Model Explainability and Causal Representation Learning

Cheng Guo

UC San Diego

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

Causal Representation Learning

- $X = f(Z)$
  - Output Vector: X
  - Latent Causal Variables Vector: Z
- Identifying Z in a DAG form (Disentanglement)
- Help with Model Explainability
- Various Assumptions

Post-Nonlinear ICA Approaches - Assumption

- PA -> parent
  - u -> latent factor
    - ε -> noise
      - X -> d-dimensional
        - Z -> n-dimensional

$$X = f(Z) + \epsilon, Z_i = g_i(\mathbf{PA}(Z_i), u_i) + \epsilon_i$$

f -> injective

Multiple Distributions

- Non-parametric Setting
    - Heterogeneous data
        - Non-stationary time series
            - No Interventions
                - Sparsity Constraint
- any true hidden causal variables can be recovered up to a component-wise transformation as long as it has no intimate neighbors

Interventions & Distribution Shifts

- Hard Interventions -> remove edges in the causal graph
- Soft Interventions -> different causal mechanisms
  - Linear Gaussian, Polynomial…
  - Latent Additive Noise Model
    - Z(i) identifiable when:
      - No Constant terms in the function g
      - Or Z(i) is a root node

g -> invertible

Temporal Condition

- Temporally Disentangled Representation Learning (TDRL)
  - Non-parametric setting
  - Distribution shifts -> modular representation
  - Extended Sequential VAE
- Mean Correlation Coefficient (MCC)

Relating to LLMs - Definition of Concept

- Concept C - Projector Matrix A
  - $AZ = b$
  - b is d-dimensional
    - -> C is d-dimensional
- Atomic Concept (Atoms) - 1-dimensional C
- Nonlinear Concepts Identified -> Linear Representations
- Environment Diversity

f -> injective + differentiable

UC San Diego

Relating to LLMs - Duality with Attention

- Causal Inference with Attention (CInA)
    - Causal Inference
    - Self-Attention
    - Covariate Balancing
- Outperform in OOD Generalization

UC San Diego

Existed Benchmarks in CRL

- CEBaB
- 3DIdent
  - ResNet-18, FCN, LeakyReLU
  - Causal3DIdent - Hare, Dragon, Cow, Armadillo, Horse, Head
  - CausalWorld, Causal Triplet, CausalCircuit…
- CausalVAE Framework
  - Synthetic: Pendulum, Flow
  - Real-World: CelebA(Beard), CelebA(Smile)
  - Shadow Dataset

Thank you for listening

References:

- Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In Advances in Neural Information Processing Systems, volume 35, 2022.
- Aapo Hyvarinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning, 2023.
- Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent neural causal models, 2024.
- Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models, 2024.
- Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions, 2023.
- Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder, 2023.
- Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 26492–26503. Curran Associates, Inc., 2022.
- Jiaqi Zhang, Joel Jennings, Cheng Zhang, and Chao Ma. Towards causal foundation model: on duality between causal inference and attention, 2023.
- Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting, 2024.
- Jiageng Zhu, Hanchen Xie, Jianhua Wu, Jiazhi Li, Mahyar Khayatkhoei, Mohamed E. Hussein, and Wael AbdAlmageed. Shadow datasets, new challenging datasets for causal representation learning, 2023.

UC San Diego