DataLab12.github.io

Hybrid Approximate Nearest Neighbor Indexing and Search (HANNIS) for High-dimensional Image Feature Search

M M Mahabubur Rahman and Dr. Jelena Tešić

Department of Computer Science

TEXAS STATE UNIVERSITY

MOTIVATION

- The brute-force k-Nearest-Neighbor Search in large and high-dimensional data is computationally expensive.
- Approximate Nearest Neighbor (ANN) search is the solution to the sluggish exact k-NN for large and high-dimensional database search.
- We compare and contrast several approximate nearest neighbors search methods in terms of recall, precision, and F1 score for several large high dimensional real datasets.
- We compare the methods for index loading into the memory and query retrieval times.

HANNIS

- We propose hybrid approximate nearest neighbor indexing and search (HANNIS) that retrieves truly similar items in the database, even if the retrieval set is large, and we find that the load items that are comparable to retrieval times.
- HANNIS outperforms all state-of-the-art methods in terms of recall, precision, and F1 score at depths of up to 100 and offers the fastest index loading and consistent retrieval performance.
- We evaluate the performance of HANNIS for efficient and effective retrieval of similar high-dimensional image features.

METHODOLOGY

❖ **Index Building:**

- HANNIS first clusters all the datapoints using k-means++ and pass each cluster with the centroid information to the next phase.
- Each cluster is then arranged into a hierarchical layer of proximity graphs for building indexes with adapted Hierarchical Navigable Small World (HNSW) algorithm.
- All the indexes are then stored into the memory along with their centroid information.

❖ **Improving effectiveness of retrieval:**

- During retrieval HANNIS only loads the index that has smallest query to centroid distance.
- The search starts at the top layer and applies greedy approach to reach local optimum then moves to the lower layer.
- Searching in the lower layer starts with the previous local optimum, and this process continues until the query is reached.

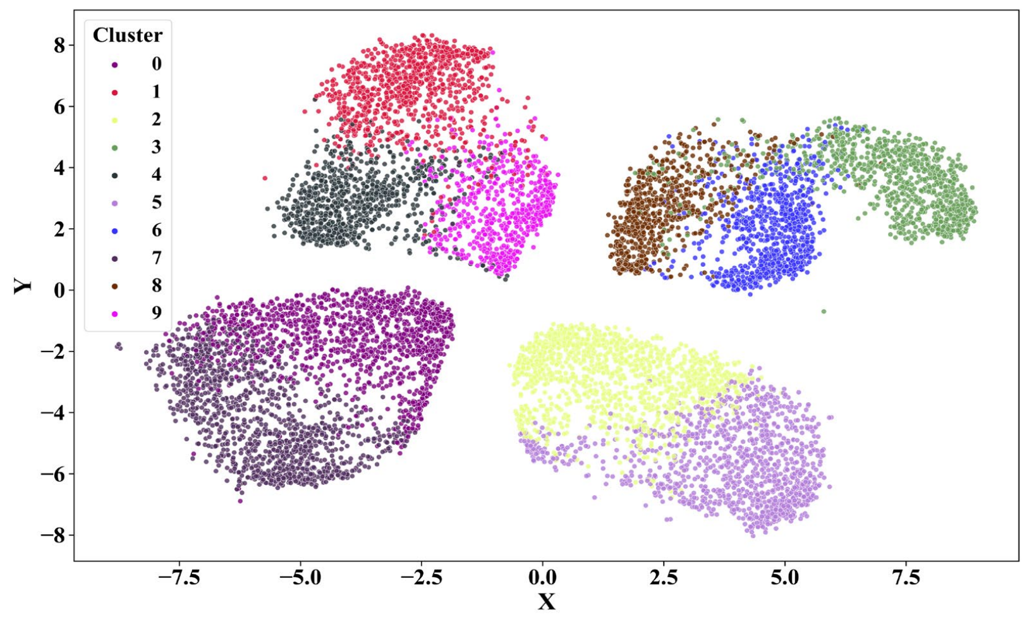

Fig. 1: Clustering with kmeans++

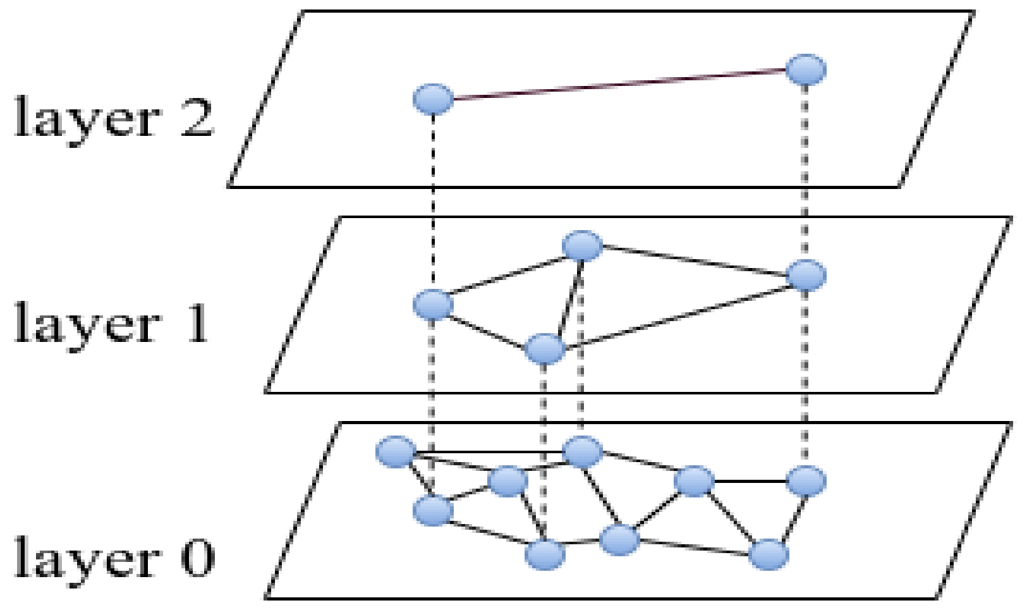

Fig. 2: Indexing with adapted HNSW approach

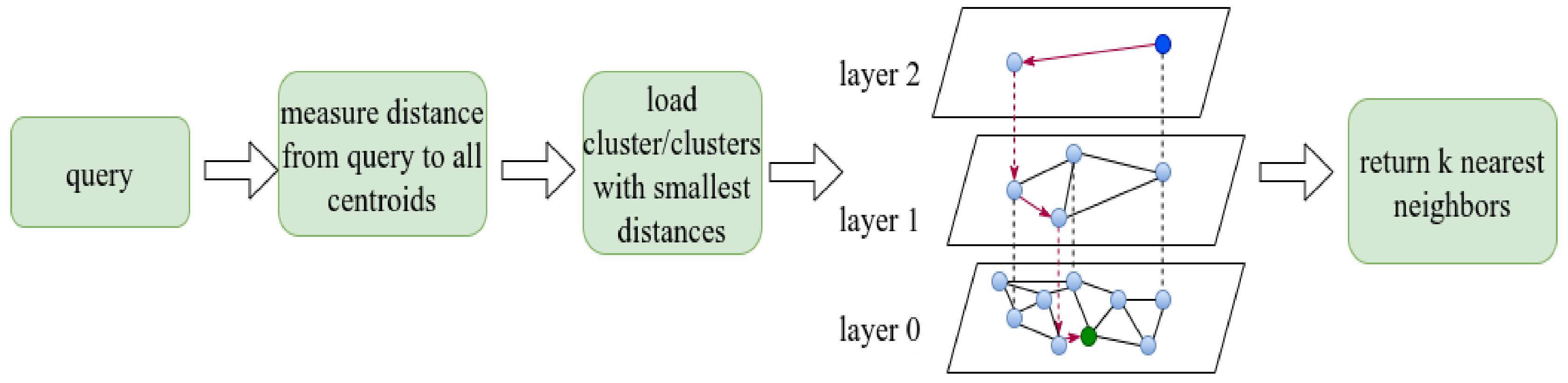
query
measure distance from query to all centroids
load cluster/clusters with smallest distances
layer 2
layer 1
layer 0
return k nearest neighbors

Fig. 3: Proposed HANNIS retrieval pipeline

DATA SETS

Dataset Name	Dimension	Number of instances
DOTA 2.0	1024	2,7 million
SIFT10M (in paper)	*128*	*10 million*
DEEP10M	96	10 million

Table 1: Data sets used for experimental analysis.

DOTA 2.0 EXPERIMENTS

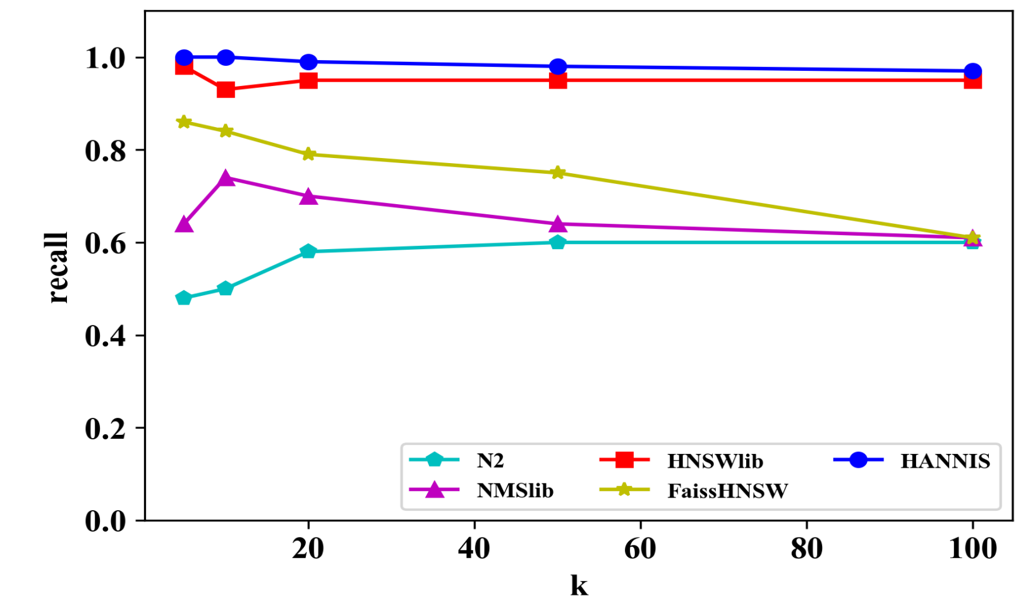

Fig. 4: Recall@k for DOTA 2.0

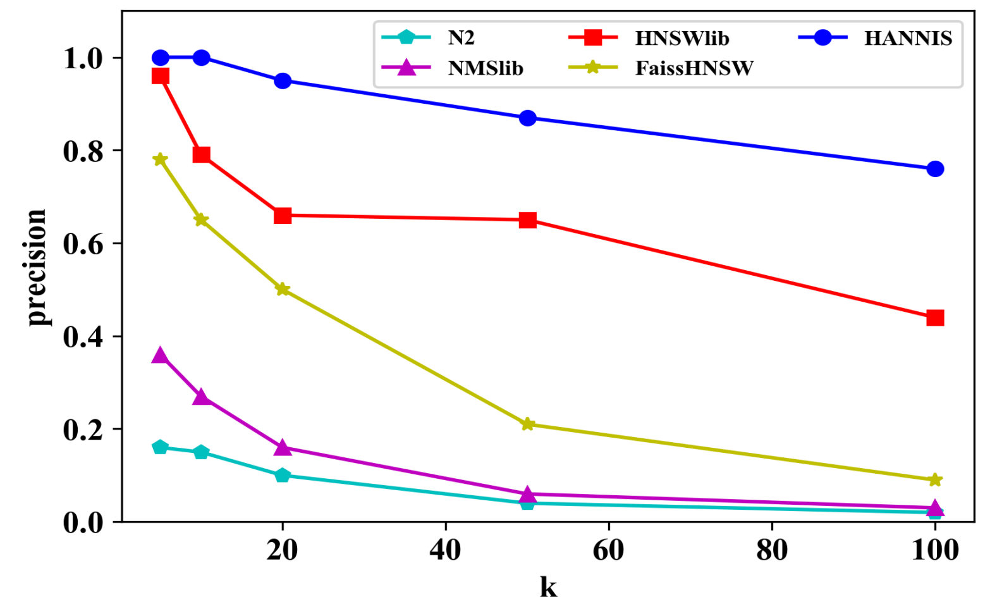

Fig. 5: Precision@k for DOTA 2.0

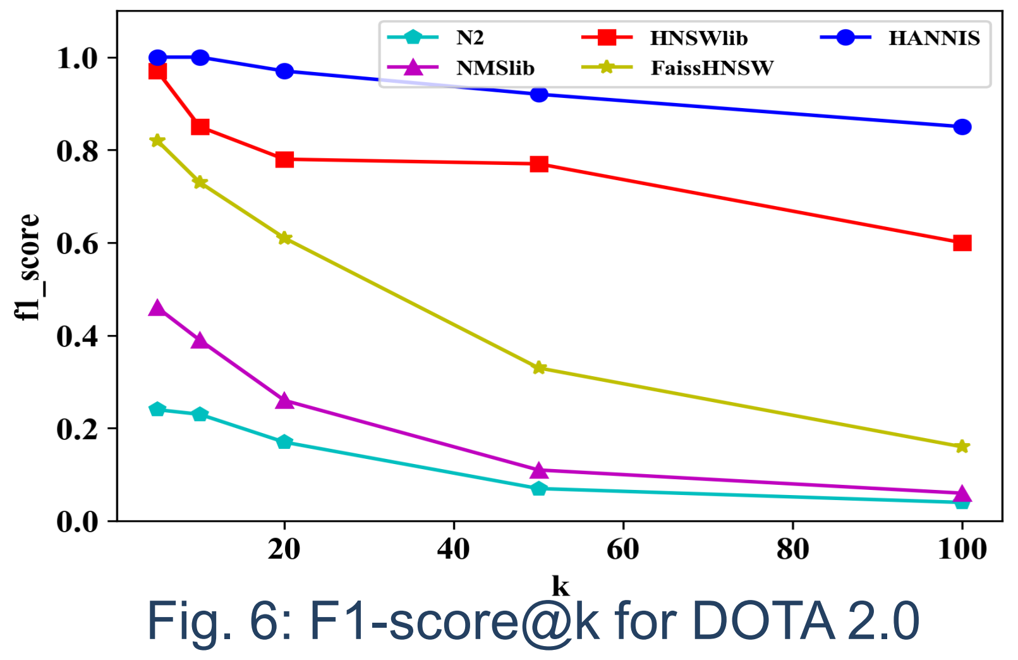

Fig. 6: F1-score@k for DOTA 2.0

DEEP10M EXPERIMENTS

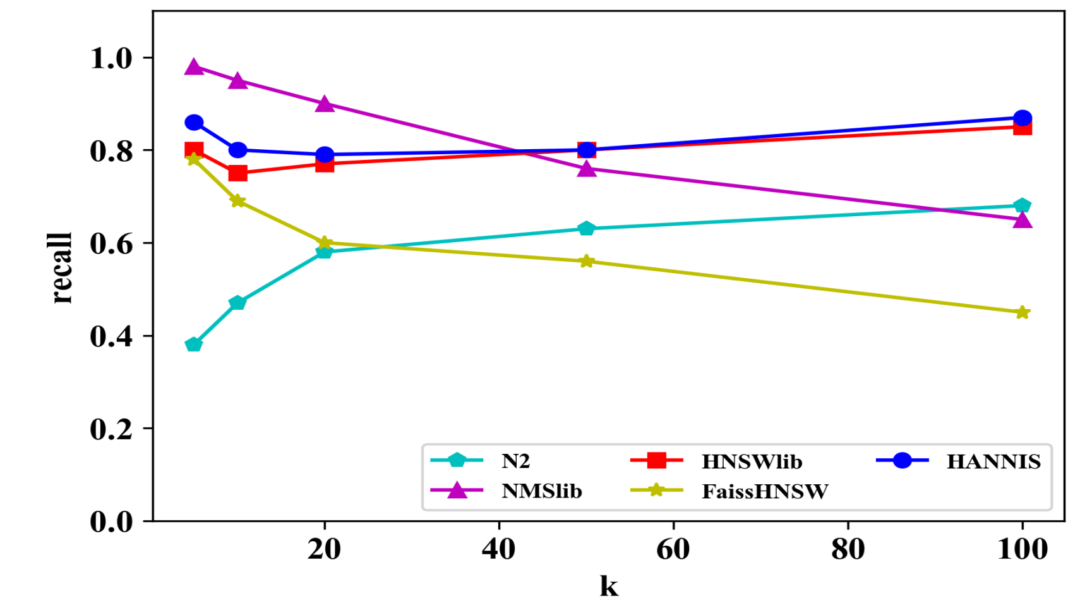

Fig. 10: Recall@k for DEEP10M

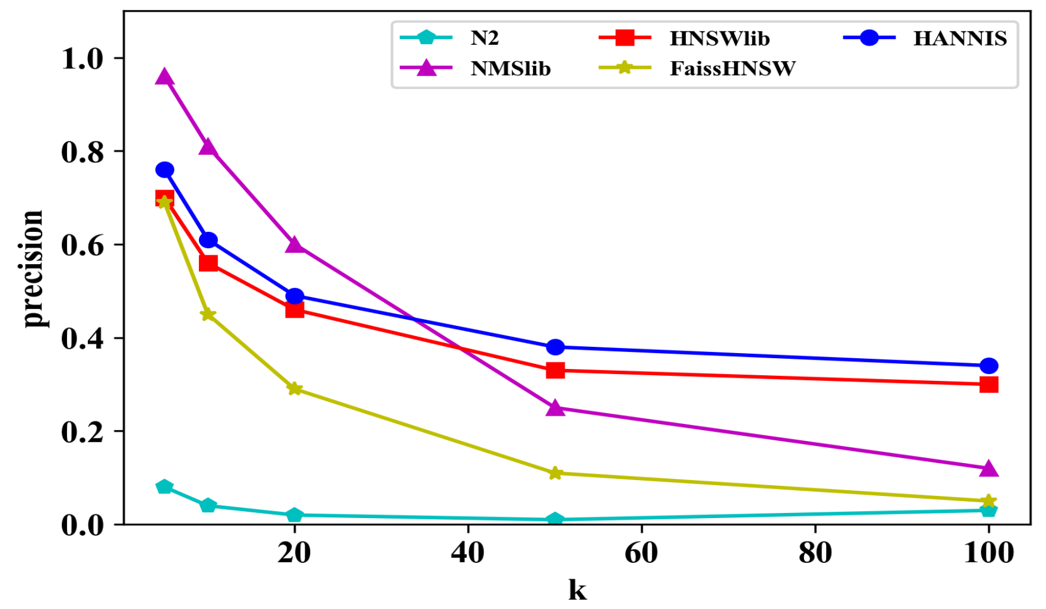

Fig. 11: Precision@k for DEEP10M

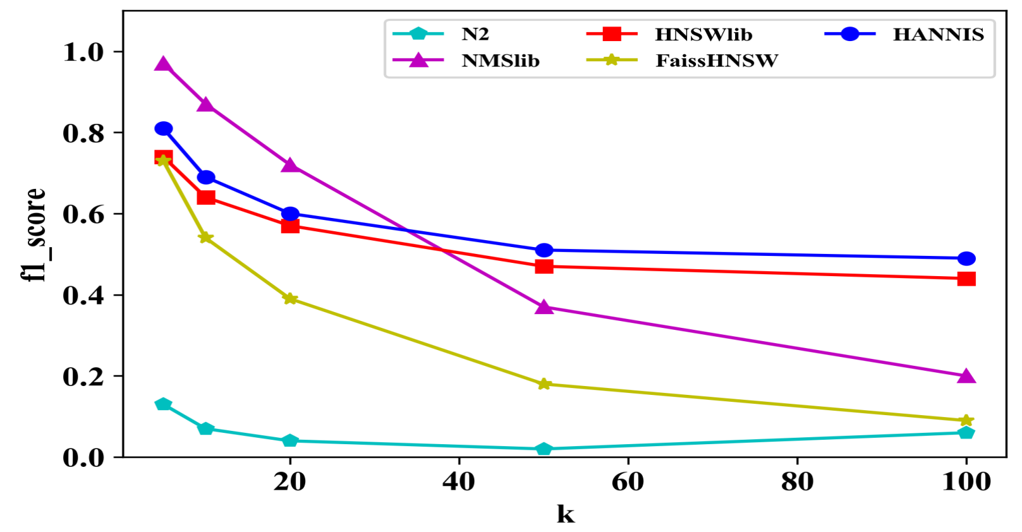

Fig. 12: F1-score@k for DEEP10M

INDEX LOAD AND RETRIEVAL TIMINGS

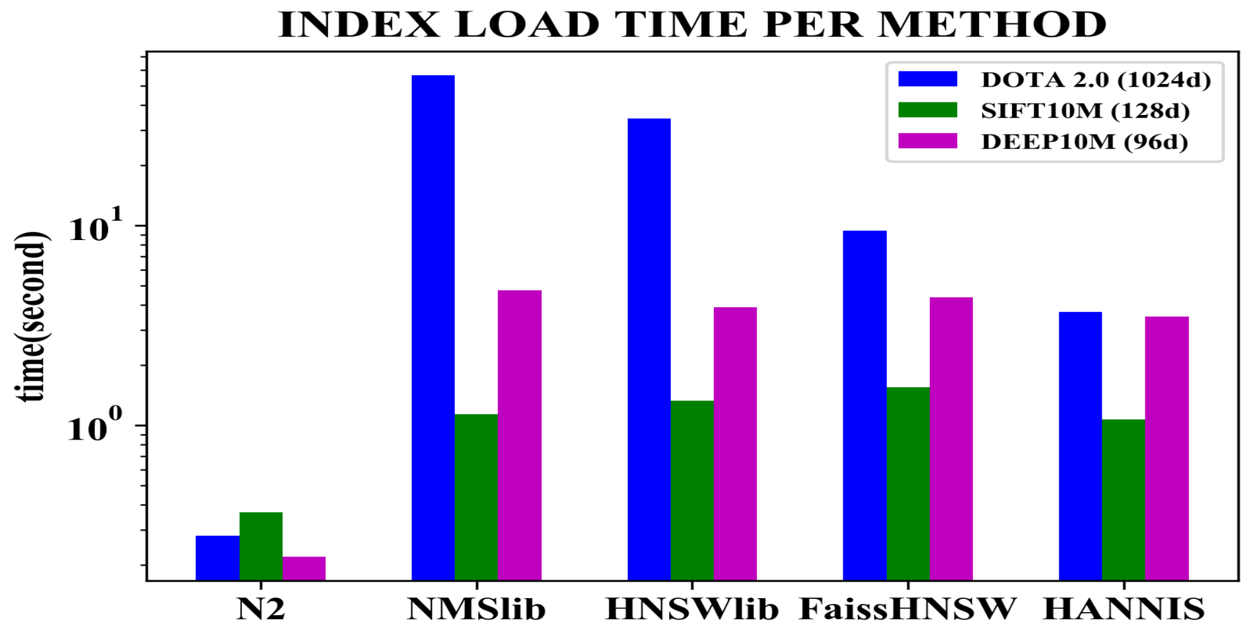

Fig. 13: Index load time

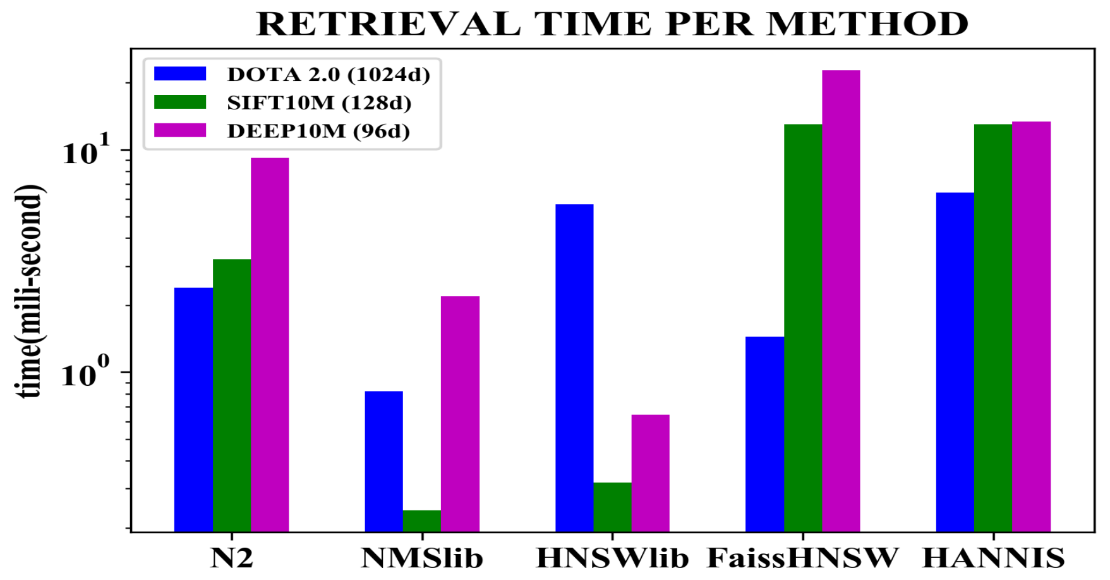

Fig. 14: Retrieval time

Conclusion and Future Work

In this paper, we have demonstrated the efficiency of our approximate nearest-neighbor indexing and search method (HANNIS) to index and search for similar image features from a high-dimensional deep-descriptor data set. HANNIS outperforms the state-of-the-art libraries built on the HNSW algorithm in terms of recall, precision, and F1 score up to 100. HANNIS offers up to 18 times faster index loading into the memory, and the retrieval times are compatible with state-of-the-art libraries.

Acknowledgments

This work is partially supported by the NAVAIR SBIR N68335-18-C-0199. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government paper references.