LMDiskANN.jl: An Implementation of the Low Memory Disk Approximate Nearest Neighbors Search Algorithm

25 January 2025

Summary

LMDiskANN.jl is a Julia package that implements the Low Memory Disk Approximate Nearest Neighbor search algorithm (LM-DiskANN) (Pan, Sun, and Yu 2023), extending DiskANN-based methods (Jayaram Subramanya et al. 2019; Singh et al. 2021) for fast, accurate billion-point nearest neighbor search while significantly reducing in-memory usage. By leveraging memory-mapped files, a dynamic graph-based index, and tunable BFS expansions, **LMDiskANN.jl** enables large-scale similarity search on commodity hardware. This package integrates well with embedding-based workflows common in vector databases and modern machine learning pipelines. It also allows for insertions and deletions after the index has been constructed with operations to keep the index pruned of unnecessary connections in the graph.

Features include:

- Low-memory adjacency storage on disk using memory maps.
- Dynamic insert and delete operations on the graph index, adapting to changes in datasets.
- Configurable search expansions (`EF_SEARCH`, `EF_CONSTRUCTION`, etc.) to tune performance vs. recall.
- Optional user-key LevelDB mapping for flexible ID to and from key lookups.

By combining these capabilities, **LMDiskANN.jl** aims to reduce the memory footprint and overhead for large-scale nearest neighbor searches. It can be incorporated into any workflow that requires efficient embedding retrieval or similarity search for vectors in high-dimensional spaces.

Statement of need

Approximate Nearest Neighbor (ANN) search is a crucial component in domains such as recommendation systems, information retrieval, and representation learning (e.g., embeddings for natural language or computer vision). Traditional approaches can suffer from excessive memory usage and slow scaling when dealing with billions of points (Nene and Nayar 1997; Wang et al. 2021). By persisting adjacency structures on disk rather than in memory, **LMDiskANN.jl** addresses some of these bottlenecks, providing:

1. **Reduced Memory Overhead**: Only a minimal fraction of data needs to reside in RAM, making it feasible to handle larger datasets on modest machines.

2. **Dynamic Updates**: Graph-based insertions and deletions support real-time or streaming scenarios where data is continually changing.

3. **High Recall**: Tuning BFS expansions and adjacency degrees can yield high-quality nearest neighbor results.

4. **Scalable Architecture**: Built on Julia's high-performance ecosystem, bridging native disk operations and advanced numeric libraries.

This approach benefits practitioners who need large-scale nearest neighbor indexing without specialized cluster infrastructures or extremely large memory capacities. The set up is made to have minimal requirements and has a simple installation proceedure. Using the package involves a few number of steps and examples are provided in the documentation.

Acknowledgements

Thanks to the Julia community for their continued support of open-source scientific computing. We also acknowledge the authors of LMDiskANN (Pan, Sun, and Yu 2023) and other works it is based on, (Jayaram Subramanya et al. 2019; Singh et al. 2021) for foundational ideas in disk-based graph indexing.

References

Jayaram Subramanya, Suhas, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. 2019. "Diskann: Fast Accurate Billion-Point Nearest Neighbor Search on a Single Node." *Advances in Neural Information Processing Systems* 32.

Nene, Sameer A, and Shree K Nayar. 1997. "A Simple Algorithm for Nearest Neighbor Search in High Dimensions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (9): 989–1003.

Pan, Yu, Jianxin Sun, and Hongfeng Yu. 2023. "Lm-Diskann: Low Memory Footprint in Disk-Native Dynamic Graph-Based Ann Indexing." In *2023 Ieee International Conference on Big Data (Bigdata)*, 5987–96. IEEE.

Singh, Aditi, Suhas Jayaram Subramanya, Ravishankar Krishnaswamy, and Harsha Vardhan Simhadri. 2021. "Freshdiskann: A Fast and Accurate Graph-Based Ann Index for Streaming Similarity Search." *arXiv Preprint arXiv:2105.09613*.

Wang, Mengzhao, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. 2021. "A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search." *arXiv Preprint arXiv:2101.12631*.