Incorporating Causal Prior Knowledge into Deep Neural Networks

Shahenda Youssef

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
shahenda.youssef@kit.edu

Abstract

Deep Neural Networks have achieved significant success in solving complex problems across various domains due to their ability to capture complicated patterns in large datasets; however, they often require large amounts of data to learn effectively and often lack transparency in their decision-making processes, relying heavily on correlation rather than causation. Such limitations have led to incorporating causal Prior Knowledge into neural network models which stands as a significant advancement in machine learning, such knowledge can mitigate this data dependency, guide the learning process, and enhance not only the robustness and generalizability of models but also their interpretability and explainability. Additionally, it enables models to adapt to new tasks and domains with greater ease and effectiveness.

This report tackles the importance of incorporating causal prior knowledge into deep neural networks and the methodologies that facilitate this incorporation. Fundamental concepts of causality are reviewed, with emphasis on its importance for advancing AI towards causal representation learning.

1 Introduction

Although Deep Neural Networks (DNNs) have shown promising results in diverse domains, they still present limitations in several aspects that are left to be resolved. The insufficient amount of training data usually hinders its performance due to the lack of generalization, and the black-box nature of deep neural networks does not allow for a precise explanation behind its mechanism, preventing a new scientific discovery. They can discover features hidden within input data together with their mutual co-occurrence. However, they are weak at discovering and making explicit hidden causalities between the features. To overcome these challenges, It is critical to incorporate causality into DNNs framework [32].

The emergence of causality in AI signifies a paradigm shift from predictive to prescriptive analytics, where machines not only forecast but also recommend actions that lead to desired outcomes. Unlike correlation, which captures coincidental patterns, causality delineates a roadmap of cause and effect, empowering AI with the ability to reason beyond the data it is trained on [25].

Incorporating causal prior knowledge into the architecture of neural networks contributes to model robustness and generalizability, allowing models to better handle changes in data distributions by focusing on causal relationships rather than correlations [31]. Embedding causal reasoning helps models provide interpretability and explanations that resonate with real-world causality, aligning more with human thinking, and extending their use to interventions and policy-making [32].

The methodologies to incorporate prior knowledge within neural networks are varied [8]. Designing network architectures that detect complex data patterns. Imposing informed constraints on the loss function directs the optimization process towards solutions that respect established relationships and theoretical frameworks. Employing data augmentation and leveraging the insights from transfer learning further exploit the breadth of existing data and pre-trained models, accelerating the learning process. Knowledge graphs are adopted to enhance neural networks with information about relations between instances [1].

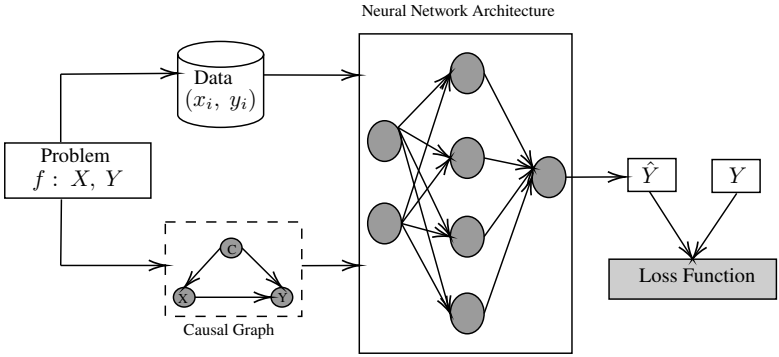

Figure 1.1: The proposed framework to incorporate causal prior knowledge into neural network.

These strategies encapsulate a concerted move towards data-driven learning with causal prior knowledge.

Current research focuses on enhancing model performance and explainability through the incorporating of prior knowledge into the learning process [36, 2, 10]. However, the development of models that incorporate causal prior knowledge continues to be a challenge. Figure 1.1 is a proposed framework to incorporate causal prior knowledge into a neural network, such a collaboration system can be achieved by involving the usual training data and additional prior knowledge that comes from an independent source, which is given by the causal graph model.

The rest of the report is organized as follows: Section 2 provides an overview of causality. Section 3 describes incorporation of prior knowledge. Section 4 addresses the related work in causal representation leaarning. Section 5 outlines some major challenges related to the incorporation of Causal prior knowledge into DNNs. In this section, we also present insights into potential directions for future research.

2 Causality

The study of causality seeks to establish the nature and strength of cause-and-effect relationships [26]. "Correlation does not imply causation" [25], two variables y and x could be correlated (statistically dependent) and, therefore, seeing x allows predicting the value of y, but if y is not caused by x then setting the value of x won't affect the distribution of y.

Causal inference is the process of concluding a causal connection based on the conditions of the occurrence of an effect. It involves establishing that a change in one variable (the cause) brings about a change in another variable (the effect). Researchers have developed methodologies to estimate causal effects from observational data. In this section, we present our interest frameworks that are introduced to causal inference.

2.1 Structural Causal Model

Structural Causal Models (SCM) provide a mathematical framework to model and infer causal relationships [24]. They are based on the idea that causal relationships can be represented by a set of structural equations and Directed Acyclic Graphs (DAGs). Each node in the DAG represents a variable, and each edge represents a causal influence from one variable to another. SCM can be represented as

$$X := f_X(PA_X, U_X), \tag{2.1}$$

a variable X in the causal graph is determined by a function f_X that could be linear or non-linear, whose inputs are its parents PA_X and a random variable U_X representing potential chaos and variables unobserved in the causal graph explicitly.

2.2 Average Treatment Effect

The potential outcomes framework [30] is used to estimate the causal effect of an intervention. Consider a binary treatment variable T, where $T = 1$ if the treatment is given and $T = 0$ otherwise. For each individual i, there are

two potential outcomes: $Y_i(1)$ is the outcome if the individual i receives the treatment, and $Y_i(0)$ is the outcome if they do not. The Individual Treatment Effect (ITE) for i would be

$$ITE_i = Y_i(1) - Y_i(0) \tag{2.2}$$

However, we never observe both potential outcomes for the same unit. This problem is known as the "Fundamental problem of causal inference" [24]. Since we cannot observe both potential outcomes for the same unit, we often focus on the average effect of the treatment across all units. The Average Treatment Effect (ATE) is defined as

$$ATE = \mathbb{E}[Y|\text{do}(T = 1)] - \mathbb{E}[Y|\text{do}(T = 0)], \tag{2.3}$$

where $\mathbb{E}[Y|\text{do}(T = t)]$ represents the expectation of the outcome Y under the intervention $do(T = t)$, do-operator allow for the identification and estimation of causal effects from observational data under certain conditions.

2.3 Propensity Score Matching

Propensity Score Matching (PSM) is a statistical technique used to estimate the effect of a treatment, policy, or other intervention by accounting for the covariates that predict receiving the treatment. The key idea is to match units that received the treatment with similar units that did not receive the treatment based on their propensity scores. The propensity score for a unit is the probability of receiving the treatment given a set of observed covariates. First, the propensity score $e(X)$ for each unit is estimated, typically using logistic regression for binary treatments

$$e(X) = P(T = 1|X) = \frac{1}{1 + e^{-(\alpha+\beta X)}} \tag{2.4}$$

where T represents the treatment assignment, X represents the covariates, α is the intercept, β is the vector of coefficients, and e is the base of the natural logarithm. After estimating the propensity scores, units are matched. The goal

is to find for each treated unit i a control unit j such that their propensity scores are as close as possible

$$\min_{i,j} |e(X_i) - e(X_j)|, \tag{2.5}$$

where $e(X_i)$ is the propensity score of the treated unit i and $e(X_j)$ is the propensity score of the control unit j.

Estimation of Treatment Effect, the Average Treatment Effect on the Treated (ATT) is often estimated by comparing the outcomes Y between matched units

$$ATT = \frac{1}{N_T} \sum_{i \in T} (Y_i - Y_{j(i)}), \tag{2.6}$$

where N_T is the number of treated units, Y_i is the outcome for treated unit i, and $Y_{j(i)}$ is the outcome for the control unit j that is matched to i.

3 Prior Knowlege Incorporation

The incorporation of prior knowledge into the construction of deep neural networks (DNNs), focuses on the nature of input data to a deep neural network, the loss function employed during training, and the model architecture or its parameters of the neural network [8, 9]. Ongoing studies are concentrated on combining methods to guarantee that the embedded prior knowledge effectively guides the learning process while still permitting the neural network to discover new data-driven patterns.

3.1 Input Data

Embedding domain knowledge into DNNs by transforming the input data, we discuss two ways to do this. One way is feature engineering is a key approach, where additional attributes derived from physics-based models are integrated with the training data. The training data is processed through domain-specific functions for embedding prior knowledge into deep learning. Feature engineering was found to be one of the most common ways of integrating prior knowledge into deep learning [16].

The other way is how to represent domain knowledge that takes the form of graph-based data as input, Knowledge graphs can be directly utilized by specialized deep network models such as Graph Neural Networks (GNNs), which process graph-structured data. These networks aggregate and synthesize information from the knowledge graph to enhance predictive tasks [6, 13, 17].

3.2 Loss Function

DNNs can be enhanced with domain knowledge by adding penalty terms to the loss function that enforce constraints derived from that knowledge [22, 12]. There are two primary types of constraints: syntactic, which are often implemented through regularization to control model complexity, and semantic, which encode domain-specific truths and logic [9]. Syntactic constraints are implemented by incorporating regularization terms into the loss function to control model complexity, such as the number of layers or parameters. It also involves an embedding approach, which is a lower-dimensional representation of discrete variables. Penalty terms based on regularizing embeddings are used to encode syntactic constraints on the complexity of embeddings to define prior parameter distributions using knowledge graphs embeddings [33]. Semantic constraints are imposed by the domain knowledge and can specify the conditions that model predictions must satisfy, such as falling within a certain numerical range.

When learning a function f from data (x_i, y_i), where x_i are input features and y_i are the actual labels, the generic hybrid loss function of the deep learning model

$$\arg\min_{f} \text{Loss}(Y, \hat{Y}) + \lambda R(Y, \hat{Y}) + \lambda_D \text{Loss}_D(\hat{Y}), \tag{3.1}$$

where $\text{Loss}(Y, \hat{Y})$ is the label-based loss, $Y, \hat{Y}$ are the actual labels and predicted values, respectively. $\lambda R(Y, \hat{Y})$ is a regularization function used to control model complexity. $\text{Loss}_D(\hat{Y})$ is the prior knowledge directly incorporated into the NN loss function and is used to enforce the model to respect the prior knowledge

while training. λ_D is a hyper-parameter determining the effect of domain loss in the objective function.

3.3 Network Architecture

DNNs can be enhanced by incorporating domain knowledge, either through constraining model parameters or by deliberate architectural choices.

Priors can be introduced on the parameters of a network. Explicitly, these would take the form of a prior distribution over the values of the weights in the network. The priors on networks and network weights represent our expectations about networks before receiving any data and correspond to penalty terms or regularizes. Two main methods have been used in DNNs, Transfer learning and Data Augmentation. Transfer learning is a technique to import weight priors in scenarios where data is scarce for the target problem. This method leverages existing models from a related source problem to inform the target model's structure or parameters, thus embedding the domain knowledge into the target domain [20, 27].

Data Augmentation based on prior information can effectively extend the original dataset with synthetic or transformed examples that reflect domain-specific insights. This method allows the integration of additional contextual or structural information, informed by prior understanding, to enrich the training process and improve the robustness and generalization of the neural network models [3].

Further, specialized structures in DNNs are enhanced when the architecture of the network is informed by domain knowledge, as the way knowledge representations are integrated into the network is largely determined by the type of architecture [16] such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph Neural Networks (GNNs).

4 Causal Representation Learning

The work by Deng et al. [11] introduces a deep learning framework for societal event forecasting that leverages causal inference, and employs Individual Treat-

ment Effects (ITE) to estimate the influence of various treatments (or events) on societal outcomes with spatiotemporal environments. The model predicts potential outcomes for various treatment scenarios, thus incorporating causal information into event predictions. This process consists of two methods:

Approximation constraints l, u apply to event prediction scores $\hat{y}$, ensuring that the training of event predictors P for a location M at time $t+\delta$ takes into account the potential outcomes estimated by the causal inference model

$$l^{t+\delta} = \min(\hat{y}^{t+\delta}), \quad u^{t+\delta} = \max(\hat{y}^{t+\delta}). \tag{4.1}$$

The constraints limit the range of the ITE by enforcing minimum and maximum values derived from causal knowledge, where $\hat{y}^{t+\delta}$ is a set of potential outcomes for all treatment events, and the defined constraint loss term

$$\mathcal{L}_{\text{CSTR}} = \sum_{t \in T} \sum_{i \in M} \text{ReLU}(l_i^{t+\delta} - \hat{y}_i^{t+\delta}) + \text{ReLU}(\hat{y}_i^{t+\delta} - u_i^{t+\delta}). \tag{4.2}$$

By minimizing the total loss while training the predictor

$$\mathcal{L}_{\text{EVT}} = \mathcal{L}_{\text{PRED}} + \mu \cdot \mathcal{L}_{\text{CSTR}}, \tag{4.3}$$

where $\mathcal{L}_{\text{PRED}}$ is the loss function defined by the predictor P and μ is a hyperparameter.

Feature reweighting involves using the ITE estimated from the causal inference model to reweight event frequency features. This reweighting is essential for capturing the importance of features for predicting events. The approach defines a gating feature $\rho^{t+\delta}$ based on ITE, where for the j-th treatment event, the estimated ITE of a location at time $t+\delta$ is computed as follows

$$\hat{\tau}_{(j)}^{t+\delta} = \hat{y}_{(j)}^{t+\delta}(1) - \hat{y}_{(j)}^{t+\delta}(0), \tag{4.4}$$

where $\hat{y}_{(j)}^{t+\delta}(1)$, $\hat{y}_{(j)}^{t+\delta}(0)$ are the predicted potential outcomes with and without the treatment, respectively. The gating variables are applied to the original event frequency features through a sigmoid function σ to obtain a soft gated signal

$$\rho^{t+\delta} = \sigma(f_\tau(\hat{\tau}^{t+\delta})). \tag{4.5}$$

The event frequency features x are reweighted using the causal feature gates. The new feature vector is the element-wise product of the original feature vector x^t and the gating variables $\rho^{t+\delta}$

$$\tilde{x}^t = x^t \odot \rho^{t+\delta} + x^t \tag{4.6}$$

Such features are fed into a predictor to perform event prediction. They conducted extensive experiments on several real-world event datasets and showed that their approach achieves the best results in ITE estimation and robust event prediction involving multiple treatments and outcomes, which is considered an advancement over traditional correlation-based forecasting methods.

It is essential to recognize the expanding number of research that highlights the integration of causal regularization strategies into the framework of predictive modeling [19, 14, 15].

The paper by Teshima [35] introduces a model-independent method for data augmentation that leverages the conditional independencies relations in the data distribution encoded in causal graphs to enhance supervised learning.

Richens et al. [29] introduces the concept of counterfactual diagnosis, which uses counterfactual reasoning to evaluate the likelihood of a disease-causing the patient's symptoms. Structural causal models (SCMs) are discussed as methods for encoding the relationships between diseases, symptoms, and risk factors for more accurate diagnostic reasoning. The authors show that incorporating knowledge into machine learning can be effective in assisting medical diagnosis to reduce diagnostic errors.

Kyono et al. [18] demonstrates the utilization of causal graphs as prior knowledge to enhance model selection to enhance the robustness of neural network performance. By embedding this knowledge within a Structural Causal Model, derive a score that assesses the compatibility of a model's predictions with the SCM and input variables.

A recent work by Terziyan and Vitko [34] presents an approach for enhancing Convolutional Neural Networks (CNNs) by incorporating causality-awareness into the architecture. The authors introduce an architecture that includes an additional layer of neurons that is engineered to estimate asymmetric causality

in images using causal disposition [21] by using convolutional layers to capture features from images and then using these features to estimate conditional probabilities of the presence of one feature given another to improve image classification and generation. The causality map which is the calculated causality estimates is integrated into the CNN architecture, and the content of this map is calculated using

$$P(F^i|F^j) = \frac{\left(\max_{l,r=1,n} F^i_{l,r}\right) \cdot \left(\max_{l,r=1,n} F^j_{l,r}\right)}{\sum_{l,r=1}^{n} F^j_{l,r}} \tag{4.7}$$

where $P(F^i|F^j)$ is the causality map of size $k \times k$ (k - number of features), the features F^1, F^2, F^k represented by $n \times n$ feature maps. The causality map provides additional inputs to the network, which are used during backpropagation, enabling the network to discover which features have significant causal relationships. They also use it as a component within Generative Adversarial Networks (GANs) to enable the generation of images with respect to causalities. They demonstrated that their suggested model not only enhances the classification effectiveness of traditional CNNs but also improves the interpretability of the model's results.

5 Challenges and Future Prospects

Current research tends to address data-driven that is independent and identically distributed (IID). However, when dealing with spatiotemporal data that does not follow this IID assumption, the task of incorporating causal models that cope with strongly correlated values over time is not trivial [7].

The utilization of deep learning within manufacturing systems remains at an early stage, not only because of its solely data-driven nature, but also due to the limited research conducted on embedded causal knowledge into deep learning models by domain experts in the field.

Figure 1.1 raises some open research questions: incorporating causal prior knowledge requires modifications to the loss function, formulating an appropriate

term for the loss function can be complex. Introducing such a term frequently leads to complex optimization problems [5, 4].

New combinations of approaches are possible which have not been investigated yet, for example, by merging the causal prior knowledge with the DNNs architecture using the attention mechanism, the model iteratively processes the knowledge by selecting the relevant content at each step. The knowledge-based attention layer helps improve the prediction and the performance of the model.

Another prospective framework is to incorporate a causal graph model with an embedding graph layer, which would then serve as input to Bayesian Neural Networks (BNNs). This integration aims to enhance causal prior knowledge by refining the prior distribution during model training. This probabilistic approach reflects uncertainty in the model's predictions, where understanding the confidence level of a prediction is as important as the prediction itself [23, 28].

References

[1] Peter Battaglia et al. "Interaction networks for learning about objects, relations and physics". In: *Advances in neural information processing systems* 29 (2016).

[2] Katharina Beckh et al. "Explainable machine learning with prior knowledge: an overview". In: *arXiv preprint arXiv:2105.10172* (2021).

[3] Andrea Borghesi, Federico Baldo, and Michela Milano. "Improving deep learning models via constraint-based domain knowledge: a brief survey". In: *arXiv preprint arXiv:2005.10691* (2020).

[4] Luiz Chamon and Alejandro Ribeiro. "Probably approximately correct constrained learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 16722–16735.

[5] Luiz FO Chamon et al. "Constrained learning with non-convex losses". In: *IEEE Transactions on Information Theory* 69.3 (2022), pp. 1739–1760.

[6] Qibin Chen et al. "Towards knowledge-based recommender dialog system". In: *arXiv preprint arXiv:1908.05391* (2019).

[7] Ricky TQ Chen et al. “Neural ordinary differential equations”. In: *Advances in neural information processing systems* 31 (2018).

[8] Tirtharaj Dash et al. “A review of some techniques for inclusion of domain-knowledge into deep neural networks”. In: *Scientific Reports* 12.1 (2022), p. 1040.

[9] Tirtharaj Dash et al. “How to tell deep neural networks what we know”. In: *CoRR, abs/2107.10295* (2021).

[10] Arka Daw et al. “Physics-guided neural networks (pgnn): An application in lake temperature modeling”. In: *Knowledge Guided Machine Learning*. Chapman and Hall/CRC, 2022, pp. 353–372.

[11] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. “Causal Knowledge Guided Societal Event Forecasting”. In: *arXiv preprint arXiv:2112.05695* (2021).

[12] Ethan Gallup, Tyler Gallup, and Kody Powell. “Physics-guided neural networks with engineering domain knowledge for hybrid process modeling”. In: *Computers & Chemical Engineering* 170 (2023), p. 108111.

[13] Manas Gaur, Keyur Faldu, and Amit Sheth. “Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?” In: *IEEE Internet Computing* 25.1 (2021), pp. 51–59.

[14] Dominik Janzing. “Causal regularization”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[15] Lucas Kania and Ernst Wit. “Causal Regularization: On the trade-off between in-sample risk and out-of-sample risk guarantees”. In: *arXiv preprint arXiv:2205.01593* (2022).

[16] Sung Wook Kim et al. “Knowledge Integration into deep learning in dynamical systems: an overview and taxonomy”. In: *Journal of Mechanical Science and Technology* 35 (2021), pp. 1331–1342.

[17] Ugur Kursuncu, Manas Gaur, and Amit Sheth. “Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning”. In: *arXiv preprint arXiv:1912.00512* (2019).

[18] Trent Kyono and Mihaela van der Schaar. “Improving model robustness using causal knowledge”. In: *arXiv preprint arXiv:1911.12441* (2019).

[19] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. “Castle: Regularization via auxiliary causal graph discovery”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1501–1512.

[20] Xuan Liu, Xiaoguang Wang, and Stan Matwin. “Improving the interpretability of deep neural networks with knowledge distillation”. In: *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2018, pp. 905–912.

[21] David Lopez-Paz et al. “Discovering causal signals in images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6979–6987.

[22] Nikhil Muralidhar et al. “Incorporating prior domain knowledge into deep neural networks”. In: *2018 IEEE international conference on big data (big data)*. IEEE. 2018, pp. 36–45.

[23] Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

[24] Brady Neal. “Introduction to causal inference”. In: *Course Lecture Notes* (2020).

[25] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[26] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[27] Maithra Raghu et al. “Transfusion: Understanding transfer learning for medical imaging”. In: *Advances in neural information processing systems* 32 (2019).

[28] Qihan Ren et al. “A Representation Bottleneck of Bayesian Neural Networks”. In: (2022).

[29] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. “Improving the accuracy of medical diagnosis with causal machine learning”. In: *Nature communications* 11.1 (2020), p. 3923.

[30] Donald B Rubin. “Causal inference using potential outcomes: Design, modeling, decisions”. In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331.

[31] Bernhard Schölkopf. “Causality for machine learning”. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 765–804.

[32] Bernhard Schölkopf et al. “Toward causal representation learning”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

[33] Naoya Takeishi and Kosuke Akimoto. “Knowledge-Based Distant Regularization in Learning Probabilistic Models”. In: *CoRR* abs/1806.11332 (2018). arXiv: 1806.11332. URL: http://arxiv.org/abs/1806.11332.

[34] Vagan Terziyan and Oleksandra Vitko. “Causality-Aware Convolutional Neural Networks for Advanced Image Classification and Generation”. In: *Procedia Computer Science* 217 (2023), pp. 495–506.

[35] Takeshi Teshima and Masashi Sugiyama. “Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 86–96.

[36] Laura Von Rueden et al. “Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.1 (2021), pp. 614–633.