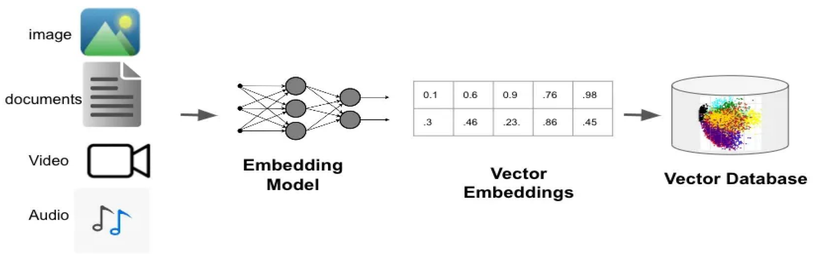
LEANN: A Low-Storage Vector Index for Personal Devices

Yichuan Wang, Shu Liu, Zhifei Li, Yongji Wu, Ziming Mao, Yilong Zhao, Xiao Yan, Zhiying Xu, Yang Zhou, Ion Stoica, Sewon Min, Matei Zaharia, Joseph Gonzalez

Berkeley sky

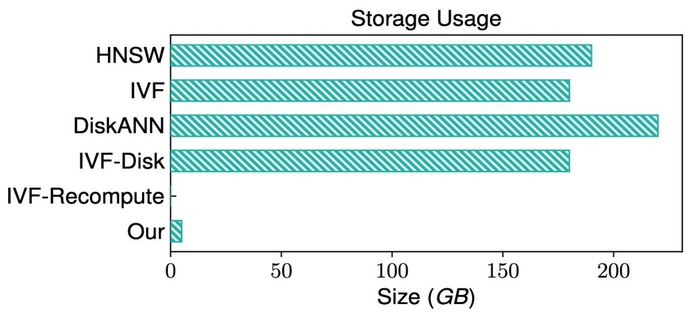
Vector Index on Edge Devices

- With generative AI advances, **semantic search** now outperform keyword-based search – power by **approximate nearest neighbor (ANN) search in vector space.**
- **Edge applications of semantic search**: personalized search, on-device assistants, and privacy-preserving retrieval (e.g., RAG over local data)



Challenges: Deploy ANN on Edge Devices

- **Huge storage (SSD) consumption**:
  - High-dimensional embedding vectors
  - Complex graph structures



75GB of personal docs require over 200GB storage, a **270% increase that is impractical for PC**.

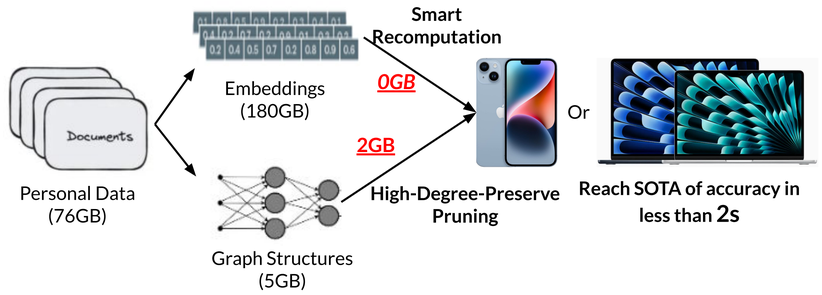
(1) *Naive Compression*: quantization or simply recomputation on-demand, greatly increasing latency.

(2) *Naive Graph Pruning*: reduce the connectivity of graph, hurt the ANNS quality w/ certain compute budget.

Naive solution hurts both **accuracy** and **efficiency** of ANNS.

Key Methods

- **Fast HNSW Graph-Based Recomputation** with **two-level search and batched execution** eliminates the need for storing pre-computed embeddings, while keeping latency low.
- **High-Degree-Preserving Graph Pruning** removes redundancy in graph, cutting storage with minimal quality loss.



Setup

Hardware: NVIDIA A10 & Apple Macbook M1

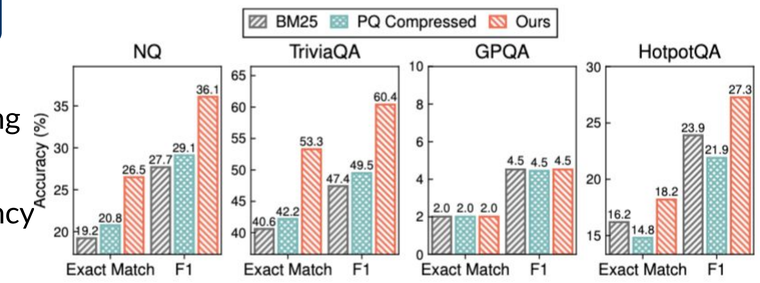
IR Datasets: NQ, Trivia-QA, GPQA, HotpotQA

Baselines: HNSW, IVF, DiskANN, IVF-Disk, PQ-Compression, EdgeRAG(IVF-Recompute)

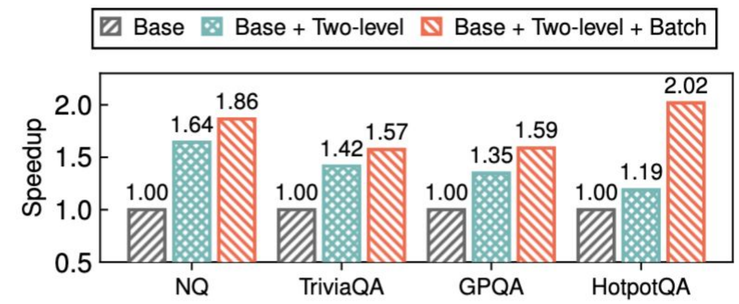
Evaluation Results

1.[Main result] Latency Storage Trade-off

2.[Main result] Accuracy



3.[Ablation study] Latency Optimization



4.[Ablation study] Graph Pruning