Custom 8-bit floating point value format for reducing shared memory bank conflict in approximate nearest neighbor search

Hiroyuki Ootomo
Tokyo Institute of Technology
NVIDIA

Approximate Nearest Neighbor Search (ANNS)
- (Exact) nearest neighbor search is computationally enormous.
- ANNS is used in information retrieval, machine learning, etc.
- The search spaces are getting larger in recent years.
- There are some ANNS libraries for many-core CPUs and GPUs.
- Several methods are proposed: Graph-based, IVFPQ, etc.

Goal
For a given a d-dimensional query vector \( q \), obtain \( k \) vectors \( V = \{x_0, x_1, \ldots, x_k\} \) from dataset vectors \( D = \{x_0, x_1, \ldots, x_n\} \) where \( i_1, i_2, \ldots, i_k = k\text{-argsmin}_i(|q - x_i|^2) \).

IVFPQ
- Can compress the dataset.
  - \( \Rightarrow \) Can treat a large dataset on a single device.
- Can compute the norm between a query vector and a dataset vector only by addition instructions once a norm2 fragment lookup table is made.
- Algorithm
  
  **Train phase**
  - Cluster center
  - Cluster vector
  - PQ codebook

  **Search phase**
  - Query vector
  - Query point quantization
  - Norm2 fragment lookup table

Custom 8-bit floating point format for norm representation

**Requirements**
- Only for data storage. No computation between the formats.
- No sign bit is needed since L2-norm is always positive.
- Low overhead for converting from/to FP32.

Proposed format: \( \text{e}5\text{m}3 \) and \( \text{e}4\text{m}4 \)

- While the representation accuracy of \( \text{e}4\text{m}4 \) is not significantly better than \( \text{e}5\text{m}3 \), representable range is only half.
  - \( \Rightarrow \) We use \( \text{e}5\text{m}3 \).

Conclusion
- We have developed custom 8-bit floating point formats for reducing bank conflict in IVFPQ on GPU.
  - The sign bit is omitted.
  - It can be converted from/to FP32 with a few operations.
- We have applied it to IVFPQ and improved the throughput with a little recall degradation.

The accuracy of ANNS

**Recall**

\[ \text{recall} = \frac{|V \cap V_{GT}|}{|V_{GT}|} \]

Where \( V_{GT} \) is the set of the ground truth for the query.

Evaluation

BIGANN 100M dataset (\( d = 128, s = 64, \text{data type} = \text{uint}8 \))

Yandex DEEP 100M dataset (\( d = 96, s = 48, \text{data type} = \text{float} \))

Parameters
- \( (*) \text{num}\_\text{probes} \): The number of clusters picked up in the first stage of the search phase.
- \text{num\_clusters} : 100,000
- \text{batch\_size} : 10,000
- \text{PQ\_bit} : \text{n} = 8

Conclusion
- We have developed custom 8-bit floating point formats for reducing bank conflict in IVFPQ on GPU.
  - The sign bit is omitted.
  - It can be converted from/to FP32 with a few operations.
- We have applied it to IVFPQ and improved the throughput with a little recall degradation.