse matrices, it compares the CNN-based method with the state of the art of sparse matrix format selection. The results indicate that the new method improves the accuracy of the best matrix format selection from 85% to 93%. The predictions rectified by the CNN model yields 1.73 average (up to 5.2) speedups to SpMV.

This work crystallizes all the explorations into a set of novel findings on the applications of CNN to sparse matrix format selection, and published them at PPOPP'2018 conference. These findings could shed insights for bridging the gap between CNN and other HPC problems.

Publication:

Y. Zhao, J. Li, C. Liao and X. Shen, "POSTER: Bridging the Gap Between Deep Learning and Sparse Matrix Format Selection," 2017 26th International Conference on Parallel Architectures and Compilation Techniques (PACT), Portland, Oregon, USA, 2017, pp. 152-153.

Yue Zhao, Jiajia Li, Chunhua Liao, and Xipeng Shen. Bridging the gap between deep learning and sparse matrix format selection. In Proceedings of the 23rd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '18). ACM, New York, NY, USA, 94-108.