Causal Discovery and Inference through Next-Token Prediction

Eivinas Butkus (eb3407@columbia.edu)
Department of Psychology,
Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA
NSF AI Institute for Artificial and Natural Intelligence (ARNI)

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Departments of Psychology, Neuroscience and Electrical Engineering,
Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA
NSF AI Institute for Artificial and Natural Intelligence (ARNI)

Abstract

Some argue that deep neural networks are fundamentally *statistical* systems that fail to capture the causal generative processes behind their training data. Here we demonstrate that a GPT-style transformer trained for next-token prediction can simultaneously discover instances of linear Gaussian structural causal models (SCMs) and learn to answer counterfactual queries about them. First, we show that the network generalizes to counterfactual queries about SCMs for which it saw *only* strings describing noisy interventional data. Second, we decode the implicit SCM from the network's residual stream activations and use gradient descent to intervene on that "mental" SCM with predictable effects on the model's output. Our results suggest that neural networks trained using statistical prediction objectives on passively observed data may nevertheless discover and learn to use causal models of the world.

Keywords: deep neural networks; causal inference; large language models, emergent representations

Introduction

Pearl (2018) has argued that deep neural networks (DNNs) trained using prediction objectives will always be fundamentally limited in their causal reasoning capacities. The argument rests on Pearl's Causal Hierarchy (PCH) (Bareinboim et al., 2022), also known as the "Ladder of Causation" (Pearl & Mackenzie, 2018). PCH describes three levels of causal capabilities— associational, interventional, and counterfactual—where queries regarding higher levels are generally underdetermined by data from lower levels. According to Pearl (2018), DNNs can only master associations because they are trained in a "statistical mode" using prediction objectives on passive observations.

We appreciate PCH's theoretical implications, but disagree with the *further* claim that DNNs trained on prediction objectives cannot go beyond the level of associations. Note that *passively* observed data does not necessarily mean *observational* data. For instance, natural language used to train large language models (LLMs) describes interventions and causal inferences (Fig. 1). LLMs may discover causal structure—the mechanisms that remain invariant under local interventions (Schölkopf et al., 2021; Pearl, 2009)—and learn causal inference engines to *predict* the next token on such strings.

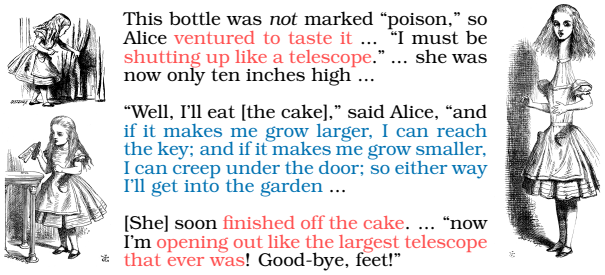

Figure 1: Natural language includes descriptions of interventions and causal inference. Examples from *Alice's Adventures in Wonderland* (Carroll, 1865).

Here we test this hypothesis empirically. We generate text (in a made-up simple language) describing interventional data and counterfactual inferences from a constrained class of linear Gaussian structural causal models (SCMs). Given snippets of this text, we train a GPT-style transformer model to predict the next token.

We found that the trained model: 1) could answer counterfactual queries about SCMs that only had noisy interventional data; 2) developed explicit internal representations of the underlying SCMs that we could manipulate with predictable effects on answers to new causal queries.

Methods

Linear Gaussian SCMs. We consider a constrained class of linear Gaussian structural causal models (SCMs) with four variables V_1, V_2, V_3, V_4, each taking the following form:

$$U_i \sim \mathcal{N}(0, 0.1)$$
$$V_j := U_j + w_{jj} + \sum_{i<j} w_{ij} V_j \quad \text{where } \forall i,j : w_{ij} \in \{-1, 0, 1\}$$

where variable V_j is a linear combination of the background variable U_j, bias term w_{jj} and a weighted combination of parent values. We generate all possible 4-variable SCMs with weights $w_{ij} \in \{-1, 0, 1\}$, resulting in 59,049 SCMs (Fig. 2).

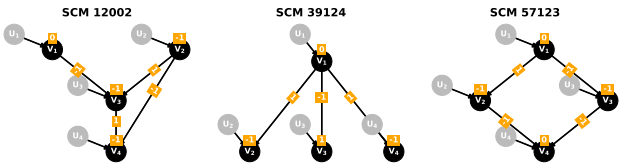

Figure 2: A few examples of the 59,049 unique SCMs.

Generating text strings from SCMs. We use the SCMs to generate two types of strings (Fig. 3). Each string begins with a token that describes its type (`DATA` or `INFERENCE`), followed by the SCM index encoded using 4 letter tokens (e.g. `A R T Q` corresponds to SCM index 12002). Observation is indicated using `OBS [variable] [value]` sequence, while intervention is indicated `DO [variable] [value]` sequence. For `DATA` strings, we sample 0-2 interventions (with values $\sim \mathcal{U}[-5, 5]$) and record the *sampled* values under that intervention. So `DATA` strings provide only noisy samples of the underlying *interventional* distributions. For `INFERENCE` strings, we sample 1-2 observations (also $\sim \mathcal{U}[-5, 5]$), 0-2 interventions. We then record the analytical solution of *counterfactual* posterior means and standard deviations (SDs) for each variable (see Pearl (2009) for details on counterfactual inference). Numbers within $[-10, 10]$ are encoded using numerical tokens with one decimal point precision (e.g. `0.3`, `-8.5`); numbers outside of that range are encoded using `-INF` and `+INF` tokens. Each string ends with an `EOS` token.

Training transformer to predict the next token. We train a GPT-style transformer model with 12 layers, hidden size 512,

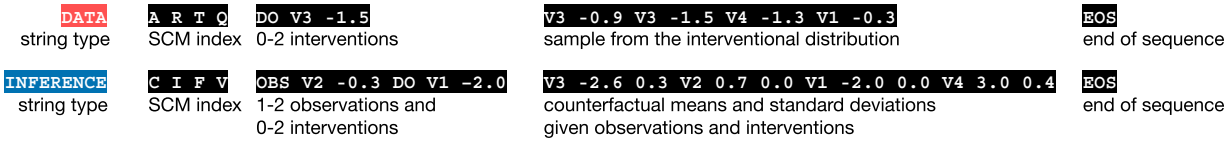

Figure 3: We generate two types of strings from the SCMs. DATA strings provide interventional data about the referenced SCM. INFERENCE strings provide examples of counterfactual inference.

8 attention heads of size 64, MLP size 2048, GELU activation function, and "Pre-LN" type layer normalization. We use the standard cross-entropy objective to train the model to predict the next token in generated text (Radford et al., 2018).

For most SCMs, the model sees both DATA and INFERENCE strings during training. However, we also randomly select 1,000 SCMs for which the model only ever sees DATA strings. A training epoch consisted of 10 DATA and 10 INFERENCE strings randomly generated per SCM (excluding INFERENCE strings for held out SCMs), resulting in around 1.2 million strings per epoch. We trained the model for 300 epochs.

Results

Transformer generalizes to SCMs with DATA strings only. We assessed causal inference capacity by asking the model to complete 10,000 strings of the following form: `INFERENCE [SCM index] [observations & interventions] [queried variable] [...]` with different SCM indices and causal queries. We then computed the mean absolute error (MAE) between *predicted* mean and SD for the queried variable (converting tokens to numerical values), and *analytically derived* result. We also considered a naive baseline that simply predicts the average mean and average SD for all queries:

SCM instances		Mean MAE	SD MAE
with DATA & INFERENCE strings	baseline	2.340 [2.306,2.375]	0.153 [0.152,0.154]
	model	**0.017** [0.015,0.019]	**0.000** [0.000,0.000]
with DATA strings only	baseline	2.237 [1.979,2.501]	0.156 [0.147,0.163]
	model	**0.016** [0.005,0.030]	**0.001** [0.000,0.004]

First, the trained model achieves near optimal performance predicting the counterfactual mean and standard deviation. Crucially, the model *generalizes* to SCM instances that only had DATA strings, ruling out the hypothesis that the model achieves causal capacities by simply memorizing the answers to all possible counterfactual queries. To sum up, a transformer trained to predict the next token can discover SCMs from interventional data and learn to answer counterfactual queries about those SCMs.

Overwriting "mental" SCMs using a linear decoder. We also wanted to investigate the internal representations that may underpin model's causal capacities. Inspired by (Li et al., 2023; Nanda, Lee, & Wattenberg, 2023), we trained linear decoders—mapping *residual activations* from each layer (at the last SCM index position) to three possible values of each weight $w_{ij} \in \{-1, 0, 1\}$. We considered only those SCMs that had both DATA and INFERENCE strings. We trained the decoders on activations from 57,049 SCMs, and report classification accuracy on held out 1,000 SCMs (Fig. 4).

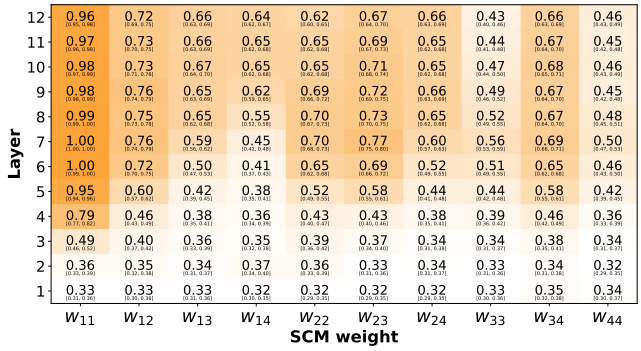

Figure 4: Linear decoder accuracy per layer and SCM weight.

Starting around layer 5, many weights of the SCM can be linearly decoded above chance (33%), suggesting that the trained transformer maps the (arbitrary) SCM index to a meaningful internal representation of the SCM.

Finally, we use the decoders to overwrite the internal SCM representation. In the example in Fig. 5, we input a query string and use gradient descent on the residual activations in layer 3 to flip the decoder prediction and change one weight in the represented SCM ($w_{12} = 0 \rightarrow 1$). After overwriting the activations, model prediction matches the expected analytical result (given the same change in the ground truth SCM), suggesting that the model is actually *using* this representation.

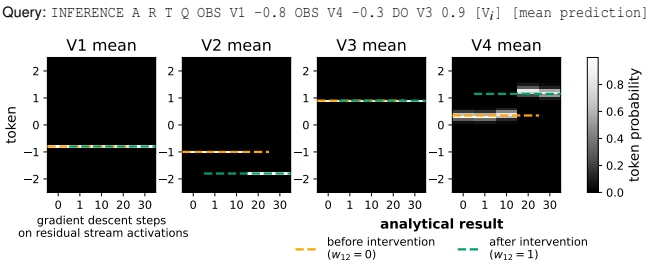

Figure 5: Intervening on internal representation of the SCM.

Conclusion

We found that a transformer model trained on next-token prediction can discover linear Gaussian SCMs from interventional data, and that it forms explicit representations of the SCMs which can be manipulated with predictable effects on model output. In future work, it is important to consider other SCMs that go beyond our toy setting. Nevertheless, our results challenge the strong claim that causal reasoning capacities cannot emerge through statistical prediction objectives.

Acknowledgments

We would like to thank Yushu Pan, Zhuofan Josh Ying, Elias Bareinboim, and members of the Causal Artificial Intelligence Lab at Columbia University for useful discussion and comments regarding this work.

References

Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl* (pp. 507–556).

Carroll, L. (1865). *Alice's Adventures in Wonderland*. London: Macmillan.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*.

Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*.

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

Pearl, J. (2018, January). *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution.* arXiv. Retrieved 2024-08-20, from http://arxiv.org/abs/1801.04016 (arXiv:1801.04016 [cs, stat])

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
(Publisher: San Francisco, CA, USA)

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634. (Publisher: IEEE)