Multi-Search

Case Study

Table of contents

01
Problem Statement

02
Architecture Framework

03
Models

04
Data Requirements & Metrics

05
Continuous Improvements

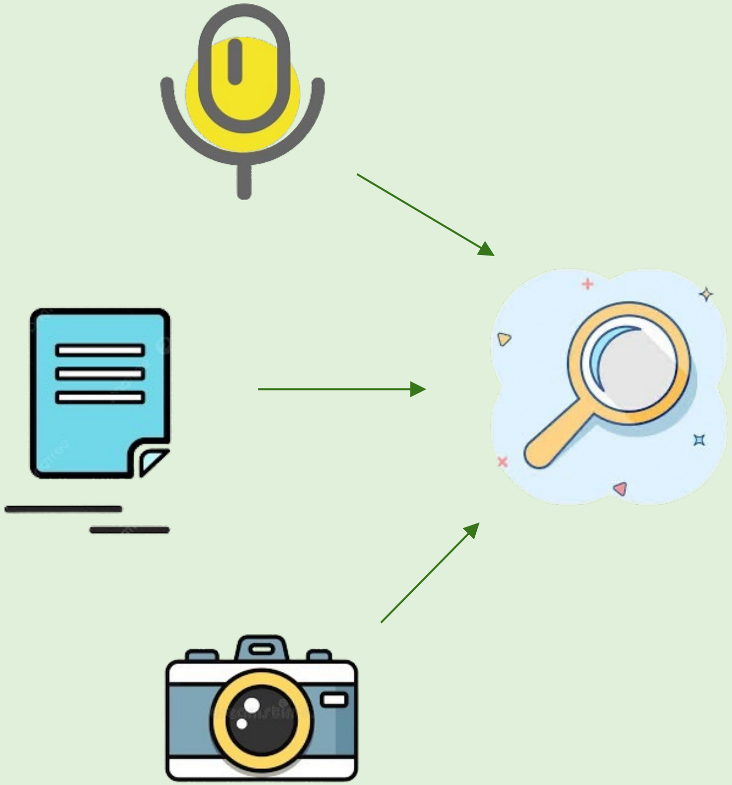
06
Trade-Off Evaluations

01

Problem Statement

Multi-Search

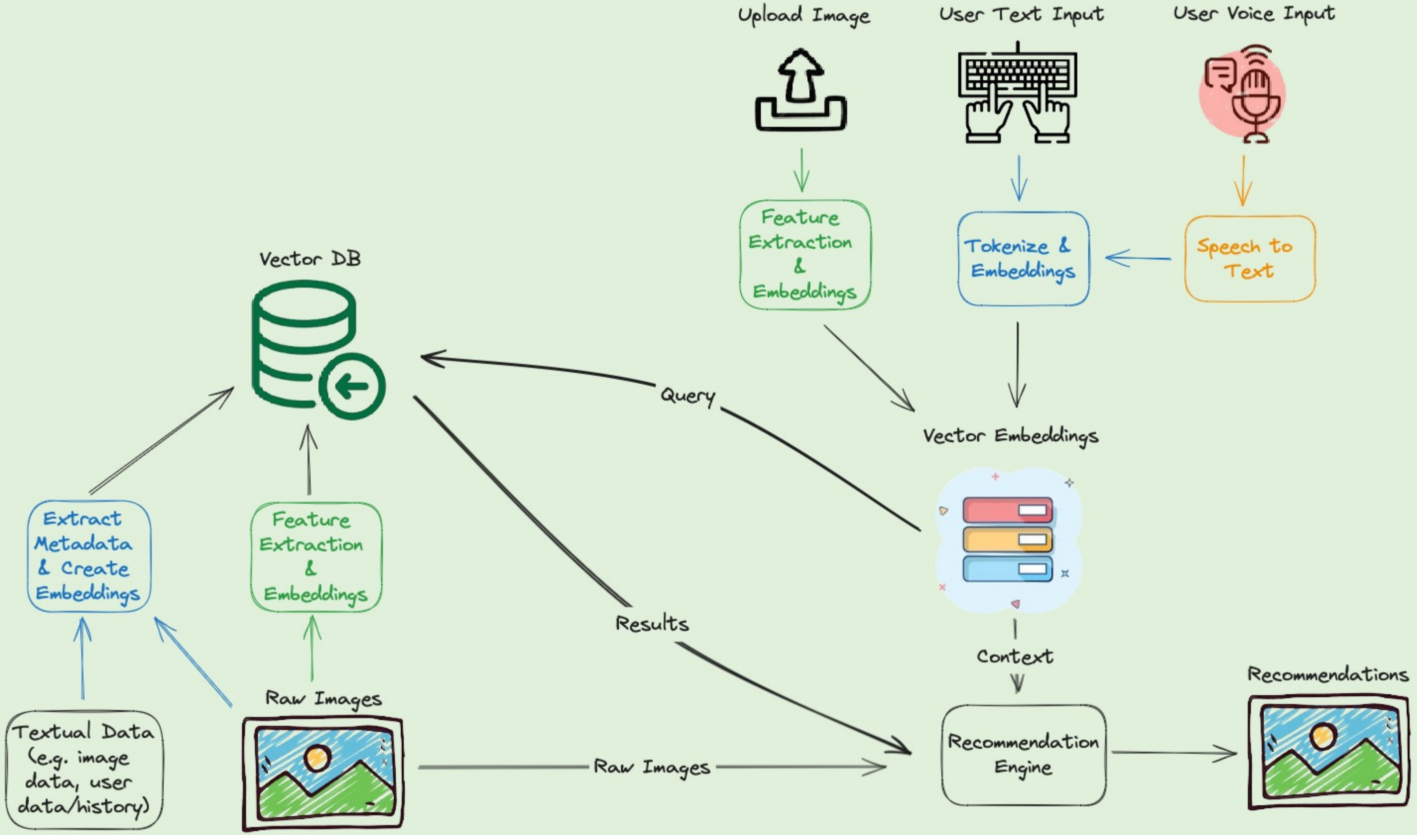
Design a product called Multi-Search, which provides enhanced search capabilities beyond traditional text searches for e-commerce websites.



02

Architecture Framework

Architecture Diagram



Main Steps

Step 1: Create Database

- Collect product data
- Extract features from image corpus and metadata
- Create embeddings and store in database

Step 2: Input Processing

- Input handling: Images, text, voice
- Extract features and create embeddings
- Normalize features for consistency (align embeddings from different modalities)

Step 3: Query

- Use a retrieval system to query the database based on user input
- Return results based on relevance

Step 4: Recommendation

- Analyze product context and query results to generate recommendations

03

Models

Models

Images

Extract features and generate textual metadata (e.g. details of the product) if required

Multimodal: GPT-4o
CNN Based: EfficientNet, ResNet

Text

Tokenize and generate embeddings

BERT, GPT-3

Voice

Convert voice to text or to text embeddings

Speech to Embeddings: Wav2Vec
Speech to Text: DeepSearch, Whisper

Engines

Query

Use the query embeddings to search for similar embeddings in the database

FAISS

Recommendation

Generate personalized recommendations for user

Collaborative filtering, content-based filtering, hybrid

04

Data Requirements & Metrics

Data Requirements (Algo)

Image Dataset

- Image data to be stored in vector database
- Ideally annotated (e.g., colour, patterns on cloths, sleeve length, collar type, buttons) to be used as product metadata
- **Ground Truth:** Dataset where each query is associated with a set of relevant results

Product Metadata

- Metadata such as title, description, material, care instruction, occasion for use, etc.
- Specific image metadata (e.g. color, sleeve length, buttons, etc.)

User Data

- User interaction data for recommendation (clicks, past purchases, user id, timestamps, etc.)

Data Requirements (Operational)

Logging of metrics from operational data:

- Latency
- Throughput
- Click-Through Rate
- Conversion Rate

Log File

Evaluation Metrics

Accuracy Metrics

- Evaluation output results

Performance Metrics

- Evaluate the efficiency of the product

User Experience Metrics

- Evaluate how satisfied users are with Multi-Search

Accuracy Metrics

Precision

$$\frac{TP}{TP + FP}$$

- Measures proportion of relevant results **among the retrieved results**
- High Precision means most of the retrieved results are relevant
- Irrelevant results leads to negative experience to users *(e.g., frustration)*

Recall

$$\frac{TP}{TP + FN}$$

- Measures proportion of relevant results **that were retrieved**
- High Recall means most of the the relevant items in the database were retrieved
- Relevant results leads to positive experience to users *(e.g., increased sales)*

F1-Score

$$2 \text{ x } \frac{Precision \text{ } x \text{ } Recall}{Precision+Recall}$$

- Trade off balanced measure when precision and recall are **equally important.**
- Retrieving more results has higher recall but lower precision
- Retrieving fewer results has higher precision but lower recall
- **Balance between presenting relevant products and ensuring users can see a wide range of options**

Performance Metrics

Latency

$$\frac{\sum_{i=1}^{N} Response\ Time_i}{N}$$

- Time taken for each query to be processed and a response generated
- Low latency ensures a positive user experience

Throughput

$$\frac{Total\ Queries}{Total\ Time}$$

- Total number of queries processed at a given period
- High throughput ensures system can handle high traffic

User Experience Metrics

Click-Through Rate

$$\frac{\textit{Number of Clicks}}{\textit{Number of Impressions}}$$

- Measures effectiveness of search results in **engaging users**
- High CTR indicates search results are **relevant** and **compelling enough to click**

Conversion Rate

$$\frac{\textit{Number of Conversions}}{\textit{Nmber of Impressions}}$$

- Measures effectiveness of search results in **driving purchases**
- High conversion rate indicates search results are **relevant** and effective in **driving sales**

User Satisfaction Survey

- Qualitative feedback
- Gauge user satisfaction *(scoring through a scale)*
- Identify areas of improvement

05

Continuous Improvements

Feedback Loop

User Feedback

- User surveys on search experience, satisfaction, issues, etc.
- Ratings and Reviews: rate relevance of results
- Track user interactions such as clicks, dwell time and conversion rate to infer satisfaction and relevance

Active Learning

- Update models and data based on user feedback
- Focus on results that show low confidence
- Add more data for queries that frequently fail to retrieve relevant results

A/B Testing

- Deploy different versions of Multi-Search or No Multi-Search vs Multi-Search to compare performance

Monitoring

Real-Time Monitoring

- Dashboards to monitor metrics such as latency, throughput, click-through rate, conversion rate.
- Alerts for deviations in performance metrics *(e.g. drop in click-through rate)*

Performance Analysis

- Analyze trends in metrics
- Analyze user feedback

Iterative Updates

Regular Updates

- Continuously update models and dataset to keep current with latest trends and user behaviours
- Model versioning and performance evaluations

Addressing Feedback

- Use user feedback to make targeted improvements
- Bug fixes from reports issues

Feature Enhancement

- Regularly introduce new features or improvements based on user needs or technological advancements

06

Trade-Off Evaluations

Accuracy vs Cost/Latency

Aim: Balance highly accurate results with the need of quick response

Possible Solutions:

- Caching frequent queries
- Edge computing to reduce latency
- Model distillation: train a smaller model using a larger model
- Dimensionality reduction in vector store
- Use model benchmarking and performance metrics to aid in assessment:
    - Test different models on their accuracy vs cost/latency and assess the trade-off between improved accuracy and increased cost/latency

Thanks!