Effective Representation to Capture Collaboration Behaviors between Explainer and User

Arjun Akula, Song-Chun Zhu
UCLA
aakula@ucla.edu, sczhu@stat.ucla.edu

Abstract

An explainable AI (XAI) model aims to provide transparency (in the form of justification, explanation, etc) for its predictions or actions made by it. Recently, there has been a lot of focus on building XAI models, especially to provide explanations for understanding and interpreting the predictions made by deep learning models. At UCLA, we propose a generic framework to interact with an XAI model in natural language.

1 Introduction

Most work on XAI typically focuses on black-box models and generating explanations about their performance in terms of, e.g., feature visualization and attribution (Sundararajan et al., 2017; Ramprasaath et al., 2016; Zeiler and Fergus, 2014). However, solely generating explanations, regardless of their type (visualization or attribution) and utility, *is not sufficient* for increasing understandability and predictability. Previous studies have shown that trust is closely and positively correlated to the level of how much human users understand the AI system — *understandability* — and how accurately they can predict the system's performance on a given task — *predictability* (Hoffman, 2017; Lipton, 2016; Hoffman et al., 2018; Miller, 2018). Therefore there has been a growing interest in developing explainable AI systems (XAI) aimed at increasing understandability and predictability by providing explanations about the system's predictions to human users (Lipton, 2016; Ribeiro et al., 2016; Miller, 2018; Yang et al., 2018). Current works on XAI generate explanations about their performance in terms of, e.g., feature visualization and attention maps (Sundararajan et al., 2017; Ramprasaath et al., 2016; Zeiler and Fergus, 2014; Smilkov et al., 2017; Kim et al., 2014; Zhang et al., 2018). However, solely generating explanations, regardless of their type (visualization or attention maps) and utility, *is not sufficient* for increasing understandability and predictability (Jain and Wallace, 2019)

In our UCLA lab, our focus is on the Explainer module. Explainer takes a natural language question from the user and identifies the intention behind it. Explainer is also responsible for controlling the dialog flow with the user. Explainable performer provides the important evidences that are necessary to answer user's question. Atomic Performer assists Explainable performer in identifying the evidences. Explainer uses this evidence to generate most acceptable and convincing explanation. We control the dialog flow inside the Explainer using discourse model called Rhetorical Structure Theory (RST). In general, RST would be an efficient and simplest way to track contextual information. Since explanations are context-dependent, we believe that RST would be the right model to capture contextual information in the Explainer (Akula and Zhu, 2019a; Akula et al., 2020a; Akula and Zhu, 2019b; Akula et al., 2021c,d,b).

Given a user's question, we first identify the dialog act of the question. We then identify the question type (contrast type) and explanation type as mentioned in the next section. Based on the explanation type, we generate the explanation and present it to the user (Akula et al., 2013, 2018, 2021a; Gupta et al., 2016; Akula et al., 2019b; Akula, 2021; Akula et al., 2019a, 2020b).

Questions posed by the user to obtain explanations from an XAI model are typically contrastive in nature. For example, questions such as "Why do the model think the people are in sitting posture?", "Why do you think two persons are sitting instead of one?", need contrastive explanations. In order to generate a convincing explanation, XAI model needs to understand the implicit contrast that it presupposes (Agarwal et al., 2018; Akula et al., 2019c; Akula, 2015; Palakurthi et al., 2015; Agarwal et al., 2017; Dasgupta et al., 2014).

Explainer's knowledge using the question types such as NOT-X, NOT-X1-BUT-X2, NOT-X-BUT-Y. Question types such as DO-X, DO-NOT-X and DO-X-NOT-Y are used by the user as intervention techniques. We now propose the following seven types of explanation types that are motivated from an algorithmic approach rather than on theoretical grounds. • Direct Explanation: Explaining based on detection scores • Part-based Explanation: Explaining based on the evidences of detected parts for the concept asked • Causal Explanation, Temporal Explanation: Explaining based on the constraints from the spatiotemporal surround • Common-sense Explanation: Explaining based on the common-sense knowledge of the concept domain • Counter-factual Explanation: Explaining based on the evidences provided for the counter-factual questions asked by the Explainer • Discourse Entailment based Explanation: Explaining based on the discourse relations among various objects/frames in the concept/videos (Akula et al., 2020c; R Akula et al., 2019; Pulijala et al., 2013; Gupta et al., 2012).

References

Shivali Agarwal, Vishalaksh Aggarwal, Arjun R Akula, Gargi Banerjee Dasgupta, and Giriprasad Sridhara. 2017. Automatic problem extraction and analysis from unstructured text in it tickets. *IBM Journal of Research and Development*, 61(1):4–41.

Shivali Agarwal, Arjun R Akula, Gaargi B Dasgupta, Shripad J Nadgowda, and Tapan K Nayak. 2018. Structured representation and classification of noisy and unstructured tickets in service delivery. US Patent 10,095,779.

Arjun Akula, Spandana Gella, Keze Wang, Song-chun Zhu, and Siva Reddy. 2021a. Mind the context: The impact of contextualization in neural module networks for grounding visual referring expressions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6398–6416.

Arjun Akula, Varun Jampani, Soravit Changpinyo, and Song-Chun Zhu. 2021b. Robust visual reasoning via language guided neural module networks. *Advances in Neural Information Processing Systems*, 34.

Arjun Akula, Rajeev Sangal, and Radhika Mamidi. 2013. A novel approach towards incorporating context processing capabilities in nlidb system. In *Proceedings of the sixth international joint conference on natural language processing*, pages 1216–1222.

Arjun R Akula. 2015. A novel approach towards building a generic, portable and contextual nlidb system. *International Institute of Information Technology Hyderabad*.

Arjun R Akula, Beer Changpinyo, Boqing Gong, Piyush Sharma, Song-Chun Zhu, and Radu Soricut. 2021c. Crossvqa: Scalably generating benchmarks for systematically testing vqa generalization.

Arjun R Akula, Gaargi B Dasgupta, and Tapan K Nayak. 2018. Analyzing tickets using discourse cues in communication logs. US Patent 10,067,983.

Arjun R Akula, Gargi B Dasgupta, Vijay Ekambaram, and Ramasuri Narayanam. 2021d. Measuring effective utilization of a service practitioner for ticket resolution via a wearable device. US Patent 10,929,264.

Arjun R. Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020a. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *ACL*.

Arjun R Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020b. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. *arXiv preprint arXiv:2005.01655*.

Arjun R Akula, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Y Chai, and Song-Chun Zhu. 2019a. X-tom: Explaining with theory-of-mind for gaining justified human trust. *arXiv preprint arXiv:1909.06907*.

Arjun R Akula, Changsong Liu, Sinisa Todorovic, Joyce Y Chai, and Song-Chun Zhu. 2019b. Explainable ai as collaborative task solving. In *CVPR Workshops*, pages 91–94.

Arjun R Akula, Sinisa Todorovic, Joyce Y Chai, and Song-Chun Zhu. 2019c. Natural language interaction with explainable ai models. In *CVPR Workshops*, pages 87–90.

Arjun R. Akula, Shuai Wang, and Song-Chun Zhu. 2020c. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2594–2601. AAAI Press.

Arjun R. Akula and Song-Chun Zhu. 2019a. Visual discourse parsing. *CVPR 2019 Workshop on Language and Vision, arXiv:1903.02252*.

Arjun R Akula and Song-Chun Zhu. 2019b. Visual discourse parsing. *ArXiv preprint*, abs/1903.02252.

Arjun Reddy Akula. 2021. *Gaining Justified Human Trust by Improving Explainability in Vision and Language Reasoning Models*. Ph.D. thesis, UCLA.

Gargi B Dasgupta, Tapan K Nayak, Arjun R Akula, Shivali Agarwal, and Shripad J Nadgowda. 2014. Towards auto-remediation in services delivery: Context-based classification of noisy and unstructured tickets. In *International Conference on Service-Oriented Computing*, pages 478–485. Springer.

Abhijeet Gupta, Arjun Akula, Deepak Malladi, Puneeth Kukkadapu, Vinay Ainavolu, and Rajeev Sangal. 2012. A novel approach towards building a portable nlidb system using the computational paninian grammar framework. In *2012 International Conference on Asian Language Processing*, pages 93–96. IEEE.

Abhirut Gupta, Arjun Akula, Gargi Dasgupta, Pooja Aggarwal, and Prateeti Mohapatra. 2016. Desire: Deep semantic understanding and retrieval for technical support services. In *International Conference on Service-Oriented Computing*, pages 207–210. Springer.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

R.R. Hoffman. 2017. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive systems engineering: The future for a changing world*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960.

Zachary C Lipton. 2016. The mythos of model interpretability. In *ICML Workshop on Human Interpretability in Machine Learning*.

Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Ashish Palakurthi, SM Ruthu, Arjun Akula, and Radhika Mamidi. 2015. Classification of attributes in a natural language query into different sql clauses. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 497–506.

Vasu Pulijala, Arjun R Akula, and Azeemuddin Syed. 2013. A web-based virtual laboratory for electromagnetic theory. In *2013 IEEE Fifth International Conference on Technology for Education (t4e 2013)*, pages 13–18. IEEE.

Arjun R Akula, Sinisa Todorovic, Joyce Y Chai, and Song-Chun Zhu. 2019. Natural language interaction with explainable ai models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 87–90.

RS Ramprasaath, D Abhishek, V Ramakrishna, C Michael, P Devi, and B Dhruv. 2016. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CVPR 2016*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

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

Shaohua Yang, Qiaozi Gao, Sari Saba-Sadiya, and Joyce Chai. 2018. Commonsense justification for action explanation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2627–2637.

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

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836.