Question Answering Mediated by Visual Clues and Knowledge Graphs

Fabrício F. Faria
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
firminodefaria@ufrj.br

Ricardo Usbeck
Paderborn University
Paderborn, Germany
ricardo.usbeck@uni-paderborn.de

Alessio Sarullo
University of Manchester
Manchester, UK
alessio.sarullo@manchester.ac.uk

Tingting Mu
University of Manchester
Manchester, UK
tingting.mu@manchester.ac.uk

Andre Freitas
University of Manchester
Manchester, UK
andre.freitas@manchester.ac.uk

ABSTRACT

This challenge focuses on the use of semantic representation methods to support Visual Question Answering: given a large image collection, find a set of images matching natural language queries. The task supports advancing the state-of-the-art in Visual Question Answering by focusing on methods which explore the interplay between contemporary machine learning techniques, semantic representation and reasoning mechanisms.

CCS CONCEPTS

• **Information systems** → **Question answering**; *Information extraction*; • **Computing methodologies** → **Visual content-based indexing and retrieval**;

KEYWORDS

Visual Question Answering, Gold-standard

ACM Reference Format:

Fabrício F. Faria, Ricardo Usbeck, Alessio Sarullo, Tingting Mu, and Andre Freitas. 2018. Question Answering Mediated by Visual Clues and Knowledge Graphs. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3184558.3192318>

1 INTRODUCTION

As of 2017, it is estimated that images and videos account for up to 73% of all consumer Internet traffic according to Cisco Visual Networking Index [2]. However, due to its non-symbolic nature, most of the content present in visual form depends on the use of associated textual content or annotations to become accessible and searchable to end-users. These associated textual information provide a limited slice to the full information content (such as entities and relations) expressed in the images. With recent advances in machine learning techniques [6, 17], in subparticular, in the field of

computer vision, the detection and classification of objects embedded in images became a very active research area [11, 13, 15]. More recently, the detection of relations between objects in an image scene [7, 9, 16, 20] defined the use of lexico-semantic graphs as a lightweight representation device for images. The emergence of richer symbolic-level representation models opened the doors to more sophisticated semantic interpretation models and applications such as Visual Question Answering (VQA).

However, addressing the problem of Question Answering over visual data requires principled semantic representation models. These models support semantic approximation and reasoning operations necessary to bridge the gap between queries and the immediate description of an image. In this process the integration of sub-symbolic methods derived from computer vision should be integrated to symbolic AI methods.

This challenge aims at advancing the dialogue between Computer Vision and Natural Language Processing, by creating a test collection which explores the semantic aspects of VQA.

2 DATASET DESCRIPTION

Most natural language queries in the dataset require the integration of one or more knowledge graphs or linguistic resources and explore different types of representation and reasoning on the top of the scene description. The test collection is derived from the Visual Genome Dataset (1.7 million object instances and 2.3 million relationships) [12].

The training dataset is composed by 4475 queries while the test dataset is composed of 514 queries. The datasets are represented in a JSON file where each JSON object is composed by a query and a list of visual genome image ids. The datasets were created by randomly selecting images from the Visual Genome dataset and asking 25 human annotators to generate natural language queries which describe the image as an answer to this query, and which explored broader types of relations, categories and aggregations present within the image. An example of an image and associated query is depicted in Figure 1.

The annotation protocol is available at [3] while the final datasets are available at [4].

The final query set contains a dataset which expresses different semantic categories and semantic relations. Examples of queries contain queries with references to: **sensorial elements** (*Are there*

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '18 Companion, April 23–27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5640-4/18/04.

<https://doi.org/10.1145/3184558.3192318>

Figure 1: Example of images, natural language queries and supporting knowledge graphs.

images of tennis players wearing red clothes?), **metonymic expressions** (Give me recipes with olives in it), **similarity relations** (Show me kites that look like parachutes), **direct spatial references** (Show me a bridge close to NYC), **indirect spatial references** (Find German people protesting, Show me a hydrant in an Asian city), **picture properties** (Give me black and white images of a skate manoeuvre), **category references** (Give me images of a snow sport), **non-essential properties** (Show me an empty kitchen), **counting references** (Show me a meeting room for 6 people), **events** (Show me a traffic accident between a double decker bus and a motorcycle), **actions** (Find a picture of a person feeding birds), **spatial relations** (Show me a sink with a view to trees outside), **normative deviation queries** (Give me images of a modified traffic sign), **affordance** (Give me images of a kind of object which can be used to extinguish fire), **internal textual references** (Show me images of trains to Chester).

3 VISUAL QUESTION ANSWERING: STATE-OF-THE-ART

In comparison to existing datasets this challenge focuses more on capturing the semantic found inside the image, with the aim to better support reasoning mechanisms in complement to statistical learning. We therefore need to investigate whether current datasets are able to capture the semantic capabilities of systems. In the past, several communities hosted workshops and challenges (e.g., ImageCLEF from 2013 - 2018¹ or VQA 2016 - 2017 [1]) creating a multitude of datasets. Also, the parl.ai initiative [14] provides a large set of benchmarks for Visual Question Answering.

For the sake of space in this paper, we just want to highlight only a few datasets, such as FVQADescription [18], VQA version 1 and 2 [1, 5] or CLEVR [8].

Kafle et al. [10] wrote an overview article in 2016 which suggests the following features to categorize datasets: number of total images, number of question-answer pairs, distinct answers, human accuracy, number of words in longest question, longest answer, average answer length, image source, annotation method and evaluation type (such as open-ended (OE) or multiple-choice (MC)) and question types. In 2017, Wu et al. [19] used a different but partly overlapping set of features for their survey paper, namely: number of images, number of questions, image source, number of

questions/image, number of question categories, question collection (aka annotation method), average question length, average answer length and evaluation metric. This challenge brings another dimension to discussion, which is the *types of inference* which can be expressed in information needs. For further information about other available methods or a detailed, structured overview over datasets, we refer the interested reader to two surveys [10, 19].

Acknowledgements This work has been supported by the BMVI projects LIMBO (project no. 19F2029C) and OPAL (project no. 19F20284) as well as by the German Federal Ministry of Education and Research (BMBF) within 'KMU-innovativ: Forschung für die zivile Sicherheit' in particular 'Forschung für die zivile Sicherheit' and the project SOLIDE (no. 13N14456). Furthermore, we would like to thank the annotators Ria Hari Gusmati, Richa Jalota, Paramjot Kaur, Hussain Abid Syed and Jan Reineke.

REFERENCES

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2425–2433. <https://doi.org/10.1109/ICCV.2015.279>
- [2] I Cisco. 2012. Cisco visual networking index: Forecast and methodology, 2011–2016. *CISCO White paper* (2012), 2011–2016.
- [3] Ricardos Usbeck Alessio Sarullo Tingting Mu Fabrício Firmino, André Freitas. 2018. Protocol for Question Answering Mediated by Visual Clues and Knowledge Graphs. <https://visual-question-answering-challenge.github.io/protocol>. (2018). [Online; accessed 04-March-2018].
- [4] Ricardos Usbeck Alessio Sarullo Tingting Mu Fabrício Firmino, André Freitas. 2018. Question Answering Mediated by Visual Clues and Knowledge Graphs - WWW 2018 LYON, FRANCE. <https://visual-question-answering-challenge.github.io>. (2018). [Online; accessed 04-March-2018].
- [5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *CoRR abs/1612.00837* (2016). [arXiv:1612.00837](http://arxiv.org/abs/1612.00837)
- [6] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2017. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* 28, 10 (2017), 2222–2232.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [8] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *CoRR abs/1612.06890* (2016). [arXiv:1612.06890](http://arxiv.org/abs/1612.06890)
- [9] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3668–3678.

¹<http://www.imageclef.org/>

- [10] Kushal Kafle and Christopher Kanan. 2016. Visual Question Answering: Datasets, Algorithms, and Future Challenges. *CoRR* abs/1610.01465 (2016). [arXiv:1610.01465](http://arxiv.org/abs/1610.01465) <http://arxiv.org/abs/1610.01465>
- [11] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3337–3345.
- [12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.
- [13] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.
- [14] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParLAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476* (2017).
- [15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
- [16] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*. 70–80.
- [17] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 157–166.
- [18] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick. 2016. FVQA: Fact-based Visual Question Answering. *CoRR* abs/1606.05433 (2016). [arXiv:1606.05433](http://arxiv.org/abs/1606.05433) <http://arxiv.org/abs/1606.05433>
- [19] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2016. Visual Question Answering: A Survey of Methods and Datasets. *CoRR* abs/1607.05910 (2016). [arXiv:1607.05910](http://arxiv.org/abs/1607.05910) <http://arxiv.org/abs/1607.05910>
- [20] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2.