Towards Cycle-Consistent Models for Text and Image Retrieval

Marcella Cornia[1] Lorenzo Baraldi[1] Hamed R. Tavakoli[2] Rita Cucchiara[1]

[1]University of Modena and Reggio Emilia, Italy [2]Aalto University, Finland

Abstract. Cross-modal retrieval has been recently becoming an hot-spot research, thanks to the development of deeply-learnable architectures. Such architectures generally learn a joint multi-modal embedding space in which text and images could be projected and compared. Here we investigate a different approach, and reformulate the problem of cross-modal retrieval as that of learning a translation between the textual and visual domain. In particular, we propose an end-to-end trainable model which can translate text into image features and vice versa, and regularizes this mapping with a cycle-consistency criterion. Preliminary experimental evaluations show promising results with respect to ordinary visual-semantic models.

Keywords: cross-modal retrieval, cycle consistency, visual-semantic models

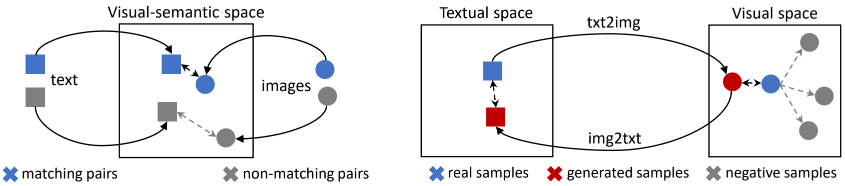
1 Introduction

Matching visual data and natural language is an important challenge for multimedia as it enables a large variety of different applications ranging from retrieval, visual question answering, image and video captioning. One of the core challenges in this scenario is that of enabling a cross-modal retrieval, *i.e.* the retrieval of visual items given textual queries, and vice versa.

Current cross-modal retrieval methods often rely on the construction of a common multi-modal embedding space in which project data from the two modalities (*i.e.* images and text) [1–3]. The retrieval, in this case, is then carried out by measuring distances in the joint space, which should be low for matching text-image pairs and higher for non-matching pairs. While this approach leads to very good results, it is not the only possible solution.

Here, we foresee a different approach and address the problem of retrieving images and captions as a translation from the image domain to the textual domain and vice versa. In the first direction, an image i (usually, represented with a feature vector x) is converted to a textual representation $\tilde{s}$ of its content; in the latter direction, a sentence s is converted into an image feature $\tilde{x}$ which reflects its meaning.

Fig. 1 visually describes the idea: a learnable architecture translates textual data to a suitable representation in a visual domain, and visual features back to the textual domain. The overall architecture is trainable end-to-end: generated



(a) Ordinary visual-semantic models (b) Our approach

Fig. 1. Instead of relying on a joint embedding space, we address the problem of cross-modal retrieval as that of learning a translation between the textual and visual domain, with a reconstruction objective which keeps the overall process cycle-consistent

visual features are required to be realistic with respect to positive and negative image samples, and a cyclic constraint is imposed to guarantee that the forward and backward translation are feasible at the same time and consistent.

2 Cycle-consistent Retrieval

We introduce a cycle-consistent text and image retrieval network which operates a translation between the textual and the visual domains. Under the model, input captions can be translated to proper image features, and image vectors can be translated back to the textual domain. Exploiting this translation capability, a reconstruction constraint makes sure that the reconstructed text is similar to the original one. The overall architecture is shown in Fig. 2.

From text to image (`txt2img`). The first part of the architecture consists of a visual-semantic model which can transform a sentence s in a meaningful vector in the image feature space, $\tilde{x}$. Word are represented with one-hot vectors that are embedded with a linear embedding, which can be either learned end-to-end together with the model, or pre-trained using another word embedding model, like Word2Vec [4], GloVe [5] or FastText [6]. Under the model, words are consumed by a GRU layer.

We train this model with a cost function which encourages the generated image vector to be close to the one of an image which has been described by the same caption. To this aim, we define a similarity function inside the image feature space (*e.g.* the cosine similarity), and apply a hinge-based triplet ranking loss commonly used in image-text retrieval [1, 2].

From image to text (`img2txt`). While sentences can be projected into an image feature space, the second component of the model translates image vectors x into the textual space by generating a textual description $\tilde{s}$. This roughly corresponds to an image captioning model in which the image is treated as the first input of an LSTM-based recurrent model.

At each iteration, the hidden state is linearly projected to the dimensionality of the vocabulary, and a softmax activation is then used to produce a probability distribution over the vocabulary. For each input image vector, the model

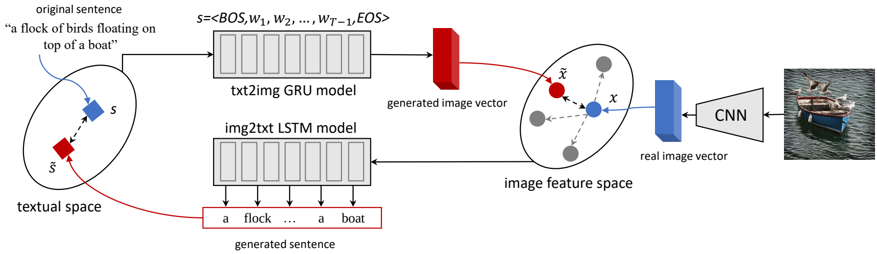


Fig. 2. Architecture of our model

generates the corresponding textual representation $\tilde{s}$ composed by the words produced at each time-step of the LSTM.

Closing the loop. The `txt2img` and `img2txt` models defined above realize the forward and backward translations between the image and the textual domain. Due to the diversity and high dimension of raw images, directly translating to and from the image domain would be intractable, therefore both models operate in the space of image feature vectors extracted from a CNN.

The mapping between the two spaces is regularized with a cycle-consistency criterion, in which we require the feasibility of the forward and backward translation at the same time. In practice, we require that the projection of a generated image vector into the textual space should be similar to the text from which the vector originated, *i.e.*

$$\texttt{img2txt}(\texttt{txt2img}(s)) \approx s. \tag{1}$$

The similarity constraint imposed by Eq. 1 could be realized by taking into account the semantics of both sentences, either by evaluating a machine translation metric or by defining a network in charge of learning the similarity between two sentences. To keep the model simple and concentrate on the evaluation of the regularization power of the proposal, we realize Eq. 1 by computing the negative log-likelihood of generated words with respect to the words in s.

Implementation details. To encode input images, we extract feature vectors from the average pooling layer of a ResNet-152, thus obtaining an image dimensionality of 2048. For encoding image captions, since we do not project images and corresponding captions in a joint embedding space, we set the output size of the GRU to the same size of image embeddings (*i.e.* 2048). The dimensionality of word embeddings is set to 300. All experiments have been performed using the Adam optimizer with an initial learning rate of 2×10^{-4}.

3 Experimental Results

We show preliminary evaluation results for the proposed approach, employing rank-based performance metrics $R@K$ ($K = 1, 5, 10$) for text and image retrieval. In particular, $R@K$ computes the percentage of test images or test sentences for

Table 1. Experimental results of our model on the Flickr8K and Flickr30k dataset using different word embeddings

Model	Word Emb.	Flickr8K						Flickr30K					
		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
`txt2img`	-	25.7	54.8	69.0	15.8	41.6	56.0	36.9	67.0	78.2	22.8	50.0	63.3
Ours	-	**28.2**	**57.4**	**71.1**	**17.5**	**44.6**	**59.0**	**41.7**	**68.9**	**78.9**	**23.8**	**51.3**	**64.0**
`txt2img`	GloVe	29.2	60.2	74.5	19.2	46.7	61.7	36.4	67.4	78.4	22.8	50.7	64.2
Ours	GloVe	**32.2**	**62.7**	**76.2**	**19.9**	**48.8**	**62.8**	**41.1**	**68.9**	**79.0**	**23.0**	**51.3**	**64.6**
`txt2img`	FastText	29.8	58.7	73.4	17.9	45.8	60.3	37.7	66.0	77.8	22.1	49.8	63.4
Ours	FastText	**32.2**	**61.4**	**74.1**	**19.2**	**47.5**	**62.0**	**40.8**	**68.5**	**79.1**	**23.5**	**51.3**	**63.8**
`txt2img`	Word2Vec	28.1	58.0	71.3	17.1	44.1	58.7	35.9	66.4	76.9	**22.3**	49.7	62.9
Ours	Word2Vec	**30.9**	**59.4**	**72.7**	**18.9**	**46.8**	**61.2**	**41.2**	**68.2**	**79.3**	**22.3**	**50.7**	**63.7**

which at least one correct result is found among the top-K retrieved sentences, in the case of text retrieval, or the top-K retrieved images, in the case of image retrieval.

As a baseline, we consider the `txt2img` model, which removes the cycle-consistency regularizer and is therefore well suited to evaluate the claims of the proposal regarding the role of the cycle-consistent constraint. This, also, is practically equivalent to a visual-semantic embedding model in which the visual projector is the identity function.

Table 1 reports the results of our model on the Flickr8K [7] and Flickr30K [8] datasets using different word embedding strategies, together with that of the `txt2img` model alone. It can been observed that the performance of the complete model is always superior to that of the baseline, thus confirming the importance of translating backwards to the textual space and demonstrating the effectiveness of our promising solution.

References

1. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
2. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
3. Wang, L., Li, Y., Lazebnik, S.: Learning Two-Branch Neural Networks for Image-Text Matching Tasks. IEEE TPAMI (2018)
4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: ANIPS. (2013)
5. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. In: EMNLP. (2014)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
7. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. JAIR **47** (2013) 853–899
8. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2** (2014) 67–78