ICML | 2022
Thirty-ninth International Conference on Machine Learning

ROCK: Causal Inference Principles for Reasoning about Commonsense Causality




Jiayao Zhang[*] **Hongming Zhang**[*†] **Weijie J. Su**[*] **Dan Roth**[*‡]

[*]University of Pennsylvania [†]Tencent AI Lab, US [‡]AWS AI Labs

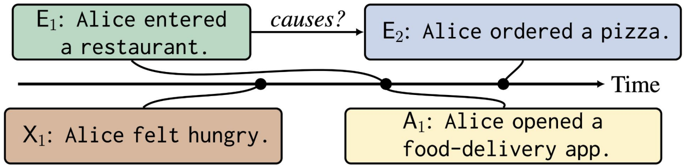
ICML | 2022
Thirty-ninth International Conference on Machine Learning

Commonsense Causality Reasoning (CCR)

Given two events (described in natural languages), reasoning about their cause-and-effect relationships in a way that corresponds to an average person's judgement.

Concrete Problems

- Estimation/Inference: does E_1 cause E_2?
- Generation/Explanation: what causes E_1?



Q: Does E_1 cause E_2?

Desiderata

- Commonsense: aligns with human's commonsense
- Zero-shot: use only pre-trained language models

Challenges

- How to account for confounders (confounding co-occurrences)?
- How to adopt formal causal inference models?

What is being done in observational studies?

Example: `E`$_1$`: Alice entered a restaurant. E`$_2$`: Alice ordered a pizza.`

First Goal: Define study units, treatments, potential outcomes, and the estimand.

Unit	Covariates				Treatment T	Observed Outcome Y
	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	...		
1	1	0	1	...	1	1
2	0	0	1	...	0	0
3	0	1	0	...	0	1

Definitions

Study Unit: Alices (i.e., humans)

Covariates $X_{i,j}$: Occurrence of the jth **context** to the ith unit

Treatment T_i: Occurrence of E_1 (to the ith unit)

Outcome Y_i: Occurrence of E_2 (to the ith unit)

The Causal Estimand (Average Treatment Effect)

$$\Delta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

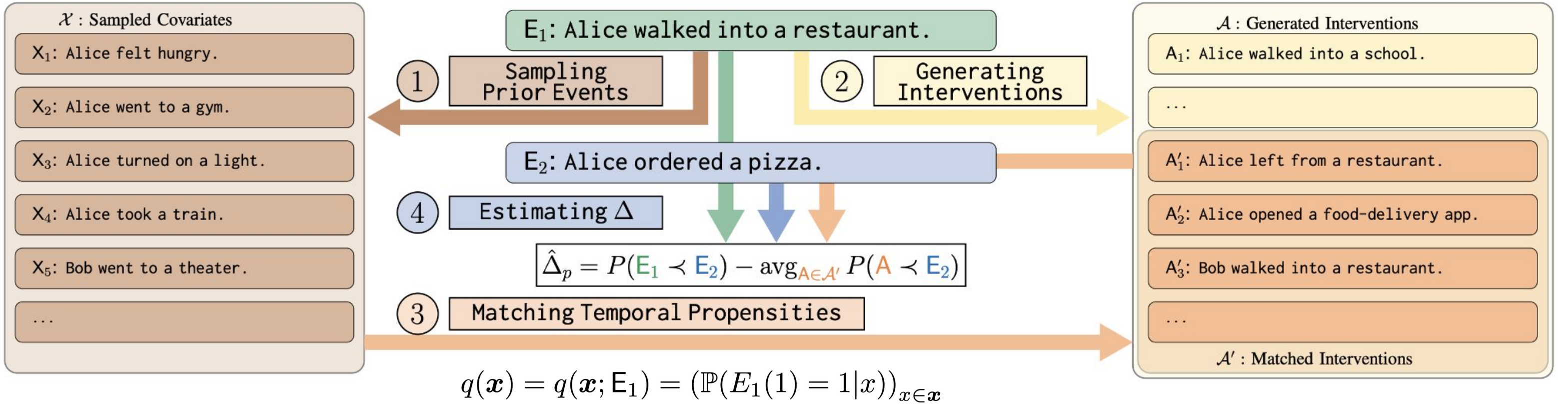
$$= \mathbb{E}_X[\, \mathbb{E}[\, Y(1) \mid X, T\,] - \mathbb{E}[\, Y(0) \mid X, T\,]\,] \quad \text{(ignorability)}$$

$$= \mathbb{E}[\, \mathbf{1}\{E_1 \prec E_2\}\,] - \mathbb{E}[\, \mathbf{1}\{\neg E_1 \prec E_2\}\,] \quad \text{(notation)}$$

$$= \mathbb{P}(\, E_1 \prec E_2\,) - \mathbb{P}(\, \neg E_1 \prec E_2\,)$$

$Y_i(T)$: the potential outcome of the ith unit corresponds to the treatment T

The ROCK Framework



1. Sample a set of events X_i (**contexts**) that occur before E_1.
2. Generate a set of **interventions** A_j based on E_1.
3. Select the **comparable interventions** by matching on **temporal propensities**.
4. Estimate the **causal estimand Δ** and report the result.

ROCK: Evaluation

- Evaluation
  - Datasets: Choice of Plausible Alternatives (COPA), and GLUCOSE.
  - Method: compute the estimand Δ for two choices, choose the choice with a higher Δ.
  - Example:

Example B.1 (Did E_1 cause $E_2^{(1)}$ or $E_2^{(2)}$?).

E_1 : The teacher assigned homework to the students.
$E_2^{(1)}$: The students passed notes.
$E_2^{(2)}$: The students groaned.

- Ablations
  - Pre-trained LM vs. a fine-tuned LM (on NYT) for temporality predictor.
  - On covariate set size.
  - On various normalization choices (e.g., how to normalize the temporal probabilities).

ROCK: Empirical Results

Performance (accuracy) on COPA and GLUCOSE

	Random Baseline	$\hat{\Delta}_1 \uparrow$ L_1-Balanced	$\hat{\Delta}_2 \uparrow$ L_2-Balanced	$\hat{\Delta}_{E_1} \uparrow$ Temporal	$\hat{\Delta}_{\mathcal{A}} \uparrow$ Unbalanced	$\hat{\Delta}_{\mathcal{X}} \uparrow$ Misspecified
COPA-Dev	0.5 ± 0.050	0.6900	**0.7000**	0.5800	0.5600	0.5300
COPA-Test	0.5 ± 0.022	**0.5640**	**0.5640**	0.5200	0.5400	0.5240
GLUCOSE-D1	0.5 ± 0.040	0.6645	**0.6968**	0.5677	0.5742	0.6581
COPA-Dev (-T)	0.5 ± 0.050	0.6200	**0.6300**	0.5300	0.4800	0.5300
COPA-Test (-T)	0.5 ± 0.022	**0.5800**	0.5740	0.4540	0.4600	0.4860
GLUCOSE-D1 (-T)	0.5 ± 0.040	0.6065	**0.6194**	0.5548	0.4387	0.3742

proposed (using ROCK)

unadjusted baselines

- Adjusted scores Δ_p are better than unadjusted scores (the last three columns).
- On COPA-Dev, the performance is similar to self-talk while being truly zero-shot.
- When computing temporal propensities (Step 3), a fine-tuned LM (first three rows) outperforms its pre-trained counterpart (last three rows).

Self-talk method: https://aclanthology.org/2020.emnlp-main.373/

Summary

- Adopt the **potential-outcomes framework** for the CCR task: find comparable interventions.
- Propose a modular framework, ROCK, to estimate the temporality-motivated *causal estimand* by **temporal propensity matching**.
- Empirical studies and ablation studies demonstrate ROCK's effectiveness in zero-shot CCR.

Future Work

- Implicit events
- Explanation generation

ROCK: Causal Inference Principles for CCR

Model

Code

Paper