北京大学高能效计算与应用中心学术报告

Invited Talk, Center for Energy-Efficient Computing and Applications

MATRIX FACTORIZATION ON GPU: A TALE OF TWO ALGORITHMS

Dr. Wei Tan

Research Staff Member
IBM T. J. Watson Research Center

2017年1月13日 星期五 3:00pm
理科五号楼410会议室

ABSTRACT: Matrix factorization (MF) is an approach to derive latent features for two categories of entities, from the observed interaction matrix between them. It is at the heart of many algorithms, e.g., collaborative filtering where two entities are users and items; topic models where two entities are documents and words; word embedding where two entities are words and words. Alternating least Square (ALS) and stochastic gradient descent (SGD) are two popular algorithms in solving MF. SGD converges fast, while ALS is easy to parallelize and able to deal with non-sparse ratings. GPU with massive cores and high intra-chip memory bandwidth sheds light on accelerating MF much further when appropriately exploiting its architectural characteristics.

In this talk, I will introduce **cuMF**, a CUDA-based matrix factorization library that accelerates both ALS and SGD to solve very large-scale MF. cuMF uses a set of techniques to maximize the performance on single and multiple GPUs. These techniques include smart access of sparse data leveraging GPU memory hierarchy, using data parallelism in conjunction with model parallelism, approximate algorithms and reduced precision. With only a single machine with up to four Nvidia GPU cards, cuMF can be 10 times as fast, and 100 times as cost-efficient, compared with the state-of-art distributed CPU solutions. Moreover, cuMF can solve the largest matrix factorization problem ever reported in current literature. In this talk I will also share lessons learned in accelerating compute- and memory-intensive kernels on GPUs.

BIOGRAPHY: Wei Tan is a Research Staff Member at IBM T. J. Watson Research Center. His research interest includes big data, distributed systems, NoSQL and services computing. Currently he works on accelerating machine learning algorithms using scale-up (e.g., GPU) and scale-out (e.g., Spark) approaches. His work has been incorporated into IBM patent portfolio and software products such as Spark, BigInsights and Cognos. He received the IEEE Peter Chen Big Data Young Researcher Award (2016), Best Paper Award at ACM/IEEE ccGrid 2015, IBM Outstanding Technical Achievement Award (2014), Best Student Paper Award at IEEE ICWS 2014, Best Paper Award at IEEE SCC 2011, Pacesetter Award from Argonne National Laboratory (2010), and caBIG Teamwork Award from the National Institute of Health (2008).