Towards a Vertical Layout for Vector Similarity Search

Leonardo Kuffo [1], Peter Boncz [1]

[1] Database Architectures, Centrum Wiskunde & Informatica, Netherlands
{lxkr,boncz}@cwi.nl

Vector similarity search has rapidly become a core component of various applications such as semantic search, recommendation systems and LLMs pipelines. Recently, an unprecedented flourish of improvements to approximate vector search on high-dimensional embeddings has been developed. Ranging from new index structures, CPU/GPU optimizations and vector size reduction (quantization). Even more recently, ADSampling [2] and BSA [3] were proposed: a novel line of research whose efforts are on pruning the number of computations in the distance evaluation by only visiting a few dimensions of the vectors during a search. However, the conventional horizontal data layout for vectors (vector-after-vector) limits their benefits due to nonoptimal data access patterns and SIMD distance kernels showing degraded performance at low dimensionalities.

We propose Partition Dimensions Across (PDX), a data layout for vectors that stores multiple vectors in one block (similar to PAX [1]), using a vertical layout for the dimensions. PDX improves search speed in exact and approximate vector similarity search thanks to our dimension-by-dimension search strategy that benefits from processing multiple-vectors-at-a-time. This improves the efficiency of SIMD distance kernels in modern architectures (avg 40% faster) only relying on scalar code that gets auto-vectorized. Using the PDX layout, we accelerated dimensions-pruning algorithms, ADSampling [2] and BSA [3], by 5.3x and 3.4x respectively without any loss of recall thanks to the more efficient distance kernels at low dimensionalities and better data access patterns of the PDX layout which can adapt per query and dataset. Furthermore, we developed PDX-BOND, a dimension-pruning algorithm with good performance on exact search and reasonable performance on approximate search that works on vector data "as-is" without preprocessing/transformation.

For future research, we aim to focus on efficient *compressed* representations of dimensions within the blocks of the PDX layout. Here, we emphasize the need for *compressed* vector representations rather than *quantized* ones. Furthermore, we are working on strategies to not only prune dimensions but entire blocks of vectors without sacrificing recall.

References

[1] Ailamaki, A., DeWitt, D. J., Hill, M. D., & Skounakis, M. (2001, September). Weaving Relations for Cache Performance. In VLDB (Vol. 1, pp. 169-180).

[2] Gao, J., & Long, C. (2023). High-dimensional approximate nearest neighbor search: with reliable and efficient distance comparison operations. Proceedings of the ACM on Management of Data, 1(2), 1-27.

[3] Yang, M., Jin, J., Wang, X., Shen, Z., Jia, W., Li, W., & Wang, W. (2024). Bridging Speed and Accuracy to Approximate K-Nearest Neighbor Search. arXiv preprint arXiv:2404.16322.