Computational Enhancements of HNSW Targeted to Very Large Datasets

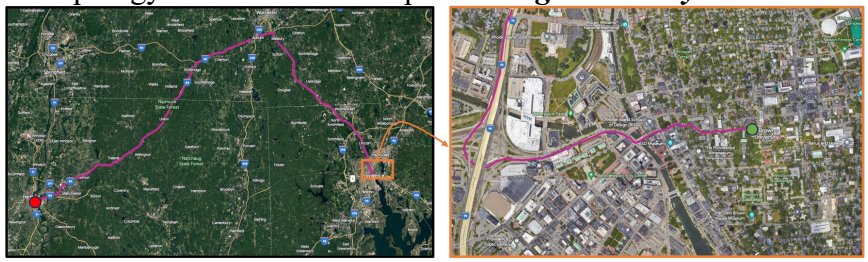
Brown University
School of Engineering

Cole Foster, Benjamin Kimia
{Cole_Foster, Benjamin_Kimia}@brown.edu

sisap

The Hierarchical Navigable Small World (HNSW) index is a graph-based approach that uses a hierarchy to separate links based on length, where:

- The long-range links of the upper layers are used to ***efficiently*** approach the approximate neighborhood of the query.
- The short-range links of the bottom layer capture the local topology of the dataset and provide ***high accuracy***.

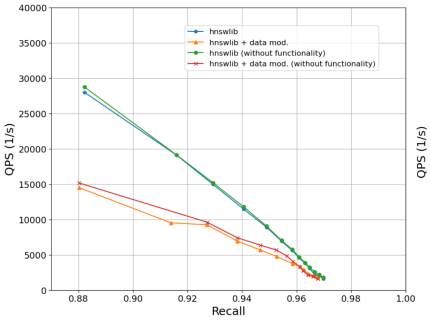


Overview: This submission to the SISAP Indexing Challenge:

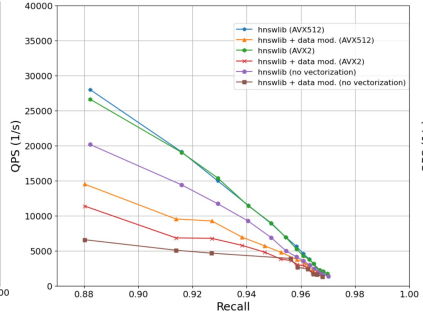
- Uses the *hnswlib* implementation of HNSW.
- Introduces a memory-modification for large datasets.
- Emphasizes the importance of computational and memory optimizations for high performance computing.
- Performed the best on all subsets of the competition.

Maximizing Search-Time Efficiency: Beyond algorithmic design, there are many performance-based optimizations that all high-performance indices should consider:

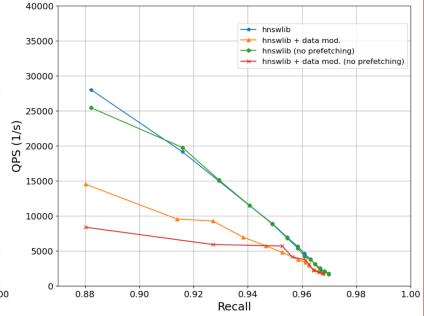
1. Our submission ***removes all unnecessary functionality*** from search time.
2. Distance computations on 768D vectors are well-suited for ***SIMD instructions,*** and *hnswlib* uses AVX-512 instructions by default.
3. Cache efficiency is improved by ***spatial locality****;* its impact is shown by the reduced efficiency of our memory modification.
4. With predictable memory accesses, the ***cache prefetching*** employed by *hnswlib* provides greater cache efficiency (right).



The impact of removing unnecessary functionality.



Several types of SIMD instructions for the vectorization of distance computations.

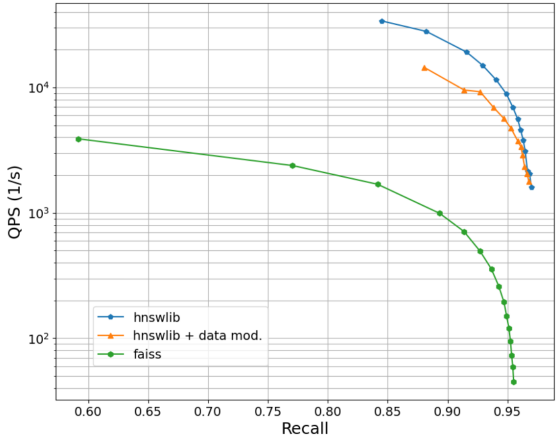


The benefit of cache prefetching.

Implementations of HNSW: the original *hnswlib* library significantly outperforms *faiss*. For performance, ***use hnswlib***.

Memory Modification for Large Datasets:

- LAION 100M dataset (single precision) takes up 286GB of memory.
- Our submission avoids a batched construction by referencing the full dataset loaded to RAM.
- This sacrifices memory contingency of vectors with their neighbors.
- This modified index can be serialized without the dataset, reducing its size from 309GB to 17GB.

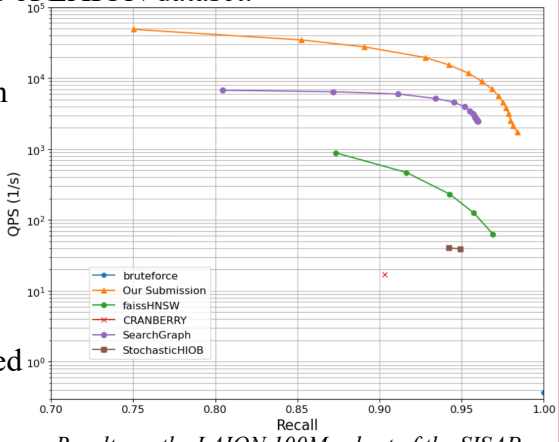


Comparison of several implementations of HNSW on LAION 100M: *faiss, hnswlib, and hnswlib* with our memory modification.

SISAP Indexing Challenge: Approximate kNN search (k=10) on 100M, 768D embeddings of LAION dataset.

- **Goal:** Fastest queries-per-second (QPS) with 90% recall.
- 24-hour time limit
- 512GB RAM limit
- 28-Core Intel(R) Xeon(R) CPU

Our submission performed the best on all subsets: 10M, 30M, and 100M.



Results on the LAION 100M subset of the SISAP 2023 Indexing Competition.