Performance Improvement of Approximate Nearest Neighbour Search on GPU: BANG Exact-distance variant

BANG [1] is the state of the art for Billion-scale approximate nearest neighbour search using a single GPU. At a billion-scale, the dataset and graph index are huge and cannot be entirely accommodated in the GPU device memory. BANG employs PQ compression techniques to compress the base dataset and keep it in the GPU memory. The graph index is kept in the Host memory and optimally accessed from the GPU.

At smaller scales of the datasets, like 100 million, the entire dataset and the graph index can be entirely accommodated in the GPU memory. To handle such datasets, we have a variant of BANG called BANG_ExactDistance [4]. CAGRA [2] (ANNS implementation by NVIDIA) outperforms BANG significantly. To perform a search operation for a batch of 10,000 queries, CAGRA takes ~ 10 ms while BANG takes 20 ms.

This project aims to improve the performance of BANG and try to match or outperform CAGRA. From a time and memory profiling comparison using NCU, we found that CAGRA effectively uses shared memory for filtering our visited neighbours and has a GPU-optimised implementation of Euclidean distance calculation. The sorting implementation in BANG can also be improved by using Bitonic sort. The CLI interface, build, and packaging of BANG_ExactDistance must also be enhanced to match the BANG_Base [3].

References:

[1] Karthik V., Saim Khan, Somesh Singh, Harsha Vardhan Simhadri, Jyothi Vedurada, "BANG: Billion-Scale Approximate Nearest Neighbor Search using a Single GPU", IEEE Transactions on Big Data (2025), PrePrints pp. 1-16

[2] Ootomo, Hiroyuki, et al. "Cagra: Highly parallel graph construction and approximate nearest neighbor search for gpus." *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024.

[3] BANG_Base Source Code, https://github.com/karthik86248/BANG-Billion-Scale-ANN/tree/main/BANG_Base

[4] BANG_Exactdistance Source Code, https://github.com/karthik86248/BANG-Billion-Scale-ANN/tree/main/BANG_Exactdistance