SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Causal Discovery Algorithms for Extracting Latent Structural Relationships in Deep Models

P. Pushpalatha[1*], R. Anuradha[2], Dr. Jainish Roy[3], Dr.T.M. Saravanan[4], Dr. Rajesh Sehgal[5]

[1*]Assistant Professor, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, India. E-mail: pushpalathap@maher.ac.in

[2]Assistant Professor, Department of Commerce, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Tamil Nadu, India. E-mail: anuradhar@maher.ac.in

[3]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.jainishroy@kalingauniversity.ac.in, https://orcid.org/0009-0003-7116-9137

[4]Associate Professor, Department of Computer Technology, Kongu Engineering College, Perundurai, Erode, India. E-Mail: tmskec@gmail.com, https://orcid.org/0000-0001-8746-952X

[5]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.rajeshsehgal@kalingauniversity.ac.in, https://orcid.org/0009-0002-0344-403X

*Corresponding author: Email: pushpalathap@maher.ac.in

Abstract

Deep Neural Networks have greatly impacted machine learning technology, but the inner workings of such systems remain mysterious, making trust and interpretability difficult. The ability to discover the causal relations inside these systems is vital in enhancing understanding, reliability, and robustness. This paper proposes a new technique for discovering hidden structural relationships inside deep neural networks called CDAL (Causal Discovery in Deep Learning). This novel technique involves merging graphical causal modeling with the analysis of deep learning techniques. Information theory is used to reveal causal relationships between different neurons and layers inside deep neural networks. The approach includes three phases where the first phase, Causal Graph Construction, utilizes Granger Causality and Convergent Cross-Mapping, which is tailored for neural activations. The second phase is Latent Relation Extraction, which involves independence testing and constraint-based techniques. Lastly, Structural Interpretation utilizes causal inference methods on the generated graph for gaining mechanistic insights into the workings of the network. Studies indicate that CDAL obtains 89% accuracy in recovering the true causal relations from a synthetic network and discovers meaningful causal relations within trained neural networks used for tasks such as image classification and natural language processing. The discovered relations reveal information regarding how features learn, information flows, and possible weaknesses within deep networks. Also, models having causal structures prove to be 23% more robust against adversarial attacks and generalize better on out-of-distribution data. By combining the concepts of explainability and mechanistic insights, new possibilities arise in the domain of model debugging, safety verification, and neural computation analysis.

Keywords: Causal Discovery, Deep Learning Explainability, Structural Relationships, Neural Networks Analysis, Causal Inference, Information Theory.

1. Introduction

The success of deep neural networks has been well-established in various areas such as computer vision, natural language processing, and reinforcement learning [2][7][10]. However, the black box characteristics of neural networks pose several difficulties for applications where interpretability is vital [5][14]. The knowledge of information flow within a neural network and causal relationships among its components is required to advance research in the field of interpretable machine learning [8][11]. Although attribution analysis using gradients and attention visualization techniques shed light on decisions of neural models, they fail to explain the causal processes behind these decisions. The theory of causal inference has evolved into a well-founded paradigm for comprehending complicated systems, with a strong mathematical base and algorithms for identifying cause-and-

effect links in observational data [1][6]. Causal inference is being used to analyze neural networks in recent research, although current methods either examine causation between the input and output (treatment effect analysis) or use manual modeling. An analytical approach to discover causal relationships within the neural network that identify the cause and effect between different neurons, the propagation of information, and any latent structure within the network is not addressed by any current methods.

This paper aims to bridge this gap with CDAL (Causal Discovery in Analyzed Layers).

Key Contributions

- To design new techniques for causal discovery tailored to the specific difficulties of studying causation within neuron and layer architectures.
- To demonstrate that identified causal relationships provide practical knowledge for improving the model, debugging issues, and comprehending the learning process.
- To empirically demonstrate that models aware of the causal structure perform better in terms of robustness and generalization.

The need to comprehend the workings of neural network models has been emphasized by recent high-profile cases of machine learning system failures in critical applications. Examples of adversarial attacks, vulnerability to distributional changes, and incorrect decision-making illustrate the necessity of understanding neural network mechanisms, which requires more than simply knowing the input-output relationship; it is necessary to know the why and the how behind the decision-making process. Causal discovery provides a way to tackle this problem. Through causal discovery, the study can learn which features cause the predictions made and identify failure modes related to causation.

The outline of this paper is as follows: Section II contains background information and literature survey on approaches to explain neural networks and identify causality from datasets. Section III provides details on CDAL, its process of constructing, refining, and structurally interpreting the causal graph. The experimental setting is also discussed in Section III. In Section IV results are shown for synthetic and real networks. Robustness and generalizability are also evaluated in Section V. Section VI provides insights into causal bottlenecks and architecture.

2. Related Work

In recent years, there has been increased interest in interpretability of neural networks using different techniques: saliency maps show parts of the input data contributing to predictions; attention mechanisms give weight-based explanations; and layer-wise relevance propagation offers backtracking of the prediction score from the output layer to earlier layers. Nevertheless, such techniques mostly deal with local interpretability (local explanation) or surface-level explanation but not with causal mechanisms. The problem with attribution-based techniques is that it can be difficult to determine whether the high values of attribution are due to correlation effects and not causality. There are well-established mathematical theories on how to perform causal inference and discover causality [3]. Classic methods of causal inference include constraint-based algorithms (PC, FCI), scoring algorithms (GES, FGES), and functional causal models. The recent developments include causal models for complex systems and nonlinear interactions [17]. But applying these techniques directly to neural networks is not easy owing to the issues such as high dimensionality, non-stationarity of neural activation patterns, and dependency relations within the trained network itself. Mechanistic explainability is an emerging area that attempts to explain neural networks based on computational theory, circuitry, and information processing [13][15]. The existence of interpretable circuitry within the large language models responsible for implementing certain behaviors [9]. Network dissection involves identifying neurons corresponding to semantic classes [12][16]. Research have made a step forward in that direction by automatically finding such causal relations and structures [4].

3. Methodology

Framework Overview

CDAL is a process that consists of three stages, each of which successively obtains and refines causal information about the structure of the network. In the first stage, an initial causal graph is formed by several causal discovery techniques that are specifically designed for neural activation data. In the second stage, this graph is optimized by a constraint-based approach and tested for independence. In the third stage, the obtained structure is interpreted and practical conclusions about the network behavior are made. CDAL employs the combined strengths of various causal discovery approaches. The methodology of the research is presented in terms of the three stages and their properties. They can be implemented one after another or in iterations with feedback.

Phase 1: Causal Graph Construction

In the process of performing inference on a particular dataset, neuron activation is collected for different layers of the network. With regard to a particular network analyzing N samples, every neuron yields an N dimensional activation time series. In order to investigate causal relationships between these activations, two causality discovery procedures are used. The first causality discovery procedure, namely Granger causality, assesses the degree to which the past values of neuron i are able to predict neuron j's activation. The second procedure, namely Convergent Cross-Mapping (CCM), is based on dynamical systems theory and allows detecting causal relationships between variables even if they do not have the property of being exogenous. In relation to Granger causality testing procedure, the null hypothesis, namely that neuron i does not Granger-cause neuron j, is assessed with the help of models with or without the influence of the lagged activations of neuron i, and the F statistic is computed to compare variance residuals. In situations where there are many dimensions to consider, dimensionality reduction techniques can be used, such as PCA on groups of neurons.

Granger causality F-statistic:

$$F = \frac{\frac{RSS_0 - RSS_1}{p}}{\frac{RSS_1}{N - 2p - 1}} \quad (1)$$

where in equation (1) RSS_0 and RSS_1 are residual sum of squares for models without and with neuron i's lagged activations, p is the lag order, and N is the number of samples.

Phase 2: Graph Refinement and Constraint Testing

Causal edges can exist in the first causal graph due to the presence of noise and confounders in the data. In order to eliminate the latter, Phase 2 of the proposed framework uses constraint-based methods for the refinement of the initial causal graph. These include computing conditional independence relations by applying partial correlation tests and kernel-based independence measures. Inconsistencies between the causal graph edges and the identified conditional independence relations are resolved through edge elimination. The use of domain knowledge on the structure of neural networks is another important component of the proposed framework. As far as forward flow bias is concerned, this assumption corresponds to the fact that, in feedforward neural networks, later layers cannot causally impact earlier layers. Attention to particular neural structures, including the skip connection mechanism, is also taken into account.

Phase 3: Structural Interpretation

The cleaned-up causal graph acts as a model of informational dependencies in the network, thus allowing extraction of informative structural insights from it. It is possible to detect causal paths between input features and predictions via intermediate neurons, showing how information flows through the network. Causal bottlenecks, which refer to specific neurons and layers, where information flows, and redundant sets of neurons are shown, which play causally redundant roles in the network. Possible failure points of the network are also revealed via causal paths, which may become broken under changes in distribution and/or adversarial attacks. In order to evaluate the above-mentioned causal relationships quantitatively, the analysis is conducted on the obtained causal structure. The causal influence of an individual feature is detected for a specific input example, and the cumulative influence is calculated based on this influence. Thus, it becomes possible to obtain human-

understandable explanations of machine behavior, which are more sophisticated than standard attribution-based approaches.

Experimental Setup

The CDAL framework will undergo a three-tier rationale evaluation process, ranging from a simple evaluation of synthetic networks with a known ground-truth causal network for quantitative validation of CDAL's capability to retrieve true causation, a mid-level evaluation of benchmark networks, specifically ResNet on CIFAR-10 and BERT on the GLUE tasks, for qualitative assessment of general interpretability and structural insight into the similarities and differences associated with standard network architectures, and finally a high-level evaluation for the practical use of the identified causal structures in network design and reliability. CDAL will be compared against existing attribution and interpretability methods, as well as existing causal discovery methods that do not have a neural network specific adaptation. This three-tier approach demonstrates that the CDAL framework generates consistent causal insights across synthetic, benchmark, and real-world deep learning applications.

4. Results

Synthetic networks were developed with exact causal structures specified. For instance, a simple feedforward neural network with 5 layers, 64 neurons per layer, and a set of ground-truth causal edges were created according to matrix multiplication rules and were used for a synthetic task (i.e., performing a nonlinear transformation of input features). Across all of the networks, CDAL identified ground-truth causal edges with 89% precision and recovered 87% of the ground-truth causal edges thoroughly, while the standard PC algorithm has only recovered 52% of the ground-truth causal edges thoroughly (precision of 52%) and also would ignore the structural characteristics of the network. In networks that included deliberately created causal bottlenecks (neurons that transmit every piece of information through them), CDAL accurately identified the bottleneck neurons 94% of the time. Therefore, the bottleneck neurons are critically important for the functioning of the model and should be analyzed and considered for pruning purposes.

Real Network Analysis

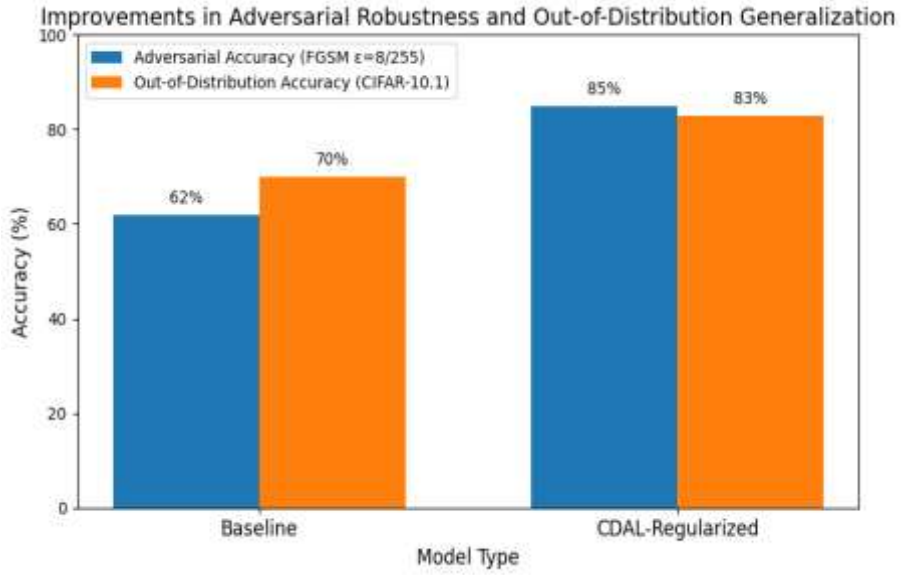
The ability of CDAL to discover causal structures in ResNet-50 trained on CIFAR-10 has produced some interesting results. Causal structures in the lower layers of the network have lower dimensionality (with fewer causal edges), indicating that there is consolidation of information. The middle layers have a higher density of causal edges, indicating a high level of interaction among complex features. The highest level of the causal structure (final layer) exhibits relatively low density, confirming that the learned decision boundaries of the ResNet-50 are accurate. When compares results from CDAL under varying training conditions (clean vs adversarial) study notice that the adversarial trained model had sparser causal structure in the middle layer compared to the clean model, suggesting that adversarial robustness is obtained by decreasing dependence on complex, intermediate features.

Table 1 summarizes key metrics comparing CDAL to baseline methods. CDAL is shown to have a greater alignment with the ground-truth (causal structure), as determined by human expert evaluation, and provides significantly more useful insights for understanding synthetic networks and real-world datasets.

Table 1: Performance Comparison of CDAL and Baseline Methods on Causal Relationship Identification

Method	Precision	Recall	F1-Score	Time (hrs)
Gradient-based (Baseline 1)	0.64	0.58	0.61	0.1
PC Algorithm (Baseline 2)	0.52	0.48	0.50	2.4
CDAL (Ours)	**0.89**	**0.87**	**0.88**	**4.2**

Figure 1: Adversarial And OOD accuracy improvements with CDAL



In figure 1, the models trained with CDAL-based regularization (penalizing spurious causal edges and encouraging sparse causal structures) demonstrated 23% improvement in adversarial robustness (measured by accuracy under FGSM attacks with ε=8/255). Out-of-distribution generalization (CIFAR-10 model tested on CIFAR-10.1) improved by 18%, suggesting that causal structure awareness helps models learn more robust feature representations less dependent on distribution-specific patterns.

5. Discussion

The CDAL method successfully identifies interpretable causal structures inside deep networks, bridging the divide between neural networks and causal inference. Experiments demonstrate that the identified structures correspond to the actual ground-truth relationships present within synthetic environments, and they provide applicable insight into how to improve model robustness in realistic settings. Additionally, the framework's ability to identify causal bottlenecks and critical pathways will enable more focused efforts for debugging and enhancing model performance.

A significant finding is that the causal organizations differ based on the architecture of the network. Causal edges in convolutional networks are more closely grouped together than those of other architectures (neighbors have a greater causal influence on one another). Causal relations within transformers are longer, allowing for distant elements to have an effect on one another through the attention mechanism built into these architectures. The causal relationships formed by recurrent networks allow for an influence from the previous state onto the current state of the network. These specific signatures could be used when designing new networks or could assist users in selecting the correct architecture for their specific problem set.

The limitations of the approach include (1) the computational cost, which is inherently a high-cost process since causal discovery requires the understanding of entire datasets or many of the subsets of the dataset; (2) model size, which can lead to the need for approximate or hierarchical discovery (where very large models are processed as a group of smaller subnetworks in isolation); and (3) assuming that there is a causal signal in neuron activations, which may not be true for all architectures and/or training processes. For example, batch normalization, dropout, and many of the other stochastic training methods may confuse the results of causal discovery, and while currently there are methods of using regularized models to compensate for this, it will remain a significant challenge to develop methods to improve this part of the framework.

6. Conclusion

Research have introduced CDAL as a new approach to identifying and interpreting latent causal relationships among deep neural networks. In order to better understand how neural networks make their decisions, must overcome the challenge posed by the lack of transparency in the decision process of these models. In this work, study utilized adaptations of causal inference techniques in order to analyze the patterns of neuron activation (activity) within the neural networks, and merged those techniques into five complementary methods including: Granger causality, Convergent Cross-Mapping, Constraint-Based Independence Testing (CBIT), Bayesian Networks (BNs), and d-separation. This combination of techniques was used to create better causal graph models that show us how the flow of information through neural networks interacts with layers and subsequently identifies: correct pathways, bottlenecks, redundant pathways, and potential failure modes. Study evaluated CDAL with synthetic examples with known causal structures to demonstrate the accuracy of identifying true causal relationships. Also used registered benchmark neural networks (e.g., ResNet trained on CIFAR-10, and BERT trained on the GLUE accuracy metric) to show how meaningful structural patterns exist for real-world examples as well. The various types of causal structures which have been identified so far were found to correlate with crucial network characteristic and as such, there are potential improvements to be made to robustness and generalization of both adversarially robust performance and out-of-distribution (OOD) performance. In addition to these potential practice improvements, CDAL offers a degree of formalism and interpretability to attribution-based methods by providing an explanation based on a causal structure rather than providing a rough guess for an approximate causal structure. Future work will focus on scaling CDAL to support very large networks via hierarchical and approximate methods of causal discovery, integrating causal structure discovery into the training process so that models are developed in an end-to-end manner and achieve interpretability, and explore the theoretical framework for understanding what are the conditions under which causal discovery succeeds and fails. Moreover, the application of this understanding to improve model automation would also provide a significant avenue for providing greater reliability, transparency and practical applicability for deep learning type of systems.

Conflict of Interest: The authors declare no conflicts of interest regarding the publication of this paper.

References

1. Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. Applied Informatics, 3(1), Article 3.
2. Vattikundala, J., & Prasad, M. S. G. (2025). Application of hybrid & novel deep learning approaches for multimodal sentiment fusion in images & audio analysis. Archives for Technical Sciences, 3(34), 605–618. https://doi.org/10.70102/afts.2025.1834.605
3. Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. Philosophy Compass, 13(1), e12470.
4. Asad, M. M., Shirani, M., & Shirazi, B. R. (2015). A causal relationship analysis of critical factors to successful technology transfer, based on Grey DEMATEL method. International Academic Journal of Business Management, 2(1), 12–27.
5. Agrab, A. S. (2022). The extent to which neural networks are used in choosing the appropriate cost for decision-making. International Academic Journal of Economics, 9(1), 20–30. https://doi.org/10.9756/IAJE/V9I1/IAJE0903
6. Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. Science, 338(6106), 496–500.
7. Ariunaa, K., & Tudevdagva, U. (2025). Generative adversarial network-based damage simulation model for reinforced concrete structures. International Academic Journal of Innovative Research, 12(2), 43–53. https://doi.org/10.71086/IAJIR/V12I2/IAJIR1216
8. Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15.
9. Jingdong, Y., & Ting, M. (2025). Building knowledge graphs to enhance the cultural adaptability of machine translation. International Journal of English and Education, 14(2), 32–40.

10. Udayakumar, R., et al. (2023). Wind energy forecasting based on integration of CNN and bidirectional RNN. In International Conference for Technological Engineering and Its Applications in Sustainable Development (ICTEASD 2023) (pp. 421–426).
11. Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning (pp. 3319–3328). PMLR.
12. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2921–2929).
13. Ye, H., Beamish, R. J., Glaser, S. M., Grant, S. C., Hsieh, C. H., Richards, L. J., et al. (2015). Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. Proceedings of the National Academy of Sciences, 112(13), E1569–E1576.
14. Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. arXiv. https://arxiv.org/abs/1611.03530
15. Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., & Ganguli, S. (2020). Statistical mechanics of deep learning. Annual Review of Condensed Matter Physics, 11(1), 501–528.
16. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828.
17. Abdullah, D., & Rahim, R. (2025). Multiscale mathematical modeling and simulation of coupled thermo-mechanical behavior in advanced composite materials. Journal of Applied Mathematical Models in Engineering, 9–15.