Causal Inference with Graphical Neural Networks

Gwen Johnson Dylan Skinner Dallin Stewart
Jason Vasquez

April 19, 2024

1 Introduction

Neural networks are at the center of computer science research and have revolutionized our ability to model complex relationships with data. However, as these networks become increasingly pervasive in decision-making processes, understanding the causality behind their predictions becomes paramount. The distinction between correlation and causation is not merely philosophical; it's foundational for building robust, interpretable, and ethically sound AI systems.

Causal inference, a field deeply rooted in statistics and philosophy, offers a framework to disentangle cause-and-effect relationships from observed data. Traditionally, causal inference has been applied in fields such as epidemiology, economics, and social sciences. However, its integration into the domain of neural networks presents a promising frontier, with implications ranging from improving model interpretability to enhancing the fairness and accountability of AI systems.

In this paper, we explore the intersection of causal inference and neural networks, focusing on the application of Graph Neural Networks (GNNs) to causal modeling. We leverage the power of GNNs to learn representations of graph-structured data and perform a supervised classification task on the LUCAS0 dataset [3]. Our goal is to demonstrate the effectiveness of GNNs in capturing causal relationships within a graph and to provide insights into the interpretability and robustness of these models.

2 Related Work

The integration of causal inference with neural networks has garnered significant interest in recent years. Several approaches have been proposed to combine the strengths of these two fields, with a particular focus on leveraging the expressive power of neural networks to model causal relationships. Koch et al. discusses ongoing work to extend causal inference to settings where confounding is non-linear, time-varying, or encoded in text, networks, and images [4]. Also, Yuan et al. compared the performance of CNN's on causal data to previous methods [5].

3 Graph Neural Networks

Originally proposed in 2005 by Mori et al. [1], Graph Neural Networks (GNNs) are a class of neural networks that operate on graph-structured data. They have gained significant attention in recent years due to their versatility in handling various types of graph data, including social networks, citation networks, biological networks, and more.

Unlike traditional neural networks, which operate on grid-like structures such as images or sequences, GNNs are specifically designed to capture and leverage the structural information present in graphs. One of their key strengths lies in their ability to learn meaningful representations of nodes in a graph, which can then be used for various downstream tasks such as node classification, link prediction, and graph classification. Their ability to capture and model complex relationships in graph data makes them invaluable tools for exploring and understanding real-world phenomena represented in graph form.

In our project, we are utilizing the power of GNNs to perform a supervised classification task. This means we are training a GNN on a labeled, tabular dataset, where each graph instance (or node) is associated with a target label.

4 do Operator

The **do** operator is a way to represent interventions in a causal model. It is a way to represent the effect of an intervention on a variable. As an example, consider the following model involving smoking.

If a person's fingernails (N) have turned yellow, this implies a higher probability that they are a heavy smoker (S) and hence have a higher probability of developing lung cancer (C). But, simply dyeing a persons fingernails yellow does not impact their probability of developing lung cancer.

So, in terms of **do** calculus, we can denote the process of setting a variable N to have a value *yellow* by $\mathbf{do}(N = \textit{yellow})$. We note that

$$P(C \mid N = \textit{yellow}) \neq P(C \mid \mathbf{do}(N = \textit{yellow})).$$

With this in mind, we now define the **do** operator.

Theorem 4.1 ([2]) *In a causal diagram Γ with nodes $X_1, \ldots, X_n$ and joint distribution $P(X_1, \ldots, X_n)$, the result of doing $X_i = x_i$ on the joint distribution is*

$$P(X_1, \ldots, X_n \mid \boldsymbol{do}(X_i = x_i)) = \frac{P(x_1, \ldots, x_n)}{P(x_i \mid \mathrm{par}(x_i))} = \prod_{j \neq i} P(x_j \mid \mathrm{par}(x_j)).$$

In this, we have $\mathrm{par}(x_i)$ represent values of the parent nodes of $\mathrm{PAR}(X_i)$ of X_i in Γ. The probabilities on the right hand side of the above equation are what we call *preintervention.* This means they use the original probabilities from the original model before doing $X_i = x_i$.

It is important to note that the above equation is how we calculate the probability of several events happening given one event has happened. What if we want to get the probability of a single event happening, given we do a single event? That leads to the following corollary.

Corollary 4.1.1 *If X and Y are random variables in a causal diagram Γ and* $\mathrm{PAR}(X)$ *are the parents of X, then*

$$P(y \mid \boldsymbol{do}(x)) = \sum_{\mathrm{par}} \frac{P(x,\, y,\, \mathrm{par})}{P(x \mid \mathrm{par})},$$

where the sum runs over all values par *that the variables* $\mathrm{PAR}(X)$ *can take. If X has no parents, then*

$$P(y \mid \boldsymbol{do}(x)) = \frac{P(x,\, y)}{P(x)} = P(y \mid x).$$

Let us now consider a basic example to see how this works. Consider the following causal diagram in Figure 1.

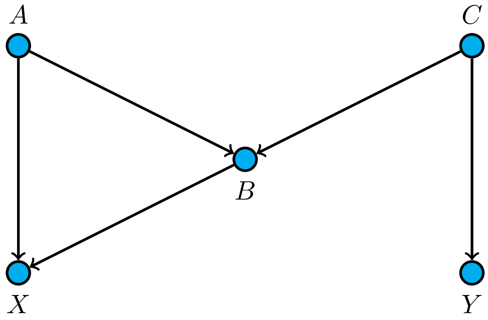

Figure 1: Basic causal diagram. Note it is in the form of a directed acyclic graph (DAG).

In this diagram, we can see that A and C are both parents of B. So, for any values of x and b, Corollary 4.1.1 tells us that

$$P(X = x \,|\, \mathbf{do}(B = b)) = \sum_{\mathrm{par}(b)} \frac{P(x,\, b,\, \mathrm{par}(b))}{P(b \,|\, \mathrm{par}(b))}$$

which, written out, is

$$\sum_{\mathrm{par}(b)} \frac{P(x,\, b,\, \mathrm{par}(b))}{P(b \,|\, \mathrm{par}(b))} = \sum_a \sum_c \frac{P(X = x\,,\, A = a\,,\, B = b\,,\, C = c)}{P(B = b \mid A = a,\, C = c)}.$$

By dependence of nodes only on their parents and the rules of probability, this turns into

$$\sum_a \sum_c \frac{P(X = x \mid A = a\,,\, B = b)P(B = b \mid A = a,\, C = c)P(A = a)P(C = c)}{P(B = b \mid A = a,\, C = c)},$$

which simplifies to

$$\sum_a \sum_c P(X = x \mid A = a\,,\, B = b)P(A = a)P(C = c).$$

Since there is only one instance where we are considering the probability with respect to c, we can simplify this to

$$\sum_a P(X = x \mid A = a\,,\, B = b)P(A = a),$$

which is our final answer.

While this introduction to the **do** operator might feel a bit abstract, it is the foundation of all current research in causal inference.

5 Data

For our project, we used the LUCAS0 dataset [3], which is a toy data set generated artificially by causal Bayesian networks with binary variables. The LUCAS0 dataset is a DAG with 11 nodes and 2000 training samples, where the DAG is represented as in Figure 2.

Each node of the graph was is associated with a specific conditional probability that the creators of the dataset used to generate the data. These probabilities can be found in Table 2 in Appendix A.

6 do Calculus Result stuff (idk where this goes)

Let's consider the following conditional probabilities with **do** operators applied:

$$\begin{aligned}
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{YF} = \mathrm{T})),\\
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{PP} = \mathrm{T})),\\
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{A} = \mathrm{T})),\\
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{AD} = \mathrm{T})),\\
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{CA} = \mathrm{T})),
\end{aligned} \tag{1}$$

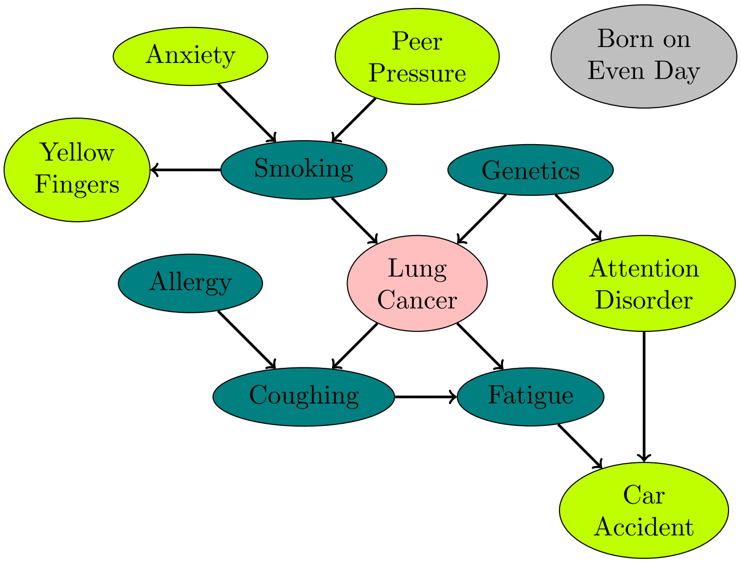

Figure 2: Basic causal diagram. Note it is in the form of a directed acyclic graph (DAG). Our target variable is shaded in **pink**, and the nodes in **teal** constitute the Markov blanket of the target variable.

where LC is lung cancer, YF is yellow fingers, PP is peer pressure, A is anxiety, AD is attention disorder, and CA is car accident. We note that we chose this probability because it does not involve any variables that are in the Markov blanket of our target variable. Thus, we should expect that there is very little predictive power in this probability.

Plugging forward with **do** calculus, we can do one as an example. We write $P(\mathrm{LC}=\mathrm{T} \mid \mathbf{do}(\mathrm{YF}=\mathrm{T}))$ (using corollary 4.1.1) as

$$P(\mathrm{LC}=\mathrm{T} \mid \mathbf{do}(\mathrm{YF}=\mathrm{T})) = \sum_{s\in\{T,F\}} \frac{P(\mathrm{YF}=\mathrm{T},\, \mathrm{LC}=\mathrm{T},\, \mathrm{S}=s)}{P(\mathrm{YF}=\mathrm{T}\,|\,\mathrm{S}=s)}, \tag{2}$$

where S is smoking. Since the Lung Cancer node has Smoking and Genetics as parents, and Smoking has parent Anxiety and Peer Pressure (see Figure 2), we can write Equation 2 to

$$\sum_{S,G,A,PP} \frac{P(\mathrm{YF}=\mathrm{T} \mid \mathrm{S})P(\mathrm{LC}=\mathrm{T} \mid \mathrm{S},\, \mathrm{G})P(\mathrm{G})P(\mathrm{S} \mid \mathrm{A},\, \mathrm{PP})P(\mathrm{A})P(\mathrm{PP})}{P(\mathrm{YF}=\mathrm{T} \mid \mathrm{S})}. \tag{3}$$

where each $\mathrm{S}, \mathrm{A}, \mathrm{G}, \mathrm{PP}$ is in terms of $\{T, F\}$.

Since not every term is in terms of every variable, we can re-write the Equation 3 as

$$
\begin{aligned}
&= \sum_{S} \frac{P(\mathrm{YF} = T \,|\, S)}{P(\mathrm{YF} = T \,|\, \mathrm{S})} \Bigg(\sum_{G} P(\mathrm{LC} = \mathrm{T} \,|\, \mathrm{S}, \mathrm{G}) P(\mathrm{G}) \Bigg) \Bigg(\sum_{A} P(\mathrm{A}) \Big(\sum_{\mathrm{PP}} P(\mathrm{S} \,|\, \mathrm{A}, \mathrm{PP}) P(\mathrm{PP}) \Big) \Bigg) \\
&= \sum_{S} \Big(\sum_{G} P(\mathrm{LC} = \mathrm{T} \,|\, \mathrm{S}, \mathrm{G}) P(\mathrm{G}) \Big) \Bigg(\sum_{A} P(\mathrm{A}) \Big(\sum_{\mathrm{PP}} P(\mathrm{S} \,|\, \mathrm{A}, \mathrm{PP}) P(\mathrm{PP}) \Big) \Bigg).
\end{aligned}
\tag{4}
$$

Using the probabilities from Table 2 on Equation 4 we get

$$P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{YF} = \mathrm{T})) = \underline{\hspace{3cm}} \text{ fix this!!}$$

Performing similar calculations on the other probabilities in Equation 1, we get

do Statements	*Probabilities*
$P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{YF} = \mathrm{T}))$	
$P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{PP} = \mathrm{T}))$	
$P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{A} = \mathrm{T}))$	
$P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{AD} = \mathrm{T}))$	
$P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{CA} = \mathrm{T}))$	

Table 1: The resulting probabilities of the **do** statements found in Equation 1. For a more detailed workthrough of these probabilities, see Appendix B.

7 Methodology

8 Experiments

8.1 Dataset Description

8.2 Experimental Setup

8.3 Results

9 Discussion

10 Conclusion

Acknowledgments

References

[1] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pages 729–734 vol. 2, 01 2005.

[2] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[3] Research Group for Lung Cancer Analysis. Lucas (lung cancer simple set) dataset. ETH Zurich Causality and Machine Learning Group, 2020. The LUCAS dataset contains CT scan images and annotations for lung cancer detection.

[4] Bernard J. Koch, Tim Sainburg, Pablo Geraldo Bastías, Song Jiang, Yizhou Sun, and Jacob Foster. Deep Learning for Causal Inference. UCLA Department of Sociology, April 2023.

[5] Ye Yuan, Xueying Ding, and Ziv Bar-Joseph. Causal inference using deep neural networks. arXiv preprint arXiv:2011.12508, 202

Appendix

A Creating the LUCAS0 Dataset

As described in the Data section, the LUCAS0 dataset is a toy dataset generated artificially by causal Bayesian networks with binary variables. The dataset consists of 11 nodes and 2000 training samples. The causal diagram for the LUCAS0 dataset is shown in Figure 2.

On the website, the authors mention that each node of the graph is associated with conditional probabilities which were used to generate the data. These probabilities are found in Table 2.

B Conditioning Outside Markov Blanket Results

B.1 $P(\mathbf{LC = T \mid do(PP = T)})$

We have $P(\mathrm{LC} = \mathrm{T} \,|\, \mathbf{do}(\mathrm{PP} = \mathrm{T}))$ yields

$$\begin{aligned}
&= P(\mathrm{LC} = \mathrm{T},\, \mathrm{PP} = \mathrm{T}) \\
&= P(\mathrm{PP} = \mathrm{T}) \left(\sum_{s,g,an} P(\mathrm{LC} = T \,|\, \mathrm{S} = s,\, \mathrm{G} = g) P(\mathrm{G} = g) P(\mathrm{An} = an) \right) \\
&= P(\mathrm{PP} = \mathrm{T}) \Big(\sum_{an} P(\mathrm{An} = an) \Big) \left(\sum_{s,g} P(\mathrm{LC} = T \,|\, \mathrm{S} = s,\, \mathrm{G} = g) P(\mathrm{G} = g) \right) \\
&= P(\mathrm{PP} = \mathrm{T}) \left(\sum_{s,g} P(\mathrm{LC} = T \,|\, \mathrm{S} = s,\, \mathrm{G} = g) P(\mathrm{G} = g) \right) \\
&= \underline{\hspace{3cm}}.
\end{aligned}$$

B.2 $P(\mathbf{LC = T \mid do(An = T)})$

We have $P(\mathrm{LC} = \mathrm{T} \,|\, \mathbf{do}(\mathrm{An} = \mathrm{T}))$ yields

$$\begin{aligned}
&= P(\mathrm{LC} = \mathrm{T},\, \mathrm{An} = \mathrm{T}) \\
&= P(\mathrm{An} = \mathrm{T}) \left(\sum_{s,g,p} P(\mathrm{LC} = T \,|\, \mathrm{S} = s,\, \mathrm{G} = g) P(\mathrm{G} = g) P(\mathrm{S} = s \,|\, \mathrm{An} = T, \mathrm{PP} = p) P(\mathrm{PP} = p) \right) \\
&= \underline{\hspace{3cm}}.
\end{aligned}$$

B.3 $P(\mathbf{LC = T \mid do(AD = T)})$

We have $P(\mathrm{LC} = \mathrm{T} \,|\, \mathbf{do}(\mathrm{AD} = \mathrm{T}))$

$$
\begin{aligned}
&= \sum_{g} \frac{P(\mathrm{LC} = \mathrm{T},\ \mathrm{AD} = \mathrm{T},\ \mathrm{G} = g)}{P(\mathrm{AD} = T \mid G = g)} \\
&= \sum_{g,s,an,p} \frac{P(\mathrm{LC} = \mathrm{T},\ \mathrm{AD} = \mathrm{T},\ \mathrm{G} = g,\ \mathrm{S} = s,\ \mathrm{AN} = an,\ \mathrm{PP} = p)}{P(\mathrm{AD} = T \mid G = g)} \\
&= \sum_{g,s,an,p} \frac{P(\mathrm{LC} = T \,|\, \mathrm{S} = s, \mathrm{G} = g) P(\mathrm{G} = g) P(\mathrm{S} = s \,|\, \mathrm{AN} = an,\ \mathrm{PP} = p)}{}.
\end{aligned}
$$

Conditional Probabilities	
Conditional	*Probability*
P(Anxiety=T)	0.64277
P(Peer Pressure=T)	0.32997
P(Smoking=T | Peer Pressure=F, Anxiety=F)	0.43118
P(Smoking=T | Peer Pressure=T, Anxiety=F)	0.74591
P(Smoking=T | Peer Pressure=F, Anxiety=T)	0.8686
P(Smoking=T | Peer Pressure=T, Anxiety=T)	0.91576
P(Yellow Fingers=T | Smoking=F)	0.23119
P(Yellow Fingers=T | Smoking=T)	0.95372
P(Genetics=T)	0.15953
P(Lung cancer=T | Genetics=F, Smoking=F)	0.23146
P(Lung cancer=T | Genetics=T, Smoking=F)	0.86996
P(Lung cancer=T | Genetics=F, Smoking=T)	0.83934
P(Lung cancer=T | Genetics=T, Smoking=T)	0.99351
P(Attention Disorder=T | Genetics=F)	0.28956
P(Attention Disorder=T | Genetics=T)	0.68706
P(Born an Even Day=T)	0.5
P(Allergy=T)	0.32841
P(Coughing=T | Allergy=F, Lung cancer=F)	0.1347
P(Coughing=T | Allergy=T, Lung cancer=F)	0.64592
P(Coughing=T | Allergy=F, Lung cancer=T)	0.7664
P(Coughing=T | Allergy=T, Lung cancer=T)	0.99947
P(Fatigue=T | Lung cancer=F, Coughing=F)	0.35212
P(Fatigue=T | Lung cancer=T, Coughing=F)	0.56514
P(Fatigue=T | Lung cancer=F, Coughing=T)	0.80016
P(Fatigue=T | Lung cancer=T, Coughing=T)	0.89589
P(Car Accident=T | Attention Disorder=F, Fatigue=F)	0.2274
P(Car Accident=T | Attention Disorder=T, Fatigue=F)	0.779
P(Car Accident=T | Attention Disorder=F, Fatigue=T)	0.78861
P(Car Accident=T | Attention Disorder=T, Fatigue=T)	0.97169

Table 2: Conditional probabilities for the nodes in the LUCAS0 dataset.