

Computational SSD For Semantic Image Retrieval

Vishwas Saxena,
Senior Technologist, Western Digital



Semantic Image Retrieval

- There are multiple approaches for image retrieval and some of these deep learning techniques are Tag based retrieval, User query on captions derived from an image, User query on Scene graphs derived from an image etc.
- The analytics derived from the image is stored in database. Different methods use different type of databases. For example, a tag-based search uses a Key Value DB, Scene graph-based retrieval uses a graph DB.

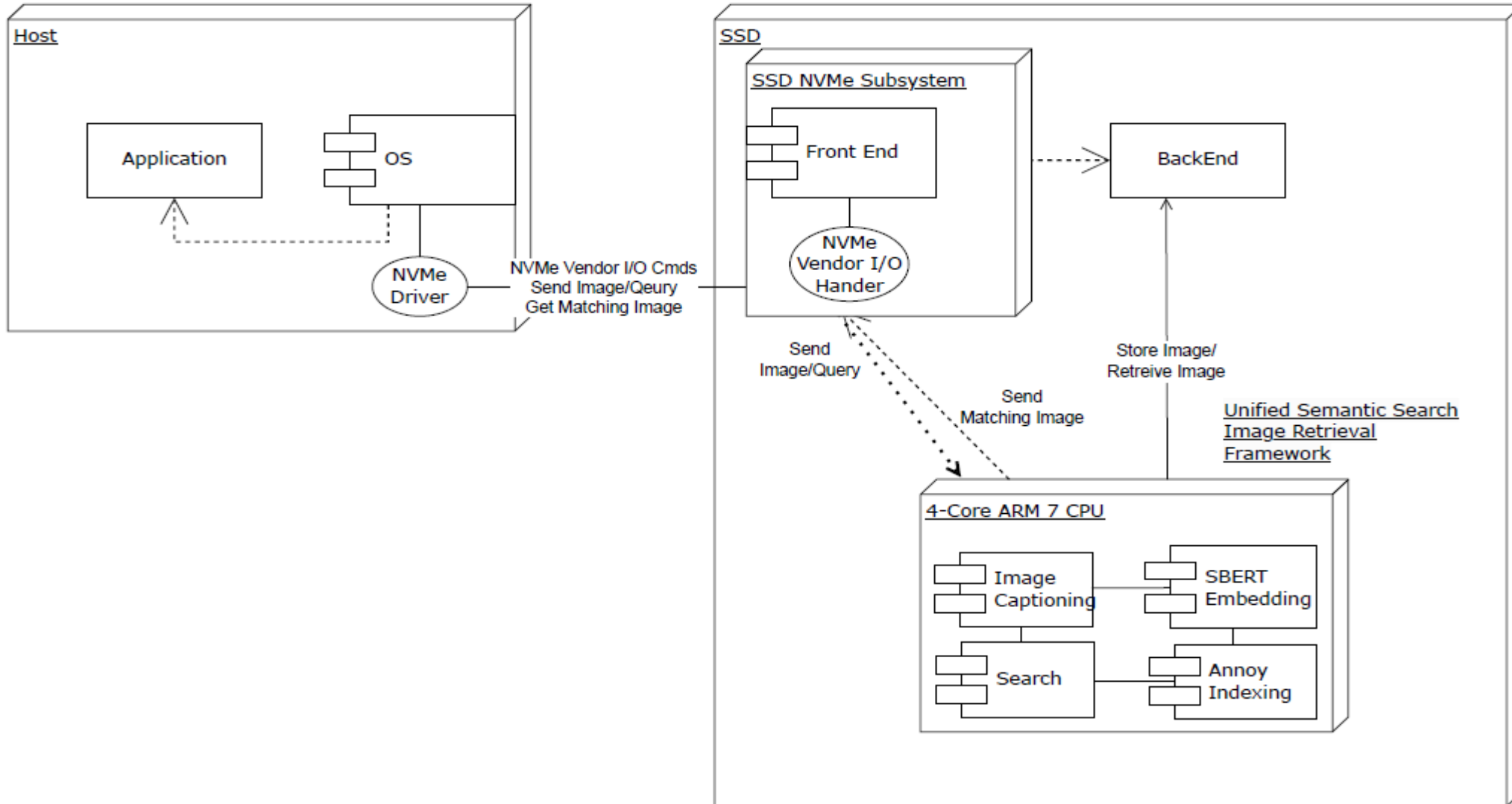
Problems in Semantic Image Retrieval

- No Standard format to store analytics from deep learning algorithms.
- Every algorithm requires a different kind of database to store the data.
- Maintaining a database on low compute device adds to the complexity of design.

Constraints

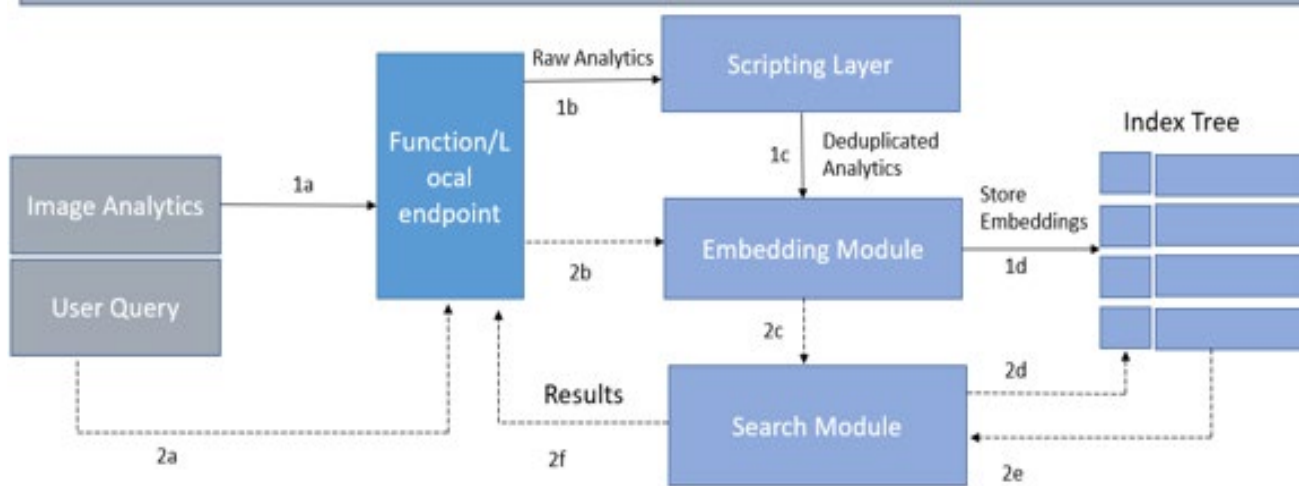
- Low Memory Footprint
 - Deep Learning models consume a lot of memory
- High Level of Accuracy
 - Higher accuracy requires usage of heavier models
- Low Latency for Search
 - RAM requirement grows exponentially as number of images in database increase, Search time takes a hit

SSD with Unified Semantic Search Image Retrieval Framework



Solution

- Image Captioning + Deduplicated Triplets
- 768-dimension encoding – SBERT Embedding Model
- Index Tree using Annoy indexing



Siamese BERT – Sentence Transformer Model

SBERT implements an additional pooling layer on top of BERT. There are three different pooling strategies implemented by SBERT.

What we get after the pooling layer is the embedding vector of a text that has 768 dimensions. This embedding then can be compared to each other with pairwise distance or cosine similarity

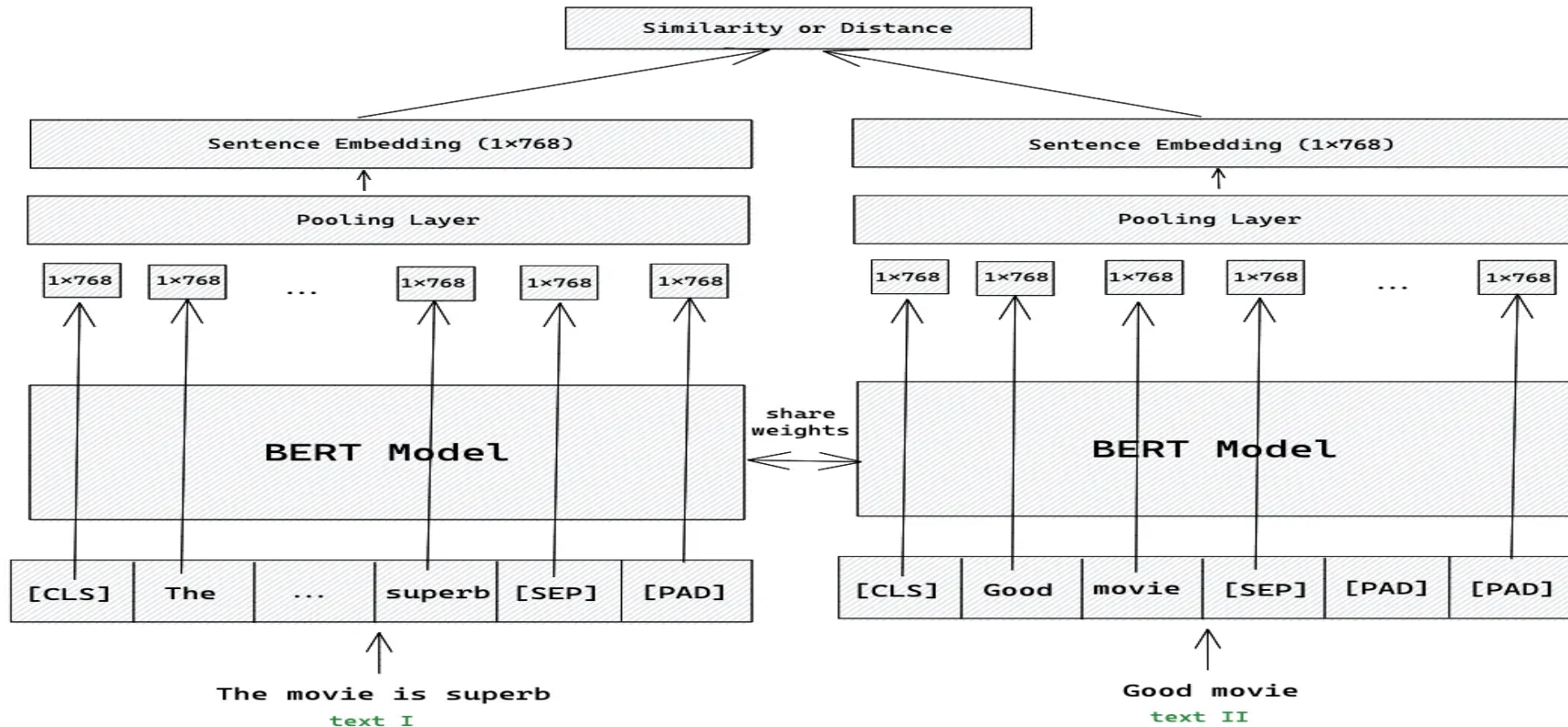
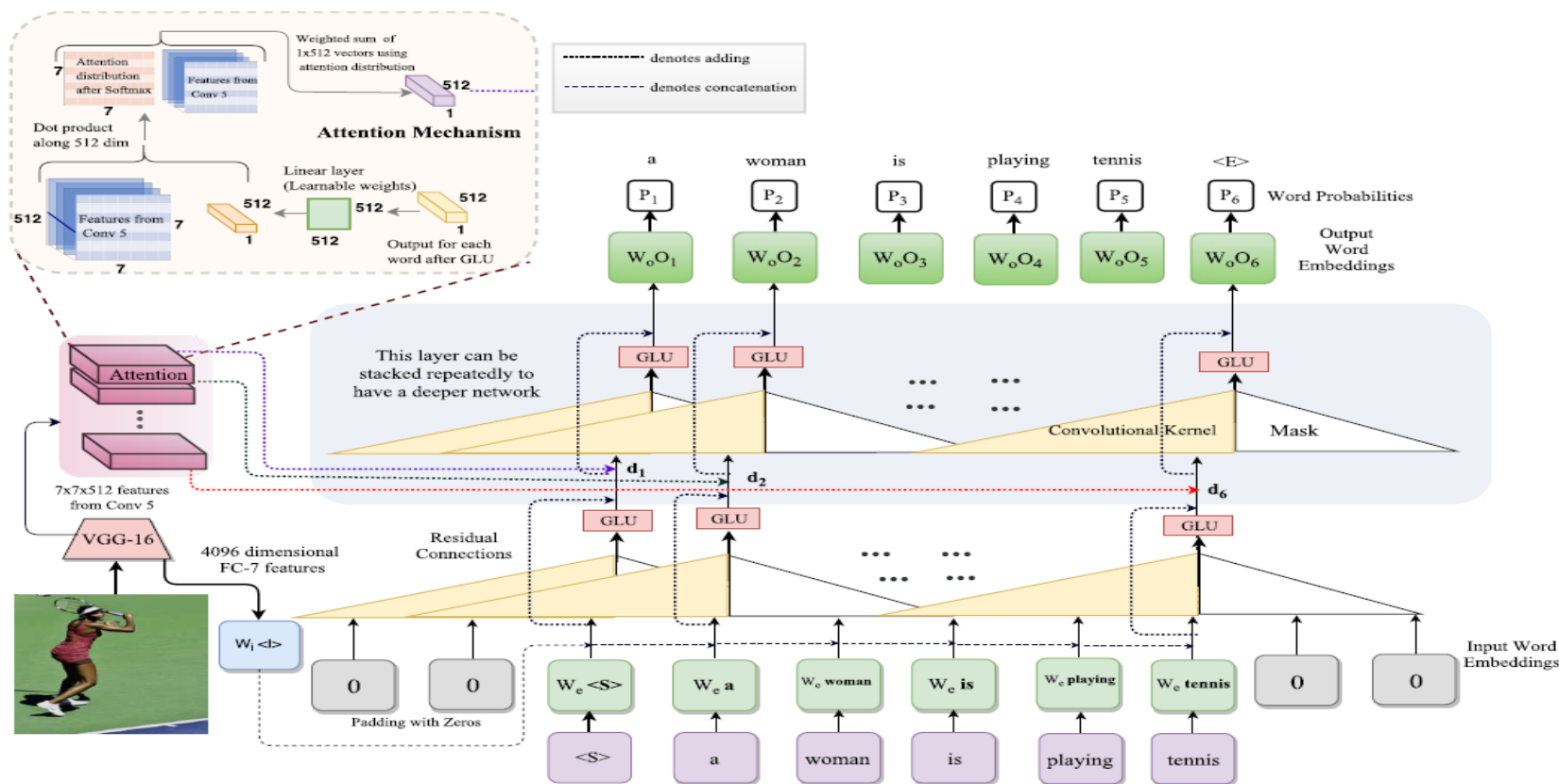


Image Captioning

A textual description of the entire image is generated and a few keywords from the textual description are tagged along with image.



Design Decisions

Decision	Rationale
Nature of Embedding	Selected context based sentence embedding model over word based embedding model. This led to better accuracy with increased RAM savings
Embedding Module	Selected “Paraphrase-MiniLM-L3-v2” SBERT Model. Consumes 375 Mb RAM. Also, for a 3.19% drop in accuracy, we could get a 50% savings in RAM
Index Tree	Store embeddings in an Index Tree to reduce storage of Embeddings in RAM. Concluded to confine the number of trees to 100 . Finalized “Angular” metric with ntree = 25
Angular Threshold	Selected Angular distance threshold = 1 to separate relevant queries from irrelevant queries

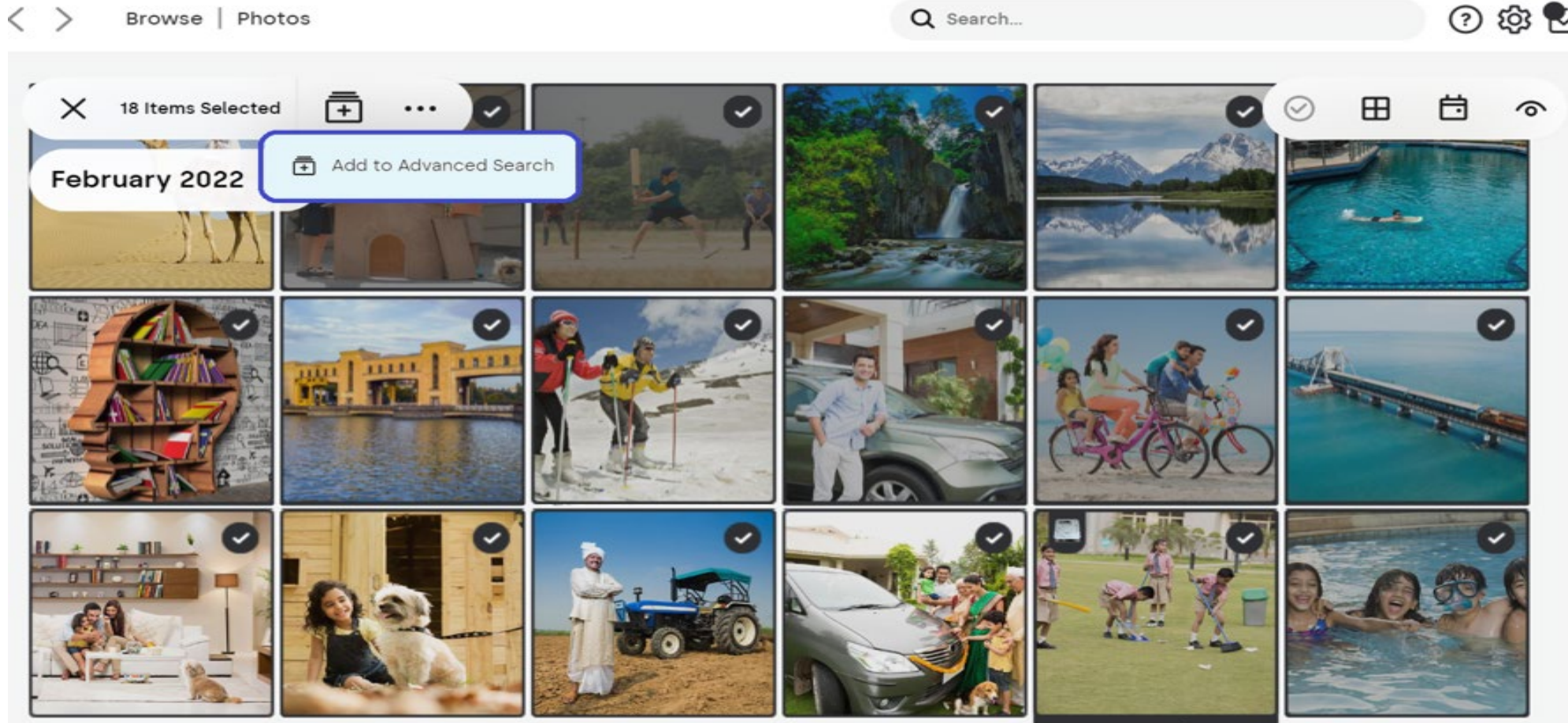
Indexing of embedding using Annoy

- An index was created from the 1.2 million embeddings and stored to SSD.
- This index was later loaded in memory of Separate 4 Core ARM CPU inside SSD and used for search functionality
- For each of the Deep Learning model, an index was created using Annoy. Sample datasets were created which had queries relevant and not relevant to the embeddings present in Image embeddings. 5 distance metrics (Angular, Euclidean, Dot, Hamming and Manhattan) with Number of Trees (10,25,50,75,100) were evaluated to arrive at final configuration

Search on embedded systems

- a. In one batch, Index stored for 1.2 million embeddings took 770 Mb disk space and when loaded in memory took less than 1Mb RAM
- b. Index Creation time and Search time was computed
- c. In one batch, Indexes beyond 1.2 million embeddings could not be created on SSD with separate 4 core ARM CPU
- d. In one batch, trees beyond 100 could not be created on SSD with separate 4 core ARM CPU


Application View – Add to Advanced Search




Application View – Semantic Image Retrieval




< > Search Results for "snow on mountains"

Q snow on mountains X ? ⚙️ ✓

More Filters 

Search Results for "snow on mountains"

Photos & Videos 

	822308	JPEG Image	Feb 1, 2022	Windows (C:)	104.19 KB
	868537	JPEG Image	Feb 2, 2022	Windows (C:)	200.89 KB
	868454	JPEG Image	Feb 2, 2022	Windows (C:)	194.83 KB



COMPUTE + MEMORY + STORAGE SUMMIT

Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023



Summary

- Uniform framework for semantic image retrieval
- Separate 4 core ARM CPU to do Image Captioning, embedding, and Searching
- Annoy Indexing for fast search on separate 4 core ARM CPU



COMPUTE + MEMORY + STORAGE SUMMIT

Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023



Please take a moment to rate this session.

Your feedback is important to us.