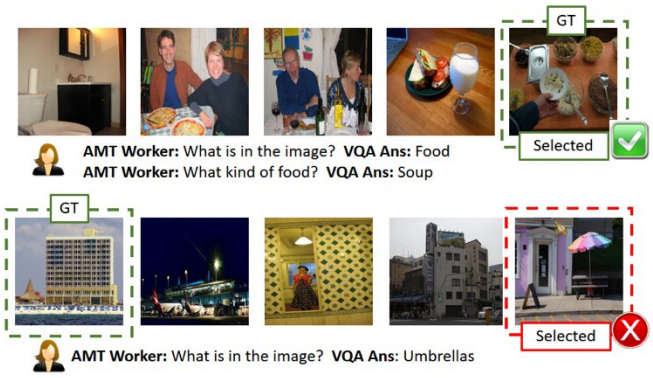
ExAG: An image-guessing game to evaluate the helpfulness of machine explanations

19 April 2019, by Ingrid Fadelli



The figure above shows two game plays without explanations (each row is a game-play example). As shown in the top row, the user (i.e., AMT Worker since we have used Amazon Mechanical Turk for large-scale crowd-source-based evaluation) is able to figure out the secret image correctly given accurate AI answers. However, as shown in the bottom row, the user may fail even when AI answers are reasonable but slightly off-point (i.e., there is an umbrella-looking structure beside the building although it is not the focus of the ground-truth image). GT means Ground Truth (i.e. the secret image), Selected is the image selected by the user after asking the questions and getting the answers and/or explanations. Credit: Ray et al.

In recent years, researchers have been trying to make artificial intelligence (AI) more transparent by developing algorithms that can explain their actions and behavior, as this could encourage greater trust in machines and enhance human-AI interactions. Despite their efforts, so far very few studies have tangibly evaluated the impact of AI explanations on the performance achieved in tasks that involve human-AI collaboration.

To address this gap in the existing literature, a team of researchers at SRI International has created a human-AI image guessing game inspired by the popular game 20 Questions (20Q), which can be used to evaluate the helpfulness of machine explanations. Their paper, recently published on arXiv, is among the first to explore the effects of developing more 'explainable' AI.

"The idea came about while we were working on a DARPA project," Arijit Ray, a computer scientist at SRI International who carried out the study, told TechXplore. "In this project, we are developing explainable AI systems, which not only generate the desired output (e.g. object detection, answers to questions, etc.) but also explanations of how they arrived at that output. We needed a mechanism to evaluate whether the additional explanations provided by AIs were useful for the user to gain a better understanding of the AI systems. To this end, we created an interactive human-AI collaborative task, Explanation-assisted GuessWhich (ExAG), which is an adaptation of the famous 20Q game, to demonstrate the effectiveness of the various machine explanation techniques that we are developing."

The image-guessing game developed by Ray and his colleagues closely resembles the popular game 20 Questions, which usually involves two players. In 20Q, one player thinks about something and the second player tries to guess what it is by asking 20 closed-ended questions (i.e., questions that can only be answered with 'yes' or 'no').

In ExAG, the adaptation of the game devised by Ray and his colleagues, a user is shown five images, one of which has been chosen by the AI system as the 'secret image.' Essentially, the user needs to figure out which one among the pictures he/she saw is the 'secret image,' by asking natural language questions about it.

In contrast with the traditional 20Q game, in ExAG human users can ask both closed and open-ended questions. For instance, they could ask 'what is in the image?,' 'where was the image taken?' and so on. The AI system answers a user's questions one

at a time and can optionally explain its answers.

Based on these answers, the user will then try to guess the image that the AI had originally selected. The overall goal of the game is to correctly identify the 'secret image' by asking as few questions as possible.

"The AI system provides two modes of explanations, visual and textual," Ray explained. "For visual explanations, the AI system generates heat maps highlighting the regions that support its answers. For example, if a user asks what is in the image, and it looks like a dog, the AI will highlight the dog region and say this is what leads to the answer 'it's a dog.' For textual explanations, on the other hand, the AI system provides answers to related questions for each of the images. So, if you ask what a person is doing and the answer is surfing, for instance, it will also answer related questions like 'what do I see in the image? A surfer.' 'Where is the picture taken? A beach.'

Due to the nature of the image-guessing game, the quality of the answers and explanations provided by the AI can significantly affect a human user's success and performance. It is worth noting that current state-of-the-art performance in visual question answering is around 65 percent, which means that the AI system generates correct answers 65 percent of the time.

Ray and his colleagues observed that users typically succeeded in ExAG by taking advantage of the AIs explanations, especially when the answers themselves were wrong. For example, if the 'secret image' portrays a dog, but the AI answers 'it's a surfer,' a visual explanation could help a human user realise the AI's mistake. According to the researchers, this proves that their game is a suitable tool for evaluating the helpfulness of AI explanations.

The figure above shows a gameplay with explanations. The heatmap visual explanation highlights the regions in the images that lead to AI answers. With such an explanation, users gain understanding that AI systems may pick up objects that are not the main focus of the image in human's perspective when answering a general question like "what is in the image". This hints the user to ask follow-up questions and finally select the secret image correctly. GT means Ground Truth (i.e. the secret image), Selected is the image selected by the user after asking the questions and getting the answers and/or explanations. Credit: Ray et al.

"In my opinion, the most interesting result of our study is that users can use just a few good explanations to win games when the AI answers are mostly wrong," Ray said. "In contrast, for games with similar answer accuracy but without explanations, users blindly trust AI-generated answers and lose the game. This supports the importance of even a few good explanations for a human-AI collaborative systems, especially when the AI system is imperfect, which it is in most cases these days."

To explain this idea better, Ray offers the example of self-driving vehicles. Over the past few years, there has been much debate about their safety, also due to accidents that occurred while the vehicles were being tested. According to Ray, effective AI explanations could encourage greater trust in the safety of self-driving vehicles, as they would allow human drivers to identify problems

beforehand and prevent accidents.

"For instance, let's assume that the AI system is experiencing trouble detecting the lanes reliably," Ray said. "Since the road is currently straight, without additional information, the user would be unable to tell whether the AI is failing. Even if he/she had some doubts, he/she would probably not do anything until the last moment, when the car has to take a turn, doesn't, and crashes, which would too late. In contrast, if a screen in the car showed explanations of how AI is perceiving the environment, such as heat maps, the user would be able to tell the AI's latent failure and take control of the wheel in advance."

The researchers found that useful explanations positively affected human users' performance in the image-guessing game. Their findings suggest that having at least one 'correct' explanation was significantly helpful, particularly in cases where the AI's answers to user questions were 'noisy' or poorly defined. Interestingly, players developed a preference for explanations over answers and often rated the AI explanations as 'helpful.'

"I think that while several lines of work tried to provide explanations for an AI system's outcomes or actions, ours is the first study to introduce a human and machine collaboration task to evaluate the effectiveness of AI explanations; thus, it brought a lot of insight into how AI explanations could enhance human-robot interactions," Yi Yao, a senior technical manager at SRI International who was involved in the study, told TechXplore.

The study carried out by Ray and his colleagues is one of the first to provide tangible evidence of the usefulness of AI explanations. The researchers hope that their research will ultimately inform the development of AI systems that can act rationally in society, thus connecting and relating better with humans.

According to Ray, AI systems that can clearly explain the reasoning and processes behind their actions would be a significant step forward in the development of intelligent machines. By effectively answering questions and rationalizing their decisions, these systems could foster a greater sense of trust in AI, as well as a deeper relationship with it.

"Many other companies, groups and research groups have been addressing explainable AI, and there have been many proposals of how to extend existing AI models and systems to provide explanations to users," said Giedrius Burachas, a senior computer scientist at SRI International and principal investigator behind the DARPA study that led to the development of the Guess Which game. "While there were lots of ideas generated, evidence that these ideas work was lacking, so one of the strong sides of our research is that it provides indisputable evidence that certain types of explanations are really very effective in improving the collaboration with the AI systems, but also in building trust in them."

Up until now, the work of Ray and his colleagues primarily focused on visual question answering (VQA) tasks, where users ask questions about images and an AI answers. They are now planning to continue their research into AI explanation techniques, applying these techniques to a broader scope of AI tasks.

"We will also continue developing protocols to evaluate the effectiveness of the AI-generated explanations with finer granularity (e.g. what explanation is more effective under what scenarios?) and from different perspectives (e.g. do explanations help users to build the mental model?)," Ray said. "To close the loop, we will use lessons learned from these evaluations to develop more effective explanation methods. We believe that the Holy Grail of explainable AI is to devise explanations that not only inform users but also improve machine performance by improving its reasoning ability."

In addition to exploring the effects of AI explanations on the performance and perceptions of human users, therefore, the researchers would like to investigate their impact on the AI systems themselves. They feel that AI explanations could also make AI systems inherently better, as they would gradually acquire reasoning and rationalization skills.

More information: Lucid explanations help: using a human-AI image-guessing game to evaluate machine explanation helpfulness. arXiv:1904.03285 [cs.CY]. arxiv.org/abs/1904.03285