A dataset and architecture for visual reasoning with a working memory

Guangyu Robert Yang (robert.yang@columbia.edu)
Columbia University (Work done as an intern at Google Brain, equal contribution)

Igor Ganichev (iga@google.com)
Google Brain (equal contribution)

Xiao-Jing Wang (xjwang@nyu.edu)
New York University

Jonathon Shlens, David Sussillo ({shlens,sussillo}@google.com)
Google Brain

Abstract

A vexing problem in artificial intelligence is reasoning about events that occur in complex, changing visual stimuli such as in video analysis or game play. Inspired by a rich tradition of visual reasoning and memory in cognitive psychology and neuroscience, we developed an artificial, configurable visual question and answer dataset (COG) to parallel experiments in humans and animals. COG is much simpler than the general problem of video analysis, yet it addresses many of the problems relating to visual and logical reasoning and memory – problems that remain challenging for modern deep learning architectures. We additionally propose a deep learning architecture that performs competitively on other diagnostic VQA datasets (i.e. CLEVR) as well as easy settings of the COG dataset. However, several settings of COG result in datasets that are progressively more challenging to learn. After training, the network can zero-shot generalize to many new tasks. Preliminary analyses of the network architectures trained on COG demonstrate that the network accomplishes the task in a manner interpretable to humans.

Keywords: Visual reasoning; visual question answering; recurrent network; working memory

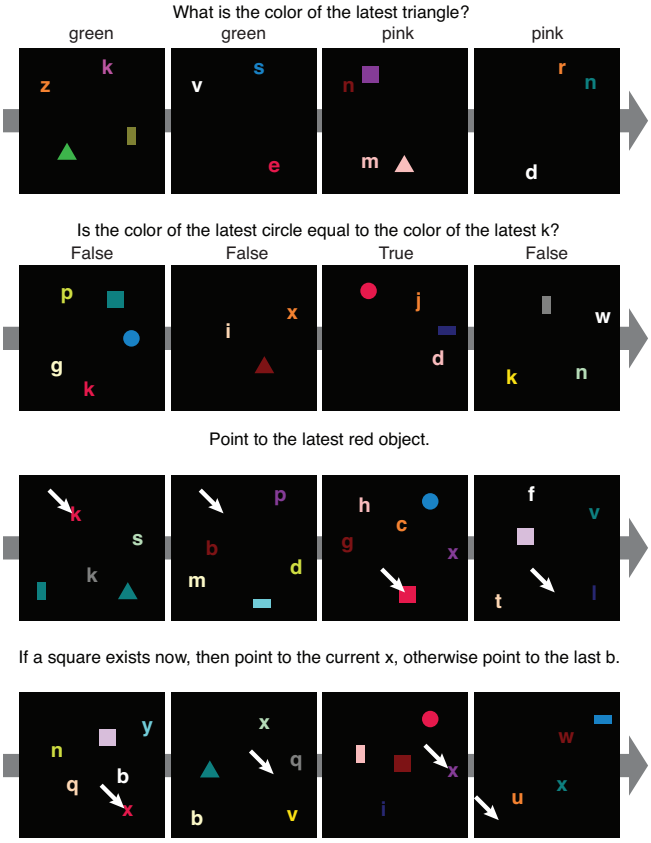

Figure 1: Sample sequence of images and instruction from the COG dataset. Tasks in the COG dataset test aspects of object recognition, relational understanding and the manipulation and adaptation of memory to address a problem. Each task can involve objects shown in the current image and in previous images. Note that in the final example, the instruction involves the *last* instead of the *latest* "b". The former excludes the current "b" in the image. Target pointing response for each image is shown (white arrow). High-resolution image and proper English are used for clarity.

[1] A longer version of this work has been submitted to the peer-reviewed conference ECCV (European Conference on Computer Vision). We believe this work should also be of interests to the cognitive science and neuroscience community.