This is the peer reviewd version of the followng article:

Artpedia: A New Visual-Semantic Dataset with Visual and Contextual Sentences in the Artistic Domain / Stefanini, Matteo; Cornia, Marcella; Baraldi, Lorenzo; Corsini, Massimiliano; Cucchiara, Rita. - (2019), pp. 729-740. (Intervento presentato al convegno International Conference on Image Analysis and Processing tenutosi a Trento, Italy nel 9-13 September, 2019) [10.1007/978-3-030-30645-8_66].

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

19/10/2024 20:32

(Article begins on next page)

Artpedia: A New Visual-Semantic Dataset with Visual and Contextual Sentences in the Artistic Domain

Matteo Stefanini[0000−0001−6153−926X], Marcella Cornia[0000−0001−9640−9385], Lorenzo Baraldi[0000−0001−5125−4957], Massimiliano Corsini[0000−0003−0543−1638], and Rita Cucchiara[0000−0002−2239−283X]

University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it

Abstract. As vision and language techniques are widely applied to realistic images, there is a growing interest in designing visual-semantic models suitable for more complex and challenging scenarios. In this paper, we address the problem of cross-modal retrieval of images and sentences coming from the artistic domain. To this aim, we collect and manually annotate the *Artpedia* dataset that contains paintings and textual sentences describing both the visual content of the paintings and other contextual information. Thus, the problem is not only to match images and sentences, but also to identify which sentences actually describe the visual content of a given image. To this end, we devise a visual-semantic model that jointly addresses these two challenges by exploiting the latent alignment between visual and textual chunks. Experimental evaluations, obtained by comparing our model to different baselines, demonstrate the effectiveness of our solution and highlight the challenges of the proposed dataset. The Artpedia dataset is publicly available at: http://aimagelab.ing.unimore.it/artpedia.

Keywords: Cross-modal retrieval · Visual-semantic models · Cultural Heritage.

1 Introduction

The integration of vision and language has recently gained a lot of attention from both computer vision and NLP communities. As humans, we can seamlessly connect what we visually see or imagine and what we hear or say, therefore building effective bridges between our ability to see and our ability to express ourselves in a common language. In the effort of artificially replicating these connections, new algorithms and architectures have recently emerged for image and video captioning [1,16,5] and for visual-semantic retrieval [13,7,15]. The former architectures combine vision and language in a generative flavour on the textual side, and in the latter common spaces are built to integrate the two domains and retrieve textual elements given visual queries, and vice versa.

While the standard objective in visual-semantic retrieval is that of associating images and *visual sentences* (*i.e.* sentences that visually describe something), the variety of sentences which can be found in textual corpora is definitely larger, and also contains sentences which do not describe the visual content of a scene. Here, we go a step beyond and extend the task of visual-semantic retrieval to a setting in which the textual

domain does not exclusively contain visual sentences, and explore the task of identifying relevant visual sentences given image queries. As such, the task establishes two challenges, the first one being that of understanding whether the sentence has a visually relevant content, and the second being that of associating elements between the two domains.

Further, we also address a second shortcoming of most visual-semantic works, *i.e.* that of dealing with photo-realistic images and simple texts. As there is a growing need of extending these algorithms to less general semantic and visual domains, we both increase the complexity on the visual and on the semantic side. To create an environment where all the aforementioned challenges live together, we focus on the case of artistic data — which surely advertise more complex and unusual visual and semantic features, and propose a new dataset with *visual* and *contextual* sentences for each visual item. In short, visual sentences deal with the visual appearance of the item, contextual ones describe either the item or its context without dealing with its visual appearance.

We also design and evaluate a model for jointly associating visual and textual elements, and identifying visual textual samples as opposed to contextual ones. Taking inspiration from state of the art models for visual-semantic retrieval, we test both traditional approaches, based on global feature vectors, and approaches that model the latent alignment between visual and textual chunks.

The rest of this paper is organized as follows: after briefly reviewing the related literature in Section 2, we present the *Artpedia* dataset in Section 3. Further, in Section 4 we propose our model for bringing visual and contextual sentences in visual-semantic retrieval, which is subsequently evaluated together with different baselines in Section 5.

2 Related work

In this section, we first give an overview of cross-modal retrieval models. Then, we review computer vision works related to the cultural heritage domain with a focus on other relevant datasets for art understanding.

2.1 Cross-modal retrieval

Cross-modal retrieval is one of the core challenges in computer vision and multimedia communities and consists in the retrieval of visual items given textual queries, and vice versa. In this context, several cross-modal retrieval models have been proposed [13,7,15], with the objective of minimizing the distance of matching image-text pairs and, on the contrary, maximizing that of non-matching elements. Among them, Faghri *et al.* [7] introduced a simple modification of standard loss functions based on the use of hard negatives that has been demonstrated to be effective in improving the performance of cross-modal retrieval and has been widely adopted by several subsequent methods [6,10,11,15].

Inspired by the use of multiple image descriptors to improve related visual-semantic tasks [1,25], Lee *et al.* [15] have recently proposed to match images and corresponding descriptions by inferring a latent correspondence between image regions and single words of the caption. In this work, we exploit a similar attentive mechanism to match

Table 1. Overview of the most relevant datasets containing artistic images.

Dataset	# Images	# Sentences	Manually Annotated	Task
Wikipaintings [12]	85,000	-	✗	Classification
Art500k [18]	554,198	-	✗	Classification and retrieval
Brueghel [21]	1,587	-	✓	Near duplicate detection
SemArt [8]	21,383	21,383	✗	Visual-semantic retrieval
EsteArtworks [3]	553	1,278	✓	Visual-semantic retrieval
BibleVSA [2]	2,282	2,271	✓	Visual-semantic retrieval
Artpedia	2,930	28,212	✓	Visual-semantic retrieval (with contextual sentences)

each painting with the sentences that actually describe the visual content of the painting itself, and we demonstrate the effectiveness of using multiple image regions in place of a single image descriptor also for visual-semantic artistic data.

2.2 Computer vision for cultural heritage

In the last years, several efforts have been done to apply computer vision techniques to the cultural heritage domain resulting in different works and applications ranging from generative models to classification and retrieval solutions. On the generative and synthesis side, up-and-coming results have been obtained by style transfer models that aim to transfer the style of a painting to a real photo [9] and, on the contrary, create a realistic representation of a given painting [23,24].

On a different note, several large-scale art datasets have been proposed to foster researches on this domain, with a particular focus on style and genre recognition [12,18]. For a comprehensive analysis, Table 1 shows a summary of the most relevant dataset related to the cultural heritage domain. To the best of our knowledge, there is a limited bunch of works that address the problem of retrieving artistic images from textual descriptions, and vice versa [2,3,8]. While [2,3] take the problem in a semi-supervised way by exploiting the knowledge from large-scale datasets containing realistic images, [8] uses additional metadata such as title, author, genre, and period of the paintings to match images and text. In this paper, we instead propose a visual-semantic model capable of discriminating *visual* and *contextual* sentences for each considered painting and, at the same time, associating the corresponding visual and textual elements.

3 The Artpedia Dataset

To foster the research on the development of visual-semantic algorithms which deal with contextual sentences, we propose a novel dataset with visual and contextual sentences describing real paintings. *Artpedia* contains a collection of $2,930$ painting images, each associated to a variable number of textual descriptions. Each sentence is labelled either as a *visual* sentence or as a *contextual* sentence, if does not describe the visual content of the artwork. Contextual sentences can describe the historical context of the artwork, its author, the artistic influence or the place where the painting is exhibited.

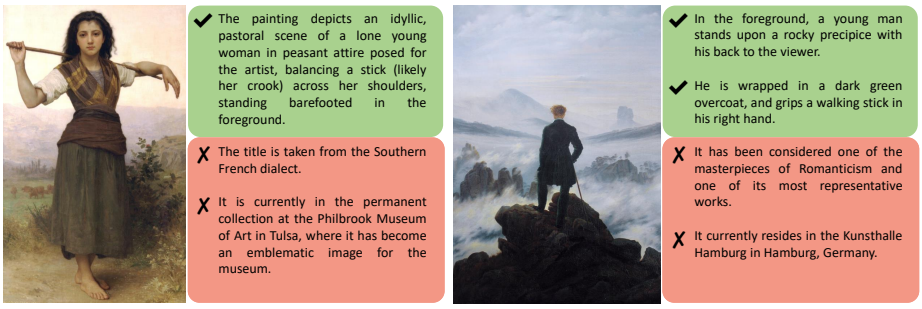


Fig. 1. Sample paintings from our Artpedia dataset with corresponding visual (green boxes) and contextual (red boxes) sentences.

As in standard cross-modal datasets, the association between sentences and painting is also provided. A sample of the dataset and its annotations is shown in Figure 1.

As the name suggests, the dataset has been collected by crawling Wikipedia pages. To this aim, our crawling strategy followed the Wikipedia category hierarchy by navigating all categories containing paintings between the 13th and the 21th century. We then extracted the textual descriptions taking into account all the summaries of each Wikipedia page and the description section whenever present. Finally, we split the text into sentences using the spaCy NLP toolbox[1] and manually annotated each sentence either as visual or contextual. As an additional product of the crawling procedure, we also release the title and the year of each painting, together with the URL of each image.

Overall, Artpedia contains a total of $28,212$ sentences, $9,173$ labelled as visual sentences and the remaining $19,039$ as contextual sentences. On average, each painting is associated with 3.1 visual and 6.5 contextual sentences. The mean length of the textual items is 21.5 words, considerably longer than those of standard image captioning datasets. For a comprehensive analysis of the visual and semantic content of our Artpedia dataset, we report in Figure 2 the distribution of paintings over the given range of centuries, the distribution of sentence lengths, and the most common object classes obtained by running a pre-trained object detector [20,14].

With respect to other visual-semantic datasets containing artistic images (reported in Table 1), Artpedia provides a larger number of sentences, divided into visual and contextual through a manual annotation procedure. Moreover, to the best of our knowledge, this is the only dataset that contains two types of artistic sentences describing both the visual content of the paintings and other contextual information. For this reason, we devise a visual-semantic model capable of jointly discriminating between visual and contextual sentences of the same painting, and identifying which visual descriptions from a subset of textual elements (*i.e.* a subset of visual descriptions from different paintings) are associated to a specific painting.

To allow the training of our model and foster researches on this domain, we also provide training, validation and test splits obtained by proportionally dividing the number of paintings. Splits have been obtained with the constraint of balancing the distribu-

[1] https://spacy.io/

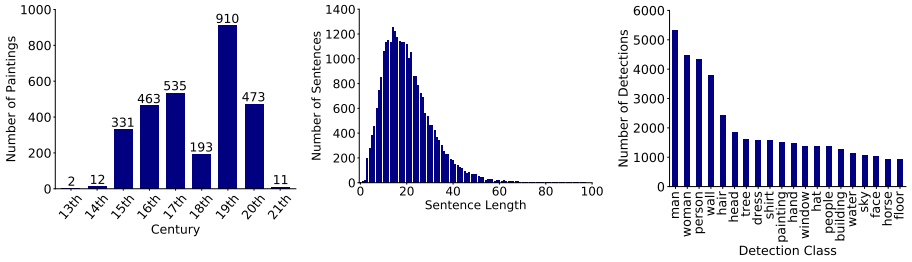


Fig. 2. Analyses on our Artpedia dataset. From left to right, we report the painting distribution over centuries, the sentence lengths distribution, and the most common detection classes.

Table 2. Number of paintings, visual and contextual sentences for each Artpedia split.

	Training	**Validation**	**Test**
Paintings	2,252	339	339
Visual sentences	7,109	1,036	1,028
Contextual sentences	14,822	2,134	2,083

tions over centuries and the number of visual sentences to maintain relevant statistics across the subsets. Table 2 reports the number of paintings for each split along with the corresponding number of visual and contextual sentences.

4 Aligning Visual and Contextual Sentences with Images

Cross-modal retrieval is characterized by two main tasks: when the query is a textual sentence, the objective is to retrieve the most relevant images, while with an image as a query, the objective is to retrieve the most relevant sentences. The goal is to maximize recall at K, the fraction of queries for which the most relevant item is ranked among the top K retrieved ones. Besides, our setting leverages the presence of visual and contextual sentences, and takes into account this difference when computing the latent alignment within a single page. In the following, we refer to a page as an element of our Artpedia dataset comprising an image and its visual and contextual sentences. Our goal is not only to maximize recall, but also to distinguish the two types of sentences associated to a painting.

In a nutshell, our model firstly maps image regions and sentence words into a joint embedding space. Then, it computes a cross-attention mechanism divided in two branches, where one attends to words with respect to each image region, while the other attends to image regions with respect to each word. This mechanism computes a similarity score for each branch between an image and a sentence. During training, the similarity score is used to minimize two loss functions: our intra-page loss, which strives to rank the sentences associated to a single image, bringing near its visual sentences and pushing away its contextual ones, and the inter-page triplet ranking loss that takes into account all images and their visual sentences as in standard cross-modal retrieval settings.

4.1 Similarity function

As mentioned before, the similarity is computed with a cross-attention mechanism that comprises two distinct branches: image-to-text and text-to-image attention, inspired by [15,25]. Since the two branches are similar, diversified only by the input order, we only describe the first one.

Firstly, given an image I, we extract salient regions such that each of them encodes an object or other entities, and project them into the joint embedding space, obtaining a final set of regions $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}, \boldsymbol{v}_i \in \mathbb{R}^D$. Also, given a sentence T composed of n words, encoded with a word embedding strategy, we project each word into the joint embedding space thus obtaining a vector $\boldsymbol{e}_j \in \mathbb{R}^D$ for each word j. Therefore, given an image I with k detected regions and a sentence T with n words, we compute the similarity matrix for all possible region-word pairs:

$$s_{ij} = \boldsymbol{v}_i^\top \boldsymbol{e}_j \quad i \in [1, k], j \in [1, n] \tag{1}$$

where s_{ij} represents the similarity between the region i and the word j. Since region and word features are ℓ_2 normalized, this product corresponds to a cosine similarity.

To attend words with respect to each image region, we compute a sentence-context vector for each region. The sentence-context vector $\boldsymbol{a}_i$ is a weighted representation of the sentence with respect to the region i of the image, where the similarities between the region i and the sentence words are used to weight each word as follows:

$$\boldsymbol{a}_i = \sum_{j=1}^{n} \alpha_{ij} \boldsymbol{e}_j \tag{2}$$

where

$$\alpha_{ij} = \frac{\exp\left(\lambda_s s_{ij}\right)}{\sum_{j=1}^{n} \exp\left(\lambda_s s_{ij}\right)} \tag{3}$$

and λ_s is a temperature parameter [4].

Finally, to evaluate the similarity of each image region given the sentence-context, we compute the cosine similarity between the attended sentence vector $\boldsymbol{a}_i$ and each image region feature $\boldsymbol{v}_i$:

$$R\left(\boldsymbol{v}_i, \boldsymbol{a}_i\right) = \frac{\boldsymbol{v}_i^\top \boldsymbol{a}_i}{\|\boldsymbol{a}_i\|} \tag{4}$$

To summarize the similarity between an image I and a sentence T, we employ average pooling between all image regions and the sentence-context vector:

$$R_{AVG}(I, T) = \frac{\sum_{i=1}^{k} R\left(\boldsymbol{v}_i, \boldsymbol{a}_i\right)}{k} \tag{5}$$

Likewise, the other branch follows the same procedure but swapping image regions and sentence words, computing a region-context vector for each sentence word, evaluating their cosine similarities and summarizing the final branch score in the same way. Finally, by averaging the similarity scores of the two branches, we obtain the final similarity score $S(I, T)$ between an image I and a sentence T.

4.2 Training

Intra-page loss. With the objective of correctly ranking visual and contextual sentences of a given image, we propose an intra-page loss function that learns the latent alignment between an image and its corresponding visual sentences within a single page of the dataset. Given an image I, a visual sentence T_V and a contextual sentence T_C, our intra-page loss is computed by taking into account the similarity score $S(I, T_V)$ between the image and the visual sentence and the similarity score $S(I, T_C)$ between the image and the contextual one:

$$L_{intra}(I, T_V, T_C) = [\alpha - S(I, T_V) + S(I, T_C)]_+ \tag{6}$$

where $[x]_+ = max(x, 0)$ and α is the margin. Note that, since this loss function is computed within a single page, both considered visual and contextual sentence are taken within the sentences of the given image I.

Inter-page triplet ranking loss. Since our final objective is not only to identify visual and contextual sentences of the same image, but also to associate matching image-visual sentence pairs within the entire dataset, we define an inter-page triplet ranking loss, which is typical of cross-modal retrieval methods.

As proposed in [7], we focus solely on the hardest negatives in the mini-batch. So that, our final inter-page triplet ranking loss with margin α is defined as follows:

$$L_{inter}(I, T) = \max_{\hat{T}} \left[\alpha - S(I, T) + S(I, \hat{T})\right]_+ + \max_{\hat{I}} \left[\alpha - S(I, T) + S(\hat{I}, T)\right]_+ \tag{7}$$

where only the hardest negative sentences $\hat{T}$ or hardest negative images $\hat{I}$ for each positive pair $S(I, T)$ are taken into account. In our case, a negative sentence $\hat{T}$ is a visual sentence of another image. Since this loss function aims to associate images and visual sentences of the entire dataset, contextual sentences are only used by our intra-page loss.

Final training objective. The final training loss is obtained by a linear combination of the two loss functions, *i.e.* $L = \lambda_w L_{inter} + (1 - \lambda_w) L_{intra}$, where $\lambda_w \in [0, 1]$ is a parameter that weights the contribution of the two losses. When λ_w is equal to 0, the training procedure only minimizes our intra-page loss, whilst when λ_w is equal to 1, all the attention is given to the inter-page triplet ranking loss.

5 Experimental evaluation

In this section, we experimentally evaluate the effectiveness of our approach by comparing it with different baselines. First, we provide all implementation details used in our experiments.

5.1 Implementation details

To encode image regions, we use Faster R-CNN [20] trained on Visual Genome [14,1], thus obtaining 2048-dimensional feature vectors. For each image, we exploit the top 20

detected regions with the highest class confidence scores. To project regions into the visual-semantic embedding space, we use a fully connected layer with a size of 512.

For the textual counterpart, we compare GloVe [19] with word embeddings learned from scratch. In both cases, the word embedding size is set to 300. Then, with the aim of capturing the semantic context of the sentence, we employ a bi-directional GRU with a size of 512, so that given a sentence with n words, the bi-directional GRU captures the context reading forward from word 1 to n and reading backwards from word n to 1, averaging the two hidden states to obtain the final embedding vector for each word.

To train our model, we use the Adam optimizer with an initial learning rate of 10^{-6} decreased by a factor of 10 after 15 epochs. In all our experiments, we use a batch size of 128 and clip the gradients at 2. Finally, the margin α and the temperature parameter λ_s are respectively set to 0.2 and 6.

5.2 Baselines

To evaluate our solution, we build different baselines to quantify both the effectiveness of using a cross-attention model and that of our intra-page loss. To this aim, we first exploit global features to encode images and sentences in place of multiple feature vectors for each image or sentence. In particular, to encode images, we extract 2048-dimensional feature vectors from the average pooling layer of a ResNet-152, while, to encode sentences, we feed word embeddings through a bi-directional GRU network and average the outputs of the last hidden state in both directions. After projecting both images and sentences into a common embedding space, the final similarity score between an image and a sentence is given by the cosine similarity between the two ℓ_2-normalized embedding vectors.

Furthermore, we compare the proposed intra-page loss function with respect to binary cross-entropy. Therefore, visual and contextual sentences are not projected into the same embedding space, but fed through a binary classification branch. In practice, each sentence is classified either as visual or contextual by concatenating the image and sentence embeddings and feeding them through two fully connected layers of size 512 and 1, respectively. For the cross-attention model, the image embedding is obtained by averaging the image region embedding vectors, while the sentence embedding is obtained by averaging the last hidden states of the bi-directional GRU in the two directions.

For both baselines, all other hyper-parameters and training details are the same as those used in our complete model.

5.3 Cross-modal retrieval results

We first evaluate the effectiveness of our model to identify and distinguish visual sentences with respect to contextual ones. Table 3 shows the results on the Artpedia test set in terms of average precision (AP). In particular, the results are obtained by training the models with λ_w equal to 0 (*i.e.* by only minimizing the intra-page loss or binary cross-entropy). As it can be seen, our intra-page loss function always obtains better performance with respect to the binary cross-entropy baseline either when exploiting global features to embed images and sentences or when using the cross-attention approach described in Section 4. Regarding the word embedding strategy, GloVe vectors

Table 3. Intra-page results in terms of Average Precision (AP).

Model	Word Embedding	AP
Global features with BCE loss	Learned	39.3
Global features with BCE loss	GloVe	40.8
Global features with intra-page loss	Learned	52.8
Global features with intra-page loss	GloVe	**55.3**
Cross-attention with BCE loss	Learned	42.6
Cross-attention with BCE loss	GloVe	41.7
Cross-attention with intra-page loss	Learned	86.3
Cross-attention with intra-page loss	GloVe	**88.5**

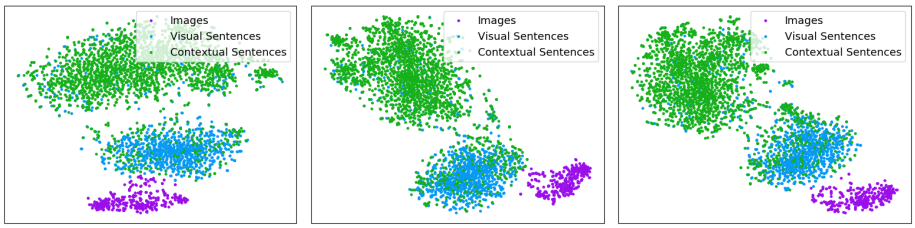
achieve better results with respect to word embeddings learned from scratch, probably due to the presence of peculiar words, typical of the artistic domain.

In Table 4, we show the performance of our complete model trained with various λ_w weights to differently balance the contribution of the two loss functions. In this case, the goal is not only to correctly distinguish between visual and contextual sentences of a given image, but also to find the corresponding visual sentences from a subset of other textual elements (*i.e.* visual sentences of different images). Results are reported in terms of recall@K ($K = 1, 5$) using a different number N of items from which perform retrieval. In details, given an image as a query, the retrieval of a textual element is performed from a subset of visual sentences of N different images (*i.e.* the visual sentences of the query and those of other $N - 1$ randomly selected images). Instead, given a textual query, the retrieval of an image is performed from a subset of N different images (*i.e.* the image linked to the query and other $N - 1$ randomly selected images from the Artpedia test set). We also report the results of identifying visual sentences with respect to contextual ones in terms of average precision. As it can be noticed, by increasing the λ_w weight, we obtain an increment of recall metrics with a slight drop of average precision values, in almost all considered combinations of features and word embeddings. Also in this case, the cross-attention mechanism and the GloVe word embeddings achieve better results than global features and learned word embeddings.

Finally, Figure 3 shows learned embedding spaces using the best model (*i.e.* cross-attention with GloVe word embeddings) using different λ_w weights. Since in this case images and sentences are composed of an embedding vector for each image region and word of the sentence, we represent each image or sentence by summing the ℓ_2-normalized embedding vectors of its image regions or words, and ℓ_2-normalized again the result. This strategy has been largely used in image and video retrieval works, and is known for preserving the information of the original vectors into a compact representation with fixed dimensionality [22]. To get a suitable two-dimensional representation out of a 512-dimensional space, we run the t-SNE algorithm [17], which iteratively finds a non-linear projection which preserves pairwise distances from the original space. As it can be observed, the higher the λ_w weight, the greater the distance between images and visual sentences in the embedding space, thus confirming the drop of average precision values when decreasing the importance of our intra-page loss during training.

Table 4. Cross-modal retrieval results with a different number N of retrievable items and with respect to different λ_w weights.

Model	Word Emb.	λ_w	AP	$N = 10$				$N = 50$				$N = 100$			
				Img-to-Text		Text-to-Img		Img-to-Text		Text-to-Img		Img-to-Text		Text-to-Img	
				R@1	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**
Global	Learned	0.25	44.9	9.4	36.3	7.6	50.0	2.4	8.3	1.3	8.9	0.6	5.0	0.5	3.6
Global	GloVe		43.1	12.4	35.4	8.5	48.7	2.7	9.1	2.2	11.1	0.6	5.3	1.8	5.5
X-Attn	Learned		85.9	15.3	40.4	17.5	61.9	2.7	13.3	4.2	16.7	2.1	8.0	2.2	9.0
X-Attn	GloVe		**88.2**	**19.8**	**44.0**	**22.7**	**69.6**	**8.6**	**22.1**	**6.1**	**23.6**	**4.4**	**15.9**	**4.0**	**14.8**
Global	Learned	0.50	50.2	9.4	38.1	9.9	50.7	1.8	10.0	2.0	10.5	0.6	6.2	1.1	5.1
Global	GloVe		46.0	8.8	37.2	9.8	48.9	1.2	10.0	2.0	10.1	1.8	4.1	1.0	4.4
X-Attn	Learned		85.2	11.5	40.1	17.4	61.0	3.2	13.6	3.8	18.7	1.2	7.7	2.4	9.9
X-Attn	GloVe		**87.5**	**26.3**	**54.3**	**21.2**	**69.7**	**8.8**	**27.7**	**7.5**	**22.9**	**6.2**	**18.6**	**4.1**	**14.1**
Global	Learned	0.75	53.4	10.6	38.3	10.4	50.0	2.4	10.6	2.3	11.6	1.5	5.6	1.4	6.2
Global	GloVe		44.9	10.9	34.2	8.9	47.7	1.8	8.6	1.8	9.3	0.9	4.4	0.7	4.6
X-Attn	Learned		84.6	10.9	37.5	18.5	64.3	2.7	10.0	5.1	20.1	1.2	7.1	2.9	11.4
X-Attn	GloVe		**86.5**	**29.5**	**57.2**	**23.7**	**71.2**	**13.6**	**31.9**	**5.8**	**23.1**	**8.6**	**22.7**	**4.1**	**13.6**



(a) $\lambda_w = 0.25$ (b) $\lambda_w = 0.50$ (c) $\lambda_w = 0.75$

Fig. 3. Comparison between visual-semantic embedding spaces obtained by training the model with different λ_w weights. Visualizations are obtained by running the t-SNE algorithm [17] on top of embedding vectors representing images and sentences (both visual and contextual).

6 Conclusion

In this paper, we have addressed the problem of cross-modal retrieval of images and sentences coming from the artistic domain. To this aim, we have collected and manually annotated a new visual-semantic dataset with visual and contextual sentences for each collected painting. Further, we have designed and evaluated a cross-modal retrieval model that jointly associates visual and textual elements, and discriminates between visual and contextual sentences of the same image. Experimental evaluations conducted with respect to different baselines have shown promising results and have demonstrated the effectiveness of our solution on both considered visual-semantic retrieval tasks.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018) 1, 2, 7

2. Baraldi, L., Cornia, M., Grana, C., Cucchiara, R.: Aligning text and document illustrations: towards visually explainable digital humanities. In: ICPR (2018) 3
3. Carraggi, A., Cornia, M., Baraldi, L., Cucchiara, R.: Visual-semantic alignment across domains using a semi-supervised approach. In: ECCV Workshops (2018) 3
4. Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: NeurIPS (2015) 6
5. Cornia, M., Baraldi, L., Cucchiara, R.: Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In: CVPR (2019) 1
6. Engilberge, M., Chevallier, L., Pérez, P., Cord, M.: Finding beans in burgers: Deep semantic-visual embedding with localization. In: CVPR (2018) 2
7. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In: BMVC (2018) 1, 2, 7
8. Garcia, N., Vogiatzis, G.: How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: ECCV Workshops (2018) 3
9. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016) 3
10. Gu, J., Cai, J., Joty, S.R., Niu, L., Wang, G.: Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: CVPR (2018) 2
11. Huang, Y., Wu, Q., Song, C., Wang, L.: Learning semantic concepts and order for image and sentence matching. In: CVPR (2018) 2
12. Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H.: Recognizing image style. In: BMVC (2014) 3
13. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. In: NeurIPS Workshops (2014) 1, 2
14. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Int. J. Comput. Vision **123**(1), 32–73 (2017) 4, 7
15. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV (2018) 1, 2, 6
16. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural Baby Talk. In: CVPR (2018) 1
17. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. J. of Machine Learning Research **9**(Nov), 2579–2605 (2008) 9, 10
18. Mao, H., Cheung, M., She, J.: Deepart: Learning joint representations of visual arts. In: ACM Multimedia (2017) 3
19. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: EMNLP (2014) 8
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015) 4, 7
21. Shen, X., Efros, A.A., Mathieu, A.: Discovering Visual Patterns in Art Collections with Spatially-consistent Feature Learning. In: CVPR (2019) 3
22. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. In: ICLR (2016) 9
23. Tomei, M., Baraldi, L., Cornia, M., Cucchiara, R.: What was Monet seeing while painting? Translating artworks to photo-realistic images. In: ECCV Workshops (2018) 3
24. Tomei, M., Cornia, M., Baraldi, L., Cucchiara, R.: Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-to-Image Translation. In: CVPR (2019) 3
25. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018) 2, 6