Wrangling with “Unsupervised Space Partitioning for ANN Search”

A Study on Performance Enhancements

Stathis Kotsis Michael Darmanis

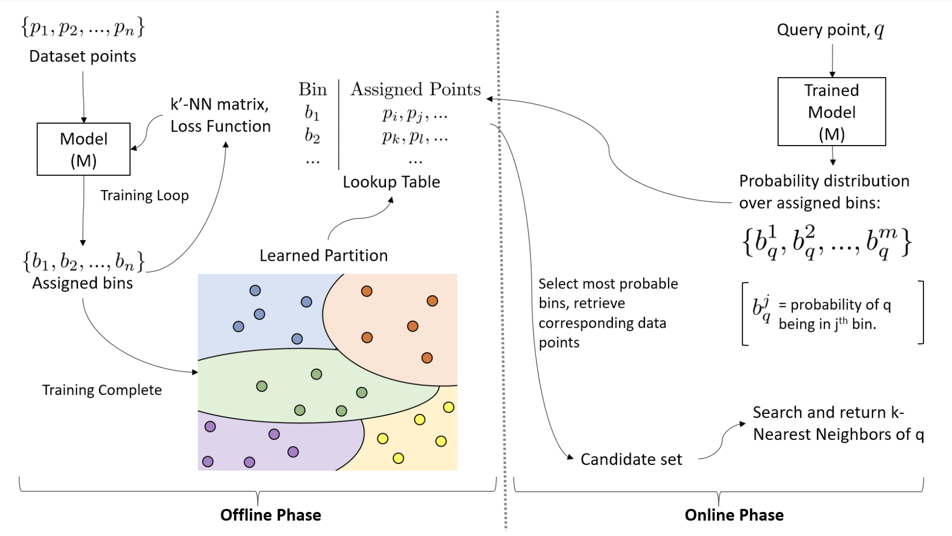
June 30, 2023

M149 Database Systems, NKUA

Introduction

- Approximate Nearest Neighbour (ANN) search is crucial in handling large datasets.
- Traditional methods may not scale well with high-dimensional data.
- This study investigates extensions to the original approach in ”Unsupervised Space Partitioning for Approximate Nearest Neighbour Search”.

Original ANN Scheme



Implementations and Integrations

Indexing

- Hierarchical Navigable Small Worlds[4] (not fully implemented)
- Product vector quantization pipeline[1] (implemented with slow training time, but significant memory savings)

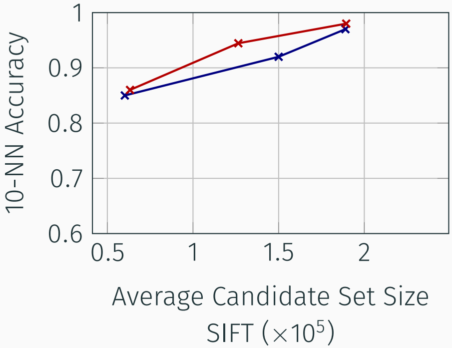
Sketching

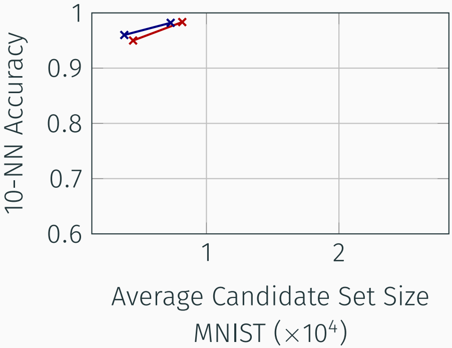
- Principal Component Analysis[5] (PCA)

Model enrichment

- Mahalanobis distance[3]
- Convolutional Neural Networks[2] (CNNs)
- Multi-ensembling paradigm[6]
- Loss function modifications (not fully implemented)

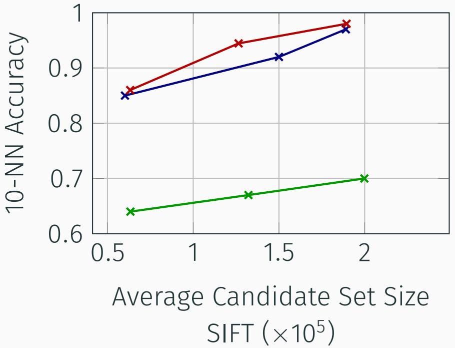
Results: Original, PCA

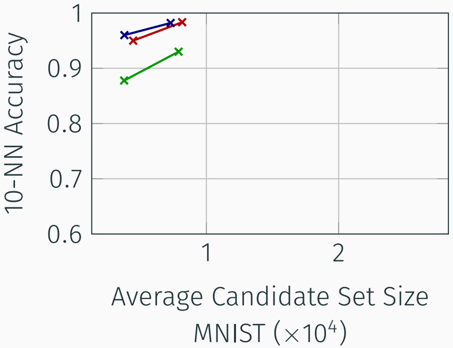




- Achieved 1% accuracy increase at candidate sizes of 190,000-195,000 on SIFT.
- Reduced search time to 0.42 ms on SIFT, a 66% improvement.
- Exhibited similar performance to the original on MNIST.
- Achieved 0.22 ms search time on MNIST, a 70% reduction.

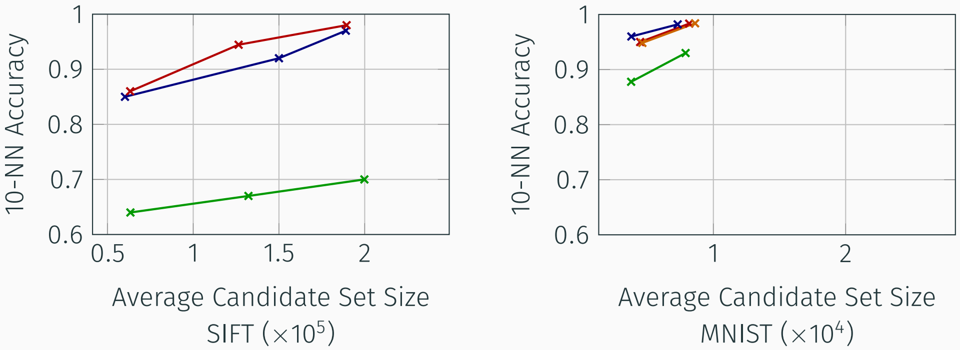
Results: Original, PCA, Mahalanobis





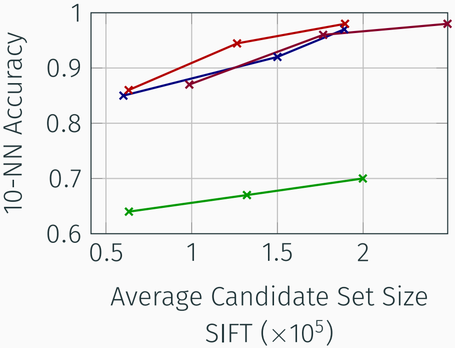
- Matched PCA with 0.42 ms search time on SIFT, a 66% improvement.
- Reduced search time to 0.21 ms on MNIST, a 70% improvement.

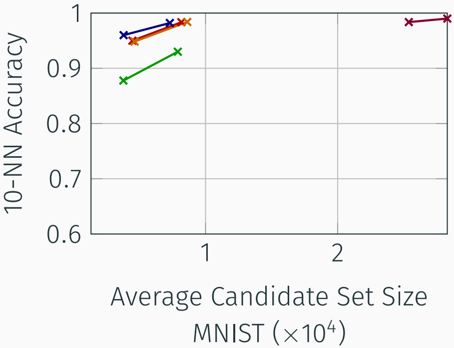
Results: Original, PCA, Mahalanobis, CNN



- Achieved high performance on MNIST with only 4-5 epochs, compared to 40+ epochs for linear models.

Results: Original, PCA, Mahalanobis, CNN, Multi-Ensembling





- Exhibited tendency of creating oversized partitions.
- Demonstrated complexity in integrating different models for high-dimensional data.

Conclusions

- Achieved notable search time reductions with PCA and Mahalanobis; difficulties presented in handling high number of partitions
- Demonstrated CNNs' efficiency; achieved high performance with minimal epochs (original used 90+ with a neural network)
- Revealed multi-ensembling complexities; smarter functions for the combination of models are required (probabilistic, genetic algorithms)
- Proposed future steps in adaptive techniques, alternative loss functions, and hybrid model and unsupervised space partitioning (hnsw and product vector quantisation)

H. Jegou, M. Douze, and C. Schmid.
Product quantization for nearest neighbor search.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(1):117–128, 2011.

A. Krizhevsky, I. Sutskever, and G. E. Hinton.
Imagenet classification with deep convolutional neural networks.
60(6), 2017.

P. C. Mahalanobis.
On the generalized distance in statistics.
Sankhyā: The Indian Journal of Statistics, Series A (2008-), 80:S1–S7, 2018.

Y. A. Malkov and D. A. Yashunin.
Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, 2018.

S. Wold, K. Esbensen, and P. Geladi.
Principal component analysis.
Chemometrics and Intelligent Laboratory Systems, 2(1):37–52, 1987.
Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

Z.-H. Zhou.
Ensemble Methods: Foundations and Algorithms.
Chapman & Hall/CRC, 1st edition, 2012.