# Visual Question Answering in a Toy World

Spring 2023

An old saying goes "1 image worths 1000 words". In matter of facts, so many questions can be answered with the information provided by one single image. This task, answering questions from images, is usually referred as Visual Question Answering and has been very popular in computer vision during the last years. Given an image, VQA systems try to answer an input question. This problem can be formulated as a multiple-choice problem or an open-ended problem.

Since the publication of the VQA dataset [1] , many models have been developed to improve the quality of the answer and provide more flexibility. Although many advances have been made in the field, it is a hard problem with so many incontrolable variables. For this reason, we propose a reduced version of the problem, instead of doing VQA on natural images, we want to perform the task in a toy world where the questions and answer refers to elements in this toy world. To further simplify the problem, we want the questions to be multiple-choice instead of open-ended. Luckily, [1] contains a subset of the dataset with this properties. We show some examples of abstract scenes and pairs of questions/answers in figure 1.

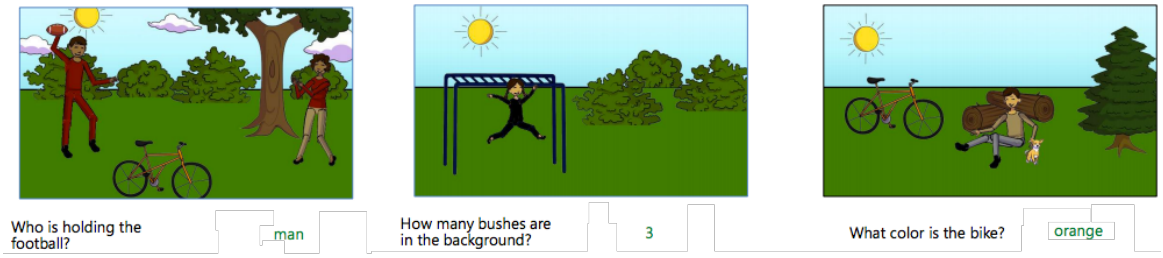


Figure 1: Examples of questions and answers for abstract scenes in [1].

In this project, we want to develop a model to take part in the VQA-Challenge [1] for abstract scenes in multiple-choice answer. Some other models tried to approach the same problem [2, 3]; we do not aim for the best performance in the challenge, we want the project to be an exploratory project where an original model is proposed, tested in the dataset and the results are reported.

## References


[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[2] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," *ArXiv preprint arXiv:1511.05099*, 2015.

[3] D. Teney, L. Liu, and A. v. d. Hengel, "Graph-structured representations for visual question answering," *ArXiv preprint arXiv:1609.05600*, 2016.


[1] http://visualqa.org/challenge.html