Visual content classifier for cultural heritage repositories

Anonymous ACL submission

Abstract

This work presents a novel approach for the automatic creation of an aligned image / text training set for the generation of descriptions of the visual content of artworks. To do this, we develop a classification tool based on a mix of heuristic rules and deep learning. This classifier is able to identify statements that describe visual art content, out of complex cultural heritage text that contains a mix of many other types of information on context, medium, author, etc. Our results are very promising when tested on texts from the Museo del Prado collections.

1 Introduction

The work we present in this paper is motivated by the problem of automatically generating visual descriptions of paintings. The current focus is 2D artwork between the 12th and the 18th century, before art currents proposed painting styles that are highly non-representational.

Datasets such as MS COCO, Open Images V4 or Flickr30k Young et al. (2014) associate manual descriptions with the images, but these depict every-day life activities and objects of the (very) recent past. This poses a problem if we were to use a model based on these datasets, given that some objects of the past are not in use any more, current (and often photographed) objects can have very similar shape to old objects, artworks may depict imaginary or symbolic objects, and they can often represent actions that are not captured in photographs (i.e. kill, decapitate, rape, etc). That means that we need a body of aligned artwork image / descriptions to be able to successfully train a model for generating descriptions using deep learning technology.

Unfortunately, descriptions of the visual content of artworks are the exception rather than the norm in cultural heritage repositories; the unspoken assumption is that one can see the artwork and thus there´s no need to describe its content. 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 scene description may be available, although it is not usually exhaustive. As a result, it is difficult to collect enough texts aligned to artworks ready to be used for training a deep learning system. To add to the problem, the relevant phrases that can be found 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.

We tackle the lack of a significant body of visual descriptions of artworks by implementing a classifier that identifies, out of complex art repository texts, those statements which refer to the visual image content. These statements will form the basis of an aligned image / description training set for description generation via deep learning.

2 Basic approach

Our goal is to create a tool that successfully discriminates between descriptive (DESC) and non-descriptive (NODESC) English statements that refer to image content. This tool will filter out sentences present in artwork descriptions that are irrelevant to the content depicted in the image. The ultimate goal is to save manual annotation work and instead extract automatically the relevant parts of the descriptions available on some museum websites and art collections (e.g. Europeana, Web Art Gallery or Wikimedia datasets). To do this, we perform the following steps:

- Pre-process complex descriptions and split them into simple statements, amenable to NLP
- Classify simple sentences via common-sense rules that likely describe visual content of an artwork rather than other types of information

- For those statements for which the common-sense rules do not apply, classify them using our deep learning models

3 Methodology

This section presents how we implement each of the three steps as part of the pipeline that forms our classification tool.

3.1 Sentence simplification

Sentences that are complex syntactically and stylistically are likely to mislead the classifier. Therefore, sentences are simplified to a basic structure of subject-verb-object (e.g. "They receive the guests discourteously, angrily, and scornfully" is transformed into "They receive the guests"). The simplification is performed in two stages: (1) parse the sentence using the Spacy dependency parser[1] in order to detect the subject, verb and object, then (2) create the simple statement by concatenating these constituents in a string.

3.2 Classify sentences using common-sense rules

This step takes the simple statements generated in the first step and starts by replacing art jargon instances of person with the concept *person*. The output sentences are passed as input to a set of rules that recognizes sentences which usually either describe or not, an artwork. The rest of this subsection explains these two phases in detail.

3.2.1 Rules of replacement

Descriptions of artworks in cultural heritage repositories contain jargon characteristic of this field. Some of these expressions have regular language equivalents, which are present in MS COCO captions. The most common visual concept happens to also be the one with the widest variety of possible instantiations, and it is *person*. After exhaustive testing, we saw that replacing some expressions in the repositories that stand for human-like concepts (e.g. *figure* or *sitter*) with the very frequent MS COCO word *person*, makes the sentence be correctly classified as DESC. Therefore, the tool first applies rules of the form *Replace X with Y*, where Y is a word in MS COCO captions. So far the words replaced by *person* are *figure*, *sitter*, in singular and plural forms, and personal pronouns, including *who*. We do not replace person-like named entities because the artwork-specific model we train takes them into consideration.

[1] https://spacy.io/api/dependencyparser

3.2.2 Rules for recognizing sentences not describing a painting

While it is true that recognising sentences describing a painting cannot be captured in simple rules, we can generally assume that these are written in the present tense. The tool therefore classifies sentences whose root verb is not in the present tense as NODESC, given that these tenses are mainly used for narratives (i.e. include the notion of a time sequence) or represent hypotheses. Examples of these narratives are about the life of the artist or the events that happened before or after the scene depicted in order to put this scene in context.

3.2.3 Rules for recognizing sentences describing a painting

Certain expressions that are characteristic of descriptions in artworks are very useful to identify DESC sentences. Expressions like *in the background*, *in the foreground*, *(the painting) depicts* appear in sentences that describe the content of the painting. Therefore, the tool classifies a sentence as DESC when the tool detects *background*, *foreground*, *depict(s)*, *portraits* as a verb, *in centre*, *(on/to) right*, *(on/to) left*.

3.3 Classify sentences using deep learning models

Due to the large body of descriptions of pictures (e.g. MS COCO) and the reduced corpus of data that allows learning what is in a painting (e.g. IconClass), we structure the task of learning which statement is likely to describe the visual content of an artwork in two sub-tasks: (1) recognize a generic image description and (2) recognize that the sentence describes the content of an artwork rather than any other image type.

3.3.1 Recognizing a sentence describing a generic image

We first train a model over a corpus including sentences from the MS COCO caption dataset as positive examples (i.e. DESC) and the English Wikipedia as negative examples (i.e. NODESC). MS COCO sentences are considered DESC because we take the MS COCO captions as canonical descriptive texts for the visual content of images. The amount of DESC sentences is around 320000. The counterpart Wikipedia sentences are also around 320000 and were randomly selected from the English Wikipedia. The resulting model, **CocoVSwiki**, fine-tunes a BERT model, concretely, distilbert-base-uncased Sanh et al. (2019).

3.3.2 Recognizing a sentence describing (or not) an artwork

The MS COCO caption dataset describes photographs, which implies that the objects and relationships are not entirely representative of the objects and relationships in artworks between the 12th and the 18th century. Additionally, photographs cannot depict fantastic creatures such as angels, dragons, unicorns, etc. Moreover, the people depicted in public photograph datasets are anonymous whereas in artwork it is important to identify the individuals (e.g. Jesus, the Virgin Mary, Abraham, Venus, etc.).

A domain-specific model for (2D) visual arts must know how to recognize a sentence describing what is going on in an artwork. For this purpose we trained **IconVSwiki**, a model that also fine-tunes distilbert-base-uncased. IconVSwiki's training set contains about 65000 DESC sentences from Iconclass notations and about 65000 NODESC randomly selected sentences from the English Wikipedia. Iconclass[2] notations provide a systematic overview of subjects, actions, entities and motifs represented in Western art. These notations are useful for art institutions to describe the works of art in their collections, and identify the significance of the scenes and elements depicted.

The difference in the number of sentences in the training set for IconVSwiki vs CocoVSwiki is due to the fact that the number of Iconclass notations is not as large as the number of captions in the COCO dataset.

3.3.3 Classification using the Deep Learning models

This classification is only triggered when the common-sense rules did not succeed. The first text classifier (CWTC) is trained with CocoVSwiki and the second classifier (IWTC) is trained with IconVSwiki. If a sentence is classified as DESC by the CWTC classifier, we simply label the sentence DESC. Otherwise, use the IWTC classifier to label the sentence as DESC or NODESC. If IconVSwiki labels the sentence DESC, this is likely to refer to iconographical content not present in the CocoVSwiki model.

[2] http://www.iconclass.org

4 Evaluation and discussion

We evaluated our visual description classification tool over a training set containing painting titles (in this case statements) and the first three sentences of the texts accompanying a subset of the paintings from the English version of the Museo del Prado collection[3]. The choice of the first three sentences is empirical and based on examining the Prado collection, in which the description of the content of the paintings is usually found at the beginning of the text. The evaluation corpus consists of 1000 sentences which we manually labeled as DESC or NODESC. The automatic labeling was performed by our classification tool, as already explained. This allowed us to calculate the F1 score of the classifications performed by the tool. For CocoVSwiki, this score is 0.22, while for IconVSwiki it is 0.801. This result marks a significant improvement on part of the classification model trained with Iconclass notations rather than everyday image descriptions.

On close inspection of the true positives and negatives returned by the classifier, we comment on several important aspects:

- Our classifier is largely successful in identifying descriptive sentences containing iconographic named entities that are present in paintings.
- IconVSwiki successfully identifies as descriptive most of those sentences that describe situations very frequent in paintings but not present in public photograph datasets, such as killings, rapes, beheadings, etc.
- The model identifies that words that contribute positively to the DESC classification label refer to entities mostly present before the 18th century. It´s interesting that some of these are also semantically close to 20th century entities from the MS COCO dataset (e.g: throne - chair, crown - hat). Using Iconclass notations makes it possible to work directly with statements including these "anachronic" entities instead of replacing them with simpler, more generic, and present-day entities by applying replacement rules.
- The classifier is able to mostly filter out sentences that refer to biographical aspects, schol-

[3] https://www.museodelprado.es/en/the-collection

ars' opinions and information that puts the painting in context.

The current implementation of the classifier has nevertheless limitations. For instance, we noticed that the evaluation corpus contains sentences where the descriptive content is embedded in a sentence expressing an opinion (e.g: *He perfectly integrates the hunter´s figure among the sinuous silhouettes of the trees* or is inferred from the explanation of the symbolic meaning or historical aspects of objects *His purple robe signifies sacrifice and martyrdom, while his rhomboidal halo echoes the Byzantine tradition*. In this case, the robe and the halo have not been previously described as part of the visual content of the artwork.

It is difficult for the classifier to infer the descriptive content in such sentences, which in fact follow the guiding principles for writing style in cultural heritage descriptions. These stylistic recommendations favor the embedding of syntactic constituents that refer to the things depicted in the painting. The consequence is that the references to these things do not depend on the root verb. In *'John the Baptist, recognisable by his clothing and by the lamb on the book, has been painted with great care.'*, the clothing, the lamb and the book do not depend on the root verb *painted*. In fact, they do not depend on any verb. This is a challenge of the sentence simplification step, which is currently based on a verb-centric syntactic parser. In the future we will address how to find references to objects depicted in a painting in verbless syntactic structures.

Another limitation we found, in this case of the common-sense rules, is that some sentences that are narrative are in fact written in the present tense. The classifier thus labels as DESC sentences that refer to events previous or following to the scenes depicted. Lastly, one of the consequences of sentence simplification is, in some cases, the creation of input with not enough information for the classifier to label the sentence correctly. This is due mostly to errors in the automatic dependency parsing. Other parsers will be tested in the future.

Future work will also test our approach on a larger and more diverse dataset. We are aware of no risks or biases that come from using these cultural heritage repositories.

5 Related work

The ultimate goal of our description classifier is to obtain a training set of aligned image-text for the automatic generation of artwork descriptions based on deep learning. Of the three perspectives Bai et al. (2021) identifies as part of a museum-like artwork description, namely content, context, and form, we focus on content.

Dognin et al. (2019) addresses three main challenges in bridging the semantic gap between visual scenes and language in order to produce diverse, creative and human-like captions. For the Cultural Heritage domain, this problem is even more significant. As far as we are aware, not many authors have tackled successfully the (visual) content description generation problem for the cultural heritage domain Sheng and Moens (2019). As it can be expected, existing methods Vaswani et al. (2017) that work well on photographs don´t generally return correct - or precise enough - descriptions for cultural heritage imagery. To generate better descriptions for artworks, some previous works use ontologies Xu et al. (2017) or hierarchical models Xu and Wang (2015), and leverage existing metadata for cultural images.

Other approaches for learning relationships between objects exist that are not language guided and thus are not based on the existence of a training set but do require scene descriptions. Raposo et al. (2017) introduces relation networks (RNs), a general purpose neural network architecture for object-relation reasoning that learn from scene description data. Johnson et al. (2018) generate images from scene graphs and use adversarial training. We are not aware of any work that has tested these approaches for cultural heritage.

Our work is different in that it takes the approach of language-guided models without requiring manual annotations, but rather relying on a combination of heuristic rules and deep learning to extract from complex text only those statements that refer to the visual content of artworks.

6 Conclusions

This paper introduces a novel approach for the automatic generation of training sets for visual description generation in the cultural heritage domain. We rely heavily on Iconclass notations, which are able to fine-tune our classifier to recognize iconographic entities, objects not in frequent use in the present, and events that are not generally depicted in pictures. Our results mark a significant improvement over what models trained on every-day life images could achieve.

References

Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *ICCV*.

Pierre L. Dognin, Igor Melnyk, Youssef Mroueh, Jarret Ross, and Tom Sercu. 2019. Adversarial semantic alignment for improved image captions. In *CVPR*.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *IEEE/CVF*.

D. Raposo, A. Santoro, D.G.T. Barrett, R. Pascanu, T. Lillicrap, and P. Battaglia. 2017. Discovering objects and their relations from entangled scene representations. In *ICLR*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Shurong Sheng and Marie-Francine Moens. 2019. Generating captions for images of ancient artworks. In *Proceedings of the 27th ACM International Conference on Multimedia*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Lei Xu, Albert Merono-Penuela, Zhisheng Huang, and Frank Van Harmelen. 2017. An ontology model for narrative image annotation in the field of cultural heritage. In *Proceedings of Workshop on Humanities in the Semantic web (WHiSe)*, pages 15–26.

Lei Xu and Xiaoguang Wang. 2015. Semantic description of cultural digital images: using a hierarchical model and controlled vocabulary. *D-Lib magazine*.

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