Causal Entropy and Information Gain for Measuring Causal Control

Francisco N. F. Q. Simoes, Mehdi Dastani, and Thijs van Ommen

Department of Information and Computing Sciences, Utrecht University
{f.simoes,m.m.dastani,t.vanommen}@uu.nl

Abstract. To effectively study complex causal systems, it is often useful to construct representations that simplify parts of the system by discarding irrelevant details while preserving key features. The Information Bottleneck (IB) method is a widely used approach in representation learning that compresses random variables while retaining information about a target variable [8]. However, traditional methods like IB are purely statistical and ignore underlying causal structures, rendering them ill-suited for causal tasks.

In this paper, we propose the Causal Information Bottleneck (CIB), a causal extension of the IB that compresses a set X of chosen variables while maintaining causal control over a target variable Y. This method produces representations T that are causally interpretable and can be effectively utilized in reasoning about interventions. We present experimental results demonstrating that the learned representations accurately capture causality as intended.

To achieve this, we extended the notion of optimal representation to the causal setting, resulting in an axiomatic characterization of *optimal causal representations*. Just as the IB Lagrangian $\mathcal{L}^{\beta}_{\text{IB}} = I(Y;X) + \beta I(Y;T)$ can be minimized to learn optimal representations [1], so can the CIB Lagrangian $\mathcal{L}^{\beta}_{\text{CIB}} = I(Y;X) + \beta I_c(Y \mid do(T))$ introduced in this paper be used to learn optimal causal representations.

The CIB, which needs to be computed during the optimization process, depends on the interventional distributions $p(y \mid do(t))$ through the term $I_c(Y \mid do(T))$. We proposed a definition of *representation intervention*, which defines the post-intervention distributions $p(y \mid do(t))$ in terms of the post-intervention distributions $p(y \mid do(x))$. As a consequence, $p(y \mid do(t))$ is identifiable when the distributions $p(y \mid do(x))$ are identifiable. We focused on scenarios where there exists a set $\mathbf{Z}$ satisfying the backdoor criterion relative to (X, Y) [6]. In those cases, the $p(y \mid do(x))$ is identifiable, and one can make use of a *backdoor adjustment formula for representations* to derive a backdoor adjustment formula for $p(y \mid do(t))$, enabling the successful application of a minimization algorithm to minimize the CIB.

For optimizing the CIB in our experiments, we introduced a local search algorithm, referred to as projected simulated annealing gradient descent (pSAGD), which integrates simulated annealing and projected gradient descent techniques. To compare different representations learned by our algorithm, we introduced a notion of *equivalence ($\cong$) of representations*,

which partitions representations into equivalence classes, termed abstractions. We demonstrated that the variation of information can be used to assess whether two representations are equivalent, provided one of them is deterministic. Our experiments showed that the learned representations in three toy models of increasing complexity align with our expectations (up to $\cong$).

Looking forward, future research directions include exploring alternative methods for incorporating causality into the information bottleneck framework, such as focusing on causal properties other than causal control, like proportionality [7]. Our approach can also be extended to scenarios where the backdoor criterion does not hold by leveraging do-calculus to facilitate the automatic computation of post-intervention distributions for interventions on representations. Another natural next step involves adapting the CIB method to continuous variables, potentially using variational autoencoders [5] to minimize the CIB Lagrangian, as previously done for the standard IB [2]. Lastly, exploring the relationship between our representation learning method and the framework of causal abstractions is also a promising avenue for future research, similar to how [3] connect the latter with the approach from [4].

References

1. Achille, A., Soatto, S.: Information dropout: Learning optimal representations through noisy computation. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2897–2905 (2018)
2. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. arXiv preprint arXiv:1612.00410 (2016)
3. Beckers, S., Eberhardt, F., Halpern, J.Y.: Approximate causal abstractions. In: Uncertainty in artificial intelligence. pp. 606–615. PMLR (2020)
4. Chalupka, K., Eberhardt, F., Perona, P.: Causal feature learning: an overview. Behaviormetrika **44**(1), 137–164 (2017)
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
6. Pearl, J.: Causality. Cambridge university press (2009)
7. Pocheville, A., Griffiths, P., Stotz, K.: Comparing causes – an information-theoretic approach to specificity, proportionality and stability. 15th Congress of Logic, Methodology, and Philosophy of Science (08 2015)
8. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)