Effective conditioned and composed image retrieval combining CLIP-based features

Alberto Baldrati[1,2] Marco Bertini[1] Tiberio Uricchio[1]
Alberto Del Bimbo[1]

[1] Università degli Studi di Firenze - MICC
[2] Università di Pisa
Firenze, Italy - Pisa, Italy
[name.surname]@unifi.it

Abstract

Conditioned and composed image retrieval extend CBIR systems by combining a query image with an additional text that expresses the intent of the user, describing additional requests w.r.t. the visual content of the query image. This type of search is interesting for e-commerce applications, e.g. to develop interactive multimodal searches and chatbots. In this demo, we present an interactive system based on a combiner network, trained using contrastive learning, that combines visual and textual features obtained from the OpenAI CLIP network to address conditioned CBIR. The system can be used to improve e-shop search engines. For example, considering the fashion domain it lets users search for dresses, shirts and toptees using a candidate start image and expressing some visual differences w.r.t. its visual content, e.g. asking to change color, pattern or shape. The proposed network obtains state-of-the-art performance on the FashionIQ dataset and on the more recent CIRR dataset, showing its applicability to the fashion domain for conditioned retrieval, and to more generic content considering the more general task of composed image retrieval.

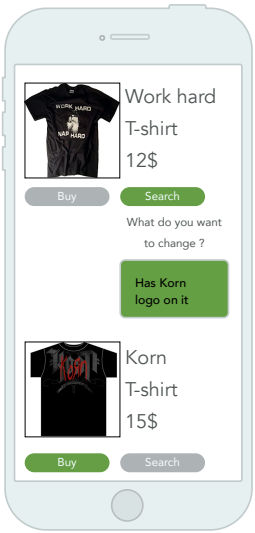

Figure 1. Example of use of conditioned image in the fashion domain for e-commerce application. The user can refine product search providing details and constraints in natural language. The system uses both visual and textual features to retrieve the desired result.

1. Introduction

Content-Based Image Retrieval (CBIR) is a basic task in computer vision and multimedia research and can be applied to general web images, as in Google Reverse Image Search, or it can be specialized to a large number of domains like landmarks [18, 37], medical images [41], cultural heritage [12,31] and e-commerce, either for general e-shopping [32,44,45] or in specific e-commerce domains like fashion [13,22,23] or interior design [36]. These CBIR systems retrieve images from a database using an input image, computing a distance between the visual features extracted from the query and the features stored in the database. Features must be discriminative enough to deal with different images and must be robust to a number of transformations to also retrieve variations of the same images. A main difficulty is to overcome the proverbial semantic gap between the low-level visual features used and the high-level meaning of the images [34].

Several variations of the basic CBIR task have been proposed to narrow this gap, requesting that the user provides some additional information regarding the intent or context

of the query. Relevance feedback is one of such mechanisms, where users refine iteratively the search results providing additional information on what is "similar" or "dissimilar" according to them [28]. More recently, CBIR systems have been extended by adding context obtained through natural language processing, where users describe what conditions must be met by the desired results in addition to the visual features of the query image. This defines the task of conditioned image retrieval, proposed to implement interactive search systems for fashion [15, 40]. But it can be effectively used in many different domains of online retail, where the retrieval of relevant products could be based on the type of product, its texture or color, shape, material or brand [30]. Composed image retrieval, instead, generalizes the approach composing the query as an image-language pair, using both visual and textual modalities to specify the user's intent [27].

In this work, we address both conditioned retrieval applied to the fashion domain and composed retrieval applied to general images. The proposed system is based on a network that combines visual and textual features derived from the OpenAI CLIP network. Despite the simplicity of the network design, the system achieves state-of-the-art results on two commonly used standard datasets, FashionIQ [40] for the fashion domain, and CIRR [27] for more general content. The system can be used to develop interactive e-commerce sites and chatbots, or to improve the performance of image search engines.

2. Related works

Several surveys provide an overview of CBIR approaches and their evolution in the past years. Zheng *et al*. [46] surveyed image search approaches from 2006 to 2016, going from methods based on Scale-Invariant Feature Transform (SIFT) to those based on Convolutional Neural Networks (CNNs). Zhou *et al*. [47] surveyed CBIR researches from 2003 to 2016, including methods based on engineered and learned features. Li *et al*. [26] reviewed both technological developments and practical applications of CBIR from 2009 to 2019. Dubey [9] has recently provided a survey on CBIR methods based on deep learning of the past decade.

Visual and language pretraining

CLIP [29] has very recently obtained remarkable results in multi-modal zero shot learning, showing feature generalization of both images and text. The approach followed by CLIP learns associations between the abundant images and natural language supervision available on the web (using 400 millions of image-text pairs for training). Despite not being directly optimized for a specific benchmark, it performs consistently well on different tasks. Although CLIP effectiveness is still subject of study [1], it has already been successfully applied to different tasks like fine-grained art classification [7], image generation [11], zero shot video retrieval [10], event classification [25] and visual commonsense reasoning [39]. This work builds upon CLIP, exploiting its potential for conditioned image retrieval. Other approaches to learn image-text alignment have been proposed in [6, 17]. ALIGN [17] uses a dual-encoder architecture and is trained on a huge dataset of 1 billion image-text pairs. Differently, the method proposed in [6] exploits contrastive distillation, resulting in a much more data-efficient process, requiring a training dataset that is $133\times$ smaller than that of CLIP.

Conditioned and combined image retrieval

This work is related to the recently introduced problem of conditioned fashion image retrieval [40], and with the very recent problem of composed image retrieval of generic images [27].

Many works have addressed the first task. In [5], a transformer that can be seamlessly plugged in a CNN to selectively preserve and transform the visual features conditioned on language semantics is presented. In [38] it has been proposed a method called Text Image Residual Gating (TIRG) that combines image and text features using gating and residual features. In [33] the authors combine graph neural networks and skip connections. In [24], they use two different neural network modules, one to deal with image style and one for image content. In [20] a Correction Network is proposed to model explicitly the difference between the reference and target image in the embedding space. In [8] is proposed a model called Modality-Agnostic Attention Fusion (MAAF), designed for composed image retrieval, treating the convolutional spatial image features and learned text embeddings as modality-agnostic tokens, that are then passed to a Transformer. An autoencoder-based model, called ComposeAE, has been proposed in [2], to learn the composition of image and text features for retrieval using a deep metric learning (DML) approach. In [42] has been proposed to measure the semantic differential relationships between images with respect to a conditioning query text using a method called CurlingNet. The main components are two networks: the so-called Delivery filter, delivers the source image to the candidate cluster according to a given query in an embedding space, while the Sweeping filter checks the attributes highlighted in the query and learns the path from the center of valid target candidates to the true target image. Conditional image retrieval has been recently extended to a multi-turn conversation in [43]. The proposed system uses ComposeAE [2] for combining image and text at each turn, feeding it into a recurrent network according to the turn order. Finally, text-conditioned image retrieval has been addressed in [15], where the authors present the SAC (Semantic Attention Composition) frame-

work that operates in two steps: firstly, the Semantic Feature Attention (SFA) module finds the salient regions of the image w.r.t. the text and then the Semantic Feature Modification (SFM) module determines how to change the relevant parts of the image compositing coarse and fine salient image features computed by SFA with text embeddings.

Regarding the second task of composed image retrieval, a new dataset called CIRR has been introduced in [27], containing generic real-world images. The authors have also proposed a baseline method, a novel model called CIRPLANT, based on transformers, that uses rich pre-trained vision-and-language knowledge to modify visual features conditioned on natural language. CIRPLANT has been tested also on the FashionIQ dataset, obtaining good results.

Differently from these previous works our method explicitly considers a learned manifold of visual and text features with the goal of learning an additive transformation in the same space, and it does not use any kind of spatial information.

3. The proposed method

The proposed method tackles the problem of conditional and composed image retrieval, i.e. the query is composed of an image and an additional textual information that expresses a request from the user with respect to the image. The goal is to find the best matching images satisfying both the similarity constraints of the reference image and the changes to the image requested in the additional text. To this end, the system must be able to understand both the contents of the image and text, and combine the textual comment to the image content.

A schema of the system training is shown in Figure 2. In contrast to previous works like [5, 20, 24, 33] that build from different image and textual model, we start from the hypothesis of having a common embedding of images and text, obtained using CLIP features. This is motivated by the fact noted in [29], that similar concepts expressed in text and images tend to share similar features, or at least be "near" in the common space.

Both image and text inputs are encoded using their respective CLIP encoders into features in the common space. The problem to be solved is that of learning a transformation from the reference image feature and input text to a combined feature that includes both the multi-modal input information and is as near as possible to the target image in the common manifold. We denote this transformation as a *Combiner* function and design a neural network architecture trained to learn the correct function.

The Combiner function, depicted in Figure 3, is simple yet more performing than more complex architectures that we tested, obtaining a new state of the art performance in conditioned and combined image retrieval; more details and ablation studies on the design of the network are available in our previous work [3]. The idea is to build an additive transformation where text, image and the combination of both are all added into the final combined feature. The training of the system is performed with triplets of: input images, relative captions and target images. Following [24, 33, 38] we employ the batch-based classification (BBC) loss.

3.1. Preprocess Pipeline

The standard preprocess pipeline of CLIP is mainly composed of two steps: a resize operation where the smaller side of the image matches the CLIP input dimension $input\_dim$ followed by center crop operation which results in a square patch $input\_dim \times input\_dim$ output. Subsequently, as the ratio between the bigger side and the smaller side increases, the area of the image lost after the preprocess increases.

To overcome such loss of information the simplest approach is to perform a zero-padding to match the smaller side to the bigger side (i.e. squaring the image). By doing this we zero out the loss of content information attributable to the center crop operation, however we lower the resolution of the useful portion of the image since the CLIP image encoder input dimension is fixed. Therefore, differently from our previous work [3], we propose a new preprocess pipeline which aims to find a compromise between the aforementioned pipelines: before applying the center crop operation we pad an image only if its aspect ratio is above a fixed target ratio. Moreover when we pad an image we do not make it square, instead we bring its aspect ratio to the target ratio. This approach has improved the performance with respect to our previous results [3].

3.2. Implementation Details

In the following experiments and in the demo we use the CLIP model denoted as RN50x4, since it outperforms the RN50 model: the visual encoder follows the EfficientNet-style model scaling and uses approximately $4\times$ the computation of a standard ResNet-50 [14]. It takes as input images of 288×288 pixels and outputs features of 640 dimensions. The text encoder is a transformer encoder with 12 layers, 10 heads and a width of 640.

In the experiments, the CLIP encoders have been kept frozen and the only trained part of the model is the Combiner function. The target ratio in the preprocess pipeline was set to 1.25. We used PyTorch in our experiments. We used Adam optimizer [21] with a learning rate set to $2e-5$. We trained the model for a maximum of 300 epochs and the batch size was set to 4096.

4. The proposed demo

The proposed demo aims to show in an interactive way how the multi-modal retrieval system described previously works. Such a demo has a twofold objective: the first one is to dynamically illustrate how the system works when

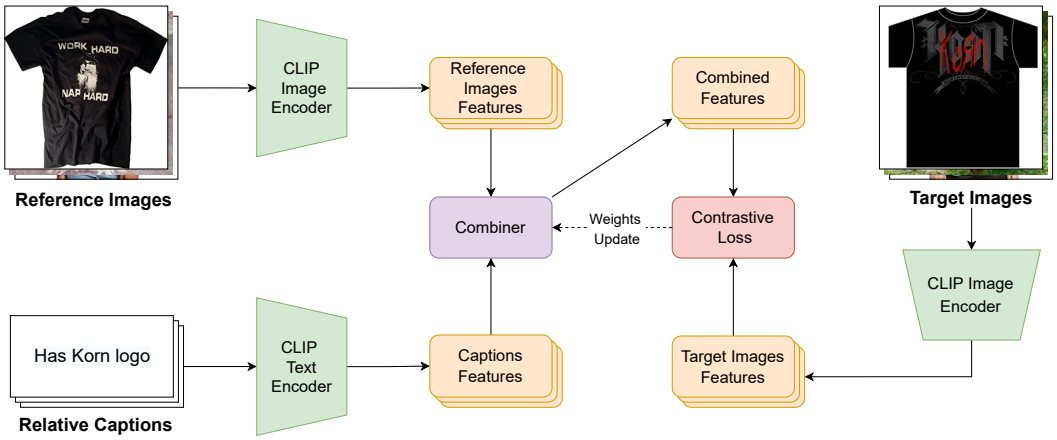

Figure 2. Training overview of the system, from the input image and captions on the left, to the target image on the right. At inference time the trained Combiner is used to produce an effective multi-modal representation used to query the database.

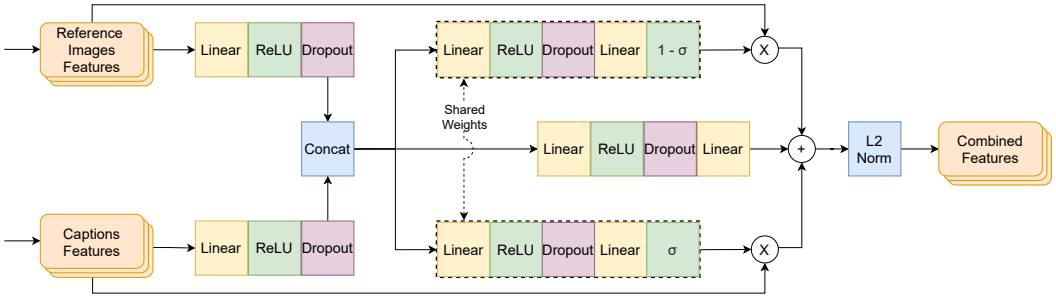

Figure 3. Architecture of the combiner network. σ represents the sigmoid function.

we use as query a pair of (reference image, relative caption) which is included in the datasets. The second objective is to simulate a real-world scenario where the user can query the system with arbitrary captions not included in the datasets. The interface of our demo is capable of handling both objectives simultaneously: it is able to suggest the relative captions associated to each reference image marking also the ground-truth target image in the results and it provides a text area where the user inputs an arbitrary caption. In the demo are included both datasets we experimented with: FashionIQ and CIRR. The demo is available at http://cir.micc.unifi.it:5000

Figure 4 shows a diagram of how the application works.

4.1. Architecture

The demo is developed as a web-app accessible through a standard web-browser, either on PC or mobile devices. Before starting the demo it is necessary to extract all the visual features of the images using the CLIP image encoder. This computation is performed off-line to avoid recalculation for every query. From a real-world perspective this precomputation makes sense, in fact, if we think for instance of an online shop, the images are not dynamically uploaded by the users but they represent the items that the shop can sell. On the other hand, the textual features are computed on-the-fly when a query is performed since in a real context the queries of the users are not known a priori. After the visual feature extraction the demo is ready to run.

The demo allows the user to choose first the dataset, then the reference image and finally to insert the caption (or choose between the default ones of the dataset). When the user selects a reference image and fills in (or selects) a relative caption, firstly the corresponding visual features are selected from the pre-computed visual features. The textual features are then extracted using the CLIP text encoder and subsequently the visual and text features are combined using the *Combiner* network, which outputs the combined features. Finally, as in standard image retrieval, the combined features are used to query the database of visual features. It is very important to notice that, once the combined features are computed, the conditioned image retrieval is totally analogous to a standard content-based image retrieval. Therefore all the techniques that are commonly used to ensure scalability of CBIR systems can be applied to the proposed system, such as hashing, approximate search, e.g. using the FAISS [19], etc. In the demo, the top 50 results

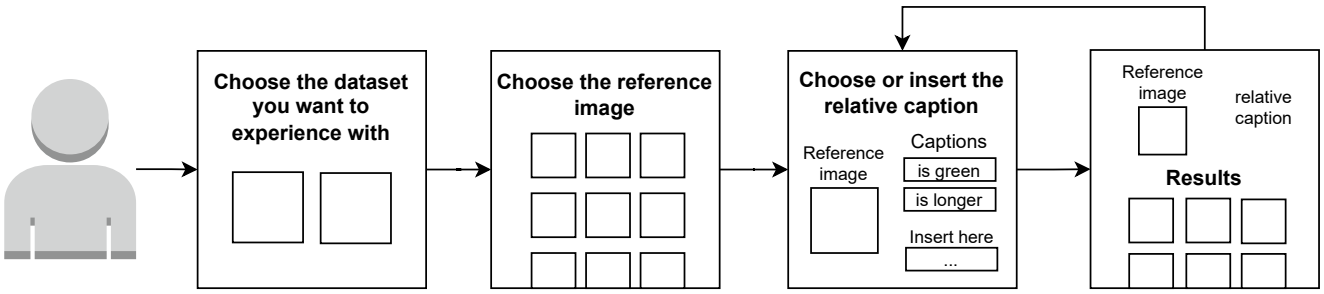

Figure 4. Demo overview. Firstly the user has to choose the dataset, there are two possible choices: fashion dataset FashionIQ and real-life images dataset CIRR. After choosing the reference image the user can insert a relative caption or select among the default ones of the dataset. Finally they can check out the results. If the user is not satisfy by the results, by clicking on a retrieved image, they can use such image as a reference image in a new query

are shown since in both datasets the broader scale metrics is R@50. Moreover, in the CIRR dataset, when a dataset caption is selected, also the subset results are displayed. Since we have two different datasets with completely different image domains we have two different Combiner networks, one for FashionIQ and the other for CIRR dataset. The right combiner network is automatically selected when choosing the dataset used in the demo.

4.2. Implementation details

The backend of the web-app is a small server written in Python with the Flask micro-framework. The frontend is written with the Bootstrap library and can be used on PCs and mobile devices. To reduce the amount of GPU memory the pre-computed features are stored in CPU RAM and loaded in the GPU only when they are needed. To further reduce the amount of required memory (and to speedup computations) both the Combiner networks and CLIP model work in half (fp16) precision. To remain consistent with standard evaluation protocol of FashionIQ we consider the dataset subdivided in three categories (Dress, Toptee and Shirt), this implies that when the reference image belongs to a category, when retrieval is carried out, only the images of the same category are taken into consideration. This is a reasonable design choice also in a real-world deployment since we can expect that a user interested in a dress does not want to look at shirts.

The suggested captions are only those included in the validation set and, when one of them is selected, the retrieved images are those of the validation set. This choice was made so that the demo could highlight the ground truth target image in the retrieved results. In fact, in both datasets, the ground truth labels are not released for the test set. On the contrary, when the user inserts a new query that is not part of the dataset, as they would in a real-world scenario, the system searches for relevant images both in the validation and test set.

We deploy our demo on a machine with an Intel Xeon E5-2620 v3 CPU, a NVIDIA Titan X 12GB GPU and 128GB of RAM. The retrieval process takes on average less than 35ms with a GPU RAM occupation of 743 MB (with a single simultaneous access). We have tested the demo also on a low-end laptop with Intel Core i7-7500U CPU, and a NVIDIA GeForce 940MX 2GB GPU and 16GB memory; also in this case the demo runs smoothly with an average retrieval time of 70ms. Obviously the number of images involved in the retrieval is relatively small (more details in Section 5), however the fact that the Combiner network is able to run almost in real-time on such a low-end device makes us believe that the system can be scaled to large-scale retrieval.

4.3. Usage and Examples

Firstly, when the demo is booted up, it is necessary to choose the dataset on which to perform the experiments. As mentioned earlier the user can make two choices: the fashion dataset FashionIQ and the real-life images dataset CIRR. Using the navigation bar it is always possible to change the choice of the dataset during the execution of the whole demo. In Figure 5 the dataset choice page is shown.

Once the dataset is chosen, the user must choose the reference image he desires. Some reference images are randomly selected from the dataset, as a suggestion to the user. Refreshing the page shows a different set each time. Figure 6 shows the interface of the demo that allows such a choice.

To complete the multi-modal query, a relative caption must also be provided. The demo allows the user both to choose among the captions included in the validation set and to insert an arbitrary caption. Figure 7 shows how the demo interface allows these two options.

Finally the user can check the results of the multi-modal query he has inserted. Furthermore, if the user wants to refine the results, a retrieved image can be used as a reference

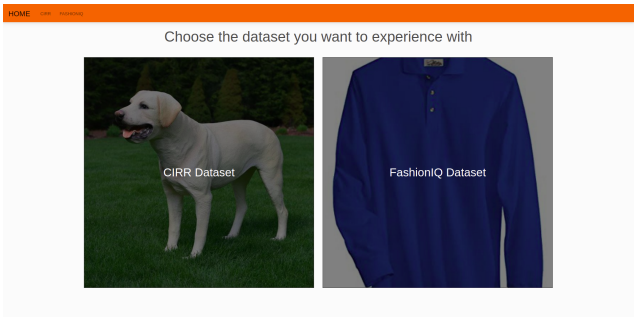

Figure 5. Dataset choice demo page. The user can either select FashionIQ or CIRR dataset.

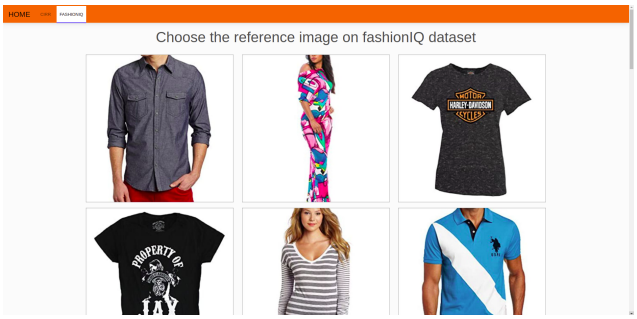

Figure 6. Reference image choice demo page. The user can select the reference image they prefer.

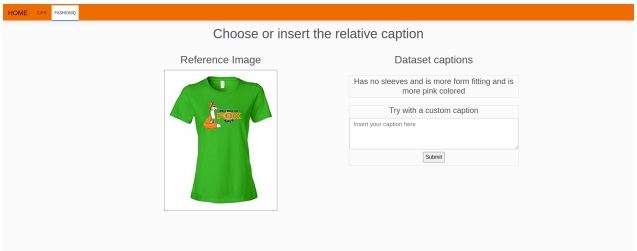

Figure 7. Relative caption insertion demo page. The user can either select or insert a relative caption.

image in a new query. This can be done by clicking on the retrieved image that the user wants to use in a new query. Such an iterative process allows a multi-step search by simulating a dialog-based search system which is more natural to use and allows the user to precisely describe what they want to search for. Figure 8 shows the demo results page.

A video showing a full example of use of the system is available at https://youtu.be/ifBQA9xAbhw.

5. Experimental results

In this section we report a comparison of the performance of the proposed system with competing state-of-the-art approaches on two standard datasets, FashionIQ and CIRR. These datasets are used also in the demo.

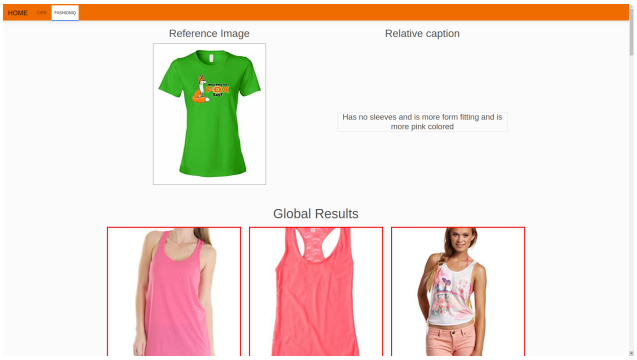

Figure 8. Results demo page. The user can check the results of the multimodal query he has inserted. Furthermore, by clicking on a retrieved image they can use it as reference image in a new query

5.1. FashionIQ

FashionIQ [40] provides 77,684 fashion images crawled from the web divided into three different categories: *Dress*, *Toptee* and *Shirt*; the authors provide standard train, validation and test splits. The training set comprises 18,000 training triplets made of a reference image, a pair of relative captions and a target image, and in total is composed of 46,609 images. The captions describe properties to modify in the reference image to match the target image. The validation set has 15,537 images and 6,017 triplets, and the test set is composed of 15,538 images and 6,119 triplets.

We follow the standard experimental setting as in [20, 24]. The evaluation metric used is the average recall at rank K (Recall@K), in particular we use Recall@10 (R@10) and Recall@50 (R@50). Note that for each triplet there is only a positive index image. Hence, each individual query has R@K either zero or one. All results reported have been computed on the validation set since at the time of writing the test set ground-truth labels have not been released.

5.2. CIRR

The CIRR (Compose Image Retrieval on Real-life images) [27] dataset is thought for overcoming two common issues that occur in conditioned image retrieval datasets (such as FashionIQ): the lack of a sufficient visual complexity due to the restricted image domain and the existence of many false-negatives since the target images cannot be extensively labeled for each (reference, text) pair. CIRR is made of 21,552 real-life images taken from the popular natural language reasoning dataset $NLVR^2$ [35]. It follows the same structure of the FashionIQ dataset and contains 36,554 triplets randomly assigned in 80% for training, 10% for validation and 10% for test. The images of the dataset are grouped in multiple subsets of six images that are semantically and visually similar. The relative captions are collected to describe differences between two images in the

same subset. This is done in order to have negative images with high visual similarity, otherwise it would be trivial to discriminate between the reference and target images.

Following previous works, the standard evaluation protocol proposed by the authors of the dataset is to report the recall at rank K (Recall@K) at four different ranks (1, 5, 10, 50). Moreover, thanks to the unique design of the CIRR dataset, it is also reported the $\text{Recall}_{\text{Subset}}$ which considers only the images in the subset of the query. This *subset* metric has two main benefits: it is not affected by false-negative samples and, thanks to negative samples with high visual similarity, it captures fine-grained image-text modifications. Of these metrics, R@5 accounts for possible false-negatives in the entire corpus, and R_{Subset}@1 illustrates the fine-grained capabilities.

5.3. Comparison with SotA

In these experiments we compare the proposed method with state-of-the-art approaches on two standard and challenging datasets.

	Shirt		Dress		Toptee		Average	
Method	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$
JVSM [4]	12.0	27.1	10.7	25.9	13.0	26.9	11.9	26.6
CIRPLANT w/OSCAR [27]	17.53	38.81	17.45	40.41	21.64	45.38	18.87	41.53
TRACE w/BERT [16]	20.80	40.80	22.70	44.91	24.22	49.80	22.57	46.19
VAL w/GloVe [5]	22.38	44.15	22.53	44.00	27.53	51.68	24.15	46.61
MAAF [8]	21.3	44.2	23.8	48.6	27.9	53.6	24.3	48.8
CurlingNet [42]	21.45	44.56	26.15	53.24	30.12	55.23	25.90	51.01
RTIC-GCN w/GloVe [33]	23.79	47.25	29.15	54.04	31.61	57.98	28.18	53.09
CoSMo [24]	24.90	49.18	25.64	50.30	29.21	57.46	26.58	52.31
DCNet [20]	23.95	47.30	28.95	56.07	30.44	58.29	27.78	53.89
SAC w/BERT [15]	28.02	51.86	26.52	51.01	32.70	61.23	29.08	54.70
Proposed approach	**36.36**	**58.00**	**31.63**	**56.67**	**38.19**	**62.42**	**35.39**	**59.03**

Table 1. Comparison between our method and current state-of-the-art models on the Fashion-IQ validation set. Best scores are highlighted in bold, second best scores underlined.

	Recall@K				R_{subset}@K		
Method	$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$
TIRG† [38]	14.61	48.37	64.08	90.03	22.67	44.97	65.14
TIRG+LastConv† [38]	11.04	35.68	51.27	83.29	23.82	45.65	64.55
MAAF† [8]	10.31	33.03	48.30	80.06	21.05	41.81	61.60
MAAF+BERT† [8]	10.12	33.10	48.01	80.57	22.04	42.41	62.14
MAAF−IT† [8]	9.90	32.86	48.83	80.27	21.17	42.04	60.91
MAAF−RP† [8]	10.22	33.32	48.68	81.84	21.41	42.17	61.60
CIRPLANT † [27]	15.18	43.36	60.48	87.64	33.81	56.99	75.40
CIRPLANT w/OSCAR † [27]	19.55	52.55	68.39	92.38	39.20	63.03	79.49
Proposed approach	**33.59**	**65.35**	**77.35**	**95.21**	**62.39**	**81.81**	**92.02**

Table 2. Comparison between our method and current state-of-the-art models on the CIRR test set. Best scores are highlighted in bold, second best scores underlined. † denotes results cited from [27]

Table 1 shows the quantitative results on the Fashion-IQ validation set. Our approach outperforms the state-of-the-art by improving up to $\sim 7\%$ in average on the R@10 metric and $\sim 5\%$ in average on the R@50 metric over the best competing methods. Our method has the highest recall in all categories, in particular we can observe how the margin is particularly large in the Shirt category.

Table 2 shows the quantitative results on the CIRR test set obtained through the official evaluation server. Also in this dataset our approach consistently outperforms current methods by a large margin especially in low rank recall measures where we achieve an improvement up to $\sim 14\%$ in R@1. Also the results of the retrieval within the subset of the queries are very good, with an improvement up to $\sim 23\%$ in R_{Subset}@1; this excellent result shows how our approach is also capable of capturing fine-grained modifications between similar images.

6. Conclusions

In this paper we tackled the problem of conditioned image retrieval using the recent CLIP model where we exploited its zero shot transfer features. Using a novel pre-process pipeline tailored for using CLIP in retrieval tasks, we developed a Combiner network that is able to compute a combined feature made from reference images integrated with a textual description. In addition we propose a pre-processing padding method that can improve the performance in datasets that have images with many different aspect ratios. We perform experiments on the challenging fashion dataset FashionIQ and the recently presented CIRR dataset. Experiments on both datasets show that our approach is able to outperform more complex state of the art methods by a significant margin.

The demo system allows users to test the proposed method using image-text pairs of the two datasets or let users provide their own texts, simulating a real-world deployment of the system. The interface allows to implement a turn-based interaction that simulates the behaviour of a user on an e-commerce site. The system can be used also on relatively low performance servers, and can be scaled to large-scale datasets using techniques commonly employed in standard CBIR systems.

6.1. Resources

Code, trained Combiner networks and instructions on how to run the demo locally are available at https://github.com/ABaldrati/CLIP4CirDemo.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 101004545 - ReInHerit.

References

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating CLIP: Towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 2

[2] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1140–1149, January 2021. 2

[3] 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. 3

[4] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 136–152, 11 2020. 7

[5] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 7

[6] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E. Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3119–3124, June 2021. 2

[7] Marcos V Conde and Kerem Turgutlu. CLIP-Art: Contrastive pre-training for fine-grained art classification. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3956–3960, 2021. 2

[8] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 2, 7

[9] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021. 2

[10] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. CLIP2Video: Mastering video-text retrieval via image CLIP. *arXiv preprint arXiv:2106.11097*, 2021. 2

[11] Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. Generating images from caption and vice versa via CLIP-guided generative latent space search. *arXiv preprint arXiv:2102.01645*, 2021. 2

[12] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proc. of European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[13] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional LSTMs. In *Proc. of ACM international conference on Multimedia (MM)*, pages 1078–1086, 2017. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[15] Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy. SAC: Semantic attention composition for text-conditioned image retrieval. In *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4021–4030, January 2022. 2, 7

[16] Surgan Jandial, Ayush Chopra, Pinkesh Badjatiya, Pranit Chawla, Mausoom Sarkar, and Balaji Krishnamurthy. Trace: Transform aggregate and compose visiolinguistic representations for image search with text feedback. *arXiv preprint arXiv:2009.01485*, 2020. 7

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. of International Conference on Machine Learning (ICML)*, 2021. 2

[18] Albert Jimenez, Jose M Alvarez, and Xavier Giro-i Nieto. Class-weighted convolutional features for visual instance search. In *Proc. of British Machine Vision Conference (BMVC)*, 2017. 1

[19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 4

[20] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proc. of AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 1771–1779, May 2021. 2, 3, 6, 7

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[22] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[23] Michal Kucer and Naila Murray. A detect-then-retrieve model for multi-domain fashion item retrieval. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 0–0, 2019. 1

[24] Seungmin Lee, Dongwan Kim, and Bohyung Han. CoSMo: Content-style modulation for image retrieval with text feedback. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 802–812, June 2021. 2, 3, 6, 7

[25] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. CLIP-Event: Connecting text and images with event structures. *arXiv preprint arXiv:2201.05078*, 2022. 2

[26] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. Recent developments of content-based image retrieval (CBIR). *Neurocomputing*, 452:675–689, 2021. 2

[27] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proc. of IEEE/CVF*

International Conference on Computer Vision (ICCV), 2021. 2, 3, 6, 7

[28] Lorenzo Putzu, Luca Piras, and Giorgio Giacinto. Convolutional neural networks for relevance feedback in content based image retrieval. *Multimedia Tools and Applications*, 79(37):26995–27021, 2020. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3

[30] Luis Armando Pérez Rey, Dmitri Jarnikov, and Mike Holenderski. Content-based image retrieval from weakly-supervised disentangled representations. In *Proc. of NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[31] Lorenzo Seidenari, Claudio Baecchi, Tiberio Uricchio, Andrea Ferracani, Marco Bertini, and Alberto Del Bimbo. Deep artwork detection and retrieval for automatic context aware audio guides. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(3s):35, June 2017. 1

[32] Raymond Shiau, Hao-Yu Wu, Eric Kim, Yue Li Du, Anqi Guo, Zhiyuan Zhang, Eileen Li, Kunlong Gu, Charles Rosenberg, and Andrew Zhai. Shop the look: Building a large scale visual shopping system at Pinterest. In *Proc. of ACM International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 3203–3212, 2020. 1

[33] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. RTIC: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*, 2021. 2, 3, 7

[34] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(12):1349–1380, 2000. 1

[35] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2019. 6

[36] Ivona Tautkute, Tomasz Trzciński, Aleksander P. Skorupa, Łukasz Brocki, and Krzysztof Marasek. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019. 1

[37] Federico Vaccaro, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Image retrieval using multi-scale CNN features pooling. In *Proc. of ACM International Conference on Multimedia Retrieval (ICMR)*, ICMR '20, pages 311–315, New York, NY, USA, 2020. Association for Computing Machinery. 1

[38] 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 *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 7

[39] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. CLIP-TD: CLIP targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*, 2022. 2

[40] 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 *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6

[41] Mussarat Yasmin, Muhammad Sharif, and Sajjad Mohsin. Neural networks in medical imaging applications: A survey. *World Applied Sciences Journal*, 22(1):85–96, 2013. 1

[42] Youngjae Yu, Seunghwan Lee, Yuncheol Choi, and Gunhee Kim. Curlingnet: Compositional learning between images and text for fashion iq data. *arXiv preprint arXiv:2003.12299*, 2020. 2, 7

[43] Yifei Yuan and Wai Lam. Conversational fashion image retrieval via multiturn natural language feedback. In *Proc. of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, Jul 2021. 2

[44] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. Learning a unified embedding for visual search at Pinterest. In *Proc. of ACM International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2412–2420, 2019. 1

[45] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. Visual search at Alibaba. In *Proc. of ACM International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 993–1001, 2018. 1

[46] Liang Zheng, Yi Yang, and Qi Tian. SIFT meets CNN: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(5):1224–1244, 2017. 2

[47] Wengang Zhou, Houqiang Li, and Qi Tian. Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064*, 2017. 2