Contribution ID: **9** Type: **Talk**

Convolutional Neural Nets for the Run Time and Energy Consumption Estimation of the Sparse Matrix–Vector Product

Tuesday, 5 November 2019 15:50 (30 minutes)

Introduction. Modeling the execution time and the energy efficiency of the Sparse Matrix-Vector product (SpMV) on a current CPU architecture is especially complex due to i) irregular memory accesses; ii) indirect memory referencing; and iii) low arithmetic intensity. While analytical models may yield accurate estimates for the total number of cache hits/misses, they often fail to predict accurately the total execution time and energy consumption. In this work, we depart from the analytic approach to instead leverage Convolutional Neural Networks (CNNs) in order to provide an effective estimation of the run time and energy consumption of the SpMV operation.

Modeling. The memory-bound nature of SpMV operation (y = Ax) is mainly due to the low non-zero densities and the irregular sparsity patterns in matrix A which, in general, dictate a considerable volume of cache misses and DRAM memory accesses to the x vector. Therefore, the vector storing the column indices (*vpos*) from the CSR format array can be regarded as a key element to understand the distinct arithmetic-to-memory access intensities and predict both the execution time and energy consumption. With this idea in mind, we design a CNN where the inputs are the values of the *vpos* array as a one-dimensional image of the sparse matrix A that captures the order in which the entries of x are retrieved from memory. Once trained, the filters in the convolutional layers should be capable of capturing meaningful features of the *vpos* array in order to yield useful run time and energy consumption estimations via the relation between flops and cache hits/misses. In the methodology proposed to tackle the SpMV modeling problem, a CNN receives the *vpos* array as an input. However, given that sparse matrices may present large variations on their size and number of nonzero entries (nnz), we split the *vpos* input array into chunks (or blocks) of size b=250 to design a CNN with a constant number of inputs (neurons). With it, the estimated total execution time and energy consumption for a matrix can be calculated as the aggregation of the time and energy estimations for each of the blocks.

Evaluation. We tackle the performance and energy consumption estimation as a regression problem, resulting in a same CNN model that is trained once per metric (time and energy). The training of the networks was performed on a set of blocks of sparse matrices from the SuiteSparse matrix collection labeled with the corresponding SpMV execution time and energy consumption drawn by an Intel Xeon Haswell core, running at 2.4 GHz. These metrics were measured using Intel RAPL performance counters. The results of the testing process reveal that the relative mean error for the set of testing matrices for the proposed models is lower than 1.7% for the execution time and less than 3.1% for the energy consumption. On a separated experiment, we extended our analysis on varying frequencies, ranging from 1.2 to 2.4 GHz. These results reveal that decreasing the processor frequency slightly improves the relative mean error.

Conclusions. We proposed a CNN models to estimate the execution time and energy consumption of the SpMV operation. The designed models incorporate a blockwise strategy which makes the CNN architecture independent of the sparse matrix size, allowing users to obtain run time and energy consumption estimations prior the execution or even when the target architecture is no accessible.

Primary authors: BARREDA, Maria (Universitat Jaume I); DOLZ, Manuel F. (Universitat Jaume I); CASTAÑO, M. Asunción (Universitat Jaume I); QUINTANA-ORTÍ, Enrique S. (Universitat Politècnica de València)

Presenter: DOLZ, Manuel F. (Universitat Jaume I)

Session Classification: Day I