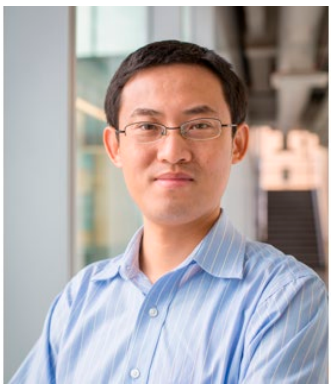
HM-ANN: Efficient Billion-Point Nearest Neighbor Search on Heterogeneous Memory

Dong Li (University of California, Merced)



Abstract

The state-of-the-art approximate nearest neighbor search (ANNS) algorithms face a fundamental tradeoff between query latency and accuracy, because of small main memory capacity: To store indices in main memory for fast query response, they have to limit the number of data points or store compressed vectors, which hurts search accuracy. The emergence of heterogeneous memory (HM) brings opportunities to largely increase memory capacity and break the above tradeoff: Using HM, billions of data points can be placed in main memory on a single machine without using any data compression. However, HM consists of both fast (but small) memory and slow (but large) memory, and using HM inappropriately slows down query time significantly. In this work, we present a novel graph-based similarity search algorithm called HM-ANN, which takes both memory and data heterogeneity into consideration and enables billion-scale similarity search on a single node without using compression. On two billion-sized datasets BIGANN and DEEP1B, HM-ANN outperforms state-of-the-art compression-based solutions such as L&C and IMI+OPQ in recall-vs-latency by a large margin, obtaining 46% higher recall under the same search latency. We also extend existing graph-based methods such as HNSW and NSG with two strong baseline implementations on HM. At billion-point scale, HM-ANN is 2X and 5.8X faster than our HNSW and NSG baselines respectively to reach the same accuracy.

Bio

Dong Li is an associate professor at EECS at University of California, Merced. Previously, he was a research scientist at the Oak Ridge National Laboratory (ORNL), studying computer architecture and programming model for next generation supercomputer systems. Dong earned his PhD in computer science from Virginia Tech. His research focuses on high performance computing (HPC), and maintains a strong relevance to computer systems. The core theme of his research is to study how to enable scalable and efficient execution of enterprise and scientific applications on increasingly complex large-scale parallel systems. Dong received a CAREER Award from the National Science Foundation in 2016, and an ORNL/CSMD Distinguished Contributor Award in 2013. His paper in SC'14 was nominated as the best student paper; His paper in ASPLOS'21 won the distinguished artifact award. He is also the lead PI for NVIDIA CUDA Research Center at UC Merced. He is a review board member of IEEE Transaction on Parallel and Distributed Systems (TPDS).