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:

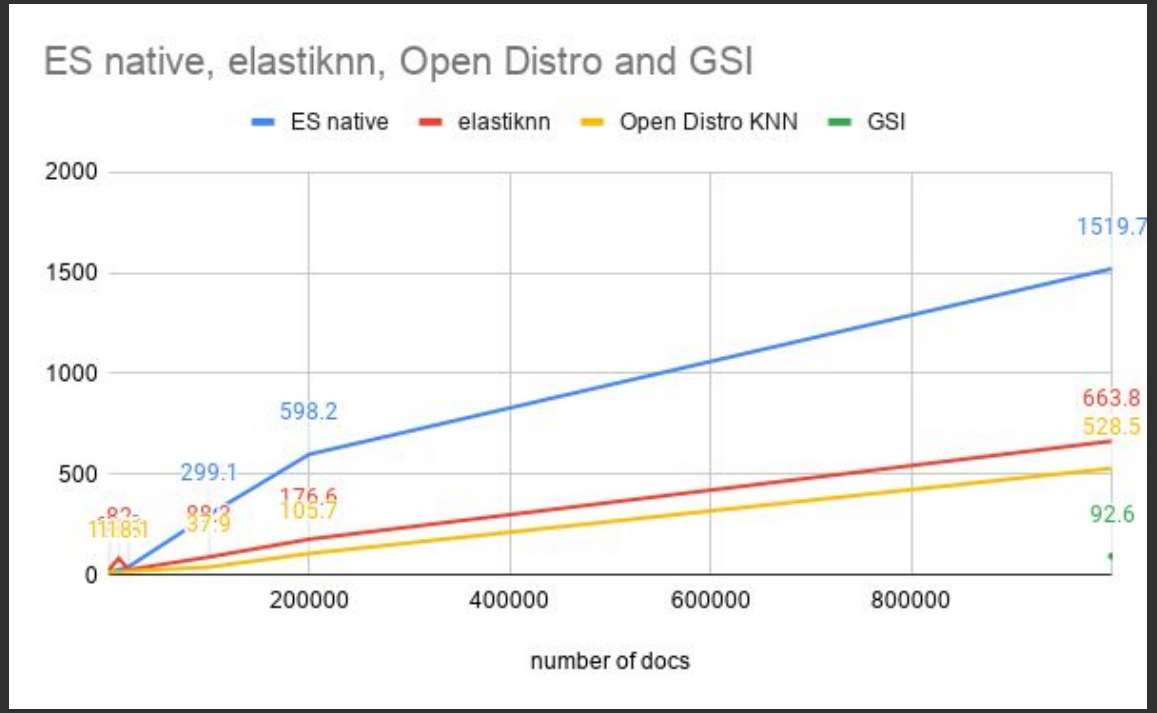
Hardware acceleration using GSI APU can significantly improve query speed

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

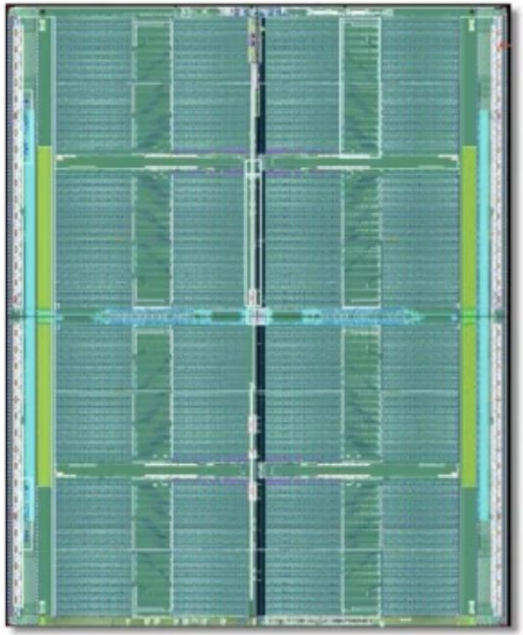
Speeding up BERT Search in Elasticsearch

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

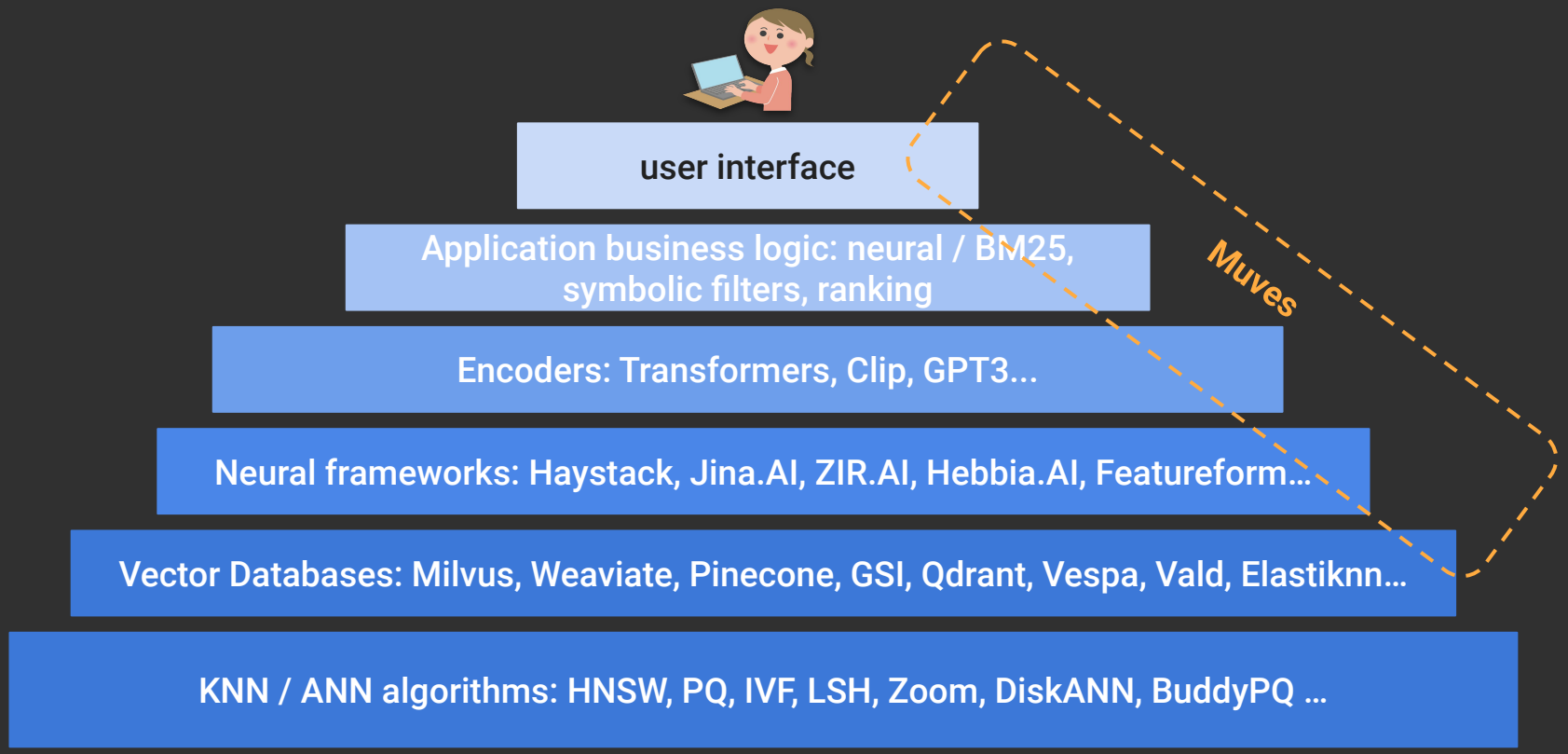


Gemini® APU Processor

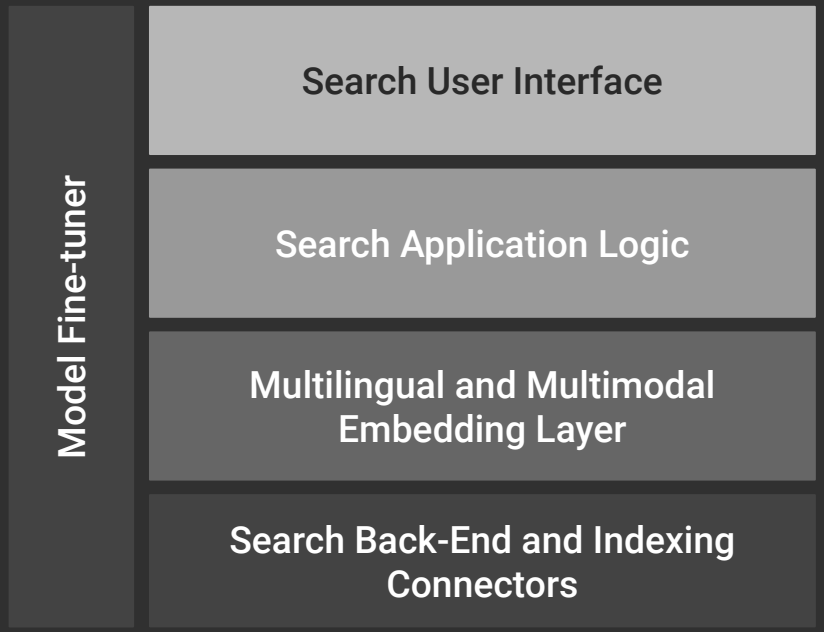


- Internal Clock
  - 200 – 500 MHz
- Compute In Memory
  - 48 million 10T SRAM cells
  - 2 million programmable "bit-processor" cores
- L1 Cache
  - 96Mb
- Algorithms
  - Similarity Search
  - Vector Processing
  - SAR BP, Image Processing, SHA-1/Password Cracking

Muves in the context of Vector Search Ecosystem



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



Search user interface *templates, including Web app, search bar, browser plug-ins, etc.*

Common search application logic*, query structures, filtering, etc.*

Embedding models*, including Hugging Face, Sentence Transformers, Clip and custom fine-tuned models, Mighty (embedding server)*

Different search back-ends*, including Elasticsearch, OpenSearch, Solr, GSI APU, etc.*

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

1

Models

Select your model…

2

Data Loader

Load your in-domain data…

3

Trainer

Fine-tune the model…

Connecting OpenSearch to GSI APU requires changes to indexing and queries

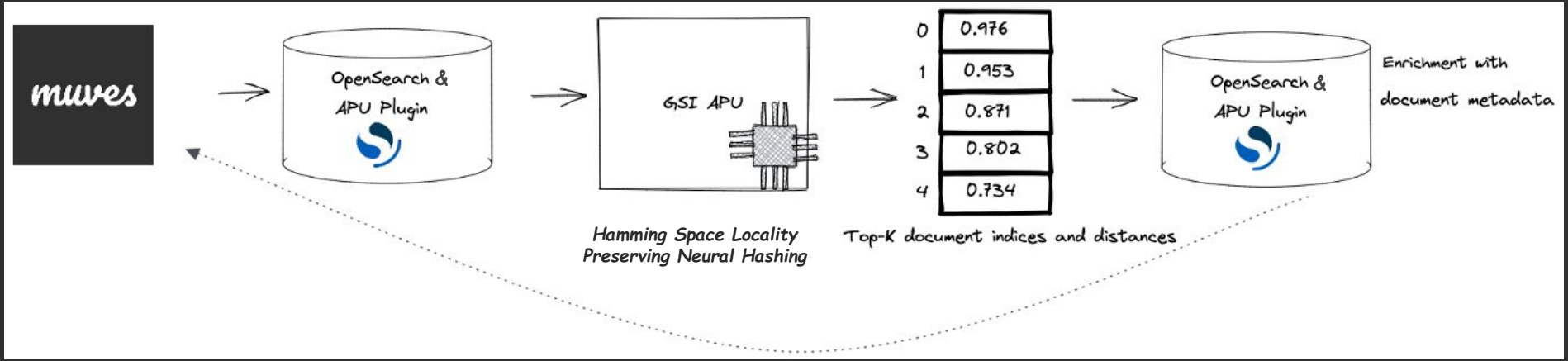
OpenSearch mappings file needs to define the vector field to be indexed by APU

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

Hugging Face

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: Image embeddings (multili

Results: 5

Safe search

Example queries:

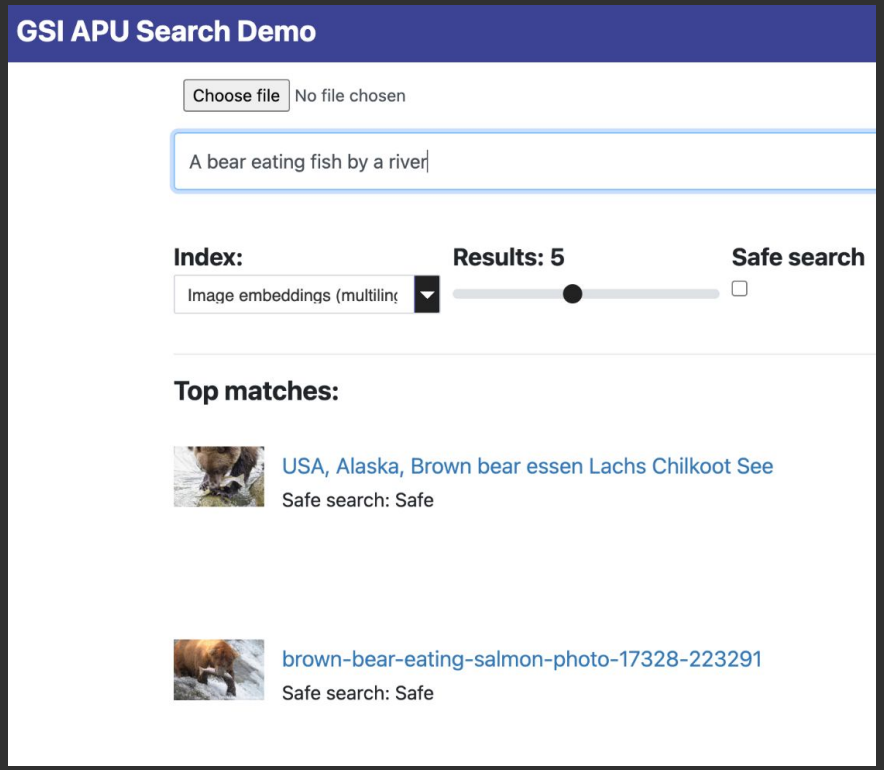
- *man walking on the beach with a dog*
- *nehir kenarında balık yiyen ayı* (a bear eating fish by a river - Turkish)
- *蓝色的桌子和椅子* (blue table and chairs - Chinese)
- *синие ботинки* (blue shoes - Russian)
- *שמלה אדומה* (red dress - Hebrew)

Indexed data in the demo are 10M images and captions from the LAION-400M dataset.

Powered by muves

Image (vector) search vs keyword search

Image Search



Keyword Search

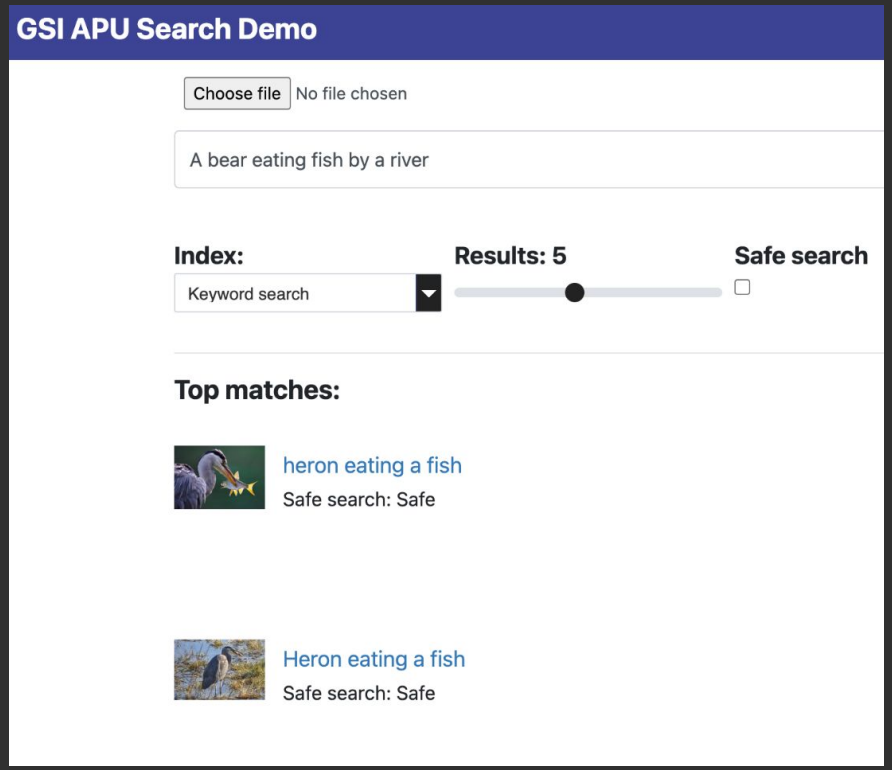
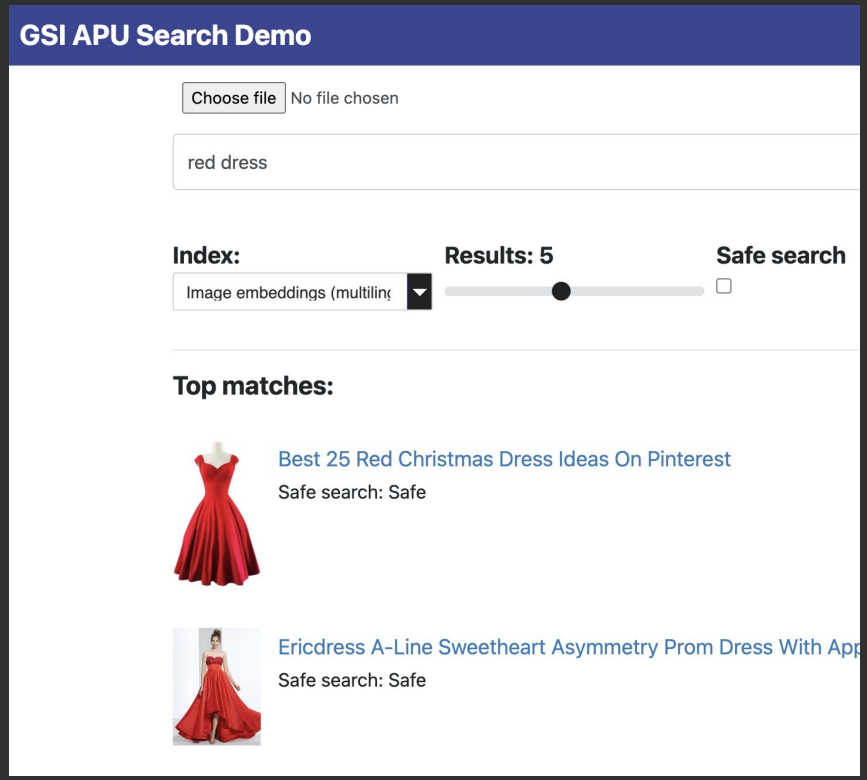
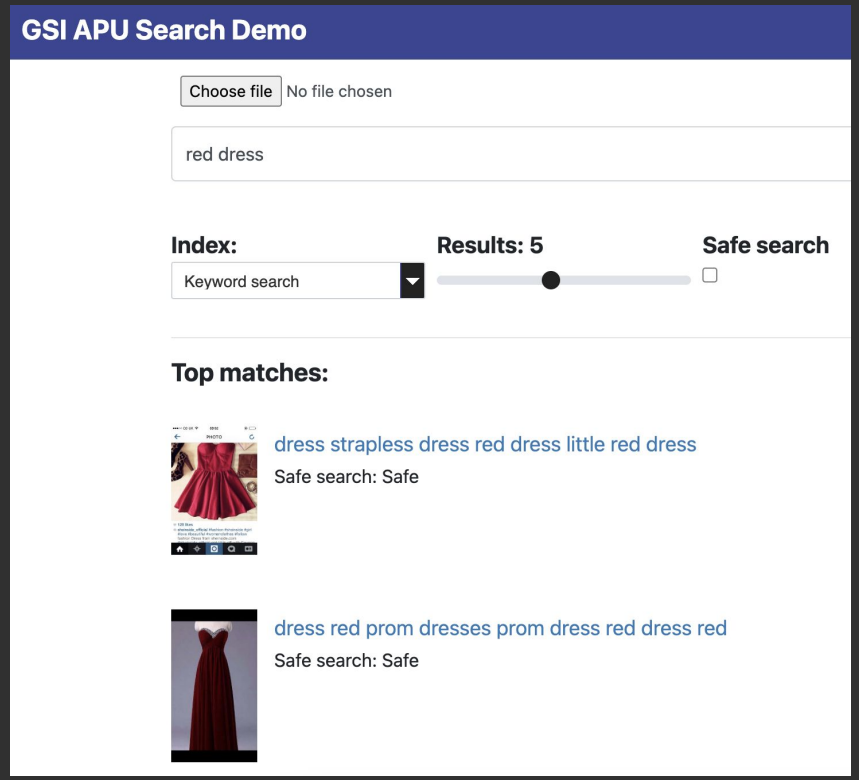


Image (vector) search vs keyword search

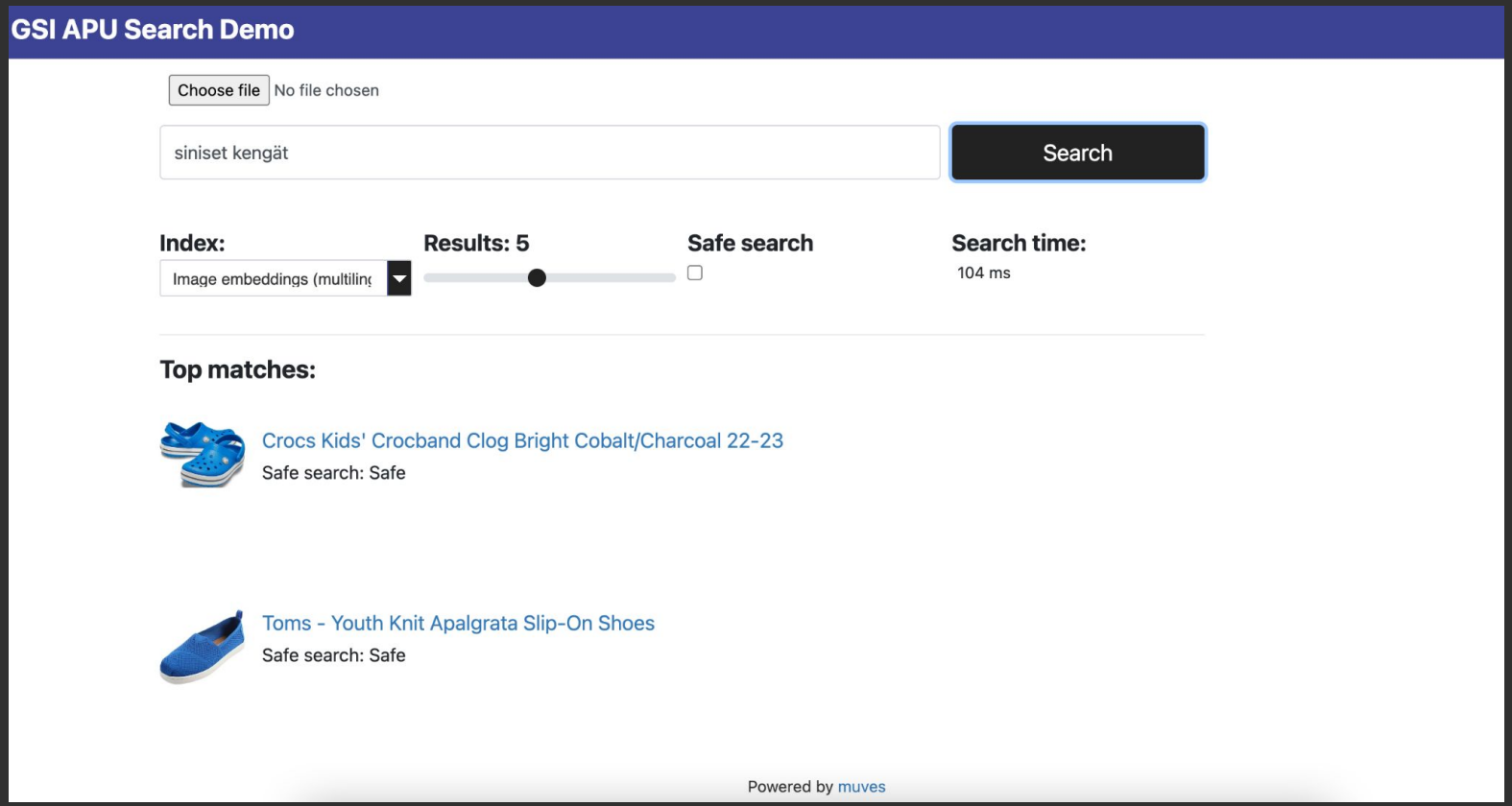
Image Search



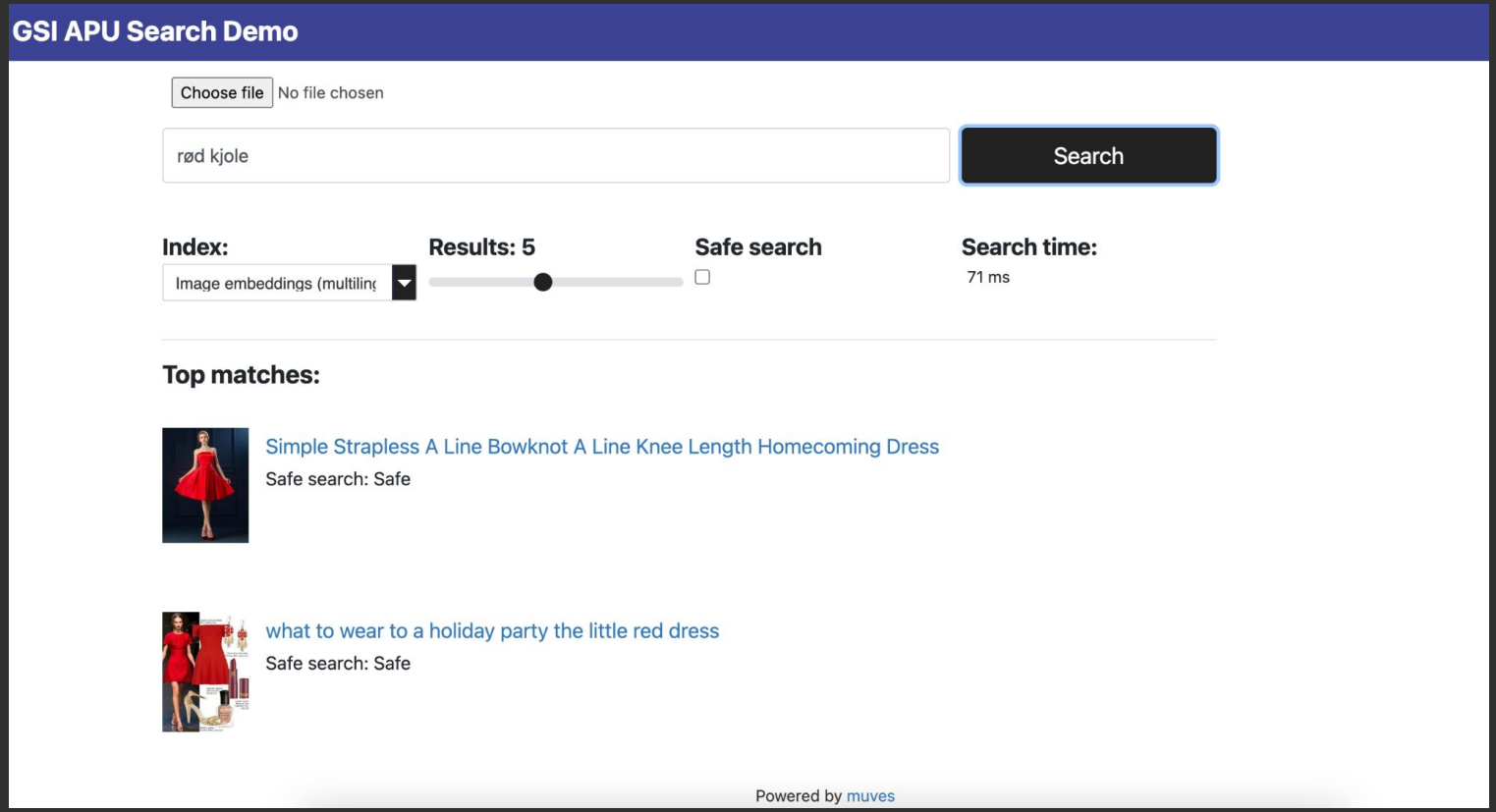
Keyword Search



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



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



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 (multiling

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

Powered by muves

Image embedding search: NDCG@10

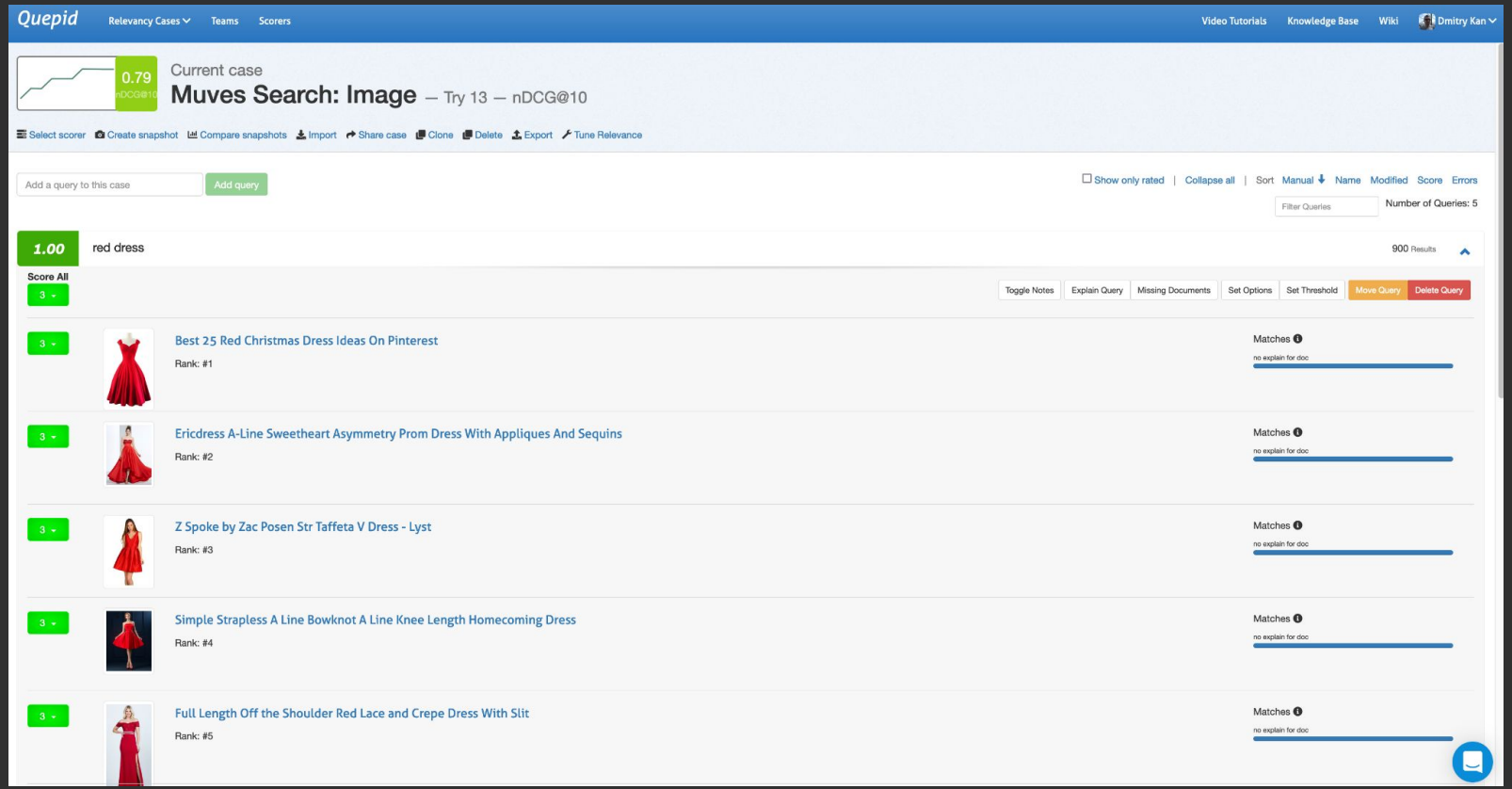


Image embedding search: NDCG@10

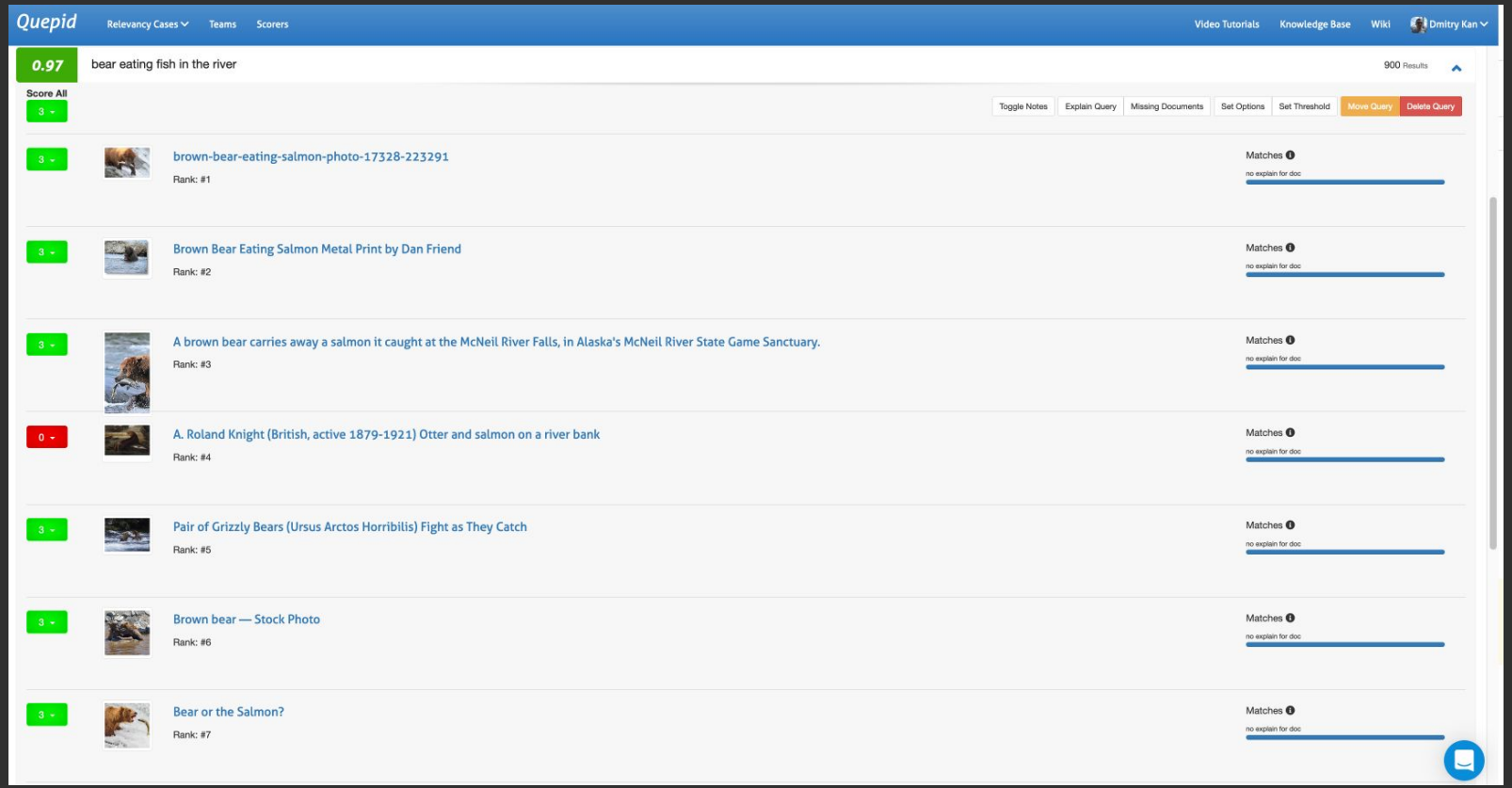
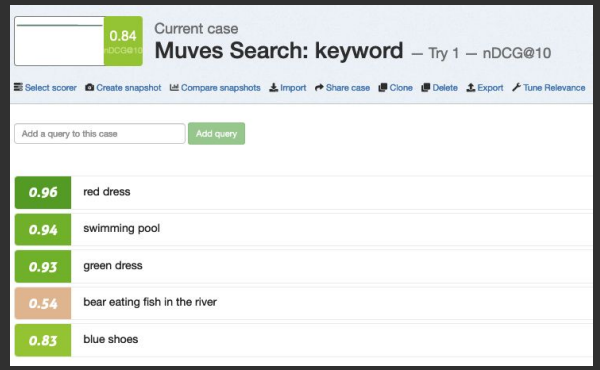
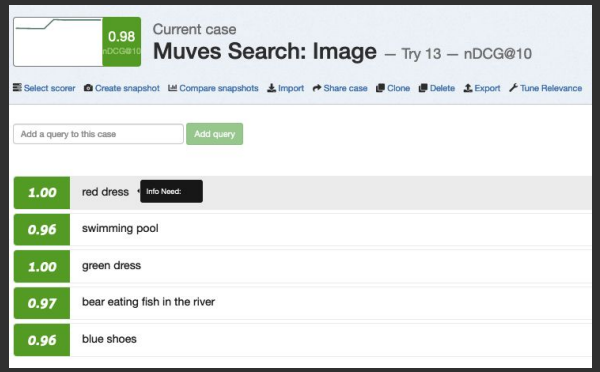
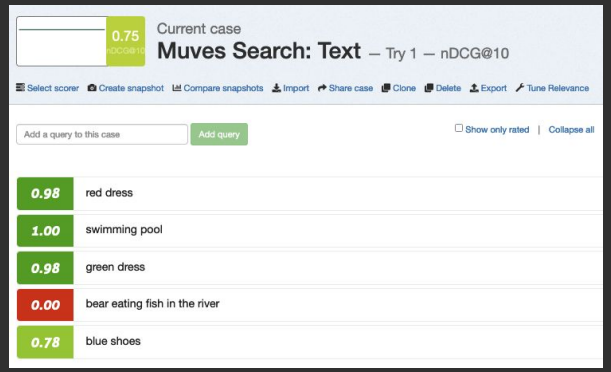


Image embedding vs text embedding vs keyword







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 **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

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

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