Reframing Explanation as an Interactive Medium: The EQUAS (Explainable QUestion Answering System) Project

Dhruv Batra [1], William Ferguson[2], Raymond Mooney[3], Devi Parikh[1], Antonio Torralba [4], David Bau[4], David Diller[2], Joshua Fasching[2], Jaden Fiotto-Kaufman[2], Yash Goyal [1], Jeff Miller[2], Kerry Moffitt[2], Alex Montes De Oca[2], Ramprasaath R. Selvaraju [1], Ayush Shrivastava [1], and Jialin Wu[3]

[1]Georgia Tech
[2]Raytheon BBN Technologies
[3]The University of Texas at Austin
[4]Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory

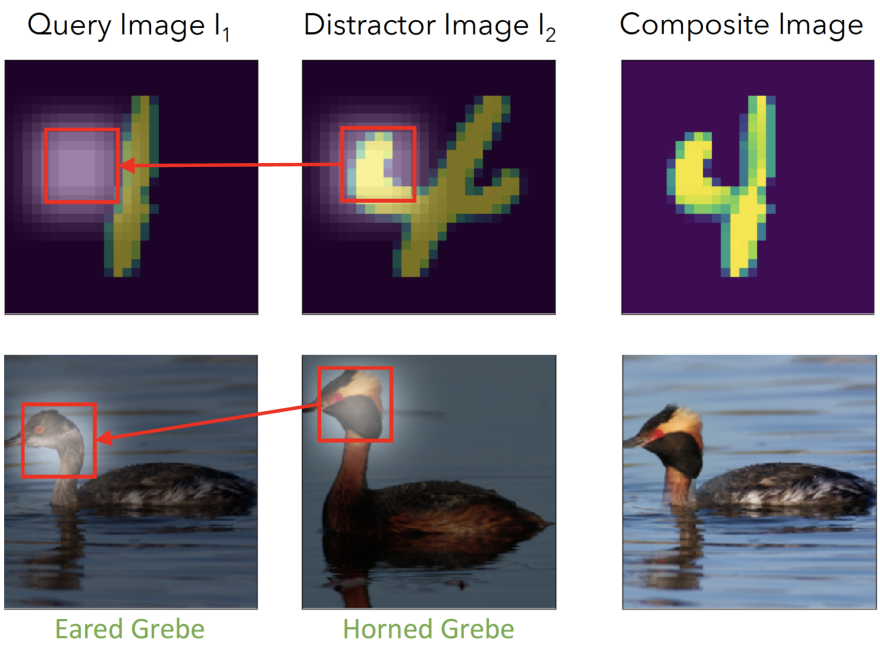
June 22, 2021

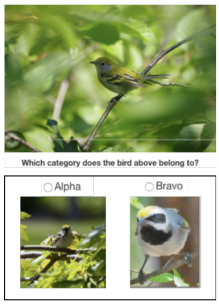
Abstract

This letter provides a retrospective analysis of our team's research performed under the DARPA Explainable Artificial Intelligence (XAI) project. We began by exploring salience maps, English sentences, and lists of feature names for explaining the behavior of deep-learning-based discriminative systems, especially visual question answering systems. We demonstrated limited positive effects from statically presenting explanations along with system answers – for example when teaching people to identify bird species. Many XAI performers were getting better results when users interacted with explanations. This motivated us to evolve the notion of explanation as an interactive medium – usually, between humans and AI systems but sometimes within the software system. We realized that interacting via explanations could enable people to task and adapt ML agents. We added affordances for editing explanations and modified the ML system to act in accordance with the edits to produce an interpretable interface to the agent. Through this interface, editing an explanation can adapt a system's performance to new, modified purposes. This deep tasking, wherein the agent knows its objective and the explanation for that objective will be critical to enable higher levels of autonomy.

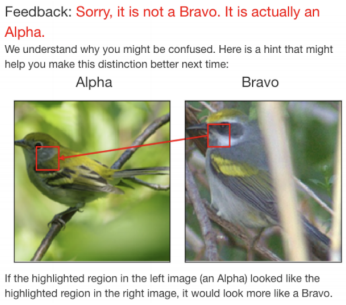
Hosted file

`Final_EQUAS paper.pdf` available at https://authorea.com/users/421190/articles/527320-reframing-explanation-as-an-interactive-medium-the-equas-explainable-question-answering-system-project

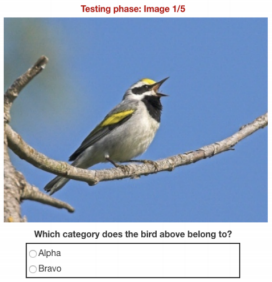




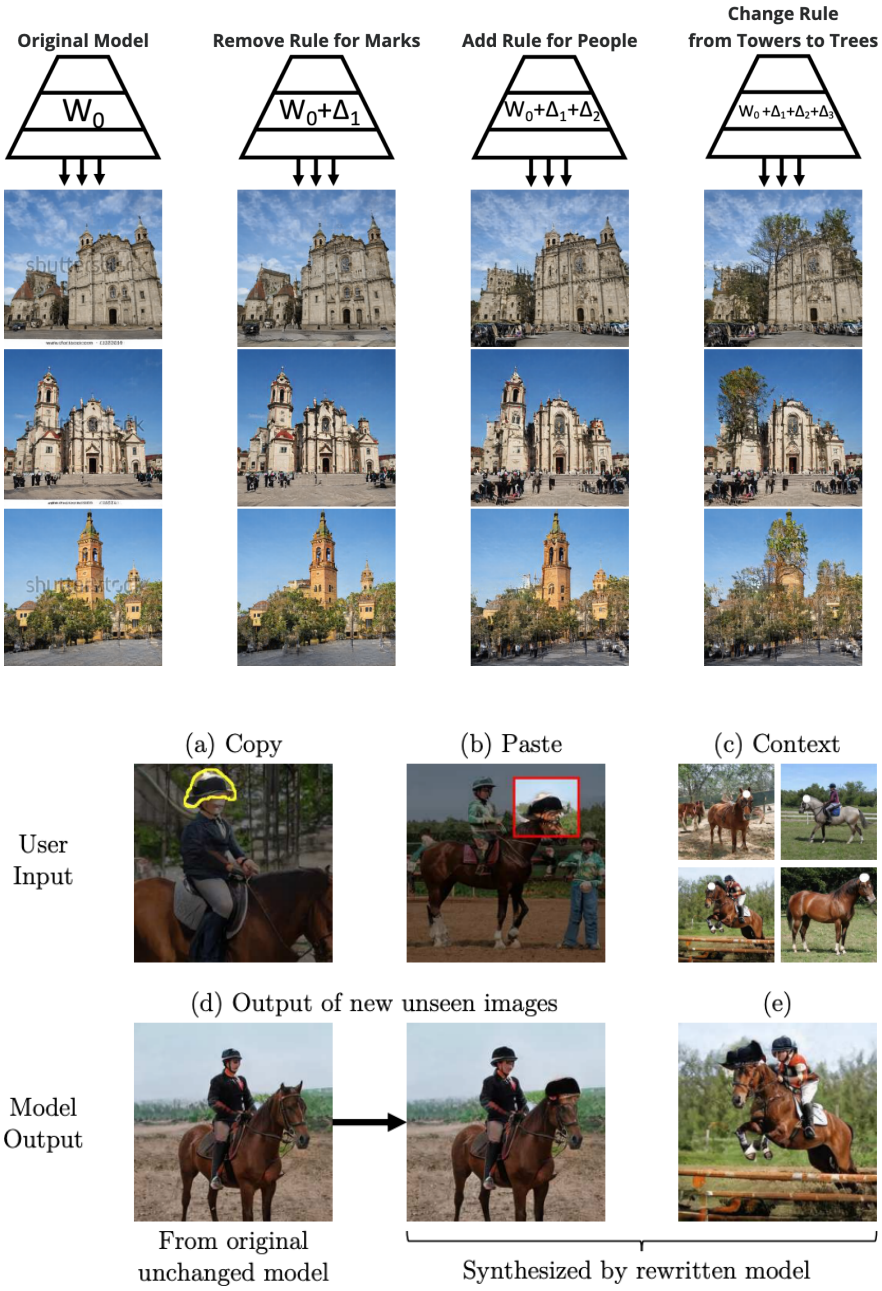
(a) Training Interface



(b) Feedback



(c) Testing Interface



Question: Is this in an Asian country?
Human Explanation: The information provided on the train's marquee is comprised of Asian characters.

Candidate 1: No VQA confidence: 0.88
Sample Retrieved Explanations:
1. The train looks European as well as the railings and surrounding area.
2. The wording on the train is in English.
3. 4.... 8...
Verification score: 0.17
Final Confidence: 0.15

Candidate 2: Yes VQA Confidence: 0.79
Sample Retrieved Explanations:
1. It does not look like a standard American train.
2. The signs are all in Japanese.
3. 4.... 8...
Verification score: 0.97
Final confidence: 0.77