MS 9. Generation of semantically aligned images and image descriptions

Version 0.2

Documentation Information

Contract Number	2018-EU-IA-0104
Project Website	https://saintgeorgeonabike.eu/
Contractual Deadline	31.05.2020
Nature	
Author	Maria-Cristina Marinescu (BSC), Joaquim More (BSC), Artem Reshetnikov (BSC)
Contributors	Antoine Isaac (EF)
Reviewer	Sergio Mendoza (BSC)
Keywords	image description, caption generation, visual relationships

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

Change Log

Version	Author	Description Change
V0.1	Joaquim More (BSC), Artem Reshetnikov (BSC)	First Iteration
V0.2	Maria Cristina Marinescu	Final Version

Table of Contents

1. Introduction

Saint George on a Bike's activity 6 focuses on the alignment of visual and textual resources that will form the training data sets. Considering the scope of the selected use cases (MS3), these resources include, as foreseen in the project description, various kinds of metadata that can be considered "textual": topics, titles, etc. Some of that metadata is already aligned with images at the level required, either in the original data sources selected in Activity 5 (e.g., in the form of subject tags in the original metadata) or as an additional activity conducted there (manual alignment between classes and image regions that show these classes), as described in MS8 “Database for the selected use cases with the description of the datasets”. There is however a data category remaining, for which sourcing aligned data has proven to be much more difficult (as discussed in MS4 “Methodology for aligning the visual and textual resources”): providing images with longer textual descriptions (captions). The deep neural networks could, at least theoretically, generate complex statements, if the training set contains this style of description. MS4 introduces three options for generating such image captions. This document describes the first two, which focus on building training datasets for deep learning algorithms, while document MS13 describes the third option, which is about the design, training and application of the description generation model per se:

1. Obtain descriptions and align them to paintings
2. Generate descriptions by manual annotation

Automatic generation of captions via deep learning relies on collecting a proper dataset of image/caption pairs for training. Manual data collection as reflected by option 2 is a complex and time-consuming process which usually involves many annotators. Additionally, not all annotators are equally good or careful, and an evaluation must be made of the collected descriptions. We describe this option in section 3. Option 1 is significantly different in the form it collects the training data, but it is fraught with other challenges we describe in section 2. Here we report mainly on the quality and quantity of the fragments of descriptions that were deemed to refer to the visual content of images, extracted from the Europeana collections based on a methodology built around the idea of the caption classifier described in Section 2.6 of MS6 “System and module-level architecture development”.

2. Option 1: Obtain descriptions and align them to paintings

Text describing the paintings can be collected from open data sources such as Europeana, Web Art Gallery, or Wikipedia as described in MS8 “Database for the selected use cases with the description of the datasets”. The main problem of this approach is the quality of this data for the task. Taking into account that such portals, which can be seen as aggregators, collect data from different sources, the resulting data is noisy. In reality, it contains a lot of information not directly related to the visual content of the paintings, such as mentions of the painter or the style, or the history of the artwork. In general descriptions suffer from a range of general issues, some of them

documented by the Europeana Data Quality Committee[1] . Therefore, it cannot generally be directly used for training deep learning models. It needs to be processed appropriately to only retain those phrases that relate directly to the visual content. The caption classifier described in section 2.6 of MS6 “System and module-level architecture development” is a tool we have developed to filter out irrelevant information from available descriptions. However, the precision of this approach needs a thorough evaluation, and it can only be done manually. We are in the process of deciding what is the best evaluation process and how to implement this in practice. The rest of this section describes the Caption seed extractor, an advanced version of the caption classifier.

The goal of the Caption seed extractor is to extract relevant caption seeds from Europeana descriptions. A caption seed is a piece of text that describes a visual relation between entities depicted in a painting. The piece of text has this format: *A_B_C*, where *A* and *C* are instances of SGoaB classes (e.g: monk, book) and *B* is a visual relation. The visual relation words come from Open Images [1] and Visual Genome [2] relation labels. Caption seeds have a dual purpose:

- As a training set for a Neural Network
- As references for the evaluation of a visual detector

The alignment process consists of four steps:

1. Prepare Europeana descriptions: The Europeana texts in languages different from English were machine translated into English by using the Google Translate API[2] .

2. Sentence tokenization: The texts that accompany the pictures are tokenized[3] , i.e. they are split into sentences, each paired up with the link to the image.

3. Sentence "tuning": Each sentence is preprocessed in order to detect visual relations between SGoaB classes,

    a. Substitution of named entities referring to persons with class ‘person’: Most of the sentences contain named entities referring to persons (e.g.: *St Demetrius on a horse*). Except a few like “*Judith*” and “*God the Father*”, these named entities are not SGoaB classes. However, they refer to the SGoaB class *person*. Therefore, the named entities referring to a person (but not in the list of SGoaB classes) are replaced with “*person*” in a sentence

St Demetrius on a horse => Person on a horse

We prepared a dataset of proper nouns present in DBpedia[4] . If one of the noun phrases present in the sentence is in this dataset, then it is replaced with *person*. The noun phrases present in the sentences are provided by the Spacy function *noun_chunks*[5] .

Noun phrases with a compound referring to the role of the person are also replaced with person (e.g: Emperor Charles V)

[1] cf https://pro.europeana.eu/project/data-quality-committee#problem-patterns
[2] Using Google Translate allows us to deploy a first version of the process relatively easily, especially relying on the existing language detection mechanism to handle the cases when language information is not present for the descriptions found in the original metadata. In the future we will try to use the European Commission's eTranslation service.
[3] The tokenizer used is Blingfire, https://github.com/microsoft/BlingFire
[4] The dataset was prepared by taking all the DBpedia articles whose title corresponds to an entity with type 'person', 'angel', 'mythological character' and so on. For example, 'Pontius Pilate'. Christian names and surnames are also listed separately. The list (very large and still in need of a last check-up) will be published on Gitlab.
[5] Natural language processing package for Python https://spacy.io/usage/linguistic-features

b. Substitution of personal pronouns with 'person': Third person singular personal pronouns such as *he* or *she* are replaced with *person*. This is how we can collect caption seeds when the SGoaB entity class is referred to with a pronoun.

c. Substitution of adnominal clauses[6] with continuous forms: Verbs in -ing form in clausal modifiers of nouns (adnominal clauses) are replaced with their present continuous form. Adnominal clauses are detected by using the Spacy dependency parser.

A person riding a horse => A person is riding a horse

The reason for this replacement is the fact that the Spacy dependency parser does not consider the verb in the adnominal clause as the root (head) of the clause, whereas the parser does this when the verb is in its continuous form. Later we will explain why it is important to identify the head of clauses.

4. Collecting seed captions for each painting: Once the sentences are preprocessed, the seed captions of each sentence are collected and paired with the link to the image. The procedure follows:

a. Segmentation: The segments of the sentence containing two entities from the SGoaB class list (E1, E2) are collected (The segment begins with a SGoaB class entity and ends with another SGoaB entity).

Miracle of person with the dragon. => person with the dragon

b. Segment parsing: The segments are parsed with the Spacy dependency parser if the syntactic head of the segment (noun, verb) is a SGoaB class or is in the visual relations items.

c. Seed caption generation: For each head, the heads of the following syntactic complements are picked up

Head part of Speech	**Syntactic complement**
Verb	Subject (in active and passive sentences) Direct object Attribute Dative Prepositional object Adjectival complement Clausal complement
Noun	Prepositional phrase

The seeds are generated by putting the head and the complements in order. The schema when the head is a verb is as follows:

[6] An adnominal clause is a clause that modifies a nominal. For example, in 'A person riding a horse', 'riding a horse' is an adnominal clause because it modifies the nominal 'person'.

<Subject - Head - Other complements> (e.g: person rides horse)

The schema when the head is a noun is the following:

<Head - Prepositional phrase> (e.g: person with dragon)

The order of the other complements follow the order in which the Spacy dependency parser identifies them.

3. Option 2: Generate descriptions by manual annotation

In the article “Overcoming Challenges In Automated Image Captioning” [3] Youssef Mroueh defines “a lack of diversity in the generated captions” as the main challenge of the caption generation task. For the Cultural Heritage domain, this problem is even more significant because of the iconographical meaning of paintings, specific types of objects and relationships between them. Therefore, a good captioning dataset should have captions that are able to represent the differences in the perceived content (i.e. diverse captions). Also, it should have multiple captions per sample in order to represent the different ways of writing the same information (i.e. rephrasing) and allowing for a more thorough assessment of the performance of the captioning method [4].

According to the article “Image Captioning in the Wild: How People Caption Images on Flickr” by Philipp Blandfort [5], most (if not all) image captioning datasets are created by employing crowdsourcing. Crowdsourcing provides several benefits over other ways of data collecting (web scraping, etc.), for instance a better quality of the captions and the possibility of simultaneous annotation. Using an existing crowdsourcing platform provides the additional benefit of having an established base of users, i.e. potential annotators. One best-of-breed example of an image captioning dataset is the Flickr 8K dataset[6], which consists of 8092 images with five captions each, which were obtained by crowdsourcing. The dataset images were hand-selected. Selected images display actions and events to be able to provide full sentence captions instead of a simple list of objects. The annotators were pre-screened (by answering questions regarding grammar and image captioning), were required to be located in the US, and had to have an approval rate of 95% on previous tasks on the crowdsourcing platform.

We have already implemented a crowdsourcing campaign for data collection for the object detection model; this is described in section 2.2.3 of MS6 “System and module-level architecture development”. Based on evaluating the accessible data mentioned in Section 2.1 “Resources for textual content” of MS4, the approaches for data cleaning and the quality of the data, we decided to launch a crowdsourcing campaign for collecting captions for Europeana paintings, most likely using Mechanical Turk. We are in the process of implementing a pilot version of it for a reduced number of images, where the annotators are the project partners. The idea is to have a first proof of concept and understand the best way to explain the annotation task and address potential pitfalls, as well as prevent future evaluation quirks. The campaign will include four steps:

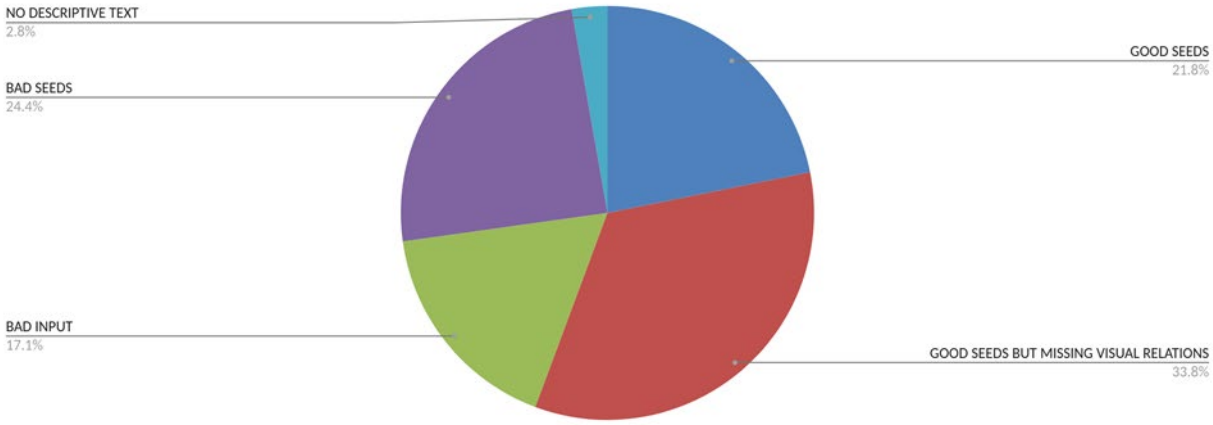
1. Prepare the subset of images that we need descriptions for
2. Choose the crowdsourcing platform and prepare the description of the task
3. Engage as potential annotators the registered users of the platform and provide the correct incentives

4. Evaluate, filter, and possibly correct errors in the captions provided by annotators in the previous step.

Additionally, we have applied for a Europeana Research Grant to implement another crowdsourcing campaign, this time with cultural heritage specialists. Besides serving for training and evaluation, we could compare the captions or caption seeds we could gather as result of this campaign with those collected from non-specialists via the crowdsourcing effort described here.

4. Discussion

Below we present the statistics for a first evaluation of the approach discussed in Section 2, i.e. the alignment of caption seeds extracted from Europeana texts and images.



We took a sample of the Europeana descriptions that contained SGoaB classes detected with object detection. We evaluated how the seed captions detected matched the real seed captions in the description. As one can see, we were able to extract good caption seeds only in about a fifth of the Europeana descriptions. In about a fourth of these, the seeds were deemed bad for one of many reasons: bad disambiguations (e.g. “hands over” vs hands, “at the head of … [an army]” vs head), missing relative pronouns, missing classes in coordinated noun phrases, anaphoras, class non-existent in SGOAB, etc. In a significant number of these cases, the English is not fully correct, so the results are partial. In almost 3% of the cases, there is no descriptive text that refers to the actual visual content of the actual image. A staggering 17% of the cases present bad inputs, i.e. sentences that are critically incomplete or don´t read at all as well formed English. Most of them are due to typos, bad punctuation, and careless writing. One important issue we encountered is that many Europeana descriptions consist of Google-style result snippets (e.g: "the bishop wears…"). The rest of the cases, almost 34%, generated good but incomplete seeds, given that relations between classes not in SGOAB are not returned. The following table details the current results: https://3.basecamp.com/4181566/buckets/12787583/uploads/3867621035.

The caption seed extractor discovered important cases for which the precision can be improved; we plan to do this in the near future. We also plan to improve the preprocessing of the sentences,

for instance by reducing the visual relations to OpenImages visual relation labels. Expressions like resting on in “The hand resting on the helmet” would be replaced with the OpenImages visual relation label on.

The hand resting on the helmet => The hand on the helmet .

5. 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] 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

[3] Youssef Mroueh, Igor Melnyk, “Overcoming Challenges In Automated Image Captioning”,2019

[4] Samuel Lipping, Konstantinos Drossos, and Tuomas Virtanen, “Crowdsourcing a dataset of captions”, Detection and Classification of Acoustic Scenes and Events, 2019

[5] Philipp Blandfort, Tushar Karayil, Damian Borth, Andreas Dengel, “Image Captioning in the Wild: How People Caption Images on Flickr”, 2017

[6] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, Serge Belongie, “Learning to Evaluate Image Captioning”, 2018