Saint George on a Bike

MS12. State of the art of parallelisation

Document Information

Contract Number	2018-EU-IA-0104
Project Website	www.saintgeorgeonabike.eu
Contratual Deadline	31/05/2021
Nature	Report
Author	Joaquim Moré (BSC)
Contributors	
Reviewer	Maria-Cristina Marinescu (BSC)
Keywords	

Co-financed by yhe Connecting Europe Facility of the European Union

Table of Contents

1. Introduction 3

2. Types of aligned texts and tasks 4

On the one hand, there are image-text pairs where the texts strictly describe what is going on in the image. On the other hand, there are pairs where the texts contain data that is not related to the content of the image. These texts can describe partially what is represented in the image but also convey information that is not related to the visual content of the image at all. 4

2.1 Descriptive aligned texts 4

In this document, we consider as ***descriptive aligned texts*** those texts that strictly describe or refer to what is going on in their corresponding aligned images. Descriptive aligned texts are also referred as *captions* by Vinyals et al. (2015) and Xu et al. (2015).... 4

In the following table, we present some of the most well known databases with images aligned with descriptive texts. 4

2.2 Accompanying texts of online images 5

3. Issues of aligned texts 6

4. Issues of aligned cultural heritage texts 7

References 8

1. Introduction

As said in Section 4.1 of the MS3 document, aligned image and metadata pairs are one of the types of data input taken into consideration. In this document, we present the state of the art regarding the alignment of images and texts. First, we will deal with the tasks where alignment of images and texts is involved. Then we will see how the content of the texts aligned depends on the task. When the task is training an image recognition system, the aligned texts describe estrictly what is going on in the image. For other tasks such as searching for images, the description of what is going on in the picture is often relaxed and the texts provide data about the author, the social and cultural contextualisation of the image and even the experience of the author when producing it.

Since the SGoaB concern is training an image recognition system we will show the resources available of aligned texts that describe what is going on in the images and the difficulties of researchers when trying to detect the sentences that describe what is going on in the picture when the aligned text also provides other kinds of data. Finally, we will deal with the available resources of aligned image and textual data when describing cultural heritage images. Most of these resources come from art collections and we will see how their aligned texts do not strictly describe what is going on in the image. The current approaches in the detection of sentences with visual content are not applied in the description of cultural heritage images so we will present some of the issues that arise when dealing with works of art.

2. Types of aligned texts and tasks

On the one hand, there are image-text pairs where the texts strictly describe what is going on in the image. On the other hand, there are pairs where the texts contain data that is not related to the content of the image. These texts can describe partially what is represented in the image but also convey information that is not related to the visual content of the image at all.

2.1 Descriptive aligned texts

In this document, we consider as ***descriptive aligned texts*** those texts that strictly describe or refer to what is going on in their corresponding aligned images. Descriptive aligned texts are also referred as *captions* by Vinyals et al. (2015) and Xu et al. (2015).

In the following table, we present some of the most well known databases with images aligned with descriptive texts.

Data source	**Textual content**	**Tasks**	**Provided by**
ImageNet	Wordnet category the image depicts (over 14 million labeled images)	Object detection	Human subjects on Amazon Mechanical Turk.
MS Coco dataset	Five natural language descriptions of each image (*captions* in the MS Coco dataset terminology)	Automatic generation of descriptions of what is in an image (*image captioning* in MS Coco dataset terminology)[1]	Human subjects on Amazon Mechanical Turk.
Visual Question Answering (VQA) dataset	614K questions and 6.1M answers associated with 205K images	Answering natural language questions about any image	Human subjects on Amazon Mechanical Turk.
Visual Genome	5.4M descriptions of visual relations between entities in different regions of the image (region	Understanding and reasoning from the contents of an image (image search, question/answering,	Human subjects on Amazon Mechanical Turk.

[1] https://github.com/ntrang086/image_captioning

	descriptions) and 1.7M visual question and answers	robotic interactions)	
Open Images	3.3M annotations indicating pairs of objects in particular relations (e.g. "woman playing guitar", "beer on table"), object properties (e.g. "table is wooden"), and human actions (e.g. "woman is jumping"[2]	Scene understanding	

Notice that the descriptive aligned texts are used for object detection, image captioning and the understanding and reasoning from the contents of an image. In order to fulfill these tasks the aligned texts must fulfill some conditions such as the ones stated in the generation of captions for the Coco Dataset. The subject that wrote captions for this dataset had to follow these instructions (Chen et al. 2015):

- Describe all the important parts of the scene.
- Do not start the sentences with “There is.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentences should contain at least 8 words

2.2 Accompanying texts of online images

Texts accompanying online images are those found in Flickr. These texts are often written by the photographers who made the photographs.

The large amount of texts accompanying online images raised the hope that descriptions of what is going on in an image can be automatically generated by retrieving the texts accompanying images similar to a query image (Farhadi et al.,

[2] https://storage.googleapis.com/openimages/web/factsfigures.html

2010, Ordonez et al. (2011), Kuznetsova et al. (2012)). Other tasks where texts accompanying online images are retrieved is image searching on the Web.

3. Issues of aligned texts

Coco dataset captions, written manually by following the instructions listed above, are regarded as the *canonical form* of a descriptive text. However, accompanying texts found in online images contain non visual content and information not related to what is going on in the image. For example, objects not found in the image such as *camera* or *lamp* may be present in the text explaining how the photographer took the photograph. When the text accompanies an image in an online art collection, visual content is often embedded in references to biographical data of the author and other extraneous information (see MSXX). Besides, aligned texts in online images often do not follow the canonical form instructions. For example texts may refer to details that are not detectable.

Dodge et al. (2012) study the cases of gaps between the content of the images and the content of texts aligned with them. They also aim to separate visual text from non-visual text in order to generate corpora with aligned texts describing the content of the image. Their definition of *visual text* is as follows: *A piece of text is visual (with respect to a corresponding image) if you can cut out a part of that image, paste it into any other image, and a third party could describe that cut-out part in the same way.*

This is also the aim of the SGoaB *descriptive sentence classifier* (see MSXXX) but instead of transforming the sentences in the original aligned text into sentences the closer to Coco canonical captions as possible, Dodge et al. (2012) formalize the definition of visual text in order to apply in use cases (training object detectors, building image search engines and automatically generating captions for images). The visual texts contain linguistic expressions that are more likely to be found in descriptive texts rather than in generic texts. An approach that transforms the original sentence into a descriptive-like sentence is the one by Kuznetsova, 2012. They apply the task called *image caption generalization* in order to release a corpus with 1 million image-texts pairs where context alignment between the two is tighter. The idea is to transform the original text into a simplified and more general version, free of noise, which is easier to be aligned to the visual content of images.

4. Issues of aligned cultural heritage texts

The efforts of creating simplified versions of original texts or selecting visual texts have been mostly applied for databases such as Flickr. However, texts aligned with images in museum collections have some peculiarities that raise new issues and challenges.

A text aligned with a cultural heritage image, even when the text is simplified or classified as visual, is unlikely to be a canonical caption in the Coco dataset sense. The texts aligned with images of Jesus, saints, kings, and so on do not fulfill the *not to use proper names* command. This is why our descriptive sentence classifier, before classifying the text as descriptive or not, generates a version of the original text where proper nouns are replaced with *a person*.

There are aligned texts such as the ones found in Europeana collections where the transformation to a Coco canonical caption has been applied quite easily. After this transformation, as explained in MSxxx, the descriptive text classifier, applied to a sample of Europena aligned texts, reached a F1 score of 0.9 (0.53 if the textual images are not transformed). However, in other collections like El Prado's, the results are not so good. The reason is that aligned texts in these collections have not been written by photographers that explain what is depicted and how, but by scholars whose main concern is to shed new light to aspects beyond what is assumed to be already seen by the spectator. So even when a sentence refers to an entity depicted in the image, the reference is in a non visual text. For example, let's see the text *Antonello´s painting is technically virtuoso, combining a meticulous calligraphy of northern origin —visible in the landscape and in Christ´s hair- with a monumental treatment of the anatomy*. In this text, the reference to Christ's hair is not in a visual text but in an argumentative text where the scholar provides an example that supports a statement: the virtuosity of the painting. The reason why Christ's hair is not found in a visual text is the fact that the scholar assumes the spectator has already seen it. How to detect references to visual contents from non visual texts and use them in the generation of a training corpus for image captioning and image searching is a very complex issue that the SGoaB project has to tackle with.

References

Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.:Microsoft COCO captions: Data collection and evaluation server. arXiv preprintarXiv:1504.00325, 2015

A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. IJCV, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár Microsoft COCO: Common Objects in Context. 2014

Peter Young, Alice Lai, Micah Hodosh, Julia Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014

Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. (2015)

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S.,Bengio, Y.: Show, attend and tell: Neural image caption generation with visualattention. arXiv preprint arXiv:1502.03044 (2015)

Anderson, Peter James, Vision and Language Learning: From Image Captioning and Visual Question Answering towards Embodied Agents, PhD Thesis, College of Engineering and Computer Science, The Australian National University, 2018

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh,D. VQA: Visual question answering*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.;Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L., Visual genome: Connecting language and vision using crowdsourced dense image annotations*, 2016*

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database*. IEEE Computer Vision and Pattern Recognition (CVPR), 2009*

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daume III, Alex Berg, and Tamara Berg. Detecting visual text*. In Proceedings of the 2012 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Montréal, Canada, June. Association for Computational Linguistics., 2012

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs*. In Neural Information Processing Systems (NIPS). 2011*

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young1, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences for images*. In European Conference on Computer Vision. 2010*

P Kuznetsova, V Ordonez, AC Berg, TL Berg, Y Choi. Generalizing Image Captions for Image-Text Parallel Corpus. ACL. 2013