This is the peer reviewd version of the followng article:
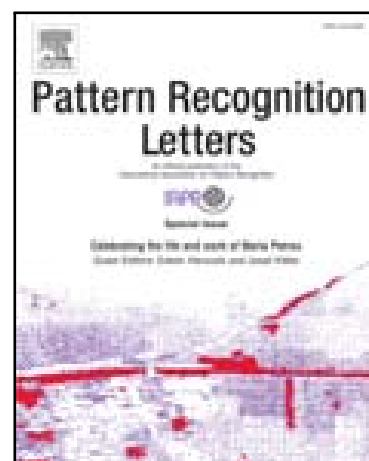
27/04/2024 10:03

(Article begins on next page)

Journal Pre-proof

Explaining Digital Humanities by Aligning Images and Textual Descriptions

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi,
Massimiliano Corsini, Rita Cucchiara

Please cite this article as: Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Massimiliano Corsini, Rita Cucchiara, Explaining Digital Humanities by Aligning Images and Textual Descriptions, *Pattern Recognition Letters* (2019), doi: https://doi.org/10.1016/j.patrec.2019.11.018

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlight

- We propose semi-supervised visual-semantic models for the Digital Humanities domain.

- Our approaches can align artistic images and text without paired supervision.

- We transfer the knowledge learned on ordinary dataset to the artistic domain.

- Experiments demonstrate the effectiveness of our distribution alignment strategy.

Keywords: Visual-semantic retrieval; Semi-supervised learning; Cultural Heritage

Explaining Digital Humanities by Aligning Images and Textual Descriptions

Marcella Cornia[a,*], Matteo Stefanini[a], Lorenzo Baraldi[a], Massimiliano Corsini[a], Rita Cucchiara[a]

[a]*University of Modena and Reggio Emilia, Department of Engineering "Enzo Ferrari", Via P. Vivarelli 10, 41125 Modena, Italy*

Abstract

Replicating the human ability to connect Vision and Language has recently been gaining a lot of attention in the Computer Vision and the Natural Language Processing communities. This research effort has resulted in algorithms that can retrieve images from textual descriptions and vice versa, when realistic images and sentences with simple semantics are employed and when paired training data is provided. In this paper, we go beyond these limitations and tackle the design of visual-semantic algorithms in the domain of the Digital Humanities. This setting not only advertises more complex visual and semantic structures but also features a significant lack of training data which makes the use of fully-supervised approaches infeasible. With this aim, we propose a joint visual-semantic embedding that can automatically align illustrations and textual elements without paired supervision. This is achieved by transferring the knowledge learned on ordinary visual-semantic datasets to the artistic domain. Experiments, performed on two datasets specifically designed for this domain, validate the proposed strategies and quantify the domain shift between natural images and artworks.

1. Introduction

As humans, we can easily link our ability to see and understand the surrounding environment with the ability to express ourselves in natural language. In the effort of artificially replicating these connections, new models have emerged for image and video captioning (Anderson et al., 2018; Lu et al., 2018; Cornia et al., 2019) and for visual-semantic retrieval (Kiros et al., 2014; Faghri et al., 2018; Lee et al., 2018). The former architectures combine vision and language in a generative flavor on the textual side, the latter build common spaces to integrate the two domains and retrieve textual elements given visual queries, and vice versa.

The leading solutions for visual-semantic retrieval have so far relied on fully supervised settings in which paired training samples are available and have been applied to

---

*Corresponding author: Tel.: +39-059-2058790; fax: +39-059-2056129;

Email address: marcella.cornia@unimore.it (Marcella Cornia)

Figure 1: Visual and textual data from the artistic domain are different from those addressed by ordinary visual-semantic datasets, posing significant challenges in the automatic understanding of arts and culture. Our approach can align illustrations and textual elements by transferring the knowledge learned on standard datasets to match images and captions coming from a target domain.
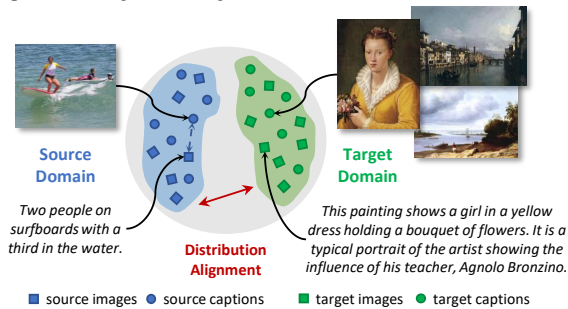
general-purpose datasets where the state of the art of concept recognition methods is useful and well assessed. In the domain of arts and culture, however, both visual and textual elements are far from those of ordinary datasets. On one side, textual descriptions often contain technical

language with symbolic reminds, metaphors and artistic or historical connections; on the other side, artworks and illustrations are characterized by visual features different from those of natural images. Beyond this domain-shift issue, the supervised training of a common visual-semantic embedding requires sufficiently large datasets. Instead, the artistic domain is often characterized by small scale datasets in which the pairing between visual and textual elements is not available or expensive to obtain.

Tackling the aforementioned setting, in this paper we propose a semi-supervised visual-semantic embedding model (`SS-VSE`) for cross-modal retrieval in the artistic domain. Our approach relies on the construction of a common semantic embedding, in which the knowledge learned on a supervised and ordinary visual-semantic dataset is transferred to an artistic dataset in which the pairing between images and sentences is not available. After using global feature vectors, we also investigate the use of auto-encoders (`SS-VSE-AE`) to obtain more compact representations of input images and sentences. Experiments are conducted on two datasets specifically designed for the artistic domain. In particular, we use the BibleVSA dataset (Baraldi et al., 2018) which contains illustrations and textual sentences extracted from the commentaries of a historical manuscript, and the SemArt dataset (Garcia and Vogiatzis, 2018) that is composed of artwork images and textual comments. Extensive experiments are presented to validate the proposed solution and to visualize the effect of the knowledge transfer between source and target datasets.

2. Related work

Deep Learning techniques often require significant efforts to be applied to the domain of Digital Humanities and Cultural Heritage, due to the presence of specific challenges. The research efforts of the past few years have resulted in various works and applications spanning from generative models to classification and retrieval solutions. On the generative and synthesis side, promising results have been obtained for transferring the style of a painting to a real photograph (Gatys et al., 2016; Sanakoyeu et al., 2018; Jing et al., 2018) and inversely, to create a realistic representation of a given painting (Zhu et al., 2017; Tomei et al., 2018, 2019a,b). On the analysis and feature extraction side, instead, several efforts have been made on

the collection and annotation of large scale datasets containing artistic images, mainly focusing on style and genre classification (Karayev et al., 2014; Mao et al., 2017; Strezoski and Worring, 2018), visual patterns detection (Shen et al., 2019), and artwork instance recognition (Del Chiaro et al., 2019).

Concerning the problem of linking textual descriptions and artistic images, there is a limited bunch of works available in the literature. In the next section, after briefly reviewing the most important works related to visual-semantic retrieval, we focus on image-text matching approaches applied to the artistic domain, and subdividing them between supervised and semi-supervised methods.

2.1. Visual-semantic retrieval

Matching visual data and natural language is a challenging task in computer vision and multimedia. Since visual and textual data belong to two distinct modalities, one of the seminal approaches (Kiros et al., 2014) has been that of generating a joint visual-semantic embedding space in which images and sentences could be compared. Even if other approaches exist, currently this is still one of the most commonly used solutions.

Following this line, Faghri et al. (2018) introduced a modification of the Hinge-based loss function to exploit hard negatives, *i.e.* worst matching pairs, during training. This has demonstrated to be effective to improve cross-modal retrieval performance and has been used in almost all subsequent works. Further, Wang et al. (2018) used a two-branch network composed of an embedding and a similarity branch: while the embedding network translates image and text into a feature representation, the similarity network predicts how well the feature representations match. Differently, Dong et al. (2018) suggested to tackle the retrieval problem exclusively in the visual space, introducing a deep neural model that learns to predict a visual feature representation from textual input.

Recently, strong improvements have been obtained by Lee et al. (2018) with a stacked cross-attention mechanism that matches images and textual descriptions by learning a latent correspondence between detected regions and words of the caption. Wang et al. (2019) extended this model by integrating an encoding of the relative position of image regions, which has proven to further enhance the learning of the joint embedding. On the same line, Li et al. (2019) proposed a reasoning model based on

3

graph convolutional networks to generate a visual representation that captures key objects and semantic concepts of a scene. All of these *supervised* methods have been proven effective when trained on large scale datasets, and are not designed to work with scarce data.

Only a few works have applied image-text matching strategies to artistic data. Among them, Garcia and Vogiatzis (2018) used additional metadata such as title, author, genre, and period of the paintings to find corresponding image-text pairs. Stefanini et al. (2019) introduced a new dataset and a visual-semantic model to discriminate visual and contextual sentences associated to artistic images and, at the same time, to align the corresponding visual and textual elements. While (Garcia and Vogiatzis, 2018; Stefanini et al., 2019) matched images and textual descriptions in a supervised way, (Baraldi et al., 2018; Carraggi et al., 2018) addressed the problem in a *semi-supervised* setting, adapting the knowledge learned on a given source domain to align images and text belonging to a different target domain and without directly training the model on the target domain. This solution, which is known as *domain adaptation*, has been used in a wide variety of applications such as image classification (Long et al., 2017), semantic segmentation (Hoffman et al., 2018; Chen et al., 2018b), object detection (Inoue et al., 2018; Chen et al., 2018a), and image captioning (Chen et al., 2017; Yang et al., 2018). 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.

3. Semi-supervised cross-modal retrieval

In the following, we describe our strategy for cross-modal retrieval in the artistic domain. Our model has a two-fold role: retrieving relevant images given textual sentences as queries, and retrieve relevant sentences when given images as queries. Parameters of the model are learned with the objective of maximizing recall at K – *i.e.* the fraction of queries for which the most relevant item is ranked among the top K retrieved ones. As training data in the artistic domain is often scarce, we build a proposal that does not need a paired training set in which the associations between images and sentences are known

in advance. Rather, our model transfers the knowledge learned on a source annotated dataset to a target dataset in which the pairing between the two modalities is unknown at training time.

In a nutshell, the paradigm of the common embedding space is exploited to learn similarities between images and sentences. In addition to using global feature vectors to encode data from both modalities, we also investigate the use of auto-encoders to learn more compact representations of images and sentences. To transfer knowledge to the artistic domain without leveraging annotated pairs, we devise a distribution alignment strategy based on the Maximum Mean Discrepancy measure, which aims at uncovering suitable cross-modal representation of cultural heritage data without supervision.

3.1. Visual-semantic embeddings

Aligning works of arts and their corresponding textual descriptions requires the ability to compare visual and textual data in this particular domain. To this end, we adopt the strategy of creating a shared multi-modal embedding space, in which both textual and visual elements can be projected and compared using a similarity function.

Formally, we denote $\phi(I, \mathbf{w}_\phi) \in \mathbb{R}^{D_\phi}$ as the feature representation computed from an image I of the dataset (such as the representation coming from a CNN), and $\psi(T, \mathbf{w}_\psi) \in \mathbb{R}^{D_\psi}$ as the representation of a textual element T, computed, for example, using a text encoder on one-hot vectors, or as a function of pre-trained word embeddings. Here, $\mathbf{w}_\phi$ and $\mathbf{w}_\psi$ indicate, respectively, the learnable weights of the visual and textual encoders.

To project those representations into a common semantic space, we perform a linear projection followed by a ℓ_2-normalization step, so that the resulting embedding space lies on the ℓ_2 unit ball:

$$f(I, \mathbf{w}_f, \mathbf{w}_\phi) = \ell_{2,norm}(\mathbf{w}_i^\mathsf{T} \phi(I, \mathbf{w}_\phi)) \tag{1}$$

$$g(T, \mathbf{w}_g, \mathbf{w}_\psi) = \ell_{2,norm}(\mathbf{w}_c^\mathsf{T} \psi(T, \mathbf{w}_\psi)), \tag{2}$$

where $\ell_{2,norm}$ is the ℓ_2 normalization function. Being D the dimensionality of the joint embedding space, $\mathbf{w}_f$ is a $D_\phi \times D$ matrix, and $\mathbf{w}_g$ is a $D_\psi \times D$ matrix.

Visual and textual elements can be compared in the joint multi-modal embedding space by computing the cosine similarity (equivalent, in this case, to a dot product)

between their projections, so that the similarity between an image I and a caption T becomes

$$s(I, T) = f(I, \mathbf{w}_f, \mathbf{w}_\phi) \cdot g(T, \mathbf{w}_g, \mathbf{w}_\psi). \tag{3}$$

Clearly, the utility of the joint embedding space is maximized when it exhibits suitable cross-modality matching properties, *i.e.* when similarities in the embedding space correspond to meaningful similarities in both modalities. In this case, the embedding space acts as a bridge between the two modalities and makes it possible to retrieve textual pieces describing a query image, and images described by a query caption by identifying the closest neighbors in both modalities.

Given a dataset annotated with matching visual-semantic pairs, a good proxy of this property is to verify that corresponding pairs are neighbours in the embedding space. As a matter of fact, classical approaches have relied on the availability of paired datasets, and have learned the joint embedding for a specific domain in a completely supervised way, *e.g.* training the parameters of the model according to a Hinge triplet ranking loss with margin, which imposes suitable similarities between matching and non-matching elements. Formally, it is defined as:

$$\ell(I, T) = \sum_{\hat{T}} \left[\alpha - s(I, T) + s(I, \hat{T})\right]_+ + \\ + \sum_{\hat{I}} \left[\alpha - s(I, T) + s(\hat{I}, T)\right]_+ \tag{4}$$

where $[x]_+ = \max(0, x)$ and α is a margin. In the equation above, (I, T) is a matching image-text pair (*i.e.*, such that T describes the content of I, and I represents the content of T), while $\hat{T}$ is a negative text with respect to I (such that $\hat{T}$ does not describe I), and $\hat{I}$ is a negative image with respect to T (such that T does not describe $\hat{I}$). The terms contained in both sums require that the difference in similarity between the matching and the non-matching pair is higher than a margin α: in the first sum, this is done by considering an image anchor and matching or non-matching captions; in the latter, instead, a caption is used as anchor.

As reported by a recent work by (Faghri et al., 2018), in a completely supervised setting it is often beneficial to replace the sums in Eq. 4 with maximum operations, so to consider only the most violating non-matching pair.

3.2. Auto-encoding images and sentences

In addition to the use of plain global feature vectors, we also investigate an alternative projection strategy in which images and sentences are fed to an auto-encoder to learn a more compact yet powerful representation of the input, which can in turn be used as the input of the projection function defined in Eq. 1.

To this end, we design a textual auto-encoder which can convert variable-length captions to fixed-length representations from which input sentences can be reconstructed. In particular, our model exploits Gated Recurrent Networks (GRUs) (Cho et al., 2014) for both encoding and decoding. Formally, given a sentence $T = (w_1, w_2, ..., w_N)$ with length N, we firstly encode it word by word through a single-layer GRU and take the last hidden state of the Recurrent layer as the encoding of the sentence. Given the recurrent relation defined by the GRU cell and the t-h word, *i.e.*

$$\mathbf{h}_t = \text{GRU}_e(w_t, \mathbf{h}_{t-1}), \tag{5}$$

the encoding of the input sentence is defined as:

$$\mathbf{h}_N = \text{GRU}_e(w_N, \mathbf{h}_{N-1}). \tag{6}$$

In the decoding stage, the input sentence is reconstructed by feeding $\mathbf{h}_N$ to a second GRU layer which is in charge of generating the reconstructed sentence. During training, at the t-th iteration the Recurrent layer is fed with $\mathbf{h}_N$ and the previous ground-truth words, and it is trained to predict the t-h word. Formally, the training objective is thus:

$$\max_{\mathbf{w}} \sum_{t=1}^{T} \log \text{Pr}(w_t | w_{t-1}, w_{t-2}, ..., w_1, \mathbf{h}_N). \tag{7}$$

The probability of a word is modeled via a softmax layer applied to the output of the decoder. To reduce the dimensionality of the decoder, a linear embedding transformation is used to project one-hot word vectors into the input space of the decoder and, vice-versa, to project the output of the decoder to the dictionary space.

Given the auto-encoder for the textual part, 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 practice, the encoder model is 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, we define the output of the encoder model $\mathbf{z}$ (*i.e.* the intermediate representation of the input image) as

$$\mathbf{z} = \tanh(\mathbf{W_e}\phi(I) + b_e), \qquad (8)$$

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

The decoder model has a symmetric structure. Therefore, starting from the intermediate vector $\mathbf{z}$, the decoder applies a single fully connected layer that transforms $\mathbf{z}$ to the size of the input image feature vector. Formally, the reconstructed image feature vector $\hat{\phi}(I)$ is defined as

$$\hat{\phi}(I) = \mathbf{W_d}\mathbf{z}_i + b_d, \qquad (9)$$

where $\mathbf{W_d}$ and b_d are the weight matrix and the bias vector of the decoder. Overall, the image auto-encoder is trained to minimize the reconstruction error for each input image. We define the decoder loss function as the mean square error between the original image feature vector $\phi(I)$ and the corresponding reconstruction $\hat{\phi}(I)$.

3.3. Aligning distributions

While the knowledge of matching and non-matching pairs on a source dataset can be exploited to train the embedding space, as discussed in Sec. 3.1, the two reconstruction losses can be applied to both the source and the target dataset, thus building encoded representations which are suitable for both datasets. However, this is not enough to transfer knowledge from the source domain to the target domain, as there is no guarantee that encoded words and sentences from the target dataset will lie together in the embedding space.

To this end, we match the distributions of textual and visual data in the target domain, while learning from pairs sampled from the source domain. Following recent works in the field (Hubert Tsai et al., 2016; Tsai et al., 2017; Yan et al., 2017), we use the Maximum Mean Discrepancy (MMD) to compare distributions. This, basically, computes the distance between the expectations of the two distributions in a reproducing kernel Hilbert space $\mathcal{H}_\kappa$ endowed with a kernel κ, and can be used as an additional loss term:

$$\mathcal{L}_{mmd} = \|\mathbf{E}_{I\sim\mathcal{I}}\left[f(I)\right] - \mathbf{E}_{T\sim\mathcal{T}}\left[g(T)\right]\|^2_{\mathcal{H}_\kappa}, \qquad (10)$$

where $\mathcal{I}$ is the distribution of the illustrations, and $\mathcal{T}$ is the distribution of captions. The kernel in the MMD criterion must be a universal kernel, and thus we empirically choose a Gaussian kernel:

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\sigma\|\mathbf{x} - \mathbf{y}\|^2\right). \qquad (11)$$

At training time, we sample two mini-batches of samples, one from the supervised set and a second one from the unsupervised dataset. The back-propagated loss is then the sum of the supervised loss (Eq. 4) on the supervised set, plus the MMD loss $\mathcal{L}_{mmd}$ approximated over the batch from the unsupervised set. Additionally, the two loss terms of the auto-encoders are evaluated over both the supervised and the unsupervised batches.

4. Experimental results

4.1. Datasets

We perform experiments on two different visual-semantic datasets containing artistic images and corresponding textual descriptions (described below). As source domains, we use Flickr30k and COCO which are composed of natural images and are commonly used to train cross-modal retrieval methods. For these two datasets, we use the splits provided by (Karpathy and Fei-Fei, 2015).

BibleVSA (Baraldi et al., 2018). The dataset consists of $2,282$ illustrations taken from the digitized version of the Borso d'Este Holy Bible, one of the most significant illustrated manuscripts of Renaissance. Each image is associated with a single textual phrase extracted from a textual commentary which describes the content of each page of the manuscript. In our experiments, we use the original training, validation, and test split, respectively composed of $1,671$, 293, and 307 image-caption pairs.

SemArt (Garcia and Vogiatzis, 2018). This dataset is composed of $21,384$ paintings extracted from the Web Gallery of Art, which contains European fine-art reproductions between the 8th and the 19th century. Each image is associated to an artistic comment and to a set of 7 different attributes comprising the title, the author, and the type of the painting. Overall, the dataset is divided in training, validation and test split with $19,244$, $1,069$ and $1,069$ elements, respectively. The average length

6

Figure 2: Comparison between the visual and textual features of ordinary visual-semantic datasets (Flickr30k, COCO) and those of BibleVSA and SemArt dataset. Visualization is obtained by running the t-SNE algorithm on top of the features. Best seen in color.
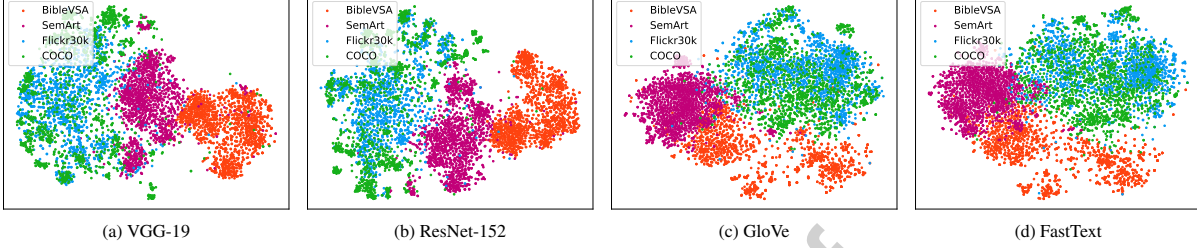

(a) VGG-19 (b) ResNet-152 (c) GloVe (d) FastText

Table 1: Semi-supervised cross-modal retrieval results using different visual features. Results are reported on BibleVSA and SemArt test set.

Method	CNN Feat.	Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
		COCO → BibleVSA					
SS-VSE	VGG-19	**13.1**	29.5	36.1	3.9	16.7	27.5
SS-VSE	ResNet-152	9.8	**31.1**	**50.8**	**6.2**	**22.3**	**30.8**
SS-VSE-AE	VGG-19	**9.8**	**27.9**	34.4	**3.6**	15.7	25.9
SS-VSE-AE	ResNet-152	6.6	23.0	**36.1**	**3.6**	**19.7**	**29.8**
		COCO → SemArt					
SS-VSE	VGG-19	3.7	11.7	19.0	2.3	10.0	19.3
SS-VSE	ResNet-152	**6.7**	**19.3**	**27.0**	**5.0**	**17.3**	**29.3**
SS-VSE-AE	VGG-19	**5.0**	**14.3**	**22.7**	1.7	9.0	15.3
SS-VSE-AE	ResNet-152	4.7	12.7	21.0	**3.7**	**11.0**	**18.0**

Table 2: Semi-supervised cross-modal retrieval results using different word embeddings. Results are reported on BibleVSA and SemArt test set.

Method	Word Emb.	Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
		COCO → BibleVSA					
SS-VSE	FastText	8.2	19.7	34.4	2.6	16.7	26.6
SS-VSE	GloVe	6.6	23.0	39.3	3.6	16.7	27.2
SS-VSE	-	**9.8**	**31.1**	**50.8**	**6.2**	**22.3**	**30.8**
SS-VSE-AE	FastText	**6.6**	**27.9**	34.4	3.3	14.4	25.2
SS-VSE-AE	GloVe	4.9	19.7	**41.0**	**3.9**	13.8	27.5
SS-VSE-AE	-	**6.6**	23.0	36.1	3.6	**19.7**	**29.8**
		COCO → SemArt					
SS-VSE	FastText	1.7	5.0	7.7	0.7	2.3	7.3
SS-VSE	GloVe	3.3	11.3	16.0	2.0	11.0	17.7
SS-VSE	-	**6.7**	**19.3**	**27.0**	**5.0**	**17.3**	**29.3**
SS-VSE-AE	FastText	3.7	10.0	17.0	3.0	9.3	11.7
SS-VSE-AE	GloVe	2.7	12.0	17.0	1.7	7.0	12.3
SS-VSE-AE	-	**4.7**	**12.7**	**21.0**	**3.7**	**11.0**	**18.0**

of each artistic comment is more than 80, with a maximum number of words equal to 830. This highlights the difference between SemArt and ordinary visual-semantic datasets (*i.e.* COCO has an average caption length lower than 11) and accentuates the challenges of this set of data. To first validate our solution in a less complex scenario, we limit the validation and test set to 300 randomly selected image-text pairs. Then, we evaluate our model using a different number of retrievable items.

4.2. Implementation details

To encode input images, we use two different convolutional networks: the VGG-19 (Simonyan and Zisserman, 2015) and ResNet-152 (He et al., 2016). We extract image features from the *fc7* layer of the VGG-19 and from the average pooling layer of the ResNet-152 thus obtaining an input image embedding dimensionality D_ϕ of 4096 and 2048, respectively.

For encoding image descriptions, we use a GRU network (Cho et al., 2014). We set the dimensionality of the GRU and of the joint embedding space D to 512, while the input size of word embeddings D_ψ is set to 300. We use either a text encoder on one-hot vectors or different pre-trained word embeddings (such as GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017)) as input of the GRU.

The model with textual and visual auto-encoders is trained using the same input and output sizes. For the training with pre-trained word embeddings, instead of using the loss function defined in Eq. 7, we compute the cosine distance between original and reconstructed embeddings of each word.

All experiments are performed by using Adam opti-

mizer with a learning rate of 0.0002 for 15 epochs and then decreased by a factor of 10. We set the margin α to 0.2, the σ parameter of the Gaussian kernel to 1 and the size of the mini-batch to 128.

4.3. Analysis of artistic visual-semantic data

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

For the visual part, we extract the activation from the VGG-19 and ResNet-152 networks, while, for textual elements, we embed each word of a caption with a word embedding strategy (either GloVe or FastText). To get a feature vector for a sentence, we sum the ℓ_2 normalized embeddings of the words, and we apply the ℓ_2-norm also to the results. This strategy is largely used in image and video retrieval literature and is known for preserving the information of the original vectors into a compact representation with fixed dimensionality (Tolias et al., 2016).

Fig. 2 shows the distributions of visual and textual features of both datasets. To get a suitable two-dimensional representation, we run the t-SNE algorithm (Maaten and Hinton, 2008), which iteratively finds a non-linear projection that preserves the statistical distribution of the pairwise distances from the original space. As it can be observed, the features of ordinary visual-semantic datasets share almost the same visual and textual distributions. BibleVSA and SemArt, on the contrary, feature a completely different distribution, according to both modalities and all feature extractors. This underlines, on the one hand, that artistic datasets define a completely new domain. On the other hand, instead, this motivates the low performance of existing models when tested on these datasets.

4.4. Cross-modal retrieval results

To evaluate the effectiveness of the visual-semantic embeddings, 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

Table 4: Semi-supervised retrieval results on BibleVSA test set.

Method	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Flickr30k → BibleVSA						
VSE	3.3	8.2	16.4	1.6	12.1	19.7
SS-VSE	9.8	23.0	39.3	4.6	16.1	26.6
VSE-AE	1.6	4.9	13.1	3.0	9.8	17.0
SS-VSE-AE	3.3	23.0	29.5	3.3	13.1	23.0
COCO → BibleVSA						
VSE	1.6	9.8	16.4	2.6	10.5	20.0
SS-VSE	9.8	31.1	50.8	6.2	22.3	30.8
VSE-AE	3.3	6.6	14.8	1.6	9.8	19.7
SS-VSE-AE	6.6	23.0	36.1	3.6	19.7	29.8

of caption retrieval, or the top-K retrieved images, in the case of image retrieval.

Firstly, we assess the performance of our full model when using different CNN features or different word embeddings, to get an insight of the role of different global feature vectors. In Table 1, we show the performance of the proposed approach on the test sets of BibleVSA and SemArt when using image features extracted, respectively, from VGG-19 and ResNet-152. Table 2 compares the use of FastText and GloVe embeddings versus a learned word embedding matrix. In this case, the results on SemArt test set are obtained by using 300 randomly selected retrievable items.
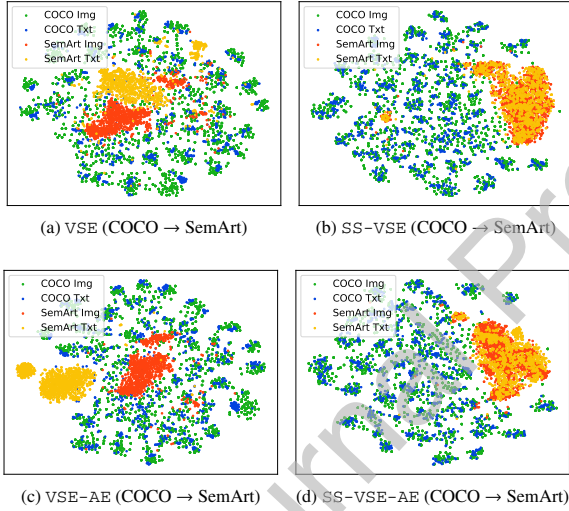
For space reasons, we limit this analysis to a single source dataset (namely, COCO), as we have observed similar behaviours on Flickr30k. The two variants of our approach are denoted as SS-VSE and SS-VSE-AE, where the first refers to the model with global feature vectors and linear projection, and the latter refers to the model with the visual and textual auto-encoder. As it can be observed, the global descriptor extracted from ResNet-152 outperforms the one extracted from VGG-19 in almost all settings. Noticeably, learned word embeddings outperform pre-trained solutions. We speculate that this performance drop is due to the the highly specialized nature of the target datasets. In this regards, word embeddings seem to offer a poor initialization point with respect to a from-scratch learning of the word embedding matrix.

Another interesting consideration is that the use of hard negatives in the triples loss function is typically beneficial in a supervised setting (Faghri et al., 2018). Instead, in our

Table 3: Semi-supervised cross-modal retrieval results on SemArt test set using a different number N of retrievable items.

| Method | N = 100 | | | | | | N = 300 | | | | | | N = 500 | | | | | | N = 1000 | | | | | |
| | Text Retrieval | | | Image Retrieval | | | Text Retrieval | | | Image Retrieval | | | Text Retrieval | | | Image Retrieval | | | Text Retrieval | | | Image Retrieval | | |
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
											Flickr30k → SemArt													
VSE	2.0	10.0	14.0	3.0	11.0	17.0	1.7	5.7	8.7	1.0	5.3	7.3	0.8	2.6	6.0	0.4	3.6	5.6	0.5	1.6	2.8	0.1	1.2	2.8
SS-VSE	**7.0**	**23.0**	**40.0**	**10.0**	**23.0**	**37.0**	**5.0**	**15.3**	**22.0**	**3.7**	**13.3**	**17.7**	**3.6**	**9.6**	**14.6**	**1.8**	**7.6**	**12.0**	**1.5**	**6.2**	**10.0**	**1.2**	**3.5**	**7.4**
VSE-AE	3.0	9.0	15.0	5.0	12.0	19.0	2.3	6.0	7.3	0.7	6.0	9.0	1.2	4.2	6.4	0.8	3.0	5.4	0.5	2.2	4.1	0.5	1.6	3.1
SS-VSE-AE	6.0	28.0	42.0	6.0	18.0	30.0	4.0	12.7	20.0	2.3	10.0	16.3	1.8	9.2	14.8	1.6	6.0	11.4	1.0	5.6	9.4	0.6	3.4	6.8
											COCO → SemArt													
VSE	5.0	13.0	21.0	3.0	8.0	19.0	1.7	8.7	15.3	1.0	8.0	12.3	1.2	3.6	6.4	1.6	3.4	6.0	1.0	2.7	3.6	0.5	2.3	3.6
SS-VSE	**16.0**	**34.0**	**52.0**	**12.0**	**32.0**	**48.0**	**6.7**	**19.3**	**27.0**	**5.0**	**17.3**	**29.3**	**3.8**	**12.2**	**19.8**	**3.4**	**11.6**	**19.4**	**2.7**	**8.9**	**14.0**	**2.3**	**6.9**	**12.9**
VSE-AE	6.0	15.0	20.0	3.0	11.0	22.0	3.0	7.3	11.7	0.3	3.7	6.7	1.6	4.0	6.2	1.2	2.8	4.0	0.8	2.6	4.0	0.8	1.6	2.3
SS-VSE-AE	**7.0**	**24.0**	**39.0**	**6.0**	**17.0**	**26.0**	**4.7**	**12.7**	**21.0**	**3.7**	**11.0**	**18.0**	**2.0**	**10.0**	**15.8**	**2.2**	**5.0**	**10.8**	**0.9**	**6.1**	**10.0**	**1.0**	**3.8**	**5.8**

Figure 3: Comparison between t-SNE projections of the embedding spaces learned with (b-d) and without (a-c) the MMD loss. Best seen in color.

(a) VSE (COCO → SemArt)

(b) SS-VSE (COCO → SemArt)

(c) VSE-AE (COCO → SemArt)

(d) SS-VSE-AE (COCO → SemArt)

semi-supervised setting, we do not report the same advantages in improving the alignment of the target domain.

4.5. Evaluation of semi-supervised embeddings

In Tables 3 and 4, we compare the performances of the two proposed semi-supervised approaches (SS-VSE and SS-VSE-AE) on SemArt and BibleVSA test set with respect to the two models trained without the distribution alignment (VSE and VSE-AE). For these experiments, we use global feature vectors extracted from ResNet-152 and learned word embeddings. Given the significant size of SemArt dataset, we report retrieval results when using different sets of database items (*i.e.* 100, 300, 500, 1000).

We notice that, when using a medium-scale source dataset like Flickr30k, the use of the auto-encoder is competitive with the use of a linear projection of the global feature vector. Instead, when transferring from a large-scale dataset like COCO, the reconstruction term is not needed and the reduced size of the representation degrades the performance. In all settings, the MMD loss gives a significant contribution to the final performance thus confirming the effectiveness of our distribution alignment strategy.

To get a better understanding of the role of the MMD loss, we also show the learned multi-modal embedding space by using t-SNE visualizations. Figure 3 shows the embedding spaces when transferring from COCO to SemArt, with and without the MMD loss. As it can be noticed, without the MMD loss the distribution of textual and visual elements on the target domain remains almost separate, as the learning signal from the source domain is not general enough on the target domain. On the contrary, when applying the MMD loss the distribution of the learned image embeddings matches that of the textual counterpart on the target domain, thus confirming the effectiveness of the proposed semi-supervised strategy. Noticeably, the distributions of the source and target domain still remain separate in the embedding space, thus underlying the diverse nature of the two sets.

Finally, Fig. 4.5 reports sample qualitative results on BibleVSA and SemArt dataset. As it can be noticed, our method can retrieve significant elements without employing any paired supervision from the artistic dataset.

5. Conclusion

We tackled the task of building visual-semantic retrieval approaches for the Cultural Heritage domain. To

Figure 4: Qualitative image-to-text (upper) and text-to-image (lower) results on BibleVSA (first and third rows) and SemArt (second and fourth rows) dataset, using the proposed semi-supervised strategy.

Query Image	Top-1 Retrieved Caption	Query Image	Top-1 Retrieved Caption	Query Image	Top-1 Retrieved Caption

A round, within a quadrangular frame of a laurel wreath, with Moses kneeling listening to the word of God appearing in the sky.

A fantastic figure with a leopard body and human head holds a spear and a shield.

A round depicting a dog hunting a heron.

This three quarter length portrait of Midshipman (later Captain) John Windham Dalling RN (1789-1853) memorialises his presence on HMS Defence at the Battle of Trafalgar (...).

This painting shows the Madonna and Child in a landscape with the Infant Saint John the Baptist. It betrays the influence of (...).

This painting depicts a still-life of flowers in a vase, with fruit on a ledge behind.

Query Caption	Top-1 Retrieved Image	Query Caption	Top-1 Retrieved Image	Query Caption	Top-1 Retrieved Image

A quadrangular vignette with Moses and Aaron, kneeling in a landscape, they listen to the word of God appearing in the form of a radiated cloud.

A landscape with the leopard with tail and dragon wings.

Quadrangular vignette with Moses preaching to the people gathered around him.

This study of a bearded man, head and shoulders, was probably made with the intention to use it in some multi-figural composition.

In this genre scene three men are depicted relaxing in a sparse interior as one plays his violin and the others jovially hold a pipe and vessels for drinking (...).

This still-life depicts Bohemian crystals, cups, and a watch.

this aim, we have proposed a semi-supervised approach which does not rely on labelled data on the artistic domain and translates the knowledge learned on ordinary visual-semantic datasets to the more challenging case of artistic data. Extensive experimental results validated the proposed strategy. Regardless, future research should consider the potential effects of semi-supervised approaches using more fine-grained methods, for example aligning detected regions and sentence words between source and target distributions instead of their global representations. As this has been proven useful in ordinary domains, its interactions with domain adaptation should be investigated. Moreover, a comprehensive comparison of domain adaptation techniques, including those employing adversarial objectives, and of their applicability to the Cultural Heritage domain is needed to further advance the research in the field.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Baraldi, L., Cornia, M., Grana, C., Cucchiara, R., 2018. Aligning text and document illustrations: towards visually explainable digital humanities, in: Proceedings of the International Conference on Pattern Recognition.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

Carraggi, A., Cornia, M., Baraldi, L., Cucchiara, R., 2018. Visual-semantic alignment across domains using a semi-supervised approach, in: Proceedings of the European Conference on Computer Vision Workshops.

Chen, T.H., Liao, Y.H., Chuang, C.Y., Hsu, W.T., Fu, J., Sun, M., 2017. Show, adapt and tell: Adversarial training of crossdomain image captioner, in: Proceedings of the International Conference on Computer Vision.

Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L., 2018a. Domain adaptive Faster R-CNN for object detection in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Chen, Y., Li, W., Van Gool, L., 2018b. ROAD: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Cornia, M., Baraldi, L., Cucchiara, R., 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Del Chiaro, R., Bagdanov, A.D., Del Bimbo, A., 2019. Webly-supervised Zero-shot Learning for Artwork Instance Recognition. Pattern Recognition Letters .

Dong, J., Li, X., Snoek, C.G., 2018. Predicting Visual Features from Text for Image and Video Caption Retrieval. IEEE Transactions on Multimedia 20, 3377–3388.

Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S., 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives, in: Proceedings of the British Machine Vision Conference.

Garcia, N., Vogiatzis, G., 2018. How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval, in: Proceedings of the European Conference on Computer Vision Workshops.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T., 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation, in: Proceedings of the International Conference on Machine Learning.

Hubert Tsai, Y.H., Yeh, Y.R., Frank Wang, Y.C., 2016. Learning cross-domain landmarks for heterogeneous domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K., 2018. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Jing, Y., Liu, Y., Yang, Y., Feng, Z., Yu, Y., Tao, D., Song, M., 2018. Stroke controllable fast style transfer with adaptive receptive fields, in: Proceedings of the European Conference on Computer Vision.

Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H., 2014. Recognizing image style, in: Proceedings of the British Machine Vision Conference.

Karpathy, A., Fei-Fei, L., 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Kiros, R., Salakhutdinov, R., Zemel, R.S., 2014. Unifying visual-semantic embeddings with multimodal neural language models, in: Advances in Neural Information Processing Systems Workshops.

Lee, K.H., Chen, X., Hua, G., Hu, H., He, X., 2018. Stacked cross attention for image-text matching, in:

11

Proceedings of the European Conference on Computer Vision.

Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y., 2019. Visual Semantic Reasoning for Image-Text Matching, in: Proceedings of the International Conference on Computer Vision.

Long, M., Zhu, H., Wang, J., Jordan, M.I., 2017. Deep transfer learning with joint adaptation networks, in: Proceedings of the International Conference on Machine Learning.

Lu, J., Yang, J., Batra, D., Parikh, D., 2018. Neural baby talk, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

Mao, H., Cheung, M., She, J., 2017. Deepart: Learning joint representations of visual arts, in: ACM International Conference on Multimedia.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Sanakoyeu, A., Kotovenko, D., Lang, S., Ommer, B., 2018. A Style-Aware Content Loss for Real-time HD Style Transfer, in: Proceedings of the European Conference on Computer Vision.

Shen, X., Efros, A.A., Mathieu, A., 2019. Discovering Visual Patterns in Art Collections with Spatially-consistent Feature Learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations.

Stefanini, M., Cornia, M., Baraldi, L., Corsini, M., Cucchiara, R., 2019. Artpedia: A New Visual-Semantic Dataset with Visual and Contextual Sentences in the Artistic Domain, in: Proceedings of the International Conference on Image Analysis and Processing.

Strezoski, G., Worring, M., 2018. Omniart: A large-scale artistic benchmark. ACM Transactions on Multimedia Computing, Communications, and Applications 14, 88.

Tolias, G., Sicre, R., Jégou, H., 2016. Particular object retrieval with integral max-pooling of CNN activations, in: Proceedings of the International Conference on Learning Representations.

Tomei, M., Baraldi, L., Cornia, M., Cucchiara, R., 2018. What was Monet seeing while painting? Translating artworks to photo-realistic images, in: Proceedings of the European Conference on Computer Vision Workshops.

Tomei, M., Cornia, M., Baraldi, L., Cucchiara, R., 2019a. Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-to-Image Translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Tomei, M., Cornia, M., Baraldi, L., Cucchiara, R., 2019b. Image-to-image translation to unfold the reality of artworks: an empirical analysis, in: Proceedings of the International Conference on Image Analysis and Processing.

Tsai, Y.H.H., Huang, L.K., Salakhutdinov, R., 2017. Learning Robust Visual-Semantic Embeddings, in: Proceedings of the International Conference on Computer Vision.

Wang, L., Li, Y., Lazebnik, S., 2018. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 394–407.

Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., Fan, X., 2019. Position Focused Attention Network for Image-Text Matching, in: Proceedings of the International Joint Conferences on Artificial Intelligence.

Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W., 2017. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition .

Yang, M., Zhao, W., Xu, W., Feng, Y., Zhao, Z., Chen, X., Lei, K., 2018. Multitask learning for cross-domain image captioning. IEEE Transactions on Multimedia 21, 1047–1061.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the International Conference on Computer Vision.

AUTHOR DECLARATION

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). She is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from **marcella.cornia@unimore.it**.

Signed by all authors as follows:

Marcella Cornia

15/07/2019

Matteo Stefanini

15/07/2019

Lorenzo Baraldi

15/07/2019

Massimiliano Corsini

15/07/2019

Rita Cucchiara

15/07/2019