What AI can bring to GLAM? Experience of "Saint George on a Bike" project

Artem Reshetnikov, Barcelona Supercomputing Center
Barcelona, Spain
artem.reshetnikov@bsc.es

Abstract—"Saint George on a Bike" project proposes several novel approaches to enrichment of metadata(captions, tags, relationships between objects, iconographic description) for the Cultural Heritage domain, which relies on combining Deep Learning and semantic metadata about paintings. Working with cultural heritage presents challenges not existent for every-day images. Models for objects detection or caption generation are usually trained with datasets that contain correct descriptions of current images or labels for objects, which were generated manually. Apart from this conceptual problem, the paintings are limited in number and represent the same concept in potentially very different styles. Finally, the metadata associated with the images is often poor or inexistent, which makes it hard to properly to generate quality metadata. Our approach can assist in generation of metadata for different tasks. By taking into account an exiting metadata of Cultural heritage objects and additional techniques, we can generate tags, relationships between objects or descriptive text which is likely to be directly related to the scene depicted in an image.

Index Terms—NLP, Cultural Heritage, Deep Learning, Metadata

I. Introduction

The application of AI(Artificial Intelligence), and in particular deep learning approaches, to the cultural heritage domain has attracted significant attention in the last time. Most of the existing work focuses on automatic metadata annotation with information such as the author, medium, image classification by style, topic, etc. or the objects that were detected in images from open datasets. However, such types of metadata is not relevant for specific tasks such as generation of descriptions, improving of search engines or improving of communication with users of GLAM sites. Focus of Saint George on a Bike project is on generation metadata, which is related more specifically to cultural heritage domain, which can help to solve these problems. First of all, rich metadata would allow a visitor of a cultural heritage site or the user of a web-page to obtain a detailed description of an artwork and would facilitate a personalized interaction with GLAM institutions. Secondly, different types of metadata could be used to automatically generate explanations in catalogs, fuel search and browse engines, or fill in rich alt-tab descriptions on websites that cater to minorities such as visually impaired citizens. Generating metadata automatically can save a lot of time and labor for manual annotators[1].

The generation of metadata for paintings or images of cultural heritage objects is challenging compared to those corresponding to real world scenes, for several reasons. First, the metadata for paintings often contain irrelevant information beyond the image content such as the life of a historical person, information about the place where the object was found, or the life of the painter. For example, the caption of the artwork in Figure 1 contains the name of the book where it has been mentioned, the language of the book and the medium of the artwork [1].

This information is obviously not relevant for the visual content of the painting. In that context, it is challenging to generate good metadata related to the scene. The second challenge is the quality of the data and the data collection process. This makes it difficult to train with a dataset similar in size to datasets containing real life images, such as MS COCO[2]. Lastly, metadata for cultural heritage objects from data providers often contain incomplete sentences or can be in different languages. Data aggregators can't distinguish such cases during data incorporating, as a result, they end up as part of the

[1]https://tinyurl.com/ypfbsr66

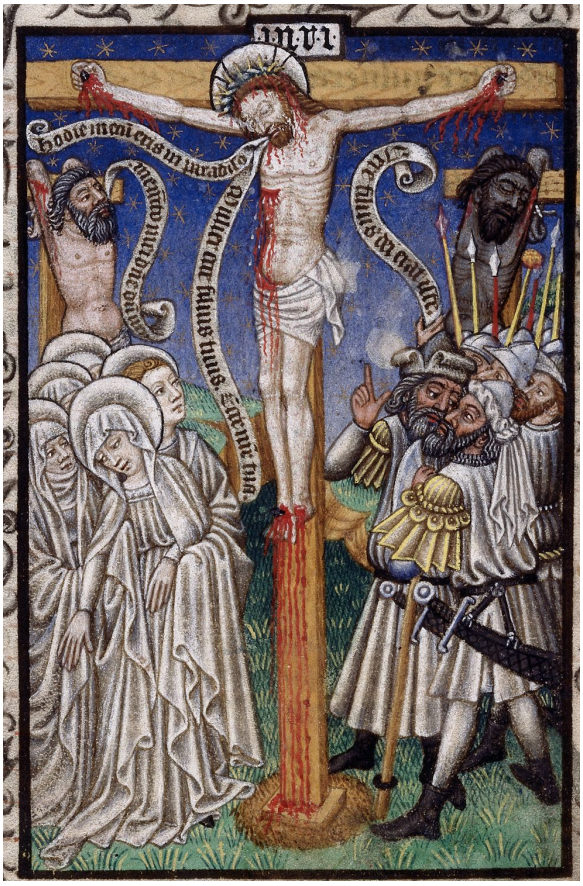

Fig. 1: Crucifixion from BL Harley

metadata and affect negatively the quality of the datasets [2]

The goal of Saint George on a Bike is to provide rich information about European cultural heritage pictorial artwork. More than one type of output may be generated, which fundamentally depends on the type of input available. The levels of semantic output that we currently contemplate are the following:

- Semantic resources in form of tags coming from existing vocabularies
- Textual captions

At this point in the project, we have designed and implemented several solutions that can generate textual tags or captions. We have identified the controlled vocabulary from which to choose semantic tags based on the Europeana Entity Collection tags, and we are in the process of:

- Refining this vocabulary
- Considering related sources such as DBpedia, Wikidata, or more specific vocabularies used by Europeana providers

In the rest of this document we explain the system and module-level architecture for each of these techniques, as well as their implementation.

II. Object detection

Object detection is a base step for several tasks, including caption generation and search. There are plenty of pretrained models (VGG-16, VGG-32, ResNet, etc.)[3][4] based on different datasets which can be used in object detection. However, object detection in cultural heritage has its own limitations. These models are usually trained with datasets whose object classes have no symbolic and iconographic dimension. However, when describing paintings, classes cannot be basic and broad-brush. For example, a bishop, Virgin Mary, or Saint George cannot be referred to as just a person when the painting contains object classes that identify them. Even a simple task such as recognizing animals and people can easily convert into a complex task if we'd like to know if the animal is a superbeing (dragon, minotaur, etc.), or what is the occupation of a person. That is why we decided to train our own model using transfer learning, which is able to detect classes with focus on cultural heritage. The detected objects are therefore labeled with our own class names.

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It is a popular approach in deep learning where pre-trained models are used as the starting point for computer vision or natural language processing tasks - given the vast compute and time resources required to develop neural network models for these problems and the huge gains it provides when applied to related problems. To improve precision of our object detection model we decided to use transfer learning.

Our implementation uses the Mask-RCNN (Kaiming et al. (2017))[5] model based on the pre-trained weights of the MS COCO dataset, as a starting point for the transfer model. The training set consists of more than 13000 manually labeled examples with annotations (source of image, file path, bounding box information, class names) in

[2]https://tinyurl.com/4rpn6vtf

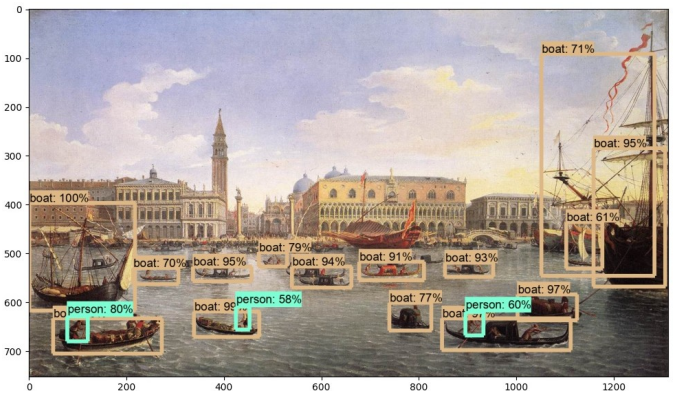

Fig. 2: Output of object detection model

VOC Pascal XML format (Figure 2). Full list of classes can be find in Appendix. We defined 69 classes based on a careful selection process that first eliminates anachronic classes from the COCO dataset, and then sets to detect the most common objects present in paintings, to further filter this set. A painting class corresponds to a category in a painting collection. The painting collection we have taken as a reference is the Wikimedia Commons collection of paintings labeled by the regular expression "Paintings of..." [3]

A. Extending the set of classes of interest

The next step extends the set of classes. Wikimedia Commons categories and subcategories are very useful to discern new painting classes when querying about basic classes. For example, if we query about paintings of people we find the subcategory angels_with_humans. In this case, humans is a general reference that covers the basic classes people, men, women and angel is a new painting class because Wikimedia has the category Paintings of people with angels. Starting from the filtered COCO dataset, new classes are added that are related via Wikimedia categories and subcategories. Among the possible classes derived from Wikimedia categories we have chosen a sample with iconographic and symbolic meanings, supernatural and metamorphosed animals (swan in Leda's paintings, cow in the rape of Europa) and devils. Apart from dragons, other fantastic animals are unicorn, centaur, minotaur. We also consider classes that help to identify people that have a social role (occupation) such as bishop, pope, knight or king.

[3]https://tinyurl.com/26dtr54j

III. Refining object class detection by using a language model

Figure 3 illustrates the caption generation technique that we have implemented, which is based on a language model (using BERT)[6]. The input to this model is the image representation of the painting containing a set of bounding boxes, one for each object class. The output is a set of statements that explain the visual relationships between the classes in the bounding boxes. In these texts, the classes corresponding to the bounding boxes are referred to with more specific denominations according to their visual relationships.

We use the detection network Mask-RCNN to identify bounding boxes in a painting and generate candidate labels for each of them. Section II explains the transfer learning process we apply to train the object detection model, starting from weights provided by a MS COCO pre-trained model. The Wikimedia Commons catalog covers iconographic classes (e.g. the Annunciation), symbolic classes (e.g. key of heaven for St. Peter), as well as imaginary beings, occupations (e.g. monks, knights), etc.

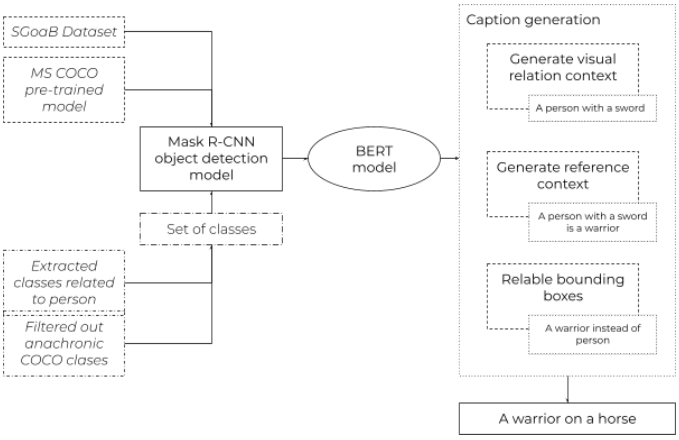

Fig. 3: Using a language model to improve object detection for caption generation

The goal of caption generation is to refine the references to the main object(s) among all salient objects of a painting. We consider that the main object is the one whose bounding box intersects the largest number of other bounding boxes in

a specific object cluster. For each object whose bounding box overlaps with the main object's, the algorithm first generates a set of sentences that describe the possible visual relations between the two. These texts are generated by using a language model to guess missing words that mask the possible relations between the two objects, and they are called visual relation contexts. We then use the language model again to generate the most appropriate completions that specialize the original object class in the visual relation context. (See figure 4) For instance, a person carrying

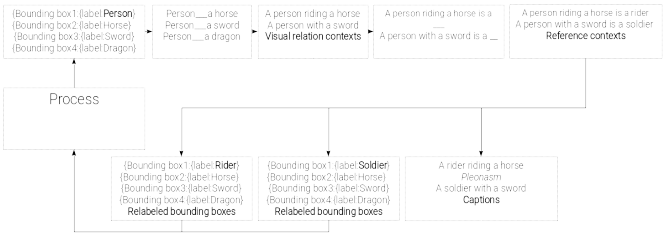

Fig. 4: Refining object class detection by using a language model

a cross is Jesus, while a person with a crown becomes a queen or a king. Only those pairs that are predicted with high accuracy by the language model, will generate a visual relation context. Each of the visual relation contexts that pass this filter are then placed in a reference context to refine the main object and generate the captions.

This tool outputs a set of textual captions and can generate basic level classes, higher-level concepts, and named entities

IV. Visual relationships between detected objects in an image

Another approach of detection of visual relationships between objects is based on analysis of bounding boxes positions. Multiple objects can be successfully detected and labeled in an image (e.g. by R-CNN). However, part of the challenge inherent in building systems for automatic image captioning is that learning the visual relationships between detected objects in an image is not trivial. In this section, we describe how a custom implementation of a bounding-box-based (bbx) analysis yields useful visual relationships between objects previously detected by R-CNN technology. We dubbed this Python based implementation "VIS-REL". Our code is applied to imagery representing sacred art produced between the 14th and the 18th centuries (both included). The context being that of sacred iconography, producing captions to enrich image annotations is a task that broadly corresponds to Panofsky's second level of interpretation of cultural heritage imagery. An example of that would be for the image beholder or the image processing system to rightfully conclude that 13 men having supper with bread and wine (primary level of interpretation) represent the figure of Jesus Christ flanked by his 12 apostles in "The Last Supper" before his crucifixion in Jerusalem (secondary level of interpretation) as described in the New Testament of the Christian Bible. The general idea of the approach is based on detection of relationships between pairs of objects. In order to assess whether any two detected objects, belonging to any two arbitrary object classes (e.g. a person and a horse), are in the same image view-plane, that is to say, at the same field depth in an image, one needs a base-reference of pairwise proportions between objects of every trained class (Figure 5) . In practice, VIS_REL computes pairwise-proportions based on common-sense measures and proportions translated as relative surface area proportions between bbxes. Those pairwise proportions between detectable objects are meant to reflect a common-sense representation of realistic pictorial proportions in paintings. Comparing of proportions and some additional measurements allows defining rules which can assume general relationships between pairwise objects:

- Stands
- Holds
- Sits
- On
- Behind
- etc.

V. Challenges

Despite the progress of the project, the technology remains significantly more primitive than human vision and cannot yet satisfactorily address all challenges of GLAM-institutions. We see a number of long-standing challenges:

- Data collection
  - Some classes are represented only in a few images

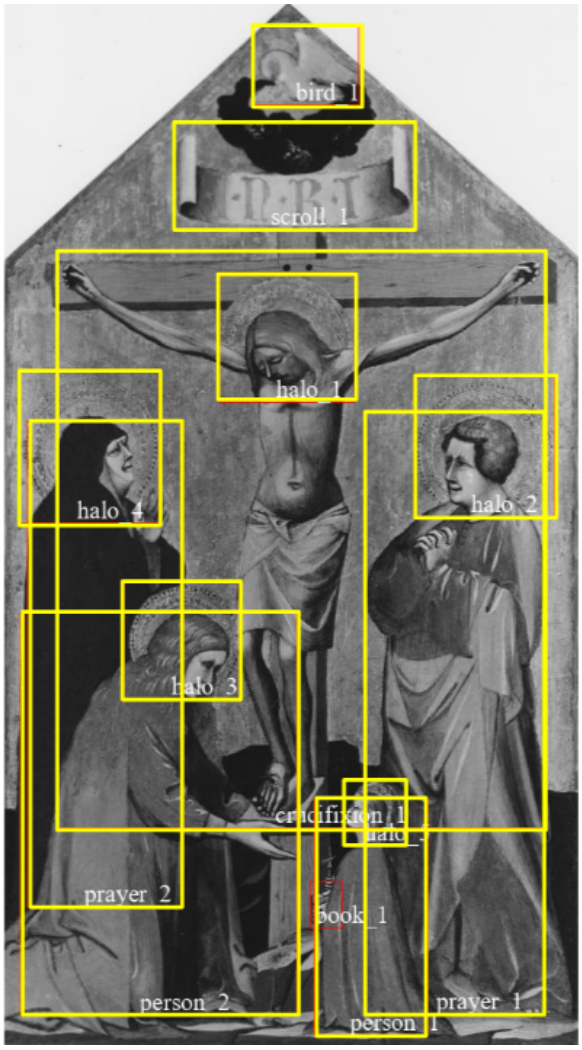

Fig. 5: Detection of relationships between pairs of objects

  - Style, medium, color may differ significantly between artists
  - Not so many paintings anyway and can't produce them when needed
- Poor metadata
  - Labeled bounding boxes
  - Descriptions of visual content
  - Labeled visual relationships
- Small dataset of paintings by data mining standards requires complementary techniques to
  - Filter out anachronisms
  - Detect imaginary objects or (unusual) actions
- Evaluation
  - Quantifying enrichments quality and usefulness to the user

VI. Future work

Future work is structured around several directions:

1) Improve the current methods for tag and caption generation
   - Increase training dataset size for object detection to about 15k pictures. This is the set that includes bounding box information and not the image/caption pairs dataset.
   - Use our own trained model (described in Section "Object detection") as an encoder for caption generation using the attention mechanism.
   - Collect training dataset size for caption generation with relevant canonical captions that can be effectively analyzed via Natural Language Processing techniques.
   - Look into the evolving meanings of a word, or homonymic meanings of words, to be able to deal with different meanings over (potentially) distinct time intervals.
   - Test other language models besides standard BERT (e.g. EuBERT). Other approaches to caption classification may be possible, such as fitting a language model over the image/caption dataset.
2) Source more data for training and/or evaluation, notably by crowdsourcing.
3) Update processes in Section III, so that resulting textual tags are 'uplifted' to semantic tags.
4) Build a knowledge graph for the domain of expertise. Complement the work on inferring visual relationships based on BBx' analysis with an approach that could start from knowledge graphs and domain axioms and refine or infer richer object labels and relationship names. These will translate in the generation of semantic graphs for the images. This task involves a thorough evaluation step, the result of which will determine the ability to generate good metadata about basic and higher level actions.
5) Extend the scope of the methods to more general topics beyond figurative and mostly

iconographic paintings

Acknowledgment

This research has been supported by the Saint George on a Bike project 2018-EU-IA-0104, co-financed by the Connecting Europe Facility of the European Union.

References

[1] Shurong Sheng, Marie-Francine Moens.2019. "Generating Captions for Images of Ancient Artworks". The 27th ACM International Conference

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár.2014."Microsoft COCO: Common Objects in Context". arXiv: 1405.0312.

[3] Karen Simonyan, Andrew Zisserman.2014. "Very Deep Convolutional Networks for Large-Scale Image Recognition". arXiv:1409.1556 .

[4] Karen Simonyan, Andrew Zisserman.2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition". arXiv:1409.1556

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. 2015. "Deep Residual Learning for Image Recognition". CVPR

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. 2017. "Mask R-CNN". ICCV

[7] YJacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ". Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Appendix A
List of classes

Crucifixion, Angel, Person, Crown of thorns, Horse, Dragon, Bird, Dog, Boat, Cat, Book, Sheep, Shepherd, Elephant, Zebra, Crown, Tiara, Camauro, Zucchetto, Mitre, Saturno, Skull, Orange, Apple, Banana, Nude, Monk, Lance, Key Of Heaven, Banner, Chalice, Palm, Sword, Rooster, Knight, Scroll, Lily, Horn, Prayer, Tree, Arrow, Crozier, Deer, Devil, Dove, Eagle, Hands, Head, Lion, Serpent, Stole, Trumpet, Judith, Halo, Helmet, Shield, Jug, Holy Shroud, God The Father, Swan, Butterfly, Bear, Centaur, Pegasus, Donkey, Mouse, Monkey, Cow, Unicorn