VISCOUNTH: A Large-scale Multilingual Visual Question Answering Dataset for Cultural Heritage

FEDERICO BECATTINI and PIETRO BONGINI, University of Florence, Italy
LUANA BULLA, Institute of Science and Technology of Cognition, National Research Council, Italy
ALBERTO DEL BIMBO, University of Florence, Italy
LUDOVICA MARINUCCI and MISAEL MONGIOVÌ, Institute of Science and Technology of Cognition, National Research Council, Italy
VALENTINA PRESUTTI, University of Bologna, Italy

Visual question answering has recently been settled as a fundamental multi-modal reasoning task of artificial intelligence that allows users to get information about visual content by asking questions in natural language. In the cultural heritage domain, this task can contribute to assisting visitors in museums and cultural sites, thus increasing engagement. However, the development of visual question answering models for cultural heritage is prevented by the lack of suitable large-scale datasets. To meet this demand, we built a large-scale heterogeneous and multilingual (Italian and English) dataset for cultural heritage that comprises approximately 500K Italian cultural assets and 6.5M question-answer pairs. We propose a novel formulation of the task that requires reasoning over both the visual content and an associated natural language description, and present baselines for this task. Results show that the current state of the art is reasonably effective but still far from satisfactory; therefore, further research in this area is recommended. Nonetheless, we also present a holistic baseline to address visual and contextual questions and foster future research on the topic.

CCS Concepts: • **Computing methodologies** → *Computer vision*; *Natural language processing*; • **Applied computing** → *Arts and humanities;*

Additional Key Words and Phrases: Visual question answering, cultural heritage

---

Authors' addresses: F. Becattini, P. Bongini, and A. D. Bimbo, Media Integration and Communication Center, University of Florence, Viale Giovanni Battista Morgagni, 65, Florence, 50134, Italy; emails: {federico.becattini, p.bongini, alberto.delbimbo}@unifi.it; L. Bulla, and M. Mongiovì (corresponding author), Institute of Science and Technology of Cognition, National Research Council, Via Paolo Gaifami, 18, Catania, 95126, Italy; emails:{luana.bulla, misael.mongiovi}@istc.cnr.it; L. Marinucci, Institute of Science and Technology of Cognition, National Research Council, Via S. Martino della Battaglia, 44, Rome, 00185, Italy; email: ludovica.marinucci@istc.cnr.it; V. Presutti, Department of Modern Languages, Literatures and Cultures, University of Bologna, Via Cartoleria 5, Bologna, 40124, Italy; email: valentina.presutti@unibo.it.

193

1 INTRODUCTION

The fruition of museum experiences, as well as the management of cultural assets, has been profoundly affected by recent technological advancements involving multimedia analysis and processing. Numerous applications have been developed to assist visitors in understanding and deepening their comprehension of the artworks exposed in a museum [6, 19, 28, 46, 54]. Interactivity is important in such applications, both to increase engagement [6, 11, 24] and to personalize the visit according to the interests of the user [32, 55]. Recently, machine learning models to enable interaction as a form of dialogue have been proposed [1, 29]. In particular, the task of **Visual Question Answering (VQA)** [36] allows users to ask questions in natural language to a machine learning model regarding the content of a visual media. Independently of the cultural heritage domain, this task has gained significant attention in the last years as a representative multi-modal reasoning task, where both visual content and natural language text need to be processed to get a result. Recent approaches have shifted from a basic formulation were the answer is directly contained in the image (e.g., how many people are there) [26, 31, 44] to the use of external or common sense knowledge for answering more complex questions (e.g., which game is she playing at) [37, 57]. Nonetheless, a domain shift exists between standard machine learning datasets used to train such models and the cultural heritage domain.

A few attempts have been made to address these tasks, specifically for art and museum visits [7, 9, 10, 27, 48]. Most of these works first collected a dataset of questions and answers relative to artwork images and then retrained a new model for VQA. However, there appears to be a general consensus regarding the fact that visual media alone are not sufficient to solve VQA in the cultural heritage domain. Most relevant information for users in fact appears to be found in contextual descriptions rather than in the visual content of the artwork itself. Whereas the artwork conveys its aesthetics, contextual information such as the name of the author, the artistic current or its allegoric meaning, requires an additional source of knowledge to be communicated to the visitor. General VQA models able to handle external knowledge are not adequate for the cultural heritage domain for multiple reasons. First, features such as painting style, architectural style, and degree of conservation are specific of the cultural heritage field and hence they cannot be learnt from out-of-domain datasets. Second, reasoning with large knowledge bases makes the task harder, therefore current state-of-the-art performances are still far from satisfactory.

Fortunately, the cultural heritage domain presents specific characteristics that might help increase performance. Traditionally, external knowledge is provided by a human expert or an informative sheet. Therefore, the additional knowledge necessary for generating the answer can be given as input, together with the image, thus avoiding the need for reasoning with or retrieving from a large knowledge base. For instance, a virtual guide in a museum might have access to both the picture of an object (e.g., a painting) and a textual description associated to it. Analogously, a virtual guide app might recognize an object from a taken picture (e.g., a church) and retrieve his corresponding description from Wikipedia or other textual sources. VQA methods for cultural heritage need to build holistic models capable of deriving answers both from an image depicting the artwork and a textual description describing the content that cannot be directly inferred by looking. Such models need somehow to combine two independently studied tasks, i.e., the classic
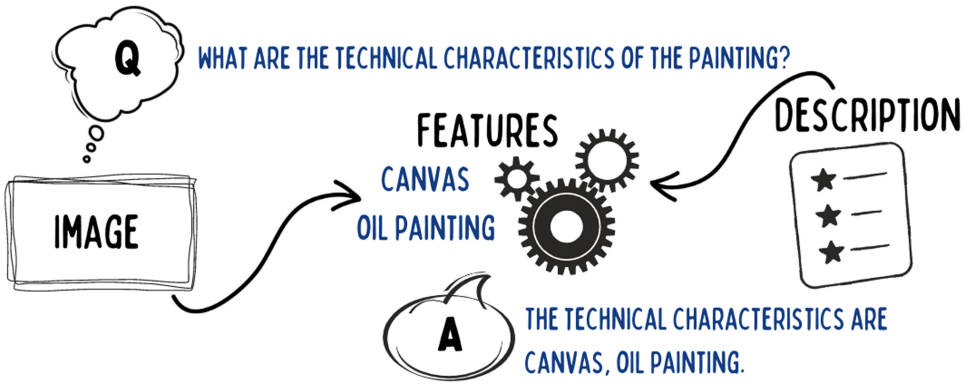
Fig. 1. Overview of our approach. The system takes as input the question, an image of the cultural asset, and a related natural language description for generating the answer.

question answering from natural language [42] and VQA, for which available approaches have interesting performances [23, 65]. Figure 1 shows the basis of our approach. Given a question (e.g., "What are the technical characteristics of the painting?"), the system considers features from both the image and a related natural language description for generating the answer (e.g., "The technical characteristics are canvas, oil painting").

However, there are no available large datasets with the characteristics discussed above, necessary to train machine learning models that jointly consider the image and the associated natural language description. In this work, we aim to fill this gap by generating a large multi-language VQA dataset for the cultural heritage domain. Difficulties are twofold: on the one hand, not only images of artworks must be collected but also accurate descriptions that require a domain expert; on the other hand, relevant questions with correct answers derivable from either the image or the description must be collected for each piece of art. In our work, we generate a large-scale dataset for cultural heritage in Italian and English by means of a semi-automatic approach that exploits data from an existing ontology-based knowledge graph. We first obtain a set of question templates asking expert and non-expert users to provide relevant questions for observed artworks. The question templates are then used to automatically extract answers from the knowledge graph, thus associating question-answer pairs with entities belonging to the cultural domain. We produce both short synthetic answers, useful for validating correctness of the prediction, and long colloquial answers, useful for user interaction through dialogue. A preliminary version of the dataset has been presented in Reference [2]. We significantly extend the dataset by considering a broader variety of question verbal forms (from 282 to 427), in particular, by considering verbal forms that are specific for certain cultural assets (e.g., "who is the author of this painting," specific for paintings) and including additional details (e.g., the span of the answer for contextual question). Furthermore, we present baselines for our proposed VQA task and discuss current state-of-the-art performances, criticality and research directions.

Overall the main contributions of our work are the following:

- We present the first complete large-scale multi-language visual question answering dataset for cultural heritage comprising approximately 500K images and 6.5M question-answer pairs in Italian and English. We detail our data collection process based on ArCO, the Italian cultural heritage knowledge graph.
- We rise the issue of domain shift in Visual Question Answering datasets for cultural heritage, which does not allow the exploitation of off-the-shelf VQA models without a re-training

phase. We also take into account visual and contextual question answering, exploring the limitations of existing image-based and text-based question answering models for artworks.

- We propose baselines for the proposed dataset, analyzing the results according to different criteria such as question type and artwork type. We believe that this will foster the advancement and development of interactive smart assistants in museum visits enabling visual and contextual question answering capabilities.

2 RELATED WORK

Since its introduction, VQA [36] has received a lot of attention from the Computer Vision and Machine Learning community. Several VQA datasets (DAQUAR [36], KB-VQA [57], COCO-QA [44], FM-IQA [26], VQA-real [1]) and methods [5, 26, 31, 44, 59, 61, 62, 65, 66, 68] have been provided since then. Other effort has been spent on grounding visual concepts to language [18, 25, 63], perhaps the most popular example being Visual Genome [33], and in general for associating images with information in natural language (Visual madlibs [64]). The interest in learning to match the visual domain with text stems from the need to address different multimodal tasks such as image captioning [63]. Large pretraining to align the two modalities are often required before addressing downstream tasks. Chen et al. [18] proposed Uniter, a joint image-text embedding learned by combining massive amounts of data from four different datasets and encouraging fine-grained alignment. The idea has then beed extended in Reference [25] leveraging adversarial learning.

In its original definition, the VQA task requires to answer questions that can be retrieved directly from the image (e.g., "How many cars are there?"). A more challenging yet valuable scenario considers questions that require external (or common sense) knowledge to be answered (e.g., "What are these people doing?" refers to a picture with snowboarders on the slope of a mountain). The external knowledge can be retrieved from a knowledge graph (e.g., ConceptNet [49], DBpedia [3], Wikidata [56]), approach employed in Ahab [57], and other works [37, 47, 58, 69], or an external textual source (e.g., Wikipedia) [38].

In some application domains the additional knowledge necessary for generating the answer can be found associated to the image in the form of natural language text. For instance, the answer to a question about a figure in a book might be contained in the surrounding text. A noteworthy scenario involves the cultural heritage domain, where external knowledge is often provided as an informative sheet associated to the cultural asset. Available datasets that contain natural language text associated with images (e.g., MS-COCO [34], ImageNet [22]) either do not contain question/answer pairs or their descriptions are not detailed enough for finding the answer to meaningful questions. Moreover, the cultural heritage domain contains specific characteristics that make models trained from other domains barely adaptable. For instance, the painting technique, the degree of conservation, the architectural style, are all specific features of the cultural heritage domain and can barely be learned from other domains.

In the cultural heritage domain most approaches have focused on classifying [16, 39, 40, 52] and recognizing [21, 30, 53] artworks. Detailed overviews of approaches for understanding and extracting patterns from artwork can be found in recent reviews [14, 17]. Del Chiaro et al. [20] provided NoisyArt, a dataset of artwork images taken from different perspectives, with their association to DBpedia entities. The dataset contains 89,095 images that refers to 3,120 artworks. Specific datasets for VQA on the cutural heritage domain are limited to AQUA [27] and an annotated subset of Artpedia [9]. AQUA contains three datasets (Train, Validation, and Test) with 69,812, 5,124, and 4,912 question-answer pairs, respectively, associated with 21,384 images of paintings. The Artpedia-based VQA dataset [9] is composed by 30 Artpedia [50] paintings, each one associated to textual descriptions from Wikipedia, with manually generated question-answer pairs. The dataset we propose in this article is orders of magnitude larger than existing datasets, since

it is composed by ~6.5M question/answer pairs, associated with ~500K images of cultural assets. Moreover it covers a much broader variety of cultural assets, that includes paintings, statues, finds, prints, and churches.

Recent work has focused on developing models able to reason on artwork images and an associated knowledge base, with the goal of answering complex questions about the artwork. Zheng et al. [67] proposed a model that generates the answer starting from embeddings of the image, the question and the knowledge graph. Yan et al. [60] considers the problem of capturing the association between artwork visual content and affective explanations. Other work [4, 15] has dealt with the problem of generating informative captions of paintings by considering style, content and contextual knowledge. Biten et al. [8] has focused on the use of the information conveyed by text within an image. None of these works consider the scenario where the external knowledge is expressed in a natural language text document associated with the image.

3 BUILDING VISCOUNTH: A LARGE VISUAL AND CONTEXTUAL QUESTION ANSWERING DATASET FOR CULTURAL HERITAGE

The need for large datasets in the Cultural Heritage domain has motivated us to exploit the large and detailed amount of structured data in the ArCo Knowledge Graph [13] to produce a comprehensive VQA dataset, useful for training and evaluating VQA systems.

ArCo consists of (i) a network of seven ontologies (in RDF/OWL) modeling the cultural heritage domain (with focus on cultural assets) at a fine-grained level of detail, and (ii) a Linked Open Data dataset counting ~200M triples, which describe ~0.8M cultural assets and their catalog records derived from the **General Catalog of Italian Cultural Heritage (ICCD)**, i.e., the institutional database of the Italian cultural heritage, published by the **Italian Ministry of Culture (MiC)**. The ArCo ontology network is openly released with a CC-BY-SA 4.0 license both on *GitHub*[1] and on the official *MiC website*,[2] where data can be browsed and acceded through the SPARQL query language.[3]

Extracting information from ArCo to generate a dataset for VQA is not free of obstacles. First, ArCo does not give us a measure of which kind of questions might be interesting for average users in a real scenario. Second, ArCo data need to be suitably transformed and cleaned to produce answers in a usable form and questions need to be associated to corresponding answers. Third, the dataset we aim at generating is huge, and therefore manual validation of produced data cannot be performed.

3.1 A Semi-automatic Approach for Generating the VQA Dataset

To create our VQA dataset, we resorted to a semi-automatic approach that involves the collaboration of expert and non-expert users and the use of text processing and natural language processing techniques to obtain an accurate list of question-answer pairs. We considered a scenario where an image is associated to available knowledge either manually (e.g., artworks in a museum can be associated with their descriptions) or by object recognition (e.g., architectural properties identified by taking pictures), and generated a dataset as a list of question-answer pairs, each one associated to an image, a description and a set of available information items. An instance of question-answer pair is: "Who is the author?"—"The author of the cultural asset is Pierre François Basan."

Our semi-automatic approach consisted in two main steps. The first part of the process focused on generating a list of question types with associated verbal forms by considering both expert

---

[1]https://github.com/ICCD-MiBACT/ArCo/tree/master/ArCo-release.
[2]http://dati.beniculturali.it/.
[3]https://www.w3.org/TR/rdf-sparql-query/.

and non-expert perspectives, the latter assessed by surveys. Then, for each question type, we automatically generated a list of question-answer pairs by combining question forms and associated answer templates with information from relevant cultural assets in ArCo, and accurately cleaning the results. This process was performed by an ad hoc tool, developed following a build-and-evaluate iterative process. At each step, we evaluated a sample of the produced dataset to propose new data cleaning rules for improving results. The process ended when the desired accuracy was achieved. Eventually, question-answer pairs from different question types were combined. Next, we first detail our question types generation process, then fully describe the question-answer pairs generation by drawing from question types.

The *question types generation* process was based on the following two perspectives carried out independently: a *domain experts' perspective*, represented by a selection of natural language **competency questions (CQs)** [41] previously considered to model the ArCo ontology network [13], and a *user-centered perspective*, represented by a set of questions from mostly non-expert (65 of 104) users, collected through five questionnaires on a set of different images of cultural assets belonging to ArCo (five cultural assets per questionnaire). In the questionnaires, the users were asked to formulate a number of questions (minimum 5, maximum 10) that they considered related to each image presented (questions they would ask if they were enjoying the cultural asset in a museum or a cultural site). In this way, we collected 2, 920 questions from a very heterogeneous group of users in terms of age (from 24 to 70 years old and 42 years average age), cultural background and interests. Subsequently, the questions were semi-automatically analyzed and annotated to recognize their semantics, associate them (when possible) with ArCo's metadata, and create corresponding SPARQL queries for data extraction.

In the clustering process, we grouped user-produced questions into semantic clusters, named *question types*, with the purpose of grouping together questions that ask for the same information. Clustering was first performed automatically by text analysis and sentence similarity, then validated and corrected manually. The automatic procedure consisted in the following steps. We initially aggregated sentences that resulted to be identical after tokenization, lemmatization, and stop words removal. Then, for each question, we identified the most semantically similar one in the whole set by Sentence-BERT [43] and aggregated sentences whose similarity was above 84% (we found empirically that this value resulted in a low error rate). Eventually, we performed average linkage agglomerative clustering with a similarity threshold of 60%. To prepare for manual validation, we extracted a list of question forms, each one associated to a numerical ID representing the cluster it belongs to. Questions in the same cluster (e.g., "Who is the author?" and "Who made it?") were placed close to each other. After removing identical sentences, we obtained about 1,659 questions, grouped in 126 clusters. Each question was then manually associated to a textual (human meaningful) ID (e.g., "AUTHOR") agreed by the annotators and a special "NODATA" ID (about 10%) was introduced for questions that refer to information that is not contained in ArCo. Table 1 gives an overview of the question types generation process, where the effort of users and experts is combined. Each question type is labeled as "Expert" if it comes from the competency questions of ArCo ontology network and has been formulates by the team of experts (counted once in column Mention), "Users" if the question was formulated by non-expert users through the questionnaire, or "Both" if both users and experts proposed such a question (possibly with different verbal forms). At the end of the process, after excluding clusters that refer to unavailable and unusable information, we obtained 43 question types, with 20 of them referred by both users and experts.

In addition, the experts grouped the question types into three categories based on their nature. Most questions (31) were labeled as "contextual," as it was not possible to find the appropriate answers in the images associated with the question type considered (e.g., "DATING"). Instead, eight question types were defined as "visual" (e.g., "BLACKANDWHITE"), since the answers can

Table 1. The 43 Question Types Associated to Their 427 Verbal Forms and to the Number of Times They Are Proposed (Column Mentions) by Experts and/or Non-expert Users

| Question type	Verbal forms	Mentions	Expert/Users
TYPE	6	18	Both
CONSERVATION	6	15	Both
DATINGCRITERION	1	1	Expert
CULTURALSCOPE	28	46	Both
DATING	81	294	Both
OWNER	6	12	Both
PREPARATORYWORK	1	1	Expert
CLIENT	19	55	Users
TITLE	8	28	Both
SUBJECT	35	166	Both
MATERIALORTECHNIQUE	4	6	Both
AUTHOR	51	320	Both
LOCATION	51	314	Both
MEASUREMENT	14	50	Both
ROLEAUTHOR	1	1	Expert
AFFIXEDTECHNIQUE	1	1	Expert
AUTHORCRITERION	1	1	Expert
AFFIXEDPOSITION	1	1	Expert
AFFIXEDELEMENT	1	1	Expert
CATEGORY	1	1	Expert
AFFIXEDTRANSCRIPT	3	6	Both
HISTORICALINFO	27	45	Users
EVENTNAME	1	1	Both
AFFIXEDLANGUAGE	1	1	Expert
USEFUNCTION	2	5	Both
TECHNIQUE	17	75	Both
USETIME	2	2	Expert
FOUNDLOCATION	2	14	Users
EVENTTIME	1	1	Expert
MOTIVATION	8	13	Users
MATERIAL	21	70	Both
SHAPE	1	1	Both
AFFIXEDAUTHOR	1	1	Expert
USECONDITIONS	1	1	Expert
DECORATIVEPURPOSE	1	1	Expert
DEDICATION	2	2	Users
STORAGE_LOCATION	2	6	Users
EXHIBITION_LOCATION	1	1	Users
BOOK	3	3	Users
PURPOSE	10	20	Both
ORNAMENTALMOTIV	1	1	Both
BLACKANDWHITE	1	1	Users
EVENTSITE	1	1	Expert
Total	427	1,604	—
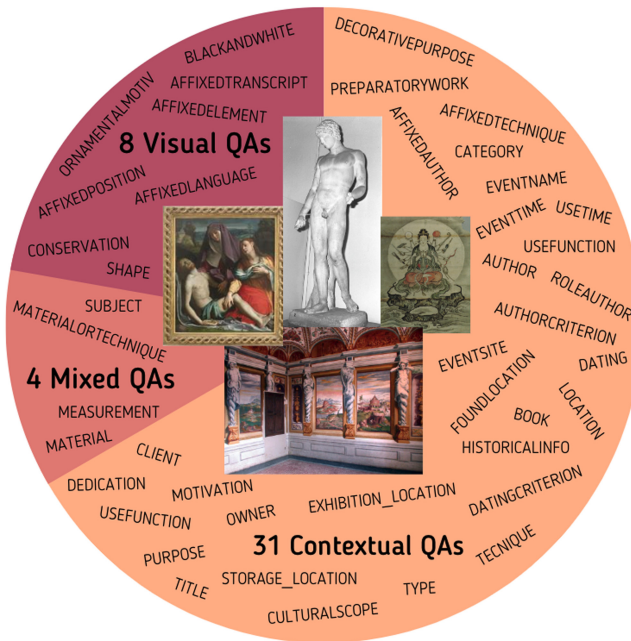
Fig. 2. Overview of the 43 question types of QA labeled as "visual," "contextual," and "mixed." At the center some images representative of the types of cultural assets (e.g., PAINTING, SCULPTURE, PRINT, FRESCO) present in VISCOUNTH.

be inferred from the images associated to the cultural asset, while for four "mixed" question types the answers derive both from visual and contextual information (e.g., "SUBJECT"). Figure 2 depicts all 43 question types of QA split into this three categories, and some examples of images of cultural assets (i.e., PAINTING, SCULPTURE, PRINT, FRESCO) to which they are associated. Eventually, the experts defined an answer template and a SPARQL query for each question type.

We employed SparqlWrapper[4] for executing the SPARQL queries and extracting textual data and pictures from ArCo. We removed cultural assets that have zero or more than one associated pictures. For each record of the query results, we generated a question-answer pair by randomly drawing a question verbal form by the set of appropriated verbal forms in the associated question cluster, with the same distribution of the results of the user questionnaires (frequently proposed questions were selected with higher probability), and building the associated answer from the answer template.

Some question verbal forms are appropriate only for specific types of cultural assets (e.g., "Who was it painted by?" makes sense only for paintings). To establish the appropriated verbal forms for a cultural assets, we mapped both question verbal forms and cultural assets with corresponding macro-categories (we defined nine macro-categories, i.e., SCULPTURE, OBJECT, PHOTO, FRESCO, CHURCH, FIND, PRINT, PAINTING, OTHER). Since this information is not available in ArCo, we considered the available textual description of the cultural asset category to build the mapping. Due to the multitude of categories, we performed a filtering and mapping operation to bring the wide range of types back into a small but explanatory set. As a state-of-the-art work on Italian cultural heritage, we took into account the controlled vocabularies defined by the ICCD-MiC,[5]

---

[4]https://github.com/RDFLib/sparqlwrapper.
[5]http://www.iccd.beniculturali.it/it/strumenti-terminologici.

which also provided the data for ArCo KG [13]. These controlled vocabularies ensure a standardized terminology for the description and cataloging of cultural heritage and help overcome the semantic heterogeneity that is often present in creating such catalogs. First, we filtered the vocabularies' elements closest to the type of artworks to which users refer in their questions. We mapped each textual description of category with an entry in the controlled vocabularies. As detailed in Reference [12], we used a string matching algorithm that takes as input a list of words from a well-defined taxonomy and a general description in free text and returns the equivalent term from the reference taxonomy.

To improve both the form of the answer itself and its rendering in its context, we adopted two approaches. First, we applied a set of cleaning rules, such as removing data with errors and changing patterns of verbal forms (e.g., from "Baldin, Luigi" to "Luigi Baldin").[6] Second, we employed pre-trained language models to improve the form of conversational answers by adapting each sentence to its associated datum (e.g., Italian prepositions and articles have to be chosen according to the gender and number of corresponding nouns or adjectives). To solve this problem, we applied the cloze task of BERT [23] on the generated answers, asking to infer words whose genre and number depend on the specific datum and cannot be previously determined.[7] Furthermore, we applied a final grammar correction task by automatic translating the sentence from Italian to English and back to Italian by means of a pre-trained language models for translation.[8]

Eventually, we automatically generated the description of each cultural asset by combining the long answers of all associated question-answer pairs, since this information is not available in ArCo.

3.2 A Large and Detailed VQA Dataset for Cultural Heritage

The generated VQA dataset contains 6.49M question-answer pairs covering cultural assets, 43 question types and 427 verbal forms. The number of question-answer pairs per template ranges from 35 to 576K. Each question-answer pair is associated with the corresponding cultural asset and its information, including its picture, a description and its URI in ArCo. The number of question types associated to each image depends on the cultural asset's type and ranges from a minimum of 1 to a maximum of 26 question types associated to a certain cultural asset, as in the example of 26 IDs associated to the "PRINT" depicted in Figure 3.

The final dataset is the largest resource available for training and validating VQA models in the cultural heritage domain. It comprises 6, 493, 867 question-answer pairs, with associated visual, textual and structured information. In Table 2, we report this data in comparison to the AQUA [27] dataset statistics. In contrast to AQUA, we consider a new dimension that incorporates mixed (contextual and visual) question types. Additionally, our dataset is two orders of magnitude larger than AQUA.

We associate each cultural asset in our dataset with a set of question-answer pairs, with both a long conversational answer and a short synthetic answer, an image, a natural language description, its URI in ArCo, the reference ontology class and its type. In addition, we provide information on the text span of the answer in the description, when possible.

We make our dataset available on GitHub.[9] We also provide two samples in Italian and English of 50 question-answer pairs per question type that we manually evaluated. Results show an overall accuracy of the long answers (percent of correct entries) of 96.6% for the Italian sample, and of 93%

---

[6]a complete list is available on https://github.com/misaelmongiovi/IDEHAdataset.
[7]https://huggingface.co/dbmdz/bert-base-italian-uncased.
[8]https://huggingface.co/Helsinki-NLP/opus-mt-it-en and opus-mt-en-it.
[9]Cf. https://github.com/misael77/IDEHAdataset.

Table 2. Comparison of Statistics from the VISCOUNTH and AQUA [27] Datasets

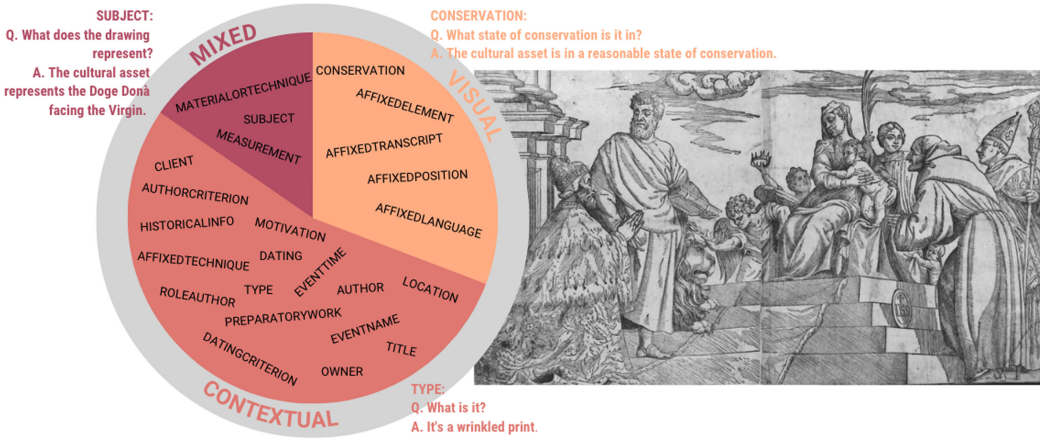|	AQUA			VISCOUNTH		
	Train	Val	Test	Train	Val	Test
Visual QA pairs	29,568	1,507	127	800,440	100,003	99,748
Contextual QA pairs	40,244	3,617	3,642	3,492,984	437,101	437,254
Mixed QA pairs	0	0	0	901,672	112,281	112,384
QA pairs	**69,812**	**5,124**	**4,912**	**5,195,096**	**649,385**	**649,386**

Fig. 3. Overview of the 26 question types associated to the PRINT representing the Doge Donà facing the Virgin. Sixteen question types are labeled as "contextual," five question types are "visual," and three are "mixed." For each group three examples of natural language question types (i.e., TYPE, CONSERVATION, and SUBJECT) are given.

for the English one. We also provide statistics that reports, for each question type, its usage, the number of associated question forms, the number of question-answer pairs generated, and the accuracy. The distribution of cultural asset types in the dataset is provided in Figure 4. The most common question type are "TYPE", "TITLE," and "MATERIALORTECHNIQUE," while "EVENTSITE," "PURPOSE," and "BLACKANDWHITE" have fewer associated cultural assets. Excluding cultural assets not classified in a specific category ("OTHER"), the macro categories with more elements are "OBJECT" (26%) and "PAINTING" (13%), while the less populated one is "FRESCO" (<1%).

Furthermore, Table 3 shows the breakdown of the number of question-answer pairs by cultural asset type and question type.

4 A VQA MODEL FOR CULTURAL HERITAGE

Visual Question Answering for Cultural Heritage requires to analyze two heterogeneous sources of information: an image depicting the artwork and a textual description providing external contextual knowledge. A model capable of effectively providing answers to both visual and contextual questions must therefore combine computer vision and natural language processing. In literature, however, most approaches deal with either one of the two modalities. To understand the challenges posed by our proposed dataset, we first propose single-modality baselines from the state of the art:

- DistilBert [45] is a very common language transformer trained by distilling the Bert base model [23]. It results to be lighter and faster with respect to Bert thanks to knowledge
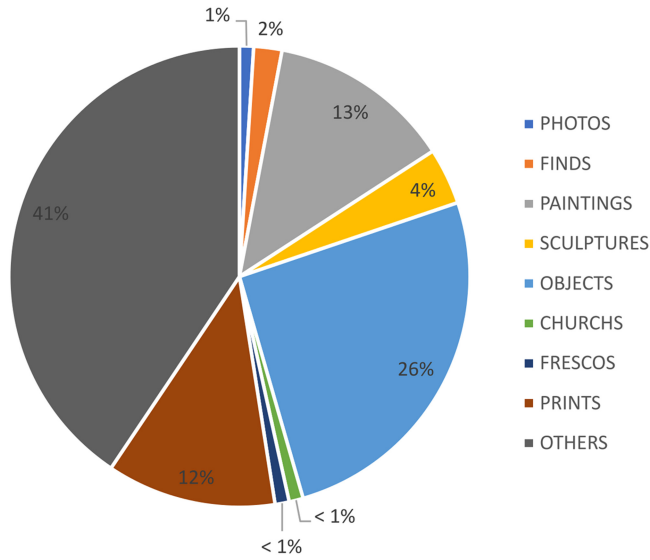
Fig. 4. Distribution of the cultural asset's typology in VISCOUNTH dataset.

distillation used at training time. For this reason the size of the DistilBert model is 40% lower, while retaining 97% of its language understanding capabilities and being 60% faster. This model can then be fine-tuned with good performances on a wide range of tasks.

- RoBERTa [35] has the same architecture of Bert [23] but is trained with optimized parameters, employs a different tokenizer and uses a different pretraining scheme.
- LXMERT [51] is a Large multimodal transformer for vision and language. It consists of three encoders: a visual encoder, a language encoder and a cross-modality encoder. This model is pretrained with large amounts of image-and-sentence pairs via diverse pretraining tasks. It has been shown that this model can achieve impressive results on different downstream multimodal tasks after an appropriate finetuning.

We then propose a multi-modality baseline model by combining DistilBert and LXMERT with a question classifier, that predicts whether the question is contextual or visual and thus if a text-based model (DistilBert) or a vision-based model (LXMERT) is required. Similar approaches have been previously adopted in VQA for cultural heritage [9, 27]. The question classifier is based on Bert [23]. We finetuned a Bert model with a binary classifier on top. The model predicts if a given question is visual or contextual. Depending on the classifier prediction, the question is passed to the most suitable branch (vision model or text-based model) together with additional information (image or textual description).

All models have been trained/finetuned using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32 on an Nvidia Titan RTX.

5 RESULTS AND DISCUSSION

5.1 Evaluation Metrics

To evaluate VQA models on the collected dataset, we follow the standard evaluation setting proposed in Reference [42]. We rely on two metrics, Exact match and Macro-averaged F1 score:

- *Exact match* measures the percentage of predictions that exactly match the ground-truth answer.

Table 3. Number of Question-answer Pairs by Cultural Asset Typology

Question type	PHOTO	FINDS	PAINTING	SCULPTURE	OBJECT	CHURCH	FRESCO	PRINT	Other	Total
TYPE	27,244	0	68,938	24,832	157,849	1,907	19	51,829	244,379	576,997
CONSERVATION	0	0	66,890	21,560	115,554	308	3	51,518	184,124	439,957
DATINGCRITERION	0	0	64,075	21,107	116,134	560	4	50,074	187,720	439,674
CULTURALSCOPE	0	0	26,744	13,765	96,606	1,828	3	9,976	140,848	289,770
DATING	25,247	0	68,589	23,343	130,031	957	4	51,598	192,023	491,792
OWNER	0	0	65,991	23,443	142,577	1,308	17	50,195	241,347	524,878
PREPARATORYWORK	0	0	14,256	4,790	33,646	15	3	18,672	37,295	108,677
CLIENT	0	0	4,310	1,170	641	0	0	1,663	4,153	11,937
TITLE	0	0	68,364	24,683	157,037	1,753	18	50,975	267,023	569,853
SUBJECT	0	0	64,307	19,904	67,791	0	3	48,102	94,791	294,898
MATERIALORTECHNIQUE	0	0	68,871	24,177	150,141	0	19	51,220	244,285	538,713
AUTHOR	21,432	0	37,994	7,523	34,128	221	0	40,507	40,105	181,910
LOCATION	0	104,210	47,797	14,580	103,088	0	0	48,426	138,830	456,931
MEASUREMENT	0	0	17,131	5,666	84,490	7	19	45,719	116,900	269,932
ROLEAUTHOR	0	0	10,207	2,949	27,387	228	0	18,014	35,828	94,613
AFFIXEDTECHNIQUE	0	0	17,987	2,721	20,012	0	0	22,817	61,846	125,383
AUTHORCRITERION	0	0	36,710	7,393	28,452	95	0	41,122	55,648	169,420
AFFIXEDPOSITION	0	0	19,864	3,235	38,381	50	0	24,442	56,950	142,922
AFFIXEDELEMENT	0	0	23,092	4,186	49,996	68	0	34,567	78,517	190,426
CATEGORY	0	0	0	1,186	29,216	12	15	0	75,102	105,531
AFFIXEDTRANSCRIPT	0	0	21,272	3,420	31,908	33	0	31,117	62,372	150,122
HISTORICALINFO	0	0	18,912	4,776	21,591	3	6	11,807	35,719	92,814
EVENTNAME	0	0	7,764	1,546	4,344	0	0	3,044	4,182	20,880
AFFIXEDLANGUAGE	0	0	6,922	1,082	15,536	0	0	5,890	26,202	55,632
USEFUNCTION	0	0	37	313	4,181	1,392	0	8	12,594	18,525
TECHNIQUE	0	0	36	315	4,016	0	0	0	13,543	17,910
USETIME	0	0	0	3	551	44	0	0	1,171	1,769
FOUNDLOCATION	0	11,173	25	1	557	0	0	16	129	11,901
EVENTTIME	0	0	7,318	1,536	4,247	0	0	3,509	3,810	20,420
MOTIVATION	0	0	2,151	960	319	0	0	1,402	2,756	7,588
MATERIAL	0	0	36	318	5,716	0	0	8	16,716	22,794
SHAPE	0	0	7,180	715	3,255	0	0	3,052	5,617	19,819
AFFIXEDAUTHOR	0	0	2,439	225	3,599	0	0	4,325	1,067	11,655
USECONDITIONS	0	0	20	299	1,878	0	0	0	3,998	6,195
DECORATIVEPURPOSE	0	0	0	6	647	0	0	0	1,349	2002
DEDICATION	0	0	0	0	914	0	0	354	1	1,269
STORAGE_LOCATION	0	0	2,412	58	411	0	0	1,185	862	4,928
EXHIBITION_LOCATION	0	0	758	24	27	0	0	4	92	905
BOOK	0	0	0	0	588	0	0	315	151	1,054
PURPOSE	0	0	0	0	8	11	0	0	104	123
ORNAMENTALMOTIV	0	0	0	0	432	0	0	0	753	1,185
BLACKANDWHITE	0	0	0	0	0	0	0	0	128	128
EVENTSITE	0	0	0	0	2	0	0	0	33	35
Total	73,923	115,383	869,399	267,810	1,687,884	10,800	133	777,472	2,691,063	6,493,867

- *Macro-averaged F1 score* measures the average overlap between the predicted answer and the ground truth. Both answers are considered as a set of unordered words among which the F1 score is computed. F1 scores are averaged over all questions in the dataset.

Note that for both metrics, we do not consider articles and punctuations.

In addition, text-based models generate variable length sentences as a subset of the textual description, whereas vision-based models pick a a candidate among a predefined dictionary of possible answers. In both cases, we take the set of words and compare it to the ground truth to compute Exact match and F1 score.

5.2 Evaluation

We carry out a quantitative evaluation by first testing off-the-shelf language pre-trained models. We do not expect such models to perform well on visual questions, but we want to assess whether such models can exploit their language understanding to comprehend questions relative to the cultural heritage domain. As detailed in Section 4, we use as text-based models RoBERTa [35] and DistilBert [45]. Both datasets have been pre-trained on SQUAD [42], a reading comprehension dataset with more than 100,000 questions-answer pairs crowd-sourced on a set of Wikipedia articles.

Table 4. F1-score and Exact Match (EM) for Different Models on Contextual Questions

| | Pretrained | | Finetuned | | | | | |
| | RoBERTa [35] | Distilbert [45] | Distilbert [45] | | LXMERT [51] | | Ours | |
Metric	F1	F1	F1	EM	F1	EM	F1	EM
AFFIXEDTECHNIQUE	0.00	0.06	**0.28**	**0.16**	0.00	0.00	0.00	0.00
CULTURALSCOPE	0.00	0.10	**0.84**	**0.40**	0.00	0.00	**0.84**	**0.40**
EVENTNAME	0.00	0.03	**0.97**	**0.86**	0.00	0.00	**0.97**	**0.86**
OWNER	0.01	0.10	**0.93**	**0.92**	0.00	0.00	0.49	0.27
TECHNIQUE	0.14	**0.58**	0.46	**0.23**	0.00	0.00	0.46	**0.23**
ROLEAUTHOR	0.00	0.15	**0.64**	**0.57**	0.00	0.00	**0.64**	**0.57**
TYPE	0.03	0.08	**0.29**	**0.20**	0.00	0.00	0.22	0.18
LOCATION	0.03	0.15	**0.96**	**0.91**	0.00	0.00	**0.96**	**0.91**
TITLE	0.03	0.21	**0.98**	**0.97**	0.00	0.00	0.93	0.90
DATING	0.01	0.40	**0.73**	**0.71**	0.00	0.00	**0.73**	**0.71**
DATINGCRITERION	0.00	0.01	**0.81**	**0.66**	0.00	0.00	**0.81**	**0.66**
HISTORICALINFO	0.00	**0.06**	0.00	**0.00**	0.00	**0.00**	0.00	**0.00**
AUTHORCRITERION	0.12	0.03	**0.52**	**0.43**	0.00	0.00	**0.52**	**0.43**
CATEGORY	0.00	0.06	**0.39**	**0.16**	0.00	0.00	**0.39**	**0.16**
AUTHOR	0.01	0.19	**0.99**	**0.91**	0.00	0.00	**0.99**	**0.91**
DEDICATION	0.24	0.38	**0.98**	**0.96**	0.00	0.00	**0.98**	**0.96**
USEFUNCTION	0.38	0.33	**0.96**	**0.92**	0.00	0.00	**0.96**	**0.92**
FOUNDLOCATION	0.01	0.29	**1.00**	**1.00**	0.00	0.00	**1.00**	**1.00**
EVENTTIME	0.03	0.03	**0.32**	**0.03**	0.00	0.00	**0.32**	**0.03**
PREPARATORYWORK	0.14	0.02	**0.99**	**0.99**	0.00	0.00	**0.99**	**0.99**
STORAGE_LOCATION	0.01	0.08	**0.96**	**0.96**	0.00	0.00	**0.96**	**0.96**
CLIENT	0.07	0.21	**0.95**	**0.91**	0.00	0.00	**0.95**	**0.91**
DECORATIVEPURPOSE	0.13	**0.18**	0.00	**0.00**	0.00	**0.00**	0.00	**0.00**
USECONDITIONS	0.04	0.07	**0.96**	**0.47**	0.00	0.00	**0.96**	**0.47**
MOTIVATION	0.01	0.13	0.89	**0.49**	0.00	0.00	**0.98**	**0.49**
EXHIBITION_LOCATION	0.01	0.03	**0.67**	**0.63**	0.00	0.00	**0.67**	**0.63**
AFFIXEDAUTHOR	0.01	0.46	0.86	**0.67**	0.00	0.00	**0.89**	**0.67**
USETIME	0.18	0.04	**0.95**	**0.75**	0.00	0.00	**0.95**	**0.75**
PURPOSE	0.00	**0.03**	0.00	**0.00**	0.00	**0.00**	0.00	**0.00**
BOOK	0.10	0.08	**0.57**	**0.54**	0.00	0.00	**0.57**	**0.54**
EVENTSITE	0.00	0.00	**0.55**	**0.55**	0.00	0.00	**0.55**	**0.55**
Mean Contextual	0.06	0.15	**0.69**	**0.58**	0.00	0.00	0.67	0.55

Interestingly, when evaluated on contextual questions, such models perform poorly as can be seen in Table 4. Both models are capable of answering with a certain degree of correctness to a few question categories, namely, "DEDICATION" and "USEFUNCTION," with DistilBert obtaining good F1 scores on an additional restricted number of categories such as "TECHNIQUE" and "AFFIXEDAUTHOR." For most of the remaining question categories, we report an F1 close to 0. This suggests the presence of a domain shift between standard question answering datasets (such as SQUAD) and VISCOUNTH. In fact, in art-related question-answers, as well as descriptions, there is often usage of domain specific jargon that is not present in generic text corpora, making the models unable to understand the question or identify the answer within the description.

Nonetheless, although unlikely given the proven capabilities of such pre-trained models, a low F1 could be caused by intrinsic limits in the architectures. To further confirm the presence of a domain shift, rather than some form of model limitation, we fine-tuned the best of the two models, DistilBert, on the VISCOUNTH dataset. This leads to a significant improvement. The model gains on average 54 points of F1-score, obtaining close to perfect results for question types such as "TITLE," "AUTHOR," "FOUNDLOCATION," and "PREPARATORYWORK." Interestingly, for other categories instead DistilBert still reports low scores, close to zero "HISTORICALINFO,"

Table 5. F1-score and Exact Match (EM) for Different Models on Mixed Questions

| | Pretrained | | Finetuned | | | | | |
| | RoBERTa [35] | Distilbert [45] | Distilbert [45] | | LXMERT [51] | | Ours | |
Metric	F1	F1	F1	EM	F1	EM	F1	EM
MATERIALORTECHNIQUE	0.00	0.27	**0.36**	**0.32**	0.27	0.16	**0.36**	**0.32**
SUBJECT	0.04	**0.13**	0.00	**0.00**	0.00	**0.00**	0.00	**0.00**
MEASUREMENT	0.00	0.04	**0.84**	**0.68**	0.00	0.00	0.00	0.00
MATERIAL	0.00	**0.39**	0.09	0.04	0.29	**0.14**	0.29	**0.14**
Mean Mixed	0.01	0.21	**0.32**	**0.26**	0.14	0.07	0.16	0.11

"DECORATIVEPURPORSE," "PURPORSE"). These categories however either are less represented in the data as shown in Table 3 or are intrinsically harder. For instance, the "HISTORICALINFO" category presents a high variability in how questions are formulated and frequently asks for generic concepts, which require a high level reasoning on the description content.

We also perform a similar evaluation with the vision-based model LXMERT [51]. However, two issues must be taken into account. First, as in most vision-based models, since they cannot rely on textual descriptions, the VQA task is treated as a classification task. Answering a question corresponds to selecting the most relevant answer among a dictionary of pre-defined words or short sentences. For this reason, the domain shift is much more emphasized: If the dictionary does not contain terms suitable for cultural heritage, then the model will not perform well. Second, whereas a text-based model could answer visual questions if the requested information is also in the description, a vision-based model cannot answer contextual questions in any way. As a consequence, we cannot apply a pre-trained vision-model due to significant differences in the answer dictionary. But even fine-tuning the model on VISCOUNTH leads to an F1-score of 0. To perform such finetuning, we create a new dictionary of answers by filtering the most frequent answers in the training set. More precisely, we selected the answers that appear more than 8 times.

Moving to mixed questions (Table 5), on the one hand, we can observe a similar behaviour for text-based models, although the overall F1-score is much lower, since visual knowledge is required to answer correctly. On the other hand, LXMERT is able to provide correct answers to some of the questions. Notably, for the "MATERIAL" question type, LXMERT surpasses text-based models by a considerable margin, yet it is unable to answer to "MEASUREMENT" questions, contrary to DistilBert.

As expected, for visual questions, we can observe an opposite trend compared to contextual questions. In Table 6, we report the results, showing that LXMERT can provide for almost all question categories a high rate of correct questions. However, after being fine-tuned on VISCOUNTH, DistilBert is capable of addressing questions related to "AFFIXEDTRANSCRIPT" and "BLACKANDWHITE." This is due to the fact that sometimes the answers can also be found in the textual description.

For most experiments, we report both the macro-averaged F1-score and the **Exact Match (EM)** metrics. It can be noticed that the F1 score is a relaxation of the EM metric in the sense that it allows an answer to be loosely compared to the ground truth, even when not all words are the same, thus accounting for synonyms or different phrasings.

Finally, we evaluate our combined model. We exploit the question classifier to understand which model is more suitable to address a specific question, without looking at the description nor the image. The BERT-based classifier, described in Section 4, obtains a question classification accuracy of 98.4% on the test set, indicating that it is fully capable of understanding the nature of the questions. We do not include mixed questions in training, and at inference time, we consider the question to be either visual or contextual based on the output of the classifier.

Table 6. F1-score and Exact Match (EM) for Different Models on Visual Questions

| Metric | Pretrained | | Finetuned | | | | | |
| | RoBERTa [35] | Distilbert [45] | Distilbert [45] | | LXMERT [51] | | Ours | |
	F1	F1	F1	EM	F1	EM	F1	EM
CONSERVATION	0.00	0.01	0.00	0.00	**0.79**	**0.53**	**0.79**	**0.53**
AFFIXEDLANGUAGE	0.13	0.66	0.01	0.01	**0.67**	**0.66**	0.66	**0.66**
AFFIXEDELEMENT	0.00	0.01	0.00	0.00	**0.83**	**0.83**	**0.83**	**0.83**
AFFIXEDTRANSCRIPT	0.02	0.08	**0.80**	**0.69**	0.05	0.04	0.04	0.04
AFFIXEDPOSITION	0.00	0.01	0.00	0.00	**0.47**	**0.32**	**0.47**	**0.32**
SHAPE	0.00	0.00	0.00	0.00	**0.68**	**0.68**	**0.68**	**0.68**
ORNAMENTALMOTIV	0.00	0.00	0.00	0.00	**0.54**	**0.54**	**0.54**	**0.54**
BLACKANDWHITE	0.00	0.00	0.70	0.70	**0.96**	**0.96**	**0.96**	**0.96**
Mean Visual	0.02	0.10	0.19	0.17	**0.62**	**0.57**	**0.62**	**0.57**

Table 7. F1-score and Exact Match (EM) for Different Models Averaged Over all Question Types

| Metric | Pretrained | | Finetuned | | | | | |
| | RoBERTa [35] | Distilbert [45] | Distilbert [45] | | LXMERT [51] | | Ours | |
	F1	F1	F1	EM	F1	EM	F1	EM
Mean Overall	0.05	0.14	0.57	0.47	0.13	0.11	**0.61**	**0.51**

As can be seen from Tables 4, 5, and 6, the model is able to exploit both models to accurately answer visual and contextual questions, with only a slight drop for language-based samples. For mixed questions, our model is able to improve compared to LXMERT but exhibits a drop compared to DistilBert. This confirms that mixed questions indeed pose a challenge yet to be solved in question answering applications.

In Table 7, we report the overall average scores in terms of F1 and Exact Match. The average is computed as the mean of all category scores, i.e., contextual, mixed, and visual together. Our combined model retains the best results, providing a baseline for future work in visual question answering for cultural heritage.

To better understand the challenges in the dataset, we show a breakdown of results divided by question category and type of cultural property in Table 8. We do this only for visual questions, since contextual questions do not exploit visual information. This table shows how the performance of our approach vary depending on the type of artwork. We can observe, as expected, that there is a gap between the score obtained for different types of artwork on specific question classes. As example the question category "CONSERVATION" (that includes questions about the conservation state of the artwork) results easier for prints than sculptures. Vice-versa, the category "AFFIXEDLANGUAGE" (that has questions about the language of the writing attached to the cultural asset) has better results for sculptures. Finally, we can observe that the category "AFFIXEDTRANSCRIPT," that refers to the text present in the artwork, obtains very low results. This is due to the fact that these kind of questions are very challenging and require the extraction and the understanding of text in images and currently this can be done only with specific networks.

5.3 Qualitative Analysis

In this section, we provide a qualitative analysis of the answers given by our approach to questions in the VISCOUNTH dataset.

The dataset is divided into three main question types: visual, contextual and mixed. For each type there are multiple question categories, which refer to different types of cultural assets. We

Table 8. F1-score Breakdown for Cultural Asset Category and Question Type

	PRINT	OBJECT	OTHER	PAINTING	SCULPTURE	FRESCO	CHURCH
CONSERVATION	0.81	0.79	0.78	0.78	0.77	1.00	0.34
AFFIXEDLANGUAGE	0.61	0.63	0.69	0.78	0.87	—	—
AFFIXEDELEMENT	0.89	0.89	0.78	0.96	0.82	—	0.57
AFFIXEDTRANSCRIPT	0.07	0.09	0.03	0.01	0.01	—	0.00
AFFIXEDPOSITION	0.54	0.61	0.40	0.32	0.22	—	0.11
SHAPE	0.81	0.73	0.71	0.59	0.46	—	—
ORNAMENTALMOTIV	—	0.56	0.54	—	—	—	—
BLACKANDWHITE	—	—	0.96	—	—	—	—

We do not report the PHOTO and FIND categories, since no visual question is present for such artworks.

Fig. 5. Qualitative Results. Answers given by our approach for different question categories/classes on different artwork types.

thus expect the answers given by our model to be affected by all this aspects. In Figure 5, we show the behaviour of our model in answering different kinds of questions for different types of cultural assets. For contextual questions, we expect that the answer has to be extracted from a natural language description, therefore a language model is sufficient to answer these questions. As we can see in Tables 3 and 4, our model is able to answer the most common contextual questions in the dataset but has lower performance for questions that appear in few examples. In Figure 5, we can observe how our model is able to answer correctly to different categories of contextual questions ("LOCATION, AUTHOR, TITLE, DATING," etc.) for different types of artworks. For these types of questions, we do not observe different performances for different types of artworks. This is due to the fact that in these cases, our question answering language model is agnostic to visual information, being solely based on textual descriptions.

Confirming the results of Table 5, we observe that our model obtains low performances on mixed questions. This kind of questions result to be very challenging, since they require both visual knowledge and contextual knowledge. For instance, for the "MATERIAL" category, the model should be able to describe the different materials the artworks are made of and learn how to recog-

nize them visually. Our model selects either the vision-based model or the textual-based model to answer a question, hence there is not a specific way to handle this kind of questions, thus leading to a lack of performance.

Regarding visual questions, we can observe from Table 8 that we have a variation in the performances based on the type of artwork for different classes of visual questions. For example, we can observe that the questions of the "SHAPE" category, that refers to the shape of the artwork, as expected, perform better for prints than for sculptures. Moreover, as shown in Figure 5, several artworks contain transcripts and there is a specific question category ("AFFIXEDTRANSCRIPT") for this detail. Our model obtains very low performance on this question class, since it does not contain a specific trained model for scene text extraction.

6 CONCLUSION AND FUTURE WORKS

We presented a large-scale heterogeneous multi-language dataset for visual question answering in the cultural heritage domain. Our dataset contains approximately 6.5M question-answer pairs in Italian and English, spanning 500K cultural assets of different types, including artworks, churches, historical objects and others. Each cultural asset is associated to an image, a natural language description and other information. We presented some baselines that employ and combine machine learning models for both contextual (natural language description) and visual processing. Our results show that fine-tuning on a domain-specific dataset is crucial for this task, thus confirming the utility of our dataset. Our best model achieves an overall accuracy (F1 average) of 0.61. Although these result is promising, we found out that certain question categories are hard to compute, especially the ones that require mixed (visual and contextual) reasoning. We believe that further research in this direction would be beneficial for the cultural heritage field, as well as for other fields where multi-modal (visual and natural language) reasoning is required.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[2] Luigi Asprino, Luana Bulla, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. 2021. A large visual question answering dataset for cultural heritage. In *Proceedings of the 7th International Conference on Machine Learning, Optimization, and Data Science (LOD'21)*. 193–197.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*. Springer, 722–735.

[4] Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5422–5432.

[5] Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. 2021. Visual question answering: Which investigated applications? *Pattern Recogn. Lett.* 151 (2021), 325–331.

[6] Federico Becattini, Andrea Ferracani, Lea Landucci, Daniele Pezzatini, Tiberio Uricchio, and Alberto Del Bimbo. 2016. Imaging novecento: A mobile app for automatic recognition of artworks and transfer of artistic styles. In *Proceedings of the Euro-Mediterranean Conference*. Springer, 781–791.

[7] Francesco Vannoni, Pietro Bongini, Federico Becattini, Andrew David Bagdanov, and Alberto Del Bimbo. 2020. Data collection for contextual and visual question answering in the cultural heritage domain. In *International Conference on Pattern Recognition*.

[8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4291–4301.

[9] Pietro Bongini, Federico Becattini, Andrew D. Bagdanov, and Alberto Del Bimbo. 2020. Visual question answering for cultural heritage. Retrieved from https://arxiv.org/abs/2003.09853.

[10] Pietro Bongini, Federico Becattini, and Alberto Del Bimbo. 2023. Is GPT-3 all you need for visual question answering in cultural heritage? In *Proceedings of the European Conference on Computer Vision (ECCV'22)*. Springer, 268–281.

[11] Mark Bugeja and Elaine Marie Grech. 2020. Using technology and gamification as a means of enhancing users' experience at cultural heritage sites. In *Rediscovering Heritage Through Technology*. Springer, 69–89.

[12] Luana Bulla, Maria Chiara Frangipane, Maria Letizia Mancinelli, Ludovica Marinucci, Misael Mongiovì, Margherita Porena, Valentina Presutti, and Chiara Veninata. 2022. Developing and aligning a detailed controlled vocabulary for artwork. In *Proceedings of the Conference on New Trends in Database and Information Systems (ADBIS'22)*. Springer, 529–541.

[13] Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019. ArCo: The Italian cultural heritage knowledge graph. In *Proceedings of the International Semantic Web Conference (ISWC'19)*. 36–52.

[14] Giovanna Castellano and Gennaro Vessio. 2021. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Comput. Appl.* 33, 19 (2021), 12263–12282.

[15] Eva Cetinic. 2021. Iconographic image captioning for artworks. In *Proceedings of the International Conference on Pattern Recognition*. Springer, 502–516.

[16] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. 2018. Fine-tuning convolutional neural networks for fine art classification. *Expert Syst. Appl.* 114 (2018), 107–118.

[17] Eva Cetinic and James She. 2022. Understanding and creating art with AI: Review and outlook. *ACM Trans. Multimedia Comput., Commun. Appl.* 18, 2 (2022), 1–22.

[18] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the 16th European Conference on Computer Vision (ECCV'20)*. Springer, 104–120.

[19] Rita Cucchiara and Alberto Del Bimbo. 2014. Visions for augmented cultural heritage experience. *IEEE MultiMedia* 21, 1 (2014), 74–82.

[20] Riccardo Del Chiaro et al. 2019. NoisyArt: A dataset for webly-supervised artwork recognition. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP'19)*. 467–475.

[21] Riccardo Del Chiaro et al. 2019. Webly—Supervised zero-shot learning for artwork instance recognition. *Pattern Recogn. Lett.* 128 (2019), 420–426.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'19)*. 4171–4186.

[24] Mihai Duguleană, Victor-Alexandru Briciu, Ionuț-Alexandru Duduman, and Octavian Mihai Machidon. 2020. A virtual assistant for natural interactions in museums. *Sustainability* 12, 17 (2020), 6958.

[25] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Adv. Neural Info. Process. Syst.* 33 (2020), 6616–6628.

[26] Haoyuan Gao et al. 2015. Are you talking to a machine? Dataset and methods for multilingual image question. *Adv. Neural Info. Process. Syst.* 28 (2015), 2296–2304.

[27] Noa Garcia et al. 2020. A dataset and baselines for visual question answering on art. In *Proceedings of the European Conference on Computer Vision*. Springer, 92–108.

[28] George Ioannakis, Loukas Bampis, and Anestis Koutsoudis. 2020. Exploiting artificial intelligence for digitally enriched museum visits. *J. Cult. Heritage* 42 (2020), 171–180.

[29] Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. 2020. Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11125–11132.

[30] Xun Jin and Jongweon Kim. 2017. Artwork identification for 360-degree panoramic images using polyhedron-based rectilinear projection and keypoint shapes. *Appl. Sci.* 7, 5 (2017), 528.

[31] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. Figure QA: An annotated figure dataset for visual reasoning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=H1mz0OyDz.

[32] Dimitrios Kosmopoulos and Georgios Styliaras. 2018. A survey on developing personalized content services in museums. *Pervas. Mobile Comput.* 47 (2018), 54–77.

[33] Ranjay Krishna et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision* 123, 1 (2017), 32–73.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[35] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettle-moyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. Retrieved from http://arxiv.org/abs/1907.11692.

[36] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. Retrieved from http://arxiv.org/abs/1410.0210.

[37] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14111–14121.

[38] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question an-swering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3195–3204.

[39] Thomas Mensink and Jan Van Gemert. 2014. The Rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of International Conference on Multimedia Retrieval*. 451–454.

[40] Federico Milani and Piero Fraternali. 2021. A dataset and a convolutional model for iconography classification in paintings. *J. Comput. Cult. Heritage* 14, 4 (2021), 1–18.

[41] Valentina Presutti et al. 2012. Pattern-based ontology design. In *Ontology Engineering in a Networked World*. 35–64.

[42] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for ma-chine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 2383–2392. https://doi.org/10.18653/v1/d16-1264

[43] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'19)*.

[44] Mengye Ren et al. 2015. Exploring models and data for image question answering. In *Proceedings of the 28th Interna-tional Conference on Neural Information Processing Systems*, Vol. 2. 2953–2961.

[45] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. Retrieved from http://arxiv.org/abs/1910.01108.

[46] Lorenzo Seidenari et al. 2017. Deep artwork detection and retrieval for automatic context-aware audio guides. *ACM Trans. Multimedia Comput., Commun. Appl.* 13, 3s (2017), 1–21.

[47] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-aware visual question answering. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19), the 31st Innova-tive Applications of Artificial Intelligence Conference (IAAI'19), the 9th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'19)*. AAAI Press, 8876–8884. https://doi.org/10.1609/aaai.v33i01.33018876

[48] Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. 2016. A dataset for multimodal question answering in the cultural heritage domain. In *Proceedings of the COLING Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH'16)*. ACL, 10–17.

[49] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[50] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Proceedings of the International Conference on Image Analysis and Processing*. Springer, 729–740.

[51] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[52] Wei Ren Tan, Chee Seng Chan, Hernán E. Aguirre, and Kiyoshi Tanaka. 2016. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'16)*. IEEE, 3703–3707.

[53] Frederik Temmermans, Bart Jansen, Rudi Deklerck, Peter Schelkens, and Jan Cornelis. 2011. The mobile museum guide: Artwork recognition with eigenpaintings and surf. In *Proceedings of the 12th International Workshop on Image Analysis for Multimedia Interactive Services*.

[54] Noelia Vallez, Stephan Krauss, Jose Luis Espinosa-Aranda, Alain Pagani, Kasra Seirafi, and Oscar Deniz. 2020. Auto-matic museum audio guide. *Sensors* 20, 3 (2020), 779.

[55] Nuria Recuero Virto and Maria Francisca Blasco López. 2019. Robots, artificial intelligence, and service automation to the core: Remastering experiences at museums. In *Robots, Artificial Intelligence, and Service Automation in Travel, Tourism and Hospitality*. Emerald Publishing Limited.

[56] Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st Inter-national Conference on World Wide Web*. 1063–1064.

[57] Peng Wang et al. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'17)*.

[58] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. FVQA: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 10 (2017), 2413–2427.

[59] Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proceedings of the European Conference on Computer Vision.* Springer, 451–466.

[60] Jianhao Yan, Wenmin Wang, and Cheng Yu. 2022. Affective word embedding in affective explanation generation for fine art paintings. *Pattern Recogn. Lett.* 161 (2022), 24–29. https://doi.org/10.1016/j.patrec.2022.07.009

[61] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 21–29.

[62] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Adv. Neural Info. Process. Syst.* 31 (2018).

[63] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans. Circ. Syst. Video Technol.* 30, 12 (2019), 4467–4480.

[64] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual Madlibs: Fill in the blank image generation and question answering. Retrieved from http://arxiv.org/abs/1506.00278.

[65] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6281–6290.

[66] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision.* 1821–1830.

[67] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. 2021. Knowledge is power: Hierarchical-knowledge embedded meta-learning for visual reasoning in artistic domains. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2360–2368.

[68] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4995–5004.

[69] Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. Retrieved from https://arXiv:2006.09073.