Automatic multi-modal processing of language and vision to assist people with visual impairments

Hernán Maina[1,2] and **Luciana Benotti**[1,2]
[1]Universidad Nacional de Córdoba
[2]CONICET, Argentina
hernan.maina@mi.unc.edu.ar
luciana.benotti@unc.edu.ar

Abstract

In recent years, the study of the intersection between vision and language modalities, specifically in visual question answering (VQA) models, has gained significant appeal due to its great potential in assistive applications for people with visual disabilities. Despite this, to date, many of the existing VQA models are nor applicable to this goal for at least three reasons. To begin with, they are designed to respond to a single question. That is, they are not able to give feedback to incomplete or incremental questions. Secondly, they only consider a single image which is neither blurred, nor poorly focused, nor poorly framed. All these problems are directly related to the loss of the visual capacity. People with visual disabilities may have trouble interacting with a visual user interface for asking questions and for taking adequate photographs. They also frequently need to read text captured by the images, and most current VQA systems fall short in this task. This work presents a PhD proposal with four lines of research that will be carried out until December 2025. It investigates techniques that increase the robustness of the VQA models. In particular we propose the integration of dialogue history, the analysis of more than one input image, and the incorporation of text recognition capabilities to the models. All of these contributions are motivated to assist people with vision problems with their day-to-day tasks.

1 Introduction

With the advent of the deep learning (DL) era, tasks related to computer vision (CV) and automatic natural language processing (NLP) have managed to deliver promising results with great potential to assist people with visual impairments (Radford et al., 2021). Despite this, most of the advances were made on research benchmarks and are far from being of practical use for people with visual disabilities.

Historically, building automated systems that are capable of exploiting multi-modal models has been considered an ambitious goal. However, this last decade has seen enormous progress in VQA systems (Antol et al., 2015; Goyal et al., 2019; Anderson et al., 2018; Zhang et al., 2015). Given its nature, this recent area of artificial intelligence tries to be the bridge that allows information and visual concepts to be converted into language, through the application of knowledge and the advances achieved in the disciplines of CV and NLP.

Answering a visual question is a task where the system receives a question about an image, and must infer the answer. This task may involve different CV problems such as image classification (Schmarje et al., 2020) *e.g.“is this a cat?”*, object detection (Jocher et al., 2022; Wu et al., 2019; Zhou et al., 2022) *e.g.“are there cats in the image?”*, image attribute extraction (Fang et al., 2014; Yu et al., 2021) *e.g.“what color is the cat?”*, scene classification (Zeng et al., 2021) *e.g.“is it a beach?”*, and counting objects (Trott et al., 2017; Chattopadhyay et al., 2016) *e.g.“how many cats are there?”*. There are also studies on the spatial relationships between objects, which are not always visible from a given point of view (Bansal et al., 2020; Qiu et al., 2020) *e.g.“what is between the cat and the sofa?”*, and on common sense questions *e.g.“why can’t the cat sleep on the couch?”*. Recent work investigates questions such as *“what is written on the player’s shirt?”* that require identifying the text associated with a particular visual object (Kant et al., 2020).

People with visual impairments would benefit from using VQA systems for daily activities such as finding products when going to a supermarket, selecting brands, checking prices or expiration dates. Currently these tasks have been addressed through platforms such as *Be My Eyes*[1] or *BeSpecular*[2] using sighted people on the other side of the application, in order to answer this wide variety of

[1]https://www.bemyeyes.com/
[2]https://www.bespecular.com/

(a) **Qs:** What color do these look?.
Ans: 'orange'.

(b) **Qs:** Could you tell me what's on this can?.
Ans: 'green beans'.

Figure 1: Examples of *answerable* visual questions.

questions. The Figure 1 shows examples drawn from the VizWiz-VQA dataset (Gurari et al., 2018), illustrating some of these needs.

The rest of this paper is organized as follows. Section 2 compares current benchmarks and the three challenges not addressed by them for people with vision problems. The next three sections describe how we propose to address such challenges.

2 Design of a CheckList oriented to vision and language

Current NLP models are often evaluated based on their performance on a series of individual tasks using benchmarks based on natural language datasets such as GLUE (Wang et al., 2018). Recently, CheckList (Ribeiro et al., 2020), proposed to perform this evaluation using a set of test cases with linguistic variability. Unlike GLUE and similar ones, CheckList evaluates linguistic capabilities independently of the NLP task *e.g."sentiment analysis or text classification"*, which allows better predicting the performance of models against data from a domain other than the one they were trained on. The CheckList analysis found that many important commercial NPL products are unable to detect ontological inconsistencies in their own responses and fail to answer questions containing co-references to previously occurring phrases, being close to 100% when a negation is found at the end of the sentence *e.g."I thought the flight would be horrible, but it wasn't"*.

To evaluate extensions and adaptations of VQA models aimed at people with visual impairments, this work will study and develop a vision and language oriented check list based on VALSE (Parcalabescu et al., 2021). This is a novel benchmark designed to test visual-linguistic capabilities on pretrained general-purpose language and vision models. In particular, we will focus on one of its six tests called *existence*. This will be potentially useful to identify *unanswerable* questions, either due to lack of information contained in the image, or because the image was poorly focused, Figure 2.

3 Towards the implementation of VQA model based on more than one image

VQA models assume a high-quality, well-framed photograph to answer a certain question. A visually impaired person often has difficulty focusing photographs on the regions that actually contain the answer. Building on the work of (Bansal et al., 2020) and (Qiu et al., 2020), we plan to explore and adapt different existing VQA architectures to start studying the problem of answering questions using multiple views. Questions such as *how will the user be prompted to provide an additional view of the image* or *how to combine several views*, they will be addressed as the development of the work progresses. We will seek to improve the responses returned by conventional systems, taking advantage of the potential of the training sets already available in the state of the art, such as VQA v2.0 (Goyal et al., 2019), Visual Genome (Krishna et al., 2016) and VizWiz-VQA. We will adapt such sets to our needs, avoiding the costly and time-consuming task of building a new and own training set. In this way, we will also reduce the biases of each dataset, with the incorporation of new knowledge, coming from the extra information provided by the different points of view used.

As a result we expect more robust models for two reasons. First, they increase the possibilities of contextualizing the question when an image does not contain the necessary information. Second, multiple views provide more spatial information about the objects in the image, possibly allowing more precise answers to reference questions.

(a) Is there any writing on this medicine bottle?.

(b) Is there a light in the room?.

Figure 2: Examples of *unanswerable* visual questions.

4 Answering visual questions that consider the conversational history

People naturally ask questions that retrieve information from what we said before in the conversation; this is known as the conversational history. For example, consider an image where there is a group of people waiting for the bus and others passing through the street in front. The following conversation occurs. ***Q1:****"Are there people at the stop?"* ***A1:****"Yes"*. ***Q2:****"How many people are there?"*. VQA models that do not consider history will respond to Q2 by counting all the people when only the ones at the stop are relevant.

In this work, we plan to analyze the type of conversation history dependency present in visual dialog datasets such as VisDial (Das et al., 2016) and GuessWhat?! (de Vries et al., 2017), and devise methods to classify them according to the type of ellipses found. Also, it is planned to extend VQA systems to try to integrate this story effectively (Agarwal et al., 2020). Finally, using works such as (Mazuecos et al., 2021) as a starting platform, we will start by solving limited domains of questions belonging to the VizWiz-VQA specialized dataset for people with visual disabilities. Initially, questions with binary answers (yes and no) will be addressed, and the study will be expanded incrementally to solve the full spectrum.

5 Study integration of Optical Character Recognition (OCR) models

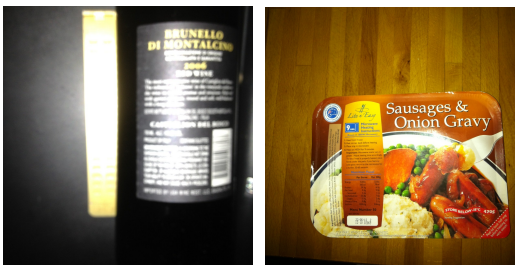
Despite the results shown in (Bigham et al., 2010), which shown that approximately 21% of the questions asked by visually impaired people necessarily involve reading or understanding the text included in the images captured from the environment, most of the current VQA systems are not prepared to be able to carry out this task. Although there are works such as (Kant et al., 2020), (Gao et al., 2020) and (Gao et al., 2021) that address the construction of VQA systems capable of reading, many focus on solving certain types of tasks and their performance are evaluated on generated sets such as TextVQA (Singh et al., 2019) and ST-VQA (Biten et al., 2019), where the images do not present quality or associated problems. Therefore, questions like *"is this product expired?" or "Can you tell me what temperature the oven is set to?"* remains an unsolved challenge. Currently, given the high accuracy offered by systems such as Keras-OCR, Tesseract and EasyOCR , we began by analyzing their performance on VizWiz-VQA samples that required reading or visual reasoning about the text contained in the image. To do this, a test pipeline was designed, where the results produced by each OCR system fed a pretrained question answering model based on the context (CoQA).

Based on the results and qualitative analysis carried out, it was identified that in many cases the incorporation of automatic systems to recognize text in the images allowed obtaining more precise answers, even more precise than those answered by humans. Figures 3(a)[3] and 3(b)[4] shows examples from the VizWiz dataset where EasyOCR was able to predict a correct answer, while ∼50% of the registered human annotations failed to do so.

6 Conclusion

Throughout this article, different lines of work and research are proposed to adapt existing VQA models to the particularities of visually impaired and blind people. The incorporation of reading capabilities and understanding of text in scenes in conventional VQA models, favor the independence of blind users, often subject to the wishes of sighted people to assist them in unfamiliar environments. The possibility of obtaining answers based on more than one photograph allows, on one hand, to reduce the complexity of the set of instructions for use and the design of the application's user interfaces, and on the other hand, it provides greater flexibility to the blind user when performs the capture of the image. Finally, a system capable of understanding incremental, conversational and spoken questions on multiple images generates a more organic and natural experience of human-machine communication not only for visually impaired people but also for those sighted.

[3]Details in https://tinyurl.com/3yvcbudp
[4]Details in https://tinyurl.com/ymhvcrmk



(a) What is this?. (b) For how long do i cook this in the microwave?.

Figure 3: Examples where the *ocr* model responds better than most people.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8182–8197, Online. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. CoRR, abs/1505.00468.

Ankan Bansal, Yuting Zhang, and Rama Chellappa. 2020. Visual question answering on image sets. CoRR, abs/2008.11976.

Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. Vizwiz: nearly real-time answers to visual questions. In Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, October 3-6, 2010, pages 333–342. ACM.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. CoRR, abs/1905.13648.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. 2016. Counting everyday objects in everyday scenes. CoRR, abs/1604.03505.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. CoRR, abs/1611.08669.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 4466–4475. IEEE Computer Society.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2014. From captions to visual concepts and back. CoRR, abs/1411.4952.

Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton van den Hengel, and Qi Wu. 2021. Structured multimodal attentions for textvqa.

Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text. CoRR, abs/2003.13962.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. Int. J. Comput. Vis., 127(4):398–414.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society.

Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. 2022. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference.

Yash Kant, Dhruv Batra, Peter Anderson, Alexander G. Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. 2020. Spatially aware multimodal transformers for textvqa. CoRR, abs/2007.12146.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. CoRR, abs/1602.07332.

Mauricio Mazuecos, Franco M. Luque, Jorge Sánchez, Hernán Maina, Thomas Vadora, and Luciana Benotti. 2021. Region under Discussion for visual dialog. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4745–4759, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. CoRR, abs/2112.07566.

Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. 2020. Multi-view visual question

answering with active viewpoint selection. Sensors, 20(8):2281.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with checklist. CoRR, abs/2005.04118.

Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. 2020. A survey on semi-, self- and unsupervised techniques in image classification. CoRR, abs/2002.08721.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. CoRR, abs/1904.08920.

Alexander Trott, Caiming Xiong, and Richard Socher. 2017. Interpretable counting for visual question answering. CoRR, abs/1712.08697.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. CoRR, abs/1804.07461.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Qiang Yu, Xinyu Xiao, Chunxia Zhang, Lifei Song, and Chunhong Pan. 2021. Extracting effective image attributes with refined universal detection. Sensors, 21(1).

Delu Zeng, Minyu Liao, Mohammad Tavakolian, Yulan Guo, Bolei Zhou, Dewen Hu, Matti Pietikäinen, and Li Liu. 2021. Deep learning for scene classification: A survey. CoRR, abs/2101.10531.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2015. Yin and yang: Balancing and answering binary visual questions. CoRR, abs/1511.05099.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. CoRR, abs/2201.02605.