XAI: Essential Components and Challenges

Zhiyuan Zhang
Department of Automation
Tsinghua University
z-zy20@mails.tsinghua.edu.cn

Abstract

Explainable AI (XAI) is of crucial significance in future Artificial General Intelligence development. Only when we build explainable AI, can we really trust AI and debug them. We first review major challenges in current XAI research. Current explainable structured modelling has low expressive power and is difficult for scaling-up learning. Interpreting intermediates results of black-box best-performing neural network models relies on specific model architectures and has made only primary steps. Then we propose essential components in the communicative XAI framework: an explainable structured model which has explicit causal chains, explicit Theory of Mind (ToM) modelling and context-aware communication message generating methods.

1 Current XAI challenges

Here we review previous XAI methods and identify their challenges. We split them into two categories: 1. explainable structured modelling and 2. inducing interpretations from black-box neural networks.

Explainable structured modelling From the beginning, we can design models that have explicit representations of perceptions and reasoning processes [1, 3]. However, these models either require lots of manual labor on specific system and knowledge representation design [1], or lack further explanations of perception before explainable parts [3]. Scaling up such systems face challenges of more prior knowledge and bigger unexplainable neural networks.

Interpreting black-box models Interpreting black-box models [4, 2, 6] rely on specific knowledge of model architectures, and the interpretation process cannot be automated or rely heavily on manual choices. And the causal chains of such knowledge is used to solve the problem, or the reasoning process, is not clear.

2 XAI: essential components for a cummunicating framework

To explain AI agents themselves, we identify three essential parts:

1. A structured model. Here we need a model with structured representations. Then we can debug the perceptions and reasoning process. When the process is fully displayed to human researchers, we can trust them [1].
2. A ToM module. A ToM module is needed, because explanation and the establishment of trust relies heavily on our estimations of the human listener. The agent need ToM modules to infer what the human knows and what he cares the most [5], to give an explanation that addresses the human's concerns.
3. A context-aware communication module. This module should be responsible for extracting most efficient and most concerned explanations from the model's structured knowledge and reasoning processl

3 Conclusion

Sorry TA I do not have any more time for this

References

[1] Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, Yixin Zhu, Ying Nian Wu, Hongjing Lu, and Song-Chun Zhu. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, Dec 2019. doi: 10.1126/scirobotics.aay4663. URL `http://dx.doi.org/10.1126/scirobotics.aay4663`. 1

[2] Wes Gurnee and Max Tegmark. Language models represent space and time. Oct 2023. 1

[3] Jingqiao Mao, Chuang Gan, Pushmeet Kohli, JoshuaB. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *Learning,Learning*, Apr 2019. 1

[4] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *Cornell University - arXiv,Cornell University - arXiv*, Apr 2017. 1

[5] Ramya Srinivasan and Ajay Chander. Explanation perspectives from the cognitive sciences—a survey. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Jul 2020. doi: 10.24963/ijcai.2020/670. URL `http://dx.doi.org/10.24963/ijcai.2020/670`. 1

[6] Quanshi Zhang, Xin Wang, Ying Wu, Huilin Zhou, and Song-Chun Zhu. Interpretable cnns for object classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence,IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jan 2019. 1