Title

Explainable QUestion Answering System (EQUAS)

Team Composition

Raytheon BBN Technologies (Raytheon BBN), William Ferguson, Principal Investigator, wferguson@bbn.com
Georgia Tech Research Corporation, Dhruv Batra
University of Texas Austin, Raymond Mooney
Massachusetts Institute of Technology, Antonio Torralba

Problem Statement

Raytheon BBN's challenge problem is the GuessWhich game where a human user/player sees a small number of images, and the AI system has been told that a randomly selected image is the right one. The user asks questions and the visual question answering (VQA) system answers them as best it can with regards to the secretly chosen image. The user then guesses which image the computer is talking about. The game is run with and without explanations provided for each answer and the user's guess accuracy is measured in each case.

Approach and Technique

The design of the EQUAS system draws heavily on pedagogical and argumentation theories to provide principles governing what elements are important for explanation, what the explanation space should look like, and what kinds of activities a user will want to perform while interacting with the system. When this is achieved, the user will have a "lucid interaction" with the system and know intuitively how to do what they want to do next.

Argumentation theory suggests that an act of explanation or argument can (and arguably should) be supported by analytics (explicit statements), prior similar cases, and visualization. Additionally, preferences against alternative choices (i.e., contrastive arguments) must be explicitly presented. Likewise, from the theory and practice of interactive pedagogy, learners need: didactic instruction (i.e., explicit presentation of knowledge); examples (i.e., prior supporting cases); problems to work through (i.e., relevant subsets of data that allow the user to construct their own model of decision-making); and counterexamples that refute common misconceptions. Argumentation and pedagogical theory align to provide a solid foundation to define the EQUAS "explanation space": analytics, visualization, cases, and rejected alternatives.

EQUAS explanations show:

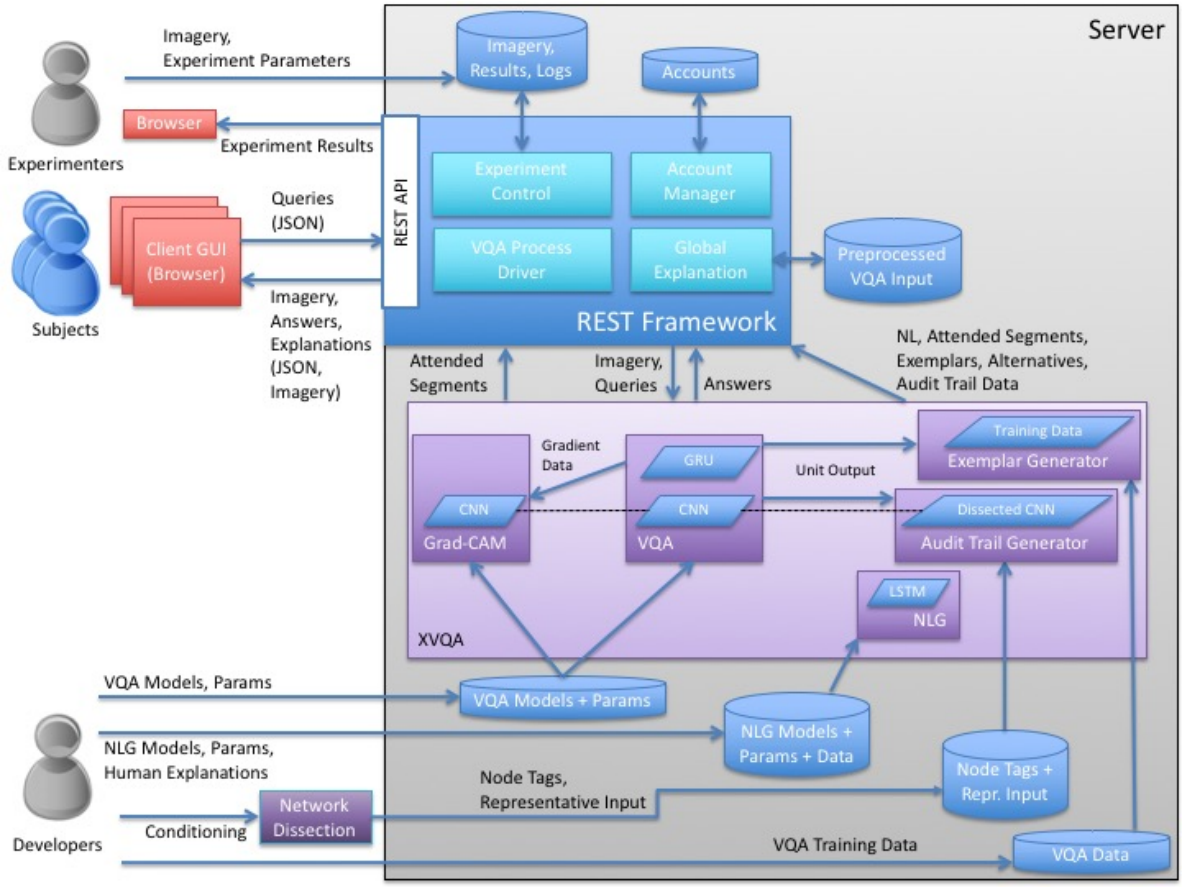
1. What recognized components support its choice;
2. What part of the data it found salient;
3. What known cases exemplify its choice; and
4. Why it rejected close alternatives.

In Phase 1, Raytheon BBN tested natural language explanations and "heat map" explanations (where the regions of the image that the system paid attention to are highlighted).

The following is the current architecture for the EQUAS system. (Note: evaluation of explanations is conducted with simpler interfaces designed for focused on-line experiments.)

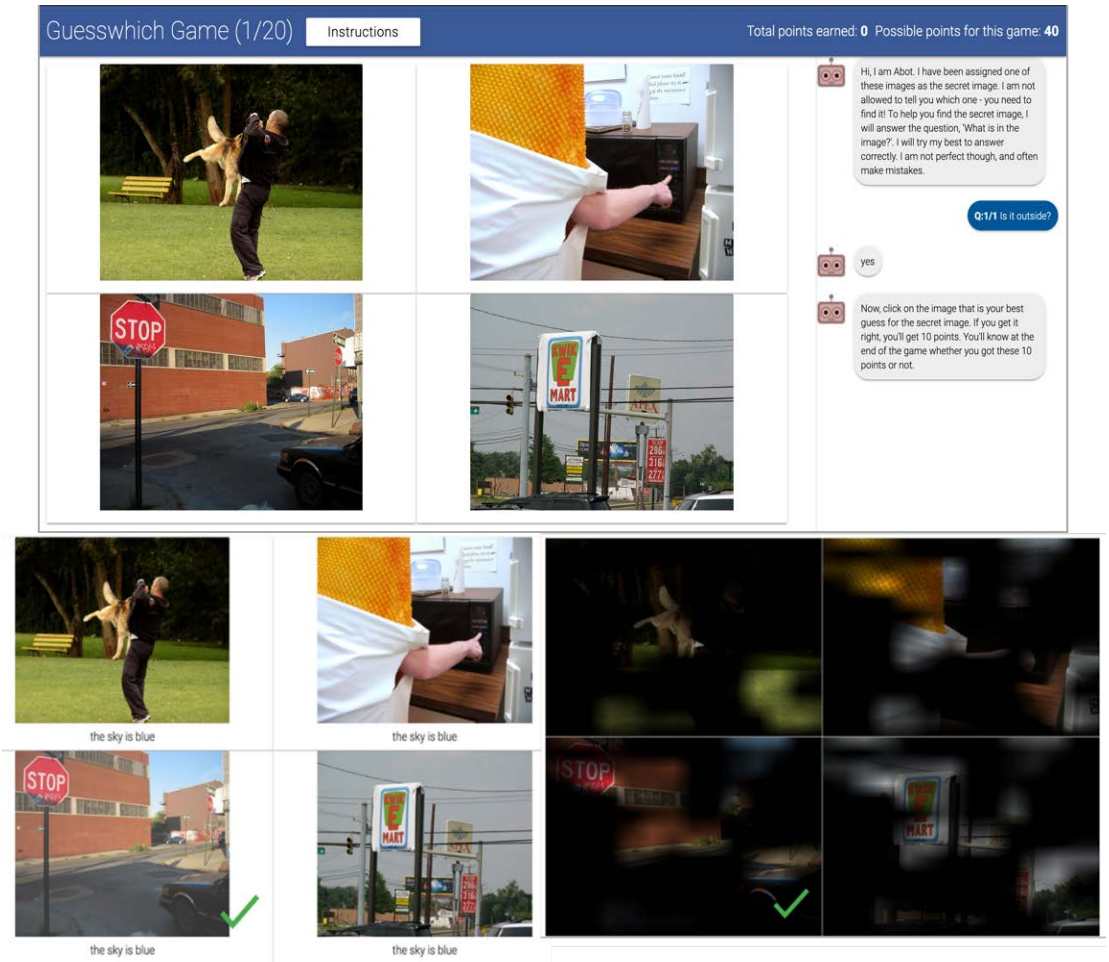
Approved for Public Release, Distribution Unlimited

Figure 1. EQUAS architecture



Below in Figure 2 are some example screenshots from the EQUAS GuessWhich studies. The top figure shows the experimental interface with no explanations. The subject has asked the question “Is it outside” and he the system has answered yes. The chosen image, the one that the system providing answers and sometimes explanation for, is the one in the bottom right quadrant of each screenshot. The bottom right image shows text explanations (the explanation has to associated with all of the images even though it is only correct for one of them, so as not to give the answer away) and the bottom left show heat map explanations.

Figure 2. EQUAS example screenshots from GuessWhich studies. (Top) Interface with no explanations; (Bottom) Interface with explanations; (Bottom left) Heat Map explanations; (Bottom right) Text explanations



Key Objectives and Current Status

While the EQUAS project has only evaluated partial explanation systems during Phase 1, by the Phase 2 the intent is to produce a unified, graphical user interface to a VQA system. Currently, there is a prototype system that can display text and heat map explanations; the prototype also allows the user to explore a large database that includes all of the images, questions, and answers on which the underlying VQA system was trained. The user interface to this database constitutes EQUAS's global explanation capability. During Phase 2, the team plans to evaluate this interface and apply it to a dataset that is an analogue to Department of Defense-relevant domains (such as overhead imagery).

Table 1 below identifies the EQUAS measures, objectives, and actuals to date.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
User/machine task performance	Improve user/machine performance at the GuessWhich task	Within subjects results show no significant effect for heat map explanations, but significant effect for text explanations
Mental model understanding	Improve user ability to predict the VQA system's answers to questions	Heat map explanations produced no effect in general, but positive effect

		when used with single questions, e.g., "*What animal is this?*"
Explanation satisfaction	Provide heat map explanations that are satisfying to the user	Satisfaction was slightly above neutral, not a very positive result
Multi-modal explanation satisfaction	Increase user satisfaction for text explanations by biasing them to refer to aspects of the image that were recognized by the system	Such explanations are statistically significantly more satisfying
Explanation goodness: faithfulness	Correlation between heat map and an "optimal" occlusion map in which pixels, when turned to noise, most alter the VQA answer	High correlation (see below); not surprising since both are computational measures of salience
Explanation goodness: correctness	Correlation between heat map explanation user "heat map" (from de-blurring experiments)	Some correlation; perfect correlations would be unlikely since people have to make guesses about where to de-blur
Appropriate trust	Use explanations to enhance a user's ability to judge which of two VQA systems was more competent, thus enabling appropriate trust	Heat map explanations provide substantial ability for users to predict relative VQA competence

	Numerical Value		**Hypothesis Results**	**User Machine Task**
	With Explanation	**Without Explanation**		
User-Machine Task Performance: heat map explanations	36%	35%	Not significant	GuessWhich
User-Machine Task Performance: text explanations	45%	31%	P < 0.05	GuessWhich

	Numerical Value		**Hypothesis Results**	**Mental Model Task**
	With Explanation	**Without Explanation**		
Mental Model Understanding	66	67	Not significant	Prediction correctness
Mental Model Understanding: single pre-determined question	80	67	P < 0.05	Prediction correctness

	Numerical Value		**Hypothesis Results**	**Satisfaction Measure**
	With Explanation	**Without Explanation**		
Explanation Satisfaction	3.2 – 3.5	Did not test	Significant, p < 0.05	Likert scale, range of responses

	Numerical Value		**Hypothesis Results**	**Goodness Measure**
	With Explanation	**Without Explanation**		
Explanation Faithfulness	0.60	N/A	P < 0.05	Faithfulness
Explanation Correctness	0.14	N/A	P < 0.05	Correctness

	Numerical Value		**Hypothesis Results**	**Trust Measure**
	With Explanation	**Without Explanation**		
Appropriate Trust/ Reliance	e.g., 1.3	1		e.g., 0 – 2 (0 is bad, 2 is good)
Appropriate Trust/ Reliance	1.3	0	Significant	0 random; 2 perfect prediction

Test and Evaluation

Phase 1 results suggest that text explanations produce significant user/machine collaboration benefits, whereas heat maps alone may not. The team has also demonstrated that heat map explanations agree

with what areas users think are relevant, and that they can help users determine which AI systems are superior in general form single examples. EQUAS text explanations were rated as relevant and useful.

Jialin Wu, Raymond J Mooney. *Faithful Multimodal Explanation for Visual Question Answering.* arXiv:1809.02805.

A. Das, H. Agrawal, C.L. Zitnick, D. Parikh, D. Batra. *Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?* Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.* arXiv:1610.02391.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, Devi Parikh. *Do Explanations Make VQA Models More Predictable To A Human?* Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

For additional information:

http://bit.ly/ganpaint (a generative adversarial network paint demonstration using named components (aspects of an image generating system) to control a created image – a painting program that works with concepts instead of paint).

Approved for Public Release, Distribution Unlimited