Conditioned and composed image retrieval combining and partially fine-tuning 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

In this paper, we present an approach for conditioned and composed image retrieval based on CLIP features. In this extension of content-based image retrieval (CBIR) an image is combined with a text that provides information regarding user intentions, and is relevant for application domains like e-commerce. The proposed method is based on an initial training stage where a simple combination of visual and textual features is used, to fine-tune the CLIP text encoder. Then in a second training stage we learn a more complex combiner network that merges visual and textual features. Contrastive learning is used in both stages. The proposed approach obtains state-of-the-art performance for conditioned CBIR on the FashionIQ dataset and for composed CBIR on the more recent CIRR dataset.

1. Introduction

Content-Based Image Retrieval (CBIR) systems search images in a database using as a query an input image, and computing a distance between the visual features extracted from the query and the features stored in the database. Research in computer vision and multimedia has addressed the problem of studying features that are discriminative enough to deal with different images and that are robust enough to deal with small variations of the same images.

Several variations of CBIR request that the user provides some additional information regarding the intent or context of the query image. For example, the interaction with a user, that provides additional information on what is "similar" or "dissimilar" according to them [27], is used in approaches based on relevance feedback. More recently, CBIR systems have been extended by asking the user to provide textual information that is added to the visual features of the image. This task is called conditioned image retrieval since the user typically adds some conditions that change some visual aspects of the query image; this task is of interest to implement interactive search systems for fashion [14, 37], Figure 1 shows such a scenario. The task has been very recently generalized as composed image retrieval, where the query is composed as an image-language pair, and using both visual and textual modalities to specify the user's intent [22].

Figure 1. Example of conditioned fashion image retrieval for e-commerce applications. Product search is refined by the user providing details and constraints in natural language. The system uses both visual and textual features to retrieve the desired result.

In this work, we address both conditioned and composed retrieval, the first applied to fashion and the latter to general images. The proposed system combines visual and textual features computed using the OpenAI CLIP network; in an initial training stage the features are simply combined with summation, with the goal of fine tuning the CLIP text encoder, then in a second training stage we learn a combiner network that merges CLIP visual and textual features; contrastive learning is used in both stages. Despite the simplicity of the approach, the proposed method achieves state-of-the-art results on two commonly used datasets, FashionIQ [37] for conditioned retrieval, and CIRR [22] for composed retrieval.

2. Related works

Several surveys provide an overview of CBIR approaches and their evolution in the past years. Zheng *et al*. [40] and Zhou *et al*. [41] surveyed image search approaches including both methods based on engineered and those based on learned features. More recently, Dubey [10] has surveyed CBIR approaches based on deep learning.

Visual and language pretraining

The OpenAI CLIP [28] network has very recently obtained remarkable results in multi-modal zero shot learning, and more in general it performs consistently well on different tasks despite not being directly optimized for a specific benchmark, thanks to its generalization capabilities of both images and text. CLIP learns associations between the images and textual descriptions using 400 millions of image-text pairs scraped from the web for training. Effectiveness of CLIP is still subject of study [1], although it has already been successfully applied to different tasks like fine-grained art classification [7], image generation [12], zero shot video retrieval [11], event classification [20] and visual commonsense reasoning [36]. Other approaches to learn image-text alignment have been proposed in [6, 16]. ALIGN [16] uses a dual-encoder architecture and is trained on a huge dataset of 1 billion image-text pairs. Instead, the method proposed in [6] is much more data efficient, exploiting contrastive distillation, and requires a training dataset that is $133\times$ smaller than that of CLIP.

Conditioned and combined image retrieval

This work is related to the recent problem of conditioned fashion image retrieval [37], and with the very recent problem of composed image retrieval of generic images [22].

The first task has been addressed in a large number of works. In [5], is presented a method based in a transformer that can be seamlessly plugged in a CNN to selectively preserve and transform the visual features conditioned on language semantics. In [35] has been presented Text Image Residual Gating (TIRG), a method that combines image and text features using gating and residual features. In [30] the authors combine graph neural networks and skip connections. In [19], the authors use two different neural network modules, one for image style and one for image content. In [17] a Correction Network is proposed to model explicitly the difference between the reference and target image in the embedding space. In [9] 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. In [2] has been proposed ComposeAE, an autoencoder-based model, to learn the composition of image and text features using a deep metric learning (DML) approach. In [38] 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 first delivers the source image to the candidate cluster according to a given query in an embedding space, while the second 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 [39]. 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. A network that learns how to combine visual and textual features computed from CLIP has been proposed in [3]. Finally, text-conditioned image retrieval has been addressed in [14], where the authors present the SAC (Semantic Attention Composition) framework 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 [22], 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.

Inspired by [3] 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.

Our contributions can be summarized as three aspects:

- We propose a novel fine-tuning scheme for conditioned image retrieval using CLIP-based features. To the best of our knowledge, we are the first to fine-tune the CLIP text encoder in a non-conventional way which breaks up the symmetry between the two encoders
- We propose a novel two-stage approach which integrates the CLIP text encoder fine-tuning with the training of a Combiner network which learns to fuse the multimodal features. Such approach achieves state-of-the-art performance on FashionIQ and CIRR datasets.
- We perform a study that tries to explain the effects of our approach on the feature distribution in the embedding spaces and how these effects are related to performance in the retrieval task.

3. The proposed method

The goal of text conditioned and composed image retrieval is to retrieve the best matching image given a multimodal input consisting of an image-text pair. I.e., given an image (named *reference image*) and a text (named *relative caption*) which expresses some modification with respect to the reference image, the aim of the retrieval is to find the best matching image which satisfy both the visual similarity constraints imposed by the reference image and integrates the changes expressed by the relative caption. In order to perform an effective retrieval, the system must be able to understand both the semantics of the image and the meaning of the text, to combine such multi domain information and finally to perform the retrieval using the fused representation.

In contrast to the most of the previous work like [9, 14, 17, 29, 38] that process text and image using different models, in our approach we follow [3] which leverages the image-text common embedding space obtained using CLIP features. As shown in [28] the CLIP encoders realize a common embedding space where analogous concepts, expressed through text or images, tend to have similar features.

Although having a common embedding space between text and images is a good starting point in the task we want to address, it is still not enough. Ideally, we would like to have a textual embedding space that contains displacement vectors in the image embedding space since the conditioned image retrieval task consists of moving between two points in image space using textual information.

We do not have any sort of guarantee that CLIP common embedding space has such additivity properties. Mathematically speaking: given an image of a dog ($\mathbf{a}$) and the text `a photo of a dog` ($\mathbf{b}$), if we denote the CLIP image encoder as ϕ_I and the CLIP text encoder as ϕ_T, the way of CLIP is trained should guarantee that $\phi_I(a) \approx \phi_T(b)$. However given an image of a black dress ($\mathbf{x}$) and the relative caption `is blue` ($\mathbf{y}$), denoting with ($\mathbf{z}$) the image of a blue dress, there is no guarantee that $\phi_I(x) + \phi_T(y) \approx \phi_I(z)$.

To overcome this mismatch between the large scale pre-training of CLIP and the downstream task we developed a two-stage approach. In the first stage we fine-tune the CLIP text encoder to adapt its embedding space to the task we want to address. In the second stage we train a Combiner network which learns to fuse the multi-modal features extracted with the CLIP image encoder and the fine-tuned CLIP text encoder. The architecture of the Combiner network is a slightly modified version of the Combiner network presented in [3].

3.1. Text encoder fine-tuning

Figure 2 shows an overview of the text encoder fine-tuning stage. In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large scale image-text pre-training and the downstream task.

Firstly we use the CLIP encoders to extract the features from the training triplets. Then we combine the query features with a simple element-wise sum followed by an L_2-normalization. Finally, with the aid of a contrastive loss which takes as inputs the combined features and the target features, the weights of the CLIP text encoder are updated.

With the fine-tuning of the text encoder using the sum as a combination of query features we break up the symmetry between the text and the image CLIP embedding spaces. Such broken symmetry is a desirable effect in a task where this symmetry does not exists: in conditioned image retrieval the objective is to move from a starting point (corresponding to the reference image) to a target point (corresponding to the target image) on the image manifold using textual information as a guide to achieve this movement. As previously mentioned we need that the vectors in the textual embedding space represent the displacement vectors between two points in the image space

We decided to keep the image encoder frozen since we are not interested in modifying the CLIP image embedding space which is proven to be high quality and discriminative [7, 20, 26, 28].

3.2. Combiner network training

Figure 3 shows an overview of the Combiner training stage. The general framework of the training is the same described in the text encoder fine-tuning stage. However, in this stage we do not backpropagate the gradients updating the weights of any CLIP encoder, instead we train from scratch a Combiner network which learns to combine the multimodal query features so that such combination is as near as possible to the target features in the image embedding space.

The Combiner network, illustrated in Figure 4, is based on architecture described in [3]. The Combiner outputs a normalized sum of multiple components: a convex combination of text and image features and a learned text-image mixture. Starting from the image and text features extracted via the CLIP encoders, firstly we project them via a linear layer followed by the ReLU function. Projected features are then concatenated and fed to two different branches with a very similar structure: two linear layers with a ReLU non-linearity between them. The aim of the first branch is to compute the coefficients of a convex combination between the image and the text features; to accomplish this, a sigmoid function is applied on the output of the branch. The other branch outputs a learned mixture of text and image features. The three contributions are finally summed and

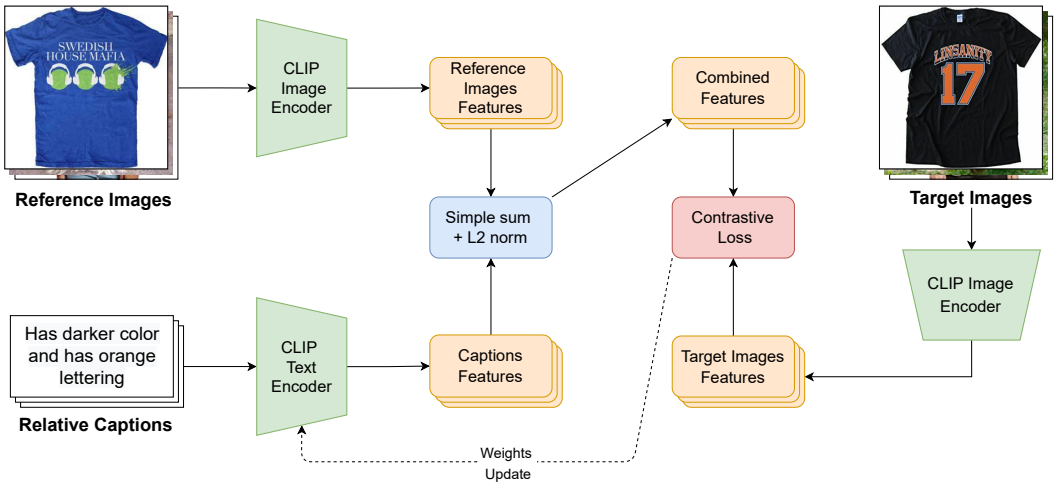

Figure 2. First stage of training overview. In this stage, using a contrastive loss, we fine-tune the CLIP text encoder to adapt its embedding space to the conditioned image retrieval task.

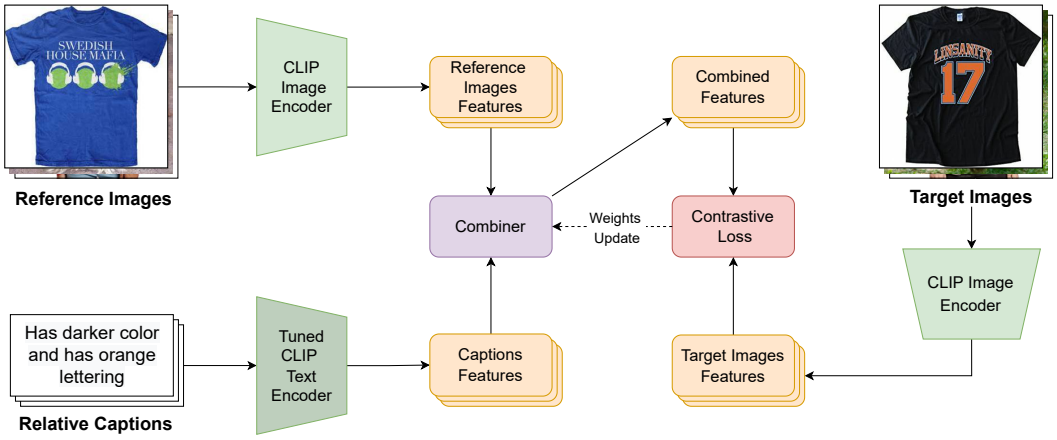

Figure 3. Second stage of training overview. In this stage, always using a contrastive loss, we train from scratch a Combiner network which learns to fuse the multimodal features extracted with the CLIP encoders. At inference time the fine-tuned text encoder and the trained Combiner are used to produce an effective multi-modal representation used to query the database.

L_2-normalized. Dropout is applied after each layer to stabilize the training process and reduce overfitting.

As illustrated in [3] the convex combination of text and image features is essential to obtain state-of-the-art performance allowing the Combiner to learn an offset from a good starting point. This is another reason why the fine-tuning of the CLIP text encoder is so important in our approach: since the Combiner network training has proven to be very sensitive to the presence and quality of a good starting point, the fine-tuning we described earlier makes this starting point even better, improving the performance of the entire approach.

3.3. Image Preprocessing

The standard image preprocess pipeline of CLIP is essentially structured 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 in our experiments we decide to use a preprocess pipeline which aims to find a compromise between the aforementioned pipelines. An image is padded only if its

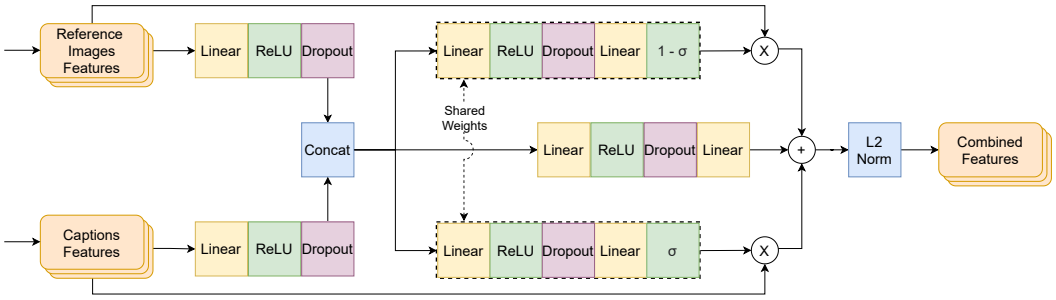

Figure 4. Architecture of the combiner network. It takes as inputs the reference image features and the captions features and outputs a fused representation. σ denotes the sigmoid function.

aspect ratio is above a fixed target. Furthermore, when such padding is performed, the image is not squared but its aspect ratio is brought to the chosen target ratio. After such adaptive padding, the center crop operation is performed.

3.4. Training Details

In both stages the training of the system is performed with batches of triplets formed by: reference images, relative captions and target images. Following [19, 30, 35], as contrastive loss, we employ the batch-based classification (BBC) loss:

$$L = \frac{1}{B}\sum_{i=1}^{B} -\log\{\frac{\exp\{\lambda * \kappa(\psi_i, \phi_i^+)\}}{\sum_{j=1}^{B}\exp\{\lambda * \kappa(\psi_i, \phi_j)\}}\} \quad (1)$$

Where $\psi_i = f_{CombiningFunction}(x_i^{query}, t_i)$ are the combined features of the query, and $\phi_i^+ = f_{ImageEncoder}(x_i^{target})$ is the representation of the target image of that query. κ is a similarity kernel implemented as a normalized dot product in our experiments and λ is a temperature parameter which controls the range of the logits. Following [28] we set the λ parameter to 100 to ensure that the logits have sufficient dynamic range in order not to penalize the training process. During the text encoder fine-tuning stage the combining function is a simple sum followed by a L_2-normalization, while during the training of the Combiner network the combining function is the Combiner itself.

We perform experiments using two different CLIP models. The smallest one has a visual encoder based on a modified ResNet-50 (RN50) [13] architecture, it takes as input images of 224×224 pixels and has an embedding dimension of 1024; its text encoder is a Transformer encoder [34] with 12 layers, 8 heads and a width of 512. The biggest one, denoted as RN50x4, has a visual encoder which follows the EfficientNet-style model scaling [32] and uses approximately $\times 4$ the computation of a ResNet-50, it takes as input images of 288×288 and has an embedding dimension of 640; its text encoder is a Transformer encoder with with 12 layers, 10 heads and a width of 640.

Following the original CLIP training pipeline, in the text encoder fine-tuning we used AdamW optimizer [23] with a learning rate of $1e-6$ and a weight decay coefficient of $1e-2$. Due to GPU memory constraints the batch size was set to 128. We fine-tuned the text encoder for 100 epochs.

In the Combiner training, both the CLIP encoders have been kept frozen and the only trained part of the model is the Combiner network. We used Adam optimizer [18] with a learning rate set to $2e-5$. We trained the model for a maximum of 300 epochs. The batch size was set to 4096.

We used the PyTorch library [25] throughout the experiments. The target ratio in the preprocess pipeline was set to 1.25. Mixed-precision [24] was used to save memory and accelerate training during both stages.

4. Experimental results

4.1. Datasets and metrics

FashionIQ [37] is a dataset for fashion conditioned image retrieval that contains 30,134 triplets from 77,684 images crawled from the web. The images are divided into three different categories: *Dress*, *Toptee* and *Shirt*.

Following the standard experimental setting for this dataset, we report as evaluation metrics the average recall at rank K (Recall@K) at two different ranks: 10 and 50. All the results are on the validation set since, at the time of writing, test set ground-truth labels have not been publicly released.

CIRR [22] (Compose Image Retrieval on Real-life images) is a dataset containing 21,552 real-life images taken from the popular natural language reasoning $NLVR^2$ dataset [31]. It contains 36,554 triplets randomly assigned in 80% for training, 10% for validation and 10% for test.

The dataset has been designed to overcome two common issues that affect conditioned image retrieval datasets (such as FashionIQ): non-complex images with too narrow domain and the high number of false-negatives.

The images of the dataset are grouped in multiple subsets of six images that are semantically or visually similar. The relative captions are collected so that they describe differences between two images in the same subset. This is done in order to have true-negative images with a high visual similarity.

The standard evaluation protocol proposed by the authors of the dataset consists of reporting the recall at rank K (Recall@K) at four different ranks (1, 5, 10, 50). Thanks to the design of the CIRR dataset, the standard evaluation protocol also includes the $\text{Recall}_{\text{Subset}}$@K metrics which consider only the images in the subset of the query. This *subset* metric has two main advantages: it is unaffected by false-negative samples and, thanks to negative samples with high visual similarity, it is able to capture the fine-grained reasoning ability of the methods.

4.2. Fine-tuning upshot

In this section we present a set of experiments which demonstrate how the fine-tuning of the CLIP text encoder brings actual benefits in terms of performance and helps the Combiner network in its task.

For each backbone we perform four different experiments varying the combining function and the CLIP text encoder (we use both the out-of-the-box CLIP text encoder and the fine-tuned one).

			Shirt		Dress		Toptee		Average	
Backbone	**FT**	**CF**	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$
RN50	✗	Sum	19.73	35.53	17.60	36.09	21.83	42.84	19.72	38.15
RN50	✓	Sum	30.71	52.21	27.52	51.66	33.61	58.95	30.61	54.27
RN50	✗	Combiner	31.80	53.38	26.82	51.31	33.40	57.01	30.67	53.90
RN50	✓	Combiner	35.77	57.02	31.73	56.02	36.46	62.77	34.65	58.60
RN50x4	✗	Sum	25.27	41.27	20.62	40.36	27.43	47.83	24.44	43.15
RN50x4	✓	Sum	35.77	57.41	31.14	55.18	38.09	61.04	35.00	57.88
RN50x4	✗	Combiner	36.36	58.00	31.63	56.67	38.19	62.42	35.39	59.03
RN50x4	✓	Combiner	39.99	60.45	33.81	59.40	41.41	65.37	38.32	61.74

Table 1. Recall at K on FashionIQ validation set varying the combining function and the CLIP text encoder. **FT** stands for fine-tuning and denotes whether the text encoder used has been fine-tuned in the first stage. **CF** stands for combining function and denotes which function has been used to combine the query features

We report the results for each combination in Table 1 for FashionIQ dataset and in Table 2 for CIRR dataset.

The effectiveness of the simple sum of out-of-the-box CLIP features for conditioned image retrieval was already known for FashionIQ dataset [3]; from our experiments we can see that, also for a broader domain dataset such as CIRR, the performance remains solid. This fact is very interesting, it makes us suppose that the CLIP common embedding maintains a certain degree of additivity, which makes the sum baseline perform quite well despite its simplicity and the absence of task-specific training.

			Recall@K				R_{subset}@K		
Backbone	**FT**	**CF**	$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$
RN50	✗	Sum	21.24	50.68	64.29	87.32	54.48	75.94	87.66
RN50	✓	Sum	31.57	65.10	77.47	94.47	65.31	84.64	93.06
RN50	✗	Combiner	31.28	64.84	77.88	94.90	62.04	81.58	91.60
RN50	✓	Combiner	37.00	70.94	82.28	96.13	67.47	85.39	93.66
RN50x4	✗	Sum	21.96	52.24	66.18	88.18	52.71	74.74	86.73
RN50x4	✓	Sum	32.62	67.02	79.74	95.31	65.41	84.67	92.54
RN50x4	✗	Combiner	33.63	67.16	80.22	95.58	63.62	82.85	92.15
RN50x4	✓	Combiner	39.75	73.71	83.90	96.87	70.92	87.42	94.19

Table 2. Recall at K on CIRR validation set varying the combining function and the CLIP text encoder. **FT** stands for fine-tuning and denotes whether the text encoder used has been fine-tuned in the first stage. **CF** stands for combining function and denotes which function has been used to combine the query features

The improvement obtained with the text encoder fine-tuning (keeping as a combining function the sum of query features) confirms our intuitions about the necessity of breaking the symmetry between the text-images embedding spaces by making the text one contains displacement vectors in the image manifold. Using both backbones and in both datasets the improvement obtained with such text encoder fine-tuning is very consistent and comes close to equaling the improvement achieved by using the Combiner network over the non fine-tuned features.

Finally it is very remarkable to notice that the Combiner network benefits from the usage of the fine-tuned textual features further improving the overall results. These results are coherent with the Combiner architecture, in fact, as illustrated in Figure 4, it is trained to learn the offset with respect to a learned convex combination of textual and visual features. The more discriminative this convex combination is, the simpler is the task of this network.

It is also worth mentioning that performing fine-tuning of the text encoder during training of the Combiner network does not result in any performance improvement. This lack of improvement was already shown by [3] in their ablation studies.

4.3. Comparison with SotA

In this section we compare the proposed method results with state-of-the-art models on FashionIQ and CIRR datasets.

Table 3 shows the comparison between our method and current state-of-the-art models on the FashionIQ validation set. Our two-stage approach manages to achieve state-of-the-art results in all recall metrics using both backbones. Compared to [3], which also uses CLIP features and a similar Combiner network, thanks to the carefully crafted two stage approach and the enhanced Combiner, we improve up to $\sim 5\%$ both on average R@10 and average R@50. Compared to the other SotA approaches the performance improvement is even greater with an average improvement of $\sim 9\%$ on R@10 metric and $\sim 7\%$ on R@50 metric over

Method	Shirt		Dress		Toptee		Average	
	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$
JVSM [4]	12.0	27.1	10.7	25.9	13.0	26.9	11.9	26.6
CIRPLANT w/OSCAR [22]	17.53	38.81	17.45	40.41	21.64	45.38	18.87	41.53
TRACE w/BERT [15]	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 [9]	21.3	44.2	23.8	48.6	27.9	53.6	24.3	48.8
CurlingNet [38]	21.45	44.56	26.15	53.24	30.12	55.23	25.90	51.01
ARTEMIS [8]	21.78	43.64	27.16	52.40	29.20	54.83	26.05	50.29
RTIC-GCN w/GloVe [30]	23.79	47.25	29.15	54.04	31.61	57.98	28.18	53.09
CoSMo [19]	24.90	49.18	25.64	50.30	29.21	57.46	26.58	52.31
AACL [33]	24.82	48.85	29.89	55.85	30.88	56.85	28.53	53.85
DCNet [17]	23.95	47.30	28.95	56.07	30.44	58.29	27.78	53.89
SAC w/BERT [14]	28.02	51.86	26.52	51.01	32.70	61.23	29.08	54.70
Baldrati et al (RN50x4) [3]	35.76	56.20	27.20	53.57	36.31	61.14	33.09	56.99
Proposed approach (RN50)	35.77	57.02	31.73	56.02	36.46	62.77	34.65	58.60
Proposed approach (RN50x4)	**39.99**	**60.45**	**33.81**	**59.40**	**41.41**	**65.37**	**38.32**	**61.74**

Table 3. 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.

Method	Recall@K				R_{subset}@K		
	$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$
TIRG† [35]	14.61	48.37	64.08	90.03	22.67	44.97	65.14
TIRG+LastConv† [35]	11.04	35.68	51.27	83.29	23.82	45.65	64.55
MAAF† [9]	10.31	33.03	48.30	80.06	21.05	41.81	61.60
MAAF+BERT† [9]	10.12	33.10	48.01	80.57	22.04	42.41	62.14
MAAF−IT† [9]	9.90	32.86	48.83	80.27	21.17	42.04	60.91
MAAF−RP† [9]	10.22	33.32	48.68	81.84	21.41	42.17	61.60
ARTEMIS [8]	16.96	46.10	61.31	87.73	39.99	62.20	75.67
CIRPLANT † [22]	15.18	43.36	60.48	87.64	33.81	56.99	75.40
CIRPLANT w/OSCAR † [22]	19.55	52.55	68.39	92.38	39.20	63.03	79.49
Proposed approach (RN50)	35.81	68.80	80.17	95.25	66.96	85.25	93.13
Proposed approach (RN50x4)	**38.53**	**69.98**	**81.86**	**95.93**	**68.19**	**85.64**	**94.17**

Table 4. 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 [22]

the best competing method [14]. Between the two backbones, we note that the bigger RN50x4 obtains the best performance, with an improvement on the smaller RN50 in the range of about 3% to 5% in all categories. The results obtained using the RN50 backbone, beside the comparable result with DCNet [17] in Dress R@50, are always the second best results w.r.t. the RN50x4 backbone.

Table 4 shows the comparison between our method and current state-of-the-art models on the CIRR test set. These quantitative results are obtained through the official evaluation server. Also in this dataset our approach manages to outperform current methods especially in low rank recall measures where we achieve an impressive improvement up to $\sim$ 19% in R@1. The results within the subset of the queries are even better, with an improvement up to $\sim$ 29% in R_{Subset}@1; this excellent result shows how our approach is also capable of capturing fine-grained modifications between similar images. Compared to FashionIQ the gap between the two backbones is smaller and stabilizes in the range of about 0.5% to 2% in all recall metrics. As in FashionIQ, the results obtained using the RN50 backbone are always the second best results w.r.t. the RN50x4 backbone.

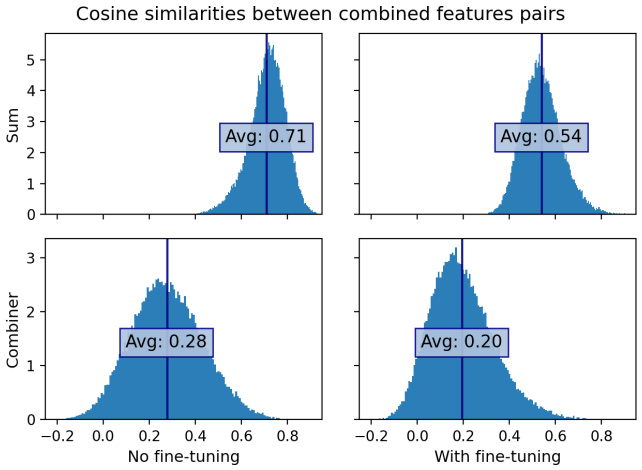

Figure 5. Histograms of cosine similarities between combined features pairs. The x-axis represents the cosine similarities while the y-axis represents the (normalized) number of pairs. In the top line plots we have used the simple sum as a combining function, in the bottom line ones we have used the Combiner network. In the left side plots we used the out-of-the-box CLIP text encoder, in the right ones we used the text encoder fine-tuned during the first stage of the training. The histogram is normalized such that the area under each curve integrates to 1.

4.4. Feature distribution study

In this section we present a few experiments which aim to provide an intuition on how the combined features are distributed in the embedding space. Specifically we focus on the text encoder fine-tuning and the Combiner network effects. Although the experiments were performed on both datasets, since the results are exactly the same we report only the experiments performed on the FashionIQ dataset. All the experiments were carried out on the validation set to avoid biased results.

We are going to report two different sets of experiments with distinct objectives. The first one focuses on studying the feature distribution in the embedding space while the second one aims to explore how such feature distribution affects the multimodal retrieval process.

For studying how the combined features are distributed in the embedding space, following [21], we compute the pairwise cosine similarities among them. Obviously, the more the features are evenly distributed, the lower their average similarity will be. In each of our experiments we randomly sample 50K pairs from the dataset since, because of the quadratic growth, it is unfeasible to use them all.

Figure 5 shows how such pairwise similarities are distributed. The reduced cosine similarity of combined features after the text encoder fine-tuning shows that it leads to a more efficient use of the embedding space with more discriminative features. This effect of a better occupation of the embedding space can be noticed, even to a greater

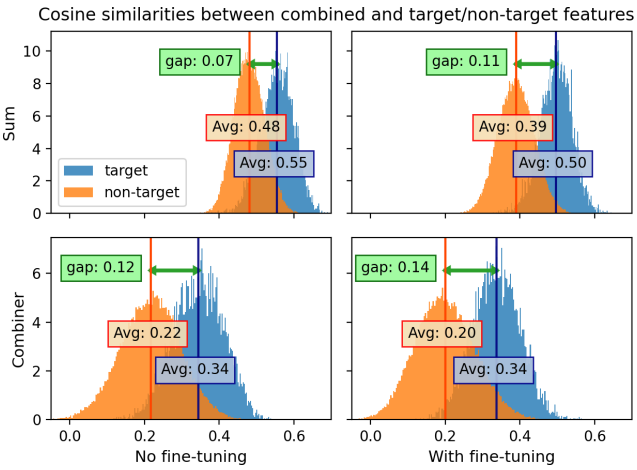

Figure 6. Histograms of cosine similarities between combined and target/non-target features. The x-axis represents the cosine similarities while the y-axis represents the (normalized) number of pairs. In orange•: cosine similarities between combined and non-target features. In blue•: cosine similarities between combined and target features. In green•: similarity gap between combined-target and combined-non target features. In the top line plots we have used the simple sum as a combining function, in the bottom line ones we have used the Combiner network. In the left side plots we used the out-of-the-box CLIP text encoder, in the right ones we used the text encoder fine-tuned during the first stage of the training. To manage different number of images the histogram is normalized such that the area under each curve integrates to 1

extent, if we consider the similarity differences between using the simple sum and using the Combiner as a combining function. It is interesting to see how both the fine-tuning of the textual encoder and the Combiner network contribute to reduce the *cone effect*: i.e. the fact that "the effective embedding space is restricted to a narrow cone for trained models and models with random weights" [21].

The previous experiments illustrate how both stages of our training pipeline affect the embedding space. However, they do not explain why such increased occupation should boost the retrieval performance. The second set of experiments deals precisely with that, i.e. studying the influence of such embedding space reshaping in the retrieval task.

We compute the cosine similarities between the combined features and the index image features. In detail, for each experiment, we perform two distinct computations: in the first one we compute the similarities between the combined features and the target features, in the second one the similarities are computed between the combined features and index features which differ from the target one. In our setting we compare each combined feature with ten non-target features which are randomly selected from the index image features.

Figure 6 highlights the similarities between combined features and target/non-target image features. It is interesting to note that we achieve the highest combined-target features average similarity using the sum of untuned CLIP features. The fine-tuning first and the Combiner network then, rather than closing the gap with the target features, they widen it with the non-target features. We formulate the hypothesis that, in the datasets we experimented with, the retrieval performances are more influenced by the similarity gap between the combined-target and combined-non target features (displayed in Figure 6 as the green arrow) rather than the absolute value of the combined-target similarity.

The two sets of experiments show two sides of the same coin. In the first experiment it is shown that, thanks to the fine-tuning and the Combiner network, the combined embedding space is used in a more efficient way. In the second experiment it is shown how this increased efficiency is fundamental to "move away" the combined features from the non-target features.

5. Conclusions

In this paper we present a novel fine-tuning scheme for conditioned image retrieval using CLIP-based features: we fine-tune the CLIP text encoder to adapt its embedding space to the task breaking up the symmetry between the encoders. We then propose a novel two-stage approach which integrates the CLIP text encoder fine-tuning with the training of a Combiner network which learns to combine the multimodal query features.

We conducted experiments on the fashion dataset FashionIQ and on the open domain dataset CIRR. Experiments on both dataset show that our two-stage approach manages to reach state-of-the-art results by a consistent margin. The performances of the proposed method are particularly solid on low rank recall measures indicating the ability of capturing fine-grained modifications among similar images.

Finally we conducted a study which aims to explain the effects of our approach on feature distribution in the embedding space and how these effects are related to performance in the retrieval task. From the experiments we can notice that both the text encoder fine-tuning and the Combiner network training led to a more efficient usage of the embedding space. Moreover it is shown that such increased sparsity in the embedding space helps to "move away" the combined features from the non-target ones improving the effectiveness of the retrieval.

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. 2, 3, 4, 6, 7

[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, 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, 3

[8] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *International Conference on Learning Representations*, 2021. 7

[9] 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, 3, 7

[10] 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

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

[12] 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

[13] 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. 5

[14] 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. 1, 2, 3, 7

[15] 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

[16] 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

[17] 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, 7

[18] 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. 5

[19] 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, 5, 7

[20] 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, 3

[21] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022. 7, 8

[22] 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. 1, 2, 5, 7

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[24] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017. 5

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proc. of International Conference on Neural Information*

Processing Systems, pages 8024–8035. Curran Associates, Inc., 2019. 5

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 3

[27] 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. 1

[28] 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, 5

[29] 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. 3

[30] 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, 5, 7

[31] 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. 5

[32] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. 5

[33] Yuxin Tian, Shawn Newsam, and Kofi Boakye. Image search with text feedback by additive attention compositional learning. *arXiv preprint arXiv:2203.03809*, 2022. 7

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 5

[35] 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, 5, 7

[36] 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

[37] 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. 1, 2, 5

[38] 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, 3, 7

[39] 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

[40] 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

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