Visual-Semantic Alignment Across Domains Using a Semi-Supervised Approach

Angelo Carraggi, Marcella Cornia(✉), Lorenzo Baraldi, and Rita Cucchiara

University of Modena and Reggio Emilia, Modena, Italy
angelo.carraggi@gmail.com,
{marcella.cornia,lorenzo.baraldi,rita.cucchiara}@unimore.it

Abstract. Visual-semantic embeddings have been extensively used as a powerful model for cross-modal retrieval of images and sentences. In this setting, data coming from different modalities can be projected in a common embedding space, in which distances can be used to infer the similarity between pairs of images and sentences. While this approach has shown impressive performances on fully supervised settings, its application to semi-supervised scenarios has been rarely investigated. In this paper we propose a domain adaptation model for cross-modal retrieval, in which the knowledge learned from a supervised dataset can be transferred on a target dataset in which the pairing between images and sentences is not known, or not useful for training due to the limited size of the set. Experiments are performed on two target unsupervised scenarios, respectively related to the fashion and cultural heritage domain. Results show that our model is able to effectively transfer the knowledge learned on ordinary visual-semantic datasets, achieving promising results. As an additional contribution, we collect and release the dataset used for the cultural heritage domain.

Keywords: Multi-modal retrieval · Visual-semantic embeddings
Semi-supervised learning

1 Introduction

Computer Vision and Natural Language Processing communities are converging toward unified approaches for pattern recognition problems, like providing descriptive feature vectors and finding cross-modality embedding spaces. As a matter of fact, architectures such as VGG [24] and ResNet [9] have been exploited for extracting representations from images, and word embeddings [2,19,22] are now a popular strategy for doing the same with text. The construction of common embeddings, on the other hand, has been proposed for solving tasks in which a connection between language and vision is needed, like automatic captioning [4,5,13] and retrieval of images and textual descriptions [1,8,11,20,27]: in this case, data from both modalities can be projected in the common space,

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11134, pp. 625–640, 2019.
https://doi.org/10.1007/978-3-030-11024-6_47

and retrieved according to distances in the embedding. While the supervised training of a common visual-semantic embedding is feasible when using sufficiently large datasets, those techniques are unlikely applicable in the case of small scale datasets, or when the pairing between visual and textual elements is not provided. In both cases, it is beneficial to transfer the knowledge learned on large-scale datasets by using domain adaptation techniques.

Following this line of research, in this paper we propose a semi-supervised model for learning visual-semantic embeddings. Given a source dataset, in which the pairing between images and captions is known, our model is able to transfer its knowledge to a target domain, in which the pairing between the modalities is either not known in advance, or not useful for learning due to the restricted size of the set. The proposed model is based on a novel combination of visual and textual auto-encoders, embedding space learning strategies and domain alignment techniques. Specifically, two auto-encoders are trained, respectively for visual and textual data, and their intermediate representations are employed as features for training the visual-semantic embedding. The alignment between the distributions of the two modalities in the common embedding space ensures that the learned representations are general enough to be applied to the target domain.

We conduct experiments by using different source and target datasets. In particular, we test our model by transferring the knowledge learned on ordinary visual-semantic datasets to the case of fashion images and to the case of cultural heritage images. Preliminary analyses will showcase the distance between the source and target distributions, while experimental results will demonstrate the capabilities of the proposed approach, in comparison with two baselines which are built by ablating the core components of the method. As a complementary contribution, we collected and annotated the visual-semantic dataset used for the domain of cultural heritage.

To sum up, the contributions of this paper are threefold: (i) we propose a semi-supervised visual-semantic model which can transfer the knowledge learned on a source domain to a target, unsupervised, domain. To the best of our knowledge, we are the first to tackle this setting in the case of a visual-semantic embedding model. (ii) Secondly, we extensively evaluate our model under different settings and by using two different target domains, namely the fashion and cultural heritage domains. Experimental results will show that the proposed approach is able to outperform carefully designed baselines, and that the contributions provided by each of the components of the model are essential for gaining the final performance. (iii) Finally, we collect and release the visual-semantic dataset for cultural heritage used in this work.

2 Related Work

Matching visual data and natural language is a core challenge in computer vision and multimedia. Since visual and textual data belong to two distinct modalities, the problem is typically addressed by constructing a common visual-semantic

embedding space in which images and corresponding sentences can be projected and compared. The retrieval, in this case, is then carried out by measuring distances inside the joint space, which should be low for matching text-image pairs and higher for non-matching pairs.

Following this line of work, Kiros *et al.* [14] introduced an encoder-decoder model capable of learning a common representation for images and text from which cross-modal retrieval can be effectively performed. Several other image and text matching methods have been proposed [7,8,11,20,27]. In particular, Faghri *et al.* [8] extended the method in [14] by exploiting the use of hard negatives and proposed a simple modification of standard loss functions obtaining a significant improvement in cross-modal retrieval performance. Wang *et al.* [27], instead, tackled the image-text matching problem using a two branch network. The network architecture consists of an embedding branch and a similarity network: while the embedding branch translates image and text into a feature representation, the similarity network decides how well the feature representations match, using logistic loss. On a different note, Dong *et al.* [6] proposed to search the visual space directly, instead of seeking a joint subspace for image and video caption retrieval. To this end, they introduced a deep neural model that encodes input captions into a multi-scale sentence embedding and transfers them into a visual feature space.

All of these methods have been proved to be effective to solve the cross-modal retrieval task, when trained with the supervision of a large dataset. None of them, however, addressed the problem in an unsupervised or semi-supervised setting. In this paper, instead, we are interested in adapting the knowledge learned on a given set of data (*i.e.* the source domain) to align images and text belonging to a different domain (*i.e.* the target domain), without directly training the network on the target domain. This solution, which is well known as domain adaptation, has been adopted in a wide variety of settings such as image classification [17], image-to-image translation [10], object detection [12], image captioning [3] and semantic segmentation [29]. Typically, it is addressed by minimizing the distance between feature space statistics of the source and target, or by using domain adversarial objectives where a domain classifier is trained to distinguish between the source and target representations.

Even though domain adaptation has been demonstrated to be effective for different computer vision and multimedia tasks, it has yet to be explored in the context of aligning images and corresponding sentences. Probably, the most important related method is that introduced in [26] which presents a semi-supervised approach to classify input images with the corresponding textual attributes. On the contrary, we aim at encoding entire sentences instead of textual attributes and at directly aligning them with the corresponding input image by addressing the cross-modal retrieval problem in a semi-supervised way.

3 Proposed Method

We propose a semi-supervised visual-semantic model which is capable of aligning images and text. In contrast to supervised cross-modal models, our proposal

does not need a paired training set, in which the associations between images and captions are known in advance, but rather transfers the knowledge learned on a source annotated dataset to a target dataset in which the pairing between images and captions is unknown at training time.

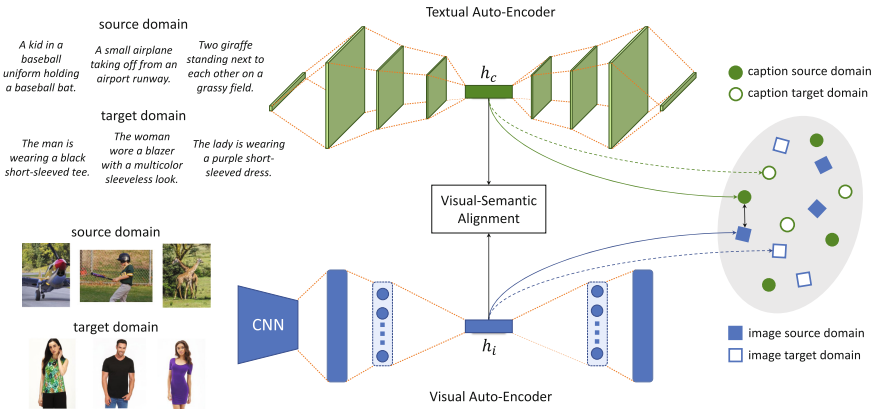


Fig. 1. Overview of our model. Two auto-encoders process visual and textual data and produce an intermediate representation for both modalities. These representations can be used to create a common embedding space in which images and corresponding sentences can be projected and compared. A semi-supervised visual-semantic alignment is exploited to match images and captions coming from a target domain, different from that used to train the model. (Color figure online)

The key element of our proposal is a network which can extract informative, discriminant and domain-invariant representations for both visual and textual data. Given a textual or visual input, this is processed by an auto-encoder which, through its reconstruction loss, naturally enforces the informativeness of its intermediate representation. Additional soft-constraints are then applied to the representation given by the auto-encoder, to ensure that the remaining desirable properties are met. Features extracted from the auto-encoder are employed to project the inputs in a joint visual-semantic embedding space, which can be trained on the source domain, so to ensure that the representation is also discriminant for cross-modal retrieval. Finally, the domain invariance of the features is enforced by applying alignment cost function between images and captions in the source and target domain. For the ease of the reader, we depict the overall architecture of the model in Fig. 1.

3.1 Textual Auto-Encoder

Recently, convolutional-based approaches for text representation have achieved competitive results in comparison to models based on recurrent neural networks [23,30]. This approach also features the additional benefit of being computationally friendly, as recurrent dependencies are removed and convolutions can

be easily parallelized. Following this line of research, we develop an encoder-decoder model based on a purely convolutional network. The auto-encoder converts variable-length captions to fixed-length representations from which input sentences can be reconstructed. In particular, our model exploits 2-d convolutional layers for encoding an input sentence and deconvolutional layers (*i.e.* transpose convolutions) to decode from a hidden representation, without relying on a recurrent architecture.

For sentence encoding, we take inspiration from the architecture proposed in [23], in which the reduction in length carried out by convolutions is exploited to project the input into a representation with lower dimensionality. Furthermore, padding is exploited to process captions with variable length, without affecting the final performance. Given a caption c, each word $\mathbf{w}^t$ is embedded into a k-dimensional word vector $\mathbf{x}^t = \mathbf{W}_e[\mathbf{w}^t]$, where $\mathbf{W}_e$ is a learned word embedding matrix, normalized so that each word embedding has unit ℓ_2-norm. A sentence of length $T^{(0)}$ is obtained by stacking word embeddings $\mathbf{x}^t$ and padding the resulting matrix when necessary, thus obtaining a structure on which 2-d convolutions can be applied.

The input sequence is then fed to a network with N convolutional layers, where each of them reduces the length $T^{(n)}$ of its input to

$$T^{(n+1)} = \left\lfloor \frac{T^{(n)} - z}{r^{(n)}} + 1 \right\rfloor, \tag{1}$$

where $r^{(n)}$ is the stride of the n-th convolutional layer along the time dimension and z is the filter size. The output of the last convolutional layer is the intermediate representation vector $\mathbf{h}_c$ of the textual auto-encoder. This is obtained by using a convolutional layer with filter size equal to $T^{(n-1)}$ thus obtaining a vector that encapsulates the input sentence sub-structures.

For the decoding phase, we exploit strided deconvolutional layers to reconstruct the original sentence starting from $\mathbf{h}_c$. The decoder is composed of N layers that symmetrically increase the spatial size of the output by mirroring the corresponding convolutional layer of the encoder model. The output of the last layer of the decoder aims at reproducing the word embedding vector of each word of the original caption.

Denoting with $\hat{\mathbf{w}}^t$ the t-th word in the reconstructed caption $\hat{c}$, the probability of $\hat{\mathbf{w}}^t$ to be word v is defined as

$$p(\hat{\mathbf{w}}^t = v) = \frac{\exp[\tau^{-1} D_{\cos}(\hat{\mathbf{x}}^t, \mathbf{W}_e[v])]}{\sum_{v' \in V} \exp[\tau^{-1} D_{\cos}(\hat{\mathbf{x}}^t, \mathbf{W}_e[v'])]}, \tag{2}$$

where $D_{\cos}$ is the cosine similarity function, τ is a positive value representing the temperature parameter [30], $\hat{\mathbf{x}}^t$ is the reconstructed word embedding vector of the t-th word, and V is the vocabulary. Note that the cosine similarity can be obtained as the inner product between $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}^0, \hat{\mathbf{x}}^1, ..., \hat{\mathbf{x}}^T\}$ and $\mathbf{W}_e$, since both matrices are ℓ_2-normed.

The overall loss function of the convolutional auto-encoder can be defined, for an input caption c, as the negative word-wise log-likelihood

$$\mathcal{L}_{\mathrm{AE}}^{\mathrm{c}}(c) = -\sum_{t} \log p(\hat{\mathbf{w}}^t = \mathbf{w}^t). \tag{3}$$

3.2 Visual Auto-Encoder

Given the auto-encoder for the textual part, we want to represent visual data in a similar way. In particular, we build an encoder-decoder model that can take an image feature vector as input and reconstruct it starting from an intermediate and more compact representation.

In detail, given an input image, we extract a feature vector from a pre-trained CNN and we feed it to an encoder model composed of a single fully connected layer. We indeed notice that a single layer leads to have a fairly informative representation of the image feature vector. Formally, let i be the input image and $\Phi(i)$ be the corresponding feature vector coming from the pre-trained convolutional network. We define the output of the encoder model $\mathbf{h}_i$ (*i.e.* the intermediate representation of the input image) as

$$\mathbf{h}_i = \tanh(W_e \Phi(i) + b_e), \tag{4}$$

where W_e and b_e are, respectively, the weight matrix and the bias vector of the encoder. Note that the output of the encoder layer is fed through a tanh non-linearity activation function.

The decoder model has a symmetric structure with respect to the encoder model. Therefore, starting from the intermediate vector $\mathbf{h}_i$, the decoder is composed by a single fully connected layer that transforms $\mathbf{h}_i$ to the size of the input image feature vector. Formally, the reconstructed image feature vector $\hat{i}$ is defined according to

$$\hat{i} = W_d \mathbf{h}_i + b_d, \tag{5}$$

where W_d and b_d are the weight matrix and the bias vector of the decoder fully connected layer. Overall, our image auto-encoder is trained to minimize the reconstruction error for each input image. Therefore, we define the decoder loss function as the mean square error between the original image feature vector $\Phi(i)$ and the corresponding reconstruction $\hat{i}$, as follows

$$\mathcal{L}_{\mathrm{AE}}^{\mathrm{i}}(i) = \|\hat{i} - \Phi(i)\|^2. \tag{6}$$

3.3 Visual-Semantic Embedding Space

The task of aligning images and corresponding sentences requires the ability to compare visual and textual data and to have a common representation of both domains. Therefore, we adopt the strategy of creating a joint visual-semantic embedding space in which visual and textual data can be projected and compared using a distance function.

Let $\mathbf{h}_i$ be the image representation coming from the encoder of the visual auto-encoder and $\mathbf{h}_c$ the corresponding textual representation coming from the convolutional auto-encoder for text. These representations can be compared in a joint embedding space by computing the cosine similarity between $\mathbf{h}_i$ and $\mathbf{h}_c$, so that the similarity between an image i and a caption c becomes

$$s(i,c) = \frac{\langle \mathbf{h}_i, \mathbf{h}_c \rangle}{\|\mathbf{h}_i\| \|\mathbf{h}_c\|}, \tag{7}$$

where, in the above formula, $\mathbf{h}_i$ and $\mathbf{h}_c$ are ℓ_2-normed to have the embedding space lying on the ℓ_2 ball.

In order to learn an embedding space with suitable cross-modal properties, we train this space according to a hinge triplet ranking loss with margin α, commonly used in image-text retrieval [8,14]:

$$\begin{aligned} \mathcal{L}_{\text{SH}}(i,c) = &\sum_{\bar{c}} [\alpha - s(i,c) + s(i,\bar{c})]_+ \\ &+ \sum_{\bar{i}} [\alpha - s(\bar{i},c) + s(i,c)]_+ \end{aligned} \tag{8}$$

where $[x]_+ = \max(0,x)$. The loss defined above comprises two symmetric terms: the first sum is taken over all negative captions $\bar{c}$ given the query image i (*i.e.* all captions that do not describe the content of i), while the other is taken over all negative images $\bar{i}$ given the query c (*i.e.* all images that do not correspond to the description reported in c). In practice, given the size of the dataset and the number of possible negative samples, the sums of Eq. 8 are taken only inside the single mini-batch.

3.4 Aligning Distributions

In order to learn relationships between visual and textual features which can be exploited in a target unsupervised domain, we use domain alignment techniques. In particular, the distributions of text and images are aligned in the common embedding space through the Maximum Mean Discrepancy (MMD) criterion. The same alignment is applied to data coming from both the source and target domain, so that the MMD criterion, together with the triplet ranking loss, implicitly enforces an alignment between text and data coming from the target unsupervised domain.

MMD, in our case, can be viewed as a two-sample test between the distributions of text and images in the embedding space, and its loss can be defined as:

$$\mathcal{L}_{\text{MMD}} = ||E_p[\xi(\mathbf{h}_i)] - E_q[\xi(\mathbf{h}_c)]||^2_{\mathcal{H}_k} \tag{9}$$

where p and q are, respectively, the distributions of the visual and textual embeddings (*i.e.*, $\mathbf{h}_i \sim p$ and $\mathbf{h}_c \sim q$) coming from both the source and target domain, ξ is a feature map defined through a kernel k, $\xi(\mathbf{x}) = k(\mathbf{x}, \cdot)$, and $\mathcal{H}_k$ is the

reproducing kernel Hilbert space of k. The kernel is empirically chosen to be a Gaussian kernel, defined as follows:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\sigma^2}||\mathbf{x} - \mathbf{x}'||^2\right) \tag{10}$$

The MMD loss is minimized to shrink the gap between visual and textual features for the supervised and unsupervised datasets. Experimental results, which will be presented in the remainder of the paper, will show that the MMD loss helps to improve the model performance on the target domain.

3.5 Training

Our training protocol aims at learning the feature representations, the alignment and the visual semantic embedding jointly from scratch. Therefore, we minimize all the objective functions defined above at the same time. Recalling that $\mathcal{L}^i_{\text{AE}}$ is the loss function for the auto-encoder on the visual domain and $\mathcal{L}^c_{\text{AE}}$ is the loss function for the auto-encoder on the textual domain, we define a joint loss function for feature learning which is applied to both the source and target domain:

$$\begin{aligned} \mathcal{J}(i,c) &= \mathcal{L}^{\text{i}}_{\text{AE}}(i) + \mathcal{L}^{\text{c}}_{\text{AE}}(c) \\ \mathcal{L}_{\text{AE}} &= \sum_{i,c \in \mathcal{S}} \mathcal{J}(i,c) + \sum_{i,c \in \mathcal{T}} \mathcal{J}(i,c), \end{aligned} \tag{11}$$

where $\mathcal{S}$ and $\mathcal{T}$ are respectively the source and target datasets. Finally, we obtain the loss function $\mathcal{L}$ for our model as:

$$\mathcal{L} = \mathcal{L}_{\text{AE}} + \mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{SH}}, \tag{12}$$

where $\mathcal{L}_{\text{MMD}}$ is the Maximum Mean Discrepancy function and $\mathcal{L}_{\text{SH}}$ is the ranking loss (applied only on the source domain). The loss is then minimized by backpropagation through Stochastic Gradient Descent.

4 Experimental Evaluation

In this section, in addition to describing employed datasets and implementation details, we provide extensive analyses and experiments to validate the proposed visual-semantic alignment model.

4.1 Datasets

For evaluating the effectiveness of our proposal, we perform experiments on different datasets. In particular, we employ two common visual-semantic datasets as source sets, and select two different target domains: fashion and artworks images.

A young man with shoulder length hair, who is wearing a dark jacket with a white collar.

A man who is wearing a golden armour is sitting on a throne and is surrounded by a large group of people.

A woman wearing a blue dress and a red cloak is sitting on a throne holding a naked baby and a book in her hands with two angels at the sides and a city in background.

A man with a light-blue cloak and a dog are looking to an angel over them.

A man with a white beard and a brown tunic is holding a crucifix.

A woman who is pointing the decapitated head of a man held by another man. On the right there are two men.

Fig. 2. Sample image-caption pairs from the EsteArtworks dataset. (Color figure online)

As source datasets, we use Flickr30K [28] and Microsoft COCO [15], which contain natural images and corresponding textual descriptions. Flickr30K is composed by 31,000 images, while COCO contains more than 120,000 images. Each image is annotated with 5 sentences describing the image content. Following the splits defined in [13], for Flickr30K we use 1,000 images for validation, 1,000 images for testing and the rest for training. For Microsoft COCO, instead, we use 5,000 images for both validation and test set.

To evaluate the generalization capabilities of our model, we employ two different target datasets containing image-sentence pairs respectively belonging to the fashion and cultural heritage domain. For the fashion domain, we employ DeepFashion [16], a large-scale publicly available dataset composed by over 800,000 fashion images ranging from well-posed shop images to unconstrained consumer photos. Only 78,979 images of this dataset are annotated with the corresponding sentences [31] which describe only the visual facts such as the color and the texture of the clothes or the length of the sleeves. These images are divided in train and test set, respectively composed by 70,000 and 8,979 images. In our experiments, we use 1,000 randomly selected training images as validation set. Following a common practice used for ordinary datasets [8], retrieval results on this dataset are reported by averaging over 8 folds of 1,000 test images each.

For the cultural heritage domain, instead, we collect 553 artworks from the Estense Gallery of Modena, which comprises Italian paintings and sculptures from the fourteenth to the eighteenth centuries. For each artwork, we collect at least one sentence describing the visual content of the artwork itself, without leveraging on personal cultural background regarding the opera or the depicted characters. Overall, we collect $1,278$ textual descriptions. Some image-sentence artwork pairs of our new EsteArtworks dataset are shown in Fig. 2. In our experiments, we split image samples in training, validation and test split according to a 60-20-20 ratio.

4.2 Implementation Details

In our experiments, we set the dimensionality of the intermediate representations for both auto-encoders to 500. For the textual auto-encoder, we set the number of convolutional and deconvolutional layers N to 3 and the word embedding dimensionality to 300. The filter size is set to $z = 4$, while the strides for each layer are set to $r = \{2, 2, 1\}$. For encoding input images, we exploit two popular CNNs: the ResNet-152 [9] and the VGG-19 [24]. In particular, we extract image features from the *fc7* layer of VGG-19 and from the average pooling layer of ResNet-152, thus obtaining an input image feature vector dimensionality of 4096 and 2048, respectively. Since we use a single encoding layer in the visual auto-encoder, its output size is set to 500.

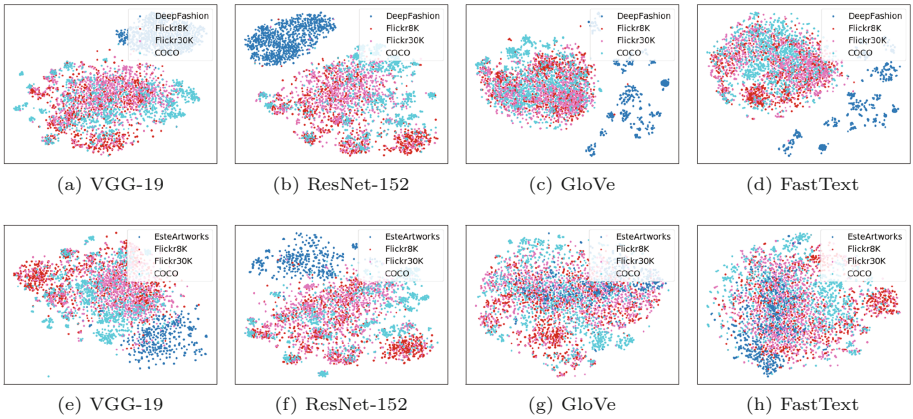
All experiments are performed with mini-batches of size 32 and using the Adam optimizer with an initial learning rate of 2×10^{-4} for 20 epochs, which is then decreased by a factor of 10 for the rest of the training. We set the margin α to 0.2 and the σ parameter of the Gaussian kernel to 1.0.

4.3 Analysis of Dataset Distributions

To get an insight of characteristics of the DeepFashion and EsteArtworks datasets, we analyze the distribution of image and textual features obtained, respectively, from CNNs and word embeddings, and compare them with those extracted from classical visual-semantic datasets.

For the visual part, we extract the activation coming from the *fc7* layer of VGG-19 and the average pooling layer of ResNet-152. For textual counterpart, we embed each word of a caption with a word embedding strategy (*i.e.* GloVe [22] and FastText [2]). To get a feature vector for a sentence, we then sum the ℓ_2 normalized embeddings of the words, and ℓ_2 normalize again the result. This strategy has been largely used in image and video retrieval, and it is known for preserving the information of the original vectors into a compact representation with fixed dimensionality [25].

Figure 3 shows the distributions of visual and textual features of the DeepFashion and EsteArtworks datasets, compared with three ordinary visual-semantic datasets (*i.e.* Flickr8K, Flickr30K and COCO). In order to obtain a suitable two-dimensional representation of a K-dimensional space (with $K = 4096$ for the VGG-19, $K = 2048$ for the ResNet-152 and $K = 300$ for both



(a) VGG-19 (b) ResNet-152 (c) GloVe (d) FastText

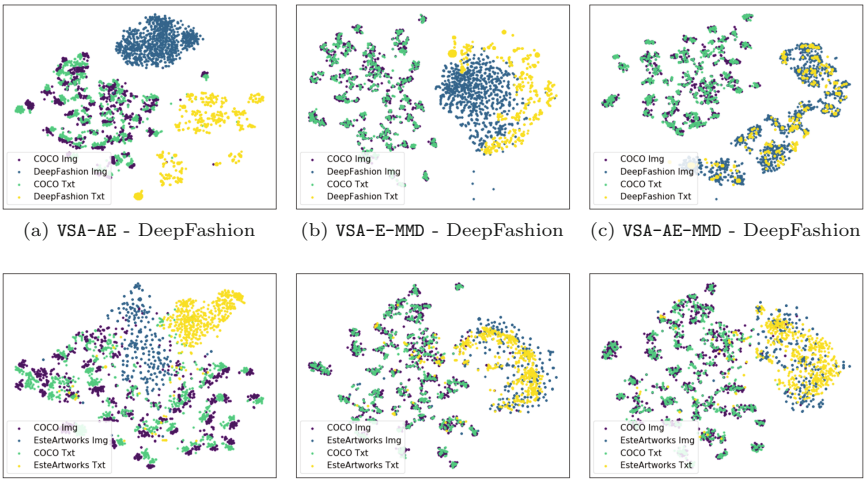
(e) VGG-19 (f) ResNet-152 (g) GloVe (h) FastText

Fig. 3. Comparison between the visual and textual features of ordinary visual-semantic datasets (Flickr8K, Flickr30k, COCO) and those of the DeepFashion (plots a–d) and EsteArtworks (plots e–h) datasets. Visualization is obtained by running the t-SNE algorithm on top of the features. Best seen in color.

Table 1. Cross-domain caption and image retrieval results.

Target	Source	Model	Caption retrieval			Image retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
DeepFashion	Flickr30K	VSA-AE (*VGG-19*)	0.3	1.2	2.1	0.6	2.8	4.7
		VSA-AE (*ResNet-152*)	1.1	3.3	5.2	1.1	4.8	8.0
		VSA-E-MMD (*VGG-19*)	6.1	6.5	6.7	0.6	2.3	4.5
		VSA-E-MMD (*ResNet-152*)	2.0	5.3	6.6	1.0	4.0	6.6
		VSA-AE-MMD (*VGG-19*)	10.7	18.8	25.0	9.5	24.3	33.9
		VSA-AE-MMD (*ResNet-152*)	**13.5**	**23.3**	**30.3**	**10.6**	**27.2**	**38.2**
	COCO	VSA-AE (*VGG-19*)	0.4	1.4	2.3	0.3	1.9	3.9
		VSA-AE (*ResNet-152*)	0.4	1.5	2.7	0.3	2.6	5.3
		VSA-E-MMD (*VGG-19*)	4.6	6.0	6.4	0.4	2.0	3.7
		VSA-E-MMD (*ResNet-152*)	4.6	5.7	6.3	0.3	2.1	3.6
		VSA-AE-MMD (*VGG-19*)	12.9	21.5	27.7	10.7	28.2	37.8
		VSA-AE-MMD (*ResNet-152*)	**18.9**	**25.3**	**30.9**	**11.4**	**28.3**	**38.0**
EsteArtworks	Flickr30K	VSA-AE (*VGG-19*)	3.6	12.7	24.5	4.5	9.1	11.7
		VSA-AE (*ResNet-152*)	10.0	23.6	**39.1**	4.2	11.4	19.3
		VSA-E-MMD (*VGG-19*)	4.5	25.5	32.7	3.8	9.5	17.8
		VSA-E-MMD (*ResNet-152*)	8.2	**28.2**	37.3	6.8	15.5	24.2
		VSA-AE-MMD (*VGG-19*)	8.2	24.5	33.6	7.2	13.3	24.2
		VSA-AE-MMD (*ResNet-152*)	**10.9**	22.7	34.5	**8.0**	**17.8**	**25.0**
	COCO	VSA-AE (*VGG-19*)	2.7	17.3	22.7	3.4	7.6	12.1
		VSA-AE (*ResNet-152*)	9.1	18.2	23.6	3.0	14.0	17.0
		VSA-E-MMD (*VGG-19*)	7.3	19.1	30.0	5.7	11.0	16.3
		VSA-E-MMD (*ResNet-152*)	6.4	21.8	30.0	6.8	14.4	22.0
		VSA-AE-MMD (*VGG-19*)	**10.9**	26.4	37.3	**7.6**	16.3	24.2
		VSA-AE-MMD (*ResNet-152*)	**10.9**	**30.0**	**42.7**	**7.6**	**17.0**	**29.2**

GloVe and FastText word embeddings), we run the t-SNE algorithm [18], which iteratively finds a non-linear projection which preserves pairwise distances from the original space. As it can be seen, both visual and textual distributions of the DeepFashion dataset are very different from those of ordinary datasets which instead almost lay in a single cluster. On the contrary, the EsteArtworks dataset shares some of the properties of ordinary visual-semantic datasets, especially in the textual domain. In fact, using either GloVe or FastText word embeddings, the distribution of this dataset is overlapped with the Flickr and COCO ones, thus highlighting a similarity in the caption style. For the visual part, instead, the distribution shift is more evident while being less separated than DeepFashion features.



(a) `VSA-AE` - DeepFashion (b) `VSA-E-MMD` - DeepFashion (c) `VSA-AE-MMD` - DeepFashion

(d) `VSA-AE` - EsteArtworks (e) `VSA-E-MMD` - EsteArtworks (f) `VSA-AE-MMD` - EsteArtworks

Fig. 4. Visualization of the embedding spaces obtained by two considered baselines (`VSA-AE` and `VSA-E-MMD`) and that of our entire model (`VSA-AE-MMD`). Visualization is obtained by running the t-SNE algorithm on top of the visual and textual embedding vectors by comparing the COCO embedding space with the DeepFashion (plots a–c) and EsteArtworks (plots d–f) ones. Best seen in color.

4.4 Cross-Domain Retrieval Results

To evaluate the results of our model, we report rank-based performance metrics $R@K$ ($K = 1, 5, 10$) for image and caption retrieval. In particular, $R@K$ computes the percentage of test images or test sentences for which at least one correct result is found among the top-K retrieved sentences, in the case of caption retrieval, or the top-K retrieved images, in the case of image retrieval.

In our experiments, we compare the results obtained by our model with two different baselines. The first one is based on the two auto-encoders without the alignment of distributions given by the maximum-mean discrepancy

function defined in Eq. 9. The second one is instead our model without the reconstruction losses for images and corresponding sentences defined in Eqs. 3 and 6 (*i.e.* our model without decoders). In the following, we refer to our complete visual-semantic alignment model as `VSA-AE-MMD`, to the first baseline without the distribution alignment as `VSA-AE` and to the second baseline without reconstruction losses as `VSA-E-MMD`.

(a) DeepFashion results

(b) EsteArtworks results

Fig. 5. Examples of top-1 retrieved images and captions on the DeepFashion and EsteArtworks dataset. (Color figure online)

Table 1 shows the caption and image retrieval results on the two considered target domains when the model is trained on two different ordinary visual-semantic datasets. In particular, we report the results of our model and the two baselines by using both VGG-19 and ResNet-152 networks. As it can be observed, the overall performances of our visual-semantic alignment model are almost always better than those achieved by the two baselines. In particular, on the DeepFashion dataset both reconstruction losses and distribution alignment give a significant contribution to the final performances which overcome by a big margin the baselines. On the EsteArtworks dataset, instead, the gain of the alignment strategy is less evident even though the entire model still obtains a better performance than the two considered baselines. The difference in performance gain on the two datasets can be easily explained by the distribution

analysis reported in Sect. 4.3. In fact, the visual and textual distributions of the EsteArtworks dataset are to some extent similar to those of Flickr30K and COCO, thus justifying the acceptable results even without using the distribution alignment or the reconstruction losses. On the contrary, the low baseline performances on the DeepFashion dataset is explained by the distance between this dataset and ordinary ones, on both visual and textual modalities.

As a further analysis, Fig. 4 shows the embedding spaces obtained by our model, compared with those of the two baselines. To obtain them, we run the t-SNE algorithm on top of the visual and textual embedding vectors (*i.e.* the outputs of the image and caption encoders). As it can be seen, our `VSA-AE-MMD` model leads to a better alignment of visual and textual embeddings on both target datasets. Finally, Fig. 5 reports some qualitative results by showing the top-1 retrieved images and captions on the fashion and cultural heritage domains.

4.5 Text Reconstruction Results

In addition to aligning visual and textual embeddings from two different domains in a semi-supervised way, our model is able to reconstruct the original input caption. To quantify the reconstruction capabilities of the model, we compute machine translation metrics between original and reconstructed sentences. In particular, we employ the BLEU [21] score, which is a modified form of precision between n-grams, to compare a candidate translation against multiple reference translations. Table 2 shows the text reconstruction results on Flickr30K and COCO when forcing the distribution alignment on the two target domains. As it can be seen, our model is able to reconstruct high quality sentences, achieving a BLUE score higher than 0.9 in all considered cases.

Table 2. Text reconstruction results.

Evaluation dataset	Unsupervised domain	BLEU@2	BLEU@3	BLEU@4
Flickr30K	DeepFashion	0.969	0.961	0.952
	EsteArtworks	0.955	0.942	0.928
COCO	DeepFashion	0.991	0.988	0.985
	EsteArtworks	0.988	0.984	0.980

5 Conclusions

In this paper, we addressed the problem of learning visual-semantic embeddings to perform cross-modal retrieval across different domains. In particular, we proposed a semi-supervised model that is able to transfer the knowledge learned on a source dataset to a target domain, where the pairing between images and corresponding sentences is either not known or not useful due to its limited size.

We applied the proposed strategy to two different target domains (*i.e.* fashion and cultural heritage) and we showed through extensive analyses and experiments the effectiveness of the proposed model. As a side contribution, given the lack of visual-semantic datasets for the cultural heritage domain, we collected artworks images and annotated them with the corresponding sentences.

Acknowledgements. This work was supported by the CultMedia project (CTN02_00015_9852246), co-founded by the Italian MIUR. We also acknowledge the support of Facebook AI Research with the donation of the GPUs used for this research.

References

1. Baraldi, L., Cornia, M., Grana, C., Cucchiara, R.: Aligning text and document illustrations: towards visually explainable digital humanities. In: International Conference on Pattern Recognition (2018)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
3. Chen, T.H., Liao, Y.H., Chuang, C.Y., Hsu, W.T., Fu, J., Sun, M.: Show, adapt and tell: adversarial training of cross-domain image captioner. In: IEEE International Conference on Computer Vision (2017)
4. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Visual saliency for image captioning in new multimedia services. In: IEEE International Conference on Multimedia and Expo Workshops (2017)
5. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Paying more attention to saliency: image captioning with saliency and context attention. ACM Trans. Multimedia Comput. Commun. Appl. **14**(2), 48 (2018)
6. Dong, J., Li, X., Snoek, C.G.: Predicting visual features from text for image and video caption retrieval. IEEE Trans. Multimedia **20**, 3377–3388 (2018)
7. Eisenschtat, A., Wolf, L.: Linking image and text with 2-way nets. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
8. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition (2016)
10. Hoffman, J., et al.: CyCADA: cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
11. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal LSTM. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
12. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: IEEE International Conference on Computer Vision and Pattern Recognition (2018)
13. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: IEEE International Conference on Computer Vision and Pattern Recognition (2015)
14. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)

15. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
16. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In: IEEE International Conference on Computer Vision and Pattern Recognition (2016)
17. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: International Conference on Machine Learning (2017)
18. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)
20. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: 40th Annual Meeting on Association for Computational Linguistics (2002)
22. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (2014)
23. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations (2016)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: International Conference on Learning Representations (2016)
26. Tsai, Y.H.H., Huang, L.K., Salakhutdinov, R.: Learning robust visual-semantic embeddings. In: IEEE International Conference on Computer Vision (2017)
27. Wang, L., Li, Y., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Trans. Pattern Anal. Mach. Intell. (2018)
28. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans. Assoc. Computat. Linguist. **2**, 67–78 (2014)
29. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: IEEE International Conference on Computer Vision (2017)
30. Zhang, Y., Shen, D., Wang, G., Gan, Z., Henao, R., Carin, L.: Deconvolutional paragraph representation learning. In: Advances in Neural Information Processing Systems (2017)
31. Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: fashion synthesis with structural coherence. In: IEEE International Conference on Computer Vision (2017)