A Cross-modal Image Retrieval Application based on SigLIP and FAISS

Jiarui Liang Ruiqi Liu Yidan Cheng
Department of Computer Science, Rice University
jl361@rice.edu rl115@rice.edu yc181@rice.edu

Abstract

This report details the development of a cross-modal image retrieval application designed to facilitate efficient image-image retrieval through advanced vectorization and indexing techniques alongside text-image retrieval. The application allows users to upload images or input text captions and receive relevant image results. Key methodologies include the adoption of Sigmoid loss for Language-Image Pre-training (SigLIP) model for converting images into vector representations and leveraging Meta's FAISS library for fast, efficient indexing and retrieval. A significant portion of the SBU dataset, comprising one million captioned images, has been processed to support the application's functionality. The application integrates these components into an existing web demo, enhanced to support both text and image searches with a user-friendly interface. Our demo is published under http://18.216.139.37/; our code is on github; we also provide slides and presentation video.

1. Introduction

In an era where the volume of digital content, particularly images, grows exponentially, the ability to quickly and accurately retrieve information based on visual data becomes important in search engine. Our project endeavors to bridge the gap between conventional text-based search mechanisms and the dynamic realm of image-based search queries. Our purpose is to process user-uploaded images to yield relevant, visually similar results alongside text-image search functionality to achieve a cross-modal image retrieval system.

The cross-modal image retrieval system is planned to be developed by combining implementation of SigLIP, a Vision-Language Model capable of image/text vectorization and FAISS (Facebook AI similarity search), a library for efficient similarity search for indexing and fast retrieval of image vectors. After that, the system is implemented through VanillaJS and Flask into a responsive user interface, where users can users can either input text captions or upload images and then obtain relevant results (similar images). The web demo is deployed using Amazon Elastic Compute Cloud (AWS EC2) to allow for live responsive user interactions.

2. Related Work

We discovered several prior work about cross-modal image retrieval system. One of them is CLIP-RS, which employs a finetuned implementation of Contrastive Language Image Pre-training (CLIP) and combines CLIP and FAISS to achieve image retrieval functions. Additionally, the platform is deployed as a Web App to perform inference tasks for text-to-image and image-to-image retrieval of remote sensing (RS) images obtained via the Mapbox GL JS API. Furthermore, the platform offers additional features such as image similarity search, image geolocation on a map, and access to images' geocoordinates and addresses [1].

3. Methodology

To implement our search engine, we integrated SigLIP to convert texts/images into vector representations and used FAISS Library to manage vectors and achieve fast image retrieval. Figure 1 shows how the entire system works.

3.1. Dataset

To achieve image retrieval function, we utilized SBU dataset [4] containing 1 million captioned images. This dataset supports the basic mechanism of fast image retrieval and provides data resources for search evaluation such as caption data; however, during vectorization, we observed that among each 10,000 images, approximately 20% of images failed loading into the model due to url not found.

3.2. Data Vectorization

Based on previous literatures, we decided to use SigLIP, an improved vision and language model based on CLIP to implement data vectorization.

3.2.1 SigLIP Model

The SigLIP model, introduced by [6], is a variant of the CLIP model[5], featuring a different loss function. SigLIP

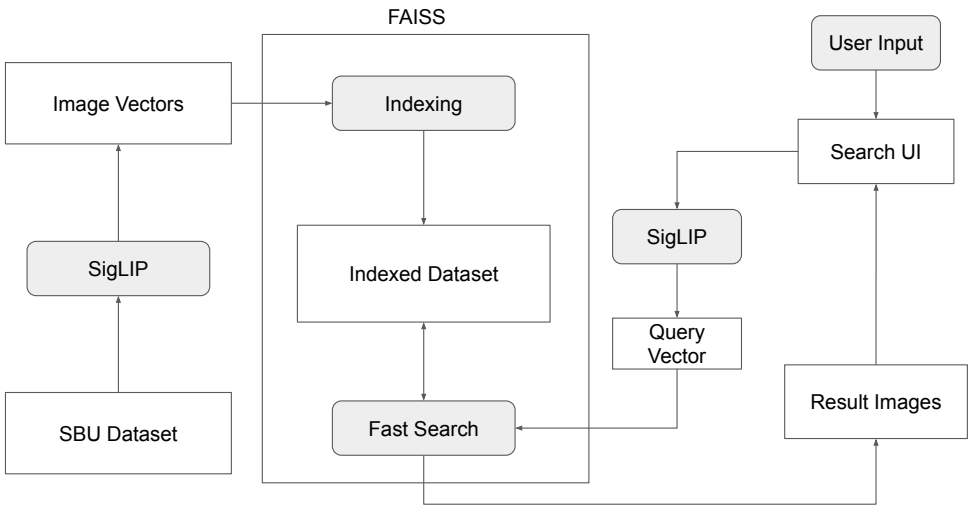

Figure 1: Process Diagram showing how the entire application works

exhibits better performance in zero-shot classification accuracy on ImageNet. Unlike conventional contrastive learning methods that utilize softmax normalization, SigLIP employs a straightforward pairwise sigmoid loss function by exclusively focusing on image-text pairs.

Researchers conducted an image-text retrieval assessment across 36 different languages and by utilizing the scaled-up mSigLIP ViT-B model, they achieved a competitive Recall rate of 42.6% for image retrieval and 54.1% for text retrieval, which supports the our requirements of using caption or image vectors to retrieve relevant images.

3.2.2 Feature Extraction

To vectorize the images from the SBU dataset and image queries, we used `get_image_feature()` method in SigLIPModel. To vectorize caption queries, we used `get_text_feature()` method in SigLIPModel. As we observed that the model outputs a 1×768 vector embedding, we set the indexer in FAISS to 768.

3.3. Indexing and Search

We adopted FAISS Library's `IndexFlatL2` and `IndexHNSWFlat` indexing methods to build up vector databases.

3.3.1 FAISS Library

Meta's FAISS Library is utilized to index image vector embeddings and search queries. FAISS Library [2] is a framework designed for vector similarity search. It facilitates the functionality of vector databases, which are responsible for handling extensive collections of embedding vectors. FAISS also comprises a toolkit of indexing techniques and associates operations utilized for tasks such as searching, clustering, compressing, and transforming vectors. During the search process, FAISS retrieves distances computed during indexing and returns the k nearest neighbors.

3.3.2 IndexFlatL2

`IndexFlatL2` is an accurate indexing method provided by FAISS that is calculated based on the Euclidean L2 distance between vectors. For a query vector $\mathbf{q}$ and a dataset of vectors $\mathbf{x}_i$, the search algorithm computes the L2 distance between the query vector and each vector in the dataset:

$$\text{distance}_i = \sqrt{\sum_{j=1}^{d} (q_j - x_{ij})^2}$$

Here, x_{ij} represents the j-th component of the i-th vector in the dataset.

During the search process, each indexed vector is decoded and compared to the query vectors one by one to guarantee precise search results.

3.3.3 IndexHNSWFlat

`IndexHNSWFlat` is a flat index topped with with a Hierarchical Navigable Small World (HNSW) structure to access elements more efficiently.

HNSW is a state-of-the-art algorithm used for an approximate search of nearest neighbours. It is a multi-layer structure consisting of a hierarchical set of proximity graphs for nested subsets of the stored elements. The maximum layer in which an element is present is selected randomly with an exponentially decaying probability distribution. This allows producing graphs similar to the previously studied Navigable Small World (NSW) structures while additionally having the links separated by their characteristic distance scales. Starting from the upper layer, the search process greedily traverses the layer nodes to find a local nearest neighbour. After that, the search switches to the lower layer and restarts from the element which was the local minimum in the previous layer and the process repeats [3].

3.4. Web Demo

We integrated the search function into a web demo where users can input text/image query. We adopted Flask to build the endpoints of search-by-caption and search-by-image. User query will be vectorized by SigLIP into embeddings; the embeddings will be utilized to retrieve top K results among FAISS index database. After that, returned images will be displayed in the search UI in the order of relevancy ranked by image embeddings similarity. The web application is deployed using AWS EC2, a scalable computing cloud service provided by Amazon Web Services (AWS).

4. Evaluation and Results

To quantify evaluation of our search engine, we proposed two evaluation methods: caption-based evaluation and user rating. We also evaluated the search-by-image function. Since it is not easy to quantify, we displayed some results in 3. During the evaluation, we observed that some images with text on them significantly influence the results. Therefore, to obtain a more accurate assessment, we filtered out the most frequent non-relevant results collected from testing queries and then evaluated the accuracy of image retrieval.

4.1. Caption-based Evaluation

The caption-based evaluation method is assessed based on Recall@K. We selected random images as groundtruth data from SBU dataset and utilized caption of the images as query to test image retrieval. And by querying N images, we can obtain Recall@K:

$$\text{Recall@K} = \frac{\sum_{i=1}^{N} I(i)}{N}$$

where $I(i)$ is a binary data (0 or 1) indicating whether the groundtruth image appears in the top K results (if it appears, return 1, otherwise 0).

We computed the Recall@K for K = 3, 5, 10, 20, 50 and N = 50 for two indexing methods - `IndexFlatL2` and `IndexHNSWFlat`. According to the results, when K gets larger, both Recall rates increases, but compared with `IndexFlatL2`, `IndexHNSWFlat` has a general lower Recall. Regarding speed of processing, we can tell that `IndexHNSWFlat` performs faster retrieval than `IndexFlatL2`. Detailed results are shown in Table1.

4.2. User Rating

Alternatively, we consider a user-centric evaluation approach, where volunteers judge the relevancy search results based on their personal judgment and then compute mean average precision for top K results(mAP@K):

$$mAP@K = \frac{\sum_{i}^{K} AP(i)}{N}$$

where K is the number of results, $\sum_{i}^{K} AP(i)$ is the average precision of top k results and N is the number of queries.

In this user rating, we involved 3 volunteers to assess top 3 and top 5 results and then we computed the Precision@3 and Precision@5 by computing average of their score. After that, we obtained Average Precision (AP) for each query. Detailed Results are shown in Table 2. Overall, the mAP of 10 queries is 0.453.

5. Discussion

Based on our evaluation statistics, the score derived from caption-based assessment is notably lower, while the user ratings score higher than expected. Upon closer examination, we found that the certain caption data fail to accurately describe what the images convey compared with users' interpretation. This misalignment between the caption and the image content not only explains the score discrepancy but also contributes to the challenge of achieving high Recall in caption-based evaluation.

Moreover, our evaluation revealed frequent instances where irrelevant images were prominently displayed. Through analysis, we identified that this phenomenon is partly attributed to images containing text. Given that SigLIP is a multi-modal vision and language model trained solely on image-text pairs, it appears to show higher sensitivity to textual elements within images during text-to-image retrieval tasks.

6. Conclusion

In this project, we developed an image retrieval application based on SigLIP, a multi-modal vision and language model and FAISS, a library for vector similarity search. We firstly vectorized images in the SBU dataset and applied IndexFlatL2 indexing and IndexHNSWFlat indexing to build up a vector databases. Then we developed a backend server using Flask and deployed the application on AWS EC2. To

Indexing Method	Recall@3	Recall@5	Recall@10	Recall@20	Recall@50	Run-Time@50 (s)
IndexFlatL2	0.08	0.1	0.12	0.14	0.18	3.085
IndexHNSWFlat	0.02	0.06	0.06	0.1	0.14	3.602

Table 1: This table shows the Recall@K results for 2 indexing method for K = 3, 5, 10, 20, 50 and their run time in seconds for K = 50. The data is computed by running 50 queries on the first 100k data.

No.	Caption Query	Precision@3	Precision@5	AP
1	A tree in a frosty field in Fernhill Heath	0.33	0.6	0.47
2	blue girl relaxes in the sunshine	0.67	0.47	0.57
3	boy cap black denim can be made in any colour or size	0.89	0.73	0.81
4	flower towers in the grass	0.67	0.40	0.54
5	Serenity playing at memere and pepe's house and being entertained by pepe and his guitar	0.44	0.27	0.36
6	The male Foo Dog or lion of Buddha at the Lama Temple in Beijing	0.33	0.40	0.37
7	street sign in HK	0.44	0.27	0.36
8	Red and Silver hearts in black box	0.22	0.20	0.21
9	Full length portrait; Seated black lady in striped dress, small white child stands beside her. Tintype, ninth plate	0.33	0.33	0.33
10	Partridge on a window sill. Missed the pear tree in the back.	0.67	0.40	0.53

Table 2: This table shows Precision@3, Precision@5 and Average Precision Results computed for each query used in user rating. The results were generated on the first 100k data. The captions were randomly chosen from SBU dataset.

No.	Image Query	Top 5 Search Results
1		    
2		    
3		    

Table 3: This table shows the results of search-by-image function under IndexFlatL2 indexing. The 3 image queries were randomly chosen from freeimages.com.

evaluate the search performance of our application, we applied the caption-based evaluation by computing the Recall@K and the user rating evaluation by computing the mean average precision. Based on what we discovered during implementation, we propose two potential improvements for future implementation: enhancing the dataset to ensure more accurate captions, or incorporating a weighting method to reduce the influence of text-heavy images when searched using captions.

References

[1] L. Djoufack Basso. *CLIP-RS: A Cross-modal Remote Sensing Image Retrieval Based on CLIP, a Northern Virginia Case Study*. PhD thesis, Virginia Tech, 2022.

[2] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. 2024.

[3] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020.

[4] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[6] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.