

**Subject:** Multimedia information retrieval

**Class:** CS336.O11.KHCL

**Lecturer:** Ngo Duc Thanh

**Student:** Ngo Xuan Cuong – 21520663

**FINAL PROJECT REPORT**

**TOPIC: CLIP-BASED COMPOSED IMAGE RETRIEVAL**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**VIETNAM NATIONAL UNIVERSITY**

Contents

[I. Introduction: 3](#_Toc156749710)

[II. Survey: 4](#_Toc156749711)

[III. Proposed method: 4](#_Toc156749712)

[1. Searching system building 5](#_Toc156749713)

[2. CLIP fine-tune 5](#_Toc156749714)

[3. Combiner building 6](#_Toc156749715)

[IV. Evaluation: 7](#_Toc156749716)

[1. Dataset 7](#_Toc156749717)

[2. Metric 7](#_Toc156749718)

[3. Result 7](#_Toc156749719)

[V. Conclusion: 8](#_Toc156749720)

1. **Introduction:**

* Image Retrieval, a foundational task in multimedia and computer vision, has evolved over time from relying on engineered features such as SIFT to the adoption of CNNs. This progression is evident in its application across diverse domains such as artworks, cultural heritage, commerce, surveillance, and nature. Essentially, the core process involves formulating a query with just an image, extracting features, and then comparing them to the features of images within a database, which can be depicted in Figure 1:

Feature matching

Feature extraction

Query

Query

Feature extraction

Image database

Figure 1: Image retrieval pipeline.

* This Image retrieval system could be further extended by using queries which is the combination of both image and text to add more information to the query. A short texture description could set more constraints on the reference image, changing its attributes or adding some specific details to the retrieved target. This kind of retrieval system is called Composed Image retrieval.

A black t-shirt with red text

Description automatically generated

A black t-shirt with white text

Description automatically generated

Has Korn logo

Figure 2: Composed image retrieval illustration.

* Composed Image retrieval plays crucial roles in diverse fields like web search, electronic commerce, and security. However, developing solutions for this task is complicated, requiring the integration of user feedback and intent, all while bridging the semantic gap between image and text content.
* In this project, I will demonstrate a system to search images based on CLIP, a deep neural network combining both visual and language modalities that can generate robust features for images and text. CLIP-based features can be effectively used to build a CIR (composed image retrieval) system that can exploit natural language as the constraint of the reference image and the changes to the image requested in the additional text.

1. **Survey:**

* To design an efficient system for image search, searching is essential to have an overview of state-of-the-art methods and the datasets. It is shown that, for image retrieval, the attention is shifted to the combination of many modalities, especially visual and text. Therefore, multimodality models such as CLIP and ALIGN are considered to be the most suitable for the Composed image retrieval task. Another criterion to look for is the feasibility and scalability of the method. The proposed method should be executable in a low-performance device with less than 8Gb CPU and 4Gb GPU, a popular configuration for computers nowadays.
* Following the above criteria, a CLIP-based method that features CIR task with respectable performance that can run on required computer properties is a proper method to be implemented.

1. **Proposed method:**

* Composed image retrieval aims to find the best matching images satisfying the constraints of the query, which consist of a reference image and relative caption that includes a descriptive request from the user about the image. To achieve that, both visual and texture information from the query should be effectively exploited and integrated before being used to retrieve target images.
* For the above reasons, it is assumed that features from images and text should come from the same vector space, which could be obtained from CLIP. CLIP functions as a vision-language model that aligns images with their corresponding texture descriptions within an embedding space. With and are respectively the image and text encoder of CLIP, features representations of both image and text, and , belongs to the same vector space . Furthermore, CLIP was trained to correlate information from images and text, it should be guaranteed that if both mention the same thing.
* However, representing the visual and texture information into the same embedding space is still not sufficient for composed image retrieval as the purpose is to consider the visual features with the aid of texture constraint. Therefore, this method aims to create a combination embedding space where image and text features could be fused through a sum operation. For instance, with an image of a black shirt , a text (“have dog logo”) and an image of a black shirt with dog logo , the goal of this approach is to create a space where + .
* To achieve such an ideal result, a two-stage process approach will be used to address the task. In the first stage, pretrained CLIP is fine-tuned for large-scale CLIP to update with the downstream task. The second stage involves around building a combiner from scratch to obtain image-text feature that can take full advantage of information provided in the query. In both stages, the model is trained with triplets (, ), where () represent the query and is the target image.

A diagram of a company

Description automatically generated

Figure 3: Method illustration.

* In the inference stage, with the query (), a fused feature generated from fine-tuned CLIP and combiner is used to retrieve the target image with cosine distances. Images with the highest similar score are returned as the result of the system.

1. **Searching system building**

* The overall process is described as follows:
* Fine-tune CLIP with the contrastive learning method with 2 datasets FashionIQ and CIRR.
* Building a combiner and training from scratch with the same method as CLIP fine-tuning.
* Extracting all the images in the database with fine-tuned CLIP.
* Extract features from images and text with fine-tuned CLIP and take them to the combiner to get the fused feature.
* Calculating the similarity score between the query and all images in the database with cosine distance.
* Take the images with the highest similarity score and return them to the user.

1. **CLIP fine-tune**

* Given the triplet (, ) represent reference image, text and target image respectively, their features are extracted by CLIP encoders, and , to obtain and , where d denotes the size of CLIP embedding space. Query feature is the element-wise sum of the above features: + = .

A diagram of a diagram

Description automatically generated

Figure 4: CLIP fine-tune.

* The goal of this stage is to minimize the cosine distance between and which can be achieve by minimize the following loss:

where is the cosine similarity and is the temperature parameter to control the logits range. Fine-tuning CLIP with this loss function helps the texture embedding space act as an additive to the visual embedding space, therefore, effectively fusing image and text information to solve the task.

1. **Combiner building**

* To have a fused feature representing the query, a combiner is needed. Overall, the combiner can be viewed as a three-branch summation, with two branches implemented as a convex combination of the two input modalities, and a third branch of a learned image-text mixture.

A diagram of a software system

Description automatically generated

Figure 5: Combiner architecture.

1. **Evaluation:**
2. **Dataset**

* FashionIQ is a fashion dataset, containing 77,684 images collected from the internet. The dataset is divided into different categories: Dress, Toptee and T-shirt. There are 18,000 triplets of a reference image, a pair of additional captions and a target image for training process.
* CIRR is an open domain dataset containing 21,552 real-life images with human-generated modification sentences. It has 36,544 triplets split into train, test and validation set with the proportion of 80%, 10% and 10% respectively.

1. **Metric**

* Recall@K evaluates the proportion of relevant items included in the top K recommendations, where K represents the total number of recommendations given for a user.
* In this report, the model is assessed with Recall@k at 2 rank (10, 50).

1. **Result**

* The model has been tested and has a respectable result compared to the state-of-the-art methods.
* The result when testing in FashionIQ dataset:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Recall@10 | Recall@50 | Released year |
| ARTEMIS | 25.25 | 50.08 | 2022 |
| BLIP4CIR | 43.49 | 67.31 | 2023 |
| CASE | 48.79 | 70.68 | 2023 |
| Candidate Set Re-ranking | 51.17 | 73.13 | 2023 |
| ***Proposed method*** | ***38.32*** | ***61.74*** | ***2022*** |

* The result when testing model in CIRR dataset:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Recall@10 | Recall@50 | Released year |
| ARTEMIS | 61.31 | 87.73 | 2022 |
| ***Proposed method*** | ***81.86*** | ***95.93*** | ***2022*** |
| BLIP4CIR | 83.88 | 96.27 | 2023 |
| CASE | 87.25 | 97.57 | 2023 |
| Candidate Set Re-ranking | 89.78 | 97.18 | 2023 |

* It is clearly seen that the proposed method outperforms all other candidates within the year 2022 by a significant margin as utilizing CLIP makes the features generated more consistent compared to other encoders at that time.
* By contrast, in 2023, novel methods have been proposed and proven to perform better than this method. However, many of them are heavily based on this combiner architecture, proving how effective this method can be.

1. **Conclusion:**

* To conclude, in this project, I have successfully built a composed image retrieval system based on CLIP. This method may not have the best result compared to state-of-the-art methods; however, it is considered to be the foundation to many of them and could be improved to perform better in CIR task, especially with its unique combiner architecture. In practice, this method could perform very well, even in low performance server and it is also scalable to large-scale datasets using techniques commonly employed in standard CBIR systems.
* All the source code and instruction to launch the system is available in the file attached to this report

**REFERENCE**

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned image retrieval for fashion using contrastive learning and CLIP-based features. In Proc. of ACM Multimedia Asia (ACMMM Asia), 2021.

[2] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In CVPR, 2022.

[4] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19305– 19314, 2023.

[5] <https://towardsdatascience.com/evaluation-metrics-for-recommendation-systems-an-overview-71290690ecba>