altice labs

Artificial Intelligence in efficient image search on MEO Cloud

Artificial Intelligence; Search; Image; Cloud platform; MEO Cloud

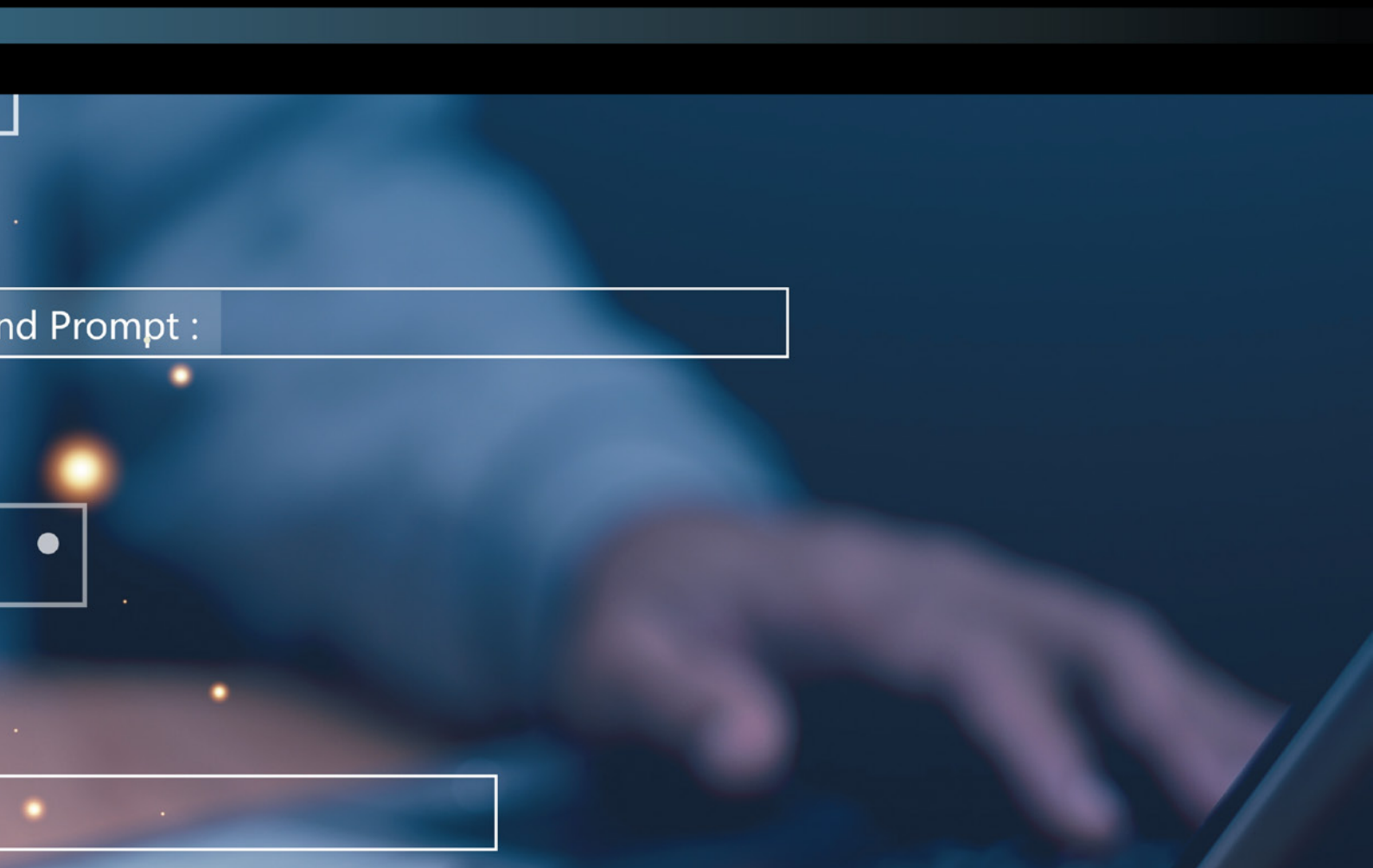
White paper

Version 1.0, July 2024

Abstract

With the technological advancement of devices, there has been an increase in the volume of data that each user accumulates. To guarantee the integrity of this data, cloud platforms make it possible to store and synchronize user data between different devices. Among Altice's offerings, MEO Cloud is particularly noteworthy in this sense. On this platform, users have several images, creating the challenge of efficiently searching for images that meet their requirements. The emergence of new approaches to image search, combined with deep learning models, makes it possible to mitigate this challenge.

The Contrastive Language-Image Pre-training (CLIP) model establishes semantic relationships by unifying text and image in the same form. Textual captions generated by the Bootstrapping Language-Image Pre-training (BLIP) generative model are used to filter out irrelevant images. The MiniLM-L6-v2 model makes it possible to compare the semantics of the captions with the user's search description. The aim of this article is to present the system developed to integrate efficient and dynamic image search into the MEO Cloud platform using artificial intelligence.



Introduction

Technological advances have increased the number of images users possess and manage in recent decades. This phenomenon is due to the spread of devices with high photographic quality, the increase in data storage capacity and the development of communication infrastructures for distribution [1]. Images capture essential moments, preserve memories, and serve as effective means of communication to convey messages. It is estimated that 5.3 billion images are captured daily; as **Figure 1** indicates, more than 1.8 trillion images were captured globally in 2023. Since 2012, the number of images captured has grown linearly, at an average rate of 10%-14%, with a forecast of 1.94 trillion images in 2024. However, there was an exception during the years of the pandemic, with a decline of approximately 20% [2].

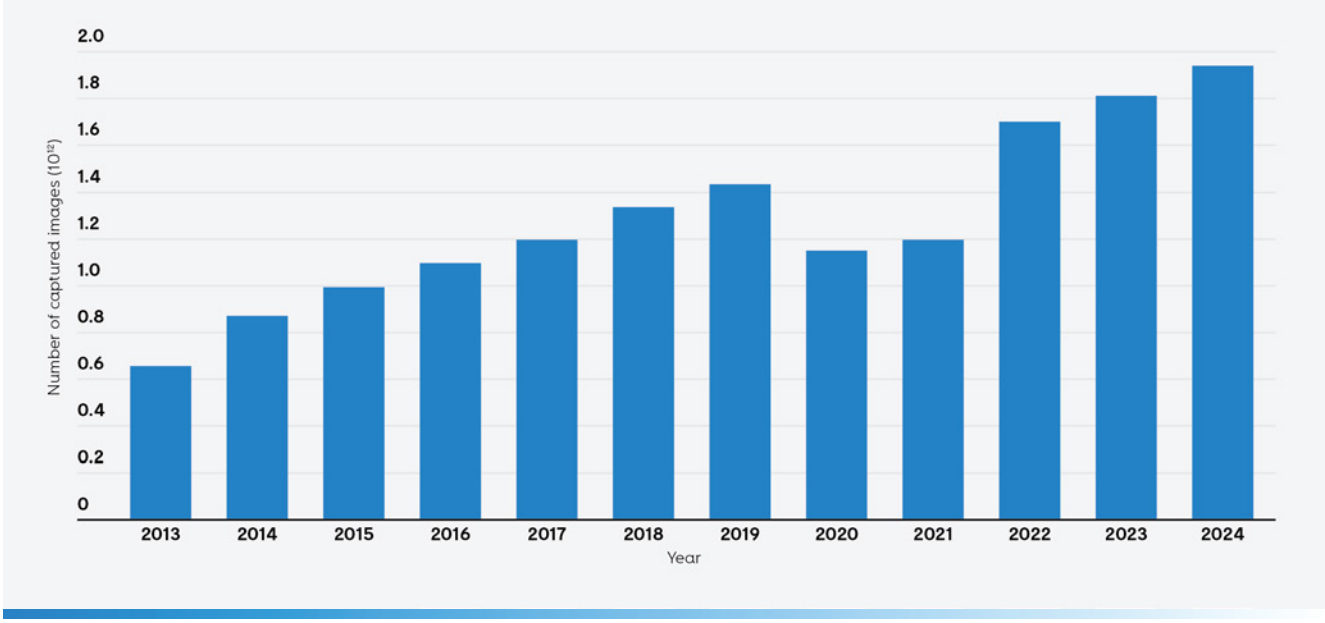


Figure 1 – Number of images captured each year [2]

To ensure the permanence and accessibility of images, users adopt cloud storage services such as MEO Cloud, which stands out as one of the products in Altice's portfolio. MEO Cloud is a cloud file storage and synchronization platform where users can create an account to store and manage images, documents, music and videos from any associated device. As users accumulate large quantities of images on these platforms, it becomes necessary to develop efficient search methods. Searching by date or file type in large sets of images is inefficient, and classifying images is a complex task, given the requirements of users who want to obtain a given visual semantic context [1].

Deep learning, a sub-area of artificial intelligence (AI), has artificial neural networks as its algorithms and compared to other algorithms, has greater performance when dealing with large volumes of data, whilst also understanding different types of data and efficiency in solving tasks. This sub-area makes it possible to improve traditional approaches and define new ones in the context of image search. This article aims to present the image search system developed for the MEO Cloud platform, which is not just adapted, but designed with user needs in mind, by implementing innovative image search approaches using deep learning models that are highly efficient in solving tasks.

Related work

A compelling image search and retrieval system must organize and search for images accurately, meeting user queries. As a result, advanced image search systems have focused on two main approaches:

- Text-Based Image Retrieval (TBIR);
- Content-Based Image Retrieval (CBIR).

Innovative AI techniques have significantly advanced, resulting in the Multi-Modal Image Retrieval (MMIR) and Cross-Modal Image Retrieval (CMIR) approaches. These approaches help to overcome limitations by refining semantic discrepancies in results [1] and are categorized as Semantic-Based Image Retrieval (SBIR).

Text-Based Image Retrieval

The TBIR approach gives the users complete control when they express themselves textually and want to obtain specific results, such as when searching for images in their gallery. The search is based on keywords or descriptions provided by the user, which are compared with the textual annotations attached to the image, such as name, tags or descriptions. The comparison is made using text-matching algorithms such as Bag-of-Words (BoW), Natural Language Processing (NLP) and Boolean retrieval, as well as indexing and search methods [5] [7]. All images with similar textual attributes are then presented to the user.

The manual description of textual characteristics requires considerable effort due to the quantity of images and the need for direct human involvement [8]. Recently, a focus has been made on automating image categorization through annotations that reflect the image's content [8]. Automatic image annotation can be categorized into three groups:

Keyword-based methods

Assigning a class or set of classes to the image using classifier models such as Convolutional Neural Networks (CNN), treating automatic annotation as a supervised classification problem [9].

Ontology-based methods

Representing keywords and their relationships in hierarchical categories, using ontological algorithms to capture semantic content [9].

Generative methods

Textual descriptions that represent the visual content are generated using generative language models that learn to associate visual characteristics of images with representative descriptions, making it possible to capture the semantic context [10].

Content-Based Image Retrieval

Image search systems based on the CBIR approach extract and store the visual feature vectors obtained from the images in storage in a database. During the search, the query image's low-level visual characteristics (such as colour, shape and texture) are determined and compared to the stored vectors [1]. These systems aim to help the user provide a query image in the search and retrieve similar images. This approach is widely used in e-commerce platforms, where users can search for visually similar products. [11].

However, the features used by these systems may not be effective when they are extracted from the entire image. The interference of different backgrounds, overlaps, obstructions and clutter makes it difficult to capture important properties of objects and regions in the images [9]. Another factor is the assumption that semantic similarity corresponds to visual similarity, which can be inaccurate [11].

Multi-Modal Image Retrieval

In the MMIR approach, textual annotations are combined with visual characteristics extracted from the image to improve the relevance of the results. These systems can be applied to e-commerce, extending the CBIR approach. The search is more precise and comprehensive, and results are selected for their textual correspondence and visual similarity to the query image. The inference on the textual description can be extended to its structure; for example, if a user searches for "beach without people", the system searches for the words "beach" and "people", also considering the term "without" [9].

The retrieved images not only match a keyword but also possess the appropriate visual characteristics, such as colour, size, or image type [1]. During an image search, the results that match the keywords in the textual query are retrieved, and the low-level visual characteristics are extracted to present the images with the most similar characteristics [1]. The approach of combining text and image in image search is particularly effective in managing to reduce some semantic discrepancies between the user's search and the obtained outcomes, providing reassurance about the system's effectiveness in handling complex queries.

Cross-Modal Image Retrieval

The CMIR approach checks between the user's natural language and the visual content of images. It is based on cross-modal comparison, in which text and image are compared. The characteristics and links of these data are analyzed, calling them multimodal text-to-image relationships [10]. The aim is to evaluate the similarity between text and image, transposing the data into the same multi-dimensional space through their embeddings (**Figure 2**). During the storage and search phases, a model capable of generating text and image embeddings is used.

The CMIR approach uses pairwise learning, in which a cross-loss function calculates the similarity between pairs of text-image embeddings. Associated text-image pairs have a higher similarity than unrelated ones [10]. In this way, the textual description provided and the images to be searched are represented similarly without resorting to textual annotations or visual characteristics.

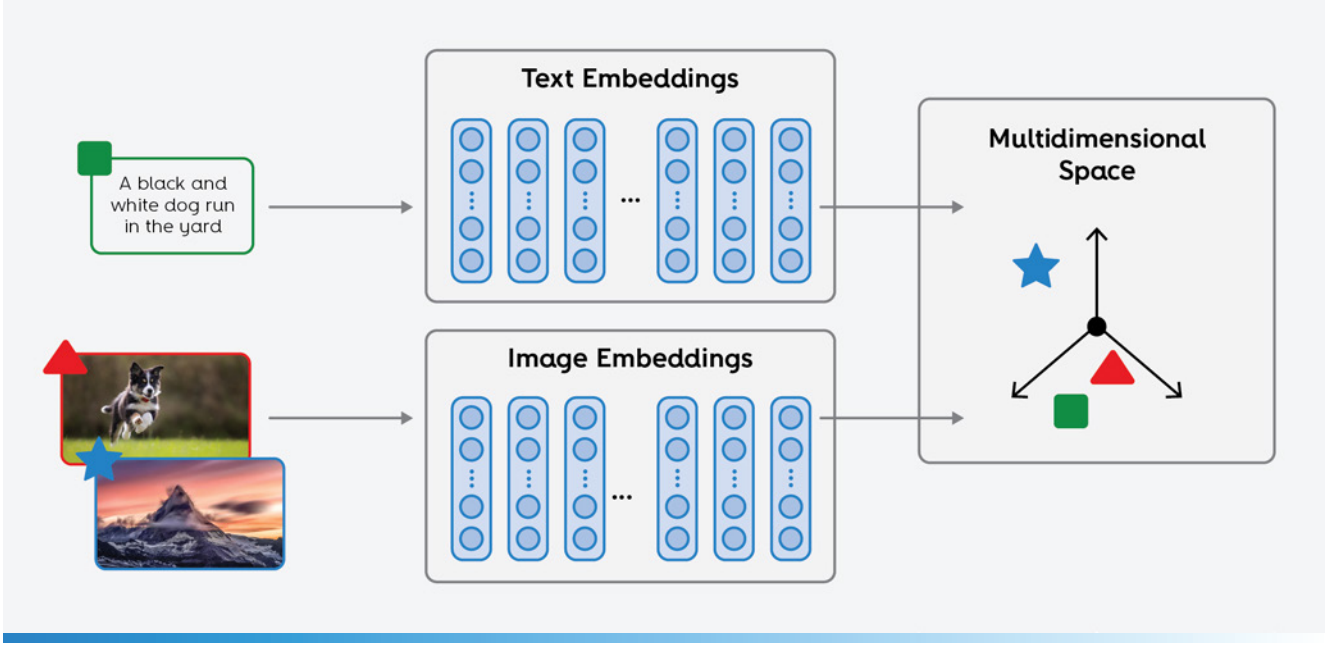


Figure 2 – CMIR approach

Embeddings are low-dimensional numeric vector representations of data, capturing semantic relationships. They make it possible to place semantically similar relationships closer together in a multidimensional space, encapsulating the characteristics of the vectors [13] [12]. The similarity between vectors is measured using similarity and distance functions, with cosine similarity being the most common, as it's more suitable for unstructured data. It allows results that do not have relationships to be ignored [13]. This function measures the cosine of the angle between two vectors, and the smaller the angle, the more similar the vectors are.

The CMIR approach is reconciled with the use of **vector databases**. Vector databases are used to store vector representations of embeddings. They are designed to store, index, and retrieve data represented in a vector space with multiple dimensions. They are suitable for AI applications where data often takes the form of embedding vectors. In the storage process, images are processed by a model, a crucial component responsible for generating image embeddings and persisting this data. When searching for images, the vector database compares the embedding vectors of the images with the embedding vectors of the textual description processed by the model, using search and indexing algorithms such as Hierarchical Navigable Small World (HNSW) [14].

Developed system

The MEO Cloud platform is essential for users to store, search, and remove images, promoting continuous and iterative interaction. To guarantee the integration of the system developed for the platform, the following functionalities are included:

- **Image storage:** this functionality allows users to retain several images in the cloud from devices such as computers, smartphones, or tablets, ensuring that the images are accessible (**Figure 3**).

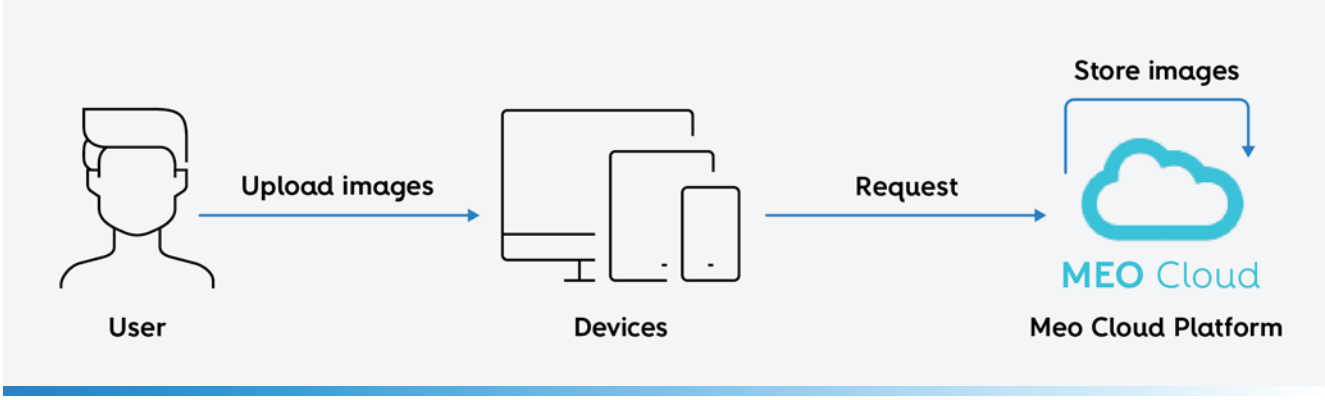


Figure 3 – Image storage

- **Image search:** this functionality locates specific images based on defined criteria, such as textual descriptions, allowing users to specify desired characteristics to find the images (**Figure 4**).

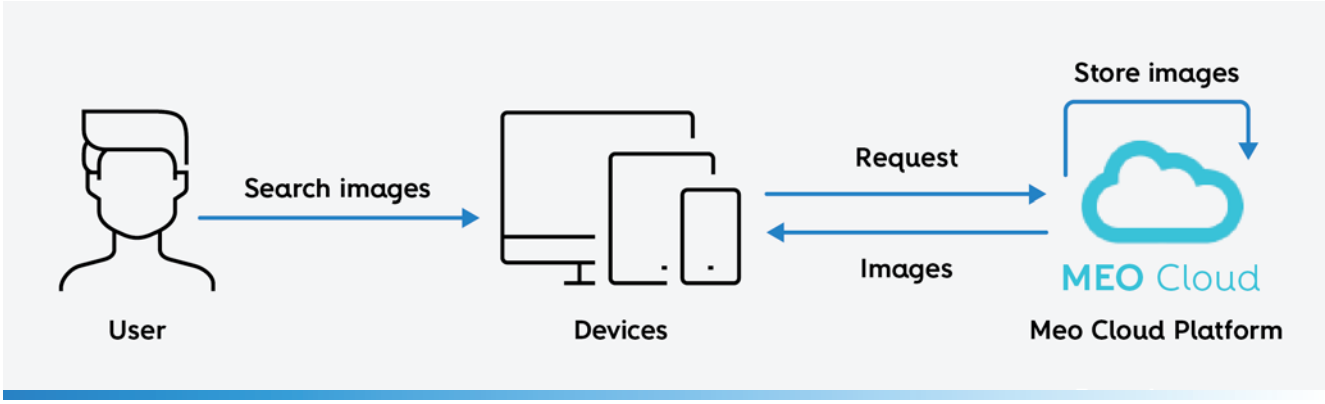


Figure 4 – Image search

- **Image removal:** this functionality allows users to delete images from the cloud, enabling efficient image gallery management (**Figure 5**).

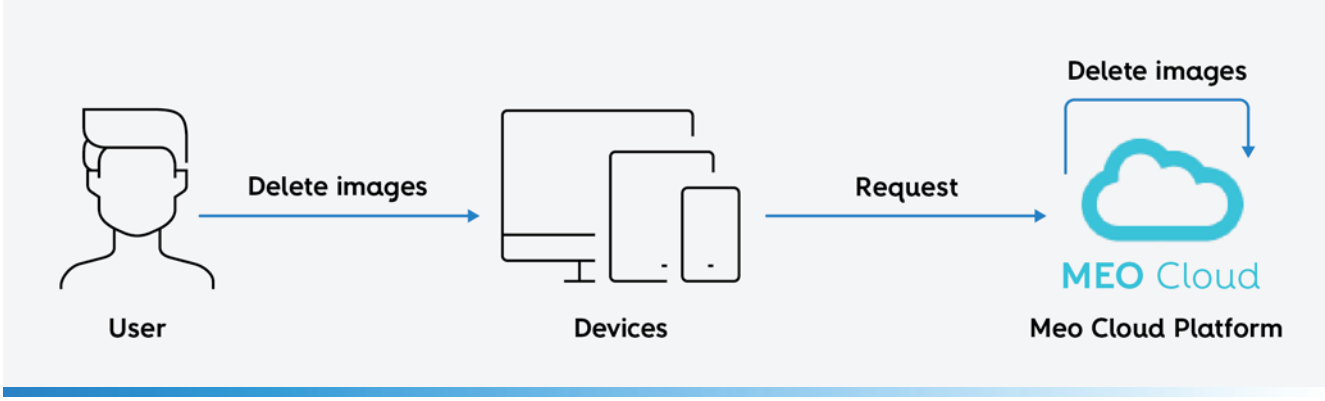


Figure 5 – Image removal

The system developed is a microservice of the MEO Cloud platform, providing a scalable image management and search solution. Being synchronized with the platform guarantees the integrity and consistency of user data. It comprises two layers: the logical layer, where the image search approaches are implemented, and the persistence layer, which stores the data for executing the approaches. Data coherence between the platform and the microservice is ensured by elements that play a crucial role in uniquely identifying the information on users and the images flowing between services. These identifiers, as follows, are established:

- **User UUID:** the data needed to search for images is stored separately in a collection of the system's vector database identified by the user's UUID.
- **Image UUID:** image data is stored as objects identified by the UUID of the corresponding image in the user's collection.

Selected approaches

The TBIR (using keywords and ontologies), CBIR and MMIR approaches all have limitations that affect the accuracy of the results. In the TBIR approach, keywords and ontologies cannot represent the semantic relationships between words and images, and using inappropriate terms results in irrelevant outcomes. In the CBIR, the visual characteristics of images often do not capture their semantic content, resulting in a gap between high-level concepts and low-level characteristics. Lastly, in the MMIR approach, by combining methods from TBIR and CBIR, there are gaps in the semantic particularities between user descriptions and images. The main limitation is the semantic discrepancy between the user's query and the images obtained, which is influenced by factors such as:

- **Ontological reasoning:** ability to analyze semantic relationships between words, such as synonyms, hyperonyms and homonyms [9].
- **Objects and regions:** identification of objects or regions of interest in the image[9].
- **Spatial context:** the spatial context of regions, objects and scenes within an image, especially for queries involving spatial prepositions such as "next to", "over", "left", and "bottom" [9].

The CMIR and TBIR approaches (with subtitles) were chosen to overcome this limitation due to their complementarity and effectiveness in capturing the semantic relationships of user search, which is carried out by textual description. Their combination makes it possible to create synergy, being implemented through state-of-the-art AI models in related tasks. The architecture (**Figure 6**) was designed for integration with the MEO Cloud platform, incorporating the selected approaches, the AI models and the vector database.

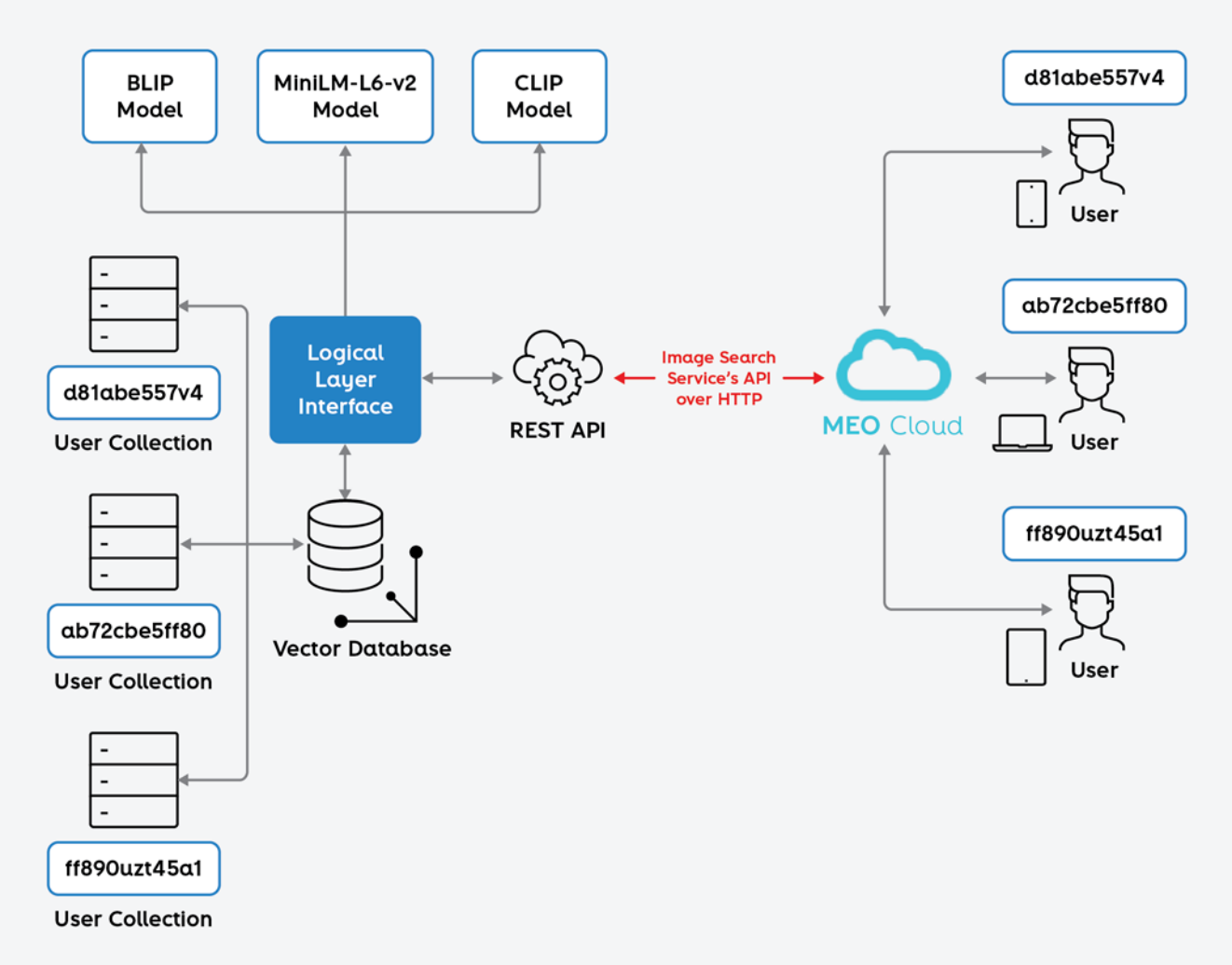


Figure 6 – System architecture

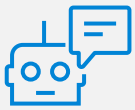
The CMIR approach allows the search to be carried out by joining text and image through embeddings and assessing the similarity between text-image pairs. The TBIR approach filters out irrelevant images obtained from the search through textual semantic analysis of the captions generated from the images, evaluating the similarity between text pairs. The system focuses on performing three main tasks:

Embeddings generation

The Contrastive Language-Image Pre-training (CLIP) model generates the embedding vectors of the images and user description to perform the search. CLIP is a multimodal Language Vision Transformer trained on a set of 400 million text-image pairs, using contrast learning to predict the correct correspondences between images and descriptions. This model can perform multimodal inference by integrating text and image in the same multidimensional space and is used for tasks such as image classification, image generation, similarity search and object tracking.

Captions generation

The Bootstrapping Language-Image Pre-training (BLIP) model automatically generates the captions that reflect the visual content of the images. These captions persist as image metadata in the vector database for filtering results. BLIP is a Visual Transformer that stands out for its flexibility in comprehending and generating tasks involving vision and language.



Semantic textual analysis

The MiniLM-L6-v2 model, known for its robustness, semantically evaluates the similarity of image captions and the user's description during an image search. Learning robust representations of textual sequences, MiniLM-L6-v2 is a Sentence Transformer that has been trained with more than a billion training pairs using contrast learning, ensuring its reliability.

The system has a vector database that facilitates the persistence of the data required for the proposed system's logic. The embeddings of the images are preserved, and the generated captions are stored as metadata, while the images remain stored on the MEO Cloud platform. Integrating the vector database guarantees the system's scalability in operations, the number of users, user data isolation, and image search efficiency.

Image store

After being stored on the MEO Cloud platform, the images are forwarded to the system along with their UUIDs and the user's UUID. The input images pass through the CLIP processor, where they are resized, normalized and converted into tensors. These tensors are fed to the CLIP model, which extracts visual characteristics from the images at different levels of abstraction. The features are transformed into 512-byte embedding vectors, representing the image in a multidimensional space. The embeddings are detached from the computation graph to ensure precision and efficiency. They are later converted into Numpy matrices for easy data manipulation and transformed into lists. The embeddings are extended to reduce precision and increase efficiency, resulting in a one-dimensional list that combines all the elements.

At the same time, the BLIP processor prepares the image, converting it into a compatible tensor. The tensor is unsqueezed, adding an extra dimension to create the batch format needed for input into the model. BLIP generates three textual descriptions for each image using the *nucleus sampling* method, which allows for more varied and creative descriptions. In storage, the objects containing the embeddings and captions as metadata, identified by the UUIDs of the respective images, are persisted in the user's collection identified by their UUID in the vector database (**Figure 7**).

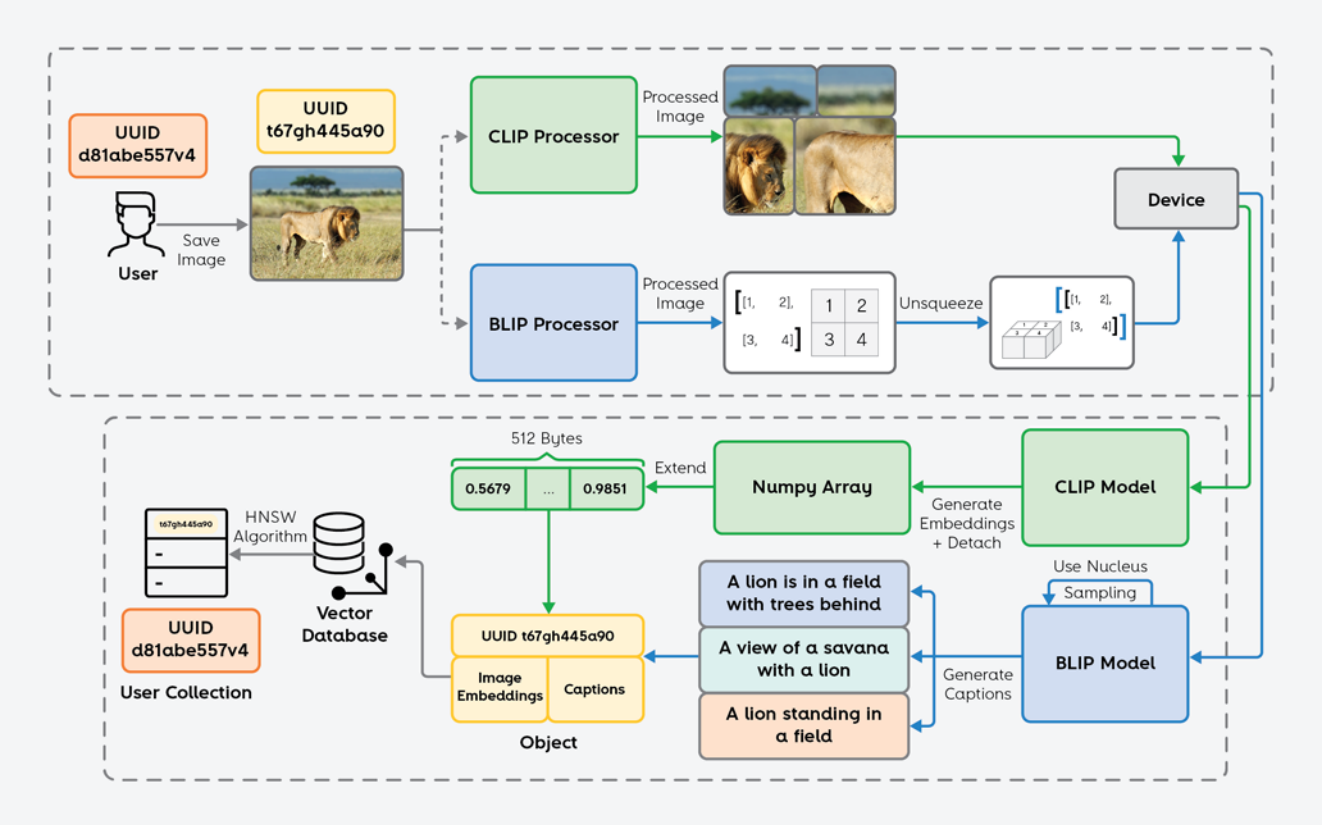


Figure 7 – Image storage pipeline

Image search

The system receives the text description and the user's UUID from the MEO Cloud platform. In the search process, the textual description is divided into tokens by the CLIP tokenizer based on the structure of natural language. Each token is given a unique identifier, allowing the CLIP model to process the textual description. The tokens are supplied to the model, which extracts semantic features and creates a 512-byte embeddings vector, representing the description in multidimensional space. The embeddings vector is provided to the vector database, which uses the HNSW algorithm to compare the similarity between the image embeddings and the query vector. The most similar objects are retrieved (**Figure 8**).

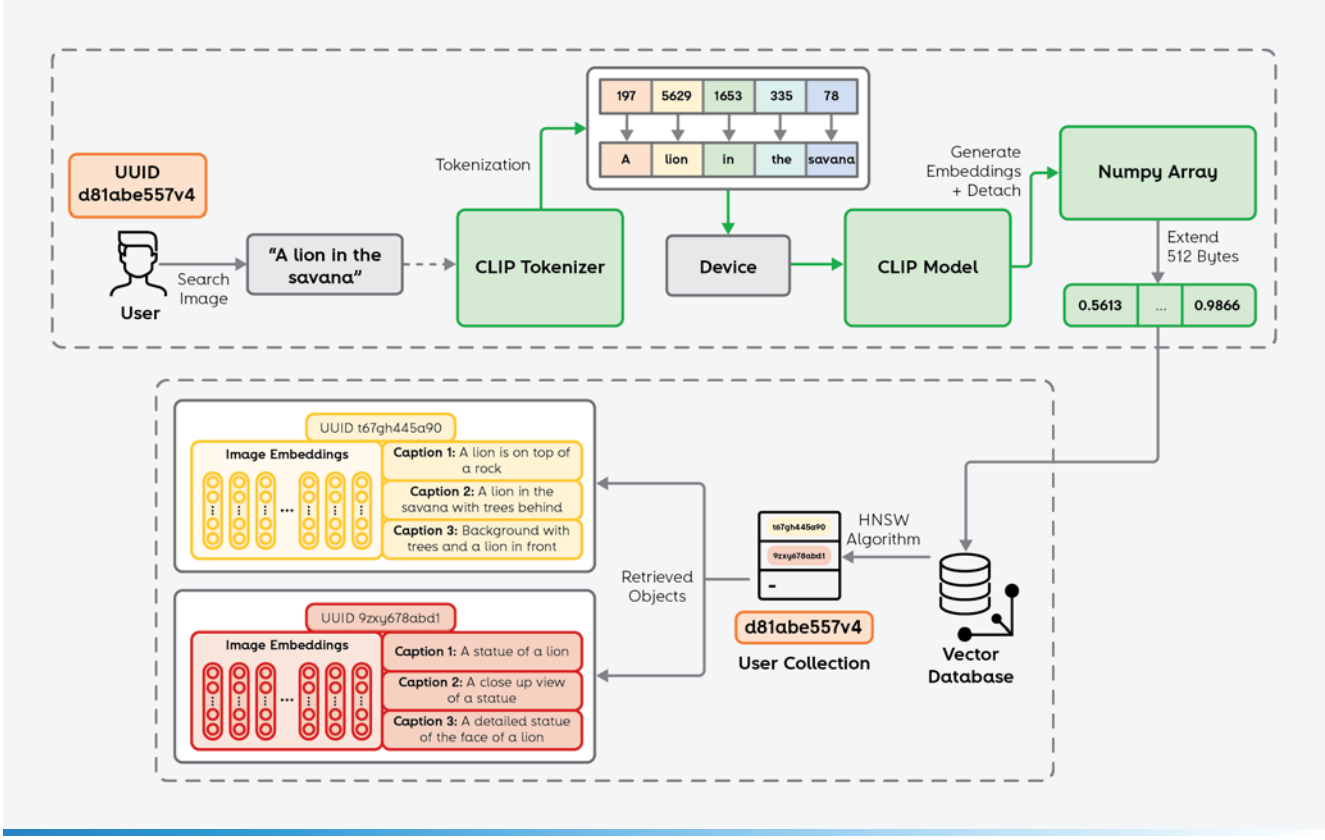


Figure 8 – Image search pipeline - retrieving results

For each retrieved object whose distance between its embedding vectors and the embedding vector of the textual description is within a specific rectification interval, the MiniLM-L6-v2 model is used for semantic analysis. The user description and image captions are encoded by the MiniLM-L6- v2 tokenizer, adjusting and truncating the sequences. The tensors are processed by the model, which generates the embeddings of the tokens extracted by the self-attention mask. The embeddings are summed and normalized by the L2 norm to ensure a unit magnitude. The similarity

between the captions and the description is calculated using the cosine distance. The image is considered relevant and kept in the results if the similarity value is high. Finally, the list of UUIDs of the objects resulting from the search is returned to the MEO Cloud platform, responsible for presenting the corresponding images to the user (**Figure 9**).

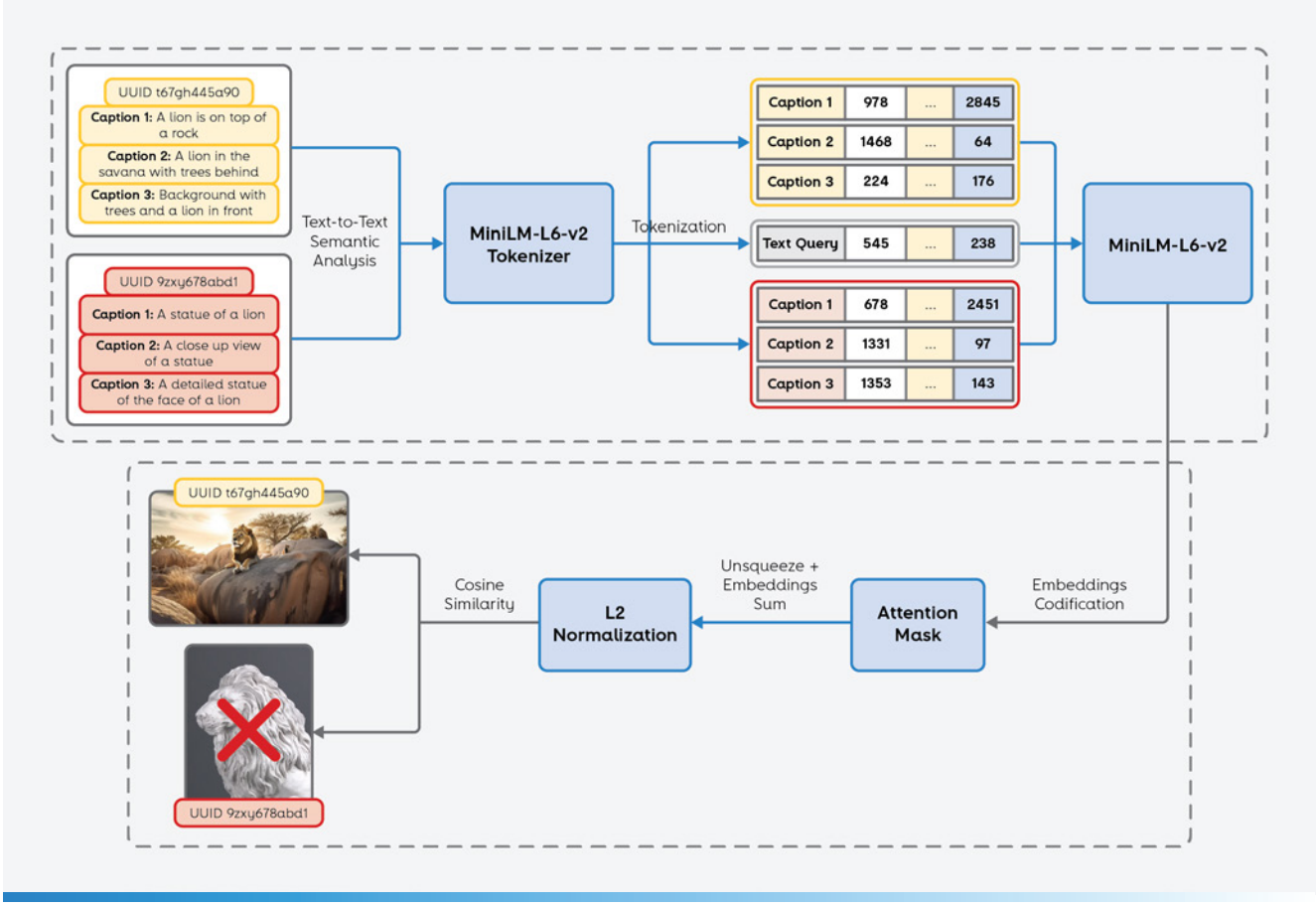


Figure 9 – Image search pipeline - filtering results

Image removal

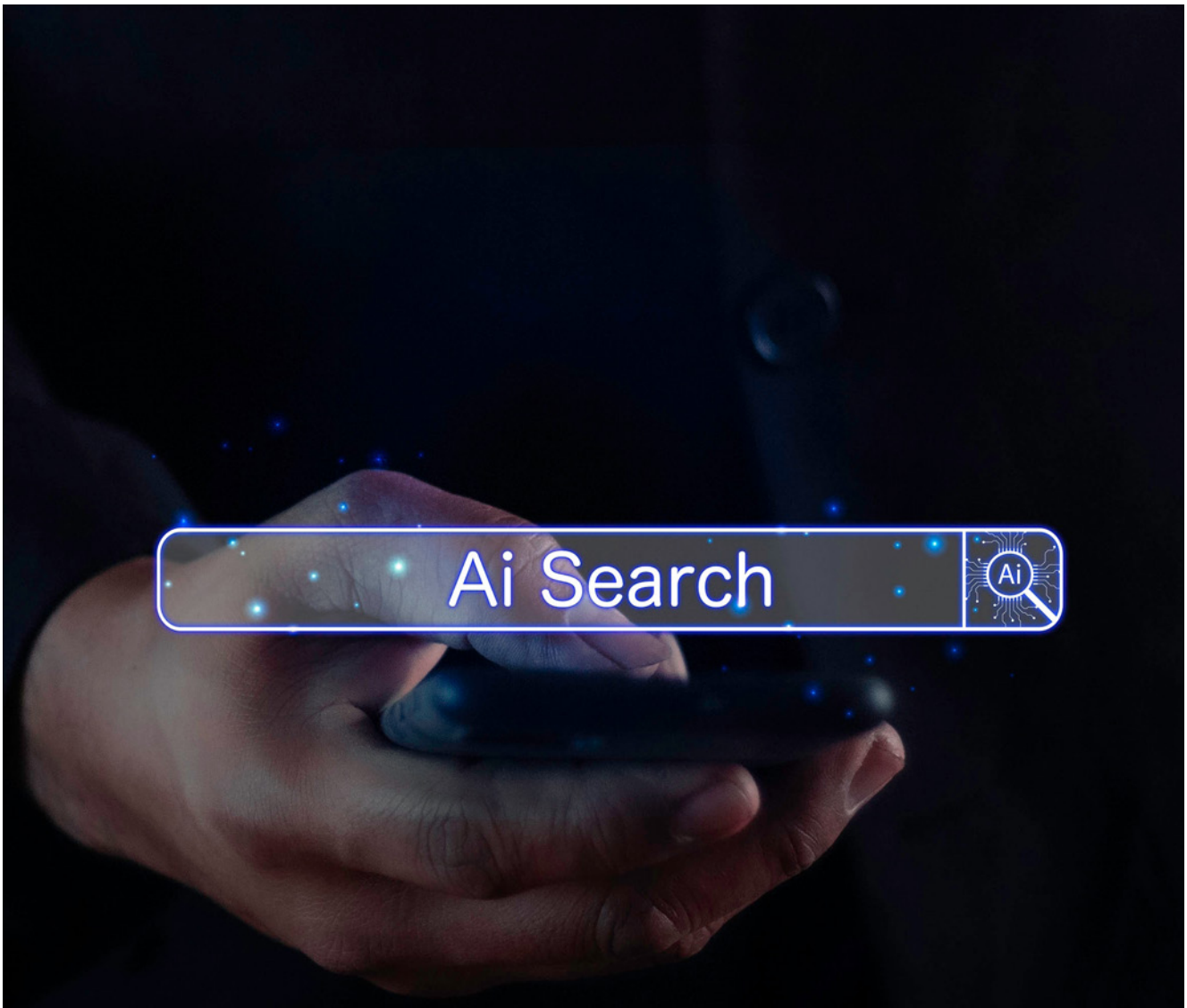
Models are not used to remove images, only the logical component is carried out using operations on the vector database. Once the images have been removed from the MEO Cloud platform, the UUID list of the images and the user's UUID are sent to the system. The vector database accesses the user's collection via their UUID. Using the UUIDs of the images in the list received, the HNSW search algorithm searches for and removes objects relating to the images containing associated information, such as embeddings and metadata. In this way, the vector database is kept up to date and synchronized with the MEO Cloud platform, ensuring that no outdated information remains stored after removing the images.

Conclusion

The solution developed significantly extends the MEO Cloud product by combining innovative image search approaches with AI models that unite text and image in the same multidimensional space, such as CLIP. The system enables efficient image searches in scenarios with many images, such as the user's cloud, overcoming traditional search limitations. In this way, the problem of the semantic discrepancy between the user's requirements and the images obtained is solved.

AI models such as MiniLM-L6-v2, which analyzes the semantics between the user's text description and the image captions, can generate more accurate results and filter out irrelevant results. Image captions are created using generative models such as BLIP, which describe the visual content in a detailed and precise way.

However, ensuring that the images align with the user's expectations is significantly affected by the subjectivity of the results. The variability in individual users' perceptions and expectations of images makes it challenging to cater for all cases uniformly. Although the intended semantics can be understood and analyzed through the models, each user's preferences and interpretations add a layer of complexity. This subjectivity makes creating a system that perfectly meets all individual needs difficult.



References

[1] Marina Ivasic-Kos. Application of digital images and corresponding image retrieval paradigm. ENTRENOVA-ENTerprise REsearch InNOVAtion, 8:350–363, 11 2022. doi: 10.54820/entrenova-2022-0030.

[2] Matic Broz. How many pictures are there (2024): Statistics, trends, and forecasts, 2022. https://photutorial.com/photos-statistics/

[3] Rohini Patil, Tanvi Nerurkar, Shivani Patil, and Asawari More. Cloud based storage system like dropbox. International Research Journal of Modernization in Engineering Technology and Science @International Research Journal of Modernization in Engineering, 864, 2582–5208. 2020.

[4] Myasar Mundher Adnan, Mohd Shafry Mohd Rahim, Amjad Rehman, Zahid Mehmood, Tanzila Saba, and Rizwan Ali Naqvi. Automatic image annotation based on deep learning models: A systematic review and future challenges. 2021. ISSN 21693536.

[5] Syed Ali Jafar Zaidi, Attaullah Buriro, Mohammad Riaz, Athar Mahboob, and Mohammad Noman Riaz. Implementation and comparison of text-based image retrieval schemes. International Journal of Advanced Computer Science and Applications, 10(1), 2019. doi: 10.14569/IJACSA.2019.0100177.

[6] P. M. Ashok Kumar, T. Subha Mastan Rao, L. Arun Raj, and E. Pugazhendi. An efficient text- based image retrieval using natural language processing (nlp) techniques. volume 1171, pages 505–519. Springer, 2021. ISBN 9789811553998. doi: 10.1007/978-981-15-5400-1 52.

[7] T Karthikeyan, P Manikandaprabhu, and S Nithya. A survey on text and content based image retrieval system for image mining, 2014. ISSN 2278-0181.

[8] Jianlong Fu and Yong Rui. Advances in deep learning approaches for image tagging. APSIPA Transactions on Signal and Information Processing, 6, 2017. ISSN 20487703. doi: 10.1017/ATSIP.2017.12.

[9] Alaa el-din mohamed Riad, Hamdy. K. Elminir, and Sameh Abd ElGhany. A literature review of image retrieval based on semantic concept. International Journal of Computer Applications, 40:12–19, 12 2012. doi: 10.5120/5008-7327.

[10] Jianan Chen, Lu Zhang, Cong Bai, and Kidiyo Kpalma. Review of recent deep learning based methods for image-text retrieval. 8 2020. doi: 10.1109/MIPR49039.2020.00042.

[11] Amit Kumar Nath and Andy Wang. A survey on personal image retrieval systems. 07 2021. doi: 10.48550/arXiv.2107.04681.

[12] Alexandre Bonnet. Image embeddings to improve model performance. 2023. https://encord.com/blog/image-embeddings-to-improve-model-performance/

[13] Ben Leder. What is a vector database? 2023. https://www.phdata.io/blog/what-is-a-vector-database/

[14] Pavan Belagatti. A deep dive into vector databases. 2023. https://www.singlestore.com/blog/a-complete-guide-to-vector-databases/

Acronyms

AI	Artificial Intelligence
BoW	Bag-of-Words
CBIR	Content-Based Image Retrieval
CNN	Convolutional Neural Network
CMIR	Cross-Modal Image Retrieval
HNSW	Hierarchical Navigable Small World
MMIR	Multi-Modal Image Retrieval
NLP	Natural Language Processing
SBIR	Semantic-Based Image Retrieval
TBIR	Text-Based Image Retrieval

Authors

Francisco Reis Izquierdo

AI software engineer intern

Altice Labs, Portugal

francisco-r-izquierdo@alticelabs.com

https://www.linkedin.com/in/francisco-izquierdo-a45856242

Contacts

Address

Rua Eng. José Ferreira Pinto Basto
3810 - 106 Aveiro (PORTUGAL)

Phone

+351 234 403 200

Media

contact@alticelabs.com
www.alticelabs.com