04

Accelerating ScaNN via pruning-based vector quantization

Do not compute: Fast approach for vector search

Abstract

Quantization, which replaces vectors with codewords, is widely used to enable fast, accurate inner-product-based similarity search over large-scale data. ScaNN is a popular approach for quantization. ScaNN computes the quantization error between each vector and all possible codewords to select the codeword with the smallest error, achieving high approximation accuracy. However, since ScaNN requires error computation with all codewords, it incurs a high computational cost, making quantization extremely slow for large-scale datasets. The proposed approach uses upper bounds on quantization errors and efficiently evaluates them to prune codeword candidates. This significantly reduces the number of error computations required. As a result, the proposed approach can substantially accelerate vector quantization while preserving ScaNN's search accuracy. Consequently, it facilitates practical large-scale data processing in applications such as image retrieval and natural language processing.

Maximum Inner Product Search

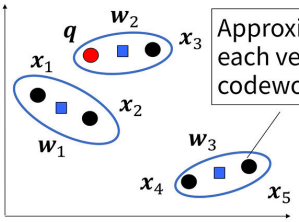
Find vector $\boldsymbol{x}$ of maximum inner product for query $\boldsymbol{q}$

⭕ Various application (e.g., image search and NLP)

❌ High CPU cost for inner products of the query and vectors in the database

Quantization: approximate vector $\boldsymbol{x}$ with codeword $\boldsymbol{w}$ for search

⭕ Improved efficiency ❌ Degraded search accuracy



Approximate each vector to a codeword

$\boldsymbol{w}$ is set according to data distribution

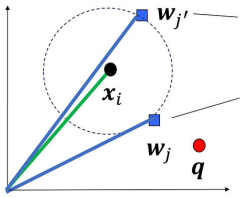
Precomputed inner products

	w_1	w_2	w_3
w_1	⋯	⋯	⋯
w_2	⋯	⋯	⋯
w_3	⋯	⋯	⋯

$O(1)$time to obtain inner product

Existing Approach: ScaNN

Idea: Preserve vector norms and reduce inner product error to improve search accuracy



Large change of the norm (length of green and blue lines)

Norm changes little, and small inner product error with $\boldsymbol{q}$

Approach: Assign to the codeword minimizing error function $l(\boldsymbol{x}_i, \boldsymbol{w}_j)$

⭕ Error function is low ⇔ Norm is preserved

❌ High CPU cost for error function

$$l(\boldsymbol{x}_i, \boldsymbol{w}_j) = \frac{h_i - 1}{\|\boldsymbol{x}_i\|^2}\left(p(\boldsymbol{x}_i, \boldsymbol{w}_j)\right)^2 - 2h_i p(\boldsymbol{x}_i, \boldsymbol{w}_j) + h_i\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{w}_j\|^2$$

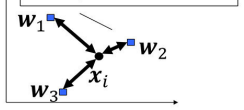
$h_i\ (>1)$: hyper parameter, $p(\boldsymbol{x}_i, \boldsymbol{w}_j)$: inner product of $\boldsymbol{x}_i$ and $\boldsymbol{w}_j$

Our approach: computation pruning

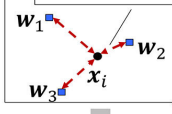
Prune unnecessary computations by efficiently obtaining upper bound $\bar{l}(\boldsymbol{x}_i, \boldsymbol{w}_j)$ and lower bound $\underline{l}(\boldsymbol{x}_i, \boldsymbol{w}_j)$

- Existing approach
- Our approach

Compute function



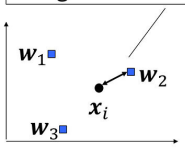
Efficiently compute bounds



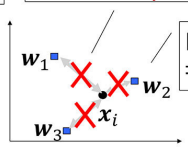
Upper bound of $\boldsymbol{w}_2$ is smaller than lower bounds of $\boldsymbol{w}_1$ and $\boldsymbol{w}_3$

$\bar{l}(\boldsymbol{x}_i, \boldsymbol{w}_2) < \underline{l}(\boldsymbol{x}_i, \boldsymbol{w}_1),\ \underline{l}(\boldsymbol{x}_i, \boldsymbol{w}_3)$

Assign codeword $\boldsymbol{w}_2$

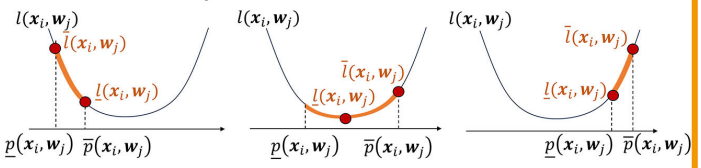


Prune computations of functions



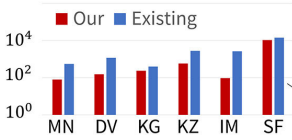
Identify codeword $\boldsymbol{w}_2$ ⇒Same results

Function $l(\boldsymbol{x}_i, \boldsymbol{w}_j)$ is quadratic for inner product $p(\boldsymbol{x}_i, \boldsymbol{w}_j)$ ⇒ $\bar{l}(\boldsymbol{x}_i, \boldsymbol{w}_j)$ and $\underline{l}(\boldsymbol{x}_i, \boldsymbol{w}_j)$ are obtained from upper/lower bounds of inner product



Achieve speedup without sacrificing search accuracy

- Processing time [s]



Up to 28 times faster

- Search accuracy [%]

	MN	DV	KG	KZ	IM	SF
Our	56.5	35.5	59.7	53.5	6.28	69.8
Existing	56.5	35.5	59.7	53.5	6.28	69.8

Same search accuracy

- Experimental data (Number of vectors ×Length of vectors)

MN：MNIST (70,000 × 780)
DV：Devnagari-Script (92,000 ×1,024)
KG：kaggle (200,000 × 202)
KZ：Kuzushiji (270,912 × 784)
IM：imagenet (2,340,373 × 150)
SF：sift10m (10,000,000 × 128)

References

[1] Y. Fujiwara, Á. López, Y. Ida, A. Kumagai, M. Nakano, M. Nakatsuji, A. Kimura, “Fast Vector Quantization Algorithm for ScaNN”, in *Proc KDD*, 2026.

Contact

Yasuhiro Fujiwara, Recognition Research Group, Media Information Laboratory