HOKEY POKEY CAUSAL DISCOVERY: USING DEEP LEARNING MODEL ERRORS TO LEARN CAUSAL STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

While machine learning excels at learning predictive models from observational data, learning the causal mechanisms behind the observed phenomena presents the significant challenge of distinguishing true causal relationships from confounding and other potential sources of spurious correlations. Many existing algorithms for the discovery of causal structure from observational data rely on evaluating the conditional independence relationships among features to account for the effects of confounding. However, the choice of independence tests for these algorithms often rely on assumptions regarding the data distributions and type of causal relationships. To avoid these assumptions, we develop a novel deep learning approach, dubbed the Hokey Pokey model, to indirectly explore the conditional dependencies among a set of variables by rapidly comparing predictive errors given different combinations of input variables. We then use the results of this comparison as a predictive signal for causal relationships among the variables. We conduct rigorous experiments to evaluate model robustness and generalizability using generated datasets with known underlying causal relationships and analyze the capacity of model error comparisons to provide a predictive signal for causal structure. Our model outperforms commonly used baseline models (PC and GES) and is capable of discovering causal relationships of different complexity (graph size, density and structure) in both binary and continuous data.

1 INTRODUCTION

The ability to learn causality is considered a significant component of human-level intelligence and can serve as one of the foundations of AI (Pearl, 2018; Bengio, 2019). Understanding the causal relationships among variables is a fundamental task that spans a broad range of disciplines including social science (Wu et al., 2010), economics (Chen et al., 2007), medicine (Cheek et al., 2018; Shen et al., 2020), and biology (Zhang et al., 2014). While properly controlled experimentation represents the most robust method of determining such relationships (Pearl, 2009), such methods can not always be applied due to cost, time, ethical considerations, or other constraints. Therefore, significant effort has been directed towards the task of causal discovery, where such relationships are inferred from observational datasets in the absence of experimental interventions. Many algorithms have been developed to address the challenges of this task, including those which rely on observations of conditional independence relationships, an approach designed to account for the effects of confounding among the observed variables (Spirtes et al., 2000). However, these algorithms often need to make assumptions about the generative causal structure and relational functions of the underlying dataset, which are impossible to know a priori, and the choice of independence test can require significant domain knowledge (Shah & Peters, 2018). To work toward the development of a more general method, we aim to leverage the flexibility of deep learning models to explore conditional relationships within observational data. In particular, we aim to observe which input variables improve predictive performance conditioned on the inclusion of other input variables. However, to directly discover all such conditional relationships would require training multiple different models to cover all possible feature combinations, which becomes intractable for large numbers of features.

In this work, we develop a novel approach for causal discovery which indirectly explores the conditional relationships in observational data without having to separately train models for each

combination of features or each possible causal structure. To do this, we develop a two stage procedure which we refer to as the Hokey Pokey (HP) approach. In the first stage, we randomly obscure information from different features in the input data during model training by applying a form of dropout to the input layer. This allows the model to learn how to predict each variable with multiple different combinations of the other variables. However, the model performance will differ depending on which inputs are obscured. We compare the performance of the model under different masking conditions to determine whether additional knowledge of a specific feature aids the predictive performance. In the second stage of the approach, we train a second model to infer causal relationships among the variables by interpreting the error patterns produced by the first model.

2 Causal Discovery from Observational Data

There are a wide variety of different methods which have been developed to address the task of causal discovery. The approaches can be broadly categorized into constraint-based methods (Spirtes et al., 2000; Yu et al., 2016), score-based methods (Chickering, 2002), and hybrid methods which incorporate aspects of both (Tsamardinos et al., 2006). Constraint-based models restrict the space of possible causal structures by placing constraints on the relationships using conditional dependencies among the variables, while score-based methods infer causal relationships based on optimization of a scoring function. Many traditional causal discovery algorithms rely on theoretical assumptions regarding the relationship of the data and the corresponding causal graph structure. Common assumptions include the Causal Faithfulness Assumption, the Causal Markov Assumption, and the Causal Sufficiency Assumption (Pearl, 2009). In addition, many methods rely on assumptions about the generative processes and functional forms of the observational dataset. For example, the CCDr (Aragam & Zhou, 2015) algorithm assumes the relationships between variables are linear and the variables are Gaussian distributed, while the LiNGAM (Shimizu, 2014) algorithm assumes linear relationships and non-Gaussian distributions. Due to their flexibility, deep learning models have the potential to generalize across different types of data without relying on these restrictive assumptions.

There have been previous deep learning-based approaches to determine causal relationships between pairs of variables (Louizos et al., 2017). Recent efforts have also applied deep learning to the problem of causal discovery among a set of variables. Goudet et al. (2018) leverage a score-based approach using Causal Generative Neural Networks (CGNN) to learn functional causal models, while Kalainathan et al. (2018) also use a generative approach but trains the model in an adversarial setting. Zhu & Chen (2019) leverage reinforcement learning to search the space of possible causal structures. Recent efforts have used deep learning to address specific challenges of causal discovery including applying attention-based convolutional neural networks for observational time-series data (Nauta et al., 2019), dealing with missing data by simultaneously imputing and learning the causal structure (Wang et al., 2020), and discovering causal signal in images (Lopez-Paz et al., 2017). Distributional shifts due to interventions have been used to infer causal relationships with Ke et al. (2019) using a learned dropout rate to represent the inferred causal relationships among input variables and Bengio et al. (2019) observing the adaptation rate of the model to interventional shifts.

To our knowledge our approach is the first to use comparisons of model errors to infer causal relationships. Additionally, most existing methods that use predictive deep learning models leverage individual models to predict each variable in the data. In contrast, we develop an approach where a single model is used to explore the prediction performance variation among all variables as a potential indicator of causal relationships which provides a potential efficiency benefit.

3 Approach

Our approach is inspired by constraint-based approaches which aim to establish causal relationships by observing the correlations among variables conditional on other variables. However, rather than explicitly evaluating the conditional independence relationships among variables, we develop a deep learning model training procedure designed to indirectly rely on such relationships among the variables in order to predict the existence of causal relationships. This method consists of two stages. In the first stage, we train a model to reconstruct an observation from itself when certain input features are randomly obscured. In the second stage, we observe the errors that are made by the first model when different sets of input features are obscured and train a second model to use these patterns to predict which edges are present in the causal graph which generated the data.

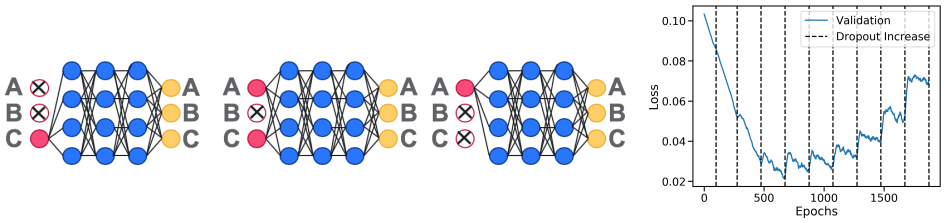

Figure 1: (Left) An illustrative schematic of the first stage HP model architecture, which uses the observational instances as both inputs and outputs with random sets of input variables obscured (crosses) during training using a dropout approach. (Right) An example learning curve for the HP model showing how the model adapts as the dropout rate on the input layer in increased step-wise during training.

3.1 Hokey Pokey Model

♫ *You put some features in, you take some features out* ♫

In the first stage, we apply a feed forward network to predict a data observation given a partially obscured version of that same observation as input, with different feature randomly obscured during training. This predictive setup, shown in the left panel of Figure 1, is designed to encourage the model to learn a flexible predictive approach that can adapt to the available inputs. In order to obscure the input information during training, we apply a mask to the variables that replaces the actual value of the masked variable with an obscured value. For binary variables, we replace the value of obscured variable with a value of 0.5. For continuous variables, we first standardize each variable (by subtracting the mean and scaling to unit variance) and then we replace each masked value with zero.

During training we apply a ramped dropout approach, starting without obscuring any features. During this phase, the model should learn to predict each output directly from its corresponding input feature. This is followed by the phase of training where each feature in the input has a probability p of being obscured where p is sampled from a uniform distribution from $p_{\min}$ to $p_{\max}$. $p_{\min}$ is fixed to 0.1, while $p_{\max}$ is increased in a step-wise fashion through training until it reaches 0.9. During this phase, the model should learn that when certain features are not available, as indicated by their assignment to the masking value, it needs to rely on other features for prediction.

While the model should become flexible to the available inputs, its performance should be reduced when information about causally relevant variables is removed. We aim to exploit the information about this reduction in performance to infer the causal relationships in the dataset. Therefore, once the model is fully trained, we generate a dataset of its predictive errors under a range of different masking conditions. During inference for this step, the feature masks are selected randomly from all possible masks with an equal probability for each mask. Multiple such masks are sampled for each observational instance in the test data such that direct comparisons can be made between the errors for a given instance under different masking conditions.

3.2 Causal Relationship Prediction

♫ *Then you shake it all about* ♫

We hypothesize that information about how the errors change under different masking conditions contains signal related to which variables are causally related to which other variables. However, determining the causal relationships directly from a large of set of observed error differences is not straightforward. In order to leverage this information for causal discovery, we develop a supervised fully connected deep learning model which takes pairs of input masks and the corresponding pairs of predictive errors as input and uses this information to predict the causal relationships among the variables. Because this model relies on supervised training, we generate a large collection of synthetic datasets using known causal graphs, which is described in detail in Section 4.1.

The structure of the inputs and outputs for this edge prediction model can be seen in Figure 2. Because a single pair of masks is unlikely to contain enough information to infer the full causal graph, we perform this prediction at the level of one causal relationship at a time. For example, we aim to predict whether an causal relationship exists between variables A and B given observations of how the error in B changes when the value of A is obscured or unobscured.

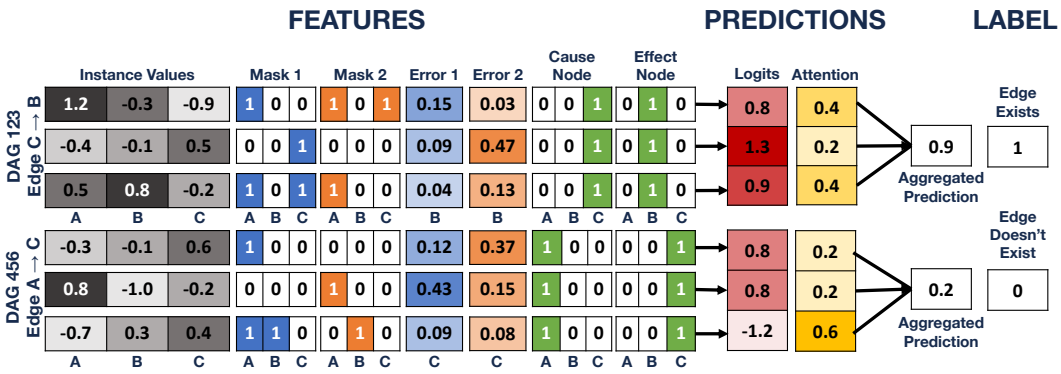

Figure 2: An illustration of the predictive signals and the batch structure used for the edge prediction model for a three node DAG example (with variables A, B, and C). The input features include the original observational variable values, a pair of input masks and their corresponding errors, and cause and effect indicators to specify the edge for which to make a prediction. Multiple mask-pair instances for the same edge of the same DAG are batched together, with the final prediction being a weighted mean across the multiple instances using learned attention weights.

The input features for the edge prediction model include a pair of masks and the corresponding predictive errors from the first-stage model. The features also include the original instance values from the observational dataset, which may provide a useful signal depending on the strategy used to obscure the inputs. If the masked value is set to the mean of the variable, as we do for the continuous, then the data instances further from the mean can likely be used more reliably to make inferences about the causal relationships. Finally, we include a set of one-hot-encoded features indicating the edge for which to make a prediction, e.g. from A to B. To train on datasets with different numbers of variables, we set values corresponding to variables that don't exist in a given dataset to zero.

In order to correctly predict whether one variable has a causal impact on another, the model needs to observe that the removal of the input variable increases the predictive error regardless of which other input variables are present. This corresponds to the detection of the conditional dependence of the variables. Therefore, rather than asking the model to predict the existence of a causal relationship based only on one observation, we provide the model with a set of multiple observations with differing mask configurations. We generate these batches by grouping together multiple mask-pair instances for the same causal edge of the same graph but which may differ in the original observational instance and the specific masks applied. The predictions for these multiple mask-pair observations are averaged to generate the final prediction. We explore the use of an attention mechanism to generate a weighted average that allows the model to focus more on certain mask-pair examples.

In order to select the mask-pairs to include as inputs for the prediction of a given causal relationship (e.g. A causes B), we include all pairs such that B is never included in the input (because predicting B from B is a trivial task) and such that A is obscured in one mask but not the other. This allows the model to observe the change in the predictive error for B when A is added or removed as a signal. The final trained model can output a prediction of the existence of a given edge for a batch of observations of mask-pair error differences. To construct a final prediction for the full causal structure of the observational dataset we must aggregate these multiple predictions. Therefore, we average the predicted probabilities of an edge across all batches that contained predictions of that edge.

4 Evaluation

We aim to evaluate the performance of the proposed approach on its ability to infer the existence of causal relationships among the variables in observational data. We evaluate the approach on two tasks of different complexities - the prediction of the causal skeleton and the prediction of the causal graph. The causal skeleton only describes whether two variables are causally linked but not the direction of that relationship, while the causal graph incorporates the direction of the causal relationships. We compare the performance of the HP method to established baseline algorithms for causal discovery including both a constraint-based approach, the PC algorithm (Spirtes et al., 2000) using a Gaussian independence test and a score-based approach, greedy equivalence search (GES) (Chickering, 2002).

4.1 DATA

To train the model and evaluate the robustness of our approach we generate a large collection of synthetic datasets with known causal structure. The use of synthetic data is common in causal discovery algorithm evaluation (Bengio et al., 2019; Kalainathan et al., 2018; Ke et al., 2019) due to the difficulty of establishing the ground truth causal structure for real-world data. To encourage generalizability, we generate 1974 random directed acyclic graphs (DAGs) with different properties using the *randDAG* function of the R *pcalg* library (Kalisch et al., 2012). We use DAGs ranging from 5 to 9 nodes, with 1 through 5 expected edges per node, and 8 different generation methods designed to target different graph topological properties. For each combination of DAG properties, we randomly generate 10 different individual graphs. Additionally, for each graph size we add every possible graph with only a single edge, as well as ten empty graphs with no edges. We divide these DAGs into training, validation, and test sets, respectively used to train the edge prediction model, perform hyper-parameter tuning, and evaluate edge prediction performance.

We use each randomly generated DAG to simulate a set of observational data that follows the given causal structure. To generate these simulations, we explore two different data types and corresponding functional forms for the causal relationships. Firstly, we generate a collection of datasets using linear Gaussian models with the edge weight and noise parameters drawn from uniform distributions for each individual edge. Secondly, we generate datasets of binary variables using Bayesian network models with the conditional probability tables randomly generated from a uniform distribution. For each observational simulation, we sample 1000 data points. Of these, 800 data points are used to train the first-stage models while the remaining 200 data points are used as the test set to generate the paired masks and errors for inputs into the edge prediction model. We apply 20 different input masks to each test instance to generate comparative pairs.

4.2 PARAMETER OPTIMIZATION

Because the approach uses a two stage procedure, the optimization of hyper-parameters for the first stage would ideally be based on the final performance in the edge prediction stage. However, performing this full optimization across many hyper-parameters would be prohibitively time and resource intensive. Therefore, we develop several heuristics for the hyper-parameter optimization of the first stage. This allows us to first optimize the parameters of the first-stage model, then generate the predictive error data for all input datasets, and finally optimize the edge prediction model on predictive error results generated from only a single dataset of errors.

To optimize the parameters of the first stage we focus on achieving models that have low average error for input masks that represent the true causal structure compared with the error for other input masks. We hypothesize that such models will provide the best predictive signal for the second stage model to infer the correct causal structure from the errors. We calculate the average level of error for each unique input mask and output variable across all predictions in the test set. Then we rank the input masks in order of average error to calculate four metrics - the ranking of the true casual mask, the correlation of a mask's ranking with the Euclidean difference of that mask to the true causal mask, the mean relative error of the true causal mask to the error of the best performing mask, and the correlation of the mask's relative error with the Euclidean difference of the mask to the true causal mask. We aim to identify model hyper-parameters that lead to better rankings and relative errors for the true causal mask and higher correlations of error and ranking with similarity to the true causal mask. We identify the model hyper-parameters that lead to the highest average value of these four metrics. This optimization is performed on a small representative sample of the input datasets.

For the edge prediction model, given a dataset of mask-pair errors spanning many different observational datasets, we perform hyper-parameter optimization using grid search, with parameters including the size of the hidden layers (128 or 256), the number of hidden layers (4 or 8), and the learning rate (0.001 or 0.0001). All edge prediction models used the Adam optimizer. The final edge prediction thresholds are selected by maximizing the F1 score on the validation data.

4.3 RESULTS

We evaluate the performance of the edge prediction task with the AUC score and F1 score, based on the binary edge existence labels for all possible causal relationships for test set DAGs. We calculate these scores both for each DAG individually as well as jointly for the full collection of edges for all

Table 1: Edge prediction F1 scores compared with several baseline approaches for the entire set of edges in the test set (Overall) and the average and standard deviation over each DAG individually for both directed and undirected edge prediction.

Data Type	Model	Undirected		Directed	
		Overall F1	DAG F1	Overall F1	DAG F1
Binary	PC	0.724	0.713 ± 0.261	0.479	0.477 ± 0.208
Binary	GES	0.503	0.457 ± 0.278	0.282	0.267 ± 0.187
Binary	HP	**0.810**	$\mathbf{0.844 \pm 0.139}$	**0.603**	$\mathbf{0.649 \pm 0.149}$
Continuous	PC	0.747	0.813 ± 0.200	**0.534**	0.571 ± 0.174
Continuous	GES	0.637	0.562 ± 0.265	0.344	0.317 ± 0.191
Continuous	HP	**0.773**	$\mathbf{0.842 \pm 0.145}$	0.497	$\mathbf{0.622 \pm 0.216}$

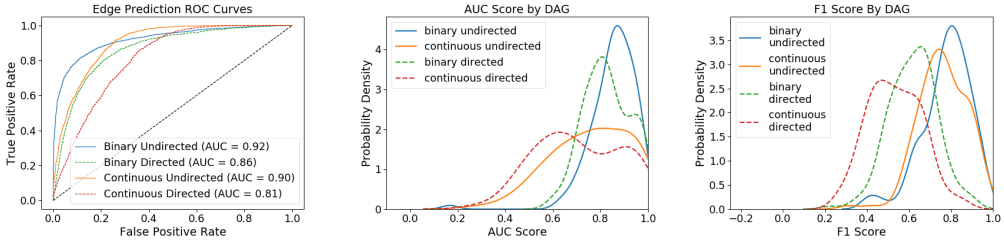

Figure 3: (left) ROC curve for the edge prediction model across all edges in the test set for binary and continuous data including the undirected and directed prediction tasks. The kernel density estimated distribution of DAG level AUC scores (middle) and F1 scores (right) across all test DAGs.

DAGs in the test set. In Table 1 we show the resulting overall and mean DAG-level F1 scores for both the binary and the continuous data in comparison with the baseline approaches. We find that the HP approach significantly outperforms the baseline approaches for the binary data for both the causal skeleton prediction (undirected) and the full DAG (directed) prediction. For continuous data, we find that HP outperforms the baselines for undirected and for the DAG-level directed predictions.

In the left panel of Figure 3, we show the overall ROC curves and corresponding AUC for each of the predictive tasks. These results show that the observed error difference patterns of the first-stage model provide a predictive signal for the existence of causal relationships among the observational variables. This figure also shows the kernel-density-estimated distributions of AUC scores and F1 scores across the DAGs in the test set. These results show that the binary datasets are a strong point for the HP approach, with high average AUC and F1 scores as well as a relatively tight distribution across DAGs. For the continuous data, the results show more variability with high average performance but with some DAGs on which the approach performs less well.

The advantage of the approach on the binary data compared with the continuous data may be related to the different methods used for masking the input variables for the two data types. For binary data we were able to use a mask value (0.5) that did not occur naturally within the actual values, while for the continuous data we replace the obscured input with the variable mean, which is likely to be similar to many observed values in the dataset. To mitigate this factor, we also explored using an extreme mask value that was outside the range of the variable distribution for the continuous data, but this leads to significantly reduced performance. We will perform further study on alternative approaches to address this issue.

To better understand model robustness and generalizability, we study model performance on datasets with different properties. In Figure 4, we show the performance of the approach as a function of properties of the causal structure and properties of the data, including the number of nodes in the underlying generative DAG, the density of the DAG, and the size of the causal effect of individual edges. For the continuous data, the causal effect size is defined as the ratio of the linear edge weight to the noise level, while for the binary data it is defined as the average difference in probability for the effect variable when the value of the cause variable is flipped. We find that the models perform slightly better on smaller DAGs compared with larger ones, but the trend is very shallow providing a hopeful indicator on the potential scalability of the method. We also find very strong performance

Table 2: Mean DAG-level edge prediction performance under ablation experiments with the change from the original model shown in parentheses and the largest performance reduction in bold.

Model Variant	Undirected DAG AUC	Undirected DAG F1	Directed DAG AUC	Directed DAG F1
Binary Original model	0.868	0.844	0.821	0.649
(a) Remove value features	0.867 (-0.001)	0.851 (+0.007)	0.828 (+0.007)	0.651 (+0.002)
(b) Remove attention	0.865 (-0.003)	0.844 (+0.000)	0.798 (-0.023)	0.631 (-0.018)
(c) Use error delta	0.867 (-0.001)	0.841 (-0.003)	0.782 (-0.039)	0.620 (-0.029)
(d) Use error sign	**0.740 (-0.128)**	**0.638 (-0.206)**	**0.654 (-0.167)**	**0.431 (-0.218)**
Continuous Original model	0.833	0.842	0.767	0.622
(a) Remove value features	0.828 (-0.005)	0.839 (-0.003)	0.764 (-0.003)	0.629 (+0.007)
(b) Remove attention	0.819 (-0.014)	0.831 (-0.011)	0.758 (-0.009)	0.620 (-0.002)
(c) Use error delta	0.800 (-0.033)	**0.796 (-0.046)**	0.715 (-0.052)	0.554 (-0.066)
(d) Use error sign	**0.788 (-0.045)**	0.800 (-0.042)	**0.703 (-0.064)**	**0.551 (-0.071)**

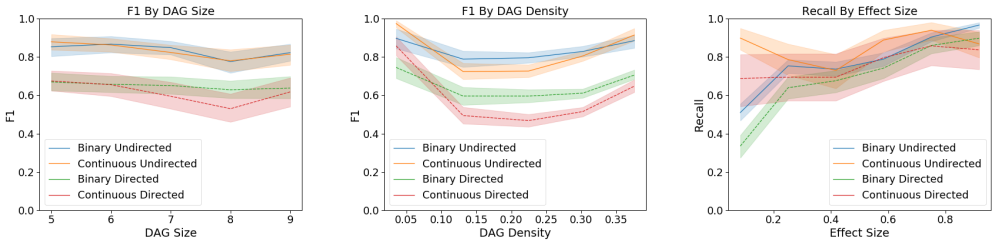

Figure 4: DAG-level F1 scores for DAGs with different numbers of nodes (left) and densities (center) and the edge-level recall values for edges with different effect sizes (right). The line shows the mean in each bin, while the band shows the standard deviation.

on the lowest density DAGs, which may be the easiest to infer due to lower levels of confounding. However, we also find that performance increases slightly on the highest density DAGs as well. Finally, for the binary data we observe the intuitive relationship that edges with larger effect sizes are more likely to be detected by the model, while the trend is less strong for the continuous datasets.

For the first-stage models, we observe a mean training time of 22 seconds for each DAG when running on CPU. To train the edge prediction models takes around 4 hours on a Tesla P100-PCIE-16GB GPU, while using the trained edge prediction model to infer the causal structure of a new dataset takes approximately 3.5 seconds.

4.4 Ablation Study

To understand the important features of the modeling approach, we perform an ablation study. The modeling decisions that we evaluate are (a) the inclusion of the observational values as features, (b) the use of the attention mechanism, (c) replacing the pair of observed errors with the difference in error, and (d) replacing the pair of observed errors with the sign of the error difference.

Consistently across both the binary and the continuous datasets, the inclusion of the original observational values (a) were not a useful predictor of the causal relationships. While this would be expected for the binary data, it is somewhat surprising for the continuous data where the difference of the variable from mean (which is used as the mask value) could be an indicator of the amount of evidence that instance can provide for a causal relationship. It is possible that the model learns to rely on the raw values of the errors as a proxy, with instances further from the mean likely having larger errors when relevant inputs are obscured. We find that the use of the attention mechanism (b) has an inconsistent level of effect across the different models, but does tend to improve the performance especially for the directed binary predictions and the undirected continuous predictions. We study the way the models leverage attention in the next section. Finally, we find that removing the individual values of the errors for each mask, either by using the difference (c) or the sign of the difference (d) as a feature, has a significant harmful impact on most of the models. This indicates that the model is relying on knowledge of the whether the individual instance is "easy" or "hard" to predict in order to

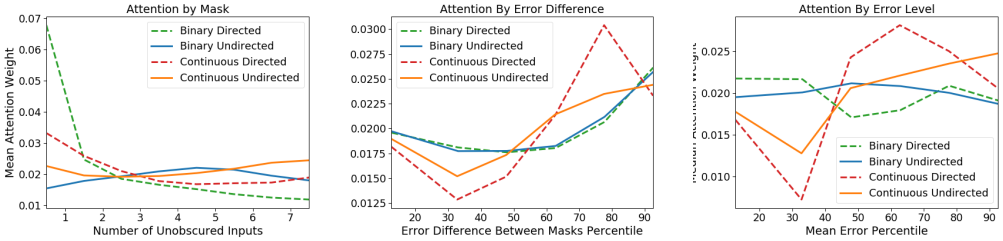

Figure 5: Mean attention weights for mask-pair observations with different properties.

interpret the evidence that the instance provides for the existence of a causal relationship rather than just focusing on whether the prediction improves or gets worse.

4.5 Attention Analysis

While we have shown that the attention mechanism has a fairly small benefit for model performance, it does have the benefit of providing a mechanism to probe the types of instances that are found to be informative by the model. To identify which mask-pair examples the model focuses on when aggregating the predictions across multiple instances, we analyze the learned attention weights for instances with different properties. Figure 5 shows how the mean attention weight varies with several instance properties for different models. Among the features we explore, we find that the number of unobscured inputs has the strongest effect on the learned attention. For the directed edge prediction, the models learn to up-weight the predictions for mask-pairs with low numbers of unobscured inputs, while for undirected prediction the models focus more evenly. We find that the models learn to up-weight instances with a larger observed difference in the errors, which likely provide larger evidence of a causal effect, especially for continuous data. Finally, we find that for continuous data the models also up-weight observations with larger mean error values between the two masks, which shows that "harder" instances are more important for inferring causal relationships.

5 Conclusions and Future Work

We propose a novel approach that uses deep learning model errors as a predictive signal to discover causal relationships among the variables from observational data. We first develop an efficient method to probe the model for differing errors when provided different sets of inputs. We then develop a predictive modeling approach to interpret the patterns in these observed errors for the inference of causal relationships. We find that the observed error differences for different input feature maps are predictive of causal edge existence for both binary and continuous datasets, outperforming commonly used causal discovery baselines. Through these results we demonstrate the feasibility of using model errors for causal structure inference.

Future efforts will focus on two key directions to enable practical application - generalizability and scalability. The supervised edge prediction model will need to be tested on data from outside the training distribution to explore whether the learned patterns generalize. We expect the deep learning based approach has the potential to generalize well to multiple data types and functional relationships among variables compared with traditional causal discovery algorithms. Therefore, future work will prioritize the application and evaluation of the approach to more mixed-type real-world datasets and will prioritize increasing the diversity of the edge prediction training set to encourage such generalizability. We will also explore the scalability of the approach to larger causal graphs which are more challenging for many causal discovery algorithms but are often encountered in real world applications. Among the questions that will need to be addressed for this is the number of mask-pair samples that will need to be generated to explore the space of the conditional dependence relationships. Finally, while the current approach relies on a two stage training procedure to generate the predictive error comparisons and then interpret them, future work will be targeted to develop a single step integrated training procedure to simultaneously explore the error patterns and infer the causal relationships.

♫ *And that's what it's all about* ♫

REFERENCES

Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse Gaussian Bayesian networks. *The Journal of Machine Learning Research*, 16(1):2273–2328, 2015.

Yoshua Bengio. From System 1 Deep Learning to System 2 Deep Learning, 2019. URL http://www.iro.umontreal.ca/~bengioy/NeurIPS-11dec2019.pdf.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Camden Cheek, Huiyong Zheng, Brian R Hallstrom, and Richard E Hughes. Application of a Causal Discovery Algorithm to the Analysis of Arthroplasty Registry Data. *Biomedical Engineering and Computational Biology*, 9:1179597218756896, 2018. doi: 10.1177/1179597218756896. URL https://doi.org/10.1177/1179597218756896. PMID: 29511363.

Pu Chen, Chihying Hsiao, Peter Flaschel, and Willi Semmler. Causal Analysis in Economics : Methods and Applications. 2007.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pp. 39–80. Springer, 2018.

D Kalainathan, O Goudet, I Guyon, D Lopez-Paz, and M Sebag. Structural agnostic modelling: Adversarial learning of causal graphs. *arXiv preprint arXiv:1803.04929*, 2018.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47 (11):1–26, 2012. URL http://www.jstatsoft.org/v47/i11/.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning Neural Causal Models from Unknown Interventions. *arXiv preprint arXiv:1910.01075*, 2019.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6979–6987, 2017.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 1 2019. ISSN 2504-4990. doi: 10.3390/make1010019.

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

Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, Gyorgy Simon, Michael Weiner, Paul Aisen, Ronald Petersen, Clifford Jack, Andrew Saykin, William Jagust, John Trojanowki, Arthur Toga, Laurel Beckett, Robert Green, John Morris, Leslie Shaw, Zaven Khachaturian, Greg Sorensen, and Maria Carroll. Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology. *Scientific Reports*, 10:2975, 02 2020. doi: 10.1038/s41598-020-59669-x.

Shohei Shimizu. LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Yuhao Wang, Vlado Menkovski, Hua Lan Wang, Xin Du, and Mykola Pechenizkiy. Causal Discovery from Incomplete Data: A Deep Learning Approach. *arXiv preprint*, arXiv:abs/2001.05343, 2020.

Qiaobing Wu, Lawrence Palinkas, and X. He. An Ecological Examination of Social Capital Effects on the Academic Achievement of Chinese Migrant Children. *British Journal of Social Work - BRIT J SOC WORK*, 40, 12 2010. doi: 10.1093/bjsw/bcq051.

Kui Yu, Jiuyong Li, and Lin Liu. A review on algorithms for constraint-based causal discovery. *arXiv preprint arXiv:1611.03977*, 2016.

Junpeng Zhang, Thuc le, Lin Liu, Bing Liu, Jianfeng He, Gregory Goodall, and Jiuyong Li. Identifying direct miRNA-mRNA causal regulatory relationships in heterogeneous data. *Journal of Biomedical Informatics*, 52, 08 2014. doi: 10.1016/j.jbi.2014.08.005.

Shengyu Zhu and Zhitang Chen. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.