



PDF Download
3726302.3730268.pdf
31 March 2026
Total Citations: 0
Total Downloads: 2288

 Latest updates: <https://dl.acm.org/doi/10.1145/3726302.3730268>

ABSTRACT

Efficient Approximate Nearest Neighbor Search on a Raspberry Pi

SILVIO MARTINICO, University of Pisa, Pisa, PI, Italy

FRANCO MARIA NARDINI, Institute of Information Science and Technologies "Alessandro Faedo", Pisa, PI, Italy

COSIMO RULLI, Institute of Information Science and Technologies "Alessandro Faedo", Pisa, PI, Italy

ROSSANO VENTURINI, University of Pisa, Pisa, PI, Italy

Open Access Support provided by:

University of Pisa

Institute of Information Science and Technologies "Alessandro Faedo"

Published: 13 July 2025

Citation in BibTeX format

SIGIR '25: The 48th International ACM
SIGIR Conference on Research and
Development in Information Retrieval
July 13 - 18, 2025
Padua, Italy

Conference Sponsors:
SIGIR

Efficient Approximate Nearest Neighbor Search on a Raspberry Pi

Silvio Martinico*

University of Pisa

Pisa, Italy

silvio.martinico@phd.unipi.it

Cosimo Rulli

ISTI-CNR

Pisa, Italy

cosimo.rulli@isti.cnr.it

Franco Maria Nardini

ISTI-CNR

Pisa, Italy

francomaria.nardini@isti.cnr.it

Rossano Venturini

University of Pisa

Pisa, Italy

rossano.venturini@unipi.it

Abstract

Approximate Nearest Neighbors (ANN) search is a core task in Information Retrieval. However, the high computational demands and reliance on expensive infrastructures limit broader contributions to ANN research. Enabling efficient and effective ANN search on low-resource devices would allow researchers in low-income countries to participate in the ANN community, thereby democratizing the field. Despite its potential, the IR literature offers little work on the feasibility of ANN search under resource constraints. In this proposal, we explore efficient solutions for large-scale ANN search on low-resource devices. We report a preliminary experimentation highlighting current limitations and outlining future challenges.

ACM Reference Format:

Silvio Martinico, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. 2025. Efficient Approximate Nearest Neighbor Search on a Raspberry Pi. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, July 13–18, 2025, Padua, Italy. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3726302.3730268>

1 Introduction

Nearest Neighbors (NN) search addresses the problem of similarity search, a fundamental task in Information Retrieval [10, 14]. In NN search, given a dataset X of vectors and a query point q , the goal is to identify the k elements of X closest to q according to a chosen distance metric. In high-dimensional spaces, exact NN search algorithms struggle to prune the search space effectively. Approximate Nearest Neighbors (ANN) search overcomes these challenges by offering a valuable trade-off between accuracy and efficiency. This approach has become increasingly important with the emergence of powerful learned vector representations (embeddings) produced by Large Language Models, which have transformed the landscape of similarity search.

Traditional ANN search relies on in-memory indexes hosted on expensive server machines with large memory. This dependency

limits access to ANN research, excluding researchers from low-income countries and independent researchers. Exploring ANN search in low-resource settings is important for two main reasons. First, it democratizes access to advanced ANN techniques by enabling experimentation on cheap hardware. Second, it unlocks new applications in environments where computational resources are limited and local processing is essential. Affordable, low-resource devices have strict memory and processing power constraints, requiring novel solutions that still deliver efficient ANN search.

This work investigates the feasibility and limitations of performing ANN search on a Raspberry Pi, a low-resource, affordable device whose cost is two to three orders of magnitude lower than that of the typically employed servers. We report a preliminary experimentation on two state-of-the-art ANN libraries on public datasets. The results allow us to identify key challenges and propose future research directions for ANN search in low-resource environments.

2 Experimental Setup

Datasets and Encoders. We conduct our experiments on two public datasets, selected because their sizes justify the use of approximate search, yet remain within the computational capabilities of the employed device. The QUORA dataset from the BEIR collections [13] consists of 523K documents and 10,000 queries; the SIFT1M [9] dataset is a classical ANN dataset consisting of one million 128-dimensional vectors derived from handcrafted visual descriptors of images, with a set of 10,000 queries. Retrieval is performed using the Euclidean distance on SIFT1M and the Inner Product on QUORA.

For QUORA, we produce both dense and sparse embeddings. The reference metric is nDCG@10. We use SPLADE-v3 [8] as our sparse encoder, which represents the most effective Bert-sized encoder, achieving 81.4 of nDCG@10 on QUORA. As our dense encoder, we employ MINILM [15] (384 dim.). It achieves 87.55 of nDCG@10.

Implementation. Our experimental evaluation relies on two main frameworks: KANNOLO [5] and SEISMIC [2–4], chosen for their performance. In particular, as recently demonstrated in [5], KANNOLO achieves comparable or superior performance compared to state-of-the-art libraries—including Meta’s library for dense ANN search, FAISS [6]—despite its simplicity and modularity. Both KANNOLO and SEISMIC proved [2, 5] to outperform the winner submissions of the “Sparse Track” of the Big ANN challenge at NeurIPS 2023 [12].

KANNOLO is a library for ANN research, designed with modularity in mind and meant to provide a tool for rapid prototyping and testing of new ideas. It supports dense and sparse vector search and uses the graph-based HNSW [11] index for retrieval.

*All authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGIR '25, Padua, Italy

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1592-1/2025/07

<https://doi.org/10.1145/3726302.3730268>

SEISMIC is a new inverted index-based retrieval machinery designed to search over learned sparse representations efficiently. It yields state-of-the-art performance in terms of accuracy/query latency trade-off. κ ANNolo and SEISMIC are written in Rust.

SEISMIC natively supports 16-bit floating point values. We extended κ ANNolo with this functionality, thus nearly halving the memory footprint of the indexes without sacrificing accuracy—a crucial improvement for resource-constrained environments.

Hardware Details. We conduct our retrieval experiments on a Raspberry Pi 3 Model B+ equipped with 4 Cortex-A53 CPUs clocked at 1,400 MHz, 1GiB of RAM, and running Raspberry Pi OS Lite (64-bit). The Rust version is 1.84.1. Experiments were conducted using single-threaded execution. Our indexes, both for the dense and the sparse domain, were previously built on a more powerful server with 512 GiB of RAM and 2 Intel Xeon Silver 4314 clocked at 2.4GHz, with 64 total threads between physical and hyper-threaded ones.

Reproducibility. Source code is available on GitHub.¹

3 Experimental Results

In our experiments, we evaluate the trade-off between search accuracy, i.e., the percentage of true nearest neighbors retrieved, and the number of queries answered in a second (QPS) at a given accuracy rate. The number of results retrieved is $k = 10$.

In Table 1, we report the retrieval results obtained with κ ANNolo on the two dense datasets in a range of accuracy between 91% and 99%. In Figure 1, we report the results of the retrieval performed with both κ ANNolo and SEISMIC on the sparse QUORA embeddings.

In all the tests, even on a low-resource device, we achieve hundreds of queries per second at high accuracy ($> 90\%$), with SEISMIC achieving over 1000 queries per second at up to 92% of accuracy. κ ANNolo gives the only exception to this trend on SIFT1M, going slightly below 100 queries per second for near-exact search ($> 99\%$).

Table 1: κ ANNolo, queries per second at different accuracy cutoffs.

Accuracy (%)	91	93	95	97	99
Quora (MiniLM)	315	288	263	216	119
SIFT1M	237	207	176	136	86

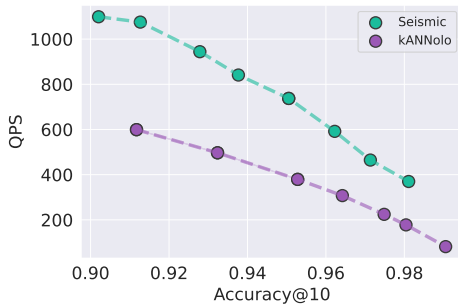


Figure 1: κ ANNolo and SEISMIC on QUORA (SPADE-v3).

¹https://github.com/TusKANNy/ann_low_resource

3.1 Challenges

Results show two main challenges in deploying ANN search on low-resource devices.

Memory Footprint of Dense Embeddings. Dense embeddings typically require a higher dimensionality to achieve the same level of expressiveness as their sparse counterparts, leading to a significant memory footprint. In our experiments, although the QUORA dataset is roughly half the size of SIFT1M and the index built on its dense embeddings employs the sparsest graph index among those adopted (with only 16 neighbors per node), the high dimensionality of these embeddings results in the largest index footprint across all the indexes evaluated. The sparse embeddings—and consequently, their corresponding index—have a significantly smaller memory footprint, which translates into better overall search performance.

Challenges with Quantization. κ ANNolo’s implementation of product quantization (PQ) [7] heavily relies on AVX instructions for pre-computing query-to-centroid distances. This table-based approach, without support for optimized libraries such as BLAS [1] and lacking the necessary hardware acceleration, becomes prohibitively slow, especially on low-performance devices. Consequently, we restricted our tests to using full vectors with half-precision floats. Addressing this limitation is critical, as the potential memory savings offered by vector quantization methods could enable the efficient use of large, learned datasets on low-resource devices.

By outlining these challenges, our work sets the stage for future improvements to better adapt ANN techniques for environments with computational and memory limitations.

4 A Proposal for Future Work

Addressing the highlighted limitations is critical for efficient and effective retrieval on low-resource devices. Improving encoder efficiency could reduce the size of dense embeddings with minimal loss of expressiveness. Potential enhancements to quantization techniques include enabling hardware acceleration on low-resource devices or developing alternative quantization methods that do not depend on precomputed tables. These improvements would further reduce the memory footprint of large, learned datasets and allow the retrieval on affordable platforms, broadening the accessibility and application of ANN techniques. Compression techniques will become fundamental in this scenario. Further investigations could explore the indexing phase, which imposes a substantial computational burden—particularly for graph indexes—thus representing a critical area for optimization in low-resource settings.

Acknowledgements

This work was partially supported by the Horizon Europe RIA “Extreme Food Risk Analytics” (EFRA), grant agreement n. 101093026, by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI” funded by the European Commission under the NextGeneration EU program, by the PNRR ECS00000017 Tuscany Health Ecosystem Spoke 6 “Precision medicine & personalized healthcare” funded by the European Commission under the NextGeneration EU programme, and by the MUR-PRIN 2022 “Algorithmic Problems and Machine Learning”, grant agreement n. 20229BCXNW.

References

- [1] Ronald F. Boisvert and Jack J. Dongarra. 2002. Preface to the special issue on the basic linear algebra subprograms (BLAS). *ACM Trans. Math. Softw.* 28, 2 (June 2002), 133–134. doi:10.1145/567806.567812
- [2] Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. 2024. Efficient inverted indexes for approximate retrieval over learned sparse representations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 152–162.
- [3] Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. 2024. Pairing Clustered Inverted Indexes with κ -NN Graphs for Fast Approximate Retrieval over Learned Sparse Representations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3642–3646.
- [4] Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, Rossano Venturini, and Leonardo Venuta. 2025. Investigating the Scalability of Approximate Sparse Retrieval Algorithms to Massive Datasets. *arXiv preprint arXiv:2501.11628* (2025).
- [5] Leonardo Delfino, Domenico Erriquez, Silvio Martinico, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. 2025. kANNolo: Sweet and Smooth Approximate k-Nearest Neighbors Search. In *Advances in Information Retrieval*. Springer Nature Switzerland, 400–406.
- [6] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281* (2024).
- [7] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. 2011. Product Quantization for Nearest Neighbor Search. *IEEE transactions on pattern analysis and machine intelligence* 33 (01 2011), 117–28. doi:10.1109/TPAMI.2010.57
- [8] Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. SPLADE-v3: New baselines for SPLADE. *arXiv preprint arXiv:2403.06789* (2024).
- [9] D.G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2. 1150–1157 vol.2. doi:10.1109/ICCV.1999.790410
- [10] Craig Macdonald and N. Tonellotto. 2021. On Approximate Nearest Neighbour Selection for Multi-Stage Dense Retrieval. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (2021). doi:10.1145/3459637.3482156
- [11] Yu A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 4 (apr 2020), 824–836. doi:10.1109/TPAMI.2018.2889473
- [12] Harsha Vardhan Simhadri, Martin Aumüller, Amir Ingber, Matthijs Douze, George Williams, Magdalen Dobson Manohar, Dmitry Baranchuk, Edo Liberty, Frank Liu, Ben Landrum, Mazin Karjkar, Laxman Dhulipala, Meng Chen, Yue Chen, Rui Ma, Kai Zhang, Yuzheng Cai, Jiayang Shi, Yizhuo Chen, Weiguo Zheng, Zihao Wan, Jie Yin, and Ben Huang. 2024. Results of the Big ANN: NeurIPS'23 competition. arXiv:2409.17424 [cs.LG] <https://big-ann-benchmarks.com/neurips23.html>
- [13] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- [14] Zhengkai Tu, Wei Yang, Zihang Fu, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy J. Lin. 2020. Approximate Nearest Neighbor Search and Lightweight Dense Vector Reranking in Multi-Stage Retrieval Architectures. *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (2020). doi:10.1145/3409256.3409818
- [15] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MINLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 485, 13 pages.