PhD Abstract

Phillip Lippe

March 2021

Title: Temporal Causality in Machine Learning

Abstract: The field of Causality is concerned with finding and understanding the effect that a change of a variable has on other variables in an environment. Meanwhile, current machine learning approaches typically rely on correlations in data but lack the clear separation of cause and effect. This can lead to problems when the correlations don't match the underlying causal relations. For instance, a neural network trained to steer an autonomous car from observations by a human expert will learn to break if it sees the break lights turning on in the car, although breaking a second earlier is the actual cause for the lights to turn on, not the other way round. The intersection of the two fields, Causality and Machine Learning, is an exciting yet difficult research direction as discovering causal relations from raw, high-dimensional data, as used in most machine learning tasks, constitutes a major challenge. Hence, the goal of this PhD is to study and develop new methods to incorporate causal understanding in machine learning models. Our focus thereby lies on Reinforcement Learning and generative models for causal representation learning on temporal data. Causality and Reinforcement Learning are closely related as an agent in an environment can perform actions to learn the effect of its interventions. Furthermore, answering counterfactual queries such as "what would have happened to the ball if the player would have moved the Pong paddle up instead of down" are crucial for learning robust policies in complex environments. Thereby, the time dimension is often neglected in Reinforcement Learning, because learning from consecutive frames breaks the i.i.d. assumption needed for optimization methods like SGD. At the same time, temporal information can be crucial for identifying causal relations as in most natural situations, the cause comes before the effect. Thus, there is great potential in incorporating causal understanding in machine learning approaches, and this project aims to bring the two fields a step closer together.