BANG Library

Contents

1. Dataset Preparation

Download the dataset files in .bin format from big-ann-benchmarks (https://github.com/harsha-simhadri/big-ann-benchmarks/blob/main/neurips21/t3/README.md).

Example: For the SIFT10M dataset, download the dataset using:

```
python create_dataset.py --dataset bigann-10M
```

The base dataset and query files are ready.

2. Graph Construction

Using the above the dataset, generate the Graph Index and Compressed Vectors using DiskANN/Vamana (https://github.com/microsoft/DiskANN/blob/main/workflows/SSD_index.md). The compression factor can be controlled by the '-B' parameter. The higher the value, lower is the compression. Set this to the memory on the GPU that can be used to store compressed vectors.

Build the Vamana Graph Index using:

```
./build_disk_index  --data_type  uint8 --dist_fn l2 --data_path
/mnt/ssd_volume/big-ann-
benchmarks/data/bigann/base.1B.u8bin.crop_nb_10000000 --
index_path_prefix sift10m_index -R 64 -L 200 -B 70 -M 48
```

Run a python script provided in the *BANG* repo (https://github.com/karthik86248/BANG-Billion-Scale-ANN/blob/main/BANG_Base/bang_preprocess.py) to extract required metadata about the Graph Index using:

```
python bang_preprocess.py /mnt/ssd_volume/diskANN-
working/build/tests/sift10m_index_disk.index
/mnt/ssd_volume/diskANN-working/build/tests/sift10m_index_disk.bin
128 1 64
```

We compute the groundtruth using:

```
/compute_groundtruth --data_type uint8 --dist_fn l2 --base_file
/mnt/ssd_volume/big-ann-
benchmarks/data/bigann/base.1B.u8bin.crop_nb_10000000 --query_file
/mnt/ssd_volume/big-ann-benchmarks/data/bigann/bigann-10M --K 10 --
gt_file /mnt/ssd_volume/diskANN-
working/build/tests/sift10m_groundtruth.bin
```

Now, we are ready to start the *BANG* Search.

3. BANG Search.

Download the code from **BANG** Repo : https://github.com/karthik86248/BANG-Billion-Scale-ANN

Navigate to *BANG_Base* directory. Build the code using:

```
mkdir build && cd build && cmake .. && make
```

For example, on the SIFT10M dataset with 10K queries, run the search for 10-recall@10 using:

```
./bang_search /mnt/ssd_volume/diskANN-
working/build/tests/sift10m_index   /mnt/ssd_volume/big-ann-
benchmarks/data/bigann/query.public.10K.u8bin
/mnt/ssd_volume/diskANN-working/build/tests/sift10m_groundtruth.bin
10000 10 uint8 l2
```

Provide various values for worklist length when prompted via the console. The values could be in the range 10 to 152 (assuming recall parameter used is 10).

4. Stock Graph Index files

For the SIFT10K dataset (http://corpus-texmex.irisa.fr/), pre-built DiskANN Graph Index files and required PQ Compressed files are packaged at the following GitHub location : https://github.com/karthik86248/BANG-Billion-Scale-ANN/blob/main/sift10kfiles.tar.gz

Extract the contents of the tarball. Provide the location of the respective files as below to *BANG* search:

```
./bang_search ./sift10kfiles/sift10k_index
./sift10kfiles/siftsmall_query.bin
./sift10kfiles/sift10k_groundtruth.bin  100 10 float l2
```