Continual Causal Abstractions

Anonymous submission

Abstract

This short paper discusses continually updated causal abstractions as a potential direction of future research. The key idea is to revise the existing level of causal abstraction to a different level of detail that is both consistent with the history of observed data and more effective in solving a given task.

Overview. [1] discusses the necessity of (a) causal abstractions for effectively solving tasks and (b) continual updates when data starts changing. [2] highlights existing approaches for (a). [3] discusses starting points for (b).

[1] Motivation

Both causality (as defined by Pearl (2009)) and continual learning (Hadsell et al. 2020) seem integral for the quest of understanding (artificial) intelligence. Causality's main contribution has been the *formalization* of key ideas such as interventions, counterfactuals and structural mechanisms. Continual learning on the other hand raised awareness for the importance of learning stable concepts *continuously over time* void of access to previous experiences much like biological systems. The goal of the AAAI 2023 Bridge Program on "Continual Causality" lies in finding answers to the question of what may be found at the intersection of the two subfields primarily studied in AI and CogSci research. This short paper envisions the extension of existing work on causal abstractions towards a continual learning setting where the task solving agent ought to revise its current model abstraction towards a new level of detail in order to be consistent with the history of observed data and also find more effective decision rules for solving the current task.

Why do we need causal abstractions? Let's illustrate with a simple example. Imagine being a chemist analyzing a particular gas. Your advisor tasks you to analyze the temperature and pressure in the volume. Using your thermo- and barometer you simply measure the desired quantities. The first task is solved. Next, your advisor tasks you to analyze the average velocity of a moving particle within the gas. It quickly becomes apparent to you that the previous level of description for the gas becomes obsolete, since the question has shifted from a macro- to a microscopic level of detail in which we suddenly need information on individual particles. Put differently, we are in need of a different level of *abstraction*. Not just that, we further want our new model to still be consistent with our previous observations, that is, if we were to measure net kinetic energy of all relevant particle combinations, then we would ideally like to see a match with the previous measurements of the thermometer. This requirement is what is being captured by the *causal* part of the abstraction transformation between the two models for each of the tasks.

Why do we need to continual abstractions? Let's illustrate again using an example. The following is a sneak peek into the lead example from the vision schematic presented in Fig.1. Imagine being a dietitian analyzing the (causal) effect of a particular diet onto the risk of heart disease. Your history of clients has taught you that the total cholesterol of the patient is characteristic of whether or not the risk of heart disease for that individual is increased or not. This is your initial, base abstraction: if the diet is balanced, the cholesterol levels will deteriorate and the risk of heart disease decreases. Now a new client, a sumo practitioner, enters your diet program but ends up overdoing it and eating three times the amount of items listed in the plan. To your surprise, although the cholesterol levels of the sumo were increased, his risk of heart disease had decreased. To cope with this new counterexample to the previous hypothesis, you decide to revise your abstraction as you found the high- and low-density lipoproteins to be more predictive of the risk of heart disease. For the sumo's case, the former increased which also increased the total cholesterol levels while still lowering the risk of heart disease. In other words, the dietitian *continually* updated the current best causal abstraction to comply with the data history while still answering the initial scientific question effectively.

[2] Existing Work on Causal Abstractions

Definitions. The study of causal abstractions is a subfield in Pearlian causality that aims at *formalizing* the philosophical concept of an abstraction such that the resulting definition is maximally "useful in practice" (commonly taken to mean that the examples such as that of the chemist from [1] work with the definition). Rubenstein et al. (2017) conducted pioneering work in establishing a formalism that discusses "exact transformations" between Structural Causal Model (SCM), which allowed the authors to (i) marginalize

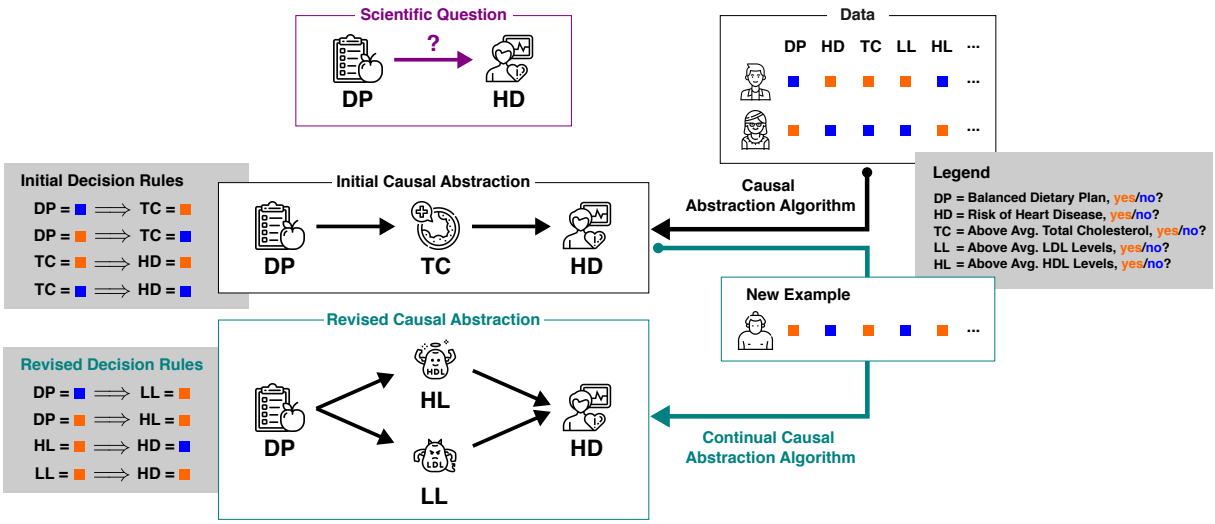

Figure 1: **Vision of Continual Causal Abstractions (CCA).** A schematic illustration for CCA. The scientific question (top purple box) asks about the causal relation (or effect) of a balanced diet onto the risk of heart disease. Based on data consisting of patient records that record various features (see right grey box for a legend), a causal abstraction algorithm provides the initial causal abstraction (middle box) that suggests optimal decision rules (middle left grey box) based on the mediator variable of total cholesterol. With the incoming new example (right teal box) the CCA algorithm provides an updated causal abstraction that uses both HL and LL as mediator variables (middle teal box) with new optimal decision rules (lower left grey box). In summary, the macroscopic view of cholesterol levels in the initial abstraction was sufficient for analyzing the initial two data points, however, the third data point required a more fine grained abstraction that considers the levels of high- and low-density lipoproteins since an increase in former also leads to an increase of TC but is actually lowering HD. (Best viewed in color.)

out 'irrelevant' variables, (ii) aggregate variables into sensible groups, and even (iii) view dynamic systems as their stationary counterpart. Following that, Beckers and Halpern (2019) fixed several shortcomings by generalizing the former formalism to "(strong) abstractions" that (i) work on SCMs directly opposed to probabilistic parameterizations and (ii) consider all possible interventions of an SCM opposed to only a selected subset. To provide a short glimpse into the state of the art formalization, given the standard formulation of an SCM $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathcal{I}, \Pr)$ with a poset $\mathcal{I}$ indicating possible perfect interventions, we have:

Definition 1 *For low-level SCM M_L, a τ-abstraction to high-level SCM M_H is given by*

1. *surjective $\tau : \mathcal{R}_L(\mathcal{V}_L) \to \mathcal{R}_H(\mathcal{V}_H)$ s.t. $\tau(\Pr_L^i) = \Pr_H^{\omega(i)}$ with $i \in \mathcal{I}_L$ and order-preserving, surjective $\omega : \mathcal{I}_L \to \mathcal{I}_H$,*
2. *surjective $\tau_{\mathcal{U}} : \mathcal{R}(\mathcal{U}_L) \to \mathcal{R}(\mathcal{U}_H)$ s.t. $\tau(M_L(\mathbf{u}_L, i)) = M_H(\tau_{\mathcal{U}}(\mathbf{u}_L), \omega(i))$, and*
3. *$\mathcal{I}_H = \omega_\tau(\mathcal{I}_L)$ where ω_τ restricts ω further to subsets of a given intervention.*

Returning to the chemist example from [1], suitable formalizations for M_H, M_L can be shown to commute via some τ-abstraction for a corresponding choice of $\tau, \tau_{\mathcal{U}}, \omega$, and ω_τ.

Learning. To the best of our knowledge, there exist no works yet on actually learning τ-abstractions i.e., no *automatic* verification of abstractions for SCM. While an end-to-end system will require such an abstraction learner, we are fortunate in that CCA can be researched independently.

[3] Future Work: Updating Existing Abstractions

While even just learning causal abstractions as in Def.1 remains an open problem, it is not a requisite for starting to work on *continual* causal abstractions since the interfaces at input and output of the to-be-developed algorithm are already provided. Furthermore, we can get access to models as in "initial- [and] revised causal abstraction" highlighted in Fig.1 by simply training neurally parameterized SCM (see Xia et al. (2021) for a formal introduction) on suitable data. In the next step, we extract decision rules from the neural network modules. While this is a hard problem in general, using suitable assumptions that still cover our problem instance (here, the "DP→HD?"-question) such as linear structural equations and binary variables, we can easily extract decision rules of the form DP=1 $\Longrightarrow$ TC=0. Given such a set of rules (as highlighted in the grey boxes on the left side of Fig.1), we can simply check our new counterexample data point (here: sumo practitioner with both TC=1, HD=0) against the extracted rules. If an inconsistency is found, then the update procedure for existing abstractions triggers. An objective for achieving the desired abstraction could be to again maximize prediction but s.t. the constraint of dropping the previous predictive variable (here: TC).

In summary we can state that the two key research questions towards CCA seem to be: (1) how do we *spot* inadequate abstractions (presented idea was checking counterexample against learned decision rules), and (2) how do we *update* our abstraction to become adequate (presented idea was an objective that maximizes prediction while discarding key variables from the previous abstraction).

References

Beckers, S.; and Halpern, J. Y. 2019. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, 2678–2685.

Hadsell, R.; Rao, D.; Rusu, A. A.; and Pascanu, R. 2020. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12): 1028–1040.

Pearl, J. 2009. *Causality*. Cambridge university press.

Rubenstein, P. K.; Weichwald, S.; Bongers, S.; Mooij, J. M.; Janzing, D.; Grosse-Wentrup, M.; and Schölkopf, B. 2017. Causal Consistency of Structural Equation Models. In Elidan, G.; and Kersting, K., eds., *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI-17)*. AUAI Press.

Xia, K.; Lee, K.-Z.; Bengio, Y.; and Bareinboim, E. 2021. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34: 10823–10836.