A Biologically Inspired Scene based Question Answering Agent

Vidhya Navalpakkam and Laurent Itti
University of Southern California, Computer Science Department

The explosive growth of information readily available through modern technology has lead to a heightened interest in automated question answering (QA) agents. Our aim is to build a biologically-inspired scene-based QA agent, and our motivation stems from two main factors. While several document-based QA systems such as Askjeeves or Geoworld exist, image-based systems are almost non existent. Current QA systems sequentially process all available data, and subsequently rank them based on some relevance metric. There is little or no interaction between the two stages. Such scheme proves futile in image-based QA, where object recognition is computationally expensive. This calls for new approaches to the problem of QA. We derive inspiration from eye tracking experiments that show a profound influence of task demands on human eye movements (Yarbus, 1967). These complement recent experiments showing high correlation between eye movements and saliency of objects in the scene (Parkhurst *et al.*, 2002; Itti & Koch, 2001). Eye fixations on an object thus appear to depends on both the semantic content or top-down relevance of the object to the task, and the bottom-up salience of the object. Building on our previous modeling of bottom-up salience, we here propose a biologically-inspired agent for the deployment of visual attention onto visual scenes, based not only on bottom-up (image-based) but also on top-down (task-based) constraints.

Our QA agent takes a video scene and question as input and generates answers with the help of a knowledge base. Our knowledge base is an ontology that is a directed graph showing relations between concepts that describe real-world entities or abstract entities. We define simple relations such as "is a", "part of", "similar", "related", "optional member" and "inherits". In order to keep track of increasing or decreasing granularity, we define complementary relations such as "includes", "is a"; "contains", "part of". Human QA abilities range from simple sensory processing to complex cognitive processing of stimuli. As human cognition is not fully understood, we confine ourselves to simpler questions of the type "who is doing what to whom?". Accordingly, our achitecture consists of 3 ontologies namely, the "object"(O), "subject" (S) and "action" (A) with "related" relations running across them. Due to the limitations of the traditional ontologies, we adopt a Bayesian probabilistic approach and associate two probabilistic measures to each of the edges denoting relations: First, a measure of the probability of co-ocurrence of the two entities linked by an edge; second, a measure of increasing or decreasing granularity as we proceed from parent to child entities. These edge weights play a crucial role in determining the relevance of entities to a given task. For instance, while looking for the "hand", a "leaf" is considered to be irrelevant since the probablity of co-occurrence of "hand" and "leaf" is low. While looking for "hand", "man" is not considered as relevant as "finger" since the former has lower granularity than the latter.

Our QA agent performs two main functions- it determines salient features in the scene and computes the relevance of those features based on the task description. To mimic the human bottom-up salience extraction processes, we use our saliency-based attention model (Itti & Koch, 2001), which computes a saliency map (SM) from the scene. In the absence of top-down influences, attention would simply scan the SM in order of decreasing saliency. To include top down influences, we first perform localized object recognition at the attended location (for the moment, in canned form, while existing automated algorithms are being evaluated in our laboratory). The object identity is linked to our OSA ontology and top-down influences are then derived from the contents of a "temporary working memory" (TWM) that predicts relevant entities from the task description as well as from entities attended to and recognized so far. The output of the TWM is a "task relevance map" (TRM) that contains regions determined to be task-relevant. Finally, a point-by-point product of the SM and TRM forms the "attentional guidance map" (AGM) that guides attention. Thus, to be attended to, an object both has to exhibit some bottom-up salience, and some relevance to the task. Initially, the task relevance map is empty, and attention is controlled by bottom-up salience. As an entity is attended to and recognized, its relevance is evaluated in the TWM. If it is relevant, more attention is paid to it and a blob is activated in the TRM around it. Objects within the blob are recognized and their relevance is computed. Based on their relevance and salience, the AGM guides the subsequent shift of attention. Our system thus aims at isolating from an incoming video input those locations and objects that are relevant to the task, and of giving those objects priority for future attentional deployment.

We aim to use our model for real time QA based on visual input. Robots can be trained to take certain actions based on the answers to questions. Thus our model can be used even for task-specific operations such as navigation by following traffic signals.

References

Itti, L., & Koch, C. 2001. Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, **2**(3), 194–203.

Parkhurst, D., Law, K., & Niebur, E. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Res*, **42**(1), 107–123.

Yarbus, A. 1967. *Eye Movements and Vision.* New York: Plenum Press.