Zero-Shot Image Retrieval with Human Feedback

Lorenzo Agnolucci*
lorenzo.agnolucci@unifi.it
University of Florence
Italy

Alberto Baldrati*
alberto.baldrati]@unifi.it
University of Florence
Italy

Marco Bertini
marco.bertini@unifi.it
University of Florence
Italy

Alberto Del Bimbo
alberto.delbimbo@unifi.it
University of Florence
Italy

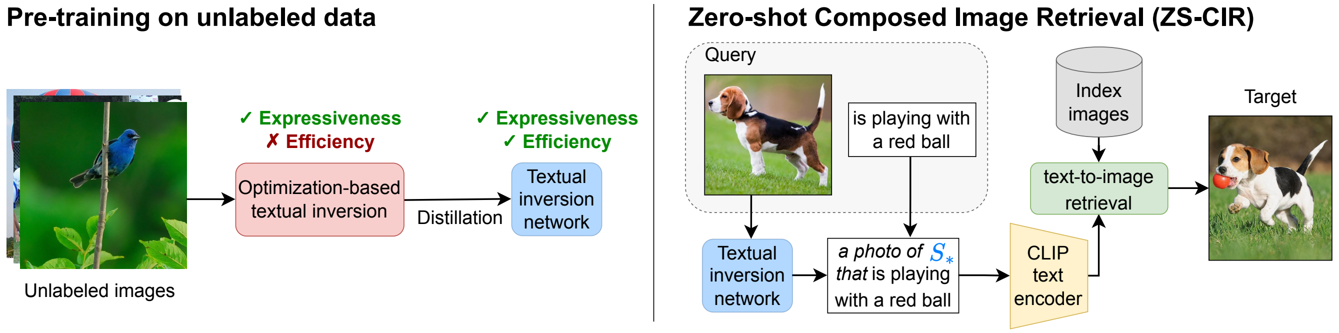

Pre-training on unlabeled data
Unlabeled images
✓ Expressiveness
✗ Efficiency
Optimization-based textual inversion
Distillation
✓ Expressiveness
✓ Efficiency
Textual inversion network
Zero-shot Composed Image Retrieval (ZS-CIR)
Query
is playing with a red ball
Index images
Target
text-to-image retrieval
Textual inversion network
a photo of S_* that is playing with a red ball
CLIP text encoder

Figure 1: Overview of the SEARLE approach [1]. The proposed demo relies on the state-of-the-art SEARLE method to let users perform image retrieval with human feedback.

ABSTRACT

Composed image retrieval extends traditional content-based image retrieval (CBIR) combining a query image with additional descriptive text to express user intent and specify supplementary requests related to the visual attributes of the query image. This approach holds significant potential for e-commerce applications, such as interactive multimodal searches and chatbots. In our demo, we present an interactive composed image retrieval system based on the SEARLE approach, which tackles this task in a zero-shot manner efficiently and effectively. The demo allows users to perform image retrieval iteratively refining the results using textual feedback.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; • **Information systems** → **Image search**.

KEYWORDS

Composed Image Retrieval, Multimodal Retrieval, Multimodal Learning, Vision and Language, Textual Inversion, CLIP

ACM Reference Format:
Lorenzo Agnolucci*, Alberto Baldrati*, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-Shot Image Retrieval with Human Feedback. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3581783.3612664

1 INTRODUCTION AND RELATED WORKS

In contrast to traditional single-modality content-based and text-based image retrieval, Composed Image Retrieval (CIR) [4, 7] incorporates both visual and textual modalities to specify the intent of the user. Specifically, given a query consisting of a reference image and a relative caption, CIR aims at retrieving target images that are visually similar to the reference image but contain the modifications indicated by the relative caption. The bi-modality of the query provides the user with more fine-grained control over the properties of the target image since some properties can be more easily described in terms of language, while others can be better expressed visually.

Existing systems, such as [2], rely on supervised learning, which limits their wider use in domains other than that of their training datasets [1, 6]. On the contrary, we employ the state-of-the-art SEARLE method, which tackles Zero-Shot CIR (ZS-CIR) [1, 6], thus removing the requirement for labeled training data. This is proven to improve the generalization capabilities of the model [6]. By employing SEARLE, we allow users to upload their images without any restriction on their domain or source and perform image retrieval with user-provided textual modifications.

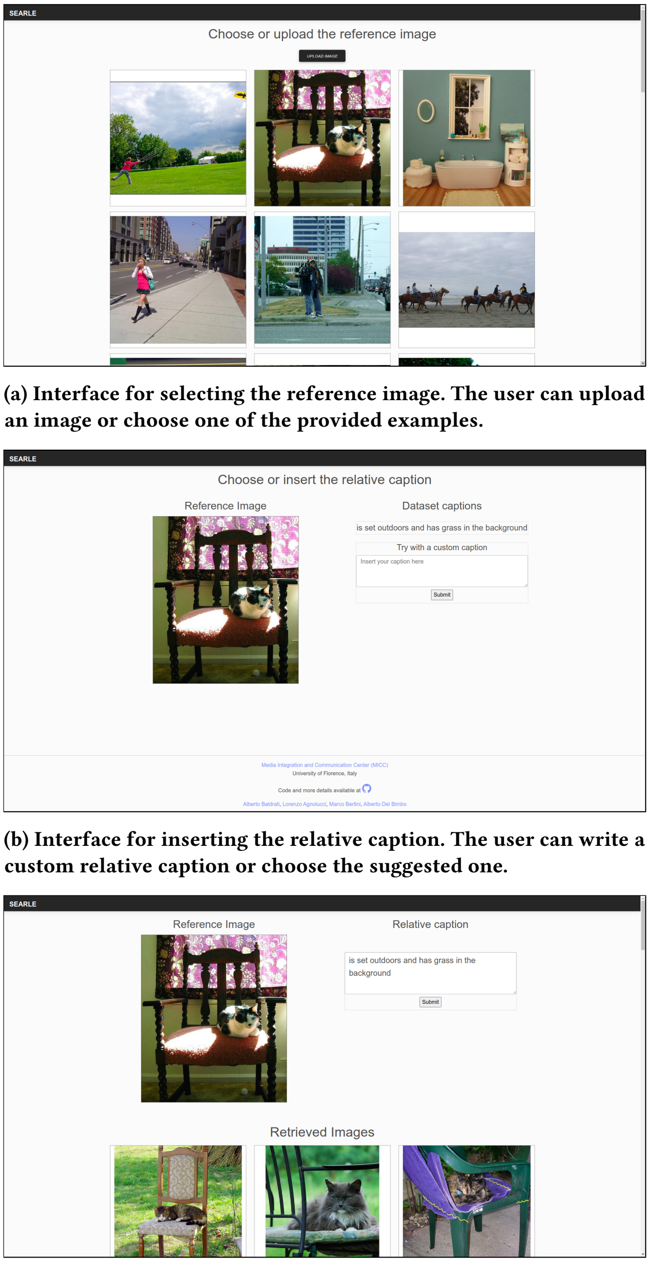
2 APPROACH

Our demo is based on the SEARLE approach [1], which relies on the CLIP [5] vision-language model. SEARLE reduces CIR to standard text-to-image retrieval by performing a *textual inversion* of the reference image, *i.e.* mapping it to a learned pseudo-word residing in CLIP token embedding space. SEARLE involves pre-training an expressive and efficient textual inversion network ϕ (consisting of an MLP) on an unlabeled image-only dataset. At inference time, given a query (I_r, T_r), ϕ predicts the pseudo-word associated with I_r, which we concatenate to T_r. Finally, we take advantage of the

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612664

[*] Equal contribution. Author ordering was determined by coin flip.


SEARLE
Choose or upload the reference image
UPLOAD IMAGE
SEARLE
Choose or insert the relative caption
Reference Image
Dataset captions
Try with a custom caption
Insert your caption here
Submit
Media Integration and Communication Center (MICC)
University of Florence, Italy
Code and more details available at
Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, Alberto Del Bimbo
SEARLE
Reference Image
Relative caption
Submit
Retrieved Images

is set outdoors and has grass in the background

is set outdoors and has grass in the background

(a) Interface for selecting the reference image. The user can upload an image or choose one of the provided examples.

(b) Interface for inserting the relative caption. The user can write a custom relative caption or choose the suggested one.

(c) Interface of the retrieval results.

Figure 2: Screenshots of our demo interface.

CLIP common embedding space to perform text-to-image retrieval. Figure 1 shows the workflow of SEARLE.

The SEARLE method is suitable for the purpose of our demonstration, as it has been shown to generalize well to images belonging to a wide range of domains, from fashion to natural images, obtaining SotA on ZS-CIR on three different datasets [1, 4, 8]. Moreover, SEARLE is really efficient, since it requires only a single forward pass of an MLP to perform textual inversion, which is carried out in a negligible time.

3 DEMO

The proposed demo allows users to leverage SEARLE to interactively perform image retrieval with user-provided textual modifications. The demo is a Flask-based web application accessible via standard web browsers on both PCs and mobile devices. We will provide an interactive live demonstration and we will also host an online server to let users independently test our system. Before running the demo, we pre-extract the visual features of the index images with the CLIP image encoder. This step removes the need of computing them for each query, thus reducing the latency and mirroring real-world scenarios such as online shops where index images represent available items. Textual features, on the other hand, are computed dynamically upon user queries, since we do not know the user-provided textual feedback in advance. Thanks to our arrangements and the efficiency of SEARLE, the demo runs on low-end CPUs with low latency times during all user interactions. We employ the 120K real-world images of COCO 2017 unlabeled set [3] as the index images. Differently from [2], which evaluates their model on the test set corresponding to the training dataset, we consider COCO even if SEARLE did not employ it for training. We use COCO as the index set for the wide content variety of its images, but we could consider any image gallery without any restriction on its domain or source thanks to SEARLE generalization capabilities.

Figure 2 shows the interface of our demo. Firstly, the user can either upload a reference image or choose one of the randomly selected examples we provide. Figure 2a illustrates the interface of the demo designated for choosing the reference image. After selecting or uploading the reference image, the user must provide textual feedback. Figure 2b shows the interface that allows the user to insert an arbitrary relative caption or to choose the suggested one if available. The suggested captions are taken from the recently introduced CIRCO dataset [1]. Finally, given the user-provided multimodal query, our demo provides the user with the retrieved images. Figure 2c shows an example of the results page of the demo. If the user is not satisfied with any of the retrieved results, they can refine their search by selecting a retrieved image as a new reference image for a subsequent query. This iterative process mimics a dialog-based search system, enabling a more natural and precise retrieval process.

4 CONCLUSION

The proposed demo system allows the user to interactively leverage the SotA SEARLE method for zero-shot composed image retrieval. The demo enables users to upload their image queries and provide user-defined textual feedback. It also facilitates a turn-based interaction that replicates the behavioral patterns of a user navigating an e-commerce website. Such an approach fosters a user-centric search experience, empowering users to effectively communicate and explore their specific search requirements. The presented system can be deployed on low-performance devices while maintaining low latency times, which makes it suitable for production environments.

ACKNOWLEDGMENTS

This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 101004545 - ReInHerit.

REFERENCES

[1] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-Shot Composed Image Retrieval with Textual Inversion. *arXiv preprint arXiv:2303.15247* (2023).

[2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21466–21474.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 740–755.

[4] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2125–2134.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of International Conference on Machine Learning (ICML)*. PMLR, 8748–8763.

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

[7] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6439–6448.

[8] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11307–11317.