Model Explainability and Causal Representation Learning: An Informal Literature Review 3

Cheng Guo
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093
c5guo@ucsd.edu

1 Introduction

Machine Learning can be seen as generating observed variables or output X based on a mixture of latent causal variables Z, or $X = g(Z)$ [10, 11, 17, 19]. The process of identifying Z in a DAG form is called causal-representation learning (CRL), which is particularly useful for researchers to explore model explainability in the age of LLMs. Most approaches are developed based on nonlinear ICA [4] and make various assumptions regarding the underlying distributions of data, including heterogeneous data with nonstationary time series [19], additive noise and distribution shifts [10], temporal and stationary condition [17]. In this review, I shall go over some of the CRL research and compare the differences in their assumptions and how the identifiability conclusion changes. I will also go over papers on how CRL is related to LLMs, including its relationships with the foundation model [11] and its duality with attention [18].

2 CRL with post-nonlinear ICA Approaches

Independent Component Analysis (ICA) has been proven applicable for linear representation learning [4]. Yet, recent researchers have generalized it to nonlinear cases that use time-contrastive learning (TCL) or variational autoencoder (VAE). They used methods beyond VAE and multi-layer perceptions (MLP) in identifying latent causal variables. In general, all these approaches have the following assumptions in the problem-setting. Let X be a d-dimensional vector representing the observations, which is generated based on a vector Z with n latent causal variables.

$$X = f(Z) + \epsilon, Z_i = g_i(\text{PA}(Z_i), u_i) + \epsilon_i \tag{1}$$

The first equation shows the observed variables X are a non-linear mixing of the latent variables Z with noise ϵ, where the function f is injective. The second equation shows each latent causal variable Z_i is generated from a function g of its parent PA(Z_i), latent factor u_i, and noise ϵ_i.

2.1 Multiple Distributions

Before digging into research using interventions to cause a distribution shift, it is worth noting that CRL can be done in a non-parametric setting with heterogeneous data and nonstationary time series without any interventions [19]. The researchers utilized sparsity constraint and graphed all latent causal variables in a Markov network. They assumed sufficient changes where all values of the latent factor u_i are linearly independent. They concluded that in this setting, the following statements are true between true hidden causal variable Z_i and Z_j in the Markov network M_Z and estimated hidden variables $\hat{Z}_k$ and $\hat{Z}_l$ in the Markov network $M_{\hat{Z}}$.

- When $\hat{Z}_k$ and $\hat{Z}_l$ are not adjacent in $M_{\hat{Z}}$, then they will not co-exist in the function for any true hidden causal variable Z_i.

- When Z_i and Z_j are adjacent in M_Z, then at most one of them will be a function of an estimated causal variable $\hat{Z}_k$.

Given the above two conclusions, they claimed that *any true hidden causal variables can be recovered up to a component-wise transformation as long as it has no neighbors that are adjacent to all of its other neighbors*, so no intimate neighbors. One remark of this conclusion is that it is not necessary that complex Markov network lead to more undetermined identifications, but as long as a variable satisfies the italic statement above, they are able to recover it to a component-wise transformation. For other variables, they are able to recover it as a function of all its intimate neighbors [19].

The researchers generalized the identifiability theory with the following two assumptions: in the DAG of all true latent causal variables, as long as two variables are adjacent, they are dependent, and any two parents of a collider are dependent. The Markov network is the moralized graph of the DAG only when the above two assumptions are held. They tested their identifiability theory on synthetic data with various causal structures and verified that they can recover latent variables up to small indeterminacies [19].

2.2 Interventions and Distribution Shifts

One prevalent method of studying CRL is through interventions that lead to distribution shifts so researchers can conclude which latent variables are related. If reflected on a causal graph, hard interventions remove all the edges to a certain causal variable and replace them with a different set of causal mechanisms [14]. On the other hand, soft interventions use a different causal mechanism and do not remove dependencies between any causal variables and their parents [14]. Regarding soft interventions, researchers have proposed latent linear Gaussian models [9] and latent polynomial models [8] and encompassed them together in a latent additive noise model [10]. One thing worth noticing is that in the assumed equation for Z_i in 1, the u_i would be a variable characterizing distribution shifts. Their identifiability result used the three assumptions developed originally in nonlinear ICA [4] and added a fourth assumption constraining the function g_i to define a boundary for beneficial distribution shifts. For example, say Z_1 is a parent of Z_2, then Z_2 can be written as $g(u)Z_1 + \epsilon_2$. if we replace $g(u)$ with $g'(u) + b$, then the constant term bZ_1 will be constant across u, which will not make Z_2 identifiable, so the fourth assumption ensures that there are no non-zero constant term in the function g [10].

The researchers also provided a partial identifiability result showing that either the fourth assumption is satisfied or Z_i is a root node, then we can recover true Z_i to Z'_j such that Z'_j is a linear transformation of Z_i [10]. They have also assumed that the function g should be invertible [10]. Beyond VAE, they used MLP for better flexibility and tested their results on synthetic data. The Mean of the Person Correlation Coefficient (MPC) shows that their identified causal variables are affirmed with true latent causal variables [10].

2.3 Temporal and Stationary Condition

When attempting to recover time-delayed latent causal variables, the researchers proposed the Temporally Disentangled Representation Learning (TDRL) framework [17], which is also applied in a non-parametric setting like the multiple distribution paper [19]. It utilizes a modular representation of distribution changes allowing it to recover causal relationships over time. The main technique is an extended sequential VAE that models distribution shifts, similar to the one mentioned in [10], but that one used MLP beyond VAE. It leverages partitioned estimated latent subspaces for capturing distribution shifts and validates it on real-world datasets in latent variable recovery with Mean Correlation Coefficient (MCC) [17].

3 Relating to LLMs

New research has been conducted in the age of LLMs to promote a causal approach beyond the foundation model. The researchers have unified the foundation approach with the causal approach by asserting definitions of the learned concepts [11]. A concept C with a projector matrix A, in short, is a linear transformation of the latent variables vector Z (as defined in section 2) to a valuation b with d_C dimensions, $AZ = b$, and the dimension of this concept C is defined by the dimension of b, so

$\dim(C) = d_C$. An atomic concept, or atoms is any concept with dimension $d_C = 1$. In this paper, the researchers further restrict function f to be differentiable. The identifiability result shows that if a concept is identified, then even if it is a nonlinear representation of the data, we can still recover them to linearity [11]. While concluding the final result for concept identifiability, the researchers assumed the Gaussian distribution of data and made two more assumptions on environmental diversity [11], much similar to the requirements of a sufficient number of environments for identifiability mentioned in [19] but with multiple distributions. In this paper, the researchers validated results by applying them to the alignment framework of LLMs and used LLaMA with the TruthfulQA dataset [6] and compared it with Inference-Time Intervention [5], another popular technique for improving alignment.

When it comes to the popular transformer architecture in LLMs, researchers discovered a duality between causal inference and attention [18]. In any causality exploration, it is important to have a balanced distribution of covariates. The self-attention mechanism, as the last layer in a neural network, effectively solves the covariate balancing problem [18]. Leveraging such duality is coined with CInA, which is proven to outperform other causal methods in out-of-distribution (OOD) generalization [18]. It also performed well on other empirical studies with datasets like IHDP [12] (resampled) and ACIC [13] in evaluating treatment effect.

4 Benchmarks and Datasets for CRL

After our meeting on Monday, I researched some existing datasets for CRL. I discovered that most existing CRL datasets are for computer vision or robotics tasks, including detecting the action taken between two photos or how would the classification results be if the researchers took certain interventions in a photo. The only text dataset related to CRL is the CEBaB dataset I mentioned last time [1], so this time, I will focus on the following computer vision and robotics datasets.

The first one is 3DIdent, a dataset with 250,000 teapot images each rendered with a 10-dimensional latent [21], which contains information on position, rotation, spotlight, and color hue. The researchers trained a convolutional feature encoder of a ResNet-18 architecture with an additional fully connected layer with a LeakyReLU hidden activation, and the performance of contrastive learning is as good as a supervised approach. Later, researchers extended it with 6 additional classes: hare, dragon, cow, armadillo, horse, and head to form the Causal3DIdent dataset [15]. They used a similar structure to test the effect of data augmentation and latent transformation on the encoder, and they concluded that augmentation and latent transformation have similar effects on a group of latents, where shifting some groups of features may cause the model to drop other aspects in style or content.

Beyond factor identification, there are also datasets for action detection. A dataset for robot manipulation is CausalWorld [2], which contains pairs of synthetic images of the current view and the goal view of 8 tasks surrounding blocks, like pushing and picking. A similar one is the Causal Triplet dataset [7], which contains the before and after pictures of daily objects on 7 actions. They are both sufficient for transfer learning tasks and achieved high accuracy for intervention classification. Another dataset on robot manipulation is the CausalCircuit dataset [3] developed by Qualcomm.

As mentioned in research with nonlinear ICA approaches [10, 11, 17, 19], researchers often use VAE as a framework to disentangle causal factors. Beyond VAE, researchers developed CausalVAE by introducing a new layer that describes a structural causal model (SCM) to the previous VAE structure [16]. They tested it on the following 4 datasets for intervention experiments and evaluated their maximal information coefficient (MIC) and total information coefficient (TIC).

- **Pendulum:** A synthetic dataset contains images of pendulums with 3 entities (pendulum, light, and shadow) and 4 concepts (pendulum angle, light position, shadow length, and shadow position).

- **Flow:** A synthetic dataset contains images of a ball and water in a container with a hole with 4 concepts (ball size, water height, hole, and water flow).

- **CelebA(Beard):** The real-world dataset contains 200k human face images, which include 4 concepts (age, gender, beard, bald). The researchers mainly used it for intervention experiments.

- **CelebA(Smile):** The real-world dataset contains 200k human face images, which include 4 concepts (gender, smile, eyes open, mouth open). The researchers mainly used it for intervention experiments.

Beyond the 4 datasets used in the CausalVAE paper, researchers created the shadow dataset, which contains object images with either Sunlight or Pointlight along with 8 other concepts like object shape or light color. They also modified the Beard and Smile datasets from CelebA to make the data distribution of factor pairs aligned with the original proposed causal graph [20].

References

[1] Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior, 2022.

[2] Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning, 2020.

[3] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[4] Aapo Hyvarinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning, 2023.

[5] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2023.

[6] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[7] Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning, 2023.

[8] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change, 2023.

[9] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models, 2023.

[10] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent neural causal models, 2024.

[11] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models, 2024.

[12] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2017.

[13] Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis, 2018.

[14] Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions, 2023.

[15] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style, 2022.

[16] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder, 2023.

[17] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26492–26503. Curran Associates, Inc., 2022.

[18] Jiaqi Zhang, Joel Jennings, Cheng Zhang, and Chao Ma. Towards causal foundation model: on duality between causal inference and attention, 2023.

[19] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting, 2024.

[20] Jiageng Zhu, Hanchen Xie, Jianhua Wu, Jiazhi Li, Mahyar Khayatkhoei, Mohamed E. Hussein, and Wael AbdAlmageed. Shadow datasets, new challenging datasets for causal representation learning, 2023.

[21] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process, 2022.