Model Explainability and Causal Representation Learning

Cheng Guo

Overview

- Causal Representation Learning (CRL)
- Post-Nonlinear ICA Approaches
  - Multiple Distributions
  - Interventions & Distribution Shifts
  - Temporal Condition
- Relationship between CRL and LLM
  - The Definition of Concept
  - Duality with Attention
- Existed Benchmarks on CRL