

muves

Multilingual and Multimodal Vector Search with Hardware Acceleration



@AarneTalman



@DmitryKan

Who are we?



Dmitry Kan

Senior Product Manager - TomTom

Principal AI Scientist - Silo AI

Host of the Vector Podcast

<https://dmitry-kan.medium.com>



Aarne Talman

Lead AI Engineer - Silo AI

Co-founder and CEO - Basement AI

PhD Student - University of Helsinki



Outline

1. Motivation and background
2. What is Muves?
3. Implementation details of our demo with GSI Technology Inc
4. Demo / Examples
5. Results from relevancy testing with Quepid
6. Lessons learned

Multimodal search

Ng also finds so-called [multimodal AI](#), or combining different forms of inputs, such as text and images, to be promising. Over the last decade, the focus was on building and perfecting algorithms for a single modality. Now that the AI community is much bigger, and progress has been made, he agreed, it makes sense to pursue this direction.



Keyword search

- *Examples: Elasticsearch, OpenSearch, Solr.*
- *Rely on matching of search terms to text in an inverted index.*
- *Makes it difficult to find items with similar meaning but containing different keywords.*
- *Not directly suitable for multimodal or multilingual search.*

EXAMPLE

Query: *A bear **eating a fish** by a river*

Result: *heron **eating a fish***



Vector search

- *Utilises neural networks models to represent objects (like text and images) and queries as high-dimensional vectors.*
- *Ranking based on vector similarity.*
- *Allows finding items with similar meaning or of different modality.*

EXAMPLE

Query: *A bear **eating a fish** by a river*

Query vector: *[0.072893, -0.277076, 0.201384, ...]*

Result vector: *[0.004142, -0.022811, 0.019714 ...]*

Result:



Hardware acceleration using GSI APU can significantly improve query speed



Published in **Towards Data Science** · Mar 15, 2021 ★

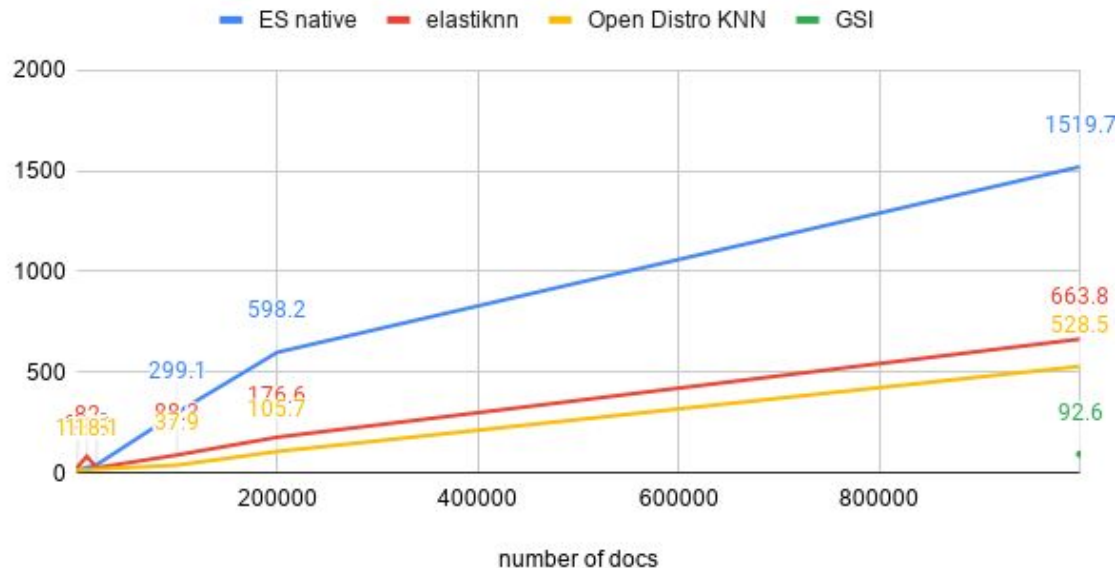
Speeding up BERT Search in Elasticsearch

Neural Search in Elasticsearch: from vanilla to KNN to hardware acceleration — In two previous blog posts on my journey with BERT: Neural Search with BERT and Solr and Fun with Apache Lucene and BER...

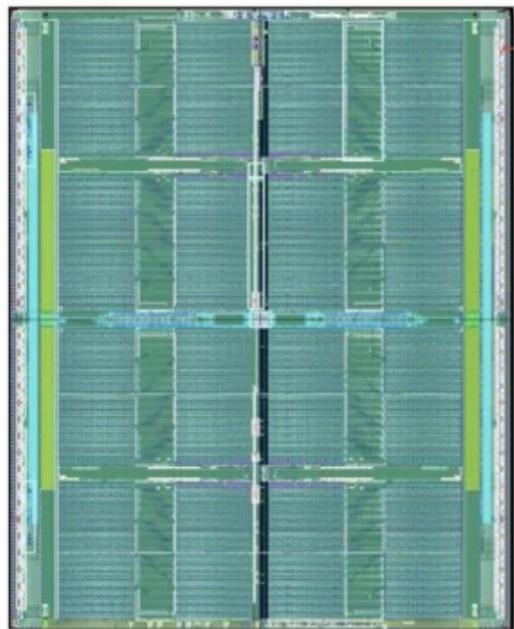
Elasticsearch 13 min read



ES native, elastiknn, Open Distro and GSI



Gemini® APU Processor



- Internal Clock
 - 200 – 500 MHz
- Compute In Memory
 - 48 million 10T SRAM cells
 - 2 million programmable “bit-processor” cores
- L1 Cache
 - 96Mb
- Algorithms
 - Similarity Search
 - Vector Processing
 - SAR BP, Image Processing, SHA-1/Password Cracking

Muves in the context of Vector Search Ecosystem



user interface

Application business logic: neural / BM25,
symbolic filters, ranking

Encoders: Transformers, Clip, GPT3...

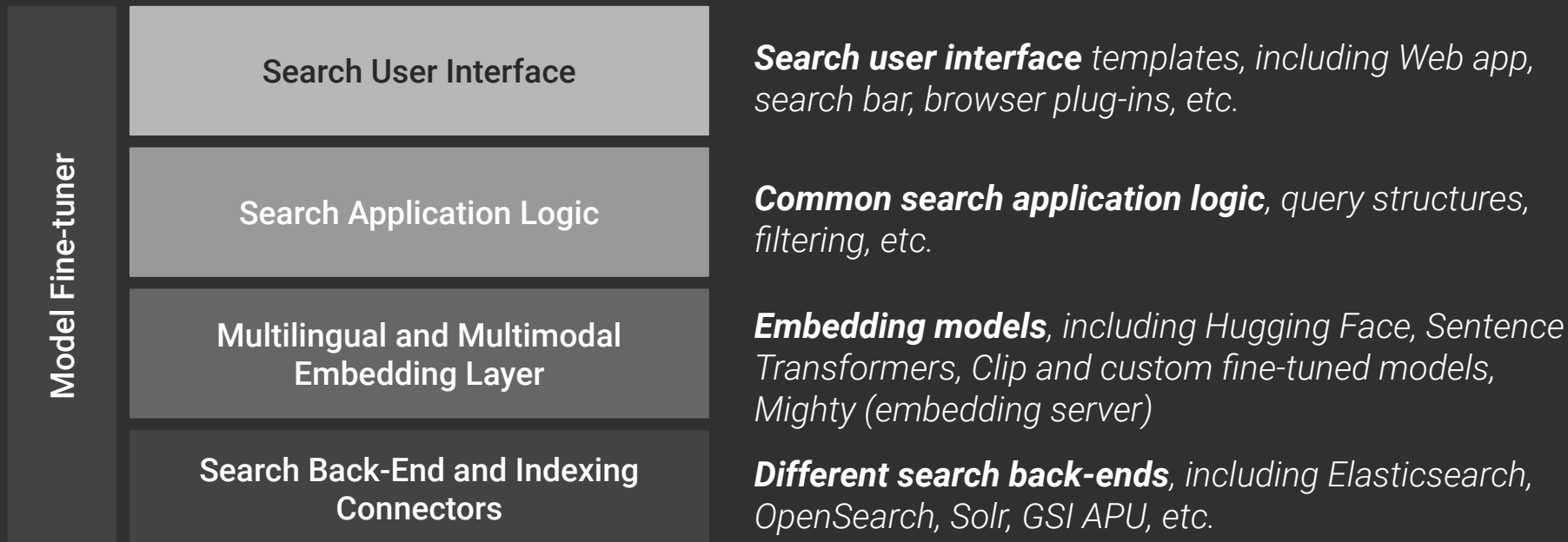
Neural frameworks: Haystack, Jina.AI, ZIR.AI, Hebbia.AI, Featureform...

Vector Databases: Milvus, Weaviate, Pinecone, GSI, Qdrant, Vespa, Vald, Elastiknn...

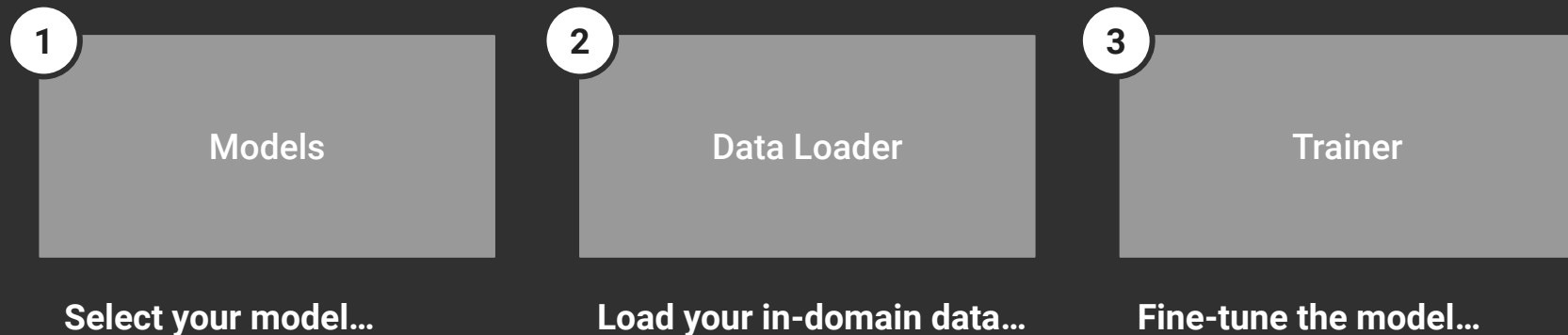
KNN / ANN algorithms: HNSW, PQ, IVF, LSH, Zoom, DiskANN, BuddyPQ ...

Muves

Muves is a search application focused on multimodal and multilingual semantic search



Muver - model fine-tuner for multilingual and multimodal vector search



Connecting OpenSearch to GSI APU requires changes to indexing and queries

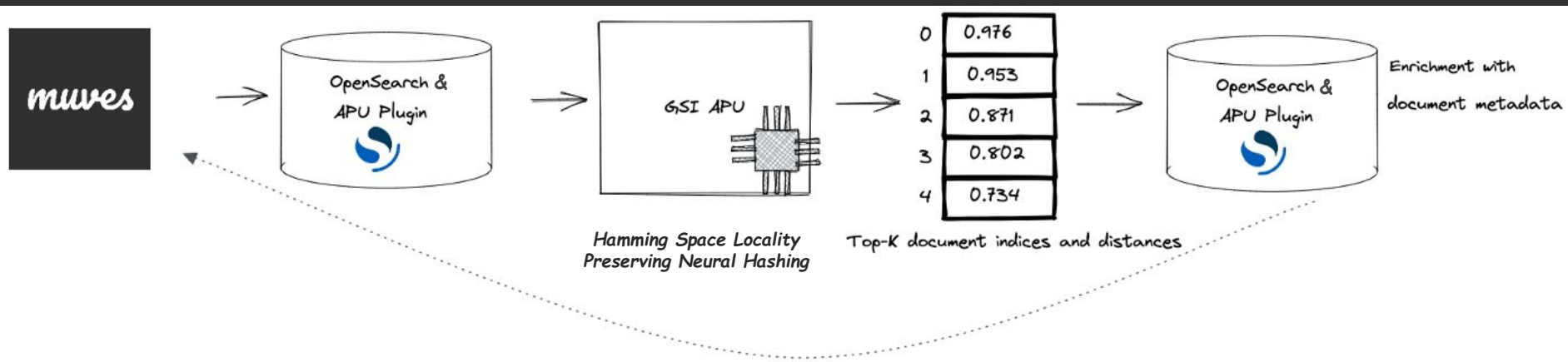
OpenSearch mappings file needs to define the vector field to be indexed by APU

```
31     "fields" : {
32         "keyword" : {
33             "type" : "keyword",
34             "ignore_above" : 256
35         }
36     },
37     "imUrl" : {
38         "type" : "text",
39         "fields" : {
40             "keyword" : {
41                 "type" : "keyword",
42                 "ignore_above" : 1024
43             }
44         }
45     },
46     "vector" : {
47         "type" : "knn_vector",
48         "dimension" : 512
49     }
50 }
```

Vectorized search query is submitted to APU as a `gsi_knn` type

```
1  {
2    "query":{
3      "gsi_knn":{
4        "field":"vector",
5        "vector":[
6          0.0015746655408293009,
7          0.025234133005142212,
8          0.0031481462065130472,
9          ...
10       ],
11       "topk":5,
12       "prefilter":{
13         "nsfw.keyword": "unlikely"
14       }
15     }
16   },
17   "size":30
18 }
```

Query workflow



For the demo we used a multilingual CLIP from Huggingface and a 10M subset from LAION-400M dataset



Multilingual CLIP model for image search

Multilingual Sentence Transformers for text embeddings

LAION-400M

The world's largest
openly available
image-text-pair dataset with
400 million samples.

Muves in action

GSI APU Search Demo

Choose File no file selected

Type a query or upload a file above (one query per line)

Search

Index:

Results: 5

Safe search

Image embeddings (multili



Example queries:

- [man walking on the beach with a dog](#)
- [nehir kenarında balık yiyen ayı](#) (a bear eating fish by a river - Turkish)
- [蓝色的桌子和椅子](#) (blue table and chairs - Chinese)
- [синие ботинки](#) (blue shoes - Russian)
- [שמלה אדומה](#) (red dress - Hebrew)

Indexed data in the demo are 10M images and captions from the [LAION-400M dataset](#).

Image (vector) search vs keyword search


Image Search

GSI APU Search Demo


No file chosen

Index: Image embeddings (multilin **Results: 5** **Safe search** ☐

Top matches:



[USA, Alaska, Brown bear essen Lachs Chilkoot See](#)
Safe search: Safe



[brown-bear-eating-salmon-photo-17328-223291](#)
Safe search: Safe


Keyword Search

GSI APU Search Demo


No file chosen

Index: Keyword search **Results: 5** **Safe search** ☐

Top matches:



[heron eating a fish](#)
Safe search: Safe



[Heron eating a fish](#)
Safe search: Safe

Image (vector) search vs keyword search

Image Search

GSI APU Search Demo

Choose file No file chosen

red dress

Index:

Image embeddings (multiling

Results: 5

Safe search



Top matches:



[Best 25 Red Christmas Dress Ideas On Pinterest](#)

Safe search: Safe



[Ericdress A-Line Sweetheart Asymmetry Prom Dress With Applique](#)

Safe search: Safe

Keyword Search

GSI APU Search Demo

Choose file No file chosen

red dress

Index:

Keyword search

Results: 5

Safe search



Top matches:



[dress strapless dress red dress little red dress](#)

Safe search: Safe



[dress red prom dresses prom dress red dress red](#)

Safe search: Safe

Support for over 50 languages - Siniset kengät (blue shoes in Finnish)

GSI APU Search Demo

Choose file No file chosen

siniset kengät

Search

Index:

Image embeddings (multiling

Results: 5

Safe search



Search time:

104 ms

Top matches:



Crocs Kids' Crocband Clog Bright Cobalt/Charcoal 22-23

Safe search: Safe



Toms - Youth Knit Apalgrata Slip-On Shoes

Safe search: Safe

Supports over 50 languages - rød kjole (red dress in Norwegian)

GSI APU Search Demo

Choose file No file chosen

rød kjole

Search

Index:

Image embeddings (multiling

Results: 5

Safe search

Search time:

71 ms

Top matches:



[Simple Strapless A Line Bowknot A Line Knee Length Homecoming Dress](#)

Safe search: Safe



[what to wear to a holiday party the little red dress](#)

Safe search: Safe

Muves supports batch queries

GSI APU Search Demo

Choose file No file chosen

Type a query or upload a file above (one query per line)

Search

Index:

Image embeddings (multilin

Results: 5

Safe search

Search time:

237 ms (20 queries)

Top matches:

Query: red dress



[Z Spoke by Zac Posen Str Taffeta V Dress - Lyst](#)

Safe search: Safe



[Simple Strapless A Line Bowknot A Line Knee Length Homecoming Dress](#)

Safe search: Safe

Image embedding search: NDCG@10

Quepid Relevancy Cases Teams Scorers Video Tutorials Knowledge Base Wiki Dmitry Kan

Current case **Muves Search: Image** — Try 13 — nDCG@10

0.79 nDCG@10

Select scorer Create snapshot Compare snapshots Import Share case Clone Delete Export Tune Relevance

Add a query to this case Add query






Show only rated Collapse all Sort Manual Name Modified Score Errors

Filter Queries Number of Queries: 5

1.00 red dress 900 Results

Score All 3 -

Toggle Notes Explain Query Missing Documents Set Options Set Threshold Move Query Delete Query

3 -		Best 25 Red Christmas Dress Ideas On Pinterest Rank: #1	Matches 1 no explain for doc
3 -		Ericdress A-Line Sweetheart Asymmetry Prom Dress With Appliques And Sequins Rank: #2	Matches 1 no explain for doc
3 -		Z Spoke by Zac Posen Str Taffeta V Dress - Lyst Rank: #3	Matches 1 no explain for doc
3 -		Simple Strapless A Line Bowknot A Line Knee Length Homecoming Dress Rank: #4	Matches 1 no explain for doc
3 -		Full Length Off the Shoulder Red Lace and Crepe Dress With Slit Rank: #5	Matches 1 no explain for doc

WhatsApp icon








Image embedding search: NDCG@10

Quepid Relevancy Cases Teams Scorers Video Tutorials Knowledge Base Wiki Dmitry Kan

0.97 bear eating fish in the river 900 Results

Score All 3 -

Toggle Notes Explain Query Missing Documents Set Options Set Threshold Move Query Delete Query

3 -		brown-bear-eating-salmon-photo-17328-223291 Rank: #1	Matches 1 no explain for doc
3 -		Brown Bear Eating Salmon Metal Print by Dan Friend Rank: #2	Matches 1 no explain for doc
3 -		A brown bear carries away a salmon it caught at the McNeil River Falls, in Alaska's McNeil River State Game Sanctuary. Rank: #3	Matches 1 no explain for doc
0 -		A. Roland Knight (British, active 1879-1921) Otter and salmon on a river bank Rank: #4	Matches 1 no explain for doc
3 -		Pair of Grizzly Bears (Ursus Arctos Horribilis) Fight as They Catch Rank: #5	Matches 1 no explain for doc
3 -		Brown bear — Stock Photo Rank: #6	Matches 1 no explain for doc
3 -		Bear or the Salmon? Rank: #7	Matches 1 no explain for doc

1

Image embedding vs text embedding vs keyword

Current case
Muves Search: keyword — Try 1 — nDCG@10

0.84
EXCEEDED

Select scorer Create snapshot Compare snapshots Import Share case Clone Delete Export Tune Relevance

Add a query to this case Add query

0.96	red dress
0.94	swimming pool
0.93	green dress
0.54	bear eating fish in the river
0.83	blue shoes

Current case
Muves Search: Image — Try 13 — nDCG@10

0.98
EXCEEDED

Select scorer Create snapshot Compare snapshots Import Share case Clone Delete Export Tune Relevance

Add a query to this case Add query

1.00	red dress	Info Needed
0.96	swimming pool	
1.00	green dress	
0.97	bear eating fish in the river	
0.96	blue shoes	

Current case
Muves Search: Text — Try 1 — nDCG@10

0.75
EXCEEDED

Select scorer Create snapshot Compare snapshots Import Share case Clone Delete Export Tune Relevance

Add a query to this case Add query ☐ Show only rated | Collapse all

0.98	red dress
1.00	swimming pool
0.98	green dress
0.00	bear eating fish in the river
0.78	blue shoes

Lessons learned

1. **Building multilingual and multimodal search is easy** — but fine-tuning is required for domain adaptation.
2. **With Muves, implementation of vector search application is much faster** than figuring out all the details from scratch.
3. **It is not the case that vector search is not high-performing and does not scale easily.**
4. **In a production setting you most likely will need filtering support.** APU allows neural search with symbolic filtering at scale.

References



Dmitry Kan · May 2

Multilingual and Multimodal Vector Search with Hardware Acceleration

Authors: Dmitry Kan, Aarne Talman We recently partnered with GSI Technology Inc. to develop a vector search demo that utilizes their APU...



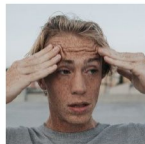
Multimodal 11 min read



Published in **GSI Technology** · 3 days ago

Natural Language Processing ... for Image Search?

Natural language processing (NLP) is a major part of search — so much so that it is even being used in image search applications. For example, Google said, when talking about its MUM model, “Eventually, you might be able ...



NLP 5 min read



Published in **Towards Data Science** · Mar 15, 2021 ★

Speeding up BERT Search in Elasticsearch

Neural Search in Elasticsearch: from vanilla to KNN to hardware acceleration — In two previous blog posts on my journey with BERT: Neural Search with BERT and Solr and Fun with Apache Lucene and BER...

Elasticsearch 13 min read



Hamming Space Locality Preserving Neural Hashing for Similarity Search

Daphna Idelson
GSI Technology
didelson@gsitechnology.com

- <https://venturebeat-com.cdn.ampproject.org/c/s/venturebeat.com/2022/03/21/andrew-ng-predicts-the-next-10-years-in-ai/amp/>
- <https://blog.muves.io/multilingual-and-multimodal-vector-search-with-hardware-acceleration-2091a825de78>
- <https://medium.com/gsi-technology/natural-language-processing-for-image-search-41fadd74221b>
- <https://www.gsitechnology.com/sites/default/files/Whitepapers/GSIT-Hamming-Space-Locality-Preserving-Neural-Hashing-for-Similarity-Search.pdf>

Powered by *muves*

<http://os-demo.gsitechnology.com:8080/>
<https://www.searchium.ai/>

muves.io