

Aligning Text and Document Illustrations: towards Visually Explainable Digital Humanities

Lorenzo Baraldi, Marcella Cornia, Costantino Grana, and Rita Cucchiara
University of Modena and Reggio Emilia
Email: {name.surname}@unimore.it

Abstract—While several approaches to bring vision and language together are emerging, none of them has yet addressed the digital humanities domain, which, nevertheless, is a rich source of visual and textual data. To foster research in this direction, we investigate the learning of visual-semantic embeddings for historical document illustrations, devising both supervised and semi-supervised approaches. We exploit the joint visual-semantic embeddings to automatically align illustrations and textual elements, thus providing an automatic annotation of the visual content of a manuscript. Experiments are performed on the *Borso d’Este Holy Bible*, one of the most sophisticated illuminated manuscript from the Renaissance, which we manually annotate aligning every illustration with textual commentaries written by experts. Experimental results quantify the domain shift between ordinary visual-semantic datasets and the proposed one, validate the proposed strategies, and devise future works on the same line.

I. 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 [1] and ResNet [2] have been exploited for extracting representations from images, and word embeddings [3], [4], [5] 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 [6], like automatic captioning [7], [8] and retrieval of images and textual descriptions [9], [10], [11], [12]. While all these strategies have been successfully applied on ordinary visual-semantic datasets, which feature natural images and text, none of them has been yet applied to the Digital Humanities domain.

To foster the research in this area, we explore the development of artificial systems capable of understanding the cross-reference between textual and visual information in documents, *i.e.* of understanding which parts of a plain text could be related to parts of the illustrations. Examples of possible applications of such systems are the alignment of commentaries with artistic books, or the alignment between textual museum guides and pictures of masterpieces.

One of the main open questions in this regard is related to the cross-domain generality, since up to now experiments and solutions have been proposed on general-purpose datasets only, where the state of the art of concept recognition methods is useful and well assessed. In the domain of arts and culture,

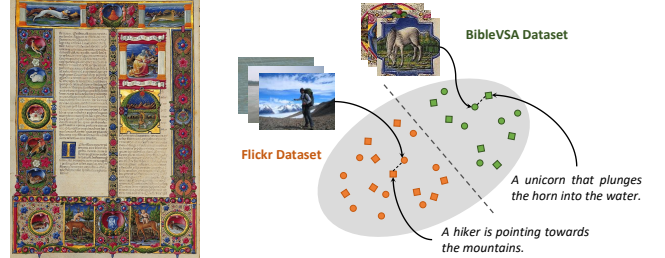


Fig. 1. Visual and textual data from the humanities are quite different from those addressed by visual-semantic datasets, posing significant challenges in the automatic understanding of arts and culture. Fostering this direction, we tackle the task of aligning *miniature illustrations* from illuminated manuscripts with *textual commentaries* written by experts in the field.

instead, both visual and textual elements are far from those of ordinary datasets. On the one hand, textual descriptions often contain technical language, with symbolic reminds, metaphors and artistic and historical connections; on the other hand, artistic illustrations are often far from naturalistic images.

In this paper, we investigate supervised and semi-supervised visual-semantic alignments in the context of historical manuscripts, with a cross-domain analysis (Fig. 1). Specifically, we consider the problem of understanding if a commentary of a digital artistic document has some parts referring to specific illustrations. In this context, we propose a new visual-semantic alignment dataset starting from the digitized version of the Borso d’Este Holy Bible, one of the most significant illustrated manuscripts of Renaissance. The dataset, which we name BibleVSA, provides the alignments between miniature illustrations and parts of text in the commentary, and can be used both to evaluate visual semantic embeddings, and to evaluate the alignment task. In the experimental section we show the challenging nature of this domain, and promising image-text alignment results, in both supervised and semi-supervised settings.

II. THE BIBLEVSA DATASET

The entire manuscript of the Borso d’Este Holy Bible consists of 320 high resolution digitized images ($3,894 \times 2,792$), for a total of 640 pages. To extract illustrations from each page, we employ the technique proposed in [13], which has been specifically tested on the same manuscript. Results have then been manually refined in order to have a highly accurate segmentation.

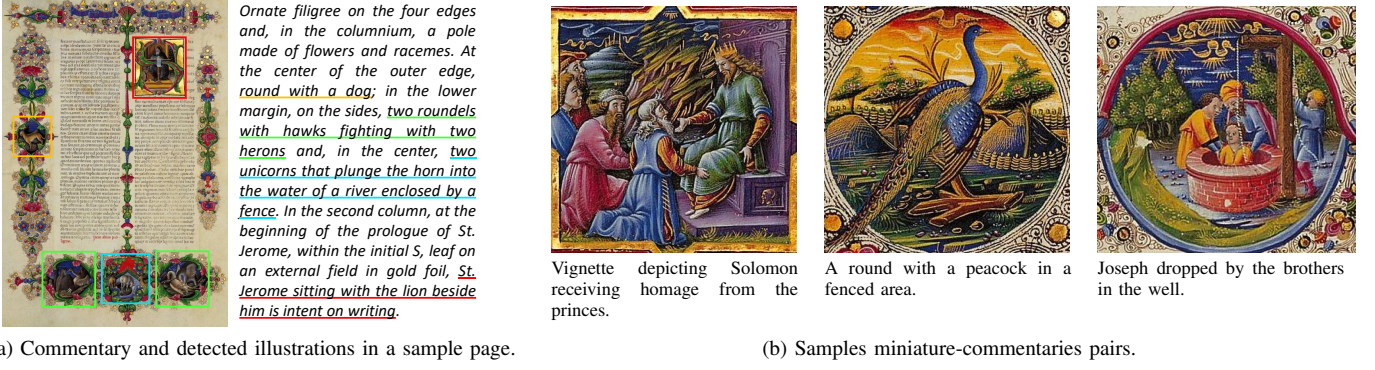


Fig. 2. Overview of the proposed BibleVSA dataset. On the left, a sample page from the Borso d’Este Holy Bible with the corresponding commentary and detected illustrations, while, on the right, samples illustration-caption pairs extracted from the commentaries.

Having a reliable annotation of the bounding box of each illustration, we then exploit an Italian commentary of the Borso d’Este Holy Bible. For each page, the commentary provides a set of paragraphs describing the visual content of each of the illustrations, the decorations of the page, and of the textual content itself. We must firstly notice that, on the one hand, the commentary provides descriptions of the Bible on a per-page basis, and it is therefore well suited as a weakly-supervised form of annotation. On the other hand, the commentary contains information which are unrelated to the task at hand, so the task of aligning each illustration with the commentary is more than just a partitioning of the text.

As an example, Fig. 2a reports the digitized version of one page, with its corresponding commentary and detected illustrations. The alignment is visualized by using the same color code for bounding boxes and textual strings. It can be noticed that the part of the text referring to illustrations is just a portion of the paragraph, while the remaining parts describe either the decorations (*ornate filigree on the four edges*) or the textual content (*at the beginning of the prologue of St. Jerome, within the initial S*). Also, descriptions jointly refer to the external frame and the content of the miniature, and often include names of people, saints and lineages.

We build a manual annotation of the alignments between each illustration and the commentary. Here we again employ a semi-automatic procedure: the original commentary is first automatically translated into English by using an off-the-shelf translator, and it is then manually checked. Each annotator is then asked to align each illustration with a piece of the commentary. The overall task is assisted by the fact that the commentary reports the position of each illustration inside the page (e.g., *at the center of the outer edge* in Fig. 2a); these parts are then removed from the final alignment, as they do not describe the content of the illustration.

The annotation process results in (a) a natural language caption of each illustration, which can be used for training visual-semantic embeddings, or caption generation architectures; and (b) the knowledge of which part of the commentary describes an illustration, which can be exploited for evaluating the alignment task. Fig. 2b reports three sample miniature-

description pairs from the dataset. As the reader can witness, the gap between the visual and the textual elements is significantly higher than in usual visual-semantic datasets, both for the complexity of the illustrations, and for the high-level semantics of the captions.

Overall, the datasets consists of 2,282 annotated illustrations. Considering its twofold application (for training visual-semantic embeddings and for solving the alignment task inside a single page), we build train, validation and test splits. Firstly, all the illustrations found in pages with a single miniature are placed in the training set, to avoid trivial validation and testing cases in the alignment scenario, and enriching the training set with useful samples for training embeddings. Then, we split the remaining pages placing them in the three sets according to a 60-20-20 ratio. This results in 1,671 training, 293 validation and 307 test image-caption pairs.

III. LEARNING VISUAL-SEMANTIC EMBEDDINGS

The task of aligning illustrations with textual elements in documents requires the ability to compare visual and textual data in this particular domain. We adopt the strategy of creating a shared embedding space, in which both textual and visual features can be projected and compared using a distance function.

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

In accordance to previous works [9], to project those representations in a common semantic space we perform a linear projection followed by a ℓ_2 -normalization step, so that the embedding space lies on the ℓ_2 unit ball:

$$f(i, \mathbf{w}_f, \mathbf{w}_\phi) = \ell_{2,norm}(\mathbf{w}_i^\top \phi(i, \mathbf{w}_\phi)) \quad (1)$$

$$g(c, \mathbf{w}_g, \mathbf{w}_\psi) = \ell_{2,norm}(\mathbf{w}_c^\top \psi(c, \mathbf{w}_\psi)), \quad (2)$$

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 then be compared in the joint embedding space by computing the dot product (*i.e.* the cosine similarity) between their projections, so that the similarity between an image i and a caption c becomes

$$s(i, c) = f(i, \mathbf{w}_f, \mathbf{w}_\phi) \cdot g(c, \mathbf{w}_g, \mathbf{w}_\psi). \quad (3)$$

Clearly, the utility of the joint embedding space is maximized when it exhibits suitable cross-modality matching properties, *i.e.* when distances in the embedding space correspond to meaningful distances in both modalities, and when corresponding pairs are matched in the embedding space. When this is verified to some extent, the embedding space acts as a bridge between the two modalities, and makes it possible to retrieve captions describing a query image, and images described by a query caption by identifying the closest neighbors in both modalities.

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. An alternative approach is that of learning cross-domain embedding spaces: in this setting, the paired supervision from one domain is exploited, together with the knowledge of the target domain, to limit the need of paired training data on the new domain.

With the joint objective of showcasing the features of the proposed dataset, and of closing the loop between images and text in such a complex domain, we explore both the aforementioned directions. In the first case, we exploit the fact that the BibleVSA dataset is sufficiently large to learn from it, while in the latter case, the strategy has the additional benefit of quantifying how much of the knowledge learned from ordinary datasets is transferable to the humanities domain.

A. The supervised way

In order to learn an embedding space with suitable cross-modality properties, we exploit the training set of BibleVSA to train the parameters of the model according to a Hinge triplet ranking loss with margin α :

$$\begin{aligned} \ell(i, c) = & \sum_{\hat{c}} [\alpha - s(i, c) + s(i, \hat{c})]_+ + \\ & + \sum_{\hat{i}} [\alpha - s(i, c) + s(\hat{i}, c)]_+ \end{aligned} \quad (4)$$

where $[x]_+ = \max(0, x)$. In the equation above, (i, c) is a matching illustration-caption pair (*i.e.*, such that c describes the content of i , and i represents the content of c), while \hat{c} is a negative caption with respect to i (such that \hat{c} does not describe i), and \hat{i} is a negative image with respect to c (such that c 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.

A recent work by Faghri *et al.* [9] has demonstrated that, in ordinary visual-semantic datasets, it is beneficial to replace the sums in Eq. 4 with maximum, so to consider only the most violating non-matching pair, leading to state of the art results on ordinary visual-semantic datasets. In Sec. V, both these approaches will be tested.

B. A semi-supervised approach

Instead of relying on the knowledge of matching and non-matching pairs on the BibleVSA dataset, we can also limit ourselves to shrinking the gap between the two modalities, while exploiting the supervision given by a second dataset. In practice, this is done by matching 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 [14], [15], [16], 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}_κ endowed with a kernel κ , and can be used as an additional loss term:

$$\mathcal{L}_{mmd} = \|\mathbf{E}_{i \sim \mathcal{I}} [f(i)] - \mathbf{E}_{c \sim \mathcal{C}} [g(c)]\|_{\mathcal{H}_\kappa}^2, \quad (5)$$

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

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\sigma \|\mathbf{x} - \mathbf{y}\|^2). \quad (6)$$

At training time, we sample two mini-batches of samples, one from the supervised set and a second one from the unsupervised dataset (*i.e.* BibleVSA). The back-propagated loss is then the sum of the supervised loss \mathcal{L} on the supervised set, plus the MMD loss \mathcal{L}_{mmd} approximated over the batch from the unsupervised set.

IV. ALIGNING COMMENTARIES AND ILLUSTRATIONS

Having obtained a distance function between visual and textual elements, either with a supervised or a semi-supervised approach, we can exploit it to tackle the alignment task between miniatures and commentaries. Given that the annotation has been provided on sub-strings, it would be necessary to match each possible sub-string of the commentary against each given illustration.

Being the aforementioned strategy computationally prohibitive, we approximate this task by ranking every sentence in the commentary given a query illustration from the same page. This roughly corresponds to the original task, under the hypothesis that each sentence describes either a single illustration, or is unrelated to the miniatures. Sentences are extracted from the original text by using an off-the-shelf NLP software.

Formally, for each illustration i found in a page, we rank the set of sentences \mathcal{C} in the commentary according to $s(i, c)$,

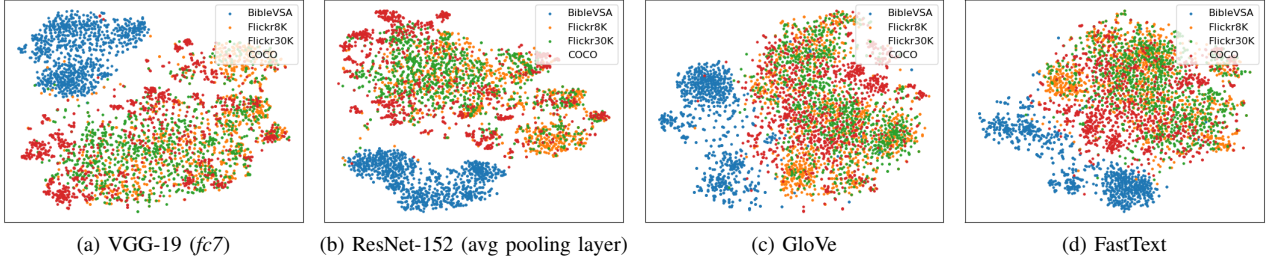


Fig. 3. Comparison between the visual and textual features of ordinary visual-semantic datasets (Flickr8K, Flickr30k, COCO) and those of the BibleVSA dataset. Visualization is obtained by running the t-SNE algorithm on top of the features. Best seen in color.

having chosen $c \in \mathcal{C}$. We refer to Section V-E for a quantitative and qualitative analysis of the alignment task.

V. EXPERIMENTAL EVALUATION

In the following, we evaluate proposed dataset and methodologies. To compare and contrast the features of the BibleVSA dataset, we also employ ordinary visual semantic datasets, which we describe in the following.

A. Datasets and implementation details

Beside the BibleVSA dataset, we employ Microsoft COCO [17] and Flickr30K [18], which both contain image-sentences pairs. We follow [7] to obtain train, validation and test splits.

To encode input images, we use two different convolutional networks: the VGG-19 [1] and the ResNet-152 [2]. We extract image features from the *fc7* 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 the fine-tuning of the image encoder, we set the input image size to 224×224 .

For encoding image descriptions, instead, we use a GRU network [19]. We set the dimensionality of the GRU and of the joint embedding space D to 1024, while the input size of word embeddings D_ψ is set to 300. In our experiments, we use either a text encoder on one-hot vectors or different pre-trained word embeddings (such as Word2Vec [3], GloVe [4], FastText [5]) as input of the GRU.

All experiments are performed by using the Adam optimizer with a learning rate of 0.0002 for 15 epochs and then decreased by a factor of 10 for other 15 epochs. We set the margin α to 0.2 and the size of the mini-batch to 128. For the semi-supervised approach, we set the σ parameter of the Gaussian kernel to 1.

B. Analysis of the BibleVSA dataset

To get an insight of characteristics of the BibleVSA dataset with respect to its visual and textual content, we analyze the distribution of features obtained from CNNs and word 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 Word2Vec, GloVe or FastText). 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 works, and is known for preserving the information of the original vectors into a compact representation with fixed dimensionality [20].

Fig. 3 shows the distributions of visual and textual features of all datasets. To get a suitable two-dimensional representation out of a (respectively) 4096, 2048 and 300-dimensional space, we run the t-SNE algorithm [21], which iteratively finds a non-linear projection which preserves 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, except for some clusters in the COCO dataset which do not overlap with the other distributions. The BibleVSA dataset, on the contrary, has a completely different distribution, according to both modalities and all feature extractors. We got very similar visualizations when using Word2Vec embeddings, so we do not report them for reasons of space. This underlines, on the one hand, that the BibleVSA dataset is not just another visual-semantic dataset, but defines a completely new domain. On the other hand, instead, this motivates the low performance of existing models when trained on this dataset (see next sections).

C. Evaluation of supervised Visual-Semantic embeddings

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

In Table I we report the performance obtained by the models described in Sec. III-A, using VGG-19 and ResNet-152, as well as all the three word embeddings strategies. *VSE* indicates the model depicted in Eq. 4, while *VSE++* indicates the model from [9]. Results on the COCO dataset are obtained by averaging over 5 folds of 1K test images. As it can be seen, the considered models perform comparably to some recent works on the field, when using ordinary datasets. For space reasons, we only report the results of some of the most interesting

TABLE I
SUPERVISED CAPTION AND IMAGE RETRIEVAL RESULTS, USING DIFFERENT IMAGE AND SENTENCE FEATURE EXTRACTORS.

Dataset	Model	Word Emb.	Caption Retrieval			Image Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
COCO	<i>sm-LSTM</i> [12]	-	53.2	83.1	91.5	40.7	75.8	87.4
	<i>Embedding Network</i> [10]	-	54.0	84.0	91.2	43.3	76.8	87.6
	<i>VSE (ResNet-152)</i>	GloVe	52.0	83.1	92.0	39.6	76.1	87.9
	<i>VSE++ (ResNet-152)</i>	GloVe	58.4	87.3	93.5	44.4	78.6	88.9
Flickr30K	<i>sm-LSTM</i> [12]	-	42.5	71.9	81.5	30.2	60.4	72.3
	<i>DAN (ResNet-152)</i> [11]	-	55.0	81.8	89.0	39.4	69.2	79.1
	<i>VSE (ResNet-152)</i>	GloVe	41.6	72.7	82.7	31.1	62.3	73.7
	<i>VSE++ (ResNet-152)</i>	GloVe	49.9	78.1	86.9	37.5	67.0	77.3
BibleVSA	<i>VSE (VGG-19)</i>	-	8.1	24.8	38.1	8.5	25.4	36.2
	<i>VSE (VGG-19)</i>	Word2Vec	7.5	24.8	38.8	8.5	25.4	38.4
	<i>VSE (VGG-19)</i>	FastText	9.8	30.3	43.6	8.1	26.4	39.4
	<i>VSE (VGG-19)</i>	GloVe	6.8	26.7	42.0	7.2	29.6	41.7
	<i>VSE (ResNet-152)</i>	-	10.4	30.0	40.7	8.8	25.4	41.0
	<i>VSE (ResNet-152)</i>	Word2Vec	12.7	30.9	44.3	10.1	30.3	43.3
	<i>VSE (ResNet-152)</i>	FastText	11.1	31.9	45.9	11.4	31.6	45.9
	<i>VSE (ResNet-152)</i>	GloVe	10.4	29.0	43.6	12.4	33.9	43.6
	<i>VSE (VGG-19 fine-tuned)</i>	-	11.4	40.4	53.7	11.4	39.1	55.0
	<i>VSE (VGG-19 fine-tuned)</i>	Word2Vec	11.4	41.0	58.3	13.4	42.7	59.3
	<i>VSE (VGG-19 fine-tuned)</i>	FastText	13.0	37.1	57.0	15.6	42.3	61.2
	<i>VSE (VGG-19 fine-tuned)</i>	GloVe	12.1	40.4	56.4	16.0	42.0	62.5

combinations of hyper-parameters. The reader, nevertheless, can find the complete set of experiments on-line¹.

It is also noticeable *VSE* performs always better than *VSE++* here, contrary to what happens in Flickr30K and COCO. Word embeddings share an important role, as shown by the performance improvement reported in the table.

Overall, the results on the proposed dataset underline that training supervised visual-semantic embedding can be a good strategy for aligning visual and textual data in this domain. The fact that the numeric results are significantly lower than those on Flickr30K and COCO underlines the challenging nature of the proposed dataset.

D. Evaluation of semi-supervised embeddings

Table II shows the results when the model is trained using the proposed semi-supervised approach. For all experiments we report the results obtained with and without the MMD loss defined in Eq. 5 and training on ordinary visual-semantic datasets; the numbers without MMD, practically, quantify the performance of models trained on Flickr30K or COCO in this particular domain. As it can be seen, the use of the MMD loss is beneficial for the model performance in all considered experiments. Also in this case, we only report the most interesting results according to $R@K$ metrics.

Figure 4 shows the learned embedding spaces for the COCO and the BibleVSA datasets when the model is trained on COCO with and without the MMD loss. As it can be noticed, by using the MMD, the distribution of the learned image embeddings of the BibleVSA dataset matches with that of the

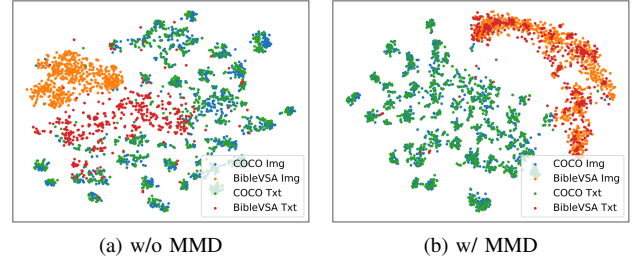


Fig. 4. Comparison between the embedding spaces learned with and without the MMD loss (t-SNE projections). Best seen in color.

textual counterpart thus confirming the effectiveness of the proposed semi-supervised strategy.

E. Ranking evaluation

Finally, we evaluate the alignment task between miniature illustrations and commentaries. Here, we consider all possible illustration-sentence pairs within each page of the BibleVSA test set. For each illustration of a given page, we rank the extracted commentary sentences according to the learned embedding space. Table III shows the ranking results on the BibleVSA test set in terms of mean average precision and accuracy top- K ($K = 1, 2, 3$), when using the best supervised model. As it can be seen, the learned embedding space allows to effectively align image and textual parts with an overall mAP greater than 85%. Also, the correct sentence is ranked first or second in the 77% and 91% of the cases, respectively. Fig. 5 shows some qualitative examples with two correct ranking results (first and second images) and a failure case (third image).

¹<http://aimagelab.ing.unimore.it/bible-vse>

TABLE II
SEMI-SUPERVISED CAPTION AND IMAGE RETRIEVAL RESULTS.

Evaluation Dataset	Training Dataset	Model	Word Emb.	Caption Retrieval			Image Retrieval		
				R@1	R@5	R@10	R@1	R@5	R@10
BibleVSA	COCO	<i>VSE (VGG-19) w/o MMD</i>	-	3.3	13.1	21.3	2.6	10.8	22.0
		<i>VSE (VGG-19) w/ MMD</i>	-	1.6	23.0	36.1	3.6	13.8	27.2
		<i>VSE (ResNet-152) w/o MMD</i>	-	3.3	6.6	14.8	3.0	11.5	17.7
		<i>VSE (ResNet-152) w/ MMD</i>	-	11.5	29.5	45.9	4.6	19.7	30.2
		<i>VSE (ResNet-152) w/ MMD</i>	GloVe	3.3	23.0	44.3	5.2	15.7	29.5
BibleVSA	Flickr30K	<i>VSE (VGG-19) w/o MMD</i>	-	1.6	1.6	8.2	1.6	10.8	19.7
		<i>VSE (VGG-19) w/ MMD</i>	-	4.9	16.4	34.4	4.6	13.1	23.6
		<i>VSE (ResNet-152) w/o MMD</i>	-	0.0	6.6	18.0	3.3	12.1	22.3
		<i>VSE (ResNet-152) w/ MMD</i>	-	4.9	27.9	39.3	3.6	15.4	25.6
		<i>VSE (ResNet-152) w/ MMD</i>	GloVe	6.6	18.0	31.1	2.6	12.8	21.3

TABLE III
ALIGNMENT OF COMMENTARIES AND ILLUSTRATIONS RESULTS.

Model	Word Emb.	mAP	Accuracy		
			top-1	top-2	top-3
<i>VSE (VGG-19)</i>	-	83.5	69.4	86.3	83.5
<i>VSE (ResNet-152)</i>	-	85.6	72.6	88.9	92.1
<i>VSE (ResNet-152)</i>	Word2Vec	86.7	75.2	89.3	93.0
<i>VSE (ResNet-152)</i>	FastText	87.6	77.5	90.6	92.6
<i>VSE (ResNet-152)</i>	GloVe	86.6	73.9	91.2	93.5



- 1) *within a circle, the Prophet Nathan speaks with King David.*
- 2) *within an oval framed by leafy scrolls, King David, after Nathan's prophecy, prays to God.*
- 3) *round with a bird resting on a branch.*



- 1) *the Estense company of the leopard with dragon tail and wings placed near a date palm.*
- 2) *the Estense company of the unicorn placed near a date palm.*
- 3) *Saint John speaks to the crowds in front of a door of the city.*
- 4) *a shield with the Este coat of arms.*



- 1) *round with a mallard in a lake landscape.*
- 2) *round with a heron in a lake landscape.*
- 3) *the Prophet Nahum announces the punishment of God against Nineveh, represented in the background.*

Fig. 5. Qualitative ranking results according the image-sentence similarities in the embedding space. The ground-truth sentence is highlighted in italics.

VI. CONCLUSION

In this work, we tackled the task of aligning miniature illustrations from illuminated manuscripts with textual commentaries written by experts in the field by providing a new visual-semantic dataset for this particular domain. We explored supervised and semi-supervised visual-semantic alignments with an extensive cross-domain analysis. Experimental results on both ordinary datasets and our BibleVSA validated the proposed strategies and confirmed that a cross-domain transfer can be possible.

ACKNOWLEDGMENTS

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] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *ANIPS*, 2013.
- [4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.
- [5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.
- [6] L. Baraldi, C. Grana, and R. Cucchiara, "Recognizing and presenting the storytelling video structure with deep multimodal networks," *IEEE TMM*, vol. 19, no. 5, 2017.
- [7] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *CVPR*, 2015.
- [8] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Visual saliency for image captioning in new multimedia services," in *ICME Workshops*, 2017.
- [9] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.
- [10] L. Wang, Y. Li, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *arXiv preprint arXiv:1704.03470*, 2017.
- [11] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *CVPR*, 2017.
- [12] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal lstm," in *CVPR*, 2017.
- [13] C. Grana, D. Borghesani, and R. Cucchiara, "Automatic segmentation of digitalized historical manuscripts," *Multimed. Tools Appl.*, vol. 55, no. 3, 2011.
- [14] Y.-H. Hubert Tsai, Y.-R. Yeh, and Y.-C. Frank Wang, "Learning cross-domain landmarks for heterogeneous domain adaptation," in *CVPR*, 2016.
- [15] Y.-H. Hubert Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning Robust Visual-Semantic Embeddings," in *CVPR*, 2017.
- [16] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," *CVPR*, 2017.
- [17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.
- [18] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, vol. 2, 2014.
- [19] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014.
- [20] G. Tolias, R. Sircé, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," in *ICLR*, 2016.
- [21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *J. of Machine Learning Research*, vol. 9, 2008.