Data Collection for Contextual and Visual Question Answering in the Cultural Heritage Domain

Francesco Vannoni Pietro Bongini Federico Becattini Andrew David Bagdanov Alberto Del Bimbo
Media Integration and Communication Center
University of Florence
Email: name.surname@unifi.it

***Abstract*—In this demonstration we propose an annotation tool to collect question-answer samples for artworks, necessary to train and evaluate visual and contextual question answering models. The tool is completely web-based, and relies on an automatic question-answer generation model to aid the annotation process. Through the annotator, users can inspect and refine the generated annotations and obtain statistics on their quality. A pre-trained visual and contextual question answering model is also provided to the final user to be able to interact with the system by asking questions about artworks.**

I. Introduction

Technology and the fruition of cultural heritage are becoming increasingly more entwined, especially with the advent of smart audio guides, virtual and augmented reality, and interactive installations. Machine learning and computer vision are important components of this ongoing integration, enabling new interaction modalities between user and museum. Nonetheless, the most frequent way of interacting with paintings and statues still remains taking pictures. Yet images alone can only convey the aesthetics of the artwork, lacking the information which is often required to fully understand and appreciate it. Usually this additional knowledge comes both from the artwork itself (and therefore the image depicting it) and from an external source of knowledge, such as an information sheet. While the former can be inferred by computer vision algorithms, the latter needs more structured data to pair visual content with relevant information.

Regardless of its source, this information still must be effectively transmitted to the user. A popular emerging trend in computer vision is Visual Question Answering (VQA), in which users can interact with a neural network by posing questions in natural language and receiving answers about the visual content.

The usage of VQA for Cultural Heritage has been explored in [2], where questions have been categorized into two categories: *visual* if they refer to the content of the artwork and *contextual* is they refer to knowledge deductible only from an external source.

Interacting through questions and dialogs will likely be the evolution of smart audio guides for museum visits and simple image browsing on personal smartphones. In this way, the classic audio guide turns into a smart personal instructor with which the visitor can interact by asking for explanations focused on specific interests. The advantages are twofold: on the one hand the cognitive burden of the visitor will decrease, limiting the flow of information to what the user actually wants to hear; and on the other hand it proposes the most natural way of interacting with a guide, favoring engagement.

However, realizing such an interactive system is not straightforward. The biggest obstacle towards this goal is the lack of specialized data for the cultural heritage domain, which will require an expensive and temporally demanding annotation campaign. In particular, there is the need for question-answer pairs related to both the visual and contextual information of artworks. To address this limitation, we propose a semi supervised approach that relies on automatic question generators to adapt textual descriptions of artworks to data that can be used to train visual/contextual question answering models.

The system we demonstrate is a web based annotation tool to browse, edit and validate datasets of automatically generated questions relative to images of artworks. The tool offers the advantage of lowering the annotation burden of building a dataset manually, while allowing the user to perform an analysis of question generation systems by showing error statistics.

II. Data collection

The purpose of the proposed system is to aid users in the annotation of artwork images with visual and contextual questions/answers.

Each artwork is paired with a picture and a textual description, which can be easily gathered from online sources such as Wikipedia or DBpedia. Users can assign a label to sentences to mark them as visual or contextual. These sentences are then fed to a text-based question generation model which converts them into questions and answers. The visual and contextual labels are automatically transfered from sentences to questions.

To obtain question-answer pairs, we first gather a collection of visual and contextual sentences relative to artworks. We use data from Artpedia [3], a dataset containing 2.930 paintings and a total of 28.212, manually labeled as visual or contextual (9.173 visual sentences and 19.039 contextual sentences). On average, an artwork is labeled with 3.1 visual sentences and 6.5 contextual sentences. The user of our system can browse all images and their textual labels and can enter new descriptions or modify existing ones. An example is shown in Fig. 1.

We then automatically generate question-answer pairs with [1], a recently proposed end-to-end trainable sequence-

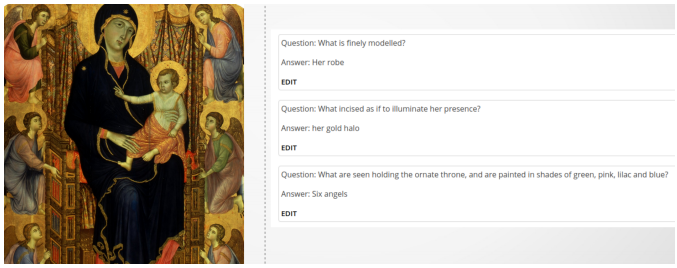

Fig. 1. A sample of automatically generated questions, previewed in our web interface.

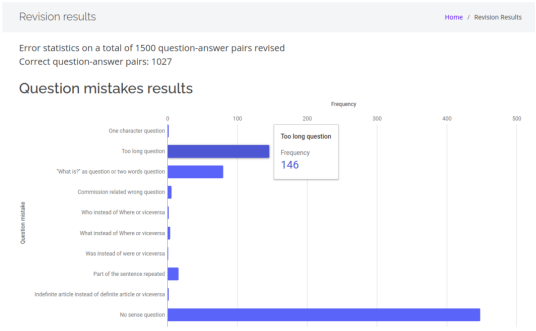

Fig. 2. Statistics of the mistakes made by the automatic question generator [1].

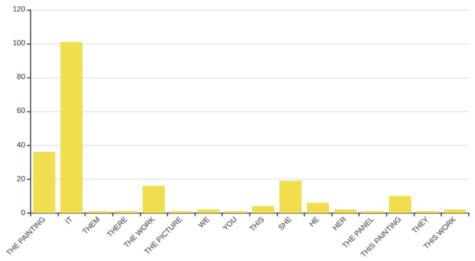

Fig. 3. Statistics of the non usable answers provided by the model [1].

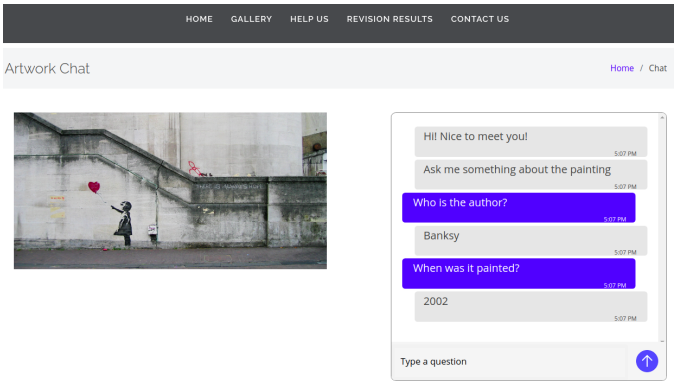

Fig. 4. Users can interact with the system by asking questions about specific artworks. Questions can be visual or contextual.

to-sequence model. We have obtained more than 100.000 generated questions and answers. Each generated item can be inspected through the web interface and can be edited by the user. Once a sample has been inspected, it is marked as "revised" and is then considered to be part of the dataset under construction. If a new question is instead added directly by the user, it is automatically flagged as revised.

To ease the data collection process, we developed a fast annotation web interface where multiple users can revise questions in parallel, inspecting a sequence of random question-answer pairs. In this way we have collected a dataset of 1027 manually revised question-answers out of a subset of 1500 automatically generated samples.

During the revision process, if a mistake is identified in the automatically generated sample, the user can label it with a customizable error category. This provides us with statistics on the quality of the question generator, which offers interesting insights about the model. We have identified 10 error categories for questions, among which the most common are *too long questions*, *two words questions* (such as *what is?*) and *nonsense questions*. Fig. 2 shows in detail a breakdown of the errors encountered in the revision process. Fig. 3 instead, show the occurrences of most frequent incorrect answers.

III. Visual and Contextual Question Answering

We have integrated in the system the model presented in [2], which answers both visual and contextual questions, relying on a question classifier to understand whether a Question Answering or Visual Question Answering submodule is better suited to answer. The user can therefore test the model by interacting with it, asking questions about an artwork and its contextual information through a chat (Fig. 4). In this way it is possible to collect data with the proposed annotation tool, train a custom model and deploy it through the web interface to the final user. In addition, the collected data can be used to test pre-trained models for visual and contextual question answering with a joint evaluation. In fact, in literature no dataset for visual question answering in the cultural heritage domain has been collected yet.

IV. Conclusion

In this demonstration we have proposed a web based annotation tool to collect in a semi automatic way questions and answers relative to artworks. The tool relies on a text-based question/answer generator. The generated samples can then be manually inspected and revised. The system also offers to inspect the quality of the generated questions by gathering error statistics and provides an interface for the user to interact with a pre-trained question answering model that answers both visual and contextual questions.

References

[1] Du, Xinya, et al. "Learning to Ask: Neural Question Generation for Reading Comprehension." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.

[2] Bongini, Pietro, et al. "Visual Question Answering for Cultural Heritage." arXiv preprint arXiv:2003.09853 (2020).

[3] Stefanini, Matteo, et al. "Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain." International Conference on Image Analysis and Processing. Springer, Cham, 2019.