Towards Causal Representation Learning

Woong-Hee Lee

2025.02.06

Overview

1. Abstract & Intro

2. Second Section

Two key topics of paper

1. Delineate some implications of causality for machine learning
2. Propose key research areas at the intersection of both communities.

Paper Structure

a. Robustness
b. Learning Reusable Mechanisms
c. A Causality Perspective

I. Intro
II. Describe different levels of modeling in physical systems
III. Present the differences between causal and statistical models
IV. Independent Causal Mechanisms (ICM)
V. Review existing approaches to learn causal relations
VI. how useful models of reality may be learned from data in the form of causal representations
VII. assay the implications of causality for practical machine learning

II - Differential equations (from physical mechanisms)

$$\frac{dx}{dt} = f(x), x \in \mathbb{R}^d$$

If we formally write this in terms of infiniesimal differentials dt and $dx = x(t + dt) - x(t)$ we get

$$x(t + dt) = x(t) + dt \cdot f(x(t))$$

II - Summary Table of Models

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

II - A. Predicting in the i.i.d setting

- Statistical models are a superficial description of reality as they are only required to model associations.
  e.g. what is the probability of heart failure given certain diagnostic measurements carried out on a patient?

- Changes in outcomes due to an intervention cannot be determined by correlation.
  e.g. There is a correlation between the number of storks and the birth rate in Europe.

  $\rightarrow$However, increasing the stork population (intervention) does not lead to an increase in the birth rate.

  $\rightarrow$In other words, correlation does not imply causation, and the intervention has no causal effect on the outcome.

II - B. Predicting Under Distribution Shifts (a)

- Interventional questions are more challenging than predictions as they involve actions that take us out of the usual i.i.d setting of statistical learning.

- An intervention not only changes the values of specific causal variables but also alters their joint distribution. As a result, classical statistical learning guarantees no longer hold.
  e.g. if smoking becomes more socially stigmatized (intervention), will the number of smokers decrease?

  $\rightarrow$The intervention of increasing social stigma against smoking is introduced.

  $\rightarrow$As a result, the joint distribution related to smoking and the number of smokers changes.

II - B. Predicting Under Distribution Shifts (b)

- What if we could learn interventions? (Here, "intervention" does not necessarily refer to deliberate actions.)

  $\rightarrow$Then, we could build AI models that are robust to naturally occurring distribution shifts in the real world.

- However, prediction performance under distribution shifts should not be evaluated based solely on accuracy.

- Moreover, if machine learning is to be integrated into decision-making processes, we must ensure that its predictions remain reliable even when experimental conditions change.

  $\rightarrow$ Causal inference can help satisfy this requirement.

II - C. Answering Counterfactual Questions - (a)

- Counterfactual reasoning involves imagining alternative possibilities based on events that have already occurred.
    - Since it requires predicting hypothetical outcomes, it is more challenging to observe in machine learning.
- In contrast, intervention is simply an experimental manipulation, making counterfactual reasoning a broader concept.
    - The Structural Causal Model (SCM) provides a mathematical framework for expressing counterfactual questions
    - Interventions are relatively easier because they involve directly manipulating variables.

II - C. Answering Counterfactual Questions - (b)

- Why is counterfactual reasoning important?
  - It allows us to imagine the outcomes of alternative actions and identify the best course of action to achieve a desired result.
  - This is crucial for decision-making in artificial intelligence.
- Counterfactual reasoning is also useful in reinforcement learning.
  - It enables agents to analyze past actions and learn better strategies.
  - It allows AI to verify experiences and learn in a way similar to the scientific method.

II - C. Answering Counterfactual Questions - (c)

- Intervention vs. Counterfactual Examples
  - Intervention
    - “How does the probability of heart failure change if we convince a patient to exercise regularly?”
    - ”If a patient is encouraged to exercise regularly, how does the probability of heart failure change?”
  - Counterfactual
    - “Would a given patient have suffered heart failure if they had started exercising a year earlier?””
    - ”If the patient had started exercising a year earlier, would they have suffered heart failure?””

II - C. Answering Counterfactual Questions - (d)

- This is why counterfactual reasoning is important in reinforcement learning.
  - It allows agents to modify past factors and estimate probabilities, leading to better feedback and improved decision-making.
  - This ultimately helps optimize an agent's decision-making process.

II - D. Nature of Data: Observational, Interventional, (Un)structured - (a)

- Data Formats
  - Observational vs. Interventional
  - Hand-Engineered vs. Raw (Unstructured)
    - e.g., Manually processed data (e.g., data tables, etc.)
    - Images, audio, and video are examples of unstructured data.

II - D. Nature of Data: Observational, Interventional, (Un)structured - (b)

- Interventions
  - Explicit Interventions
    - Since direct experimental interventions are performed (e.g., A/B testing), the effects of the intervention can be analyzed.
  - Domain Shift / Unknown Interventions
    - Interventions exist, but it is unclear what specific intervention has taken place.
    - e.g., Market changes: Over time, consumer purchasing patterns shift, but the exact cause of the intervention is unknown.

II - D. Nature of Data: Observational, Interventional, (Un)structured - (c)

- Hand-Engineered Data Raw Data
  - Hand-Engineered Data
    - Assumes that past data is structured at a high level.
    - If the data is structured in this way, it is easier to match with causal variables and infer causal structures.
  - Raw Data
    - If the data is not structured in this way, forming causal structures becomes more challenging.
    - While statistical models are weaker than causal models, they have the advantage of being learnable from raw data.

II - D. Nature of Data: Observational, Interventional, (Un)structured - (d)

- Even if causal relationships can be learned from limited observational data,
    - data collected from diverse environments is required, and
    - a method for performing interventions is also necessary.

II - D. Nature of Data: Observational, Interventional, (Un)structured - (e)

- Future Direction: Reducing Expert Involvement
  - Currently, feature selection itself requires a significant amount of prior knowledge to determine which variables to measure.
  - In the future, it will be necessary to reduce direct expert-driven data collection and instead leverage techniques such as meta-learning and self-supervision to learn causal relationships.
- The Utility of Causal Models Depends on the Environment and Task
  - To train a causal model suitable for a specific environment and task, the appropriate granularity of high-level variables is required.
  - In other words, the complexity of a causal model depends on what types of interventions can be applied and what data is available.