MS8. Database for the selected use cases with the description of the datasets

Version 0.4

Documentation Information

Contract Number	2018-EU-IA-0104
Project Website	www.saintgeorgeonabike.eu
Contratual Deadline	31/08/2020
Nature	Report
Author	Artem Reshetnikov(BSC)
Contributors	Maria Cristina Marinescu (BSC)
Reviewer	Antoine Isaac (EF)
Keywords	Database, dataset, use cases

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

Change Log

Version	Author	Description Change
V0.1	Artem Reshetnikov (BSC)	First version
V0.2	Maria Cristina Marinescu (BSC)	Reviewed
V0.3	Antoine Isaac (EF)	Reviewed
V0.4	Ariadna Lobo (BSC)	Final

Table of Contents

1. Introduction ---3
2. Dataset ---3
  2.1 Europeana Collection ---3
  2.2 WikiArt ---3
  2.3 IconClass AI Test set---3
  2.4 Pharos ---3
  2.5 Museum d'Orsay ---4
  2.6 The British Museum ---4
  2.7 Wikimedia Commons---4
  2.8 Web Gallery of Art ---4
  2.9 MS COCO dataset---4
  2.10 Wikipedia---4

1. Introduction

This document presents a summary of the data sources that we have used so far to obtain images and captions for training our deep learning algorithms and classifier. Given that the techniques that we implemented so far focus on textual labels and captions, but we haven't yet tackled the task of semantic labels and graphs, we will probably have to consult additional data sources. More importantly, we need to enhance the size of our current datasets to at least 10K images, and image / caption pairs. This will certainly require identifying more (or better) datasources. We are starting to explore possible direct collaborations with museums and cultural heritage institutions, such as, for instance, RKD - one of the world's leading documentation and research institutes in The Netherlands.

2. Dataset

Our main dataset contains 4k images with annotation and additional metadata which was done manually. The dataset contains images from several databases. As was mentioned before we need to enhance the size of our current datasets to at least 10K images. The dataset is not published yet. Provided link is for internal use.

Link: https://github.com/areflesh/sgoab_dataset

2.1 Europeana Collection

Europeana is a web portal created by the European Union and contains digitized museum collections from more than 3,000 institutions across Europe. In the case of the *Saint Goerge on a Bike*, the Europeana collections were used to train the object detection model for iconographic classes, and also for a set of first experiments for caption generation.

Link: https://www.europeana.eu/es

2.2 WikiArt

WikiArt (formerly known as WikiPaintings) is an online, user-editable visual art encyclopedia. It contains both public domain and copyright protected artworks. Materials from this site were used for the object detection model for iconographic classes.

Link: https://www.wikiart.org/

2.3 IconClass AI Test set

Iconclass is a classification system that is designed for art and iconography. It is used for the description and retrieval of subjects in images such as artworks, book illustrations, and photographs. Iconclass is the most widely accepted classification system for visual documents and it is used by museums and art institutions worldwide. It contains 28,000 hierarchically ordered definitions that are divided into ten main categories, each of which has subcategories. Images of this datasource were used for the object detection model for iconographic classes.

Link: https://labs.brill.com/ictestset/

2.4 Pharos

PHAROS is an international consortium of fourteen European and North American art historical photo archives committed to creating a digital research platform allowing for comprehensive consolidated access to photo archive images and their associated scholarly documentation. Materials from this data source were used for the object detection model of iconographic classes.

Link: http://pharosartresearch.org/

2.5 Museum d’Orsay

Musée d'Orsay is a museum in Paris that aims to encourage research into the history of art in the second half of the 19th century. On this website, in addition to the collections catalog, the researcher can find resources which are continually updated: an index of memoirs and theses, an index of works (texts and images), etc. Materials from the data source were used for object detection of iconographic classes. This source of data is the only one that - as of today - provides paintings of 18 and 19 centuries which are necessary for the proper training of object detection taking into account the style of paintings.

https://www.musee-orsay.fr/

2.6 The British Museum

The British Museum is a public institution dedicated to human history, art, and culture. Its permanent collection of some eight million works is among the largest and most comprehensive in existence, having been widely sourced during a long time. It documents the story of human culture from its beginnings to the present. The main advantage of this source of data is an open online collection which can be scraped using an API provided by services of the museum. Materials from this data source were used for the object detection model of iconographic classes.

Link: https://www.britishmuseum.org/collection/galleries/prints-and-drawings-virtual-gallery

2.7 Wikimedia Commons

Wikimedia Commons (or simply Commons) is an online repository of free-use images, sounds, other media, and JSON files. The main advantage of the source is the size of the potential dataset which can be used in training. Materials from this data source were used for the object detection model of iconographic classes.

Link: https://commons.wikimedia.org/wiki/Main_Page

2.8 Web Gallery of Art

Web Gallery of Art is a searchable database of European fine arts and architecture (3rd-19th centuries), currently containing over 49.500 reproductions. The main advantage of datasets is captions provided manually and which can be used in captions generation in the future. We used their images as a testset for the caption classifier.

Link: https://www.wga.hu/

2.9 MS COCO dataset

We used The MS COCO dataset to select good caption examples for the training of the caption classifier. The MS COCO dataset is a large-scale object detection, segmentation, and captioning dataset. COCO has about 330K images, more than 200K of them labeled.

Link: https://cocodataset.org/#home

2.10 Wikipedia

We used Wikipedia to extract bad caption examples for the training of the caption classifier. Wikipedia is a free online encyclopedia, created and edited by volunteers around the world and hosted by the Wikimedia Foundation.

Link: https://www.wikipedia.org/