

# **A SENSITIVITY ANALYSIS BASED APPROACH IN INTERPRETING CAUSAL INFERENCE IN NEURAL NETWORK AND ITS RELATION WITH PCA**

**Ankit Yadu**

**Master of Technology Thesis**  
June 2020



International Institute of Information Technology, Bangalore

**A SENSITIVITY ANALYSIS BASED APPROACH IN  
INTERPRETING CAUSAL INFERENCE IN NEURAL  
NETWORK AND ITS RELATION WITH PCA**

Submitted to International Institute of Information Technology,  
Bangalore  
in Partial Fulfillment of  
the Requirements for the Award of  
Master of Technology

by

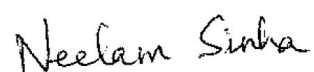
**Ankit Yadu**  
**SMT2017001**

International Institute of Information Technology, Bangalore  
June 2020

*To my parents and siblings for their love and support,*  
*To Prof. Neelam Sinha for her invaluable guidance throughout the work,*  
*To Judea Pearl and the research community for their contributions*  
*towards Causality.*

## Thesis Certificate

This is to certify that the thesis titled **A sensitivity analysis based approach in interpreting causal inference in neural network and its relation with PCA** submitted to the International Institute of Information Technology, Bangalore, for the award of the degree of **Master of Technology** is a bona fide record of the research work done by **Ankit Yadu, SMT2017001**, under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma. The thesis conforms to plagiarism guidelines and compliance as per UGC recommendations.



---

Dr. Neelam Sinha

Bangalore,

The 12<sup>th</sup> of June, 2020.

# **A SENSITIVITY ANALYSIS BASED APPROACH IN INTERPRETING CAUSAL INFERENCE IN NEURAL NETWORK AND ITS RELATION WITH PCA**

## **Abstract**

A singular problem that mars the wide applicability of machine learning (ML) models is the lack of generalizability. Recent research trends in interpretability of ML models rely on robustness of the model to a varied feature space. In this work, we propose a new method based on causal inference and sensitivity analysis for verifying the causal inference results. The generation of counterfactual scenarios aids in comprehending the machine learning model. The proposed method is verified on IRIS data set and hand-crafted numerals, and the principal features and their variability are obtained which shows how the principal component analysis can be used to capture the gist of causal analysis. The method helped to capture the sensitivity parameter which is used to understand the model's causal behaviour by interpreting the causal effect of each input feature on the model's prediction and provides a formal method to analyse the local behaviour of the input feature on the model prediction.

## Acknowledgements

I would first like to thank my thesis advisor, Dr. Neelam Sinha. She has always guided me with her knowledge on the subject and provided her invaluable insights. For the time I have worked under her guidance, she had always found time to review the work and provided her detailed comments for me to improve upon. She has always given me an opportunity to step up my work and for this I will always be grateful to her. I feel fortunate to work under her and could learn ways of approaching research problems.

I shall take this opportunity to thank Judea Pearl for his contribution towards the field of Causality and inventing do-calculus. I was inclined to work and contribute in this field with my thesis work due to the robustness and indispensability of causal inference in the field of A.I. I will be ever thankful to my advisor for introducing this field to me. During my thesis work, I have explored do-calculus and inference rules, confounding bias, counterfactuals, estimation of causal effect and tools for causal inference.

I would also like to acknowledge Samsung Research Institute - Bangalore for sponsoring my master's (M.Tech) and my team SAIT-India for their guidance and support. Their research exposure has inspired me to always look into the problem from different view - research perspective and industrial perspective, as well. I must also appreciate Suhas PK, for always helping me during this work and co-working on several problem related to thesis work.

Finally, I must express my gratitude to my parents and siblings for their continuous encouragement during my master's and thesis work and for their constant faith in me in accomplishing this work.

## Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>v</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>x</b>
<b>List of Abbreviations</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Prior Work . . . . .	2
1.2 Contribution . . . . .	4
<b>2 Causal Inference</b>	<b>6</b>
2.1 Graphical Model . . . . .	6
2.1.1 d-Separation . . . . .	8
2.1.2 <i>do</i> -operator . . . . .	9

2.2	Causal Bayesian Network . . . . .	9
2.3	Confounding Bias . . . . .	13
2.3.1	Back-door criterion . . . . .	14
2.3.2	Front-door criterion . . . . .	14
2.4	Interventional ( <i>do</i> ) calculus - Rules of Inference . . . . .	15
2.5	Dissolving Simpson's Paradox using the <i>do</i> operator . . . . .	16
2.6	Seven Tools of Causal Inference . . . . .	19
<b>3</b>	<b>Proposed Framework</b>	<b>21</b>
3.1	Individual Feature Perturbation . . . . .	21
3.2	Sensitivity Analysis - Effect of noise on Prediction . . . . .	22
3.3	Interpreting Average Causal Effect (ACE) . . . . .	22
3.3.1	Area Under ACE plot . . . . .	24
3.3.2	Exploring relation between PCA and ACE . . . . .	25
<b>4</b>	<b>Experiments and Results</b>	<b>26</b>
4.1	Dataset Processing and Training . . . . .	26
4.1.1	IRIS Dataset . . . . .	26
4.1.2	Handcrafted Numerals . . . . .	27
4.2	Methodology for IRIS dataset . . . . .	27
4.3	Methodology for handcrafted numerals . . . . .	29



<b>5 Conclusion and Future Work</b>	<b>34</b>
-------------------------------------	-----------

<b>Bibliography</b>	<b>35</b>
---------------------	-----------

## List of Figures

FC2.1 d-separation on a DAG . . . . .	9
FC2.2 Original and Interventional DAG . . . . .	12
FC2.3 DAG with intervention for inference rules . . . . .	16
FC2.4 Simpson's Paradox Example . . . . .	17
FC3.1 Original and Interventional SCM . . . . .	23
FC4.1 ACE of IRIS data on interventional values . . . . .	29
FC4.2 Classification accuracy due to noise corrupting the data . . . . .	30
FC4.3 Handcrafted digits on $\mathbb{R}^{8 \times 8}$ space . . . . .	31
FC4.4 Snapshot of ACE on handcrafted digits 3 and 8 . . . . .	31
FC4.5 Area under absolute ACE and total area . . . . .	32
FC4.6 Snapshot of ACE on handcrafted digits 1 and 7 . . . . .	33
FC4.7 Area under absolute ACE and total area . . . . .	33

## List of Tables

TC4.1 Result of PCA on IRIS dataset . . . . .	28
TC4.2 KL-divergence between original and perturbed distribution . . . . .	28

## List of Abbreviations

<b>ACE</b>	.....	Average Causal Effect
<b>BN</b>	.....	Bayesian Network
<b>CBN</b>	.....	Causal Bayesian Network
<b>CE</b>	.....	Causal Effect
<b>DAG</b>	.....	Directed Acyclic Graph
<b>NN</b>	.....	Neural Network
<b>PCA</b>	.....	Principal Component Analysis
<b>SCM</b>	.....	Structural Causal Model

## CHAPTER 1

### INTRODUCTION

Machine learning models have enormous potential in solving sophisticated problems in variety of domains be it object detection [1], autonomous driving [2], DNA sequence generation [3], speech recognition [4] and language processing [5]. These models are capable of learning complex representations. However, human interpretability of these models has been very challenging due to the Black Box nature of it, thus making these models untrustworthy to be used in critical scenarios like healthcare applications as in disease discovery and diagnosis [6], drug discovery [7], autonomous driving [2] etc. Trusting these models in critical application requires us to be cognizant about the pertinent features and their effectiveness, that the model has learnt. These models would have been validated only on the perceived scenario and lacks to accommodate the unseen scenarios - drifting distribution. Hence there is a high need to strengthen the modelling methods for meaningful predictions from such models. The emerging field in machine learning, Explainable AI [8], aims to address this problem of discovering the Black Box decisions in deep neural networks.

Various Methods have been proposed to overcome the problem of interpretability, each having its own perspective towards Explainable AI [8]. Arguably, the machine learning model takes raw input features and learns the mapping between input and output. The learnt mapping doesn't help to understand how changing a feature would affect

model's prediction. The existing method [9] does an exhaustive search on the feature space, where the permutation of the features each with its own range of values are used to assign the feature with an importance score. This method exponentially increases the computational time and thus, is unscalable. Hence, a much simpler approach, causal inference method has been proposed in this paper which helps us to analyse the causal relation of each feature and has been validated with the feature's sensitivity.

## 1.1 Prior Work

The advancement of Deep Neural Networks has motivated many researchers to investigate feature attribution. Some efforts in this regard DeepLIFT [10], Layerwise Relevance Propagation (LRP) [11], Locally Interpretable Model (LIME) [12], Causal Framework for sequence to sequence to models [13] and Probabilistic based Causal approach [14].

For the analysis of attribution-based approach, [15], contributed axioms that could be used to understand model's prediction. [15] explains fundamental axioms for attribution-based methods. The explainable method must satisfy below axioms to make the prediction transparent and trustworthy -

- **Sensitivity:** A feature is sensitive to the model if it leads to a different prediction from its baseline.
- **Implementation Invariance:** Networks are functionally equivalent, if for every input, the output is same, irrespective of the network implementation.
- **Completeness:** Sum of all the attributions is equal to the difference between model's output at a particular input and its baseline.
- **Symmetry Preserving:** Two or more features would get same attribution if swap-

ping them doesn't change the function.

- **Integrated Gradients:** Instantaneous Gradients averaged along a path from the baseline to a particular feature.

A perturbation-based method, LIME framework, generates local explanation for the decision function. A particular instance is perturbed to get the original model's prediction and based on the output, it attempts to explain the prediction with an interpretable model like – linear model, decision tree etc. The explanation would depend on the particular instance being explained; hence the framework is a local model, generating explanations around an instance. On the other hand, in gradient based approach, one would take the gradient of model's output w.r.t to the feature to assign relevance score. This method could not hold (i) because feature importance is relative to the model and neural-network model fails to reveal how a feature changed the prediction relative to the model not seeing the feature at all, *baseline*, and (ii) the gradient based approach breaks for non-linear models. An ideal baseline is a point on the decision boundary where the predictions are neutral, [15]. To overcome the limitation of gradient based approach, DeepLIFT, takes the model's gradient to explain how output changes, based on change in input from its baseline. [10] have approximated the gradients for explaining the prediction w.r.t baseline, which violates implementation invariance as in [15].

Authors in [16] have categorized interpretability algorithms as-

1. **Traditional interpretability :** This method comprises of inherently interpretable models like decision tree, linear regression, where we extract the explanation from the weights or node values as the model has learnt the data and post-hoc interpretability like LIME, feature visualization, saliency maps methods where we perform the interpretation on the already trained model.
2. **Causal interpretability :** This method comprises of explaining the causal effect of

a model component on the target variable, explanation for counterfactual scenarios, guaranteeing fairness, discovering causal relationship in the data.

Probabilistic based casual approach, [14], a recent work contributed on a causal based attribution method for neural network models which helps to compute the causal effect of each feature on the output. It relies on the work of considering DNN models as an SCM, [17]. Additionally, it includes efficient algorithm to compute the causal effect.

## 1.2 Contribution

We have tried to address the problem of interpretability by proposing perturbation-based framework for causal inference. The objective of the proposed work is *(i) to understand the impact of the input feature individually on the model prediction*, and *(ii) to perform sensitivity analysis for verification of causal attribution*. The results will show that how causal attribution is aligned with known methods such as PCA and other sensitivity analysis methods. Achieving this would help in improving the reliability of the model, thus making it better suitable for deploying in critical scenarios like health-care applications, autonomous driving etc. We also report observation obtained using principal component analysis (PCA) with those obtained using causal analysis.

Authors in [16] have suggested following characteristics in interpreting counterfactual explanation which are related to this work -

- Model prediction on the counterfactual sample must be close to the output from original sample.
- Number of sample perturbed to generate counterfactual sample should be small.



- A counterfactual sample is interpretable if it lies close to the training data distribution.
- Counterfactual instance generation should be fast.
- Each counterfactual samples generated must be different from others.

In order to have a method both causal and interpretable, it is necessary to provide an explanation of how the results have been achieved [16]. There are certain benchmark dataset which are utilized to evaluate the interpretability like ImageNet (ILSVRC) [18], MNIST [19]. However, extending the same on other dataset which does not comprises of human-grounded truth is difficult to verify without formal methods. In the proposed work we attempt to incorporate sensitivity analysis on the trained model to verify the causal inference results.

## CHAPTER 2

### CAUSAL INFERENCE

This chapter lays the foundation for inferring causal relationship. Section 2.1 exhibit the need of graphical models and DAG for causal inference and few properties that would become a building block for any graphical model to be factorized into a conditional distribution. Section 2.2 presents a Causal Bayesian Network and necessary definitions and theorems for a causal model and estimating causal effect. Section 2.3 illustrate on the criterion for adjustment of the variables which would aid in discovering the causal influence among the variables and finally, section 2.4 shows how we can use the probabilistic calculus in refactorizing an interventional distribution into a traditional conditional distribution. Section 2.5 briefly proves the efficacy of do-operator in dissolving famous confounding bias problem - Simpson's paradox. Section 2.6 summarizes with several tools of causal inference with its efficacy in automated reasoning.

### 2.1 Graphical Model

A graph comprises of set of vertices,  $V$ , and set of edges,  $E$ , that connects the pair of vertices. For a graphical model,  $V$  denotes the variables and  $E$  denotes the relationship between the vertices to which it connects. A directed graph includes a directed edge ( $X \rightarrow Y$ ), hence is called directed graph. Additionally, a directed graph without any cycles ( $X \rightarrow Y, Y \rightarrow Z, Z \rightarrow X$ ) is a directed acyclic graph (DAG). Figure FC2.1 illustrate

three different valid structure for DAGs.

Graph can also be used to represent a joint distribution of  $n$  variables –  $P(x_1, x_2, \dots, x_n)$ . This representation is a much efficient approach as a tabulated structure would comprise on  $2^n$  entries when the variables are binary; entries increasing exponentially as the state of the variables increases. Also, a graphical representation would aid in depicting relevant relations between the variables.

Graphical model can be categorized as follows-

- **Undirected Graph** - also known as Markov networks.
- **Directed Graph** – Of these, DAGs are often called Bayesian network because of its ability to decompose joint distribution into conditional distributions as in (Eqn 2.1)

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i | x_1, \dots, x_{i-1}) \quad (\text{Eqn 2.1})$$

Local Markov property suggests that  $X_i$  is independent of other non-descendants once we know the predecessors,  $X_i$  depends on. Hence, r.h.s of (Eqn 2.1) can be rewritten as

$$P(x_i | x_1, \dots, x_{i-1}) = P(x_i | Pa_i) \quad (\text{Eqn 2.2})$$

Here,  $Pa_i$  is called Markovian parents of  $X_i$ .

This leads us to make a formal definition for Markovian parents and a Markov compatibility between a DAG and a joint probability distribution.

**Definition 1.** (Markovian Parents) In a joint distribution,  $P(v)$ , where we have a proper ordering of the variables,  $v$ ;  $Pa_i$  satisfies to be termed as Markovian parents of  $X_i$ ,

iff,  $Pa_i$  is a minimal set of predecessor that induce  $X_i$  independent of all its other predecessor:

$$P(x_i|Pa_i) = P(x_i|x_1, \dots, x_{i-1}) \quad (\text{Eqn 2.3})$$

**Definition 2.** (Markov Compatibility)) A DAG,  $G$ , and a joint probability distribution,  $P$ , are said to be compatible, if factorization of  $P$  as in (Eqn 2.1) is captured by  $G$ . This also means that  $G$  represents  $P$ , or that  $P$  is Markov relative to  $G$

### 2.1.1 d-Separation

A DAG helps in evaluating conditional independencies among the set of variables,  $X$ , and  $Y$ , given  $Z$  by testing if the path between  $X$  and  $Y$  is blocked by  $Z$ . Here, the path depicts sequence of edges, irrespective of directionality and “block” is to be interpreted as blocking the flow of information between the variables that are connected by such paths. Any such path is called a d-separated path. In contrast, the path which allow the flow of information is called an active trail.

**Definition 3.** (d-Separation) Two disjoint set of nodes,  $X$ , and  $Y$ , are said to be d-separated by  $Z$ , iff,

1. the path between  $X$  and  $Y$  contains either a chain  $a \rightarrow b \rightarrow c$  or a fork  $a \leftarrow b \rightarrow c$ , where  $a \in X, b \in Z, \text{ and } c \in Y$
2. the path between  $X$  and  $Y$  contains a collider (inverted fork)  $a \rightarrow b \leftarrow c$ , such that  $b$  or its descendant  $\notin Z$

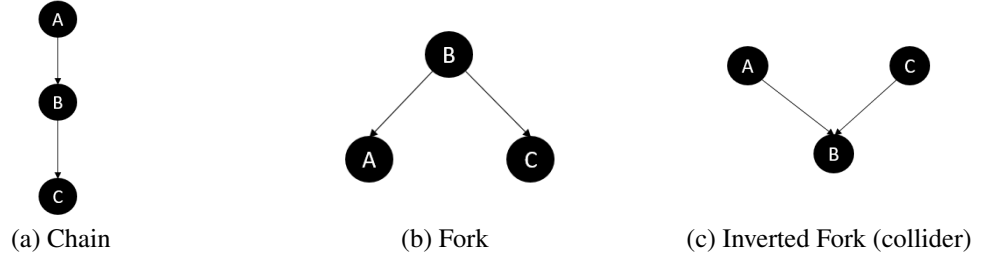


Figure FC2.1: illustrate different structure of DAG to evaluate d-separation. (a) and (b) depicts A and C are d-separated given we know about B, i.e., when we condition on B. (c) shows block in flow of information between A and C in absence of B

### 2.1.2 *do*-operator

Evaluating a conditional probability,  $P(Y = y|X = x)$ , seems a straight forward task by conditioning the distribution of  $Y$  on observation,  $X = x$ , which can be evaluated from the Bayesian network. However, the interventional probability,  $P(Y = y|do(X = x))$ , can be evaluated from the augmented graph, where all the links to  $X$  are removed and its value is set to  $x$ ; this is called ‘*doing*’, which is captured in the probabilistic equation via the ‘*do*’ operator.

## 2.2 Causal Bayesian Network

As we have seen, Bayesian network, though, aids in decomposing joint distribution to conditional one and using  $d$ -separation we can further factor the distribution into conditional independencies, however, these factorizations don’t suffice in inferring causal relationship. They merely render upon associational knowledge among the variables. Hence, we need a sound construct over Bayesian network to interpret causation

**Definition 4.** Given a joint distribution,  $P(v)$ , on set of variables,  $V$ , and  $P_x(v)$  represents an interventional distribution after clamping a subset of variables,  $X$ , to a constant,  $x$ . Let  $P_*$  represents all interventional distribution,  $P_x(v)$ , including  $P(v)$  which resembles no intervention. Under this consideration, a DAG  $G$  is said to be a causal

Bayesian network with  $P_*$ , iff, following conditions holds true:

- (i)  $P_x(v)$  is Markov relative to  $G$
- (ii)  $P_x(v_i = 1, \forall V_i \in X)$  i.e., interventional node still satisfies probability distribution
- (iii)  $P_x(v_i|pa_i) = P(v_i|pa_i), \forall V_i \notin X$  i.e.,  $P(v_i|pa_i)$  remains unaffected to intervention that does not involve  $V_i$

Extending condition (iii), we can write the factorization of an interventional distribution as,

$$P_x(v) = \prod_{\{i|V_i \notin X\}} P(v_i|pa_i) \quad (\text{Eqn 2.4})$$

Based on the above definitions, we can decompose the causal Bayesian network into a functional causal model comprising of set of equations as

$$x_i = f_i(pa_i, u_i), \forall i \in \{1, 2, \dots, n\} \quad (\text{Eqn 2.5})$$

Here,  $pa_i$  (parents of  $i$ ) are set of variables that directly influence  $X_i$  and  $U_i$  represents random disturbances. If these unmeasured quantities are mutually independent, then the model is termed as Markovian. The equation of the above form where each variable,  $x_i$ , has a functional equation on its set of parents and disturbances is called a structural causal model and the network which was decomposed is called a structural model.

The structural model is the basis for causal model, which reveals how each variable is influenced by its predecessors in the network represented by a DAG. The unmeasured random disturbances play a vital role in altering the causal relationship among the variables.

**Definition 5. (Causal Model)** A causal model is a tuple,  $M = \langle D, \Theta_D \rangle$ , where,  $D$  depicts causal structure and  $\Theta_D$  represents parameter set compatible with  $D$ . The parameter set

assigns function,  $x_i = f_i(pa_i, u_i), \forall X_i \in V$  and a probability distribution,  $P(u_i)$ , to each  $u_i$ , where  $PA_i$  denotes set of parents of  $X_i$  in  $D$  and  $U_i$  denotes random disturbances according to  $P(u_i)$ , and  $u_i$  are mutually independent.

A causal model aids in discovering a path,  $(X \rightsquigarrow Y)$ , from any variable, say  $X$ , to any other variable, say  $Y$ , to establish a causal relationship between them i.e., to state if  $X$  influences  $Y$ . Although, there can be many causal structures which would generate a same joint distribution, hence, any minimal structure which could guarantee the consistency of the structure and also comprises of the path,  $(X \rightsquigarrow Y)$ , could help in inferring causal relationship between them.

Now, once we have presumed the causal structure among the variables, we would want to estimate the effect of a particular variable,  $X_i$ , on another variable,  $X_j$ . Evaluating causal effect depends on the type of intervention – changing the mechanism by which  $X_i$  influences. For example, as in (Eqn 2.5), altering the functional mechanism of  $X_i$  with a random value,  $x_i$ , would augment the causal structure in the DAG by eliminating all the immediate causes of  $X_i$ ; also called an atomic intervention.

**Definition 6.** (Causal Effect) For a disjoint set of variable,  $X$ , and  $Y$ , the causal effect of  $X$  on  $Y$  is denoted as  $P(y|do(x))$ . This interventional probability reveals a change in functional model in (Eqn 2.5) by substituting all the equations corresponding to variables in  $X$  by  $X = x$ .

The definition 6 of causal effect illustrate on an augmented graph where all the immediate cause (directed arrows) to  $X$  would be clamped to a fixed value,  $x$ . Thus, evaluating a causal effect (C.E) w.r.t another realization of  $X$ , say  $x'$ , would be estimating the average deviation on  $Y$  given  $X$  is intervened with  $x$  and  $x'$ , valid realization of  $X$ .

$$C.E = \mathbb{E}[Y|do(x)] - \mathbb{E}[Y|do(x')] \quad (\text{Eqn 2.6})$$

Apparently, another type of intervention is as a variable. This can be achieved by encapsulating the functional model in (Eqn 2.5) with another function, such that, in absence of any intervention, the original functional relationship between  $X_i$  and  $pa_i$  becomes evidentially visible as

$$x_i = I(pa_i, f_i, u_i) \quad (\text{Eqn 2.7})$$

where  $I$  is a three-argument function satisfying  $I(a, b, c) = f_i(a, c)$  whenever  $b = f_i$ .

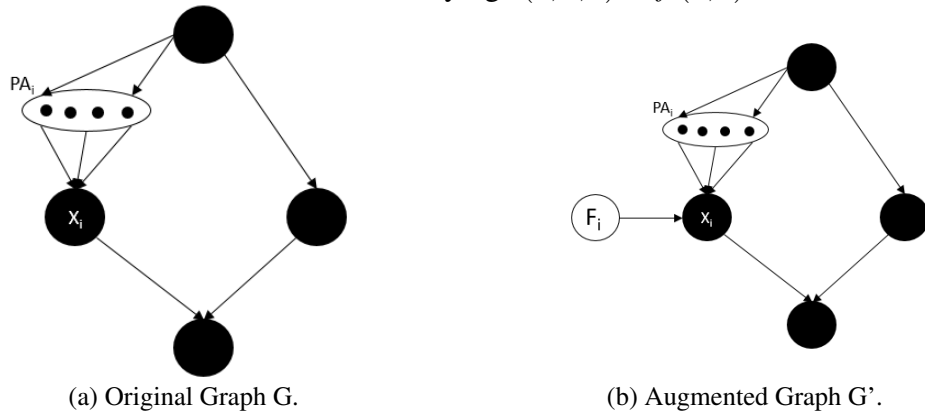


Figure FC2.2: illustrate original (a) and augmented graph (b), where  $X_i$  is intervened with  $F_i$

The effect of an intervention [20] is illustrated in the Figure FC2.2. The augmented graph depicts an additional link to show the effect of an intervention in the causal structure. Furthermore, in an augmented graph,  $Pa'_i = Pa_i \cup \{F_i\}$ , and the conditional probability in  $X_i$  can be estimated by

$$P(x_i | pa'_i) = \begin{cases} P(x_i | pa_i) & \text{if there is no intervention, i.e., } F_i = \textit{idle} \\ 0 & \text{if different node is intervened, i.e., } F_i = \textit{do}(x_i \text{ and } x_i \neq x'_i) \\ 1 & \text{if evaluated for an interventional node, i.e., } F_i = \textit{do}(x_i \text{ and } x_i = x'_i) \end{cases} \quad (\text{Eqn 2.8})$$

The augmented model due to an intervention represents a transformed distribution



called as an interventional distribution which can thus be formally factorized as below

$$P(x_1, \dots, x_n | do(X = x')) = \begin{cases} \prod_{j \neq i} P(x_j | pa_j) & \text{if } x_i \text{ is intervened with } x'_i, \text{ i.e., } x_i = x'_i \\ 0 & \text{if } x_i \text{ is intervened with } x'_i, \text{ i.e., } x_i \neq x'_i \end{cases} \quad (\text{Eqn 2.9})$$

Further, the effect of an intervention on a particular variable,  $X$ , on any other variable,  $Y$ , which is neither parent nor descendants of  $X$  can be formulated only in terms of its direct cause as -

$$P(y | do(X = x_i)) = \sum_{pa_i} P(y | x_i, pa_i) P(pa_i) \quad (\text{Eqn 2.10})$$

As in (Eqn 2.10),  $P(y | x_i, pa_i)$  and  $P(pa_i)$  represents pre-interventional distribution

Although, we are able to adjust an interventional distribution in terms of its direct cause (Eqn 2.10), we must take into account that the process of discovering causal relationship between any pair of variables is only possible when all the parents of the variable which is intervened are observable. This is concisely captured in the theorem as

**Theorem 2.2.1. (Causal Identification)** *In a Markovian model, provided a causal graph,  $G$ , with a set of vertices,  $V$ , the causal effect,  $P(y | do(X = x))$ , is identifiable, iff,  $\{X \cup Y \cup Pa_X\} \subseteq V$ , i.e., all the variables are observable, and hence, can be adjusted by (Eqn 2.10)*

## 2.3 Confounding Bias

As we have seen in the previous definitions, we can adjust the effect of an influence by intervening on its predecessors. However, there are certain situations where it is difficult to label a disjoint set of parents to estimate an interventional effect – confounding

bias between the two variables; among which we intend to find an interventional effect. We will see in section 2.5 how the confounding bias could lead to a paradox - Simpson's paradox and how it can be resolved using *do*-operator from 2.1.2. This bias can be resolved by adjusting an appropriate set of variables as indicated by the back door and front door criteria, depending on the causal model.

### 2.3.1 Back-door criterion

For any pair of variable,  $X_i$ , and  $X_j$ , where  $X_j$  being the descendant of  $X_i$ , there exists a back door path comprises of set of variables,  $Z$ , iff,

1.  $z$  is not a descendant of  $X$ ,  $\forall z \in Z$
2.  $Z$  blocks every path between  $X_i$ , and  $X_j$  that contains an arrow into  $X_i$

The causal effect of  $X$  on  $Y$  when  $Z$  comprises of a set of variables that satisfies back door criterion relative to  $(X, Y)$  can be estimated by -

$$P(y|do(x)) = \sum_z P(y|x, z)P(z) \quad (\text{Eqn 2.11})$$

Figure FC2.2 (a) illustrate back door path relative to  $(X, Y)$  comprises of nodes in  $U$ . (b) and (c) depict no back door path relative  $(X, Y)$ .

### 2.3.2 Front-door criterion

For any pair of variable,  $X_i$ , and  $X_j$ , where  $X_j$  being the descendant of  $X_i$ , there exists a back door path comprises of set of variables,  $Z$ , iff,

1. All directed paths from  $X_i$  to  $X_j$  are intercepted by  $Z$

2. No back-door paths from  $X_i$  to  $Z$  are unblocked
3. All back-door paths from  $Z$  to  $X_j$  are blocked by  $X_i$

Furthermore, the causal effect of  $X$  on  $Y$  when  $Z$  comprises of a set of variables that satisfies front door criterion relative to  $(X, Y)$  can be estimated by-

$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z) P(x') \quad (\text{Eqn 2.12})$$

Figure [FC2.2](#) (a) and (b) illustrate front door path relative to  $(X, Y)$  comprises of nodes in  $Z$ . (c) depicts no front door path relative to  $(X, Y)$ .

## 2.4 Interventional (*do*) calculus - Rules of Inference

Till now, we have seen what causal model is capable of and how an intervention on a particular set of variables could aid us in grasping its influence on a target variable. The effect of intervention leads to estimation in an augmented graph which is factorized to an interventional distribution comprising of  $do(\cdot)$  operator. In this section, we will see how we can use the rules of inference to transform the post-interventional distribution into a traditional conditional distribution.

For a given causal model,  $G$ , which is compatible with a joint distribution,  $P(\cdot)$ , we have following inference rules for a disjoint subset of variables,  $X, Y, Z, W$

Rule 1 (Insertion or deletion of observations)

$$P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if } ((Y \perp\!\!\!\perp Z)|X, W)_{G_{\bar{X}}}$$

Rule 2 (Action or observation change)

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \quad \text{if } ((Y \perp\!\!\!\perp Z)|X, W)_{G_{\bar{X}Z}}$$

Rule 3 (Insertion or deletion of actions)

$$P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}, \overline{Z(W)}}}$$

where,  $Z(W)$  is the set of nodes in  $Z$  which are not ancestors of any  $W$  nodes in  $G_{\bar{X}}$

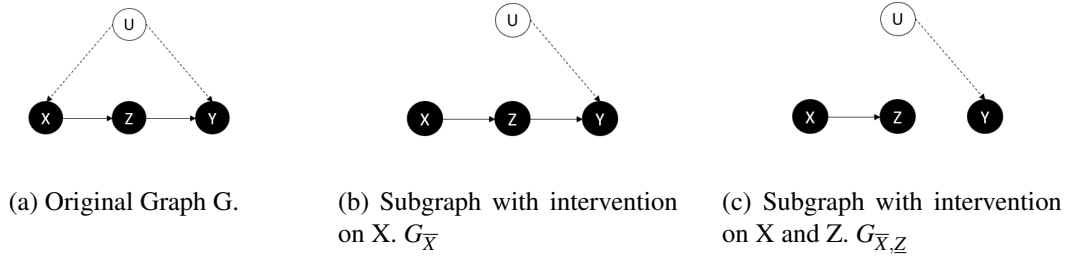


Figure FC2.3: illustrate original (a) and augmented graph due to an intervention on (b)  $X$  and (c)  $X$  and  $Z$ . Here,  $U$  represents an unobserved variable, hence connected via dashed arrows.

Rule 1 suggests the application of d-separation. By intervening on  $X$ , the causal effect on  $Y$  becomes irrelevant due to  $Z$  as in Figure FC2.3 (b)  $G_{\bar{X}}$  which eliminates back-door path between  $Z$  and  $Y$ , hence can be removed from the distribution as in r.h.s of rule 1.

Rule 2 assures that intervening on  $Z$  has same effect on  $Y$  as observation on  $Z$ . This is due to the fact that  $G_{\bar{X}, \underline{Z}}$  eliminates all the paths between  $Z$  and  $Y$  as in Figure FC2.3 (c)

Rule 3 claims that external intervention on  $Z$  won't affect  $Y$ , in case,  $Y$  and  $Z$  are d-separated in an augmented graph  $G_{\bar{X}}$ , which would result in  $G_{\bar{X}, \overline{Z(W)}}$

## 2.5 Dissolving Simpson's Paradox using the *do* operator

Simpson's paradox [21] refers to statistical observation where an event, say  $C$ , increases the probability of an event, say  $E$ , in a given population,  $p$ , however, the same event,  $C$ , decreases the probability of the same event,  $E$ , in every subpopulation of  $p$ .

This can be stated mathematically in below equations -

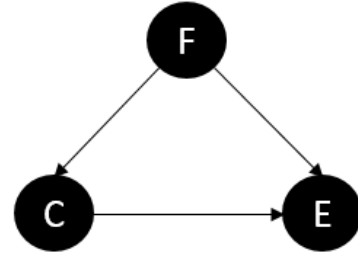
$$P(E|C) > P(E|\neg C) \quad (\text{Eqn 2.13})$$

$$P(E|C, F) < P(E|\neg C, F) \quad (\text{Eqn 2.14})$$

$$P(E|C, \neg F) < P(E|\neg C, \neg F) \quad (\text{Eqn 2.15})$$

Combined		$E$	$\neg E$	Recovery Rate	
(a)	Drug ( $C$ )	20	20	40	50%
	No drug ( $\neg C$ )	16	24	40	40%
		36	44	80	
Males		$E$	$\neg E$	Recovery Rate	
(b)	Drug ( $C$ )	18	12	30	60%
	No drug ( $\neg C$ )	7	3	10	70%
		25	15	40	
Females		$E$	$\neg E$	Recovery Rate	
(c)	Drug ( $C$ )	2	8	10	20%
	No drug ( $\neg C$ )	9	21	30	30%
		11	29	40	

(a) Recovery rate (*E*) under drug (*C*) for (a) combined population, (b) male subpopulation and (c) female subpopulation.



(b) Causal Model of Simpson Paradox

Figure FC2.4: illustrate (a) an example of Simpson's Paradox on drug-effect on population and (b) represents a possible causal model for example in (a)

As an example, Figure FC2.4 illustrate on usage of drug (*C*) in treatment (*E*) of a population (*F*). (a) depicts drug (*C*) being effective in treatment of combined population. However, (b) and (c) clearly shows male and female, subpopulation of (*F*), having lesser recovery rate on using drug (*C*). The interpretation of data in the Figure FC2.4 leads to a paradoxical statement, i.e., a drug is effective in the treatment of a population when the gender is unknown, however, the same drug has an opposite effect when treated on male or female subpopulation.

Since, we are interested in the effect of the drug, we must be careful with the probabilistic terms as in (Eqn 2.13 - Eqn 2.15). These equations merely reflect upon the

statistics observed so far while using the drug (C). The effect is actually captured when we intervene on the distribution to see the impact on the recovery (treatment, E), which is captured by the *do* operator -

$$P(E|do(C)) > P(E|do(\neg C)) \quad (\text{Eqn 2.16})$$

The causal structure in Figure FC2.4 (b) suggests the gender(F) being a confounder. [20] suggests to hold the confounder fixed in such cases for proper estimation of the causal effect. In this example, we can evaluate for both sides of inequality in (Eqn 2.16) using male ( $\neg F$ ) and female (F) population separately. For an unbiased effect due to gender, let's assume drug (C) does not affect either of the subpopulation's distribution, i.e.,

$$P(F|do(C)) = P(F|do(\neg C)) = P(F)$$

Based on the explanation above on how effective the drug is on either of the subpopulation as compared to no drug at all, we can write them as following inequality for each of the subpopulation.

$$P(E|do(C), F) < P(E|do(\neg C), F)$$

$$P(E|do(C), \neg F) < P(E|do(\neg C), \neg F)$$

Probability of recovery when drug is used -

$$\begin{aligned} P(E|do(C)) &= P(E|do(C), F)P(F|do(C)) + P(E|do(C), \neg F)P(\neg F|do(C)) \\ &= P(E|do(C), F)P(F) + P(E|do(C), \neg F)P(\neg F) \end{aligned}$$

Probability of recovery when drug is not used -

$$\begin{aligned} P(E|do(\neg C)) &= P(E|do(\neg C), F)P(F|do(\neg C)) + P(E|do(\neg C), \neg F)P(\neg F|do(\neg C)) \\ &= P(E|do(\neg C), F)P(F) + P(E|do(\neg C), \neg F)P(\neg F) \end{aligned}$$

The data in the Figure FC2.4 (a), thus, clearly suggests that the drug has an adverse effect on the population as a whole which dissolves Simpson's paradox.

$$P(E|do(C)) < P(E|do(\neg C))$$

## 2.6 Seven Tools of Causal Inference

Authors in [22] offers seven tools which summarizes the capability of causal inference towards automated reasoning -

1. Encoding causal assumptions : Graphical model presumed with a causal structure provides a way to infer the causal relationship and d-separation aids in locating the dependencies within the model, aligned with the data.
2. *Do*-calculus and the control of confounding : We have seen how confounding bias could lead to paradoxical statements which brings spurious causal inference and how *do*-operator handles the confounding bias by focusing on 'doing' rather than 'seeing'.
3. The algorithmization of counterfactuals : Counterfactual deals in altering the situation to estimate its probability, as in counter-world. It is related to finding "causes of effects" rather than "effects of causes".
4. Mediation analysis and the assessment of direct and indirect effects : Used for generating explanations of the sort - what fraction of the effect of X on Y is mediated by Z?

5. Adaptability, external validity, and sample selection bias : Bias in model training and deployment environment is addressed by *do*-calculus; as it can adjust the variable to render the change in distribution for desired effect.
6. Recovering from missing data : Causal model offers to recover causal relationship from missing data under formalized conditions.
7. Causal discovery : d-separation, functional decomposition and spontaneous local changes offers to detect set of causal models compatible with the data.



## CHAPTER 3

### PROPOSED FRAMEWORK

We have considered three- and four-layer neural network to demonstrate the causal inference of the input. The representation of neural network is of a directed acyclic graph (DAG) and hence, can be modelled as an SCM [17]. The causal attribution for each feature has been evaluated using definition 6 in (Eqn 2.6), where baseline has been taken as the mean of the individual features. Since model's behavior depend on the realization of how it has been trained given the stochastic nature of weight initialization, learning parameters and convergence to final parameters, we have demonstrated the effectiveness of the causal inference with sensitivity analysis to validate our findings and solidify on how causal inference can be used to uncover the “black-box” nature of neural network like models.

Following methods are employed to help to interpret and verify the causal inference in machine learning domain.

#### 3.1 Individual Feature Perturbation

Perturbing an individual feature while keeping other features intact would help to analyze how a particular feature causes model to predict the output and helps to understand the particular feature distribution. We have perturbed each feature with a Gaussian

noise as in equation (Eqn 3.1)

$$\mathbb{X} = \mathbb{S} + \varepsilon, \quad \varepsilon \in \mathcal{N}(\mu, \sigma^2). \quad (\text{Eqn 3.1})$$

In (Eqn 3.1),  $\mathbb{S}$  is the original signal,  $\mathbb{X}$  is the signal after perturbation and  $\varepsilon$  is the noise. Different value of  $\sigma$  would result in different noise energy. We have used *KL-Divergence* to measure level of corruption by additive noise, higher the value, greater the distance between the original distribution and the new distribution after perturbation.

### 3.2 Sensitivity Analysis - Effect of noise on Prediction

This is inline with the counterfactual principles of causal inference, according to Pearl's highest level of interpretability [22],  $p(y'|x', x, y)$ , where we aim to predict the output,  $y'$ , with the perturbed sample,  $x'$ , given, we know the model's output,  $y$ , for the input sample,  $x$ . The neural network-based DL models,  $f: \mathbb{R}^n \rightarrow \mathbb{R}^k$ , learn the mapping between the n-dimensional feature and k-dimesnsional output. Though the function could have implicitly learnt some tolerance, in the process of learning the mapping, it depends on the validation set used for cross examining the parameters during the backpropagation process. Hence, the robustness of the training dataset severely affects the model performance. For a counter approach, adding noise to the relevant feature could help us to understand the model's sensitivity on a particular feature.

### 3.3 Interpreting Average Causal Effect (ACE)

The rules of causal inference aid in estimating the Average Causal Effect (ACE) of an individual feature on the model function,  $f(x)$ , by intervening the particular input feature node with a distribution function as in Figure FC3.1(a). The illustration in

Figure FC3.1(b) indicates one of the input feature,  $x_2$ , has been intervened by clamping it to  $x_0$  (interventional value).

$$f(x) = p(y|x_1, x_2, x_3) \quad (\text{Eqn 3.2})$$

$$f(x) = p(y|x_1, do(X_2 = x_0), x_3) \quad (\text{Eqn 3.3})$$

(Eqn 3.2) represents the conditional probability of  $y$  when  $x_1$ ,  $x_2$ , and  $x_3$  are observed. (Eqn 3.3) represents the interventional probability of  $y$  when  $x_1$  and  $x_3$  are observed, and  $x_2$  is intervened, by clamping its value to  $x_{fixed}$ , which can be evaluated using inference rules of *do*-calculus [23].

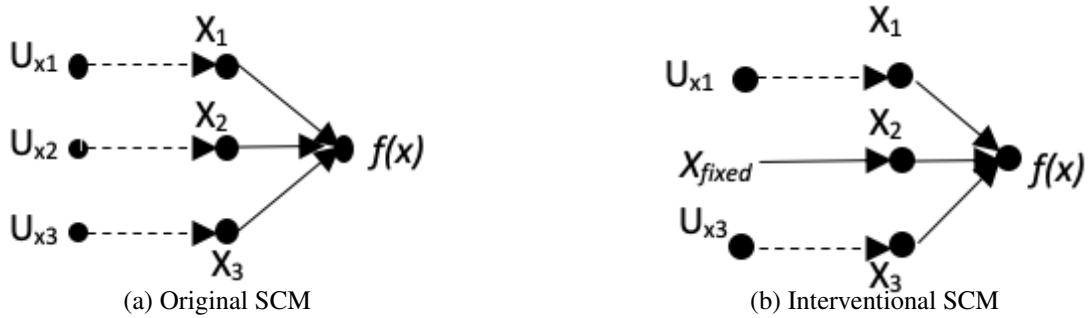


Figure FC3.1: SCM. Unobserved (exogenous) variables  $U_i$  are connected by dashed arrows. Each input feature  $X_i$  has a corresponding unmeasurable influencing variable,  $U_i$ . (a) illustrates original SCM encoding the presence of causal influence of  $X$  on  $f(x)$  by  $X \rightarrow Y$  and absence of any influence of  $f(x)$  on  $X$ . (b) depicts, one of the feature  $X_2$  intervened with a random value  $x_{fixed}$ , and hence called an interventional SCM.

An intervention on a particular input feature with a continuous domain would, in effect, change the distribution of the output function. The analysis of this interventional distribution would help us to understand the causal attribution of the particular feature like - how the causal attribution changes over the intervention and what is the strength of the causation in a particular interventional range. The causal attribution is captured by estimating the ACE for each feature w.r.t output. The ACE plot depicts feature-wise

average deviation of the causal attribution from the baseline for a range of interventional values.

### 3.3.1 Area Under ACE plot

Correlation is a measure of the degree to which two variables are related. *For example* - regression, conditional independence. However, causation is a process of exploring the cause-effect relationship between the variables. *For example* - intervention, explanation, attribution. Hence, causal analysis can't be explored from the observational distribution alone. In other words, correlation does not imply causation [24, 25]. Hence, we would like to explore some of the measures using ACE plot. Here, we have considered following 2 measures-

1. Total area under ACE (T\_ACE) as in (Eqn 3.4a).
2. Area under absolute value of ACE, (A\_ACE), using only the magnitude of ACE as in (Eqn 3.4b).

$$T\_ACE = \int_{\alpha_1}^{\alpha_2} y d\alpha \quad (\text{Eqn 3.4a}) \quad A\_ACE = \int_{\alpha_1}^{\alpha_2} |y| d\alpha \quad (\text{Eqn 3.4b})$$

(Eqn 3.4a) represents total area under ACE and (Eqn 3.4b) represents area under absolute value of ACE, from  $\alpha_1$  to  $\alpha_2$ . We are interested in these  $\alpha$  values as they don't alter the model's prediction. These values are computed for each feature separately, i.e, as many features those many ACE plots.

### 3.3.2 Exploring relation between PCA and ACE

PCA computes decorrelated transformed axes - each of which is a linear combination of the original features; this is in contrast to causal inference that measures significance by d-separating each individual feature from all of the others. On employing PCA on a dataset yields the principal directions in the feature space, indicated by the vectors along which maximal values of variance is observed. Each of these principal directions is indeed a linear combination of the features themselves. The first principal component captures the direction of first highest variance across the entire dataset (collapsing all classes) and is not expected to carry any class-specific contributions. On the contrary, the least principal component captures least variance and could be expected to contain class-specific contributions. In this work, we attempt to look at the relation between PCA and ACE, as both are used for feature analysis.

## **CHAPTER 4**

### **EXPERIMENTS AND RESULTS**

In our work, we have attempted to uncover the black-box nature of neural network-based model with the open dataset: IRIS [26] and handcrafted numerals, to reduce the feature space as in MNIST [19].

#### **4.1 Dataset Processing and Training**

##### **4.1.1 IRIS Dataset**

1. Dataset is normalized using MinMaxScaler [27]
2. Target variable is transformed via OneHotEncoding [28]
3. A 3-layer dense network trained with a relu() activation function
4. A stochastic gradient descent optimizer is used with a learning rate of 0.01 with a cross-entropy loss function.
5. An epoch of 1000 is used for training
6. Training-Validation Split : Random split with training size of 80%

#### 4.1.2 Handcrafted Numerals

1. Each digits are handcrafted in  $\mathbb{R}^{8 \times 8}$  space by enabling the corresponding pixels for each digits as illustrated in Figure FC4.3 (b) - (k).
2. Employed random rotation of maximum 15 degree and random shifting of 0.2 the pixels to create 51 images for each digits
3. A 4-layer dense network trained with a `relu()` activation function at hidden layers and `logsoftmax()` activation function at the output layer.
4. A stochastic gradient descent optimizer is used with a learning rate of 0.03 and momentum of 0.9
5. A negative log likelihood loss function is used
6. An epoch of 50 is used for training
7. Training-Validation Split : Random split with training size of 35 images and validation set of 16 images.

#### 4.2 Methodology for IRIS dataset

We have taken IRIS dataset [26] as a starting problem to validate the causal inference principles, since this is an open dataset with small dimensional feature space,  $\mathbb{R}^4$ , and hence easy for validation and visualization. The four features are – Sepal Length (SL), Sepal Width (SW), Petal Length (PL), and Petal Width (PW). We have attempted to reproduce the results from [14] for sensitivity validation. Figure FC4.1 shows PW's low value, has highest causal attribution for Iris-setosa class, moderate values of PW causes model to label samples to Iris-versicolor class and higher values of PW attributes to Iris-virginica class.

Table TC4.1: Result of PCA on IRIS dataset. 1<sup>st</sup> row represents the 1<sup>st</sup> principal component vector,  $V_1$ , and the  $i^{th}$  row depicts the  $i^{th}$  principal component vector,  $V_i$ , with the respective contributive weights from each of the feature.

Principal Component (PC)	Feature Weights			
	Sepal Length (SW)	Sepal Width (SW)	Petal Length (PL)	Petal Width (PW)
$V_1$	0.36	-0.08	0.85	0.35
$V_2$	0.65	0.73	-0.17	-0.07
$V_3$	-0.58	0.59	0.07	0.54
$V_4$	0.31	-0.32	-0.47	0.75

Table TC4.2: Illustrates KL-divergence between original and perturbed distribution. Clearly “PW” suffers the maximum increase in distance for the same perturbation level, as compared to, other features, when the feature is corrupted with the noise variance =1. “PL” closely follows “PW” when perturbed. “SW” shows minimum change in distribution among all others.

Feature	KL-divergence
Sepal Length (SL)	0.42
Sepal Width (SW)	0.29
Petal Length (PL)	0.73
Petal Width (PW)	0.75



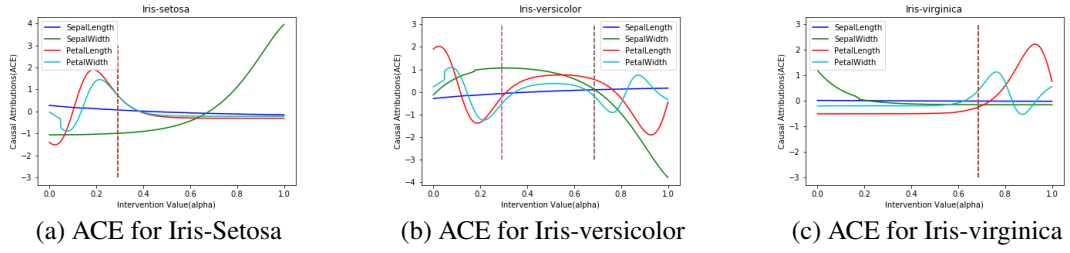


Figure FC4.1: Causal attribution for various interventional values. The vertical line in each plot illustrate the valid interventional range of "PW",  $\alpha$ , for each class. (a) Intervening "PW" between 0 - 0.29 would make the sample being classified as Iris-setosa, (b) interventional range of "PW" from 0.30 to 0.69, would classify the sample as Iris-versicolor and (c) "PW's" value from 0.69 to 1.0 would classify the sample as Iris-virginica.

The set of four feature defines IRIS dataset whose PCA, [27], analysis yields 4 principal directions,  $principalcomponents(PC) \in \{V_1, V_2, V_3, V_4\}$  as tabulated in table TC4.1. The initial set of principal components,  $\{V_1, V_2\}$ , capture the highest variance. It is interesting to know that the least principal component,  $V_4$ , clearly shows, "PW" has maximum contribution which has also been observed in the ACE plot in Figure FC4.1. "PL" being the second highest contributor in  $V_4$  and subsequently, "SW" and "SL". This order of the feature contribution is inline with the ACE observation in Figure FC4.1 Illustration of sensitivity of each feature in Figure FC4.2 confirms that "PW" is the most significant feature. Furthermore, table TC4.2 depicts KL-divergence between original and perturbed distribution after addition of noise on each feature individually. "PW" suffers maximum deviation from its original distribution and "PL", "SL", "SW" follows respectively. This observation is also aligned with other sensitivity analysis approach we have discussed earlier and supports our observation from causal attribution.

### 4.3 Methodology for handcrafted numerals

The purpose of creating the handcrafted numerals is to inject human level understanding of digits, the significant pixels for classification of a particular digit, as an implicit ground truth in the dataset. The steps to create the dataset and the model to

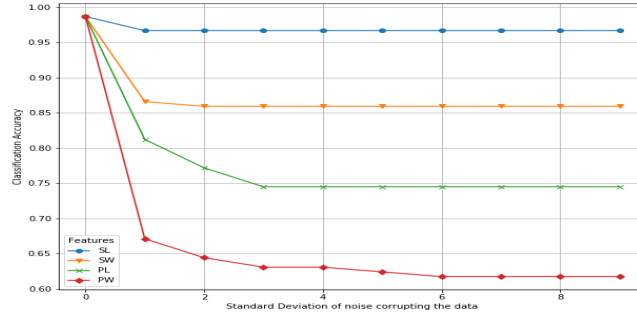


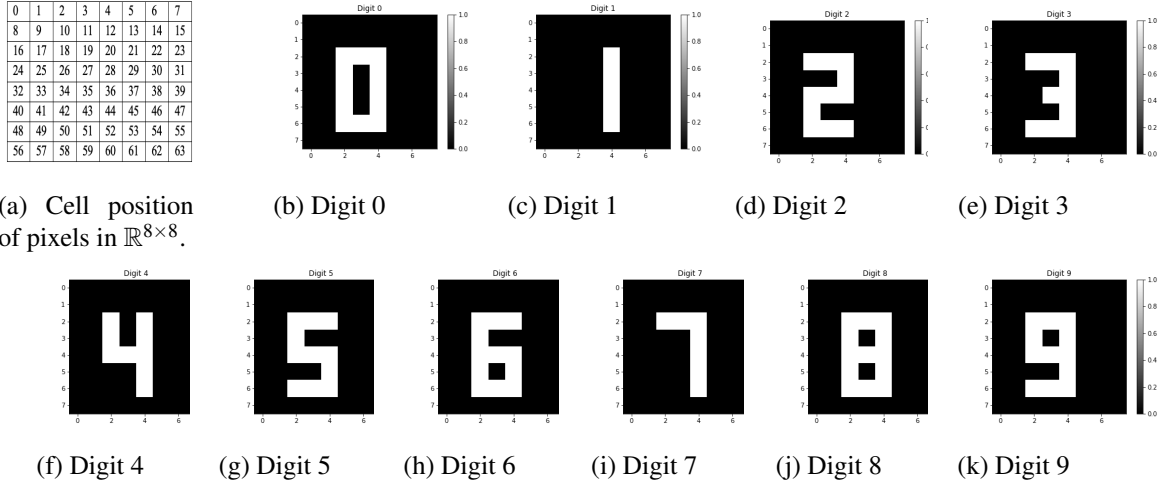
Figure FC4.2: Illustrates the impact of perturbing each feature individually on Classification accuracy. X-axis shows the perturbation level (Noise Energy) while the Y-axis shows the corresponding classification accuracy, for each feature individually. Among the 4 features, the feature "PW" suffers the maximum drop in accuracy, from 0.97 to 0.68, when the feature is corrupted with Noise variance = 1. This could be indicative of the critical importance of this feature for the task of classification, as compared to the other features. The drop-in classification accuracy for similar perturbation of "PL" is from 0.97 to 0.82, for "SW" the drop is from 0.97 to 0.86, and for "SL" the drop is from 0.97 to 0.96. The minimal drop in classification accuracy is exhibited by the feature "SL", indicating that it is the least important of features, for the task of classification. For each of the features, the classification accuracy monotonically decreases as the Noise variance increases, as expected.

perform the training is mentioned in section 4.1.1.

Figure FC4.3 illustrate (a)  $\mathbb{R}^{8 \times 8}$  space with cell position for each pixels used to generate the digit images. (b) - (k) represents image for each digit 0 - 9 by enabling certain pixels as in (Eqn 4.1)

From the generated dataset, we have derived list of significant pixels as in (Eqn 4.1). All other pixels are considered as background pixels. For simplicity of analysis, we have focused on 2-class classification problem, i) 1 vs 7 and, ii) 3 vs 8. We have deliberately chosen these classes because they have certain pixels which contribute to both the classes, only handful of them could differentiate among the classes, hence are called significant pixels for the corresponding class,  $S$ .

$$S = \{p_{ij} | p_{ij} = 1 \forall i, j \in \mathbb{R}^{8 \times 8}\} \quad (\text{Eqn 4.1})$$

Figure FC4.3: Handcrafted digits on  $\mathbb{R}^{8 \times 8}$  space

Using (Eqn 4.1), we have derived union of significant pixels for the digits, as,

$$S_{8,3} \in \{18, 19, 20, 26, 28, 34, 35, 36, 42, 44, 50, 51, 52\}$$

$$S_{7,1} \in \{18, 19, 20, 28, 36, 44, 52\}$$

Here,  $S_{8,3}$  represents set of significant pixel for digits 8 and 3, and  $S_{7,1}$  represents set of significant pixel for digits 7 and 1.

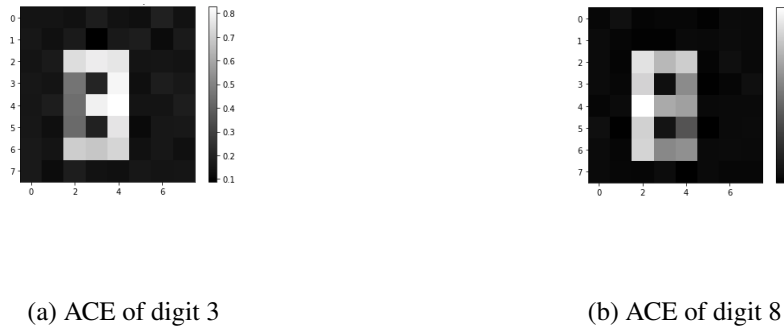


Figure FC4.4: Snapshot of ACE on handcrafted digits 3 and 8.

Figure FC4.4 illustrate the snapshot of ACE of each pixel after intervening on class neuron 3 as in (a) and class neuron 8 as in (b). The colorbar indicates the causal attribution of each pixel which clearly reveals the significant pixels for digits 3 and 8.

It has been observed that the highest weightage to the most significant pixel (42) has been assigned to a lower principal component (beyond  $V_{15}$ ), on performing PCA on digit dataset.

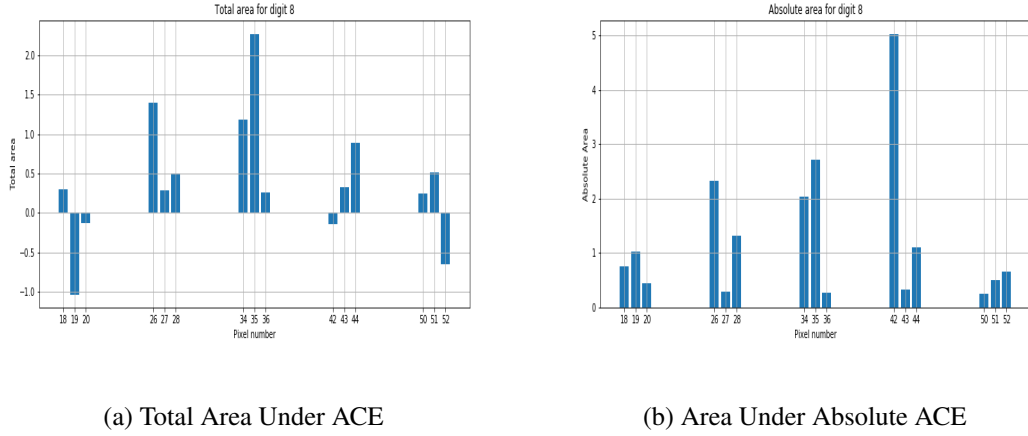
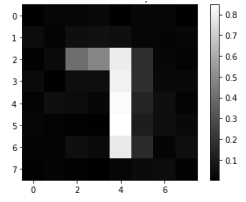


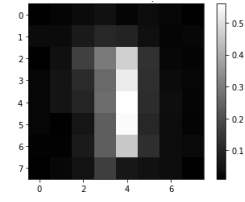
Figure FC4.5: Area under absolute ACE and total area.

Figure FC4.5 illustrates (a) total area and (b) area under absolute ACE of the list of significant pixels for digit 8. Pixel 20 has near to smallest  $A_{ACE}$  and is comparable to  $T_{ACE}$  of pixel 42. However, perturbing pixel 42 leads to most misclassification while perturbing pixel 20 leads to none, which can be associated with their respective area under absolute ACE,  $A_{ACE}$ . This leads to an implication that  $T_{ACE}$  could be misleading in an effort to discover the significant feature. The observation signifies that area under absolute ACE,  $A_{ACE}$ , is more important than the total area,  $T_{ACE}$ . Additionally, we have used the decision tree method for classification which reveals pixel 42 as the root node. Also, pixel 35 shows the highest total area,  $T_{ACE}$ , and also the highest causal attribution in ACE analysis.

Figure FC4.6 illustrate the snapshot of ACE of each pixel after intervening on class neuron 7 as in (a) and class neuron 1 as in (b). The colorbar indicates the causal attribution of each pixel which clearly reveals the significant pixels for digits 7 and 1.

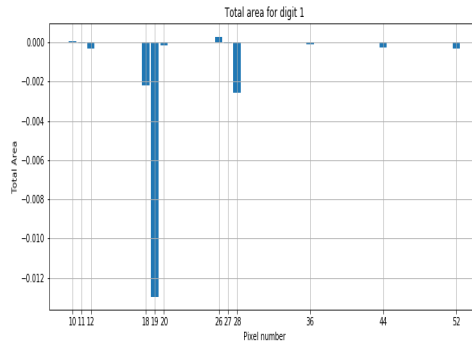


(a) ACE of digit 7

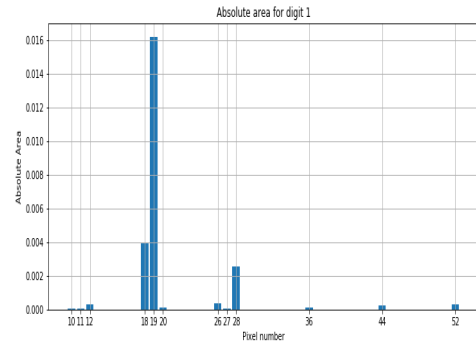


(b) ACE of digit 1

Figure FC4.6: Snapshot of ACE on handcrafted digits 1 and 7.



(a) Total Area Under ACE



(b) Area Under Absolute ACE

Figure FC4.7: Area under absolute ACE and total area.

Figure FC4.7 illustrates (a) total area and (b) area under absolute ACE of the list of significant pixels for digit 1. Pixel 18 and 19 are the most significant pixels in contribution to classifying the digits as 1 or 7. Pixel 28 is a key pixel for the digit 1 and can be seen to hold a decent causal attribution.

## CHAPTER 5

### CONCLUSION AND FUTURE WORK

Our work presents the way of interpreting the causal inference on a “black-box” like model e.g. neural network. This is the first attempt to interpret the causal inference results and explore the relationship between PCA and ACE, as far as our literature survey is concerned. Additionally, we have observed that area under absolute ACE is more significant than considering total area under ACE. The method has been used for the dataset with features varying from 4 (IRIS) to 64 (handcrafted numerals). This helps to better understand and adopt the causal analysis in the machine learning domain. We have also attempted to address the counterfactual like questions for a neural network model, which reveals the degree of generalization, the model can possess.

The future work will be in the direction to perform the feasibility study of the proposed framework on real world high dimensional dataset. Additionally, we will be interested in exploring causal attribution for transfer learning approach as well as to validate the causal attribution in a time series model like LSTM, where variables preceding in time can affect the current input (input can cause other inputs).

## Bibliography

- [1] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, “A survey of deep learning-based object detection,” *arxiv preprint arXiv:1907.09408*, 2019.
- [2] J. Chen, Z. Xu, and M. Tomizuka, “End-to-end autonomous driving perception with sequential latent representation learning,” *arxiv preprint arXiv:2003.12464*, 2020.
- [3] K. Raza and S. Ahmad, “Recent advancement in next generation sequencing techniques and its computational analysis,” *arxiv preprint arXiv:1606.05254*, 2017.
- [4] M. Anusuya and S. Katti, “Speech recognition by machine: A review,” *arxiv preprint arXiv:1001.2267*, 2009.
- [5] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavvaf, and E. A. Fox, “Natural language processing advancements by deep learning: A survey,” *arxiv preprint arXiv:2003.01200*, 2020.
- [6] V. Prabhu, A. Kannan, G. J. Tso, N. Katariya, M. Chablani, D. Sontag, and X. Amatriain, “Open set medical diagnosis,” *arxiv preprint arXiv:1910.02830*, 2019.
- [7] Y. Li, J. Hu, Y. Wang, J. Zhou, L. Zhang, and Z. Liu, “Deepscaffold: a comprehensive tool for scaffold-based de novo drug discovery using deep learning,” *arxiv preprint arXiv:1908.07209*, 2019.

- [8] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *arxiv preprint arXiv:1910.10045*, 2019.
- [9] H. S., “Shapley value,” in *Game Theory. The New Palgrave.*, J. Eatwell, M. Milgate, and P. Newman, Eds. Springer: Palgrave Macmillan, London, 1989, pp. 210–216.
- [10] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” *Proceedings of the 34th International Conference on Machine Learning*, 2017.
- [11] S. Bach, A. Binder, G. Montavon, F. Klauschen, M. K-R, and Samek, “Pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLoS ONE*, vol. 10(7): e0130140, 2015. [Online]. Available: <https://doi.org/10.1371/journal.pone.0130140>
- [12] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- [13] D. Alvarez-Melis and T. S. Jaakkola, “A causal framework for explaining the predictions of black-box sequence-to-sequence models,” *arxiv preprint arXiv:1602.04938*, 2017.
- [14] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. Balasubramanian, “Neural network attributions: A causal perspective,” *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- [15] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” *arxiv preprint arXiv:1703.01365*, 2017.



- [16] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, “Causal interpretability for machine learning – problems, methods and evaluation,” *arxiv preprint arXiv:2003.03934*, 2020.
- [17] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. C. Vishwanath, “Learning causal implicit generative models with adversarial training,” *arxiv preprint arXiv:1709.02023*, 2017.
- [18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115(3), 2015.
- [19] Y. LeCun and C. Cortes, “MNIST handwritten digit database,” 2010. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [20] J. Pearl, *Causality*. Cambridge university press, 2009.
- [21] E. Simpson, “The interpretation of interaction in contingency tables,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 13, no. 2, pp. 238–241, 1951. [Online]. Available: <http://www.jstor.org/stable/2984065>
- [22] J. Pearl, “The seven tools of causal inference, with reflections on machine learning,” *Commun. ACM*, vol. 62, no. 3, pp. 54–60, 2019. [Online]. Available: <https://doi.org/10.1145/3241036>
- [23] J. Pearl, “A probabilistic calculus of actions,” *arxiv preprint arXiv:1302.6835*, 02 2013.
- [24] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search 2nd ed.* MIT Press, Cambridge, MA, 2000.
- [25] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009, 2000a.

- [26] D. Dua and C. Graff, “UCI:Machine Learning Repository,” 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [28] J. Reback, W. McKinney, jbrockmendel, J. V. den Bossche, T. Augspurger, P. Cloud, gfyoung, Sinhrks, A. Klein, S. Hawkins, M. Roeschke, J. Tratner, C. She, W. Ayd, T. Petersen, MomIsBestFriend, M. Garcia, J. Schendel, A. Hayden, V. Jancauskas, P. Battiston, D. Saxton, S. Seabold, alimcmaster1, chris b1, h vetinari, S. Hoyer, K. Dong, W. Overmeire, and M. Winkel, “pandas-dev/pandas: Pandas 1.0.5,” Jun. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3898987>