OBilet Case Documentation

İlhan Sertelli

14 May 2025

1 Introduction

In this case, I implemented a search technique that filters and lists hotels depending on the visual preferences of the users. The goal was to allow users to submit free-text queries, such as “double room with sea view” or “room with air conditioning and city view”, and receive matching hotel images in response by using some AI solutions. The system combines various techniques such as image captioning, keyword-based search, and vector-based semantic search to provide accurate results. This document summarizes the solution architecture, tools used, search methods applied, limitations encountered, and final observations.

2 Tools

Several tools and technologies were utilized to build the solution. The most critical ones include the following:

- **OpenAI GPT-4o API**: This API was used to perform image captioning by analyzing the content of hotel room images. The model provides detailed descriptions including room capacity, view, and amenities.
- **Sentence-Transformers (all-mpnet-base-v2)**: This pre-trained embedding model was used to encode both image captions and user queries into vector representations for semantic similarity calculations.
- **Scikit-learn Cosine Similarity**: Cosine similarity from Scikit-learn was used to compare query and caption embeddings to perform semantic search.

Initially, the BLIP (Bootstrapped Language-Image Pretraining) model was considered for the captioning task due to its open-source feature. However, after several tests, it was observed that BLIP’s descriptive accuracy did not meet the desired level for this specific use case. Hence, the system was adapted to use OpenAI’s GPT-4o Vision API, which provided significantly more reliable and descriptive captions.

3 Keyword and Vector Based Semantic Search

The implemented solution combines both keyword-based and vector-based semantic search methods to maximize retrieval accuracy.

Keyword-Based Search

Keyword-based search operates by checking if the user's query appears as a direct substring within the generated image captions. This simple matching strategy allows the system to quickly find exact textual matches. In other words, it checkes a string if a part of the query directly takes part in the image caption done by OpenAI API.

Vector-Based Semantic Search

To handle queries that do not exactly match the wording of the captions, vector-based semantic search was implemented using the *all-mpnet-base-v2* sentence transformer model. Both user queries and image captions are converted into high-dimensional vector embeddings. Cosine similarity is then calculated between these vectors to determine semantic closeness. If the similarity score exceeds a predefined threshold, the image is considered a valid match.

Hybrid Scoring

The system combines the results of both methods by taking the maximum of the keyword and semantic similarity scores. This hybrid approach improves the robustness of the search by capturing both exact and contextual matches.

4 Agent-to-Agent Communication

The system architecture is designed using an agent-based communication approach in order to contruct modularity in the system. Each agent is responsible for a specific task, and they communicate through well-defined data exchanges.

- **Captioning Agent:** This agent is responsible for generating descriptive text captions for each hotel image using the OpenAI GPT-4o Vision API. These captions include information such as room capacity, view type, and amenities.
- **Indexing Agent:** After captions are generated, the Indexing Agent collects and stores them in an index structure, associating each caption with its corresponding image URL.
- **Query Matching Engine:** This agent receives user queries and matches them with the indexed captions. It applies both keyword-based and vector-based semantic search methods, leveraging sentence embeddings and cosine similarity.

The agents operate in sequence, passing data to each other in a pipeline:

Captioning Agent → Indexing Agent → Query Matching Engine

This agent-to-agent communication structure ensures that each component is isolated yet cooperative, making the system easier to extend or adapt in future work. For example, the captioning agent could be replaced with another model without affecting the search logic.

5 Limitations

While the system provides effective results in most scenarios, several limitations were encountered throughout the development process:

- **API Access and Cost:** Initially, the OpenAI GPT-4o API was considered to generate highly descriptive captions. However, when it was tried to do so, it is realized that there was no free credit. To continue the development without incurring costs, the BLIP model was evaluated as an alternative.
- **Captioning Accuracy:** Although BLIP is open-source and cost-effective, it did not produce sufficiently detailed or accurate descriptions for this specific hotel room classification task. Its captions were often too generic or inconsistent with the visual content, leading to poor matching performance even the largest parameter version of it was used.
- **Reconsidering GPT-4o API:** Given BLIP's underperformance, the solution was redirected to switch back to the GPT-4o Vision API despite its cost. This decision was made to ensure higher caption quality and more reliable search results.
- **Model Limitations:** Even with GPT-4o, the captioning process is not guaranteed to be perfect in all cases. Visual ambiguities, low-quality images, or occlusions can lead to missing or inaccurate descriptions.
- **OpenAI API Url Access Problem:** Having switched back to OpenAI API, it is observed that the API was unable to access the image links provided in the case study document. Thus, in order to solve that, the images are downloaded to local one by one and pushed to the Github repository so that the API can access it.
- **Threshold Sensitivity:** The semantic similarity threshold directly impacts the number of results returned. Setting it too low may return irrelevant results, while setting it too high may miss relevant ones. Fine-tuning this value remains a balancing challenge.

6 Result

The final implementation successfully integrates image captioning with keyword-based and vector-based semantic search to retrieve hotel images based on user preferences.

The OpenAI GPT-4o Vision API demonstrated strong performance in generating accurate and descriptive captions for the provided hotel images. The captions typically included essential details such as room capacity, view type, and available amenities, making them highly suitable for the matching process.

However, some classification inaccuracies may still occur due to the limitations of the sentence transformer model used for semantic search. While the system generally performs well, vector-based similarity scoring is not flawless and can occasionally produce false positives or overlook certain matches. Despite these minor limitations, the system is capable of providing mostly accurate search results that align with user queries.

The complete implementation, including the agents, search logic, and a sample interactive interface, is publicly available on GitHub for further review and testing:

- **GitHub Repository:** `https://github.com/ilhansertelli/ObiletCase`