Sample-efficient Learning of Concepts with Theoretical Guarantees: from Data to Concepts without Interventions.

Hidde Fokkema (University of Amsterdam)

Concept Bottleneck Models (CBM) address some of the challenges modern ML approaches face by learning interpretable concepts from high-dimensional data, e.g. images, which are used to predict labels. In this talk, I will describe a new framework that provides theoretical guarantees on the correctness of the learned concepts and on the number of required labels, without requiring any interventions. Our framework leverages causal representation learning (CRL) methods to learn latent causal variables from high-dimensional observations in a unsupervised way, and then learns to align these variables with interpretable concepts with few concept labels. We propose a linear and a non-parametric estimator for this mapping, providing a finite-sample high probability result in the linear case and an asymptotic consistency result for the non-parametric estimator. We evaluate our framework in synthetic and image benchmarks, showing that the learned concepts have less impurities and are often more accurate than other CBMs, even in settings with strong correlations between concepts.