Reproducibility Report for ACM SIGMOD 2023 Paper: "High-Dimensional Approximate Nearest Neighbor Search: with Reliable and Efficient Distance"

HELEN XU, Georgia Institute of Technology, USA
HONGSHI TAN, National University of Singapore, Singapore
KYLE DEEDS, University of Washington, USA

This report describes the reproducibility process and results for the SIGMOD '23 paper "High-Dimensional Approximate Nearest Neighbor Search: with Reliable and Efficient Distance," which introduces a randomized algorithm called `ADSampling` for approximate k-nearest-neighbor search. We have been able to faithfully reproduce the original paper's finding that `ADSampling` reduces the number of accessed dimensions with little harm to the accuracy.

1 INTRODUCTION

The original paper [1], by Jianyang Gao and Cheng Long from Nanyang Technological University, studied efficient approximate k-nearest-neighbors (AKNN) search. The main observation is that most of the time in almost all AKNN algorithms is spent on distance comparison operations. To speed up the AKNN search, this paper proposes a randomized algorithm called `ADSampling` which reduces runtime wrt the dimensionality and succeeds with high probability.

2 SUBMISSION

The reproducibility submission includes:

- a README file (including a list of dependencies and workflow instructions),
- scripts to download and pre-process data,
- scripts to index the datasets,
- scripts to test performance of KNN search, and
- scripts to plot the datasets.

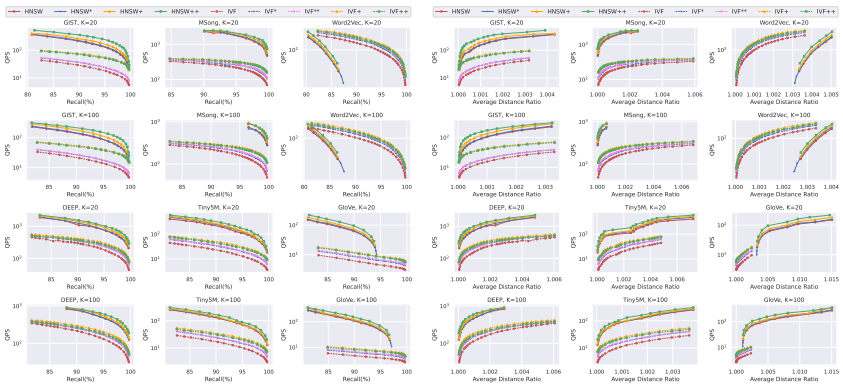
The paper also contains a link to a GitHub repository with code and scripts at https://github.com/gaoj0017/ADSampling, but we did not test this, as it was not mentioned in the reproducibility submission README.

3 HARDWARE AND SOFTWARE ENVIRONMENT

Table 1 compares the original setup of the authors with our local machine. The README estimated a running time of several days for the complete set of experiments, which matches what we observed during our review. Most of the time was spent reproducing Figure 1.

Table 1. Hardware & Software environment

	Paper	Repro Review
CPU	AMD Threadripper PRO 3955WX @ 3.9GHz	Intel Xeon Gold 6248R CPU @ 3.0GHz
Cores	16C/32T	48C/96T
RAM	64GB	384GB
Storage	SSD	SSD
Operating System	Ubuntu 20.04 LTS	Ubuntu 20.04 LTS



(a) Recall (%). (b) Average distance ratio.

Fig. 1. Reproduced variant of Figure 5 from the paper [1] (The figure is transposed in the actual paper).

4 REPRODUCIBILITY EVALUATION

4.1 Process

We followed the instructions in the README to set up the dependencies, prepare the datasets, run the experiments, and plot the results.

We had some difficulty setting up the `faiss` library due to a mismatch in the Python version (it requires Python 3.8-3.10). After some correspondence with the authors to set up the right Python version in Conda, we were able to install all of the dependencies.

In terms of running the experiments, the scripts mostly run smoothly after some modification to point to the correct location for the binaries. Furthermore, we ran into some slight issues with the bash version that required additional changes to the loop syntax in the high-level bash scripts. We corresponded with the authors to make the aforementioned changes.

Once the issues in the scripts were resolved, we were able to run the experiments and produce the expected output. The log files (tsvs) are in the `results/` directory under the top-level `ADSampling-ARI/` folder. Running the plotting scripts directly produces figures in PDF and PNG format and are stored in `ADSampling-ARI/` directory.

4.2 Results

The artifact reproduces Figures 5, 6, and 12 of the paper [1]. Due to space limitations, we only include an example of our reproduced version of Figure 5 (the main results) in the paper in Figure 1. Overall, the reproduced results align with those presented in the paper. The discrepancy in Queries Per Second (QPS) can primarily be attributed to our utilization of a more powerful server.

5 SUMMARY

Overall, the reproducibility submission is comprehensive and successfully replicates the main claims of the paper. We hope that the authors will resolve the slight issues in the dependencies setup and paths in the scripts in their final reproducibility setup.

REFERENCES

[1] Jianyang Gao and Cheng Long. 2023. High-Dimensional Approximate Nearest Neighbor Search: With Reliable and Efficient Distance Comparison Operations. *Proc. ACM Manag. Data* 1, 2, Article 137 (jun 2023), 27 pages. https://doi.org/10.1145/3589282