LeanVec: Searching vectors faster by making them fit

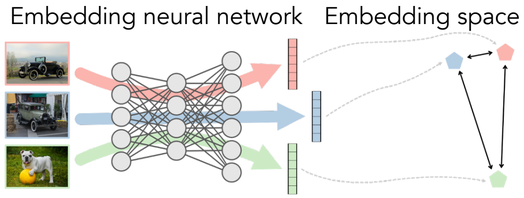
[1] DataStax
[2] Intel Labs
[3] Microsoft

Mariano Tepper[1], Ishwar Singh Bhati[2], Cecilia Aguerrebere[2], Mark Hildebrand[3]*, Ted Willke[1]**

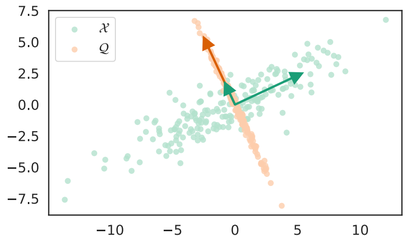
** Work done while at Intel Labs*

Problem Statement

Finding semantically related vectors among millions or billions is crucial for applications like recommender systems, RAG, NLP, among many others.

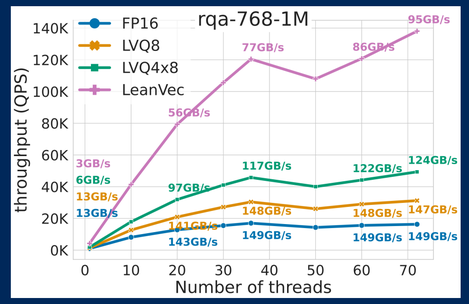


Graph-based indices are SOTA, but random access patterns and increasing vector dimensionality limit their potential.



Considering the distribution of queries is key for dimensionality reduction

New SOTA for vector search for in-distribution and out-of-distribution queries



LeanVec

$$\min_{\mathbf{A},\mathbf{B}\in \mathrm{St}(D,d)} \left\| \mathbf{Q}^\top \mathbf{A}^\top \mathbf{B}\mathbf{X} - \mathbf{Q}^\top \mathbf{X} \right\|_F^2$$

A, B row-orthonormal matrices
Q matrix with query vectors
X matrix with database vectors

We propose two solvers:

Frank-Wolfe block-coordinate descent	Eigen-search
Convex relaxation of the orthonormality constraints on **A, B**	Finds the best convex combination of $\mathbf{Q}^T\mathbf{Q}$ and $\mathbf{X}^T\mathbf{X}$
✓ Theoretical guarantees	✓ Fast algorithm

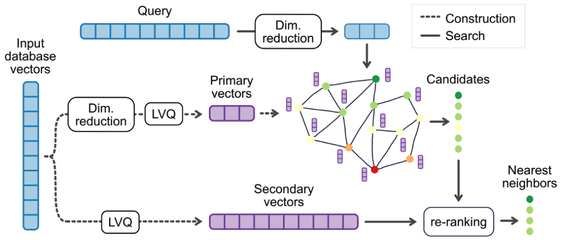
Using LeanVec for graph search

- Two level for speed and accuracy

$$\langle \mathbf{q}, \mathbf{x} \rangle \approx \langle \mathbf{Aq}, \mathrm{quant}(\mathbf{Bx}) \rangle$$

Computed at search. Linear approx for speed

Precomputed during indexing

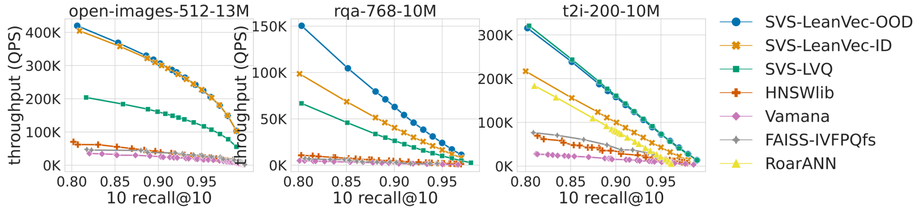


Contributions

- High-quality graphs up to 4.9x faster than SOTA
- LeanVec-ID (query distr. agnostic) up to 3.6x faster search
- LeanVec-OOD, new optimal projection with two fast optimization algorithms, up to 5.4x faster search
- Integrated into Scalable Vector Search for reproducibility

Results

Faster search



Faster construction