Energy-Efficient Ranking on FPGAs through Ensemble Model Compression (Abstract)

Veronica Gil-Costa[1], Fernando Loor[1], Romina Molina[1,2,3], Franco Maria Nardini[3], Raffaele Perego[3] and Salvatore Trani[3]

[1]*Universidad Nacional de San Luis, San Luis, Argentina*
[2]*Università degli Studi di Trieste, Trieste, Italy*
[3]*ISTI-CNR, Pisa, Italy*

In this talk we present the main results of a paper accepted at ECIR 2022 [1].
We investigate novel SoC-FPGA solutions for fast and energy-efficient ranking based on machine-learned ensembles of decision trees. Since the memory footprint of ranking ensembles limits the effective exploitation of programmable logic for large-scale inference tasks [2], we investigate binning and quantization techniques to reduce the memory occupation of the learned model and we optimize the state-of-the-art ensemble-traversal algorithm for deployment on low-cost, energy-efficient FPGA devices. The results of the experiments conducted using publicly available Learning-to-Rank datasets, show that our model compression techniques do not impact significantly the accuracy. Moreover, the reduced space requirements allow the models and the logic to be replicated on the FPGA device in order to execute several inference tasks in parallel. We discuss in details the experimental settings and the feasibility of the deployment of the proposed solution in a real setting. The results of the experiments conducted show that our FPGA solution achieves performances at the state of the art and consumes from 9× up to 19.8× less energy than an equivalent multi-threaded CPU implementation.

References

[1] V. Gil-Costa, F. Loor, R. Molina, F. M. Nardini, R. Perego, S. Trani, Ensemble model compression for fast and energy-efficient ranking on fpgas, in: European Conference on Information Retrieval, Springer, 2022, pp. 260–273.
[2] R. Molina, F. Loor, V. Gil-Costa, F. M. Nardini, R. Perego, S. Trani, Efficient traversal of decision tree ensembles with FPGAs, Journal of Parallel and Distributed Computing 155 (2021) 38–49.

IIR2022: 12th Italian Information Retrieval Workshop, June 29 - June 30th, 2022, Milan, Italy
gvcosta@unsl.edu.ar (V. Gil-Costa); fernandoloor1@gmail.com (F. Loor); mromy00@gmail.com (R. Molina); francomaria.nardini@isti.cnr.it (F. M. Nardini); raffaele.perego@isti.cnr.it (R. Perego); salvatore.trani@isti.cnr.it (S. Trani)
0000-0003-4637-9725 (V. Gil-Costa); 0000-0002-8552-1221 (F. Loor); 0000-0001-7688-6248 (R. Molina); 0000-0003-3183-334X (F. M. Nardini); 0000-0001-7189-4724 (R. Perego); 0000-0001-6541-9409 (S. Trani)