Proceedings of the 7th International Conference on Statistics: Theory and Applications (ICSTA'25)
Paris, France - August, 2025
DOI: 10.11159/icsta25.004

Causality for Trustworthy Artificial Intelligence

Sheng Li
University of Virginia
1919 Ivy Road, Charlottesville, VA, USA
shengli@virginia.edu

Extended Abstract

Artificial Intelligence (AI) systems have significantly transformed numerous domains, yet their susceptibility to backdoor attacks poses critical threats to reliability and trustworthiness. This talk explores the role of causal analysis in enhancing the trustworthiness of AI models by mitigating such vulnerabilities. The talk begins with a comprehensive introduction to causality, covering basic notations and key assumptions. Then I will introduce our recent work that leverages causal analysis for trustworthy AI. First, we study the inference-stage black-box backdoor detection problem, where defenders aim to build a firewall to filter out the backdoor inputs in the inference stage, with only input samples and prediction labels available. We provide a novel causality-based lens to analyze heterogeneous prediction behaviors for clean and backdoored samples in the inference stage, considering both sample-specific and sample-agnostic backdoor attacks [1]. Motivated by the causal analysis and do-calculus in causal inference, we propose a new approach to distinguish backdoor and clean samples by analyzing prediction consistency after progressively constructing counterfactual samples. Second, we analyze the anti-backdoor learning problem from a causal perspective and propose a new approach to train clean models directly from poisoned datasets [2]. Our approach leverages both the image and the attack indicator to train the model. Based on this training paradigm, the model's perception of whether an input is clean or backdoored can be controlled. These advances highlight the essential role causality in developing safe and trustworthy AI systems.

The reference section at the end of the paper should be edited based on the following:

References

[1] M. Hu, Z. Guan, J. Guo, Z. Zhou, J. Zhang, and S. Li, "BBCaL: Black-box Backdoor Detection under the Causality Lens", *Transactions on Machine Learning Research (TMLR)*, 2024.

[2] M. Hu, Z. Guan, Y. Zeng, J. Guo, Z. Zhou, J. Zhang, R. Jia, A. Vullikanti, and S. Li, "Mind Control through Causal Inference: Predicting Clean Images from Poisoned Data," in *International Conference on Learning Representations (ICLR)*, 2025.