COMPUTATIONAL RESEARCH in BOSTON and BEYOND SEMINAR

Leveraging Heterogeneous Hardware Resources for Efficient Machine Learning Inference Service

BAOLIN LI

Northeastern University

ABSTRACT:

Machine learning (ML) model inference is a key service in many businesses and scientific discovery processes. In the meantime, modern HPC systems and clouds are integrating more and more heterogeneous resources into the system. In this talk, I will discuss various techniques we develop to efficiently serve ML inference workloads using heterogeneous hardware resources. I will first introduce a simplified version of the inference serving problem, of which we apply a integer linear programming solver to minimize the energy consumption. Next, I will emphasize on a few challenges in a production environment where the inference queries display high variety and the users demand strict quality-of-service (QoS). To solve this complicated problem, we propose RIBBON, a cost-effective and QoS-aware inference server deployed on heterogeneous cloud computing instances. RIBBON formulates this as a black-box optimization problem and devises a Bayesian Optimization-driven strategy to allocate the heterogeneous resources. Compared to existing approaches, RIBBON saves up to 16% of the inference serving cost on various representative workloads.

FRIDAY, AUGUST 5, 2022
12:00 PM – 1:00 PM

https://math.mit.edu/sites/crib/

ZOOM Link…

https://mit.zoom.us/j/96155042770
Meeting ID: 961 5504 2770

Massachusetts Institute of Technology