A Dataset for Multimodal Question Answering in the Cultural Heritage Domain

Shurong Sheng
Department of Computer Science
KU Leuven
Shurong.Sheng@cs.kuleuven.be

Luc Van Gool
Department of Electrical Engineering
KU Leuven
luc.vangool@kuleuven.be

Marie-Francine Moens
Department of Computer Science
KU Leuven
Sien.Moens@cs.kuleuven.be

Abstract

Multimodal question answering in the cultural heritage domain allows visitors to museums, landmarks or other sites to ask questions in a more natural way. This in turn provides better user experiences. In this paper, we propose the construction of a golden standard dataset dedicated to aiding research into multimodal question answering in the cultural heritage domain. The dataset, soon to be released to the public, contains multimodal content about the fascinating old-Egyptian Amarna period, including images of typical artworks, documents about these artworks (containing images) and over 800 multimodal queries integrating visual and textual questions. The multimodal questions and related documents are all in English. The multimodal questions are linked to relevant paragraphs in the related documents that contain the answer to the multimodal query.

1 Introduction

Multimodal Question Answering (MQA) invokes answering a query that is formed using different modalities. This topic combines Computer Vision (CV), Natural Language Processing (NLP) and possibly Speech Recognition (SR). With the increasing use of mobile devices, taking pictures becomes an easy and natural way for people to interact with cultural objects. Therefore, we consider a question composed of a textual query combined with a picture of a cultural object or part thereof, where the picture and the natural language question can provide complementary information. Interest in MQA dramatically increased in recent years (Antol et al., 2015; Malinowski et al., 2015; Zhu et al., 2016).

Malinowski et al. (2015) show techniques for the joint processing of images and natural language sentences, Zhu et al. (2016) introduce object-level visual question answering research, which has some similarity with the visual question answering that we propose in this paper. Yet, no studies exist that regard MQA in the cultural heritage domain, with the objective of improving the user experience. Moreover, state-of-the-art research in the area of cultural heritage is uni-modal and either only demonstrates the recognition of images of a cultural object (Bay et al., 2005; Bay et al., 2006), or is only concerned with the processing of a vocal input (Santangelo et al., 2006; Ardito et al., 2009).

This paper reports on a new dataset constructed to facilitate MQA in the cultural domain. The dataset includes images of 16 fascinating artworks from the old-Egyptian Amarna period, documents with regard to this period, and 805 multimodal questions composed of both the full image and part-of-image level queries with regard to the artworks. A part-of-image level query is a question posed in natural language that regards part of the image of a cultural heritage object. In our dataset, the multimodal questions correspond to linked paragraphs in documents that are part of a larger document collection. Part-of-image level queries ask for more detailed information about an image than the identity of the object it shows. Their resolution requires the joint processing of the images and the natural language questions. All images and their related documents are collected from external web sources and the multimodal questions are collected from people of different ages and backgrounds, via a survey. The dataset is stored as a database using the unrelational database management system MongoDB.

Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pages 10–17, Osaka, Japan, December 11-17 2016.

The remainder of the paper is organized as follows: Section 2 discusses some advantages of using multimodal queries during cultural heritage visits. Section 3 describes how we have built the golden standard dataset and offers some analysis. Section 4 discusses a particular application of the dataset, Section 5 provides concluding remarks.

2 Why Multimodal Question Answering ?

MQA constitutes an innovative way of learning about cultural artefacts and offers a novel user experience as it provides a more natural way of interacting. Currently, cultural heritage information offered to the public is often still:
1) One way communication of predefined content;
2) A one-suits-all supply of information, not at all personalized;
3) Linear in that people are supposed to follow a pre-specified tour, with a fixed sequence in which predefined information is provided, or random, but therefore lacking of context and coherence;
4) Offered through dedicated hardware (e.g., as guided tours with audio devices), not belonging to the user and not providing a lasting souvenir of the visit.

MQA research in the cultural heritage domain intends to change this state of affairs:
1) The querying becomes personal, with users actively exploring parts and aspects of artworks or landmarks.
2) Consequently, the provided answers are personalized.
3) The visit does not have to follow a predefined tour. Instead, the user can influence the sequence through questions and queries, the answers to which may have to be found distributed in a variety of unstructured sources such as full texts and images described by text. Moreover, such system could be made to learn from earlier, multimodal interactions.
4) People can use their own mobile phones, tablets and other personal devices.
5) The MQA will have to bridge the fields of question processing, document content extraction and linking, and information search.

A visitor's multimodal query combined with her context helps turn general information about cultural heritage into a personalized guided tour. Such a framework is still rarely incorporated. A museum or landmark tour can be dynamically altered based on personal multimodal queries and artwork variance, resulting in an adaptive guidance system.

3 Dataset

In the context described so far, personalized and important questions of visitors are supposed to be expressed as a multimodal query. Such query is composed of a photo taken by the visitor of an artwork or of a detail of it, augmented by a question expressed in natural language, as Figure 1 shows.

3.1 Data Collection

The MQA dataset that is constructed consists of the following parts:
1) A selection of artworks of the old-Egyptian Amarna period;
2) A set of multimodal queries built by users, each composed of a natural language question and the photo of the cultural object or of part of it;
3) Textual documentation on the Amarna period;
4) The set of relevant paragraphs extracted from the documentation that answer a multimodal query.

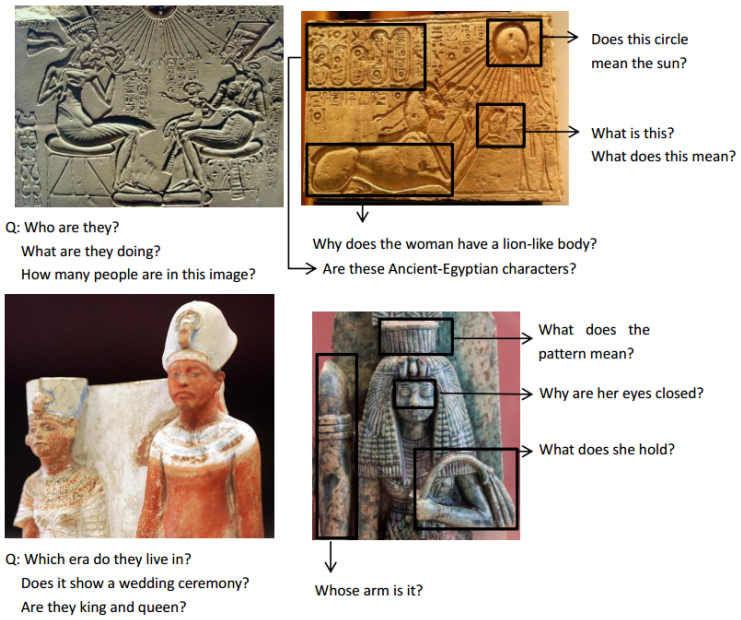


Figure 1. The images in the left column are two examples of full images with corresponding textual questions, and in the right column are two examples of visual queries specified by bounding boxes with corresponding textual questions.

We imagine a museum room where a set of artefacts have been collected for an exhibition. In our case the artefacts regard statutes, relief sculptures and paintings involving famous characters such as queen Nefertiti and pharaoh Akhenaten.

Figure 2. Example artwork images in our dataset, each shown from multiple viewpoints.

Images of Typical Artworks: The picture sets of the objects and their related documents were mainly collected from web sources such as Wikipedia, online collections of some museums, and Google Arts&Culture. Photos for 16 objects from the Amarna period were collected, with for each object up to 10 photos taken from different viewpoints. Example images are shown in Figure 2.

Multimodal Questions: To collect multimodal questions about the selected artworks, we have made a survey. This involved a document that contains photos of 16 artworks and a guideline document that were sent to all members of the department and all e-mail correspondents of the first author. The users were asked to give two kinds of questions for each artwork. The first was a textual question going with the full image and the second was a textual question along with a subpart of the image, marked by the user with a bounding box. Examples are shown in Figure 1. The guideline document explained the on-site scene of these artworks, the role of the users or respondents, and how they should draw a bounding box in the image. Also, the respondents were instructed to pose relevant and diverse questions about the artworks, with the help of some example questions. We finally received 42 responses from respondents with ages ranging from 20 to 60 and with varied backgrounds, and 1142 questions in total with regard to the artworks.

Documents on the Amarna period: We have downloaded multiple websites (we call this 'document collection on the Amarna period' further in this paper) relevant to the Amarna period from the highest ranked web pages by entering the keywords 'Amarna period','Nefertiti','Akhenaten','Tiye','Amarna style', respectively, using the Google search engine, and online collections of some museums and institutes including the Brooklyn Museum[1], Museum of Fine Arts, Boston[2], Egyptian Museum Berlin[3], Europeana[4] and Google Arts&Culture[5] with regard to the Amarna period. There are 204 documents in this collection. We have manually identified those documents that are related to the collected multimodal queries and have identified the paragraphs therein that are relevant to answer the queries.

3.2 Data Processing

The data collected from the Web and obtained from the users was noisy. Hence, several methods have been implemented for data cleaning.

Multimodal Questions: Different users may ask the same questions with regard to an artwork, so we first filtered the data by removing duplicate questions.

We wanted the questions to be of good quality, i.e. exhibiting a strong link between the images and the textual parts. For example, 'Who is the person in the picture?' is a good question, whereas 'Does not everyone like travelling?' clearly is not. To assure this we asked three annotators to check on the questions validity. We keep a question in our dataset when at least two annotators have labeled the query as 'valid', thus considering the textual question as relevant to the corresponding image.

These two steps reduced the number of queries from 1142 to 805. We noticed that quite a few questions were repeated by different users. So we assume that we have collected questions that are representative of what users would typically ask when interacting with the collected artworks in the simulated museum exhibition.

No.	Links	Number of related pictures
1	http://www.egyptian-museum-berlin.com/c53.php	5
2	https://en.wikipedia.org/wiki/Amarna_period	17
3	https://en.wikipedia.org/wiki/Amarna_art	20
4	http://www.heptune.com/art.html	15
5	http://www.touregypt.net/featurestories/amarnaperiod.htm	7
6	https://en.wikipedia.org/wiki/Nefertiti	19
7	http://www.crystalinks.com/nefertiti.html	9
8	http://www.ancient.eu/Nefertiti/	3
9	https://en.wikipedia.org/wiki/Tiye	4
10	http://quatr.us/egypt/art/amarna.htm	4

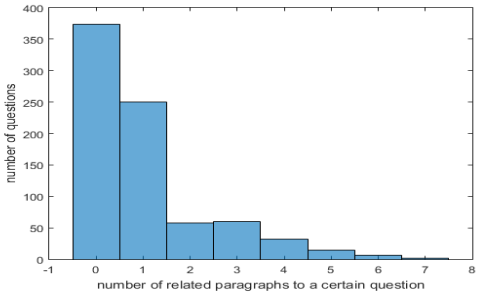
Table 1. List of example sources that form the related documents for the artworks.

Document Labeling: From the query collecting phase, we then get to the document collection. The documents in the collection are annotated as relevant to the collected artworks or not. As it is quite obvious whether a document is related to the collected artworks, a single annotator was asked to label them. With 204 documents about the Amarna period in the dataset, 101 documents (coined the 'related document collection') were labeled as related to at least one of the collected artworks. The remaining 103 unrelated documents were kept in the dataset and they will still be searched by the MQA system.

Question and Paragraph Linking: For our domain-specific question answering system, some of the questions are rather open-ended and the answers may be quite long. For example the answer to the question 'Why is the man's body so strange?' refers to the analysis of Akhenaten's body by several

[1] https://www.brooklynmuseum.org/opencollection/collections
[2] http://www.mfa.org/collections
[3] http://www.egyptian-museum-berlin.com/
[4] http://www.europeana.eu/portal/en
[5] https://www.google.com/culturalinstitute/beta/

experts and contains more than one paragraph from the same document. Therefore, within the related document collection, three annotators were asked to link paragraphs that are relevant for each question, with one of them acting as a judge if there is inconsistency between the other two annotators. Some of these paragraphs are part of the full text of a related document and some are obtained from the captions of pictures in the document (the latter are listed as related pictures in Table 1). Documents are uniquely defined by their URL and paragraphs (including image captions and titles) by their start and end word position in the HTML-document. The number of related paragraphs obtained from the related document collection ranges from 0 to 6 and a histogram of the number of related paragraphs forming an answer is shown in Figure 3(a). An example of a linked question with its sole related paragraph is given in Figure 3(b). In the task that we propose we have now only identified relevant paragraphs in which the answer to the multimodal question can be found and we have not yet identified relevant sentences or phrases that provide the answer. So far, the focus is on the difficult task of interpreting multimodal questions and not (yet) on extracting the answer from the document collection in its condensed form.



(a)

Image id	00801
Textual question	who is the person?
Related paragraph	Unfinished head of a statue of Queen Nefertiti.
Related document	https://www.flickr.com/photos/menesje/8507350000

(b)

Figure 3. (a) Histogram with the number of related paragraphs for each multimodal query; (b) Example of a textual question and related paragraph.

We show the whole processing procedure in Figure 4.

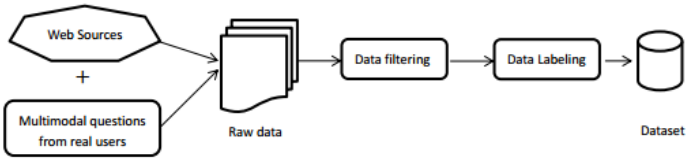


Figure 4. The diagram of the data processing procedure.

3.3 Dataset Analysis

Number of artworks	Number of questions	Number of documents	Document collection size	Number of related docs	Number of related paragraphs
16	805	204	261127	101	139

Table 2. Statistics of the dataset.

Table 2 shows some statistics for the dataset. 'Related paragraphs' denote the number of paragraphs linked to the multimodal questions, and the 'document collection size' describes the size of the document collection by the number of English words it contains. 55% of the questions in the dataset can be answered purely by the related document collection in the dataset. On the other hand, several questions such as 'How many colors does the image contain?' can be answered separately by low-level image analysis. Additionally, questions about cultural heritage objects are sometimes difficult to answer due to the lack of historical knowledge. Also, a lot of questions such as 'Why is the woman so ugly?' 'Do they

really love each other?' are difficult to answer even for a human. We still keep these questions in our dataset as these questions are examples of what humans would ask.

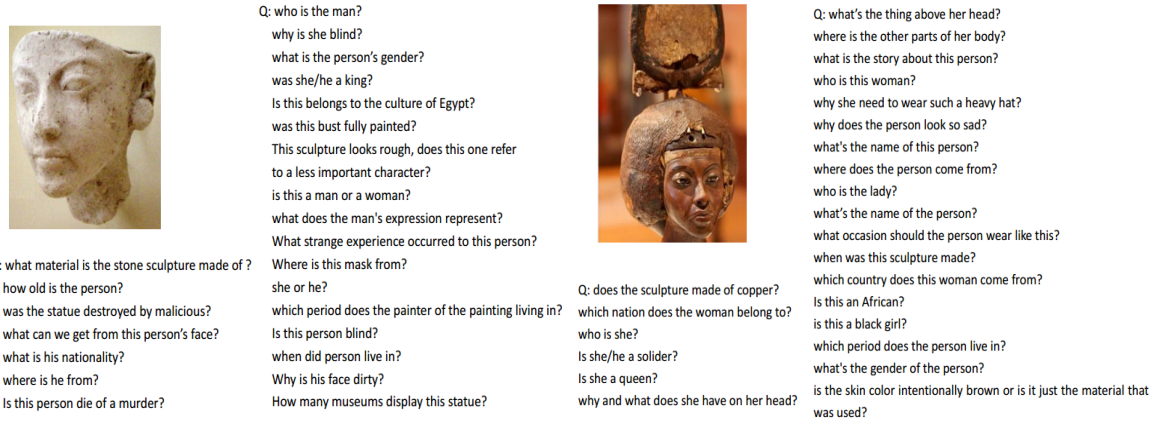
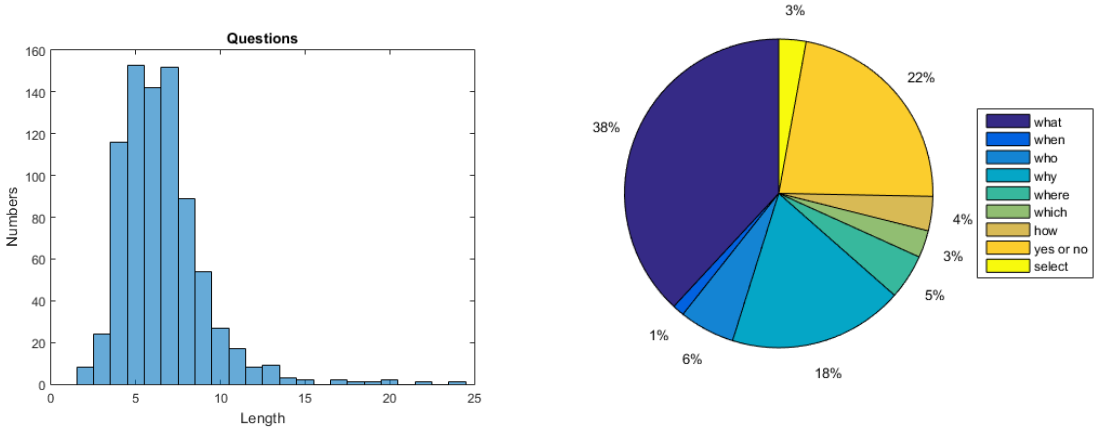


Figure 5. Sample textual questions with their corresponding images.

Some example questions are shown in Figure 5. The questions in this dataset are diverse and can be used for various artificial intelligence tasks, especially for natural language processing. They contain simple image understanding questions with regard to the object class (e.g., 'Is this a woman or man?') and the attributes of the objects (e.g., 'What material is the sculpture made of?'). Some questions are complex and need deep and common sense reasoning. For example, to answer the question 'Why do they give each other a hand?', we should know that this question is about the two persons who wear a similar crown in the image. Based on our common sense, the two persons wearing a similar crown and giving each other a hand should be a couple or more specifically king and queen, and a couple in this ancient culture like to give each other a hand to show their feelings and relationship.



(a) Histogram of the length of the questions in words. (b) Type distribution of the multimodal questions in the dataset.

Figure 6. Statistics related to the collected multimodal questions.

We categorize the questions based on the method of (Gao et al., 2015) into 9 types, the statistics of which are shown in Figure 6(b):
1) What: questions about the attributes and features of the object.
2) When: questions related to time with regard to the subject.
3) Who : questions about the identity of a person.
4) Why: questions about the reason of some phenomenon.
5) Where: questions about the location of the object.
6) Which: questions that need reasoning about the object.

7) How: questions about the methods related to the object.
8) Yes or No: questions that you can answer with Yes or No.
9) Select: Selective questions.

The questions' length in the dataset ranges from 2 to 24 English words and the average length of the questions is 6.56 words. A histogram of the questions' length in the dataset is shown in Figure 6(a).

Compared to other application areas of MQA such as in-door scenes and wild-life animals, MQA tasks in the cultural heritage domain are more difficult. Due to uniqueness of the artworks and historical reasons, it is hard to obtain a large amount of data for each artwork and thus few training data. As no MQA dataset in the cultural heritage domain has been released so far, our dataset, which will be made public soon, can be regarded as the first benchmark for MQA research in that field.

4 Dataset Application

This dataset is explicitly constructed for facilitating MQA. In this case, the answer estimation problem can be formulated as the probability of an answer $\boldsymbol{a}$ conditioned on a multimodal question $\boldsymbol{q}$ composed of an image $\boldsymbol{q_i}$ and its corresponding textual question $\boldsymbol{q_t}$, as shown in the formula below.

$$\boldsymbol{P(a|q) = P(a|q_i, q_t, \theta)}$$

In this formula, $\boldsymbol{\theta}$ denotes a vector of all parameters to learn. $\boldsymbol{q_i}$ and $\boldsymbol{q_t}$ can be represented as real valued vectors: $\boldsymbol{q_i} = [\boldsymbol{w_{1i}, w_{2i}, ...w_{pi}}]$, $\boldsymbol{q_t} = [\boldsymbol{w_{1t}, w_{2t}, ...w_{nt}}]$ or other forms that the computers can directly use, $\boldsymbol{q}$ is a joint representation of $\boldsymbol{q_i}$ and $\boldsymbol{q_t}$. With this formula, answers with the highest $\boldsymbol{k}$ probabilities will be retrieved as a ranked answer set $\boldsymbol{A}$. Note that $\boldsymbol{A}$ can be an empty list if all probabilities are zero or are considered too low to yield a valuable answer.

Compared to other popular multimodal datasets composed of visual and language data (e.g., DAQUAR (Malinowski and Fritz, 2014), VQA Antol et al. (2015)), which provide only question-answer pairs that can be easily collected by online crowdsourcing platforms such as Amazon Mechanical Turk (AMT)[6], the dataset described in this paper, which contains a large set of documents and diverse multimodal questions, can give better perspectives for MQA research:

1) Natural language processing of the questions and textual documentation

   Query formation involves keywords selection from raw textual questions. To detect answer types in the dataset, named entity recognition and more specifically the recognition of person and location names and other types of entity information in the question and textual documents can be implemented. Also, the natural language questions and documents can be used to research coreference resolution.
2) Information retrieval models for searching documents and ranking paragraphs

   The dataset can also be used to study suitable information retrieval and answer ranking models for MQA, or to adapt existing retrieval models such as vector space and language models that allow to make inferences over textual and visual data.
3) Cross-modal semantic representation

   The multimodal questions in our dataset can be used to find suitable joint semantic representations of the multimodal data provided by the photo of the object or its part and the textual content of the question, cross-modal distributional semantics integrating text based representation of meaning with information coming from vision and from cross-modal coreference resolution.
4) Cross-modal coreference resolution

   The multimodal queries and documents can be used to research cross-modal coreference resolution by linking mentions of entities in the language and the visual data.
5) Additional image processing

   Several questions in our dataset can be used for variant image analysis tasks including image recognition (e.g., 'What does she hold in her hand?') and object detection (e.g., 'How many persons are in this picture?') and activity recognition (e.g., 'Is the man dancing?').

[6]https://www.mturk.com/mturk/welcome

5 Conclusion and Future Work

This paper describes the dataset manually built for multimodal question answering on a cultural heritage collection. The dataset concerns images of artworks from the ancient Egyptian Amarna period, a document collection relevant to the Amarna period and 805 multimodal questions composed of both natural language questions and images. The data is collected from web sources and processed in several steps including data cleaning, document labeling and question-paragraph linking. By analyzing and classifying the questions, we have proved that the multimodal questions in this paper are very diverse and this dataset can be used for research on many natural language processing tasks such as named entity recognition and coreference resolution, better information retrieval models, cross-modal semantic gap reduction and additional image processing tasks.

In the next stage of our research, we will design and develop real-time processing methods for analyzing the natural language questions and their corresponding pictures. We will also develop methods to instantly retrieve a relevant answer to a question. In the future we might expand the dataset with annotations of fine-grained answers in the form of text phrases or image segments that answer the multimodal queries, if they prove to be useful in our MQA research.

Acknowledgements

The authors would like to thank the respondents of the questionnaire for this dataset. This work is funded by the KU Leuven BOF/IF/RUN/2015, and is a part of the project "Mobile Augmented Reality and Vocal Explanations for Landmarks and visual arts (MARVEL)".

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Carmelo Ardito, Paolo Buono, Maria Francesca Costabile, Rosa Lanzilotti, and Antonio Piccinno. 2009. Enabling interactive exploration of cultural heritage: An experience of designing systems for mobile devices. *Knowledge, Technology & Policy*, 22(1):79–86.

Herbert Bay, Beat Fasel, and Luc Van Gool. 2005. Interactive museum guide: Accurate retrieval of object descriptions. In *Proceedings of the Seventh International Conference on Ubiquitous Computing UBICOMP, Workshop on Smart Environments and Their Applications to Cultural Heritage*.

Herbert Bay, Beat Fasel, and Luc Van Gool. 2006. Interactive museum guide: Fast and robust recognition of museum objects. In *Proceedings of the First International Workshop on Mobile Vision*.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? Dataset and methods for multilingual image question answering. In *Advances in Neural Information Processing Systems (pp. 2296-2304).*

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 1682–1690, Cambridge, MA, USA. MIT Press.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9.

Antonella Santangelo, Agnese Augello, Antonio Gentile, Giovanni Pilato, and Salvatore Gaglio. 2006. A chat-bot based multimodal virtual guide for cultural heritage tours. In *Proceedings of the International Conference on Pervasive Systems and Computing*, pages 114–120.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question answering in images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.