Saint George on a Bike

MS4. Methodology for aligning the visual and textual resources

Document Information

Action Number	2018-EU-IA-0104
Project Website	https://santgeorgeonabike.eu/
Contractual Deadline	01/03/2021
Milestone	MS4
Document type	Report
Author	Maria-Cristina Marinescu (BSC)
Contributor(s)	Artem Reshetnikov (BSC)
Reviewer	Joaquim Moré (BSC)
Keywords	Visual description, classifier, symbolic content, manual annotation, crowdsourcing

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

Table of Contents

1. Introduction

As described in Section 4.1 of the MS3 (Use cases) document, the types of data input that we take into consideration are either images of paintings, or aligned image and metadata pairs, when available. Metadata may be of many types, e.g. labels, captions or, in the best of cases, descriptions, and it may, or may not be exclusively related to the visual content of the image. For instance, metadata may refer to the painter, the epoch or style, the medium, and descriptions may talk about the context of the painting such as details about the painter, the location, or the reason for which the piece was executed.

On the other hand, the output may consist of complex semantic representations (graphs or textual captions), but simpler levels of representation should also be present in the final output and are in fact the main focus of the methodologies we have been developing until now. Most of these techniques center on object classification and object detection, which can generate textual labels or semantic resources. More recently we have started to work on visual relation detection; this would eventually allow us to generate knowledge graphs, besides textual labels for relations. We are in the process of testing various approaches for caption generation, although this - as well as, to some extent, knowledge graphs - may turn out exceedingly complex for the available data.

The next section explains in detail what data is - or could become - available for training or even testing. Based on that, we explain the main challenge we face and some of the options we see going forward. Lastly, we focus on the current approach for aligning textual descriptions and images. This is a key step to be able to generate good captions from a cultural heritage viewpoint which are also correct in a language sense - although the resulting phrases may be simple rather than sophisticated.

2. Textual resources corresponding to paintings

2.1 Resources for textual content

We have downloaded and analyzed many potential sources of metadata associated with paintings between the 12th and the 18th century. Below we present a summary with the corresponding types of metadata available, and the issues the metadata in description form presents for our alignment and caption generation tasks.

Data source	**Metadata types**	**Issues with descriptions**
Europeana collection	Tags, descriptions	Symbolic descriptions Different languages Not descriptive enough
IconClass Dataset	Tags	

WikiArt Online Gallery	Tags, descriptions based on articles from Wikipedia	Symbolic descriptions Not descriptive enough
Pharos collection	Tags	
Wikimedia Commons	Tags, descriptions are based on articles from Wikipedia	Symbolic descriptions Not descriptive enough
VisArt v.2	Tags	
Pinterest	Tags	
BING	Tags	
Rijksmuseum	Tags, descriptions	Different languages Not descriptive enough
PeopleArt dataset	Tags	
British Museum	Tags, descriptions	Symbolic descriptions Not descriptive enough Phrases are complex
Musee d'Orsay	Tags, descriptions	Different languages Not descriptive enough
Museo del Prado	Tags, descriptions	Symbolic descriptions Different languages Not descriptive enough Phrases are complex
National Gallery of Art	Tags, descriptions	Symbolic descriptions Not descriptive enough
Web Gallery of Art	Tags, descriptions	Symbolic descriptions Different languages Not descriptive enough
WikiData	Tags, descriptions are based on articles from Wikipedia	Not descriptive enough Phrases are complex
Wikipedia	Tags, descriptions are based on articles from Wikipedia	Not descriptive enough Phrases are complex
Getty Museum	Tags	

Out of all these sources, we are only using Europeana, British Museum, National Gallery of Art, Museo del Prado (the English descriptions and machine translations of the Spanish descriptions), and the English descriptions from Musee d´Orsay.

We cannot use Web Gallery of Art because of license issues, both for images and for descriptions. Likewise, we cannot use Wikipedia articles because they have very little embedded structure, the texts are complex and extensive, and there is no apparent pattern to extract only the image description part. Additionally, some museums only provide descriptions in the official language of the country of origin, which may not be handled well by automatic machine translation.
For the purpose of training a classifier to filter out statements that are likely not describing the visual content of images out of more generic descriptions - explained in section 2.3 - we additionally use manually generated descriptions from the MS COCO[1] and Open Images V4[2] datasets. The advantage of these data sources is that they are good and quite complete descriptions; the disadvantage is that they do not correspond to cultural heritage images, with the issues we comment in section 2.3, mainly that they handle a different set of objects or the same objects with different (modern) shapes.
As a last mention, Rijksmuseum is a data source that we just started to explore, so we haven't done any experiments with it yet.

2.2 Issues with the existing descriptions

We now explain in some detail what are the concrete issues we found with the descriptions in the various datasets. These refer to the last column in the above table.

- Different languages (e.g. Louvre, Prado, Europeana): The descriptions may be provided only in languages other than English and automatic machine translation is not usually good enough. For generating image descriptions it is necessary to have the training descriptions all in one language that NLP methods can handle properly within the same language framework. Generation of a Knowledge Graph does not have this problem.
- Not descriptive enough: The metadata is about the historical context, the painter, the medium or style, rather than being a description of the visual content.
    - A secondary issue is exhaustiveness, which refers to the fact that even for visual descriptions, usually not all entities and actions are mentioned in the text.
- Phrases are complex: NLP techniques don't perform well for complex phrase structure.
- Symbolic descriptions: When the paintings have symbolic or iconographic content, many of the visual descriptions refer to the symbols that are represented rather than the concrete entities and actions.

The bottom line is that the sheer number of descriptions that do not have these issues seems really quite small for paintings between the 12th and the 18th century.

2.3 Current approach for alignment

In this section we focus on textual descriptions as a form of input to be aligned with the corresponding images. These are not captions in the traditional usage of the word, as they are meant to describe the visual content of the image, while captions are usually shorter, not exhaustive, and sometimes may not even fully relate to the image.

As it became clear from the previous sections, descriptions are the exception rather than the norm in cultural heritage repositories. Datasets such as MS COCO, Open Images V4 or Flickr30k[3] do have descriptions associated with the images, but these depict every-day life activities and objects. Given that a neural network model generates the most likely next word (of a description) based on the sequence of words so far, it will not be able to output words outside the training context. That is, it won't be able to correctly generate descriptions about past objects and actions using a model trained on new concepts. This is the reason for which the descriptions that exist - for pictures rather than paintings - cannot be used as input. We need to generate this body of descriptions to train on, or generate them by other means and test them - most likely manually. To summarize, the options we see at this point for description generation are the following:

1. Obtain descriptions and align them to paintings
2. Generate descriptions by manual annotation
3. Generate descriptions automatically and test them manually

Only options (1) and (2) allow training based on image-text pairs. Option (3) relies on the visual relation detection mechanisms - in addition to NLP technology - to generate short sentences describing the existing entity-relation triples.

The rest of this document focuses on option (1), which fits our original intention. Further iterations of this document will go into more details and possibly also describe options (2) and (3), depending on the course that the work will take. Section 2.6 in MS6 (System and module-level architecture development) gives additional details of the methodology and tool we are describing below. The purpose of this tool is to filter out sentences present in painting descriptions that are irrelevant to the content depicted in the image. Concretely, it is implemented as a classifier that either accepts or rejects individual statements as possible descriptions. This is an approach for reusing extensive descriptions available from some museums, Europeana, and Wikimedia datasets.-

These descriptions often talk about the historical context, the life of the artist, or give information about the technique, medium, or style of the painting. Some description of the scene is also available, although it's usually not exhaustive. To add to the problem, the phrases are often stylistically complex and typical of art professionals rather than normal speech. This presents a challenge to Natural Language Processing models, which are

best applied to relatively simple statements and syntax. Our goal is to build a classifier which can successfully discriminate between descriptive and non-descriptive statements that refer to image content.

We are now at the second version of this classifier. The first version was developed using a Long Short Term Memory (LSTM) architecture and it was framed as a binary classification task (good/bad), where a good example was similar to manually generated MS COCO and Open Images descriptions. Following this approach, we were able to select a total of about 4000 good image/caption pairs from about 60K on the Europeana and Web Gallery of Art websites.

Due to the small proportion of resulting good pairs, in the second version of the classifier we changed the classification methodology by introducing a general language model (BERT) adapted to this classification task. The workflow follows the traditional stages in a neural network with a BERT model adapted to a text classification task similar to that of sentiment analysis.

The accuracy of the classifier based on this transformer model was very good when classifying sentences in a test set containing COCO, Open Images, and Wikipedia sentences (F1=0.99) - the three sources for good and bad descriptions in the training set, but it didn't work well for the 1000 Europeana sentences tested (F1=0.53). To improve this score, we automatically replace concepts from the cultural domain with the more frequently used synonym or hypernym in the model. This type of tuning is based on the fact that the training dataset mostly contains words frequently used in common language, while Europeana is a cultural heritage website that uses language and concepts that are specific to this domain and not to real life. After this pre-processing step, we obtain an F1 score of 0.90. All sentences classified as descriptive are then replaced by the original version using cultural heritage vocabulary.

3. Discussion: future work

We structure future work to advance in two different directions:

1. Improving the classifier tool

   We will be looking at new ways to improve the precision and recall of our tool. One idea is to evaluate whether visual content descriptions are always in present tense, while the rest of the descriptions are usually not, as it seems at first inspection. We will also look at ways of incorporating information about the image to better filter out statements.

2. Obtaining more painting descriptions

We are actively looking for new sources of descriptions on updated cultural heritage websites and are in the process of contacting various museums. Concretely, we have started to analyze data from the Rijksmuseum museum, we have scheduled meetings with two of the largest museums in Spain, and we are exploring the possibility of downloading data and metadata from the Netherlands Institute for Art History (RKD) via our Europeana partners. About one week ago, the Louvre museum opened its entire digitized art collection to the public; there are about 500K works and at a first sight, some have descriptions associated with them although it's not clear how many. There doesn´t seem to be an API available on their site, but we will take a close look at this to assess the interest and feasibility of exploiting this metadata.

The other approach we are aware of to obtain image descriptions is manual annotation or correcting automatically generated annotations by humans. An implementation of this are crowdsourcing campaigns. Our intention is to set up this type of action in the future, most probably originating with Europeana.

References

1. *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.
2. *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
3. P*eter Young, Alice Lai, Micah Hodosh, Julia Hockenmaier,* From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014