Causal Abstractions and Causal Representation Learning

Sander Beckers

University of Tübingen – Cluster of Excellence in Machine Learning
sanderbeckers.com

Causal Representation Learning Workshop, 05/08/2022

Outline

Outline

Goals

1. Formulate Causal Abstraction Learning (CAL)

2. Interpret CRL as lying somewhere between CAL and RL

Big Picture: Goal 1

RL = Representation Learning: use $\vec{X}$ to learn $\mathcal{V}_H$

CRL = Causal Representation Learning: use $\vec{X}$ to learn

- $\mathcal{U}_H$, $\mathcal{V}_H$,
- M_H over $\mathcal{U}_H$, $\mathcal{V}_H$.

CAL = Causal Abstraction Learning: use $(\mathcal{U}_L, \mathcal{V}_L, \mathcal{I}_L)$ to learn

- M_L over $\mathcal{U}_L$, $\mathcal{V}_L$,
- $\mathcal{U}_H$, $\mathcal{V}_H$,
- M_H over $\mathcal{U}_H$, $\mathcal{V}_H$,
- such that M_H is a causal abstraction of M_L.

Big Picture: Goal 2

RL:

- understandable (Auto Encoders)
- realistic (data exists)
- non-causal

CAL:

- understandable (Causal Abstractions + Auto Encoders)
- unrealistic (data does not exist, M_L too complex)
- very causal (both M_L and M_H)

CRL:

- understandable *if we view it as simplified CAL*
- realistic *if we view it as complicated RL*
- causal enough (M_H)

Outline

CRL using Auto Encoders

- Schölkopf, B. Locatello, F., Bauer, S., Ke, NR., Kalchbrenner, N., Goyal, A., and Bengio, Y.: Towards Causal Representation Learning, IEEE, 2021
- von Kügelgen, J., and Schölkopf, B.: From Statistical to Causal Learning, Preprint, 2022

Data: Low-level, high-dimensional, entangled $\vec{X}$

Target:

- High-level, low-dimensional, disentangled $\mathcal{U}_H$ and $\mathcal{V}_H$
- Causal model M_H over $\mathcal{U}_H$ and $\mathcal{V}_H$

Standard Auto Encoder

$\vec{X} = p(\mathcal{V}_H)$ with p = Decoder

$\mathcal{V}_H = q(\vec{X})$ with q = Encoder

Choose distance function d over $\vec{X}$, a suitable α, and consider the reconstruction loss (expected, worst-case, etc.,):

$$d(\vec{X}, p(q(\vec{X}))) < \alpha$$

But our $\mathcal{V}_H$ are not independent...

Reduced Form Auto Encoder (RFAE)

Reduced form of a causal model: $\mathcal{V}_H = m_H(\mathcal{U}_H)$

So we learn $\mathcal{V}_H = m_H(\mathcal{U}_H)$, $\vec{X} = p(m_H(\mathcal{U}_H))$, and $\mathcal{U}_H = q(\vec{X})$

such that

$$d(\vec{X}, p(m_H(q(\vec{X}))) < \alpha$$

Limitation of RFAE

Of course reduced form entirely ignores *interventions*!

M_H can be seen as a function $M_H : \mathcal{U}_H \times \mathcal{I}_H \to \mathcal{V}_H$

m_H is simply M_H for empty intervention: $m_H(\mathcal{U}_H) = M_H(\mathcal{U}_H, \emptyset)$

What is CRL? The challenge of learning M_H for *all* $\mathcal{I}_H$.

Outline

Outline

Literature

Rubenstein, P.K., Weichwald, S., Bongers, S., Mooij, J.M., Janzing, D., Grosse-Wentrup, M., Schölkopf, B.: Causal Consistency of Structural Equation Models, UAI 2017

Beckers, S. and Halpern, J.: Abstracting Causal Models, AAAI 2019

Beckers, S., Eberhardt, F., and Halpern, J.: Approximate Causal Abstraction, UAI 2019

Causal Models

Causal Model $M = (\mathcal{V}, \mathcal{U}, \mathcal{F}, \mathcal{I})$:

- $\mathcal{V}$: endogenous variables
- $\mathcal{U}$: exogenous variables
- $\mathcal{F}$: set of structural equations (one for each $X \in \mathcal{V}$):
- $\mathcal{I}$: set of allowed interventions, i.e., the interventions that *we care about*, or *that are possible*.
  - Innovation by Rubenstein et. al. (2017):
  - E.g., $A := X_1 + X_2$. Then can we allow an intervention like $X_1 \leftarrow 5$? What choice of $A \leftarrow a$ would work?

Probabilistic Causal Models: add Pr over $\mathcal{U}$, this induces $\Pr_{\mathcal{V}}$ over $\mathcal{V}$.

Causal Abstraction

Say we have $M_L = (\mathcal{V}_L, \mathcal{U}_L, \mathcal{F}_L, \mathcal{I}_L)$, $M_H = (\mathcal{V}_H, \mathcal{U}_H, \mathcal{F}_H, \mathcal{I}_H)$, and an *abstraction function* τ: $\mathcal{V}_H = \tau(\mathcal{V}_L)$

Challenge: can we extend the interpretation of τ so that we make sense of: $M_H = \tau(M_L)$?

Remember: $M_L : \mathcal{U}_L \times \mathcal{I}_L \to \mathcal{V}_L$

To extend τ, we need:

- $\tau_{\mathcal{U}_L} : \mathcal{U}_L \to \mathcal{U}_H$
- $\omega_\tau : \mathcal{I}_L \to \mathcal{I}_H$

$$
\begin{tikzcd}
(\vec{u}, C \leftarrow c) \arrow[r, "M_H(\,.\,)"] & E = e' \approx e \\
(\vec{u}, \vec{W} \leftarrow \vec{w}) \arrow[u, "{\tau_U(\,.\,), \omega_\tau(\,.\,)}"'] \arrow[r, "M_L(\,.\,)"] & \vec{T} = \vec{t} \arrow[u, "\tau(\,.\,)"]
\end{tikzcd}
$$

How well does M_H approximate M_L?

$=$

How close are the predictions of M_H and M_L?

$=$

distance between e' and e.

Causal Abstraction

$$M_H = \tau(M_L)$$

iff

for all $\vec{u}_L$, $\vec{W} \leftarrow \vec{w} \in \mathcal{I}_L$:

$$\tau(M_L(\vec{u}_L, \vec{W} \leftarrow \vec{w})) = M_H(\tau_{\mathcal{U}_L}(\vec{u}_L), \omega_\tau(\vec{W} \leftarrow \vec{w})).$$

τ-α approximate abstraction: play around (expected, worst-case, etc.,) with probabilities of

$$d(\tau(M_L(\vec{u}_L, \vec{W} \leftarrow \vec{w})), M_H(\tau_{\mathcal{U}_L}(\vec{u}_L), \omega_\tau(\vec{W} \leftarrow \vec{w}))) < \alpha$$

Where do $\tau_{\mathcal{U}_L}$ and ω_τ come from?

- Just look for best $\tau_{\mathcal{U}_L}$ that does the job
- $\omega_\tau(\vec{Y} \leftarrow \vec{y}) = \vec{Z} \leftarrow \vec{z}$ if "$\tau(\vec{y}) = \vec{z}$", and not defined else.

Outline

Problem Formulation

Data: Low-level, high-dimensional, disentangled $(\mathcal{U}_L, \mathcal{I}_L, \mathcal{V}_L)$

Target:

- High-level, low-dimensional, disentangled $\mathcal{U}_H$ and $\mathcal{V}_H$
- Causal model M_H over $\mathcal{U}_H$ and $\mathcal{V}_H$
- Causal model M_L over $\mathcal{U}_L$ and $\mathcal{V}_L$
- $\tau : \mathcal{V}_L \rightarrow \mathcal{V}_H$ so that $M_H = \tau(M_L)$

Part 1: RL for CAL

Apply Standard Auto Encoders to learn the functions and representations (ignoring causality):

$\mathcal{U}_L = p_{\mathcal{U}}(\mathcal{U}_H)$ and $\mathcal{U}_H = \tau_{\mathcal{U}}(\mathcal{U}_L)$ such that $d(\mathcal{U}_L, p_{\mathcal{U}}(\tau_{\mathcal{U}}(\mathcal{U}_L))) < \alpha$.

$\mathcal{V}_L = p_{\mathcal{V}}(\mathcal{V}_H)$ and $\mathcal{V}_H = \tau(\mathcal{V}_L)$ such that $d(\mathcal{V}_L, p_{\mathcal{V}}(\tau(\mathcal{V}_L))) < \alpha$.

$\mathcal{I}_L = p_{\mathcal{I}}(\mathcal{I}_H)$ and $\mathcal{I}_H = \omega(\mathcal{I}_L)$ such that $d(\mathcal{I}_L, p_{\mathcal{I}}(\omega(\mathcal{I}_L))) < \alpha$.

Part 2: Causal Abstraction Constraints

1. $\omega \approx \omega_\tau$
   - Do they have similar domains?
   - $d(\omega(\mathcal{I}_L), \omega_\tau(\mathcal{I}_L)) < \alpha$?

2. Find M_L and M_H such that M_H is a τ-α approximate abstraction of M_L:

$$d(\tau(M_L(\vec{u}_L, \vec{W} \leftarrow \vec{w})), M_H(\tau_{\mathcal{U}_L}(\vec{u}_L), \omega_\tau(\vec{W} \leftarrow \vec{w}))) < \alpha$$

Outline

CAL does not match CRL context

Low-level is entangled

Low-level causal model is too complex

We don't have data/knowledge of low-level interventions

We don't observe the low-level exogenous variables

AE's only work when the high-level variables are independent

Solution

Move away from CAL towards RL

Suggestion:

- We don't need M_L
- Acquire separate data sets $\vec{X}_i$ under specific (but unknown) high-level interventions
- Solve RFAE problem for each set i
- Require that the combination of solutions are consistent (i.e., are all derived from a single M_H)

Realistic CRL?

Data: for each $i \in \{1, \ldots, n\}$:

- low-level, high-dimensional, entangled $\vec{X}_i$
- where i corresponds to unknown unique high-level $\vec{C}_i \leftarrow \vec{c}_i$

Target:

- High-level, low-dimensional, disentangled $\mathcal{U}_H$ and $\mathcal{V}_H$
- Causal model M_H over $\mathcal{U}_H$ and $\mathcal{V}_H$

Part 1: RFAE

For each i we learn:

$\mathcal{V}_H = m_i(\mathcal{U}_H)$, $\vec{X}_i = p_i(m_i(\mathcal{U}_H))$, and $\mathcal{U}_H = q_i(\vec{X}_i)$

such that

$$d(\vec{X}_i, p_i(m_i(q_i(\vec{X}_i)))) < \alpha$$

Part 2: Causal Constraints

Remember: $M_H : \mathcal{U}_H \times \mathcal{I}_H \to \mathcal{V}_H$

So find M_H and $\vec{C}_i \leftarrow \vec{c}_i$ such that for all i:

$$d(m_i(\mathcal{U}_H), M_H(\mathcal{U}_H, \vec{C}_i \leftarrow \vec{c}_i)) < \alpha$$

Conclusion

1. Defined Causal Abstraction Learning in the image of CRL.
2. Causal Abstraction Learning can help in understanding CRL.