The Rise of Causality in Robustness of Computer Vision

Haohan Wang, Maheep Chaudhary

Abstract

Causality has emerged as a revolutionary technique in the field of deep learning aiming to eradicate the spurious correlation using different methods like making unobserved confounders observed and removing them using back-door or front-door methods. Additional a lot of work has been done in making the unexplainable or partially explainable machine learning models to express their decision making power using more discriminator features using the third ladder of causation, i.e. counterfactual inference which harness the imaginative power of the machine learning model. In this paper, we aim to provide a comprehensive survey of the causal techniques used to improve the robustness and explainability of present-day computer vision systems first we provide a taxonomic different causal technique like Directed Acyclic Graph(DAGs) to make a Structural Causal Model(SCM). Secondly, we shed some light on some of the present works of causality. Third, we state the summarised version of works focusing on increasing the robustness of computer vision models using causality. Fourth, we give an overview of the works in computer vision to make it more explainable using counterfactual inference.

1 Introduction

Some tens of thousands of years ago a mere spark in the neurons of a human changed the course of the planet forever. From the spark came organized societies than towns and cities and eventually the science and technology-based civilization we enjoy today all because of a simple Spark that caused us to ask: **Why?**. This was the only questions that helped humans to conquer the planet and live a superior life as compared to other species on the planet Earth. In this era where AI is taking over every domain rapidly lacks the general capability of imagination and intervention.

Today every second model claims to surpass human performance but crumbles when it comes to real-world deployment or generalization, the rise of causality in machine learning has played a key role to make models more generalized as most of the machine learning models increase their performance or accuracy using the spurious correlation or patterns in the collected data. As mentioned in the paper **?** sometimes human annotation or the environment can act as a confounder for the model affecting both the label and the image. For example, **?** a model which aims to classify images of cows and camels both type of images contains cows and camels are labeled but due to environmental/domain bias most of the pictures of cows are taken in green pastures, and on other hand, most of the pictures of camels are taken in deserts. After training a simple Convolutional Neural Network model gives a very impressive performance but fails disastrously when the examples of cow images with the beach in it as a background are encountered by the model. It happens as the model only focuses on the green pastures to classify the image as Cow and desert to classify an image as Camel as it is easy to learn the features of green pasture and desert to learn, as the feature of cow and camels vary a lot as compared to the background.

To encounter this many works by applying the concepts of causality in them have been proposed such as Counterfactual augmentation **?** in which the author remove the main foreground feature in our example cow and regenerates the image by adding noise in that region of various kinds for regenerating the image using Generative Adversarial Networks(GANs), also the authors do the vice-versa technique to apply the Noise in the background of the image and leave the main portion of the image untouched. The process helps in therefore eradicating the confounder effect of the background affecting both the label y and the image X.

There have also been previous works that propose to eradicate unobserved confounders coming from human and annotators to make the model more robust. In the work **?** where the author focus to get unobserved using the behavior in the observed events, i.e. in the annotated data of the annotator and estimating the unobserved confounder. The author in this particular case handles it by observing the answers annotated by the annotator in the task of Visual Dialogue by arguing a_i(answer) is a sentence observed from the "mind" of user u during dataset collection. Then, $\sum P(A) * P(u|H)$, where H is history and A is answer and can be approximated as $\sum (P(A)P(a_i|H))$.

The other method ?, ?, ?, ? to make the model robust is to make the model focus more on the portion of the image where the real feature lies by identifying where the model is paying most attention tampering with the main features of the image and analyzing if the model changes its prediction and penalizes it if the model does it. To promote the generalization of the model in different domains many of the works ?, ?, ?, ?, ?, ?, ?, ? use causality and argues that the class contains objects containing the characterize special causal features, where the domains act as the intervention on objects that change non-causal features. These kind of works also include works ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ? ? focusing on creating or augmenting counterfactual image or feature so as to challenge the model to learn the samples that are not in the sample space given but can be encountered therefore promoting the concept of mixup? where it is focused to generate more and more samples in the boundary of desired feature space.

Additionally, many works have been proposed which have focused to make the ML models more explainable by highlighting the more discriminatory features where most of the explainable AI(XAI) models have focused on just highlighting the region with which the model things is necessary to make the classification using the Grad-CAM technique which extends the applicability of class activation maths procedure by including the information of gradient in it. But by applying the third ladder of causation that is Counterfactual Inference with a focus on the imaginative power of the models to perturb images minimally so as to alter its prediction which provides the most discriminatory features as done the works ?, and ?.

Based on the above works, most of the present work ? argues the question, i.e. What is the difference between Casuality and Adversarial Attacks? As they propose some of the points to answer the above question such as:

- Adverserial Examples(AEs) are used to fool the classifier whereas Counterfactual Examples(CEs) are used to generate constructive explanations.
- AEs show where an ML model fails whereas the Explanations sheds light on how ML algorithms can be improved to make them more robust against AEs
- CEs mainly low-dimensional and semantically meaningful features are used, AEs are mostly considered for high-dimensional image data with little semantic meaning of individual features.
- Adversarial must be necessarily misclassified while counterfactuals are agnostic in that respect
- Closeness to the original input is usually a benefit for adverserials to make them less perceptible whereas counterfactuals focus on closeness to the original input as it plays a significant role in the causal interpretation

The other works ?, ? and ? aim to apply the concepts of causality to discover causal features in less time and accurately like ? where the author use the concepts of causality and proposes a novel counterfactual framework for both Zero-Shot Learning (ZSL) and Open-Set Recognition (OSR), whose common challenge is generalizing to the unseen-classes by only training on the seen-classes. Some works use causality in ? zero-shot semantic segmentation where the author proposes a counterfactual method to avoid the confounder in the original model. In the spectrum of unsupervised methods, zero-shot learning always tries to get the visual knowledge of unseen classes by learning the mapping from word embedding to visual features. Additionally, the works aim to solve the problem of Adversarial Robustness and Few-Shot learning ? where the author argues that the pre-trained knowledge of the models is used which is indeed a confounder that limits the performance. There proposes to develop three effective Interventional Few-Shot Learning(IFSL) algorithmic implementations based on the backdoor[1] adjustment, the fine-tuning only exploits the D's knowledge on "what to transfer", but neglects "how to transfer".

This survey is structured as follows. In Section II, we first define some notations and give an introduction about Causality, with some diagrams explaining Structural Causal Models(SCMs). In the third section, we start by discussing the present works that are being done in the field of Causality and shift to different approaches to how concepts of causality is being applied in the domain of Computer Vision

[1] Given an ordered pair of variables (X, Y) in a DAG G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X and the causal effect of X on Y is given by

$$P(Y|do(X)) = \sum_z P(Y|X,Z)P(Z) \tag{1}$$

field to make it more robust and explainable. Finally, the conclusion of this paper and discussion of future works are presented in Section 5.

2 Overview

Will write about the SCM's and other common terminologies used while applying Causality in Machine Learning.

3 Causality

3.1 Notations

In this section we will define the notations to maintain consistency across the whole Survey. Capital letters denote the Variables and the small letters represents the specific values of the Variables. In particular, X is used to denote Dataset, where $X = \{x_1, x_2, ..., x_n\}$. The x_i represents the data points inside the dataset and $P(X)$ represents the probability distribution. Y is used to denote label/outcome of a data-point and $\overline{Y}$ or $\overline{X}$ is used for denoting the counterfactual label or counterfactual image, i.e. Y changes it's value to another class after a perturbation in original image, denoted by $\overline{X}$. In general we can learn $P(Y|X)$ from labeled data $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$.
There are several variables to consider when we are estimating different variables for an event, if we see an environment simply we find that there are only X and Y where $X \rightarrow Y$. But a complete environment does not consist only of the an Factor X and Outcome Y. Behind that there are a lot of variables that effect not only indirectly and directly effects X but also Y. The variables which effect both X and Y are known as *Confounders*(C), which mostly occurs due to the correlation between X and Y and given in detail in 3.2. There are other variables which effect the X directly are known as "*Parent*"(P_{ai}) and the ones effecting indirectly can be divided into 2 categories, i.e. the ones that are observed are known as *Endogenous Variables*"$(\mathbf{V})$, and the ones which are unobserved are known as *Exogenous Variables*$(\mathbf{U})$. Sometimes during intervention the feature is passed after passing noise in it or empty, therefore changing the overall function. All the variables intervened upon including functions are denoted by $*$ symbol in the power of the variable. For ex: Before Intervention a function may look like $f(v, q, p)$ but when v and q are intervened upon then the function is defined as $f^*(v^*, q^*, p)$. They can also be denoted using the *Do-Calculus*? making the function as $f(do(V = v_i, Q = q_i))$. The resulting effect is denoted by Y_{effect}. The images of a class say u, I_u are made up of the Style(S) or Context(c) and the main object, where in most of the cases S or c acts as C. The functions and equations used in the survey has Loss defined by L, Generator by G, Discriminator by D, Cross Entropy by CE, and Mean Squared Error by MSE. θ defines parameters of the model, z or Z defines the latent space, ϕ defines a function, KL-Divergence is define by the variable KL and E denotes the expectation, weight of the model is denoted by W or w, and distance by d.

3.2 Background

In this section we will give an introduction about Causality and about the recent works which might have a non-trivial impact in the Machine Learning Field in the coming years. Variables have always been a friend to predictions or either coming to a conclusion but Causality which is a science of cause and effects say otherwise by proposing that as every coin has two sides, the same is the case for variables therefore concluding that not all variables are friendly but are also enemies to our conclusion and prediction, preventing a system to identify the true data generating principles. The variables are known as Confounders, and occur due to the correlation between them and the outcome. The are merely correlated and do not effect Y, i.e. the outcome would be same if they would have not been there. Hence making the quote *Correlation is not Causation* more prevalent. In mathematical terms they can be defined as the variables which effect both the factor X and the outcome Y. They do not have any significance in predicting the outcome but to only hamper the performance by confusing the models using association as the method to identify true data generating principles. There have been various efforts to De-confound a model by eradicating the confounders but it is nearly impossible to identify all the confounders therefore putting the confounders into the category of Unobserved Confounders. Therefore dividing the Confounders into two categories of Observed Confounders and Unobserved Confounders.
Various techniques have been proposed to De-Confound the model, the most effective of them all is

Randomized Controlled Trial(RCT), which has it's origin in the pioneering work of the English scientist and statistician Sir Ronald Fisher who proposed to randomly assigns the participants into control group or treatment group so as the difference in both the groups come only because of the experiment and not due to any confounder. But, the technique has many down sides, i.e. many experiments cannot use the RCT technique as it may require performing experiments on the living beings making the whole process unethical, costly and time-consuming. Therefore instead of applying the RCT many techniques use the Intervention on Observational Data so as to identify confounders and also proving the basic hypothesis of identification of confounder. It could be easily done using the *Do-Calculus* which gives a powerful way to intervene on the assumed variables as confounders and analyse the outcome of the model. The *Do-Calculus* is used for Interventions and counterfactuals which are defined through a mathematical operator called $do(x)$, which simulates physical interventions by deleting certain functions from the model, replacing them with a constant $X = x$, while keeping the rest of the model unchanged, where many of the works used it to intervene on the original data to observe the confounder and eradicate it. One of the intervention techniques used to mitigate the effect of Confounding variable is to use *Instrumental Variable(IV)* which has it's roots from the book *The Tariffon Animal and Vegetable Oil*?. One of it's key advantages were that even in the presence of unmeasured confounding, may consistently estimate different kind of effects ?, ? such as the *Average Causal Effect(ACE)* or *Average Treatment Effect(ATE)* or Total Effect(TE), which can be defined as the mean of the difference between the outcome of treated variable, i.e. when $X = x^*$ and outcome of no treated variable, i.e. when $X = x$ for a sample i, when the mediator is M, i.e. $X \rightarrow M \rightarrow Y$.

$$ATE = Y_{x,M_x} - Y_{x^*,M_x^*} \tag{2}$$

The TE consist of the equation which shows it's relationship with *Total Indirect Effect(TIE)* and *Natural Direct Effect(NDE)*, where TIE reflects the effect of X on Y through the mediator M, i.e., $X \rightarrow M \rightarrow Y$, while the direct effect on $X \rightarrow Y$ is blocked by setting $X = x^*$. TIE is denoted as:

$$TIE = Y_{x^*,M_x} - Y_{x^*,M_{x^*}} \tag{3}$$

NDE denotes the effect of X on Y with the mediator M blocked, and expresses the increase in Y with X changing from x^* to x, while M is set to the value it would have obtained at $X = x^*$, i.e., the response of M to the treatment $X = x$ is disabled:

$$NDE = Y_{x,M_x^*} - Y_{x^*,M_x^*} \tag{4}$$

Therefore ATE can be concluded as:

$$ATE = NDE + TIE \tag{5}$$

The *Effects* are used to eliminate the bias in the modules of the Network or model in Computer Vision. To make all the interventions and counterfactual mechanisms transparent and easy to implement, *Structural Causal Models(SCM)* ?, ? were proposed as the whole model or environment is needed to make a conclusion or assumption about a model, making the process simple, transparent and effective to express the Causality. The SCMs use the Directed Acyclic Graph(DAG) which consists of vertices and edges, with each edge directed from one vertex to another, such that following those directions will never form a closed loop. A directed graph is a DAG if and only if it can be topologically ordered, by arranging the vertices as a linear ordering that is consistent with all edge directions.

Definition 1: Structural Causal Model(SCM) M consist of a set 4-tuples $< \mathbf{U}, \mathbf{V}, F, \mathbf{P(V)} >$ where *Exogenous Variables* , denoted by $\mathbf{U} = \{U_1, U_2, U_3, ..., U_n\}$ are unobserved variables, that are determined by factors outside the model. $\mathbf{V}$ denotes the *Endogenous Variables*, which are observed and are determined by other variables in the model and are denoted by $\mathbf{V} = \{V_1, V_2, V_3, ..., V_n\}$. F is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup P_{ai}$ to V_i, where $U_i \subseteq U, P_{ai} \subseteq V \; V_i$, and the entire set F forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. That is, for $i = 1, ..., n$, each $f_i \in F$ is such that

$$v_i \leftarrow f_i(p_{ai}, u_i) \tag{6}$$

$\mathbf{P(U)}$ is simply a probability function defined over the domain of $\mathbf{U}$.

The uni-directional edge provides the direction in which a Variable(V_1) is affecting the other Variable(V_2). The bi-directed edge dashed-edges denotes the unobserved variables, i.e. *Exogenous Variables* which affect the *Endogenous Variables.* In this case the *Endogenous Variables* seems to be affecting each other but in reality they are merely correlated and are affecting by a same variable.

Other interventional techniques have been proposed so as to mitigate the effect of Confounding variable, such as *BackDoor***?**, and *FrontDoor***?**, where *Backdoor* and *Frontdoor* methods could only be implemented if there is assumption of a confounder and we know the structure of *Structural Causal Model(SCM),* as for the IV it could be used with partial knowledge of the SCM. The *Backdoor* criteria is satisfied by a variable C relative to an ordered pair of variables (X, Y) in a DAG if no node in C is a descendant of X and C blocks every path between X and Y that contains an arrow into X. If C satisfies the backdoor criteria then the $X \rightarrow Y$ can be identified as:

$$P(Y|do(X)) = \sum_{c} P(Y|X, c)P(c) \tag{7}$$

But sometimes backdoor criteria does not get satisfied by the DAG therefore when there is a medium M in between the variable X and Y. The frontdoor criteria is satisfied if

- C intercepts all directed paths from X to Y
- There is no backdoor path from X to Z
- All backdoor paths from C to Y are blocked by X

The frontdoor works by first estimating the effect of X on M, and then estimating the effect of M on Y holding X constant. Both of these effects are unbiased because nothing confounds the effect of X on M and X blocks the only back-door path between M on Y, making the equation

$$P(Y|do(X)) = \sum_{M} P(M|X) \sum_{X^*} P(Y|X^*, M) * P(X^*) \tag{8}$$

There are various properties regarding SCM stated in **?** and are based on the Ladders of Causation. The SCM is Markovian if the variable in the *Exogenous Variables* are independent and it can be easily observed that SCM agrees on all lower layers but disagrees on all higher layers. A typical data-generating SCM encodes rich information at all three layers but even very small changes might have substantial effect, which is generally seen in the higher layers. The property is based on the Corollary, i.e.
Corollary: 1 *It is generally impossible to draw higher-layer inferences using only lower-layer information.*
The property holds generally but in today's scenario we can move from layer using the information of layer using the Causal Bayesian Network that uses *Do-Calculus* for intervention to get the insight from layer 2 using the layer 1 data.
Also the Factorization implied by the semi-markovian model does not act like chain rule, i.e.

$$P(e|a, b, c, d) = P(a)P(b|a)P(c|b, a)P(d|c, b, a) \tag{9}$$

But the factorization looks something like:

$$P(e|d, c, b, a) = P(a)P(b|a)P(c|a)P(e|b, c) \tag{10}$$

which implies that b and c are only affected by a also seen by a direct edge in SCM.

3.3 Causality and it's three ladder of Causation

From the very start machine learning had a deep-rooted relationship with statistics but classical statistics only summarise to data by taking into account the correlation between the variables due to which the models were not able to generalize outside their domain and therefore was less robust. As soon as these models will start to realize that the "*Rooster does not cause the Sun to rise*", i.e. "*Correlation is not Causation*" until then the problem of generalizability and robustness is not going anywhere. The Ladder of Causation **?**, **?**, **?** plays a key role in defining every Causation technique spanning three rungs which are based on common tasks of Observing, Doing and Imagining.

3.3.1 I^{st} Ladder Of Causation

The first Ladder of causation, i.e. Association in simple terms can be defined as seeing or observing, which entails detection of regularities in our environment. It is on a large scale shared by animals and was also shared by humans before the Cognitive Revolution. Today's Artificial Intelligent Systems works on seeing or association, i.e. by observing data without taking into consideration the concept of Causality between the features. The Ladder calls for prediction based on the passive observation of data. For example: If there are a large number of crimes and there is high sales of ice-cream then by logics of classical statistics and probability, it can be concluded if we let crime be denoted *"C"* and sales of the ice cream by *"S"* that $P(C|S)$ is very high and indirectly can be concluded that ice-cream sale is on of the main reasons for crime. But that cannot be the case, can it be? The 2^{nd} ladder of Causation may have an answer for it.

3.3.2 II^{nd} Ladder of Causation

The second Ladder of Causation focuses on "doing", i.e. on intervention. It ranks higher than the association, i.e. the first Ladder of Causation. This is because it involves not just observing/seeing but also changing the observed data so as to validate the basic hypothesis of the model regarding the True Data Generating Process. The most important route to implement the second Ladder of Causation is Do-Calculus []. It denotes intervention and counterfactuals and is defined by mathematical operators called do(x) which simulates physical interventions by deleting certain function from the model replacing them with a constant (X = x) while keeping the rest of the model unchanged. The post-intervention model is denoted by M_x. The resulting distribution from intervention do(X = x) is given by the equation

$$P_M(y|do(X = x)) = P_{M_x}(y) \tag{11}$$

In other words, the outcome, i.e. Y of the model M in the post-intervention distribution is defined as the probability when the outcome is $Y = y$ for each outcome when assigned by the model M_x. Extending the above example the $P(C|S)$ is quite high when the data is analyzed, the scenario is solved by taking "background features"(another word for confounder) yielding the criteria

$$P(C|S, K = k) > P(C|K = k) \tag{12}$$

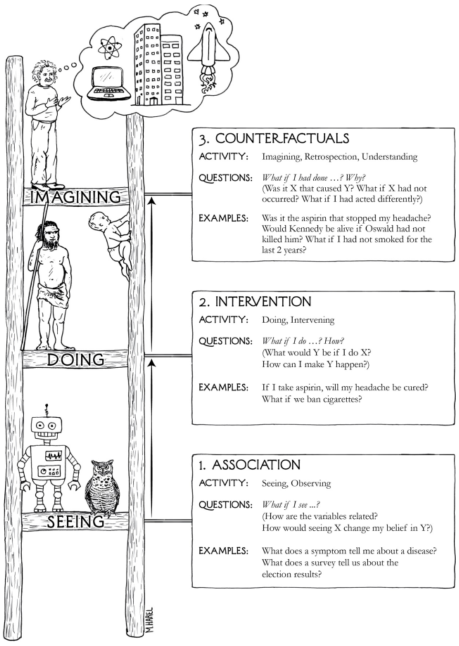
whereas, K is the background feature, in our case is temperature. When we take K, i.e. temperature in our particular case, and analyze applying to *Do-calculus* making *Temperature or K = 90°*. We will find out that there is no association between ice-cream Sales and crime.

3.3.3 III^{rd} Ladder Of Causation

The third Ladder of Causation which provided us humans with super-evolutionary speed through the power of imagination. The intervention or 2^{nd} Ladder of Causality helps in intervening on existing observations and therefore validating the model's hypothesis for *true data-generation principles*. But intervention cannot alone be sufficient to get to the true data-generating principles, as it may require going past in time or is sometimes morally unethical. Therefore the *third Ladder of Causation* proposes *Counterfactuals* which answers the queries in the closest alternate word based on the observed and interventional data. In Computer Vision Counterfactuals are used to generate synthetic examples which tries to make the model more adversarially robust and domain independent. It has also been used in Computer Vision in explaining the models decision by highlighting more discriminatory features in images as explained in the Section 5.3.

3.4 Challenges of Robustness in Computer Vision

Deep convolutional networks have achieved an incredible performance sometime surpassing the human performance when the training and testing data distributions match, but in recent times it has been observed when the model is deployed in real-world or when testing data is little different from training data. According to our definition robustness is a property of machine learning models to perform on testing data which is different from training data in terms of its non-causal features(features that should not contribute to models prediction) or upto a minimal point to it's causal features. An image consists of image features and confounders. If we make Structural Causal Model(SCM) then the Directed Acyclic Graph(DAG) will look as shown in Fig.2 at left side, where the D represents the *Confounder*, the X represents the *image features* and Y represents the *label*. The works mentioned in this survey proposes to use different methods to eridicate the confounder and therefore obtaining the Fig.2 at right side.



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Figure 1: Ladder Of Causation

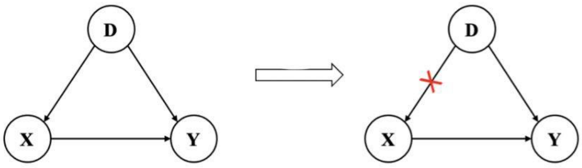


Figure 2: Left: Structural Causal Model(SCM) of the original scenario in which usually training is done. Right: Intervened Scenario, which is used by Causal techniques to eradicate the confounders

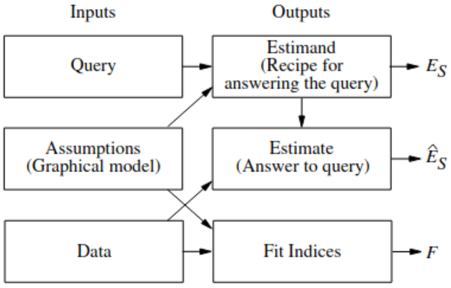

Inputs
Outputs
Query
Estimand (Recipe for answering the query)
E_S
Assumptions (Graphical model)
Estimate (Answer to query)
$\hat{E}_S$
Data
Fit Indices
F

Figure 3: How the SCM "inference engine" combines data with causal model (or assumptions) to produce answers to queries of interest.

3.5 Present Works

In this section, we will shed some light on some biggest leaps of innovation in causality and some of the present incredible works which are pushing the field forward. The Ladder of Causality had a very profounding effect on several fields and on the advancement of Causality as a concept, but the work **?** advances the scope of Causality in Machine Learning, where the author proposes 7 Tools of Causal Inference which are based upon the III Ladders of Causality, i.e. Association, Intervention and Counterfactual. The author proposes a hypothesis as given in Fig 3 which states that we all have some assumptions from which we answer our query and from our data we validate our assumptions, i.e. "Fit Indices". The author proposes the 7 tools as :-

- *Transparency and Testability* : Transparency indicates that the encoded form is easily usable and compact. The testability validates that the assumption encoded are compatible with the available data
- *Do-Calculus and the control of Confounding* : It is used for intervention, mainly used when we are trying to shift from 1st layer to 2nd.
- *The Algorithmization of Counterfactuals* : When we can analyse the counterfactual reasoning using the experimental or observational studies.
- *Mediation Analysis and the Assessment of Direct and Indirect Effects* : We find out the direct and indirect effects, such as what fraction of effect does X on Y mediated by variable Z.
- *Adaptability, External Validity and Sample Selection Bias* : Basically it deals with the robustness of the model and offers do-Calculus for overcoming the bias due to environmental changes.
- *Recovering from Missing Data* : The casual inference is made to find out the data-generating principles through probablisitic relationship and therefore promises to fill the missing data.
- *Causal Discovery* : The d-separation can enable us to detect the testable implications of the casual model therefore can prune the set of compatible models significantly to the point where causal queries can be estimated directly from that set.

Some works **?**, **?** have extended the application of Causality by using the *III* Ladder of Causation, i.e. *Counterfactuals* in Unit Selection. The until selection problem entails two sub-problems, *Evaluation* and *Search*. The author focuses on the latter, which focuses on solving the problem so as to find an objective function that, ensure a counterfactual behaviour when optimized over the set of observed characteristics C for the selected group. The author suggests to have two theorems, i.e. "Monotonicity"[2] and "Gain equality"[3] to optimize the problem statement, solved by A/B Testing ans statistical analysis **?**, **?**, **?**, **?** and **?** previously. Therefore proposing an equation as:

$$argmax\beta * P(complier|c) + \gamma * P(always - taker|c) + \theta * P(never - taker|c) + \rho * P(defier|c) \quad (13)$$

[2] Monotonicity expresses the assumption that a change from $X = false$ to $X = true$ cannot, under any circumstance make Y change from true to false.

[3] Gain-equality states that the benefit of selecting a *Complier* and a *Defier* is the same as the benefit of selecting an *Always-taker* and a *Never-taker* (i.e., $\beta + \rho = \gamma + \theta$)

where c is the subset taken and α, β, γ and ρ denotes the benefit of *Compiler, Always-taker, Never-taker* and *Defier.*

As we have seen until now, i.e. SCMs are quite important to get a view on causality and solve the problem through them, but it consist of some problem. It is where the work **?** shows it significance which tries to solve the problem by highlighting and solving the two kind of problems, i.e. "*Causal Effect Identification*" and *"Estimation"*. The *"Causal Estimation"* is the process of identifying the effect of different variables. *"Identification"* is obtained when we apply backdoor criterion or any other step to get a better insight. The power of identification has been shown by us in this survey to a large extent. The author proposes Neural Causal Models, which are SCM but are capable of amending Gradient Descent into it.
The work **?** focuses on estimating the Local Treatment Effect(LTE) which measures the effect among compilers under assumption of *Monotonicity.* The author argues that by obtaining the PDF may give very valuable information as compared to only estimating the Cumlative Distribution Function(CDF). The paper tries to focus on estimating LTE by deconfounding the effects of unobserved confounders using the *Binary Instrumental Variables*, which are the variables to counteract the affect of unobserved confounders. To be an instrumental variable these are the following conditions it should consist of:
Relevance: The instrument Z has a causal effect on the treatment X.
Exclusion restriction: The instrument Z affects the outcome Y only through the treatment X.
Exchangeability (or independence): The instrument Z is as good as randomly assigned (i.e., there is no confounding for the effect of Z on Y).
Monotonicity: For all units i, $X_i(z_1) \geq X_i(x_2)$ when $z_1 \geq z_2$(i.e., there are no units that always defy their assignment). The author develops two methods to approximate the density function, i.e. *kernel-smoothing* and *model-based approximations.* For both approaches the author derive double/de-biased machine learning estimators. *Kernel Smoothing method* smooths the density by convoluting with a smooth kernel function, whereas the *Model-based approximators* projects the density in the finite-dimensional density class based on a distributional distance measure. The work **?** introduces a loss function known as *Causal Loss* which aims to get the intervening effect of the data and shift the model from rung1 to rung2 of ladder of causation. Also the authors propose a *Causal sum Product Network(CaSPN).* The causal loss measures the prob of a variable when intervened on another variable. They extend CaSPN from *Interventional sum-product networks(iSPN)* be reintroducing the conditional variables, which are obtained when they intervene on the observational data. They argue that the CaSPN are causal losses and also are very expressive. The author suggests a way(taken from iSPN) conditional variables will be passed with adjacency matrix while weight training and target variables are applied to the leaf node. They train the *CaSPN, Neural Network(NN)* with causal loss, standard loss and standard loss + $\alpha * causal\_loss$ and produce the results. Also they train a Decision tree to argue that their technique also works on Non-Differential Networks, therefore they propose to substitute the Gini Index with the Causal Decision Score which measures the average probability of a spit resulting in correct classification.

4 Robustness Challenges in Computer Vision

5 Causality in Computer Vision

We have divided all the papers read in 3 categories as given below and therefore summary of the papers will be written there.

5.1 Counterfactual Synthesizing or Causal augmentation

A machine learning model when trained in an environment performs well in it, but is not expected to perform the same in the different environment. This section deals with those paper taking this approach to promote causality in Machine Learning models. Some of the works been proposed make use of the counterfactual mechanism to generate images so as to make the model "domain adaptable" and invariant to various environments. To solve this problem some of the works take the approach of synthetically generating more data by perturbing/augmenting the original image and sometimes generating data from scratch to make the model more robust. We have divided the section into two parts where the *Generating Novel Data* encounters those scenarios in which the additional data is created using counterfactuals, i.e. mostly using Generative Adversarial Networks(GANs) **?** to make the model robust. The second section *Counterfactual Augmentation* deals with the data generation technique in which the new/ additional data is created by perturbing the original data.

5.1.1 Generating Novel Data

$$\overline{X} = \phi(G(X, \overline{Y}), \gamma(X, \overline{X})) \tag{14}$$

Most of the works while applying Causality in Machine Learning make the model robust by generating more data out of the existing one given a counterfactual class $\overline{Y}$ by changing it minimally $\gamma(X, \overline{X})$ so as to change it's label or output as in 14, using various methods such as altering them ?. The focus on the minimal change had been a prime focus of works in this area, where ? minimize it by penalizing unrealistic looking images. Additionally also applying l2 loss between generated and original image while ? minimize it by minimizing counterfactual loss. ? takes this concept a step further by generating counterfactual images using *Independent Mechanism(IM)* based on object's shape, texture and background using cGAN.

The application of this method have been applied to various domains in Computer Vision including *Unsupervised learning* ?, *Few-Shot Learning with Open Set Recognition* ?, *Zero-shot semantic segmentation* ? and *Dataset Preparation* ?. In Unsupervised Learning the data in the source and target vary in terms of style and therefore to limits the Co-variate Shift $P(X|S = s) \neq P(X|S = t)$ and Conditional Shift assumptions, i.e. $P(Y|X, S = s) \neq P(Y|X, S = t)$ the ? generates images while adding *U(Content)* in both target and source domain using image generation to learn more non-discriminative semantics in source domain, which is however discriminative in target domain in source and target domain using the k pairs of end-to-end functions $(M_i, M_i)^k$ in unsupervised fashion, where $M(X_s{}^{4}) = (X_t{}^{5})$. To handle the challenge of identifying unseen images and therefore to generate images between the space of seen and unseen images the ? generate samples from the class attribute of an unseen-class, that lie in the sample domain between the ground truth seen and unseen. Interestingly ? generate images so as to de-confound the model which will be discussed in detail in next paragraph, in this particular case it generate the fake features so as to de-confound the real feature representation which indirectly affects the label using the word embedding with real features of the seen class and will generate fake images using word embedding. Rather than to make the model more complex by using different techniques to de-confound data, the ? generate dataset using the concepts of causality by asking some simple perturbations, i.e. specifically by asking the 4 major questions, i.e. *descriptive (e.g., 'what color'), explanatory ('what's responsible for'), predictive ('what will happen next'), and counterfactual ('what if')* and generating data from it.

The approach to de-confound the model using data-driven approach is not always optimal as the process of generating more data and training on it may increase the complexity exponentially. Therefore ? only generates data and optimize the model's performance using the contrastive learning technique to increase the mutual information between the joint embedding of Sample(Q) and image(V), i.e. $mm(Q, V) = a$, joint embedding of Q and factual image(V_+), i.e. $mm(Q, V_+) = p$ and decreases the mutual information between Q and Counterfactual Image(V_-), i.e. $mm(Q, V_-) = n$ and a by taking cosine similarity $s(a, n)$ and $s(a, p)$, where (V_-) denotes counterfactual, factual image is denoted by (V_+) and original samples(Q). Thereby concluding

$$L_c = E[-log(\frac{e^{s(a,p)}}{(e^{s(a,p)} + e^{s(a,n)})})] \tag{15}$$

$$Total\_Loss = \lambda_1 * L_c + \lambda_2 * L_{vqa} \tag{16}$$

The ? de-confounds the model with the help of Deep Neural Network(DNN), by arguing that the second last layer(just behind the softmax layer) has a very linear relationship with the labels and can be used to intervene and generate counterfactual example to make the model robust, while there are other models which try to generate counterfactual image so as to give the DNN the effect to increase it's distance from the prediction or effects of counterfactual image.

5.1.2 Counterfactual Augmentation

$$X^* = \phi(a * X + (1 - a) * \overline{X}, \overline{Y}) \tag{17}$$

In addition to generating counterfactual images from scratch, some works use the augmentation to generate it by perturbing manually in the same original data X using the image $\overline{X}$ of counterfactual class $\overline{Y}$ to generate X^* to make the model more robust as in 17. ? and ?, ? and ? focus

[4] Source Domain Sample

[5] Target Domain Sample

on increasing the data to train the model to de-bias the model by applies the technique which perturb original image's causal or non-causal features, identified using human annotation of important parts so as to annihilate the behaviour of background as C. While the above method may increase the number of samples irrespective of the number of biased images and unbiased images, **?** masks the main object and confounded object by measuring the probability distribution while dividing the dataset as Both, Just Main, Just Spurious, and Neither using the pattern the dataset is redistributed as: $P(Spurious|Main) = P(Spurious|notMain) = 0.5$ To minimize the potential for new SPs by setting the $P(Main|Artifact) = 0.5$), it moves images from Both, Neither to Just Main, Just Spurious if $p > 0.5$, i.e. $p = P(Main|Spurious)$ but if $p < 0.5$ then it moves images from Just Main, Just Spurious to Both, Neither. **?** applies the method in VQA task augmenting questions by converting them using the rule-based criteria, i.e. quesitons having *"number"* and *"other"* questions by converting them to "yes/no" questions, i.e. What color is the $<$ *Subj* $>$? $<$ *Color* $>$ is changed to Is the color of $<$ *Subj* $>$ is $<$ *Color* $>$?. While rule-based methods, i.e. to generate more data by shuffling causal and non-causal feature might improve the performance but sometimes hampers the performance generating unrealistic images that might not be helpful when creating model for real-world. **?** generates more data by applying the augmentation in *Vision-and-Language Navigation*(VLN) task by sampling batch of paths P, augments them and reconstruct instructions I using Speaker. With the pairs of (P, I), so as to maximize the navigation loss L_{NAV}, while the other module NAV, i.e. navigation model trains so as to minimize the L_{NAV} making the whole process more robust and increasing the performance.

While the method of generating additional data and then training neural network(NN), might be computationally expensive and time taking, the methods **?** and **?** generate data so as to help the model focus more on causal features. **?** use it to improve VQA robustness to let the model focus on Visual Features and also to make it question sensitive by replacing a part of the original image with some noise or simply masking. While **?** increases the attention map $do(A = \bar{A})$ by imagining non-existent attention maps $\bar{A}$ to replace the learned attention maps and keeping the feature maps X_{map} unchanged.

$$Y_{effect} = E[Y(A = A, X = X)] - E[Y(A = \bar{A}, X = X) \tag{18}$$

$$Total\_Loss = L_{CE}(Y_{effect}, Y) + L_{others} \tag{19}$$

where L_{others} represents the original objective such as standard classification loss.

While the above methods may require human annotation to mask out or perturb the causal features. **?** augments the images using multiple retinotopic centres which act as instrumental variable technique to achieve casual intervention to eliminate the minute confounders in the image proposing $maxP(Y = \bar{y}|X = x+\rho) - P(Y = \bar{y}|do(X = x+\rho))$, which subject to $P(Y = \bar{y}|do(X = x+\rho)) = P(Y = \bar{y}|do(X = x))$. But De-biasing by generating more data or using different functions may also be a bad thing in terms of performance and can hamper it as the eradicating bad bias may also sometimes eradicate good bias, **?** avoids it using the backdoor criteria to only remove bad bias. All the above methods apply the augmentation according to the author's assumption, i.e. *What will be best for the performance of the model* but it might not always be correct, therefore **?** optimizes the augmentation process by identifying the optimum augmentation process in result increasing the overall performance and reducing time complexity as proposed by dividing all samples from the training domains into a training and validation set and by training a classifier to predict the domain d from input x. In reiterative fashion it applies the first data augmentation in the list to the samples of the training set. Saving the domain accuracy on the validation set after training. This step is repeated all data augmentations in the list. The data augmentation is selected with the lowest domain accuracy averaged over five seeds.

5.2 Eliminating bias using Causality

The section discusses about the causal methods to make the Machine Learning model robust using Causal Intervention. We have divided this section into three parts, i.e. *Eliminating bias using backdoor method* which discusses about the works that have used Backdoor method to eliminate the bias in the model. The second part *Eliminating bias using Counterfactual Subtraction* which discusses about the works that have used counterfactual logic to create a scenario where only confounder leaves its effects and therefore subtracting it from the original output, therefore eradicating the confounder from the original model prediction. The third part *Others* consist of other methods used to make the model robust by eradicating the confounder.

5.2.1 Eliminating bias using backdoor method

$$P(Y|do(X)) = \sum_{C} P(Y|x, C)P(C) \tag{20}$$

In order to remove the variables or factors confounding the Machine Learning model so as to remove the bias is one of the prevalent field of research in Machine Learning, one of the common and effective methods has been to find the direct effect that one variable X, has on another one Y and isolate the effect with spurious correlations present in them, also regarded as "Backdoor method" using the *Do-Calculas* as given in 20. **?** uses to it deconfound the $< subject, object >$ in Video Visual Relation Detection (VidVRD), to fairly incorporate each possible predicate prototype into consideration by first learn the set of predicate prototype or relation references with the same predicate, comprising of 2 losses, i.e. : $L = L_{obj} + \lambda * L_{pred}$, where L_{obj} is the cross entropy loss function to calculate the loss of classifying video object trajectories and L_{pred} is binary cross entropy loss used for predicate prediction. But it is not always the case to deconfound the model while training as recently *Transfer Learning* has emerged as one of the popular techniques in Machine Learning many of the techniques in Deep Learning, but the models are biased and it is not possible to train the model from scratch to de-confound it, **?** approaches this problem de-confounding the pre-training model in *Few Shot Learning(FSL)* so as to eliminate the *pre-trained knowledge of the models(D)* using backdoor which act as confounder. The fine-tuning only exploits the D's knowledge on "what to transfer", but neglects "how to transfer". Though stronger pre-trained model improves the performance on average, it indeed degrades that of samples in *Query Set(Q)* dissimilar to *Support Set(S)*. By proposing 4 variables, i.e. *"D", "X", "B", "Y"* where X is the feature representation of the image, B is the low dimension representation of X and Y are the logits. D affects both the X and B, also X affects B, X and B affects the logit Y and deconfounds it by removing effect of D on X using backdoor.

But most of the confounders as given above cannot be easily separated from the features, therefore some works **?**, **?** and **?** use *"Disentanglement"* to separate of the deconfounding feature from the overall features to make it observable and *"Contrastive Learning"* mixed with *"Causal Intervention"* to deconfound it and also to increase the mutual information between the positive features and query, while decreasing mutual information between the negative features and query, as done by **?** in which two contrastive losses L_{qv} optimizes *QV-CL* module increasing the Mutual information of the positive frames of video and the query and L_{vv} optimizes *VV-CL* increase the mutual information between the start and end boundaries of the video. While the textual and visual features is mitigated using causal interventions $P(Y|do(X))$ with event as surrogate confounders to learn representations. **?** disentangles emotions and context from image, having emotion dicriminator(D_e) and context discriminator(D_c) where the loss comprises as :

$$L = CE(D_e(G_e(f_b(x))), y_e) + MSE(D_c(G_e(f_b(x))), 1/n) \tag{21}$$

where g_e is emotion generator and y_e is the emotion label and n is the number of counfounder and the same loss is for context replacing D_e, G_e and D_c by D_c, G_c and D_e, here n represents number of emotions and separated features fall within reason-able domains as IERN should be capable of reconstructing the base feature $f_b(x)$, i.e. $L = MSE(G_r(G_e(f_b(x)), G_c(f_b(x))), f_b(x))$ where builder is to combine each emotion feature with different context features so as to avoid the bias towards the observed context strata. **?** uses Deconfounded Cross-modal Matching(DCM) method to capture the true effect of query and video content on the prediction to remove the confounding effects of moment location by disentangling the moment representation from visual content, and applying causal intervention on the disentangled multi-modal input based on backdoor adjustment.

As using the backdoor by assuming a single confounder may not be a very good method as there can be many confounders which needs to be deconfounds, therefore some works use memory **?** bank while some **?**, **?** generate and then store the potential deconfounders to make them observable while applying the backdoor method, where **?** deconfounds the object-context entanglement in the class activation maps(CAM) where context acts as a confounder in *Weakly Supervised Object Loacalization*(WSOL) consisting of 2 *CAMs* and *Causal Context Pool* in between them storing the context of all images of every class debaising the *CAM*. The **?** applies to localize objects described in the sentence to visual regions in the video by deconfounding the *Style*(S) by generating counterfactual examples after taking the vectors from a memory bank by taking the top selected top regions for described object. The selected regions and frames are grouped together into frame-level content(H_c) and region-level content(R_c), and the rest of the regions are grouped as R_s and H_s and "Z" occurs due to some specific objects occurring frequently in the frames. The most similar one and replaces the original one, to generate examples to have them hard to distinguish from real ones contrastive learning is used. The equation looks like:

$$IE(p|do(R_s = R_{s\_generated})) < IE(p|do(R_c = R_{c\_generated})) \tag{22}$$

$$IE(p|do(H_s = H_{s\_generated})) < IE(p|do(H_c = H_{c\_generated})) \tag{23}$$

where the IE is Interventional Effect The second confounder C is eliminated by taking the textual embedding of the object as the substitute of every possible object z and apply backdoor adjustment. While applying backdoor method on many features or single feature has been possible while they are observed but unobserved variables may sometimes are very important to get deconfounded which may come during annotation of the data so as to get the debiased model. The similar concept is handled by **?** where it focuses to eliminate the spurious correlations, i.e. *dialog history* which could be easily eliminated by eliminating the direct effect of history using backdoor method and *Unobserved* that comes from the annotator in the task of Visual Dialogue and can be seen in the 'a_i'(answer), i.e. is observed from the *"mind"* of user u during dataset collection. Therefore $\sum(P(A) * P(u|H))$, where H is history and A is answer can be approximated as $\sum(P(A)P(a_i|H))$. They further use $p(a_i|QT)$, where QT is Question Type to approximate $P(a_i|H)$ because of two reasons: First, $P(a_i|H)$ essentially describes a prior knowledge about a_i without comprehending the whole Q, H, I triplet.

5.2.2 Eliminating bias using Counterfactual Subtraction

$$Y_{effect} = Y(a, b, c) - Y^*(do(A = a^*, B = b^*)) \tag{24}$$

The section discusses the works that use the counterfactual situation to get the only biased prediction and subtracting the effect of that counterfactual situation to get the unbiased model, also known as *Total Direct Effect*(TDE) in various application as mentioned in 24. The works uses the concept in various sub-fields of Machine Learning, **?** uses it *Visual Question Answering*(VQA), **?** in scene graph generation (SGG) task, **?** in Human Trajectory Prediction, and **?** in Compositional Action Recognition, while creating the counterfactual scenario. **?** mitigate the language bias using different layers for different sub-modules, i.e. for visual(v), question and visual+question which is denoted by Knowledge base K. The model is trained normally denoted by Z_q, k, v, another counterfactual scenario is created by training the model without passing the visual and Knowledge base Z_q, v^*, k^* and are subtracted from each other to eliminate the biasness of the language model. The v^* and k^* denotes that while inferencing visual features and knowledge in not passed. Similarly **?** removes the context bias, where the true label is influenced by Image(whole content of the image) and context(individual objects, where the model make a bias that the object is only to sit or stand for and make bias for it). Therefore using the

$$TDE = y_e - y_e(x_{bar}, z_e) \tag{25}$$

, where the first term denote the logits of the image when there is no intervention, the latter term signifies the logit when content(object pairs) are removed from the image, therefore giving the total effect of content and removing other effect of confounders. **?** cuts off the inference from environment to trajectory by constructing the counterfactual intervention on the trajectory itself and compares the factual and counterfactual trajectory clues to alleviate the effects of environment bias and highlight the trajectory clues. They Y_{causal} is defined as $Y_{causal} = Y_i - Y_i(do(X_i = x_i))$. A generative model is defined, which generates trajectory by a noise latent variable Z indicated by $Y*_i$. Finally the loss is defined as:

$$Y_{causal} = Y_i^* - Y_i^*(do(X_i = x_i)) \tag{26}$$

$$L_{causalGAN} = L2(Y_i, Y_{causal}) + log(D(Y_i)) + log(1 - D(Y_{causal})) \tag{27}$$

,where D is the discriminator. **?** inhibits the co-occurrence bias in the same action with distinct objects and also to deconfound the direct effect of appearance by taking prediction from only visual appearance, making the counterfactual scenario and subtracting it from the output of the normally trained model. The only losses which gets constituted in the model are: Appearance loss, Structural Loss and fusion Loss by using the cross-entropy. **?** annihilates the effect by directly subtracting the logits of counterfactual model and biased model which aims at predicting the near future based on past observation in first-person vision by deconfounding the visual effect and therefore get logits *"A"* from the pipeline without making any changes to the model and also getting the logits *"B"* when they provide a random value to visual feature denoting the question of counterfactual and getting unbiased model.

$$Unbiased\_logit = A - B \tag{28}$$

5.2.3 Others

As the extension of above methods used to implement the backdoor method for de-biasing is used to implement other debiasing methods, i.e. intervention using disentanglement **?**, **?** to separate the features needed for prediction and the confounder in it. The **?** separates the spuriously correlated Z from S using the latent variable V and therefore observes it using intervention $p(y|do(s^*)) = p(y|s^*)$. Similarly **?** disentangles factors of variation and proposes two new metrics to study causal disentanglement and one dataset named *CANDLE*, i.e. **Unconfoundess metric**: If a model is able to map each G_i(Generative Factor) to a unique Z_I ,the learned latent space Z is unconfounded and the second metric **Counterfactual Generativeness** which proposes a counterfactual instance of x with respect to generative factor G_i , $\bar{x}$, generated by intervening on latent space of x, i.e. Z_{I_x} corresponding to G_i and any change in the latent dimensions of Z that are x not responsible for generating G_i , i.e. Z_I, should have no influence on the generated counterfactual instance $\bar{x}$ with respect to generative factor G_i, which can be computed using the Average Causal Effect(ACE). Generative factors G is said to be disentangled only if they are influenced by their parents and not confounders we obtain a z for every g and acts as a proxy for it.
Some of the works **?** generate a dictionary of confounders, where the image objects are fed into two sibling branches: a Self Predictor to predict its own class, e.g., x_c , and a Context Predictor to predict its context labels, e.g., y_c, where it is used to caluculate the $E[G(z)]$ to get the top confounders from the dictionary. The complexity arose from the dictionary in some cases when the confounders act as the colliders is mitigated through the use of Neural Causation coefficient(NCC). But as defined in **?** the confounder is not always observable therefore it exploits the unlabelled background to model an observed substitute for the unobserved confounder, to remove the confounding effect in *Weakly-supervised Temporal Action Localization*(WTAL) by detecting the action segments with only video-level action labels in training and extracts the distribution of the input video features to identify the unobserved confounder Z, using the equation

$$P(x_1, x_2, \ldots.., x_n|Z = z) = \Pi P(x_t|Z = z) \tag{29}$$

The additional confounder c left could be identified with blessings of weak ignorability by replacing the expectation over C with a single z in $E[E[A|X = x, C = C_i]] = A$. **?** use constrastive learning to increase mutual information between points of one class and decreases mutual information of different class label as if there are 3 data-points (x_{di}, y), (x'_{dj}, y) and (x_{dk}, y') then the distance in causal features between x_i and x_j is smaller than distance between x_i and x_k or x_j and x_k. While **?** eliminates the di[U+FB00]erence between the natural distribution and the adversarial distribution, which can be assumed that the difference b/w $P_\theta(Y, s|X)$ and $P(Y, s|X)$ is the main reason of the adversarial in robustness, where s is the spurious correlation. Therefore defines the loss as:

$$minCE(h(X + E_{adv}; \theta), Y) + CE(h(X; \theta), Y) + CE[P(Y|G(X, s)), P(Y|G(X + E_{edv}, s))] \tag{30}$$

where E_{adv} adversarial perturbation, θ is the parameters of the model, and g represents the parameter optimized to minimize the CE, i.e. Cross Entropy loss.
? uses instrumental variable-based approach to learn the domain-invariant relationship between input features and labels as the input features of one domain are valid instrumental variables for other domains. Therefore proposes a model Domain-invariant Relationship with Instrumental VariablE (DRIVE) which learns the conditional distribution of input features of one domain given input features of another domain with *Maximum Mean Discrepancy*(MMD) that minimizes the distance b/w the feature representation of two different domains. **?** focuses on maintain the robustness of Federated Learning during the test phase and therefore propose to use *Test-Specific and Momentum Tracked Batch Normalization* (TsmoBN) which argues that the d_{si} for datasets of different domain[6], are used in training, X are the samples, R are the raw extracted features of X, F is the normalized feature representation of R and Y is the classifier. So to remove the confounding effects brought by d_u[7], causal intervention is done on normalized features (i.e., do(F)) to let the feature distribution similar to training distributions. This intervention is done by introducing the surrogate variable s_v, which is test-specific statistics of raw features R during testing by obtaining the test normalized features that have similar distributions as the training normalized features and momentum is introduced to integrate relations among different batches.

[6] Datasets coming from different users

[7] Dataset coming during test phase

5.3 Explainability

The section is divided into 2 parts, namely *Generating Counterfactual Image for Explanation* and *Others.* The former discuses about those works which generates the counterfactual image using minimal change in the original image so as to explain what are the most discriminatory regions in the original image from which the model is interpreting the class of the image, whereas the latter discusses the works which focus on explainability and transparency of the system but does not take the above defined path.

5.3.1 Generating Counterfactual Image for Explanation

$$I_u^* = F(I_u, Y_u) - mm(F(I_u, Y_u), F(I_{u'}, Y_{u'})) \tag{31}$$

Generating Counterfactual Image for explanation deals with the methods which are mostly build upon the method of attribute explanation but it only highlights the regions from which the model is estimating it's output but by generating counterfactual image we can easily take a look at the regions to which the model is paying attention but for a particular class highlighting discriminatory features, it could be easily done as given in the equation 31 by finding most important features to which the Deep Neural Network(DNN) is paying most attention using F function. Similarly the most important features from counterfactual class are also obtained using F. By taking the same features of both the class using the mm function are subtracted from original class features $F(I_u, Y_u)$ to obtain the most discriminatory features in the image I_u making it I_u^*. Many works have contributed to solve this problem by different methods but the most prevalent method used by **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?** and **?** is to perturb the original image I by taking the image of another class $\bar{I}$(generally user chosen) as reference so as to explain the most important part or regions which are contributing to model's output by modifying the original image to create counterfactual image I^*. **?** uses a predictor $h(I)$, and a confidence predictor $s(I)$ to get the activation tensors so as to get the segmented region of the image which is discriminative of the counter class. **?** only focuses to replaces minimal attributes unchanged, i.e. A = a1, a2, a3, a4, a5....an using

$$min(lambda * loss(I(\bar{A})) + ||I - I(\bar{A})) \tag{32}$$

where loss is cross-entropy for predicting image $I(\bar{A})$ to label $\bar{u}$. **?** uses cycle-GAN to translate real images x of class i to counterfactual images $\bar{x}$ and both images are fed into Discriminative Attribution model finding most discriminative features which first masks the region in $\bar{x}$ to get x. **?** uses Target Attribution Network(TAN) to generate the counterfactual image using the loss *Counterfatual Map loss* for minimal change, *Adverserial loss* to stabilize the adversarial training and *Cycle Consistency loss* for producing better multi-way counterfactual maps. The feature selection plays a very important role in generating the counterfactual image therefore the **?**, **?** and **?** optimizes the feature selection to generate counterfactual image as **?** uses

$$f(I^*) = (1 - a) * f(I) + a * P(f(\bar{I})) \tag{33}$$

where I^* represents the image made using the I and $\bar{I}$, $*$ represents the Hamdard product, $f(.)$ represents the spatial feature extractor and $P(f(.))$ represents a permutation matrix that rearranges the spatial cells of $f(\bar{I})$ to align with spatial cells of $f(I)$.It uses two greedy sequential relaxations – first, an exhaustive search approach keeping a and P binary and second, a continuous relaxation of a and P that replaces search with an optimization. **?** uses iterative masking and composition steps that optimize the selected features by perturbing them to produce the target class by changing minimally using the function:

$$argmin(\sum(\overline{f_k}(\overline{x}) - \frac{1}{N} * \sum(\overline{f_k}(X_i, u_t))) + \lambda(\overline{X} - X)) \tag{34}$$

? uses transformer for counterfactual explanation, consisting of 5 losses, i.e. having *Adversarial Loss* and Perturbation Loss, i.e. $L = E[G(x, \bar{y}) + G(x + G(x, \bar{y}), y)]$ for maximum similarity in generated and original image. *Domain classification loss* generates counterfactual image using $L = E[-log(D(\bar{y}|x + G(x, \bar{y})))]$ where $G(x, \bar{y})$ is the perturbation introduced by generator to convert image from x to $\bar{x}$. *Reconstruction Loss*, i.e. $L = E[x-(x+G(x, y')+G(x+G(x, y'), y))]$ and *Explanation loss* which guarantee that the generated fake image produced belongs to the distribution of H. $L = E[-(y'|x+G(x, y'))]$. Although optimizing feature selection improves generation of counterfactual image but it does not make the model transparent as much as it should be and does not provide with any substantial power to the user as **?** provides it by creating "isofactuals" [8] by changing W in orthogonally so as to discover hypotheses in the input

[8] Image interpolations with the same outcome but visually meaningful different features.

space themselves with faithful counterfactuals using an invertible deep neural network $z = \phi(x)$ with a linear classifier $y = W_t * \phi(x) + b$ by altering a feature representation of x along the direction of weight vector, i.e. $\bar{z} = z + \alpha * w$ where $\bar{x} = \phi(z + \alpha * w)$. The non-discriminatory properties $= e(x)$, $e(x) = K_t * z$, where K is orthogonal to W. $e(\bar{x}) = K_t * (z + \alpha * w) = K_t * z = e(x)$. To measure the difference between the counterfactual and image intermediate feature map h, i.e. $m = |delta_h| * cos(angle(delta_h, h))$ for every location of intermediate feature map. In addition to make the model more transparent, **?** also decrease the computational cost non-trivially by generating counterfactual image using the text and not by re-iterating **?** or applying exhaustive-search **?** in image classification. Using an image of class A to get counterfactual feature to obtain a counterfactual text. The counterfactual text contains the minimal B-type features from conversion from A which then used for text-to-image GAN model to generate a counterfactual image using the AttGAN and StackGAN, evaluating through $\log(P(B)/P(A))$ where $P(.)$ is the classifier probability of a class for obtaining the highest-scoring counterfactual image.

Explanability is integrated into the model by generating counterfactual image but the process of generation remains as of a black-box. Although the transparency can be improved by using high-level features but generating latent space increases transparency of the model significantlly and make the process less computationally expensive as **?**, **?**, **?** and **?** use latent space for explanation technique for visual models. While **?** focus on taking limitations of current Conditional Image Generators in account. Using the Distribution Causal Graph(DCG) where the causal graph is made but the nodes is represented the MLP, i.e. $\log P(X = (x_1, x_2, x_3 ... x_n)) = \sum(\log(P(X = x_i|\theta_i)))$ and the Counterfactual Image Generator which translate the latent factor into the image using the original image as anchor while generating it which is done using Fader Networks which adds a critic in the latent space and AttGAN adds the critic in the actual output. But sometimes the prediction of the model is not well calibrated to obtain low-dimensional, continuous latent space for the training data. In addition to that there is no distinction between the images with high-confidence and ones near the boundary. To address these problems **?** proposes TraCE which focus on irrelevant feature manipulation based on lack of well-calibrated model's prediction to obtain low-dimensional, continuous latent space for the training data and a predictive model and outputs a desired target attribute with prediction uncertainty based on the latent space. Uncertainty-based calibration objective is achieved using a counterfactual optimization strategy to reliably elucidate the intricate relationships between image signatures and the target attribute. The evaluation techniques like *Validity* focus on optimizing the ratio of the counterfactual with desired target attribute and total number of counterfactual, *Sparsity* focus on ratio of number of pixels altered to total no of pixels. The other 2 metrics are *Proximity* focus on average l2 distance of each counterfactual to the K-nearest training samples in the latent space and *Realism score* so as to have the generated image is close to the true data manifold. TraCE reveals attribute relationships by generating counterfactual image using the different attribute like age "A" and diagnosis predictor "D". $\delta_{Ax} = x - \bar{x_a}; `delta_{Dx} = x - x_d$ The $\bar{x_a}$ is the counterfactual image on the basis for age and same for $\bar{x_d}$. $\bar{x} = x + \delta_{Ax} + \delta_{Dx}$ and hence at-last we evaluate the sensitivity of a feature by $F_d(\bar{x}) - F_d(\bar{x_d})$, i.e. F_d is the classifier of diagnosis. Evaluating these kind of methods and also information loss may sometimes act as a big obstacle resulting in the loss of some non-trivial features but Invertible Neural Networks(INNs) **?** generative capacities overcome this for image classification by generating counterfactual examples efficiently, as it is fast and invertible[9]. INNs only change class-dependent features as the INNs have their latent spaces semantically organized. It generates the desired label image by $\overline{x} = f_{inv}(f(x) + \alpha * \delta_x)$ where, $f(x)$ converts the image into a latent space in which $\alpha * \delta_x$ counterfactual features are added to alter the label of image x and therefore f_{inv} inverses the latent space to generate the counterfactual image $\overline{x}$. $||z + alpha_0 * \delta_x - \mu_p|| = ||z + \alpha_0 * \delta_x - \mu_q||$ where the $z + alpha_0 * \delta_x$ is the line separating the two classes and μ_q and μ_q are the mean distance from line.

As the latent space creation provide more transparency to the model but generation of high-level features increase the transparency exponentially and can also be used with the help of latent space generation $Z = z1, z2, ..., zn$. The **?** uses latent space to reveal hidden using intervention mechanism, obtained using discrete variational autoencoders based on high level concepts that shift predicted class[10] which gives most discriminatory features after intervention using the loss $L = l(G(\phi(\bar{x}), x))$, where $\phi(\bar{x})$ is the counterfactual model, $\bar{Z}$(intervened Z) is made human interpretable using Visualization. The **?** and **?** make the model more transparent by inserting more transparency in the counterfactual image generation which focus on creating high-level features using fault-lines to that defines main features from which the humans differentiate between the two similar classes. The **?** uses 2 concepts: *PFT* and *NFT*, where *PFT* add those *xconcepts* to input image and *NFT* subtracts it to change model prediction. The *xcon-*

[9] it has full information preservation between input and output layers, where the other networks are surjective in nature

[10] human interpretable and not low level features like pixels

cepts are those semantic features that are main features extracted by CNN and from which fault-lines are made by selecting from them. While **?** use fault-lines using a dialogue between the user and the machine, where an image is given to machine and same blurred image is given to a person, machine decides what information should be provided based on the blurred image so the person can understand. The real image is revealed and if the person is able to predict the parts that it was missing before then the machine gets a positive reward and functions in a RL training technique way.

Counterfactual images may provide a Counterfactual map so as to come to a conclusion about the information in the original image. But it is not always the case as in images of biological or mechanical system may have different components or parts that indicate different functionalities and roles. The **?** and **?** approaches this problem by segmenting the original image to obtain the contribution of the segments in the model's output. **?** implements it by searching small set of segments that alters the classification in case of removal which are identified by segmenting the image with l segments and using is searched using the best-first search repetitively to avoid a complete search through all possible segment combinations and is selected based on the highest reduction in predicted class score until one or more same-sized explanations are found after an expansion loop. An additional local search can be performed by considering all possible subsets of the obtained explanation. If a subset leads to a class change after removal, the smallest set is taken as final explanation. When different subsets of equal size lead to a class change, the one with the highest reduction in predicted class score can be selected. Sometimes there is an absolute need to only take the explanations from the image on which the model is confident enough, as **?** proposes concept of "*Overdetermination*", which is given when the model is more than sure that the label is something in counterfactual image. Image(x) is segmented into different segments $S = s1, ..., sn$ and then applies the same segmentation to $\bar{x}$ creating $\bar{S} = \bar{s_1}, ..., \bar{s_n}$. The system determines the contributions that different subsets of S make to y by substituting with the corresponding segments of $\bar{S}$ by generating a counterfactual image through those segments in large number using GAN, where each perturbed image is then passed through the model M to identify the classification probability of all the classes and therefore the significance of every segment is obtained that is contributing in the layer.

As the problem of information loss was approached by **?**, proved that fine details play a very important role in defining the output of the model, focusing on this problem **?** uses disentangled latent space, obtained using an encoder with Diversity-Enforcing loss to uncover multiple valuable explanations about the model's prediction so as to learn a perturbation to produce counterfactual image using Counterfactual Loss minimally using Proximity Loss. The system is based on 3 metrics, i.e. Diversified, Sparse and Valid. Disentangled latent representation leads to more proximal and sparse explanations and also Fisher information matrix of its latent space to focus its search on the less influential factors of variation of the ML model as it defines the scores of the influential latent factors of Z. The counterfactual play a very important role as described by the above works but faces difficulty in when they are generated from High-Dimension data, limiting their application. **?** improves the existing methods as they are not able to properly generate counterfactual image in the high-dimensional data, unsemantic raw features and also in scenario when the effective counterfactual for certain label are not guaranteed, proposing Attribute-Informed-Perturbation(AIP) which convert raw features are embedded as low-dimension and data attributes are modeled as joint latent features using two losses: $Reconstruction\_loss$(used to guarantee the quality of the raw feature) + Discrimination loss,(ensure the correct the attribute embedding) i.e.

$$min(E[\sum_{a} *(-a * log(D(\overline{x})) - (1-a) * (1 - D(x)))]) + E[||x - \overline{x}||] \tag{35}$$

where $D(\bar{x})$ generates attributes for counterfactual image. To generate the counterfactual 2 losses are produced,one ensures that the perturbed image has the desired label and the second one ensures that the perturbation is minimal as possible, i.e. $L_{gen} = CE(F(G(z,a)), y) + \alpha * L(z, a, z_0, a_0)$ where CE is the Cross Entropy Loss and $L(z, a, z_0, a_0)$ is the l_2 norm between the attribute and the latent space. Counterfactual images provide a non-trivial information about the image from which the model is giving the output, but faces limitation when the original image is on the edge of the boundary, i.e. very similar to counterfactual image but of different label. In this case one of the potential solutions is to generate counterfactual image of same label by exaggerating the effect of features belonging to the label. Based on the concept the works **?**, **?**, and **?** produce counterfactual image generation to modify the image so as it gives it's original prediction but with more confidence. **?** analyze when the model has given wrong label by identifying changes leading to a correct classification, in the process making us aware about the decision boundary. It uses Epistemic uncertainty, i.e. the useful features using the Gaussian Mixture Model and therefore only the target class density is increased and the prediction is changed using a subtle change therefore the most salient pixel, identified using the gradient. The **?** use this concept as a metric to judge the model's robustness in robotics application so as helping it to reach it's objective by minimal and realistic modification, using the equation: $min\ d_g(x, \bar{x}) + d_u(F(\bar{x}), t_u)$, where d_g is the distance b/w the modified and original image as $d_g \propto Model\_Robustness$, d_u is the distance b/w the class space and F is the predictor that $\bar{x}$ belongs to t_u class. The loss defines as: $total\_loss = (1 - \alpha) * L_g(x, \bar{x}) + (\alpha) * L_u(x, t_u)$, where L_u is the loss x belongs to t_u class. where $\bar{f}$ gives the logits for class k, X_i, u_t represents the i^{th} training data that is classified into u_k class and the N is the number of modifications. But only using this concept to generate exaggerated image of the original label limits the transparency of the model therefore **?** along with exaggerating the semantic effect of the given outcome label, it also show a counterfactual image to explain the decision of the classifier while minimally changing it using:

$$L_{cgan} = \log \frac{P_{data}(x)}{q(x)} + log(\frac{P_{data}(u|x)}{q(u|x)}) \tag{36}$$

where $P_{data}(x)$ is the data distribution and learned distribution q(x), whereas $\frac{P_{data}(u|x)}{q(u|x)} = r(u|x)$ is the ratio of the generated image and the condition and for giving desired output the condition-aware loss is introduced, i.e. $L := r(u|x) + D_{KL}(f(\bar{x})||f(x) + \delta)$, where $f(\bar{x})$ is the output of classifier of the counterfactual image is varied only by δ amount. For self-consistency $G(x, 0) = x$, where G is the generator and here $\delta = 0$ also reverse perturbation of $\bar{x}$ should recover x. To mitigate small or uncommon details, usually ignored by GAN generated images are compared using semantic segmentation with object detection combined in identity loss. $L_{identity} = L_{rec}(x, G(x, 0)) + L_rec(x, G(G(x, \delta), -\delta))$

5.3.2 Others

The interpretation and transparency of the model is significantly increased if focused on the high-level/global features to make the changes in the model human interpretable. High-level or Global features are the features which are not pixel based but are based on the concepts from which a human eye views an image, for ex:- In a dog image the human mind will see it as consisting of the nose, eye, ears, mouth, tails etc. to comprehend the image which oftenly discusses by several works **?** which summarizes a discussion based on *eXplanation by Design (XbD)*: given a dataset of training decision records, how to develop a machine learning decision model together with its explanation and *Black Box eXplanation (BBX)*: given the decision records produced by an obscure black box decision model, how to reconstruct an explanation for it. It emphasises on expressive logic rule languages for inferring local explanations(by local they mean the explanation of data point), together with bottom-up generalization algorithms to aggregate an exhaustive collection of local explanations into a global one, optimizing jointly for simplicity and fidelity in mimicking the black box. The more informative causal explanation should be provided and the local level information availability can be quite beneficial for the progress of the field and concludes that a model should be *Model-agnostic*, *Logic-based*, *provides Local and Global explanability* and *High-Fidelity*: provides a reliable and accurate approximation of black-box behaviour. Based on the concept **?**, and **?** propose to increase the transparency of the model. **?** proposes

$$Effect = E(F(I)|do(C = 1)) - E(F(I)|do(C = 0)) \tag{37}$$

where F gives output on image I and C is the concept and proposes a *VAE* which can calculate the precise CaCE by generating counterfactual image by just changing a concept and hence computing the difference between the prediction score. Whereas **?** proposes causal graphical models to identify causal features by testing it on type-level[11] explanation rather token-level[12] explanations of particular events by learning a *Partial Ancestral Graph*(PAG) G, using the *FCI* algorithm **?** and the predicted outcome $\bar{Y}$ whereas Z are the high-level which are human interpretable and not like pixels using the equations $V = (Z, \bar{Y})$ and $\bar{Y} = g(z_1, ...z_s, \epsilon)$. On the basis of possible edge types, they find out which high level causes, possible causes or non-causes of the black-box output $\bar{Y}$. **?** focus on intervention on pixel-wise masking and adversarial perturbation elucidating the importance of portions in the image using the equation:

$$Effect(x_i \to x_j, Z) = P(x_j|do(\bar{x}_i), Z_{X_i}) - P(x_j|Z_{X_i}) \tag{38}$$

and casual effect can be defined as:

$$E_{X_i}[Effect(x_i \to x_j, Z)] = (P(X_i = x_i|Z) * (38)) \tag{39}$$

[11] links between kinds of events, or equivalently, variables

[12] links between particular events

Deep reconstruction loss is applied in 39 using the KL-divergence between the output probability distribution of original and auto-encoder inserted network. The work ? generates counterfactual explanation using a deep inversion approach when there is only availability of trained deep classifier and not actual training data and uses metrics for semantic preservation by applying methods such as Image Space Optimization(ISO) ? and Latent Space Optimization(LSO) ?. The author also focuses on manifold consistency for the counterfactual image using the Deep Image Prior model. -

$$argmin(\lambda_1 * \sum_{l}(layer_l(\bar{x}), layer_l(x)) + \lambda_2 * L_{mc}(\bar{x}; F) + \lambda_3 * L_{cf}(F(\bar{x}), \bar{y})) \tag{40}$$

where, $layer_l$:The differentiable layer l of the neural network, it is basically used for semantic preservation. L_{mc}: It penalizes $\bar{x}$ whcih do not lie near the manifold. L_mc can be *Deterministic Uncertainty Quantification* (DUQ). L_{fc}: It ensures that the prediction for the counterfactual matches the desired target. ? uses Proactive Pseudo-Intervention (PPI), a contrastive learning strategy and intervenes to find causal features consisting of a saliency mapping module that highlights causally relevant features which are obtained using the WBP using the loss $L = \sigma(l(x^*, \neg y; f(\theta)))$, an intervention module that synthesizes contrastive samples and prediction module which is encouraged to modify its predictions during causally-relevant synthetic interventions. To prevent saliency maps covers the whole image while optimization uses L1-norm of saliency map is used to encourage succinct (sparse) representations and loss $L = sigma(l(\bar{x}, y; f(\theta)))$ to prevent model to always give $\neg y$ as prediction, also giving images with random masks on them.

5.4 Others

The section discusses the works which varies in concepts and do not belong to any of the above section, where ? focuses on fine-tuning of pre-trained language models has a great success in many NLP fields but it is strikingly vulnerable to adversarial examples, as it suffers severely from catastrophic forgetting: failing to retain the generic and robust linguistic features that have already been captured by the pre-trained model. The proposed model maximizes the mutual information between the output of an objective model and that of the pre-trained model conditioned on the class label. It encourages an objective model to continuously retain useful information from the pre-trained one throughout the whole fine-tuning process.

$$I(S; Y, T) = I(S; Y) + I(S; T|Y) \tag{41}$$

two models overlap, i.e. the objective model and the pretrained model. S represents the features extracted the model by the objective model and T is the features extracted by the pretrained model. Sometimes models train themselves on OOD(out-of-distribution) but sacrifice their performance on the ID(in-distribution) data to sound more robust but a totally robust model has good accuracy on both the distributions. ? tackles this problem by teaching the model both about the OOD and ID data points and take into account the P_{OOD} and P_{ID}, i.e. the predictions of ID and OOD. Based on the above predictions the it can be easily introspected that which one of the distributions is the model exploiting more and based on it they produce the second branch of the model that scores for S_{ID} and S_{OOD} that are based on the equation $S_{ID} = \frac{1}{XE(P_{GT}, P_{ID})}$, where XE is the cross entropy loss. Further these scores are used to compute weights W_{ID} and W_{OOD}, i.e. $W_{OOD} = \frac{S_{OOD}}{S_{OOD}+S_{ID}}$ to train the model to blend the knowledge from both the OOD and ID data points. The model is then distilled using the knowledge distillation manner, i.e. $L = KL(P_T, P_S)$, where P_T is the prediction of the teacher model and the P_S is the prediction of the student model. ? focuses on eradicating the algorithm-based counterfactual generators which makes them inefficient for sample generation, because each new query necessitates solving one specific optimization problem at one time and consider the causal dependence among attributes to account for counterfactual feasibility by taking into account the counterfactual universe for rare queries, by employing novel umbrella sampling technique, i.e. by using the weighted-sum technique, calculating the weight of each biased distribution, and reconstructing the original distribution and conduct evaluations with the umbrella samples obtained. The counterfactual can be generated by giving a specific query q_0, instead of a label using the hypothetical distribution. ? focuses on finding the causal direction between pairs of random variables, given samples from their joint distribution by using causal direction classifier to effectively distinguish between features of objects and features of their contexts in collections of static images. Causal relations are established when objects exercise some of their causal dispositions, which are sometimes informally called the powers of objects. Based on it the author provides two hypothesis:

- Image datasets carry an observable statistical signal revealing the asymmetric relationship between object categories that results from their causal dispositions.

- There exists an observable statistical dependence between object features and anticausal features, basically anticausal features are those which is caused by the presence of an object in the scene. The statistical dependence between context features and causal features is nonexistent or much weaker.

Neural Causation Coefficient(NCC) is proposed to learn causation from a corpus of labeled data. As joint distributions that occur in the real world, the different causal interpretations may not be equally likely, i.e. the causal direction between typical variables of interest may leave a detectable signature in their joint distribution. Additionally they assume that whenever X causes Y, the cause, noise and mechanism are independent but we can identify the footprints of causality when we try to Y causes X as the noise and Y will not be independent. **?** discusses about the usefulness of Causality in Medical Imagining by taking sub-fields as *Data Scarcity*, *Data Mismatch* with some examples and therefore focuses on establishing the causal relationship between images and their annotations as it provides a clear and precise framework for expressing assumptions about the data. As in case of *Data Scarcity* Semi-supervised learning (SSL) is generally used and leverage readily available unlabelled data in the hope of producing a better predictive model than is possible using only the scarce annotated data but a model trained on image-derived annotations will attempt to replicate the (most often manual) annotation process, rather than to predict some pre-imaging ground truth therefore consisting of a confounding variable that comes from the annotator. In *Data mismatch* the mismatch between data distributions, typically between training and test sets or development and deployment environments, tends to hurt the generalizability of learned models and therefore it can be said that Dataset shift is any situation in which the training and test data distributions disagree due to exogenous factors. Moreover when analysing dataset shift, it is helpful to conceptualise an additional variable Z, representing the unobserved physical reality of the subject's anatomy. There are also another types of shifts like manifestation shift(under which the way anticausal prediction targets (e.g. disease status) physically manifest in the anatomy changes between domains), acquisition shift which result from the use of different scanners or imaging protocols and Data mismatch due to sample selection bias where the indicator variables in sample selection concern alterations in the data-gathering process rather than in the data-generating process. Some works **?** focus on providing the literature regarding the difference between the Counterfactual and Adverserial Example constituting of the points:

- AEs are used to fool the classifier whereas the CRs are used to generate constructive explantions.

- AEs show where an ML model fails whereas the Explanations sheds light on how ML algorithms can be improved to make them more robust against AEs

- CEs mainly low-dimensional and semantically meaningful features are used, AEs are mostly considered for high-dimensional image data with little semantic meaning of individual features.

- Adversarials must be necessarily misclassified while counterfactuals are agnostic in that respect

- Closeness to the original input is usually a benefit for adversarials to make them less perceptible whereas counterfactuals focus on closeness to the original input as it plays a significant role for the causal interpretation

6 Conclusion

References

E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. van den Hengel. Counterfactual vision and language learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10041–10051, 2020. doi: 10.1109/CVPR42600.2020.01006.

A. Akula, S. Wang, and S.-C. Zhu. Cocox: Generating conceptual and counterfactual explanations via fault-lines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2594–2601, Apr. 2020. doi: 10.1609/aaai.v34i03.5643. URL https://ojs.aaai.org/index.php/AAAI/article/view/5643.

A. R. Akula, K. Wang, C. Liu, S. Saba-Sadiya, H. Lu, S. Todorovic, J. Chai, and S.-C. Zhu. Cx-tom: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models, 2021.

J. P. Ang Li. Unit selection based on counterfactual logic. 2019.

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2(3):4, 2020.

S. Beery, G. van Horn, and P. Perona. Recognition in terra incognita, 2018.

M. Blumenthal, C. Christian, and J. Slemrod. Do normative appeals affect tax compliance? evidence from a controlled experiment in minnesota. *National Tax Journal*, 54(1):125–138, 2001.

D. C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1), Jul 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. URL http://dx.doi.org/10.1038/s41467-020-17478-w.

C.-H. Chang, G. A. Adam, and A. Goldenberg. Towards robust classification model by counterfactual and invariant data generation, 2021.

G. Chen, J. Li, J. Lu, and J. Zhou. Human trajectory prediction via counterfactual analysis, 2021a.

L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang. Counterfactual samples synthesizing for robust visual question answering, 2020.

Y. Chen, X. Yang, T.-J. Cham, and J. Cai. Towards unbiased visual emotion recognition via causal intervention, 2021b.

X. Dong, L. A. Tuan, M. Lin, S. Yan, and H. Zhang. How should pre-trained language models be fine-tuned towards adversarial robustness?, 2021.

N. Eckstein, A. S. Bates, G. S. X. E. Jefferis, and J. Funke. Discriminative attribution from counterfactuals, 2021.

T. Freiesleben. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, Oct 2021. ISSN 1572-8641. doi: 10.1007/s11023-021-09580-9. URL http://dx.doi.org/10.1007/s11023-021-09580-9.

T.-J. Fu, X. E. Wang, M. Peterson, S. Grafton, M. Eckstein, and W. Y. Wang. Counterfactual vision-and-language navigation via adversarial path sampling, 2020.

I. Gat, G. Lorberbom, I. Schwartz, and T. Hazan. Latent space explanation by intervention, 2021.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations, 2019.

Y. Goyal, A. Feder, U. Shalit, and B. Kim. Explaining classifiers with causal concept effect (cace), 2020.

F. Hvilshøj, A. Iosifidis, and I. Assent. Ecinn: Efficient counterfactuals from invertible neural networks, 2021.

B. Höltgen, L. Schut, J. M. Brauner, and Y. Gal. Deduce: Generating counterfactual explanations efficiently, 2021.

M. Ilse, J. M. Tomczak, and P. Forré. Selecting data augmentation for simulating interventions, 2020.

J. J. Thiagarajan, K. Thopalli, D. Rajan, and P. Turaga. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Scientific Reports*, 12, 01 2022. doi: 10.1038/s41598-021-04529-5.

M. Jiang, X. Zhang, M. Kamp, X. Li, and Q. Dou. Tsmobn: Generalization for unseen clients in federated learning via causal intervention, 2021.

H.-G. Jung, S.-H. Kang, H.-D. Kim, D.-O. Won, and S.-W. Lee. Counterfactual explanation based on gradual construction for deep networks, 2021a.

Y. Jung, J. Tian, and E. Bareinboim. Double machine learning density estimation for local treatment effects with instruments. 2021b.

R. A. Lewis and D. H. Reiley. Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on yahoo! *Quantitative Marketing and Economics*, 12(3):235–266, 2014.

A. Li. *Unit selection based on counterfactual logic*. PhD thesis, UCLA, 2021.

Y. Li, X. Yang, X. Shang, and T.-S. Chua. *Interventional Video Relation Detection*, page 4091–4099. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450386517. URL https://doi.org/10.1145/3474085.3475540.

Z. Liang, W. Jiang, H. Hu, and J. Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.265. URL https://aclanthology.org/2020.emnlp-main.265.

S. Liu, B. Kailkhura, D. Loveland, and Y. Han. Generative counterfactual introspection for explainable deep learning, 2019.

Y. Liu, J. Chen, Z. Chen, B. Deng, J. Huang, and H. Zhang. The blessings of unlabeled background in untrimmed videos, 2021.

D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering causal signals in images, 2017.

D. Mahajan, S. Tople, and A. Sharma. Domain generalization using causal matching, 2021.

A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks, 2015. URL https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu. Interventional video grounding with dual contrastive learning, 2021.

E. C. Neto. Causality-aware counterfactual confounding adjustment for feature representations learned by deep models, 2020.

Y. Niu and H. Zhang. Introspective distillation for robust question answering, 2021.

Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen. Counterfactual vqa: A cause-effect look at language bias, 2021.

K. Oh, J. S. Yoon, and H.-I. Suk. Born identity network: Multi-way counterfactual map generation to explain a classifier's decision, 2021.

J. Pan, Y. Goyal, and S. Lee. Question-conditioned counterfactual image generation for vqa, 2019.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995a. ISSN 00063444. URL http://www.jstor.org/stable/2337329.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995b.

J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

J. Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini. Meaningful explanations of black box ai decision systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):9780–9784, Jul. 2019. doi: 10.1609/aaai.v33i01.33019780. URL https://ojs.aaai.org/index.php/AAAI/article/view/5050.

G. Plumb, M. T. Ribeiro, and A. Talwalkar. Finding and fixing spurious patterns with explanations, 2021.

J. Qi, Y. Niu, J. Huang, and H. Zhang. Two causal principles for improving visual dialog, 2020.

W. Qin, H. Zhang, R. Hong, E.-P. Lim, and Q. Sun. Causal interventional training for image recognition. *IEEE Transactions on Multimedia*, pages 1–1, 2021. doi: 10.1109/TMM.2021.3136717.

Y. Rao, G. Chen, J. Lu, and J. Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification, 2021.

A. G. Reddy, B. G. L, and V. N. Balasubramanian. On causally disentangled representations, 2021.

P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The value of reputation on ebay: A controlled experiment. *Experimental economics*, 9(2):79–101, 2006.

J. M. Robins. Semantics of causal dag models and the identification of direct and indirect effects.

P. Rodriguez, M. Caccia, A. Lacoste, L. Zamparo, I. Laradji, L. Charlin, and D. Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations, 2021.

D. Rosenberg, I. Gat, A. Feder, and R. Reichart. Are vqa systems rad? measuring robustness to augmented data with focused interventions, 2021.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

N. Sani, D. Malinsky, and I. Shpitser. Explaining the behavior of black-box prediction algorithms with causal learning, 2021.

A. Sauer and A. Geiger. Counterfactual generative networks, 2021.

F. Shao, Y. Luo, L. Zhang, L. Ye, S. Tang, Y. Yang, and J. Xiao. Improving weakly-supervised object localization via causal intervention, 2021.

F. Shen, J. Liu, and P. Hu. Conterfactual generative zero-shot semantic segmentation, 2021.

S. Singla, B. Pollack, S. Wallace, and K. Batmanghelich. Explaining the black-box smoothly- a counterfactual approach, 2021.

L. Sixt, M. Schuessler, P. Weiß, and T. Landgraf. Interpretability through invertibility: A deep convolutional network with ideal counterfactuals and isosurfaces, 2021. URL https://openreview.net/forum?id=8YFhXYe1Ps.

S. C. Smith and S. Ramamoorthy. Counterfactual explanation and causal inference in service of robustness in robot control, 2020.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*, volume 81. 01 1993. ISBN 978-1-4612-7650-0. doi: 10.1007/978-1-4612-2748-9.

P. Sun, B. Wu, X. Li, W. Li, L. Duan, and C. Gan. *Counterfactual Debiasing Inference for Compositional Action Recognition*, page 3220–3228. Association for Computing Machinery, New York, NY, USA, 2021a. ISBN 9781450386517. URL https://doi.org/10.1145/3474085.3475472.

X. Sun, B. Wu, X. Zheng, C. Liu, W. Chen, T. Qin, and T. yan Liu. Latent causal invariant model, 2021b.

S. S. Sundar, S. Narayan, R. Obregon, and C. Uppal. Does web advertising work? memory for print vs. online media. *Journalism & Mass Communication Quarterly*, 75(4):822–835, 1998.

K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang. Unbiased scene graph generation from biased training, 2020.

K. Tang, M. Tao, and H. Zhang. Adversarial visual robustness by causal intervention, 2021.

J. J. Thiagarajan, V. Narayanaswamy, D. Rajan, J. Liang, A. Chaudhari, and A. Spanias. Designing counterfactual generators using deep model inversion, 2021.

T. Vermeire and D. Martens. Explainable image classification with evidence counterfactual, 2020.

D. Wang, Y. Yang, C. Tao, Z. Gan, L. Chen, F. Kong, R. Henao, and L. Carin. Proactive pseudo-intervention: Causally informed contrastive learning for interpretable vision models, 2021a.

P. Wang and N. Vasconcelos. Scout: Self-aware discriminant counterfactual explanations, 2020.

T. Wang, J. Huang, H. Zhang, and Q. Sun. Visual commonsense r-cnn, 2020.

W. Wang, J. Gao, and C. Xu. Weakly-supervised video object grounding via causal intervention, 2021b.

A. White, K. H. Ngan, J. Phelan, S. S. Afgeh, K. Ryan, C. C. Reyes-Aldasoro, and A. d'Avila Garcez. Contrastive counterfactual visual explanations with overdetermination, 2021.

M. Willig, M. Zečević, D. S. Dhami, and K. Kersting. The causal loss: Driving correlation to imply causation, 2021.

R. S. Winer. A framework for customer relationship management. *California management review*, 43

(4):89–105, 2001.

P. G. Wright. *Tariff on animal and vegetable oils.* Macmillan Company, New York, 1928.

K. Xia, K.-Z. Lee, Y. Bengio, and E. Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference, 2021.

C.-H. H. Yang, Y.-C. Liu, P.-Y. Chen, X. Ma, and Y.-C. J. Tsai. When causal intervention meets adversarial examples and image masking for deep neural networks. *2019 IEEE International Conference on Image Processing (ICIP)*, Sep 2019. doi: 10.1109/icip.2019.8803554. URL http://dx.doi.org/10.1109/ICIP.2019.8803554.

F. Yang, S. S. Alva, J. Chen, and X. Hu. Model-based counterfactual synthesizer for interpretation. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, Aug 2021a. doi: 10.1145/3447548.3467333. URL http://dx.doi.org/10.1145/3447548.3467333.

F. Yang, N. Liu, M. Du, and X. Hu. Generative counterfactuals for neural networks via attribute-informed perturbation, 2021b.

X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua. Deconfounded video moment retrieval with causal intervention, 2021c.

K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020.

H. Yin, P. Molchanov, Z. Li, J. M. Alvarez, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion, 2020.

J. Yuan, X. Ma, K. Kuang, R. Xiong, M. Gong, and L. Lin. Learning domain-invariant relationship with instrumental variable for domain generalization, 2021.

Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua. Interventional few-shot learning, 2020.

Z. Yue, Q. Sun, X.-S. Hua, and H. Zhang. Transporting causal mechanisms for unsupervised domain adaptation, 2021a.

Z. Yue, T. Wang, H. Zhang, Q. Sun, and X.-S. Hua. Counterfactual zero-shot and open-set visual recognition, 2021b.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.

T. Zhang, W. Min, J. Yang, T. Liu, S. Jiang, and Y. Rui. What if we could not see? counterfactual analysis for egocentric action anticipation. In Z.-H. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1316–1322. International Joint Conferences on Artificial Intelligence Organization, 8 2021a. Main Track.

Y. Zhang, M. Gong, T. Liu, G. Niu, X. Tian, B. Han, B. Schölkopf, and K. Zhang. Adversarial robustness through the lens of causality, 2021b.

W. Zhao, S. Oyama, and M. Kurihara. Generating natural counterfactual visual explanations. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 5204–5205. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Doctoral Consortium.

Y. Zhao. Fast real-time counterfactual explanations, 2020.

Álvaro Parafita and J. Vitrià. Explaining visual models by causal attribution, 2019.