Learning Joint Visual Semantic Matching Embeddings for Language-guided Retrieval

Yanbei Chen[*1] and Loris Bazzani[✉2]

[1] Queen Mary University of London
yanbei.chen@qmul.ac.uk
[2] Amazon
bazzanil@amazon.com

Abstract. Interactive image retrieval is an emerging research topic with the objective of integrating inputs from multiple modalities as query for retrieval, e.g., textual feedback from users to guide, modify or refine image retrieval. In this work, we study the problem of composing images and textual modifications for language-guided retrieval in the context of fashion applications. We propose a unified Joint Visual Semantic Matching (JVSM) model that learns image-text compositional embeddings by jointly associating visual and textual modalities in a *shared* discriminative embedding space via compositional losses. JVSM has been designed with *versatility* and *flexibility* in mind, being able to perform multiple image and text tasks in a *single* model, such as text-image matching and language-guided retrieval. We show the effectiveness of our approach in the fashion domain, where it is difficult to express keyword-based queries given the complex specificity of fashion terms. Our experiments on three datasets (Fashion-200k, UT-Zap50k, and Fashion-iq) show that JVSM achieves state-of-the-art results on language-guided retrieval and additionally we show its capabilities to perform image and text retrieval.

1 Introduction

Text-based image retrieval methods have been the foundation of many advances and developments in different domains, such as search engines, organization of documents, and more recently natural language processing-based technologies. On the opposite spectrum, content-based image retrieval approaches have demonstrated great success in various tasks in the past decade, such as image search, face recognition and verification, and fashion product recommendation. Given the growing maturity of these two research fields, in the recent years we are witnessing the cross-pollination and conjunction of these areas. One of the main motivations is that documents often contain multimodal material, including images and text.

A user-friendly retrieval interface should entail the flexibility to ingest various forms of information, such as image (Figure 1, top) or text (Figure 1, middle),

[*] Work done during an internship with Amazon.
[✉] Corresponding author.

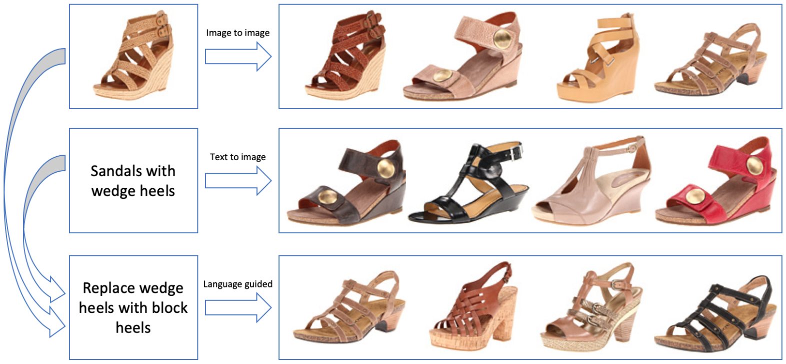

Fig. 1. Different image retrieval pipelines: 1) image-to-image retrieval which focuses on retrieving visually similar images but it includes images with other types of heels, 2) text-to-image retrieval by specifying the type of heels (no visual similarity guaranteed) and 3) language-guided retrieval of image, where the modification text is used to obtain images visually similar to the original one but replacing one aspect (type of heels).

and empowers users to interact with the system (Figure 1, bottom). Interactive retrieval is therefore becoming the core technology for improving the online shopping experience via automated shopping assistants, which help the user to search or discover products to purchase. Interactions can be found in different forms: relevance [26] (e.g., similar/dissimilar inputs), drawing or region selection [44, 22] (e.g., sketching, spatial layout, in-painting, clicking), and textual feedback [15, 46, 1, 36, 9] (e.g. attributes, language, including speech to text).

In this work, we explore the textual form of interaction with a specific focus on language-guided retrieval via modification text [36] for images in the fashion domain. As sketched in Figure 1, the idea is to augment the query image with a modification text describing how to modify the image, then the method should retrieve visually similar images as defined by the modification, e.g., by replacing wedge heels with block heels. Our main motivation is that refining the search results with modification text in form of natural language or attribute-like descriptions is the key for a user-friendly interactive search experience, especially in the context of fashion where visual cues are important and it is typically difficult to express keyword-based queries given the specificity of fashion terms. We present a unified Joint Visual Semantic Matching (JVSM) approach that has the capability of learning image-text compositional embeddings. JVSM has been designed with *versatility* and *flexibility* in mind, being able to perform multiple retrieval tasks in a *single* model, including language-guided retrieval of image or text, and text-image matching.

Existing image retrieval models are generally optimized for the image-to-image retrieval task, which has its limitations given that images are often associated to multimodal information. To bridge the gap between the textual and imagery modalities, recent work considers learning visual semantic embeddings [5], such that image and text are semantically comparable in a shared common space. In this way, it is possible to train image retrieval models to perform text-image matching tasks. In order to provide retrieval methods the ability to deal with language guidance, recent work [36] proposes to compose image and modification text as search input query, which allows to refine the search results tailored to the additional textual input.

To the best of our knowledge, JVSM is the first attempt to jointly learn image-text embeddings as well as compositional embeddings in a unified embedding space, which enables us to perform language-guided retrieval of image or text, and text-image matching with a single model. The key technical challenges that we tackle in this work are: (1) learning a visual semantic embedding space shared by image and text; and (2) learning the mapping functions that allow to compose image and modified text for refining image retrieval results. Although these two aspects have been examined separately in [5] and [36], the problem of how to jointly address them in a unified solution for fashion search has not been systematically investigated or addressed. Another advantage of the proposed framework is that it can be trained using privileged information, which is exclusively available at training time, and it functions to constrain the solution space for the image-text compositions.

JVSM is trained using an extension of the visual semantic embedding loss [5] with two new loss components that act in the compositional embedding space. The objective of those loss components is to encourage synergistic alignment between the compositional embeddings and the target images, target textual descriptions to be retrieved. We demonstrate the benefits of JVSM with respect to the state-of-the-art methods by conducting a comprehensive evaluation on three fashion datasets: Fashion-200k [11], UT-Zap50K [43, 42], and Fashion-iq [10].

The contributions of our work are summarized in the following:

- We present a unified model (JVSM) to learn a visual semantic embedding space and compositional functions that allow to compose image and modification text. The key novelty of JVSM lies in its versatility to perform multiple image and text retrieval tasks using a single unified model, including language-guided retrieval of image or text, and text-image matching.
- We introduce novel loss formulations which define a unified embedding space where image embeddings, text embeddings and compositional embeddings are synergistically tied and optimized to be fully comparable.
- We demonstrate that JVSM can ingest textual information in different forms of composition: attribute-like modifications (e.g., "replace wedge heels with block heels") and natural language form (e.g., "the dress I am searching for has a floral pattern and is shorter").
- We show that JVSM effectively uses privileged information that is only available at training time. The advantages are not only boosts in performance, but also its task-agnostic property, *i.e.*, to be flexibly used for processing different inputs, which is desirable in many retrieval interfaces.
- We advance the state-of-the-art of language-guided retrieval and text-image matching on different fashion datasets.

2 Related Work

Interactive Retrieval aims at incorporating user feedback into an image retrieval system to guide, modify or refine the image retrieval results tailored to the users' expectations. User feedback can be given in different formats such as modification text [36,35], attribute [15,46,1,27], natural language [9,10], spatial layout [22], and sketch [44]. Since text naturally serves as an effective modality to express users' fine-grained intentions for interactive retrieval, we focus on language-guided retrieval. In this problem, *compositional learning* [23,24,14,31,3,13] plays a fundamental role to integrate various forms of textual feedback (e.g., attribute-based modification text, and natural language) with the imagery modality [36,10]. Vo et al. [36] proposes residual gating to modify the image only when the attribute feedback is relevant. Guo et al. [10] propose a *multi-turn* model with a simple compositional module and a new fashion dataset for natural language-based interactive retrieval (Fashion-iq). In this work, we tackle *single-turn* retrieval with a multi-task learning model: JVSM, which facilitates a user-friendly retrieval interface to process both unimodal and multimodal inputs.

Text-Image Matching, also known as a text-to-image or image-to-text retrieval [6,40,38,5,37,4,45,19], aims at learning a cross-modal visual-semantic embedding space, in which closeness represents the semantic similarity between image and text. Typically, a two-branch network is designed to learn the projections of image and text into a common embedding space via metric learning [5,37,45]. Existing works along this line of research generally study the design of network architectures [4,19] or the formulation of learning constraints [40,5,37,45]. Compared to these works, JVSM has the advantage of using the semantically meaningful association of image and text as a form of auxiliary supervision to guide the learning of another task, such as language-guided retrieval. Instead of building multiple task-specific models inefficiently, JVSM underpins a task-agnostic retrieval interface to flexibly ingest various forms of information (e.g., image, text, or their combination), which is the first attempt in the literature.

Learning Using Privileged Information [33,32] is originally proposed as a learning paradigm to use additional information only available at training time with the purpose of improving model performance on related tasks [30,17,20,21,12,41,7,16,18,2]. It is first considered for image retrieval [30] and web image recognition [20,8] based upon the SVM+ formulation, but now is ubiquitous in many machine learning models. Most deep learning models use some kind of privileged information from a secondary task to guide the learning of a model for the primary task. To mention a sample of recent methods, Hoffman et al. [12] use depth images to guide the learning of RGB image representation for object detection. Yang et al. [41] leverage on bounding boxes and image captions for multi-object recognition. Lee et al. [18] use labelled synthetic images to constrain the learning on unlabelled real-world images for semantic segmentation. We propose to use text associated to images in form of attribute-like descriptions as privilege information to constrain the solution space for image-text compositions. Rather than train extra privileged networks heavily as previous works, JVSM retains the same model size to be more efficiently trained.

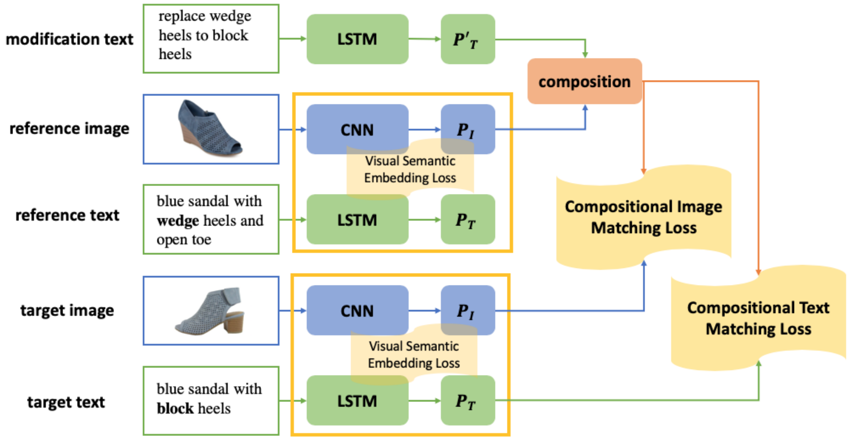
modification text
replace wedge heels to block heels
LSTM
P'T
composition
reference image
CNN
PI
Visual Semantic Embedding Loss
reference text
blue sandal with wedge heels and open toe
LSTM
PT
Compositional Image Matching Loss
target image
CNN
PI
Compositional Text Matching Loss
Visual Semantic Embedding Loss
target text
blue sandal with block heels
LSTM
PT

Fig. 2. Proposed Joint Visual Semantic Matching model. The reference and target image-text pairs are fed into the VSE model while the modification text is fed into an LSTM with individual semantic projection layers. The embedding of the reference image is composed with the modification text as the image-text compositional embeddings, which are further tied synergistically to the image embeddings and text embeddings in a ***common*** embedding space, by jointly optimizing a visual-semantic matching loss (Section 3.1) and two compositional matching loss (Section 3.2). At test time, different types of embeddings are fully comparable in the share space, thus facilitating to process both unimodal (e.g., image, or text) and multimodal (e.g., image with text) inputs for flexible retrieving either images or text descriptions in the database.

3 Proposed Approach

We focus on building a versatile model that tackles a *primary task* of single-turn language-guided retrieval, which facilitates an *auxiliary task* of text-image matching. In our primary task, we are given a reference image and a modification text that describes what content should be modified in the image. The objective of our primary task is to learn an image-text compositional embedding that encodes the information required to retrieve the target image of interest, which should reflect the changes specified by the modification text.

To achieve this goal, we leverage on an auxiliary task of learning a visual semantic embedding space to align image embeddings and text embeddings. In this auxiliary task, we are given an image which is associated to its related text (e.g., attribute-like description: "sandals with block heels"). The objective is to build an embedding space where image and text are close to each other if they represent the same image-text pair, while being far away if they are a negative pair. The auxiliary text which describes the content of the related image is considered as privileged information and it is used exclusively during training of the model to learn a more expressive visual-semantic embedding space that minimises the cross-modal gap between the vision and language domain. As privileged information is not always available for all the examples in the training set, e.g., an image may not have a description associated to it, we propose soft semantic matching to overcome such issue which we discuss in Section 3.2.

JVSM integrates the aforementioned two tasks in a unified *multi-task* learning framework. The proposed model consists of four trainable modules as shown in Figure 2: (1) the visual embedding module (blue CNN blocks), (2) the textual embedding module (green LSTM blocks), (3) the semantic projection modules (P_I, P_T and P'_T blocks), and (4) the compositional module (orange block). For optimization, the model is trained with three loss functions (yellow blocks): (a) the visual semantic embedding loss, (b) the compositional image matching loss, and (c) the compositional text matching loss.

Section 3.1 describes the model components for the auxiliary task of learning a generic visual semantic embedding space used for text-image matching. Section 3.2 describes the components for learning the image-text compositional embedding space used for language-guided retrieval.

3.1 Visual Semantic Embedding

The property that we would like to obtain from learning a Visual Semantic Embedding (VSE) space is to encode the semantic similarity between visual data (i.e., input images) and textual data (i.e., attribute-like descriptions). The main advantage is that pairwise image and text are closely aligned, therefore it enables JVSM to perform text-image matching. To achieve this goal, we construct our VSE model as a two-branch neural networks for image-text matching similarly to [37, 45]. As Figure 2 shows, the VSE model consists of three basic components.

Visual Embedding Module. A standard Convolutional Neural Network (CNN) pre-trained on ImageNet projects the input images to image embeddings. In our experiments, we used MobileNet and remove the classification layer as the backbone network for its quality-speed trade-off.

Textual Embedding Module. This module encodes words from tokenized sentences (attribute-like descriptions or modification text) into text embeddings. We defined it as an LSTM which is trained from scratch. In the case of attribute-like descriptions, sentences are interpreted as privileged information, since it provides additional useful information that is available only during training but not at at test time as discussed in the previous section.

Semantic Projection Layers. The projection layers are responsible to project the image and text embeddings to the common visual-semantic embedding space, where image and text can be compared. P_I and P_T in Figure 2 are defined as linear mappings of the outputs of the visual embedding and textual embedding modules. We define as $\mathbf{v}$ and $\mathbf{t}$ the feature representation of the visual module and textual module after the respective projection modules P_I and P_T.

We train the VSE model by optimizing for the *bi-directional triplet ranking loss* [38], formally defined as follows:

$$L_{vse} = [d(\mathbf{v}, \mathbf{t}) - d(\mathbf{v}, \mathbf{t}^-) + m]_+ + [d(\mathbf{v}, \mathbf{t}) - d(\mathbf{v}^-, \mathbf{t}) + m]_+ \tag{1}$$

where the positive (negative) textual embedding for an image $\mathbf{v}$ is denoted as $\mathbf{t}$ ($\mathbf{t}^-$), the positive (negative) visual embedding for a text $\mathbf{t}$ as $\mathbf{v}$ ($\mathbf{v}^-$), $d(\cdot, \cdot)$

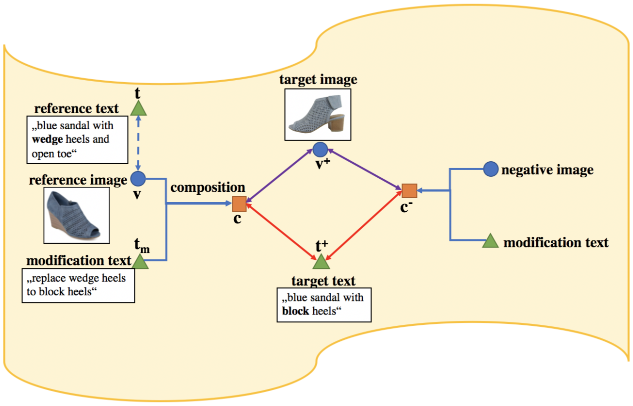

Fig. 3. Sketch of the common visual-semantic embedding space where: 1) the reference text $\mathbf{t}$ and image $\mathbf{v}$ are encouraged to be close to each other; 2) the composition of the image-modification text pair $\mathbf{c}$ on the left side is encouraged to be close to the target image $\mathbf{v}^+$ and text $\mathbf{t}^+$; 3) the negative composition $\mathbf{c}^-$ is encouraged to be far from the target image $\mathbf{v}^+$ and text $\mathbf{t}^+$.

denotes the L2 distance, $[\cdot]_+$=max$(0, \cdot)$, and m is the margin between positive and negative pairs.

Negative sample selection ($\mathbf{t}^-$ and $\mathbf{v}^-$) plays a fundamental role for training [39, 5]. When using the hardest negative mining method proposed in [5], we observed that the loss only decreases with a very small learning rate, thus leading to slow convergence. Inspired by the robust face embedding learning [29], we adopt mini-batch semi-hard mining with the conditions $d(\mathbf{v}, \mathbf{t}) < d(\mathbf{v}, \mathbf{t}^-)$ for $\mathbf{t}^-$ and $d(\mathbf{v}, \mathbf{t}) < d(\mathbf{v}^-, \mathbf{t})$ for $\mathbf{v}^-$, which select the semi-hard negative samples to ensures more stable and faster convergence.

Remark. The key intuition of introducing the VSE space is to ensure that image and text are semantically tied in a shared embedding space (Figure 3). This is beneficial for further learning an image-text compositional embedding: (1) the two-branch networks are jointly optimized to associate each image with its corresponding semantic information, thus leading to a more discriminative and expressive embedding space; (2) within this VSE space, we can formulate objectives that align image-text compositional embeddings to the visual and textual modalities jointly; (3) it enables JVSM to perform text-image matching, as well as language-guided retrieval of either image or text.

3.2 Image-Text Compositional Embedding

After pre-training the model with the VSE loss, the image-text compositional module has the objective of learning encodings of reference image and the respective modification text to retrieve either the target image or text, which should contain the changes specified by the text.

We encode the reference and target image into the embeddings $\mathbf{v}$ and $\mathbf{v}^+$ using the visual embedding module followed by the semantic projection layer P_I as showed in Figure 2. The modification text is encoded into the vector $\mathbf{t}_m$ via the textual embedding module and a new projection layer P'_T which is initialized with P_T pre-trained using the VSE loss. Optionally, some training examples contains auxiliary privileged information in the form of attribute-like descriptions, which are encoded into $\mathbf{t}$ via the textual embedding module and its semantic projection P_T.

In order to compose the visual and textual representations into new semantic representations that resemble the visual representations of the target image, we use the state-of-the-art Text Image Residual Gating (TIRG) model proposed in [36]. The main advantage of TIRG is that it leverages on gated residual connections to modify the image feature based on the text feature, while retaining the original image feature in the case that the modification text is not important. We define as $\mathbf{c} = f_c(\mathbf{v}, \mathbf{t}_m)$ the compositional embedding, which is the result of applying TIRG $f_c(\cdot, \cdot)$ on the visual embedding $\mathbf{v}$ and the modification text embedding $\mathbf{t}_m$.

We train JVSM using L_{vse} and two proposed loss functions defined on the compositional embedding space: the *compositional image matching loss* and the *compositional text matching loss*. We define the compositional image matching loss as bi-directional triplet ranking loss as follows:

$$L_{im} = [d(\mathbf{c}, \mathbf{v}^+) - d(\mathbf{c}^-, \mathbf{v}^+) + m]_+ + [d(\mathbf{c}, \mathbf{v}^+) - d(\mathbf{c}, \mathbf{v}^-) + m]_+ \quad (2)$$

where $\mathbf{c}^-$ is the negative TIRG composition of the image embedding $\mathbf{v}^-$ and its modification text $\mathbf{t}^-$ selected via semi-hard mining. The goal of Eq. 2 is to encourage alignment between the compositional embedding and the target image, while pushing away other negative compositional and image embeddings.

The compositional text matching loss has access to the privileged information and is defined as follows:

$$L_{tm} = [d(\mathbf{c}, \mathbf{t}^+) - d(\mathbf{c}^-, \mathbf{t}^+) + m]_+ + [d(\mathbf{c}, \mathbf{t}^+) - d(\mathbf{c}, \mathbf{t}^-) + m]_+ \quad (3)$$

The goal of Eq. 3 is to encourage alignment between the compositional embedding and the target text while pushing away other negative compositional and text embeddings.

The final loss function of JVSM is the combination of the VSE loss and the proposed compositional losses: $L = L_{vse} + L_{im} + L_{tm}$. The intuition underlying the proposed loss function is depicted in Figure 3: 1) the reference text $\mathbf{t}$ (privileged information) and image $\mathbf{v}$ are encouraged to be close to each other; 2) the composition of the positive image-modification text pair $\mathbf{c}$ on the left side should be as close as possible to the target image $\mathbf{v}^+$ and text $\mathbf{t}^+$; 3) the composition of the negative image-modification text pair $\mathbf{c}^-$ on the right side should be as far as possible to the target image $\mathbf{v}^+$ and text $\mathbf{t}^+$.

In the case that privileged information (i.e., image and description pairs) is available for every training example, we use the same semi-hard mining procedure for negative sample selection defined for the VSE model in Section 3.1.

However, in many applications it is often the case that privileged information, used in the compositional text matching loss, is not available for all training examples (e.g., the Fashion-iq dataset). In order to overcome such issue, we propose a *soft semantic matching* procedure. First we sample a minibatch of k sentences $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k\}$, and compute the set of distances between $\mathbf{c}$ and every $\mathbf{t}_i$, that is, $d_{1;i} = ||\mathbf{c} - \mathbf{t}_i||$. As the target image embedding $\mathbf{v}^+$ is supposed to match the missing $\mathbf{t}^+$ under the constraint of Eq. 1, its semantic distances with respect to other sentences can serve as references, measured by $d_{2;i} = ||\mathbf{v}^+ - \mathbf{t}_i||$. By minimizing $L_{tm} = ||d_1 - d_2||^2$, relative distances with respect to sentences are encouraged to be similar for $\mathbf{c}$ and $\mathbf{v}^+$.

Note that semi-hard and soft semantic matching give comparable results when the dataset is fully annotated. However, it is not possible to perform semi-hard matching when image-text pairs are not available for a all training samples (e.g., either the image or text is missing). Therefore using semi-hard matching in this scenario would require to discard a significant percentage of training samples (the ones containing image only and text only) which significantly decreases performance compared to soft semantic matching.

4 Experiments

We study the performance of JVSM against the state-of-the-art on the task of language-guided retrieval of image using three fashion datasets: Fashion-200k [11], UT-Zap50K [43, 42] and Fashion-iq [10]. We explored two types of modification text: 1) provided in the form of attribute-like modifications as proposed in [36] (Fashion-200k and UT-Zap50K) and 2) provided in the form of natural language feedback as presented in [9] (Fashion-iq). Results are measured in terms of the standard recall at K ($R@K$) defined as the percentage of test queries for which we correctly retrieved the targets in the top-K retrieved samples.

In addition, we perform an ablation study to understand how the different losses have impact on the results. Moreover, we show the flexibility of the proposed method on (1) language-guided retrieval of text (a complementary version of language-guided retrieval of image); and (2) text-image matching.

4.1 Implement Details

The backbone CNN of the visual embedding module is mobilenet-v1 pretrained on ImageNet, which represents one of the best trade-offs between quality and speed. This CNN can be easily replaced with more powerful but slower networks. We used the final layer before the classifier, with dimensionality of 1024. We performed data augmentation of the input images consisting of random flipping. As for the textual embedding module, we used a single-layer LSTM with 1024 units. The projections P_I, P_T and P'_T are linear layers with 512 units, that is the dimensionalty of the joint embedding space. The visual embedding module is finetuned, while all other networks are trained from scratch. Training consists of

Table 1. Language-guided retrieval performance (%) on Fashion-200k. * indicates our implementation of TIRG.

Method	R@1	R@10	R@50
Han et al. [11]	6.3	19.9	38.3
Show and Tell [34]	12.3	40.2	61.8
Relationship [28]	13.0	40.5	62.4
FiLM [25]	12.9	39.5	61.9
TIRG [36]	14.1	42.5	63.8
TIRG* [36]	15.1	41.9	62.0
JVSM (ours)	**19.0**	**52.1**	**70.0**

two stages: we first train the VSE model using L_{vse}, and then we train the JVSM model using $L = L_{vse} + L_{im} + L_{tm}$. We empirically found that the proposed two-stage training protocol helps to converge to a better minimum with lower loss value compared to optimizing directly the final loss L.

4.2 Fashion-200k

Fashion-200k [11] is a popular dataset of fashion products consisting of about 200k images. Images are accompanied by 4,404 concepts that were automatically extracted from product descriptions which we used as privileged information. We followed the protocol for creating queries as [11] and modification text as [36]: pairs of products with one word difference are selected as reference-target pairs and therefore the modification text has one word of difference (e.g., “replace blue with yellow”). We used the same experimentation protocol of [36] on how to create training and testing splits.

Table 1 shows the results of our method compared with the most recent state-of-the-art methods which are available in the empirical Odissey of Vo et al. [36]. These numbers indicate that TIRG was the best performing methods on Fashion-200k when compared to other methods in [36]. However, JVSM significantly outperforms TIRG by a margin of +4.9, +9.6 and +6.2 at different recalls. This shows the importance of properly leveraging on privileged information during training.

Denoted as TIRG* in Table 1 is our implementation of TIRG using the same backbone networks as our method. The relative difference between the original version of TIRG and our implementation is marginal. Therefore we can use our implementation of TIRG as reference method to enable the evaluation on the UT-Zap50K and Fashion-iq datasets which were not used in the original paper [36].

Figure 4 shows some qualitative results on Fashion-200k. The first three rows report success cases on the categories dress, jacket and skirt where the provided modification is on color, style and length, respectively. The second last row shows a failure case: JVSM is able to focus on the right concept (from beaded to lace) while preserving the color from the query (black dress). One can notice that the ground truth (ranked by JVSM at position 793) is ambiguous in this case,

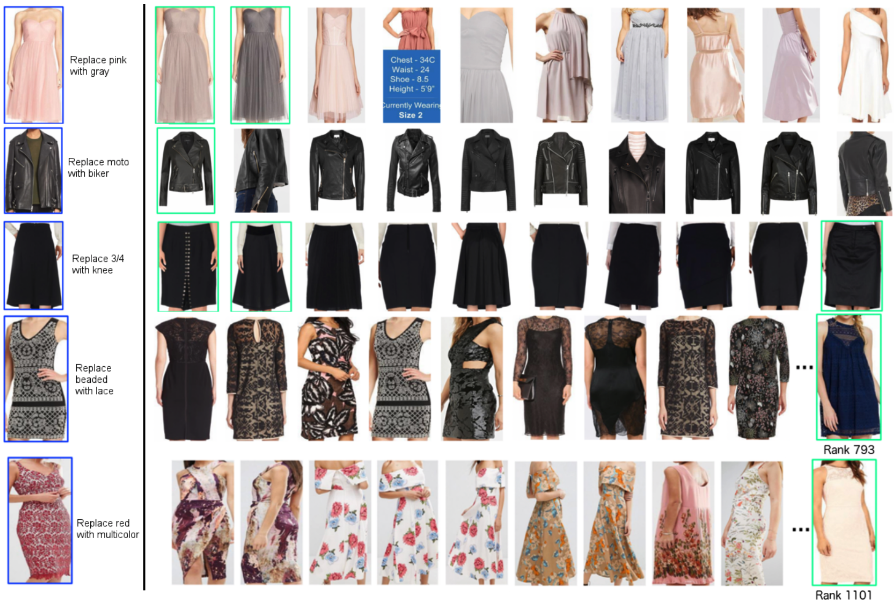

Fig. 4. Qualitative results of language-guided retrieval on Fashion-200k. The query image (blue contour) and modification text are on the left. The retrieved images are on the right and ranked from left to right (ground-truth is in green contour).

because two properties are changed at the same time (color and style) although the change of color was not specified in the modification text. This demonstrates the limitation of the protocol of automatically generating the modification text from single attributes as proposed in [36]. Therefore, it motivates us to carry out a proper evaluation using modification text generated by human annotators as provided by the Fashion-iq dataset [10] (see Section 4.4). The last row shows another failure case where the ground-truth annotation is wrong since the target image is defined as multicolor, while it has clearly a single color. JVSM is anyway able to retrieve relevant multicolor dresses in the first ranks.

4.3 UT-Zap50K

The UT-Zap50K [43, 42] dataset consists of 50,025 images divided in 4 categories (shoes, sandals, slippers and boots), further annotated with 8 fine-grained attribute-like descriptions including category, sub-category, heel height, insole, closure, gender, material and toe style. The dataset was introduced for the task of pairwise comparisons of images, however given the presence of attribute-like annotations, it suits well with the task of language-guided retrieval. Thus it is possible to create the modification text in the same way as described for the Fashion-200k dataset. We generated the training and testing splits with 80% and 20% of the data, respectively.

Table 2. Language-guided retrieval performance (%) on UT-Zap50k. * indicates our implementation of TIRG.

Method	R@1	R@10	R@50
TIRG* [36]	4.5	25.4	56.4
JVSM (ours)	**10.6**	**37.1**	**63.5**

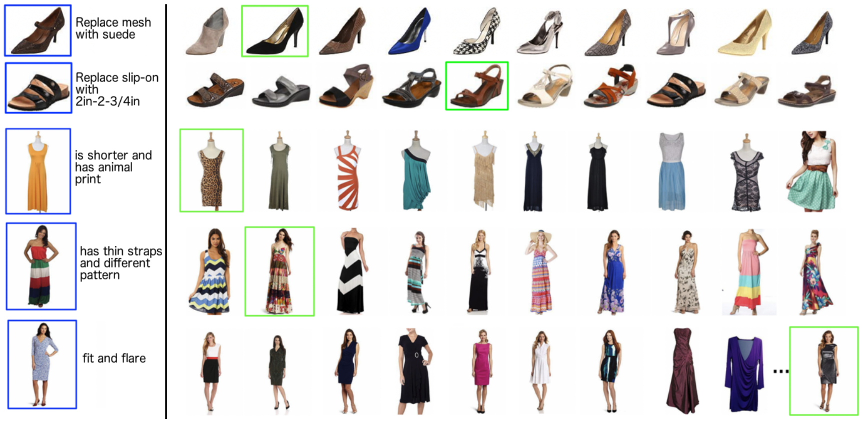
Replace mesh with suede
Replace slip-on with 2in-2-3/4in
is shorter and has animal print
has thin straps and different pattern
fit and flare
...

Fig. 5. Qualitative results of language-guided retrieval on UT-Zap50K (first two rows) and Fashion-iq (last three rows). See text for comments on the results.

Table 2 shows the results of the proposed method in comparison with the approach which was best performing on the Fashion-200k dataset. Our method outperforms TIRG by a margin of +6.1, +11.7 and +7.1 at different recalls. Leveraging privileged information during training is the key to such improvement. Figure 5 (first two rows) shows the qualitative results on high heels shoes and sandals where the provided modification text is on material and style, respectively. The target images are on second and fifth position for the two cases. In addition, JVSM is able to retrieve relevant images for the given modifications on other ranks too.

4.4 Fashion-iq

The Fashion-iq dataset [10] was proposed for multi-turn dialog-based image retrieval. It consists of 77,684 images of 3 categories (dress, shirt and top&tee). A subset of 49,464 images are annotated with side information derived from product descriptions, i.e., attributes, which we use as privileged information. Moreover, 60,272 pairs of images are also annotated with relative captions, which are natural language descriptions of the difference between reference and target images. Therefore, they can be used as modification text to retrieve a target image given the reference image and the relative caption. Since not every image is annotated with attribute information, it becomes important to use the soft semantic matching procedure presented in Section 3.2, otherwise the size of the

Table 3. Language-guided retrieval performance (%) on Fashion-iq. * indicates our implementation of TIRG.

Method	Dress		Shirt		Toptee	
	R@10	R@50	R@10	R@50	R@10	R@50
TIRG* [36]	7.3	18.1	10.1	21.8	10.5	23.8
1-turn [10]	7.7	23.9	5.0	17.3	5.2	17.3
JVSM (ours)	**10.7**	**25.9**	**12.0**	**27.1**	**13.0**	**26.9**

training set would be significantly smaller thus affecting the results. We used the training and validation splits proposed in [10] and we train a JVSM model for each individual category.

The results of the Fashion-iq dataset are reported in Table 3. We compare JVSM with TIRG* and report the results from the paper which introduced the dataset [10], named "1-turn". Note that we do not include results for multiple turns, since our model is neither trained nor adapted to perform multi-turn dialog-based retrieval and thus it would an unfair comparison. The proposed method outperforms both methods on all three categories, showing the effectiveness of our approach on natural language-based modifications.

Figure 5 (last three rows) shows the qualitative results for the dress category. Since modification text is created by human annotators, one can notice that they are more realistic and expressive (multiple modifications) compared to the Fashion-200k and UT-Zap50K datasets, where a single attribute at a time was modified. This setup is closer to a real-world scenario where the user is allowed to express the modifications in textual form, which can include abstract concepts. JVSM is able to learn multiple and more articulated modifications, such as the concept of "animal print" and "different pattern" in the third row and the forth row of Figure 5, which was not possible on other datasets. A failure case is shown in the last row of Figure 5. In this case the modification text ("fit and flare") includes quite a broad list of solutions, in which JVSM is able to capture a subset of them (e.g., at rank 4 and 6), however not the one labelled by the annotator.

4.5 Ablation Study and Other Tasks

In this section, we explore the advantages of using the proposed compositional losses in JVSM by an ablation study. Importantly, we show the flexibility of JVSM, trained for language-guided retrieval of image, to perform (1) language-guided retrieval of text, and (2) text-image matching (that is, text-to-image retrieval). In the case of language-guided retrieval of text, we are given the same inputs as our primary task of language-guided retrieval of image, however we retrieve the target text descriptions accompanied the target images. In the case of text-to-image retrieval, we are given a sentence which is encoded into its textual embedding and used to retrieve the most similar visual embeddings. Potentially, JVSM is able to perform other tasks in the joint space (e.g., retrieve images given an image, or retrieve compositions given an image). We did not explore them due to the lack of relevant groundtruth in test set.

Table 4. Ablation study on Fashion200k showing different tasks (see text for details).

Method	Language-guided retrieval of image			Language-guided retrieval of text			Text-to-image retrieval		
	R@1	R@10	R@50	R@1	R@10	R@50	R@1	R@10	R@50
baseline TIRG*(L_{im})	15.1	41.9	62.0	18.1	32.3	51.8	-	-	-
baseline VSE (L_{vse})	-	-	-	-	-	-	22.7	48.7	69.4
$L_{vse} + L_{im}$	15.6	44.0	63.7	34.2	46.9	65.5	21.3	49.8	70.4
$L_{vse} + L_{im} + L_{tm}$	**19.0**	**52.1**	**70.0**	**50.4**	**66.7**	**82.9**	**23.4**	**51.7**	**72.4**

Table 4 reports our ablation study on Fashion-200k. We trained different models: (1) the baseline L_{im} for language-guided retrieval of either image or text, or L_{vse} for image-text matching; (2) $L_{vse} + L_{im}$; and (3) our final loss $L_{vse} + L_{im} + L_{tm}$. We can notice that adding L_{im} to L_{vse} improves the results (+2.1% of R@10) for the task of language-guided retrieval of images (first 3 columns). We obtain a more significant improvement when further adding L_{tm} (+3.4% of R@1). This demonstrates the benefits of introducing an auxiliary task and the use of privilege information.

Table 4 (middle 3 columns) reports the results for language-guided retrieval of text. We observe a significant improvement by adding our loss components. In addition, we find that language-guided retrieval of text is more effective than language-guided retrieval of image. This result is expected: when a rich textual description of the image is available, text is more discriminative than image due to the concrete language semantics specified discretely.

Table 4 (last 3 columns) reports the results for text-to-image retrieval. The results show a similar behavior that we have seen for language-guided retrieval tasks with improvements by adding both compositional losses. It is worth noting that the primary task helps the auxiliary task of text-image matching. We think that textual modifications encode how two images differ, and thus this relative information helps to reshape the embedding space to be more discriminative.

5 Conclusion

We presented a novel multi-task model: JVSM, which to the best of our knowledge is the first attempt to construct a visual-semantic embedding space and compositional functions that allow to compose image and modification text. JVSM underpins a user-friendly retrieval interface to perform both language-guided retrieval of either image or text, and text-image matching. We demonstrated the benefits of JVSM with respect to the state-of-the-art methods by conducting a comprehensive evaluation on the fashion domain achieving new state-of-the-art for language-guided retrieval, and provided interesting observation in multiple retrieval tasks. Promising future directions include learning spatial-aware image-text embeddings, and integrating various forms of interaction (e.g., clicks or sketches) to learn multimodal embeddings.

References

1. Ak, K.E., Kassim, A.A., Hwee Lim, J., Yew Tham, J.: Learning attribute representations with localization for flexible fashion search. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
2. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S.: Text2shape: Generating shapes from natural language by learning joint embeddings. In: Asian Conference on Computer Vision (2018)
3. Chen, Y., Gong, S., Bazzani, L.: Image search with text feedback by visiolinguistic attention learning. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
4. Engilberge, M., Chevallier, L., Pérez, P., Cord, M.: Finding beans in burgers: Deep semantic-visual embedding with localization. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
5. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
6. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems (2013)
7. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. In: European Conference on Computer Vision (2018)
8. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: 2010 IEEE Computer society conference on computer vision and pattern recognition. pp. 902–909. IEEE (2010)
9. Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., Feris, R.: Dialog-based interactive image retrieval. In: Advances in Neural Information Processing Systems. pp. 678–688 (2018)
10. Guo, X., Wu, H., Gao, Y., Rennie, S., Feris, R.: The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. arXiv preprint arXiv:1905.12794 (2019)
11. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: IEEE International Conference on Computer Vision (2017)
12. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
13. Hosseinzadeh, M., Wang, Y.: Composed query image retrieval using locally bounded features. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
14. Kato, K., Li, Y., Gupta, A.: Compositional learning for human object interaction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 234–251 (2018)
15. Kovashka, A., Parikh, D., Grauman, K.: Whittlesearch: Image search with relative attribute feedback. In: IEEE International Conference on Computer Vision (2012)
16. Lambert, J., Sener, O., Savarese, S.: Deep learning under privileged information using heteroscedastic dropout. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
17. Lapin, M., Hein, M., Schiele, B.: Learning using privileged information: Svm+ and weighted svm. Neural Networks **53**, 95–108 (2014)

18. Lee, K.H., Ros, G., Li, J., Gaidon, A.: Spigan: Privileged adversarial learning from simulation. In: International Conference on Learning Representation (2018)
19. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: European Conference on Computer Vision (2018)
20. Li, W., Niu, L., Xu, D.: Exploiting privileged information from web data for image categorization. In: European Conference on Computer Vision (2014)
21. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. In: International Conference on Learning Representation (2015)
22. Mai, L., Jin, H., Lin, Z., Fang, C., Brandt, J., Liu, F.: Spatial-semantic image search by visual feature synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
23. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1792–1801 (2017)
24. Nagarajan, T., Grauman, K.: Attributes as operators. In: European Conference on Computer Vision (2018)
25. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence (2018)
26. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Transactions on circuits and systems for video technology **8**(5), 644–655 (1998)
27. Sadeh, G., Fritz, L., Shalev, G., Oks, E.: Joint visual-textual embedding for multimodal style search. arXiv preprint arXiv:1906.06620 (2019)
28. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: Advances in Neural Information Processing Systems 30, pp. 4967–4976 (2017), http://papers.nips.cc/paper/7082-a-simple-neural-network-module-for-relational-reasoning.pdf
29. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
30. Sharmanska, V., Quadrianto, N., Lampert, C.H.: Learning to rank using privileged information. In: IEEE International Conference on Computer Vision (2013)
31. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: Advances in Neural Information Processing Systems (2019)
32. Vapnik, V., Izmailov, R.: Learning using privileged information: similarity control and knowledge transfer. The Journal of Machine Learning Research (2015)
33. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks (2009)
34. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
35. Vo, N., Jiang, L., Hays, J.: Let's transfer transformations of shared semantic representations. arXiv preprint arXiv:1903.00793 (2019)
36. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval - an empirical odyssey. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)

37. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
38. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
39. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2840–2848 (2017)
40. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
41. Yang, H., Tianyi Zhou, J., Cai, J., Soon Ong, Y.: Miml-fcn+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
42. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: IEEE Conference on Computer Vision and Pattern Recognition (Jun 2014)
43. Yu, A., Grauman, K.: Semantic jitter: Dense supervision for visual comparisons via synthetic images. In: IEEE International Conference on Computer Vision (Oct 2017)
44. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
45. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: European Conference on Computer Vision (2018)
46. Zhao, B., Feng, J., Wu, X., Yan, S.: Memory-augmented attribute manipulation networks for interactive fashion search. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)