Addressing Multimodal Information Loss in Composed Image Retrieval

Aymane El Firdoussi
ENS Paris-Saclay
Paris, France
January 13, 2025
aymane.elfirdoussi@telecom-paris.fr

Abstract

Composed Image Retrieval (CoIR) [13] has recently gained significant attention in computer vision research. It involves retrieving images based on a complex multi-type query comprised of a reference image $\mathcal{Q}$ and a text-based description or modification $\mathcal{T}$ of this latter. This task is inherently challenging as it requires effectively integrating both visual and textual information into a unified representation. In this paper, we investigate the efficiency of the multimodal encoding in BLIP, a vision-language model used for CoIR. We show that BLIP's current encoding does not fully capture all the necessary information from both modalities. To address this, we propose a novel approach that enhances multimodal representations by incorporating additional unimodal information. Our findings provide valuable insights into the limitations of existing methods and suggest directions for future improvements in CoIR models. Code for experiments, intially forked from CoVR repository[1], can be found here: https://github.com/elfirdoussilab1/RecVis-project .

1. Introduction

Image retrieval [5, 9] has long been a fundamental problem in computer vision due to its wide-ranging applications in query-based systems. A key challenge in image retrieval lies in formulating queries that accurately capture the user's intent. Traditional approaches, such as content-based retrieval [10], which relies purely on visual similarity, or text-based retrieval [4], which searches based on textual descriptions, often fall short in fully expressing complex user queries. To address this limitation, Composed Image Retrieval (CoIR) has emerged as a powerful paradigm that leverages multimodal queries, combining both visual and textual prompts, to specify the target image more precisely. In this setting, the reference image provides a broad contextual representation, while the accompanying text refines the query by highlighting specific modifications or attributes of interest. A fundamental challenge in CoIR is determining which aspects of the reference image should be preserved and which should be disregarded, typically distinguishing the primary object of interest from background elements or irrelevant details. This inherent ambiguity makes CoIR a particularly compelling research problem, bridging the gap between vision and language understanding.

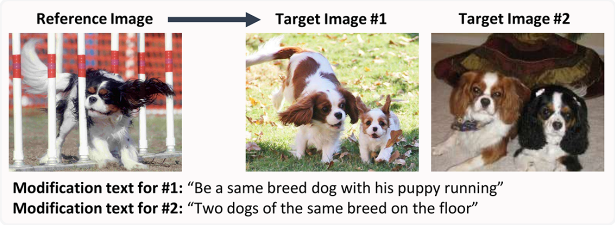
Reference Image
Target Image #1
Target Image #2
Modification text for #1: "Be a same breed dog with his puppy running"
Modification text for #2: "Two dogs of the same breed on the floor"

Figure 1. A sample from the CIRR dataset: The input consists of a reference image and a modification text. The goal is to retrieve the correct target image that best matches the specified transformation.

2. Training BLIP for CoIR

Numerous models have been employed for Composed Image Retrieval (CoIR), including BLIP [2] and its enhanced version BLIP-2 [3], OpenAI's CLIP [8], and CIRPLANT [6], among others. In parallel, several benchmark datasets have been developed to facilitate research in this area, such as CIRR [6], FashionIQ [14], and the recently introduced WebVid-CoVR datasets [11, 12].
In this work, we primarily focus on BLIP-based models and assess their performance on the CIRR dataset.

2.1. Model description

BLIP-CoIR. Our CoIR model architecture builds on BLIP (described in Figure 2), a pre-trained image-text model. Since BLIP is not intentionally trained for composed visual retrieval, we therefore adapt it to our task as

[1] https://github.com/lucas-ventura/CoVR

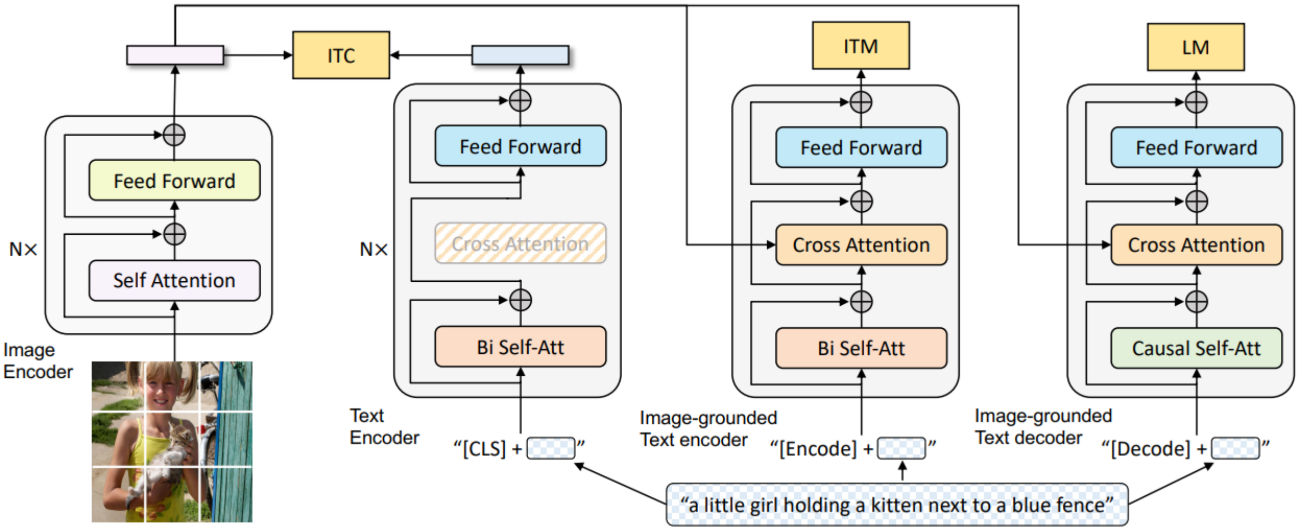

Figure 2. BLIP Architecture as of [2]: it consists of a unified vision-language model which can operate in one of the three functionalities: (1) **Unimodal encoder** (image or text) is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) **Image-grounded text encoder** uses additional cross-attention layers to model vision-language interactions, and is trained with a image-text matching (ITM). (3) **Image-grounded text decoder** replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder.

follows:
We use the BLIP **image encoder**, which is a Vision Transformer [1], to encode the query image $\mathcal{Q}$ into visual features q of dimension $N \times d$, where N is the number of feature vectors produced by this block. These latter, together with the **encoded** modification text t are then forwarded to the BLIP **image-grounded text encoder** which outputs a multi-modal embedding $f(q,t) \in \mathbb{R}^d$. Finally, the retrieved image $\mathcal{Z}$ is the one that maximizes the cosine similarity between the multimodal embedding of the query $f(q,t)$ and the embedding of a target image $z \in \mathbb{R}^d$, i.e.:

$$\arg\max_{z} \langle z, f(q,t) \rangle \tag{1}$$

where $\langle , \rangle$ represents the normalized dot-product in $\mathbb{R}^d$, and is defined as: for all $a, b \in \mathbb{R}^d$:

$$\langle a, b \rangle = \frac{a^\top b}{||a|| ||b||}$$

2.2. Training

CIRR dataset. The Compose Image Retrieval on Real-life images (CIRR) includes over $36,000$ annotated query (image and text) and target triplets $(\mathcal{Q}, \mathcal{T}, \mathcal{Z})$, where 80% of this data is used for training, 10% for validation and 10% for test. The modification text of each triplet has been collected using Amazon Mechanical Turk (AMT). An example of a sample from CIRR is shown in Figure 1. Table 1 summarizes the number of triplets and images forming each set.

Set	Nb. of triplets	Nb. of images
`Train`	28,225	16,939
`Validation`	4,184	2,297
`Test`	4,184	2,316

Table 1. Statistics of the CIRR dataset [6], showing the number of triplets and unique images in the training, validation, and test sets.

Loss function. We optimize our BLIP-based models by minimizing the HN-NCE loss [7], which increases the weight of most similar samples and uses as negatives all target images z_i in the batch $\mathcal{B}$. More precisely, given a training batch $\mathcal{B}$ of triplets (q_i, t_i, z_i), we define the quantity $S_{i,j} = \langle f_i, z_j \rangle$ being the cosine-similarity between the multimodal embedding f_i and the target image z_j, and we aim to minimize the following loss function:

$$\begin{aligned}\mathcal{L}(\mathcal{B}) = & -\sum_{i \in \mathcal{B}} \log \left(\frac{e^{S_{i,i}/\tau}}{\alpha s^{S_{i,i}/\tau} + \sum_{i \neq j} e^{S_{i,j}/\tau} w_{i,j}} \right) \\ & -\sum_{i \in \mathcal{B}} \log \left(\frac{e^{S_{i,i}/\tau}}{\alpha s^{S_{i,i}/\tau} + \sum_{i \neq j} e^{S_{j,i}/\tau} w_{j,i}} \right)\end{aligned}$$

where $\alpha = 1$, the temperature $\tau = 0.07$ and the weights $w_{i,j}$ are set as in [7] with $\beta = 0.5$.

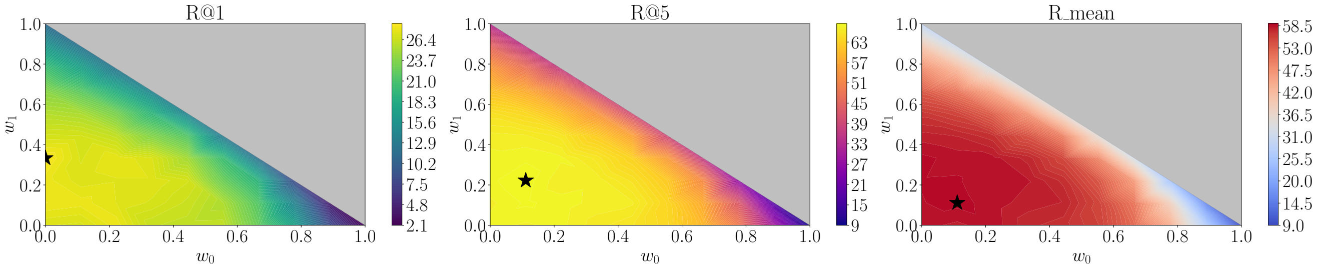

Figure 3. Heatmaps showing BLIP-Large recall performance on CIRR after applying the Mixed embedding (Equation (2)). We observe that The optimal recall is achieved at non-zero values of (w_0, w_1), which demonstrates the sub-optimality of BLIP's multimodal encoding block. The star (*) in each plot marks the optimal weight combination (w_0, w_1).

Implementation details. We train two variants of our **BLIP-CoIR** model:

- **Base model**: which has an image encoder (Vision Transformer) of 12 transformer layers, each with a hidden size of 768 and 12 attention heads. We use the Capfilt checkpoint as initial state.
- **Large model**: that features a more complex architecture, with 24 transformer layers, a larger hidden size of 1024, and 16 attention heads. We initially use a checkpoint of this model finetuned on COCO.

The Large model has approximately more than the double of parameters of the base model, making it significantly larger and more capable of capturing complex relationships between images and text, but it comes with the cost of higher compute resources for both training and inference. For more computational efficiency, we freeze the BLIP (both Large and Base) image encoder (ViT) during training. Experiments are conducted on 4 NVIDIA P4 8-GB GPUs, and all the details concerning the deployed hyperparameters are summarized in Table 5 in the Appendix. Figure 5 shows the training loss evolution of both models.

2.3. Results

Evaluation metrics. We use the standard evaluation metrics [6] which consists on reporting the recall at ranks 1, 5, 10 and 50. The recall at rank k (denoted R@k) quantifies the number of times (percentage) the correct image is present among the top k predictions of the model.

Model	R@1	R@5	R@10	R@50
`BLIP-L [11]`	49.16	79.76	88.65	97.49
`BLIP-L⋆ (ours)`	27.03	67.42	80.08	95.07
`BLIP-B⋆ (ours)`	21.84	59.02	73.60	93.23

Table 2. Retrieval performance (R@K) comparison between different BLIP models. Our results (marked with (*)) show a performance gap compared to the baseline from the CoVR paper [11], mainly due to differences in batch size during training.

Analysis of the results. We remark that we get lower performance results of our trained BLIP-CoIR models compared to the baseline listed in the reference paper [11], which is mainly due to the decrease in the batch-size, as we only use 16 samples in each iteration, compared to 2048 in [11], to update the weights of the models during training. Also, as expected, BLIP-Large (initially finetuned on COCO) outperforms BLIP-Base on CIRR, showing its higher capacity to capture more interesting multimodal features in the data. Therefore, we use the Large BLIP model for all the upcoming experiments.

3. On the impact of the unimodal embeddings

As stated in equation (1), the predictions (retrievals) made by our BLIP models relie solely on the multimodal embedding $f(q,t)$, as it is directly compared to the target images embeddings. However, we are not guaranteed that this multimodal encoding might not miss some important information present in each mode q and t. So how can that be verified in practice ?

3.1. Fixing by Mixing

We want to assess whether the model's multimodal encoding $f(q,t)$ captures almost all the necessary information present in the user's query (image and modification text). This is equivalent to showing that this embedding is optimal in terms of the performance recalls. And to validate this assumption, we propose to evaluate the model on a convex mixture of the three modal embeddings: q, t and $f(q,t)$, and we define a new **"mixed"** embedding as follows:

$$m(q,t) = w_0 q + w_1 t + w_2 f(q,t), \quad \sum_{i=0}^{2} w_i = 1 \qquad (2)$$

The idea behind this new multimodal embedding $m(q,t)$ is to include (linearly) some additional information from q and t, and see whether there exists some couple $(w_0, w_1) \neq (0,0)$, such that the performance of BLIP-CoIR is better using this new embedding. We call the resulting model that uses $m(q,t)$ instead of $f(q,t)$ the **Mixed** model.

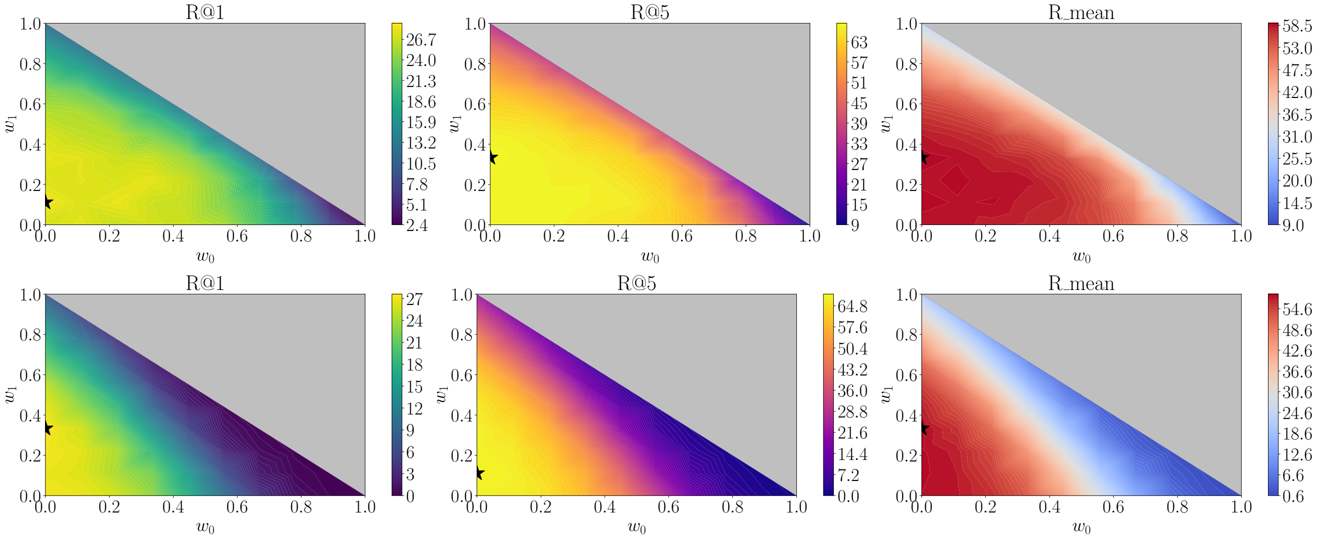

Figure 4. Comparison of aggregation strategies for Mixed embedding: (Top) Median aggregation, (Bottom) Max-pooling. The median approach yields superior recall performance across all ranks.

Model	R@1	R@5	Avg
`baseline`	27.03	67.42	58.18
`Mixed`	**28.32**	**68.14**	**58.69**

Table 3. Effect of incorporating Mixed embedding on BLIP retrieval performance. The slight improvements in recalls confirm that additional unimodal information is needed.

Table 3 and Figure 3 summarize the results gotten with this new strategy. We clearly observe from that BLIP's original multimodal embedding $f(q,t)$ is sub-optimal, as the optimal recall value is not achieved for $(w_0, w_1) = (0,0)$, meaning that indeed BLIP's **image-grounded text encoder** does not capture the entire necessary information provided by the query, but is still close to optimality.

Additionally, we observe that the optimal weights $(w_i)_{i=0}^2$ vary across different recall levels, which is an interesting and counter-intuitive result as one would initially expect them to be the same.

3.2. The impact of the aggregation rule:

As we have stated earlier in section 2.1, the Vision Transformer forming BLIP's image encoder encodes the query image into N visual features of dimension d. Additionally, the text encoder (Bert model) also gives a sequence of d-dimensional vectors (encoding of each token in the text). Thus, to compute the **mixed** embedding $m(q,t)$ (2), we need to aggregate the features of each unimodal encoding (q and t) into one d-dimensional vector. Therefore, the performance of our model using $m(q,t)$ will depend on the type of aggregation rule we're deploying.

Model	**Aggregation**	R@1	R@5	Avg
`baseline`	-	27.03	67.42	58.18
`Mixed`	`Mean`	28.32	68.14	58.69
	`Median`	**28.53**	**68.26**	**58.81**
	`Max-pooling`	27.34	68.07	58.44

Table 4. Comparison of the performance of Mixed BLIP model using different aggregation rules. Optimal performance accross all reported recalls were achieved by aggregating the vectors of q and t using the Median.

We then evaluate our approach using three different aggregation rules: Mean (Figure 3), Median (Figure 4 (top)) and Max-pooling (Figure 4 (bottom)). Table 4 summarizes our findings.

We observe that the Median aggregation rule consistently achieves the best performance across all recall ranks, while Max-pooling proves to be the least effective, which can be explained by the fact that the Median is **robust** to outliers compared to the other rules.

Furthermore, and interestingly, the optimal recalls across the three methods are obtained for $w_0 = 0$ (the weight assigned to q) in most cases. This essentially means that the multimodal embedding $f(q,t)$ already captures all the important information present in the query image, but fails to fully encode the details conveyed by the modification text!

4. Conclusion & Future work

In this work, we investigated the efficiency of BLIP's multimodal encoding for Composed Image Retrieval (CoIR) and explored its limitations. Our findings reveal that BLIP's current multimodal representation does not fully capture all the necessary information from the input modalities, particularly from the textual modification. Through our proposed mixed embedding approach, we demonstrated that incorporating additional unimodal information improves retrieval performance, which highlights the sub-optimality of BLIP's multimodal encoding block. However, as we only conducted our experiments on a single random seed, this result is still inconsistent, and need further statistical guarantees (running on multiple seeds and reporting the standard deviation of the recalls).

Future directions could involve exploring more sophisticated mixing strategies beyond the linear convex approach, developing advanced techniques for encoding multi-type queries, as well as evaluating the proposed methods across diverse datasets beyond CIRR.

References

[1] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 2

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[4] Wen Li, Lixin Duan, Dong Xu, and Ivor Wai-Hung Tsang. Text-based image retrieval using progressive multi-instance learning. In *2011 international conference on computer vision*, pages 2049–2055. IEEE, 2011. 1

[5] Bin Liu, Yue Cao, Mingsheng Long, Jianmin Wang, and Jingdong Wang. Deep triplet quantization. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 755–763, 2018. 1

[6] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 1, 2, 3

[7] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6967–6977, 2023. 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[9] Rishab Sharma and Anirudha Vishvakarma. Retrieving similar e-commerce images using deep learning. *arXiv preprint arXiv:1901.03546*, 2019. 1

[10] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, 2001. 1

[11] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5270–5279, 2024. 1, 3

[12] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[13] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 1

[14] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 1

Addressing Multimodal Information Loss in Composed Image Retrieval

Supplementary Material

Training loss evolution:

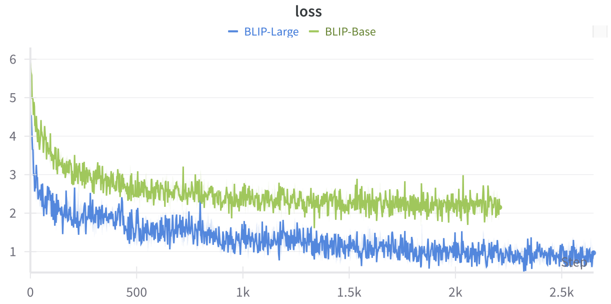

Figure 5. Training loss curve of BLIP-Base and BLIP-Large.

Training Hyperparameters

Parameter	**Value**
BLIP-Base	
`optimizer`	AdamW
`batch-size`	16
`learning rate`	$9.5 \times 10^{-7} < \text{lr} < 10^{-5}$
`learning rate scheduler`	StepLR
`epochs`	5
`weight decay`	0.05
`precision`	bf16
`seed`	1234
BLIP-Large	
`optimizer`	AdamW
`batch-size`	16
`learning rate`	$7 \times 10^{-6} < \text{lr} < 10^{-4}$
`learning rate scheduler`	StepLR
`epochs`	6
`weight decay`	0.05
`precision`	bf16
`seed`	1234

Table 5. Training hyperparameters of experiments reported in Table 2.