Explainable AI as Collaborative Task Solving

Arjun Akula[1], Changsong Liu[1], Sinisa Todorovic[2], Joyce Chai[3], Song-Chun Zhu[1]
University of California, Los Angeles[1], Oregon State University[2], Michigan State University[3]
aakula@ucla.edu[1], liucs81@gmail.com[1], sinisa@oregonstate.edu[2],
jchai@cse.msu.edu[3], sczhu@stat.ucla.edu[1]

Abstract

We present a new framework for explainable AI systems (XAI) aimed at increasing human trust in the system's performance through explanations. Based on the Theory of Mind, our framework X-ToM explicitly models machine's mind ($\mathbf{pg^{M}}$), human's mind as inferred by the machine ($\mathbf{pg^{UinM}}$), as well as machine's mind as inferred by the human ($\mathbf{pg^{MinU}}$). These mental representations are incorporated to (1) learn an optimal explanation policy that takes into account human's perception and beliefs; and (2) quantitatively evaluate human's trust of machine behaviors. We have applied X-ToM in the context of visual recognition. Compared to the most popularly used attribution based explanations (saliency maps), our X-ToM significantly improves human trust in the underlying vision system.

1. Introduction

How to increase human trust and reliance in AI systems is of great concern in a wide range of applications that depend on the machine's predictions and decisions. Previous studies have shown that trust is closely and positively correlated to the level of how much human users understand the system (**understandability**) and how accurately they can predict the system's performance on a given task (**predictability**) [4]. Despite an increasing amount of work on XAI [10, 8, 9], providing explanations that can increase understandability and predictability remains an important research problem. To address this problem, we propose **X-ToM**, a new explainable AI (XAI) framework based on the Theory-of-Mind [1]. The ability to reason about other's perception and beliefs, in addition to one's own perception and beliefs, is often referred to as the Theory-of-Mind (ToM). Our X-ToM framework, using ToM, provides explanations by taking into account the user's mind to increase understandability and predictability.

Cognitive studies [5] have shown an explanation can only be optimal if it is generated by taking user's perception and belief into account. As humans can easily be overwhelmed with too many or too detailed explanations, XAI systems need to understand the user and identify user-specific content for explanation. Explanation is also not one shot and often involves interaction between the human and the system. The context of such interaction plays an important role in determining follow-up explanations. Motivated by these findings, our X-ToM generates interactive and collaborative explanations by incorporating machine's understanding of human's mind; and evaluates human's trust in the machine by explicitly measuring human's understanding of machine's mind.

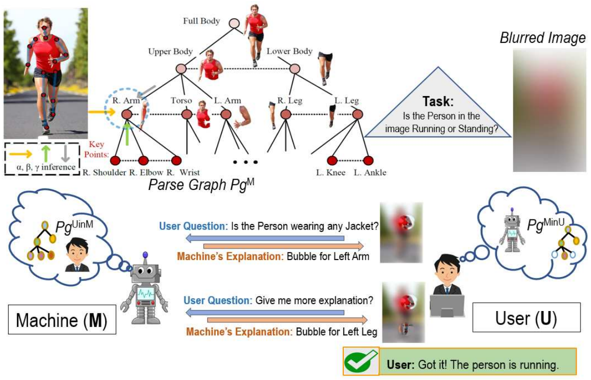
Full Body
Upper Body
Lower Body
R. Arm
Torso
L. Arm
R. Leg
L. Leg
Key Points:
α, β, γ inference
R. Shoulder
R. Elbow
R. Wrist
L. Knee
L. Ankle
Parse Graph Pg^{M}
Task:
Is the Person in the image Running or Standing?
Blurred Image
Pg^{UinM}
Pg^{MinU}
User Question: Is the Person wearing any Jacket?
Machine's Explanation: Bubble for Left Arm
User Question: Give me more explanation?
Machine's Explanation: Bubble for Left Leg
Machine (M)
User (U)
User: Got it! The person is running.

Figure 1. **XAI as Collaborative Task Solving**: Our interactive and collaborative XAI framework based on the Theory-of-Mind. The human user seeks visual explanations through a dialog, in the form of bubbles, from XAI agent for solving a given collaborative task.

As part of this framework, we have designed a new collaborative task-solving game for visual recognition. As illustrated in Figure 1, in our X-ToM game, the machine (M) and the user (U) are positioned to solve a collaborative task. The machine is given an original image and is supposed to detect and localize objects and parts of interest or a human activity appearing in the image. The user is given a blurred version of the original image, and the user seeks the machine's help essentially through the explanations generated by the machine in order to recognize objects/parts in the blurred image. This game provides a unique collaborative setting where the system is motivated to provide human-understandable explanation for its visual recognition and the user is motivated to seek the system's recognition and explanation to help his/her own understanding.

To facilitate this collaborative game, X-ToM explicitly models mental states of visual understanding ("minds") of the machine and user using parse graphs (pg) in the form of And-Or Graph (AOG). In a pg, nodes represent objects and parts detected in the image, and edges represent spatial relationships identified between the objects. X-ToM mind models include:

- $\mathbf{pg^M}$: the machine's own inference about objects and their locations in the image.
- $\mathbf{pg^{UinM}}$: the human's mind as inferred by the machine.
- $\mathbf{pg^{MinU}}$: the machine's mind as inferred by the human.

Our empirical results show that X-ToM allows the user to achieve a high success rate in visual recognition on blurred images. We also found that the most popularly used attribution based explanations (saliency maps) [8] are not effective to improve human trust in AI system, whereas our Theory-of-Mind inspired approach significantly improves human trust in AI by providing optimal explanations.

2. X-ToM Framework

Our X-ToM consists of three main components:

- A **Performer** that generates image interpretations (i.e., machine's mind represented as pg^M) using a set of computer vision algorithms;
- An **Explainer** that generates maximum utility explanations in a dialog with the user by accounting for pg^M and pg^{UinM} using reinforcement learning;
- An **Evaluator** that quantitatively evaluates the effect of explanations on the human's understanding of the machine's behaviors (i.e., pg^{MinU}) and measures human trust by comparing pg^{MinU} and pg^M.

2.1. X-ToM Game

An X-ToM game consists of two phases. The first phase is the collaborative task phase. The user is shown a blurred image and given a task to recognize what the image shows. X-ToM has access to the original (unblurred) image and the machine's (i.e. **Performer's**) inference result pg^M. The user is allowed to ask questions regarding objects and parts in the image that the user finds relevant for his/her own recognition task. Using the detected objects and parts in pg^M, X-ToM **Explainer** provides visual explanations to the user, as shown in Figure 1. This process allows the machine to infer what the user sees and iteratively update pg^{UinM}, and thus select an optimal explanation at every turn of the game.

The second phase is specifically designed for evaluating whether the explanation provided in the first phase helps the user understand the system behaviors. The **Evaluator** shows a set of original (unblurred) images to the user that are similar to (but different from) the ones used in the first phase of the game (i.e., the set of images shows the same class of objects or human activity). The user is then given a task to predict in each image the locations of objects and parts that would be detected by the machine (i.e., in pg^M) according to his/her understanding of the machine's behaviors. Based on the human predictions, the **Evaluator** estimates pg^{MinU} and quantifies human trust in the machine by comparing pg^{MinU} and pg^M.

2.2. X-ToM Explainer (for Explanation Generation)

The explainer, in the first phase of the game, makes the underlying α, β, and γ inference process of the performer more transparent to the human through a collaborative dialog. At one end, the explainer is provided access to an image and the performer's inference result pg^M on that image. At the other end, the human is presented a blurred version of the same image, and asked to recognize a body part, or pose, or human action depicted (e.g., whether the person is running or walking). To solve the task, the human may ask the explainer various "what", "where" and "how" questions (e.g., "Where is the left arm in the image"). We make the assumption that the human will always ask questions that are related to the task at hand so as to solve it efficiently. The explainer answers these questions using pg^M and justifies the answers by showing the corresponding visual explanations in the image.

As visual explanations, we use "bubbles", where each bubble reveals a circular part of the blurred image to the human. The bubbles coincide with relevant image parts for answering the question from the human, as inferred by the performer in pg^M. For example, a bubble may unblur the person's left leg in the blurred image, since that image part has been estimated in pg^M as relevant for recognizing the human action "running" occurring in the image. Following the "principle of least collaborative effort" and the aforementioned findings [5] that explanations should *not* overwhelm the human, our X-ToM explainer utilizes pg^M and pg^{UinM} (i.e., the contextual and hierarchical relationships explicitly modeled in the AOG) for controlling the depth and breadth of explanations. To enable this control, each bubble is characterized by a number of parameters, including the amount of image reveal (i.e., the unblurring level), size, and location in the image, to name a few. We use reinforcement learning to train the explainer to optimize these parameters and thus provide optimal visual explanations.

2.3. X-ToM Evaluator (for Trust Estimation)

The second phase of the X-ToM game serves to assess the effect of the explainer on the human's understanding of the performer. This assessment is conducted by the evaluator. The human is presented with a set of (unblurred) images that are different from those used in the first phase. For every image, the evaluator asks the human to predict the performer's output. The evaluator poses multiple-choice questions and the user clicks on one or more answers. Based on responses from the human, the evaluator estimates pg^{MinU}. By comparing pg^{MinU} with the actual machine's mind pg^{M} (generated by the performer), we have defined the following metrics to quantitatively assess human trust in the performer:

Justified Positive and Negative Trust: It is possible for humans to feel positive trust with respect to certain tasks, while feeling negative trust (i.e. mistrust) on some other tasks. The positive and negative trust can be a mixture of justified and unjustified trust [4]. We compute justified positive trust (JPT) and negative trust (JNT) as follows:

$$\text{JPT} = \frac{1}{N} \sum_{i} \sum_{z=\alpha,\beta,\gamma} \Delta\text{JPT}(i,z),$$

$$\Delta\text{JPT}(i,z) = \frac{\|pg^{MinU}_{i,z,+} \cap pg^{M}_{i,+}\|}{\|pg^{M}_{i,+}\|},$$

$$\text{JNT} = \frac{1}{N} \sum_{i} \sum_{z=\alpha,\beta,\gamma} \Delta\text{JNT}(i,z),$$

$$\Delta\text{JNT}(i,z) = \frac{\|pg^{MinU}_{i,z,-} \cap pg^{M}_{i,-}\|}{\|pg^{M}_{i,-}\|},$$

where N is the total number of games played. z is the type of inference process. $\Delta\text{JPT}(i,z)$, $\Delta\text{JNT}(i,z)$ denote the justified positive and negative trust gained in the i-th turn of a game on the z inference process respectively.

Reliance: Reliance (Rc) captures the extent to which a human can accurately predict the performer's inference results without over- or under-estimation.

3. Experiments

We conduct human subject experiments to assess the effectiveness of the X-ToM Explainer, that is trained on AMT, in increasing human trust through explanations. We recruited 120 human subjects from our institution's Psychology subject pool. These subjects have no background on computer vision, deep learning and NLP. We applied between-subject design and randomly assigned each subject into one of the three groups. One group used X-ToM Explainer, and two groups used the following two baselines respectively:

- $\Omega_{\textbf{QA}}$: we measure the gains in human trust only by revealing the answers for the tasks without providing any explanations to the human.
- $\Omega_{\textbf{Salience}}$: in addition to the answers, we also provide saliency maps generated using attribution techniques to the human as explanations.

Within each group, each subject will first go through an introduction phase where we introduce the tasks to the subjects. Next, they will go through familiarization phase where the subjects become familiar with the machine's underlying inference process (Performer), followed by a testing phase where we apply our trust metrics and assess their trust in the underlying Performer. Figure 2(a) compares the justified positive trust (JPT), justified negative trust (JPT), and Reliance (Rc) of X-ToM with the baselines. As we can see, JPT, JNT and Rc values of X-ToM are significantly higher than Ω_{QA} and Ω_{Salience} ($p < 0.01$). *Also, it should be noted that attribution techniques ($\Omega_{Salience}$) did not perform any better than the Ω_{QA} baseline where no explanations are provided to the user.* This could be attributed to the fact that, though saliency maps help human subjects in localizing the region in the image based on which the performer made a decision, they do not necessarily reflect the underlying inference mechanism. In contrast, X-ToM Explainer makes the underlying inference processes (α, β, γ) more explicit and transparent and also provides explanations tailored for individual user's perception and understanding. Therefore X-ToM leads to the significantly higher values of JPT, JNT and Rc.

3.1. Gain in Reliance over time

We hypothesized that, human trust and reliance in machine might improve over time. This is because, it can be harder for humans to fully understand the machine's underlying inference process in one single session. Therefore, we conduct an additional experiment with eight human subjects where the subjects' reliance is measured after every session. Note that each session consists of a familiarization phase followed by a testing phase. The results are shown in Figure 2(b). As we expected, subjects' reliance increased over time. Specifically, reliance with respect to α inference process significantly improved only after 2.5 sessions. Reliance with respect to β and γ inference processes significantly improved after 4.5 sessions.

4. Related Work

Most prior work has focused on generating explanations using feature visualization and attribution.

Feature visualization techniques typically identify qualitative interpretations of features used for making predictions or decisions. For example, gradient ascent optimization is used in the image space to visualize the hidden feature layers of unsupervised deep architectures [2]. Also, convolutional layers are visualized by reconstructing the input of each layer from its output [9]. Recent visual explanation

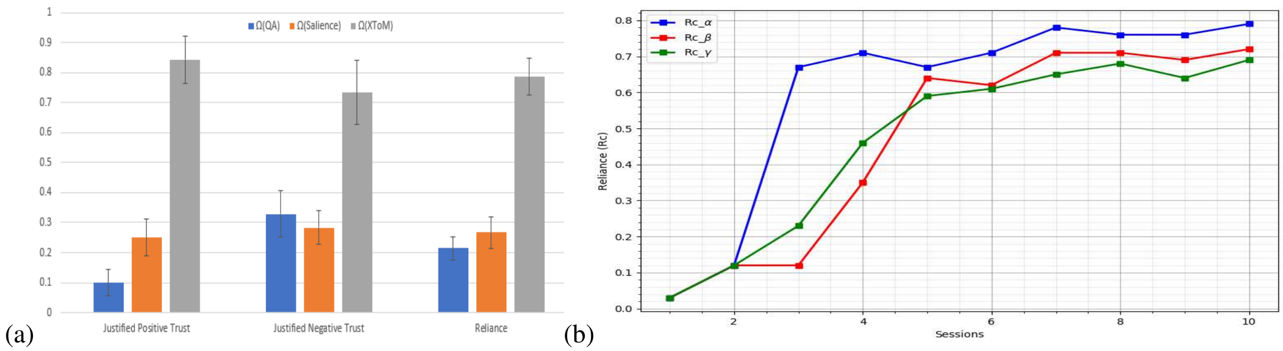

Figure 2. (a) Gain in Justified Positive Trust, Justified Negative Trust and Reliance: X-ToM vs baselines (QA, Saliency Maps). Error bars denote standard errors of the means. (b) Gain in Reliance over sessions w.r.t α, β and γ processes.

models seek to jointly classify the image and explain why the predicted class label is appropriate for the image [3].
Attribution is a set of techniques that highlight pixels of the input image (saliency maps) that most caused the output classification. Gradient-based visualization methods [11, 7] have been proposed to extract image regions responsible for the network output. The LIME method proposed by [6] explains predictions of any classifier by approximating it locally with an interpretable model.

5. Conclusion

This paper presents X-ToM – a new framework for Explainable AI (XAI) and human trust evaluation based on the Theory-of-Mind (ToM). X-ToM generates explanations in a dialog by explicitly modeling, learning, and inferring three mental states based on And-Or Graphs – namely, machine's mind, human's mind as inferred by the machine, and machine's mind as inferred by the human. We demonstrated the superiority of X-ToM in gaining human trust relative to baselines.

References

[1] Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*, pages 319–326. IEEE, 2016. 1

[2] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical report, University of Montreal*, 1341(3):1, 2009. 3

[3] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016. 4

[4] R.R. Hoffman. A taxonomy of emergent trusting in the humanmachine relationship. *Cognitive systems engineering: The future for a changing world*, 2017. 1, 3

[5] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018. 1, 2

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. 4

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *ICCV*, 2017. 4

[8] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *34th International Conference on Machine Learning*, 2017. 1, 2

[9] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1, 3

[10] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8827–8836, 2018. 1

[11] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016. 4