UNIVERSITY OF CALIFORNIA SAN DIEGO

Contrastive, Causal, and Game Theoretic Explanations to Understand Multi-Modal Models

A Thesis submitted in partial satisfaction of the
requirements for the degree of Master of Science

in

Computer Science

by

Aditya Lahiri

Committee in charge:

Professor Babak Salimi, Chair
Professor Jingbo Shang, Co-chair
Professor Barna Saha

2023

The Thesis of Aditya Lahiri is approved and is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Babak Salimi for his support as the chair of my committee. I am also grateful to him for providing me with the opportunity to do research in the Trustworthy Data Management Lab under his guidance with a great set of collaborators. I would also like to thank my friends and professors from my undergraduate studies at BITS Pilani, Goa. Special thanks also to my Manager and colleagues at American Express, AI Labs, India who first introduced me to the world of model explainability and gave me immense guidance and support in my first employment. I would like to extend my thanks to my colleagues at the Trustworthy Data Management Lab who have always engaged with me in thoughtful discussion and pushed me to do better. Finally, none of this would be possible without the efforts and hard work of my parents, my elder brother and my partner. They have always believed in me and encouraged me to do things I never thought I could do. They have invested a lot of time and energy in me, and in turn I have been able to re-invest it into this work.

Chapter 2, in full, is a reprint of the material as it appears Alipour, K., Lahiri, A., Adeli, E., Salimi, B., and Pazzani, M. (2022). Explaining Image Classifiers Using Contrastive Counterfactuals in Generative Latent Spaces. arXiv preprint arXiv:2206.05257.

Chapter 3, in full, is a reprint of the material as it appears in Lahiri, A., Alipour, K., Adeli, E., and Salimi, B. (2022). Combining counterfactuals with shapley values to explain image models. arXiv preprint arXiv:2206.07087. Thedissertation/thesis author was the primary investigator and author of this paper.

Chapter 4, in part is currently being prepared for submission for publication of the material. Lahiri, Aditya; Galhotra, Sainyam; Shanmugam, Karthik;Salimi,Babak. The dissertation/thesis author was the primary investigator and author of this material.

To my mummy, papa & brother for their love and belief.

ABSTRACT OF THE THESIS

Contrastive, Causal, and Game Theoretic Explanations to Understand Multi-Modal Models

by

Aditya Lahiri

Master of Science in Computer Science

University of California San Diego, 2023

Professor Babak Salimi, Chair

With the surge in use of machine learning and deep learning models for critical decision making in fields like healthcare, social justice and finance, the need to understand, debug and audit these blackbox systems has become pertinent. It has given rise to an entire field called Trustworthy and Responsible AI. Explainable AI is a subfield which focusses on explaining the decision making process of these highly complex blackbox systems. Our work concentrates on explanations that are causal and contrastive in nature. Instead of capturing spurious correlations, we aim to capture the underlying causal interactions in the model. We also want our explanations to be contrastive since as humans, we primarily understand outcomes and decision by comparing them to other similar situations. We rarely look at things in isolation, and rely on comparison

and contrast to deepen our understanding. We incorporate notions from Game Theory to tackle the feature attribution problem of ascertaining how much difference an input feature makes on the outcome of an instance. Our methods address blackbox models that can be either trained on images or tabular data. They can help understand model decision making from the lens of multiple stakeholders- model builders, model auditors or model users. The comprehensive coverage of our proposed methods enable us to produce a toolkit for model explainability that is widely applicable and also generalizable.

Chapter 1

Introduction

Complex and blackbox machine learning models have proliferated almost every aspect of our daily lives. Whether it be getting recommended media or getting approved for a loan, these models have found a place in every aspect of discovery and decision making. This has given rise to the need of developing methods and tools to understand their underlying decision making processes. This is essential for multiple reasons - assessing fairness, model debugging, ensuring trust in the model, and increasing its acceptance, among others. A number of approaches have been developed to explain model decision making. One of the most popular class of methods are post-hoc explainability methods. These explain the model after it has been trained as opposed to during model training itself. This has the advantage that one does not need to retrain their models which is usually an expensive process and one can also understand and explain legacy models that could be running in production environments. A specific kind of post-hoc explainability method is a model agnostic post-hoc explanation technique. Being model agnostic implies that these methods can be applied to any kind of underlying machine learning or deep learning model. In fact, these methods treat the model like an API call, where, the input features X are fed into this API, and the model prediction y is obtained as an input. These do not require any knowledge of the internals of the model. This is specially useful in cases such as model auditing, where usually the auditor does not get access to the organization's internal model along with its learnt parameters. Instead, the model can just be made available to the auditor

who can validate the model for appropriate behaviour. Our theses focusses on these model agnostic post-hoc explanation techniques. We use a concept from Game Theory called Shapley values to understand feature impact on the model outcome. We modify the vanilla shapley value framework to allow it to be contrastive so as to facilitate comparison and more intuitive model understanding.

1.1 Overview and Contributions

In this thesis we present our work to explain and understand both computer vision and tabular data based machine learning and deep learning models. Our methods build on top of Game Theoretic notion of Shapley Values. We also incorporate probabilistic and causal notions of direct, indirect and total effects along with sufficiency and necessity scores to enable a more nuanced understanding of the model decision making. First, we present our work on explaining Computer Vision based models using counterfactuals and probabilistic causal scores of sufficiency and necessity. Next, we use contrastive shapley values to explain the difference in outcomes between a pair of images in terms of their high level interpretable features. Finally, we present a framework that incorporates the well studied causal notions of Direct, Indirect and Total effects into Shapley values, this enabling more granular understanding of the underlying blackbox model.

1.2 Open source tools

Our developed methods are made public and has attracted attention from the open-source community. The following code repositories have been created and made public to help others use our research-

Counterfactual Contrastive Shapley for Computer Vision Models

Causal Contrastive Shapley to Capture Direct, Indirect and Total Effects.

Chapter 2

Explaining Image Classifiers Using Contrastive Counterfactuals in Generative Latent Spaces

In this chapter, we present our proposed method to explain Computer Vision based black-box models with the aid of counterfactuals and probabilistic causal scores of sufficiency and necessity.

2.1 Motivation & Overview

Despite their high accuracies, modern complex image classifiers cannot be trusted for sensitive tasks due to their unknown decision-making process and potential biases. Counterfactual explanations are very effective in providing transparency for these black-box algorithms. Nevertheless, generating counterfactuals that can have a consistent impact on classifier outputs and yet expose interpretable feature changes is a very challenging task. We introduce a novel method to generate causal and yet interpretable counterfactual explanations for image classifiers using pretrained generative models without any re-training or conditioning. The generative models in this technique are not bound to be trained on the same data as the target classifier. We use this framework to obtain contrastive and causal sufficiency and necessity scores as global explanations for black-box classifiers. On the task of face attribute classification, we show how different attributes influence the classifier output by providing both causal and contrastive feature

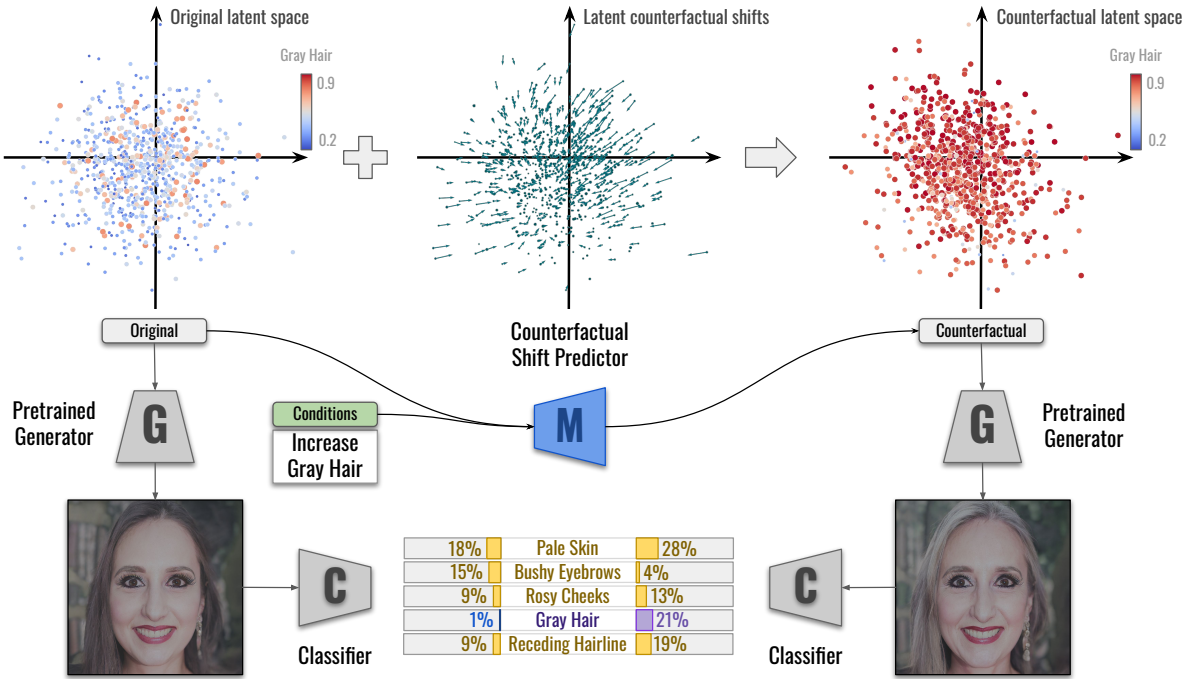

Figure 2.1. We introduce a method to learn counterfactual generation for a black-box classifier C. In this process, a shift-predictor M is trained in the latent space of a generative model G.

attributions, and the corresponding counterfactual images. Regardless of their accuracy, AI algorithms have yet to provide a level of interpretability to be accepted as trustable assets by their lay users in real-world applications. In recent years, eXplainable AI (XAI) has made an outstanding effort towards bringing transparency to AI with a focus on fairness and bias. Among different methods in this field, counterfactual explanations have received a lot of attention due to their scalable, intuitive, and logical approach [1, 2, 3, 4].

A counterfactual explanation for a black-box AI should provide transparency to the inner functionality of the algorithm through causal arguments and yet be interpretable to human users. Some methods specifically focus on the task of generating counterfactuals with a very high chance of changing AI's output such as adversarial examples [5, 6]. On the other hand, studying the impact of interpretable attributes in the input on AI output usually goes as far as minimal correlations with model output and fails to meet the next two steps in Pearl's ladder of causation [7]: intervention and counterfactuals.

While we acknowledge the fact that AI machines are not necessarily trained to follow causal reasoning based on interpretable features, we are interested in attributes that can provide the best of both worlds: to be as *interpretable* and as *causal* as possible. Learning such attributes can bridge the gap between causality and interpretability and lead to generating counterfactual explanations by changing meaningful attributes and still have a high causal influence on AI outcome. Learning the interactions between causality and interpretability in feature space can bring us closer to the true definition of explainability of AI. Such a framework can also provide us with means to measure whether an AI machine is following any human-understandable pattern to produce its output or not.

Comparing and contrasting target points by observing their differences along a fixed set of understandable dimensions has been one of the primary ways in which humans have always explained and understood concepts [8, 9]. These notions are also a natural way of explaining image classifiers. We can see and understand the difference between a pair of images that are distinctly different from each other in certain known attributes. This enables us to reason about how those differences may cause them to obtain separate outcomes in some downstream tasks performed on them. Therefore, we aim to generate explanations of the following general form: "For images with attributes having value for which the algorithm made decision outcome, the decision would have been foil-outcome with probability score had the attribute been counterfactual-value" [10]. In this work we seek to bring this contrastive framework for counterfactual explanations for image classification.

Notions of sufficiency and necessity build on this general form and allow us to reason about the necessary and sufficient conditions for a specific outcome. For an individual who received a positive outcome, necessity captures the importance of the existing value of an attribute in obtaining this outcome. On the other hand, for individuals who received a negative outcome, sufficiency reflects the ability of an attribute to flip the negative result into a positive one by modifying its existing value to some new value. Using a probabilistic interpretation of contrastive counterfactuals, we quantify the sufficiency and necessity of attributes to compute their causal

responsibility towards the classifier's output.

Obtaining these probabilities of sufficiency and necessity is a challenging task. They require generating counterfactual images that reflect the exact changes we desire in a set of human-understandable attributes. In our work, we formalize the definition of these scores in the context of images and traverse the latent space of generative models to obtain these counterfactual images which correspond to user-defined values of the set of these interpretable attributes. This enables us to compute these scores efficiently. An added benefit of our method is that instead of going through the expensive process of creating new datasets, it allows to sample a large number of inputs through pre-trained generative models and estimate contrastive counterfactuals over this sub-population for explaining any black-box image classification model (Fig 2.1).

In summary, the contributions of this work are as follows: (1) We introduce a method to produce contrastive counterfactuals for an image classifier; (2) Our method can use generative models pre-trained on any dataset and independent of the classifier training dataset; (3) We propose contextual, contrastive, and causal explanations in the form of sufficiency and necessity scores to explain the black-box model. (4) We use our method to provide global explanations for a black-box classifier trained on the CelebA dataset [11].

2.2 Related works

Previous work in this area generally approaches the problem from several different perspectives. Some of the prior work take a fundamental approach and revolve around exposing the causal roots and achieving causal models. A group of recent work take more rigorous approach by implementing contrastive counterfactuals for various applications. On the other hand, some of the methods in the vision area are centered around the use of generative models such as auto-encoders or GANs to produce interpretable counterfactuals. These generative approaches are divided to supervised and unsupervised techniques based on whether they involve annotations or classifiers in the process.

Causality

When estimating the causal effect of annotations, it's important to consider the confoundedness between these attributes for the purpose of any intervention. In that regard, most of the previous work attempts to learn a form of a structural causal model (SCM) [12, 13]. Parafita *et al.* [14] use causal counterfactuals to provide explanations by obtaining attributions for known latent factors. Dash *et al.* [15] use a conditional GAN to generate counterfactuals. Bahadori *et al.* [16] use a causal prior graph and existing annotations to explain the predictions. Khademi and Honavar [17] compare different methods of causal effect estimation such as CBPS and NPCBPS to interpret predictive models and explain their prediction based on inputs average causal effect (ACE). In a different approach, Zaeem and Komeili [18] introduce interpretable attributes as "concepts" and propose learning the presence of each "concept" in different layers of the classifier. Ghorbani *et al.* [19] also develop a systemic framework to automatically identify higher-level concepts which are both human-interpretable and important for the ML model. However, these concepts are neither necessarily causal nor require any prior interpretable attributes as input. While these techniques provide comprehensive explanations on the causal effect of attributes on model output, yet the causality is often quantified over a population and in correlation metrics. Such correlations are represented as global explanations and satisfy the first step of Pearl's ladder of causation [7], however, they cannot guarantee a causal impact on a case-by-case intervention (local explanation). In this work, we aim to go beyond correlations and provide a framework for a complete implementation of the causation ladder.

Contrastive counterfactuals

Contrastive counterfactuals have been the building blocks of ideas in philosophy and cognition that guide people's understanding and dictate how we explain things to one another [20, 9], and have been argued to be central to explainable AI [21]. To quantify these notions, probabilistic measures have been formalized and applied to a variety of fields including AI, epistemology, and legal reasoning [22, 23]. Recent work has also focused on using them in the

field of Explainable AI [10, 24, 25]. Our work is following a trend of ongoing research into generating counterfactual explanations for AI algorithms [26, 27, 28, 29, 2, 30, 4, 3, 31]. We are specifically interested in the implementation of this framework in the image classification problem. This topic is inherently challenging as it demands a probabilistic causal model based on the algorithm's output.

Explainable Autoencoders

Due to their strong abilities in representation learning, auto-encoders have received a lot of attention in the XAI community. Variational auto-encoders (VAEs) have shown promising results in learning causal [32] and interpretable [33] representations or interception of interpretable attributes [34]. Similarly, Castro *et al.* [35] propose a framework to measure how much the latent features in a VAE represent the morphometric attributes in the MNIST images. Some studies propose a feature importance estimation based on Granger causality [36, 37]. While auto-encoders are very strong in representation learning, they tend to lose detail in regenerating highly complex data. We focus our approach on the use of pre-trained generative models that can produce high-resolution counterfactuals with accurate shifts in interpretable features.

Unsupervised disentanglement in GANs

Within recent related work in this area, a number of them are dedicated to unsupervised disentanglement of latent space of GAN models to interpretable feature spaces. Some approaches use fundamental techniques such as projection [38], PCA [39], and orthogonal regularization [40], while others use self-supervised techniques to learn interpretable representations [41]. Another popular approach is the unsupervised discovery of linear [42, 43] or nonlinear [44] directions that correlate with interpretable features. Moreover, adversarial methods [45] alongside contrastive learning-based and intervention-based approaches [46, 47] have also been studied for the purpose of interpretable direction discovering for GANs. Despite their impressive results, these techniques may combine multiple distinguishable attributes in one detected direction. On the other hand, the detected interpretable directions are not always easy to label and hence cannot be used for

any label correction or model improvement. In this work, we propose learning latent directions that correspond to actual labels so the explanation results can be used in improving datasets and training procedures.

Supervised direction discovery

On the other hand, a large number of contributions in this area are focused on supervised discovery of interpretable directions in the latent space of generative models. The majority of these models are implemented for GANs such as StyleGAN [48] due to their resounding success and high quality. A classic solution in this area is finding the class boundary hyper-plane in the latent space of GAN [49]. There are approaches that attempt to find counterfactuals with the use of gradient descent in a GAN's latent space [50]. Some of the existing approaches train GANs to either apply residuals [51] or masked transformations [52] on images for the purpose of counterfactual generation. Moreover, some prior work experiment with incorporating the classifier [53] or contrastive language-image models [54] into GAN to accommodate attributes into the latent space. Another novel approach is the use of energy-based models (EBMs) for a controllable generation with GAN, however, this technique requires manual labeling of the latent samples [55]. Styleflow [56] introduces counterfactuals with conditional continuous normalizing flows in the latent space, however, their solution is specifically tailored for the extended latent space of StyleGAN. In our proposed approach, we seek a minimal training process by utilizing only pretrained GANs. Our methodology follows a simple and scalable implementation to be compatible with different generative models.

2.3 Method

In this section, we first discuss the kind of explanations that can be obtained using contrastive counterfactuals and provide some necessary background. Following that, we formalize the probability of sufficiency and necessity, which are at the core of our explanations. We also describe how we can compute those scores in the setting of image classifiers. We later explain

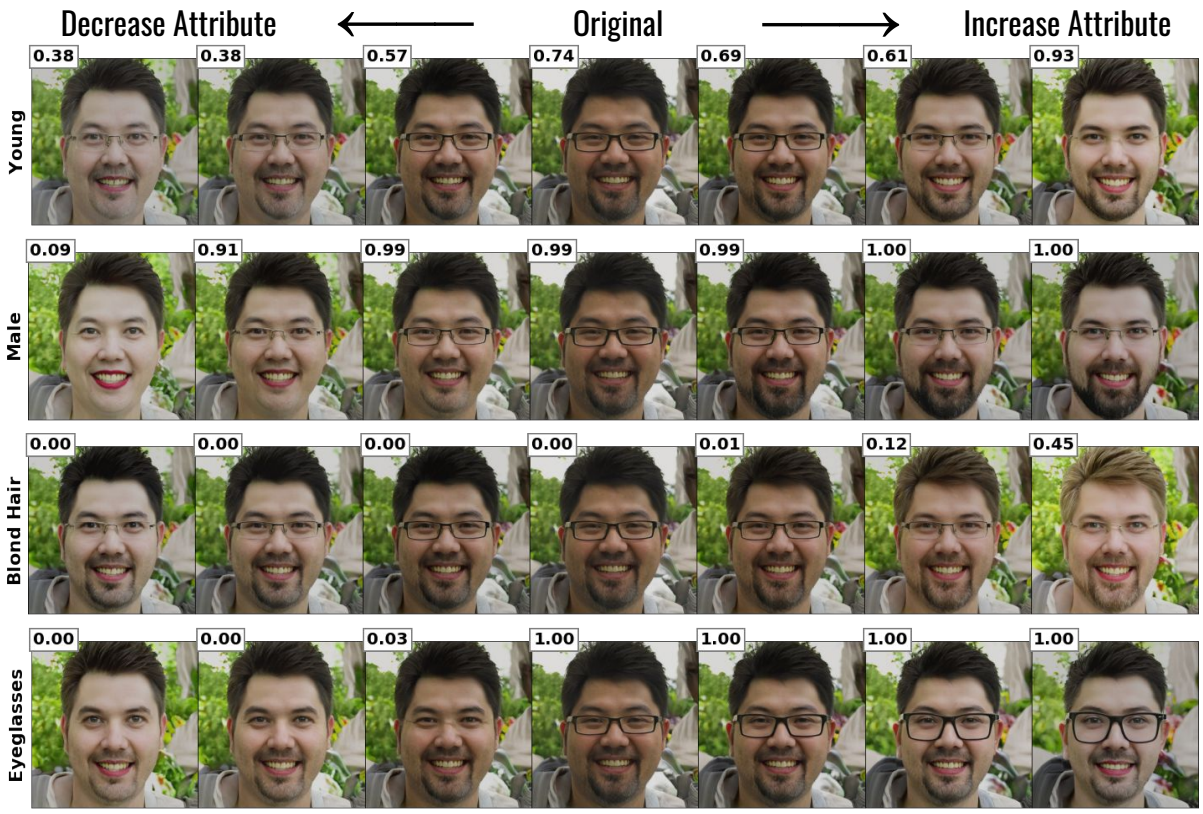

Figure 2.2. A shift predictor learns to predict optimum shifts in proximity of any input and manipulate the input for different attributes and different directions.

the algorithm pipeline, which consists of a pretrained generative model G to produce realistic images and introduce a shift predictor to achieve counterfactual latent vectors for G (Fig 2.2). These allow us to compute the necessity and sufficiency scores for explaining black-box image classifiers.

2.3.1 Contrastive Counterfactual Explanations

Our goal is to use probabilistic contrastive counterfactuals and develop a feature attribution method that generates explanation for an image classifier. These feature attributions quantify the causal contribution of a set of interpretable attributes on the outcome of the classifier. Specifically, for an image classifier which predicts the output Y for input images with an interpretable attribute A, our framework generates explanations of the following form: "For an input image

with an attribute a for which the classifier outcome is y, the classifier outcome would be $\hat{y}$ with probability s, had the input attribute been $\hat{a}$ instead of a". In the task of attractiveness classification for face images, these explanations pertain to images that had the positive outcome of being classified as attractive. For those cases, such explanations measure the probability with which increasing an interpretable attribute such as baldness could lead to a negative outcome instead. Therefore, they measure the extent to which the original value of the attribute is necessary for positive outcomes, hence called *probability of necessity*. Complementary to that we provide sufficiency scores for input images that receive the negative output. Such explanations compute the extent to which changing an attribute is sufficient to flip a negative outcome to a positive one, hence called *probability of sufficiency*.

In this work, we rely on Pearl's probabilistic causal models [57] to formalize and evaluate the notions of probability of sufficiency and necessity. Next, we briefly review probabilistic causal models and then build on that to mathematically define *Necessity* and *Sufficiency* scores.

Causal models and counterfactuals. A probabilistic causal model (PCM) consists of (1) a set of observable *endogenous* attributes $\mathscr{A}$, (2) a set of latent *background or exogenous* variables $\mathscr{U}$, (3) a set of structural equations $\mathscr{F}$ that capture the causal dependencies between the attributes by associating a function $F_A \in \mathscr{F}$ to each endogenous attribute $A \in \mathscr{A}$ that expresses the values of each endogenous attribute in terms of $\mathscr{U}$ and $\mathscr{A}$, and (4) a probability distribution $P(\mathbf{u})$ over the exogenous variables $\mathscr{U}$. Given a probabilistic causal model, an intervention on an endogenous attribute $A \subseteq \mathscr{A}$, denoted $A \leftarrow a$, is an operation that modifies the underlying causal model by replacing F_A, the structural equations associated with A, with a constant a. The *potential outcome* of an attribute Y after the intervention $A \leftarrow a$ in a context of exogenous variables u, denoted $Y_{A \leftarrow a}(u)$, is the solution to Y in the modified set of structural equations.

The distribution $P(\mathbf{u})$ induces a probability distribution over endogenous attributes and potential outcomes. Considering proper PCMs, one can express counterfactual queries of the form $P(Y_{A \leftarrow a} = \hat{y})$, or simply $P(\hat{y}_{A \leftarrow a})$; this reads as "What is the probability that we would

observe $Y = \hat{y}$ had A been a?" and is given by the following expression:

$$P(\hat{y}_{A \leftarrow a}) = \sum_{u} P(\hat{y}_{A \leftarrow a}(u))\, P(u). \tag{2.1}$$

Probability of Necessity and Sufficiency. We are given a binary image classifier with the output $Y = \{y, \hat{y}\}$, where $y = 1.0$ and $\hat{y} = 0.0$ denote the positive (favorable) and negative (unfavorable) outputs respectively, and a binary attribute $A = \{a, \hat{a}\}$ associated with the input images, where $a = 1.0$ and $\hat{a} = 0.0$ respectively denote the presence or absence of the attribute. The probability of SUFficiency and NECessity of A for Y measures as follows:

$$NEC = P(\hat{y}_{A \leftarrow \hat{a}} \mid a, y), \tag{2.2}$$

$$SUF = P(y_{A \leftarrow a} \mid \hat{a}, \hat{y}). \tag{2.3}$$

Given a sub-population of input images with attributes a, for which the classifier returns the positive output, the notion of probability of necessity (2.2) captures the probability that on changing the attribute A from its default value of a to the intervened value of $\hat{a}$, the classifier will return a negative outcome instead. In other words, it measures the extent of positive classifications that are attributable to the *original state* of the attribute $A = a$. The probability of sufficiency (2.3) is the dual of the probability of necessity. It applies to sub-population of input images with default attribute value $\hat{a}$, for which the classifier produced a negative output. It measures the effect of changing the attribute by intervention to a from its default state of $\hat{a}$. It computes the probability that this change could cause the classifier to return a positive outcome for these cases which were originally handed out a negative outcome. Hence, it measures the *capacity* of setting A to a to *flip* the negative outcome from the classifier.

We can choose to *change* the attribute from its default state by moving in its direction of increase or decrease. This would allow us to measure the sufficiency and necessity of changing the attributes in both directions and give a more in-depth understanding of how features are influencing the outcome of the black-box classifier. We denote necessity scores of increasing the

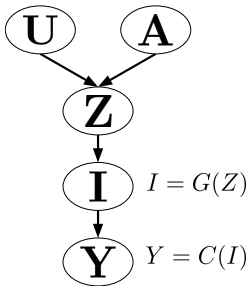

Figure 2.3. Causal model.

attribute with NEC^+ and the necessity scores of decreasing the attribute with NEC^-. We follow a similar notation for sufficiency scores.

Computing Necessity and Sufficiency.

We assume a causal model as shown in Fig. 2.3 that has the following components. *a*) Unobserved attributes ($\mathscr{U}$) - These are all the attributes that we do not have any observations for, and cannot account for. They constitute the exogenous variable in our causal model. *b*) A set of interpretable attributes ($\mathscr{A}$)- These are a set of attributes that we have known values for. They are our endogenous variables in the causal model. *c*) A latent space (**Z**) - The unobserved (exogenous) attributes, and observed set of interpretable attributes (endogenous) together directly affect the value of the latent space. *d*) A generative model (G) - It takes as input the above mentioned latent space **Z** and transforms it to an image I. *e*) The classifier to be explained (C) - It takes as input the image I and produces the target label, Y.

We generate counterfactuals, and also pass the produced images (both originals and counterfactuals) to the black-box classifier to obtain the classifier's output. We use this information to compute the sufficiency and necessity scores. We generate counterfactuals by following Pearl's three-step procedure [57].

- **Abduction** Given the prior distribution of the latent variable of a generative model G and the set of attributes $\mathscr{A}$, train a model M that estimates the updated probability of latent

variable conditioned on any subset of attributes:

$$P(\mathbf{z} \mid \mathscr{A}^{\bar{}} = \mathbf{a}),\ \mathscr{A}^{\bar{}} \subseteq \mathscr{A}. \tag{2.4}$$

- **Action** Take a set of interpretable attributes $\mathscr{A}^{\bar{}}$. Based on the causal model, perform an intervention by setting a subset of attributes $\mathscr{A}^{\bar{}} \subseteq \mathscr{A}$ to their determined values $\mathscr{A}^{\bar{}} \leftarrow \hat{\mathbf{a}}$.

- **Prediction** Given the model M, obtain the modified latent vector probability that corresponds to the new value of the attribute(s). Pass this modified latent vector into the generative model to obtain the corresponding image. Finally, run black-box classifier inference for this counterfactual image to obtain target output:

$$\begin{gathered} \hat{I} = G(\hat{\mathbf{z}}),\ \hat{\mathbf{z}} = M(\mathbf{z}, \mathscr{A}^{\bar{}} = \hat{\mathbf{a}}), \\ \hat{Y} = C(\hat{I}). \end{gathered} \tag{2.5}$$

2.3.2 Counterfactual Generation

Conducting the three steps of counterfactual generation is a non-trivial task due to the complication that arises when setting the attributes in the Abduction and Prediction steps. Specifically, generative models tend to have very complex latent spaces, where finding a path from $\mathbf{z}$ to $\hat{\mathbf{z}}$ for the purpose of attribute change is intractable. To reconcile for it, we propose training a MLP model M that serves as the shift predictor in our pipeline and can provide prediction for $\hat{\mathbf{z}} = M(\mathbf{z}, \mathscr{A}^{\bar{}} = \mathbf{a})$. With the use of M, we can now update the probability of latent variable $P(\mathbf{z})$ to the probability of counterfactual latent variable $P(\mathbf{z} \mid \mathscr{A}^{\bar{}} = \mathbf{a})$ in the prediction step and follow Pearl's procedure. In the following, we provide the details on the generative model and shift predictor algorithm.

Generative model.

Generative models are vastly popular in different fields of AI, and their recent advance-

Figure 2.4. Examples of counterfactual images that are generated during computation of the explanations scores.

ments in creating realistic images have made them a viable approach to producing a latent representation of an image dataset. In our experiments, we utilize StyleGAN2 [48] as a state-of-the-art generative model which can be used to generate high resolution and realistic images in different domains. StyleGAN feeds the latent variable into a mapping network that transforms it into an intermediate latent variable. Aside from its ability to produce styles, this transformation also provides the intermediary latent space as a more regulated domain to learn and traverse through interpretable attributes.

Algorithm 1: Training a shift predictor for binary attributes

Data: Classifier C. GAN model G. Counterfactual faithfulness ratio γ.
Result: Parameters Θ_M for the shift predictor model M.
$\Theta_M \leftarrow$ Random initialization
for *number of iterations* **do**
  Sample a batch of b noise variables and target outputs:
  $\{\mathbf{z}^{(1)},...,\mathbf{z}^{(b)}\} \leftarrow p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0},I)$
  $\{\hat{\mathbf{y}}_i^{(1)},...,\hat{\mathbf{y}}_i^{(b)}\} \leftarrow Bern.(p=0.5),\ \forall i = 1..m$
  Predict counterfactual latent codes:
  $\hat{\mathbf{z}}^{(j)} \leftarrow M(\mathbf{z}^{(j)},\hat{\mathbf{y}}^{(j)}),\ \forall j = 1..b$
  Generate images from the noise variables and predict attributes by the classifier:
  $I^{(j)} \leftarrow G(\hat{\mathbf{z}}^{(j)}),\ \forall j = 1..b$
  $\mathbf{y}^{(j)} \leftarrow C(I^{(j)}),\ \forall j = 1..b$
  Compute attribute conditioning and faithfulness loss:
  $\mathscr{L}_a \leftarrow \frac{1}{b}\sum_{j=1}^{b}\sum_{i=1}^{m} -\hat{\mathbf{y}}_i^{(j)} log(\mathbf{y}_i^{(j)})$
  $\mathscr{L}_f \leftarrow \frac{1}{b}\sum_{j=1}^{b} ||\hat{\mathbf{z}}^{(j)} - \mathbf{z}^{(j)}||$
  Update the shift predictor parameters: $M \leftarrow \nabla\Theta_M(\mathscr{L}_a + \gamma\mathscr{L}_f)$
end

Shift predictor. A shift predictor model is an MLP model that can take the latent variable of an image from a generative model G and generates the latent variable for its counterfactual based on the attributes produced by a classifier (Fig. 2.4). For a generative model $G : \mathscr{R}^d \rightarrow \mathscr{R}^n$ that has a latent space with dimension d and a classifier $C : \mathscr{R}^n \rightarrow \mathscr{R}^m$ that predicts m attributes, we define our shift predictor as $M(\mathbf{z},\hat{\mathbf{y}}) : \mathscr{R}^d \times \mathscr{R}^m \rightarrow \mathscr{R}^d$, where $\mathbf{z} \in \mathscr{R}^d$ is the latent variable for the input image and $\hat{\mathbf{y}}$ denotes the attributes for the intended counterfactuals. In the training process, shift predictor learns the directions in the latent space of G that correspond to changes in the attributes predicted by the classifier. Without the need for any manual labeling, the training procedure only requires the latent variables of images from G to input the shift predictor and supervise it with the labels generated by the classifier (see Alg. 1).

During the training, shift predictor learns to produce a counterfactual latent variable that satisfies any combination of attributes defined by $\hat{\mathbf{y}}$. In other words, if the classifier predicts a set of attributes $\mathscr{A} = \{A_1,A_2,...,A_m\}$, shift predictor can provide a counterfactual latent variable

compatible with any selected subset of attributes $\bar{\mathscr{A}}$:

$$\hat{\mathbf{z}} = M(\mathbf{z}, \{A_i = \hat{a}_i \mid A_i \in \bar{\mathscr{A}}\}). \tag{2.6}$$

Under the assumption of proper training, the shift predictor is an approximation of latent variable distribution conditioned by the subset of attributes $\bar{\mathscr{A}}$:

$$\hat{\mathbf{z}} \sim P(\mathbf{z} | \{A_i = \hat{a}_i \mid A_i \in \bar{\mathscr{A}}\}),\ \bar{\mathscr{A}} \subseteq \mathscr{A}. \tag{2.7}$$

The loss function in the training process pursues two objectives: 1) minimizing the error in prediction of attributes $\bar{\mathscr{A}}$ for the counterfactual image, 2) assuring a level of faithfulness and similarity between to the original input image and its newly generated counterfactual. The attribute loss $\mathscr{L}_a$ is defined as a cross entropy between the conditioned attributes and the attributes predicted by the classifier. In the training process, the conditioned attributes $\bar{\mathscr{A}}$ are distinguished from unset attributes so the loss will be only calculated for them. On the other hand, the faithfulness loss $\mathscr{L}_f$ is calculated as the normal distance between the original latent variables and their counterfactuals. The overall loss in the training process is defined as a combination of these two losses with a faithfulness factor γ which defines a balance between attribute accuracy of counterfactuals and their faithfulness to the original input:

$$\mathscr{L} = \mathscr{L}_a + \gamma \mathscr{L}_f = \sum_{A_i \in \bar{\mathscr{A}}} -\hat{\mathbf{y}}_i log(\mathbf{y}_i) + \gamma ||\hat{\mathbf{z}} - \mathbf{z}|| \tag{2.8}$$

2.4 Experiments and Results

We run our experiments to explain black-box classifiers that are trained on the task of classifying face images. We have annotations for the set of interpretable attributes A that we will choose to use to explain the model's behavior. We train a multi-task classifier built on top of a pretrained VGG [58] backbone to predict the set of interpretable attributes A for new unseen

images. The CelebA dataset is used as the training set and provides 39 binary attributes including attractiveness which we use as the target output Y for any black-box classifier of choice. As the set of explanatory interpretable attributes(A), we choose six other labels: blonde hair, heavy makeup, baldness, mustache, youngness, and maleness. We make a simplifying assumption to use an underlying causal model in which the explanatory attributes are independent of each other. We model attractiveness as the positive class $(y = 1.0)$ and unattractiveness as the negative class $(\hat{y} = 0.0)$ for the black-box classifier to predict. In the set of interpretable attributes(A), an attribute has its default value as (a) when it is not explicitly set. Otherwise, during intervention, it is set to value $(\hat{a})$ which can be $+1$ if we want to move in the direction of its increase, and -1 if we want to move in the direction of its decrease. We intervene and set $\hat{a}$ to 0 if we do not wish to modify the attribute. Our initial dataset consists of 200 images randomly produced by the generative model. We pass the images through the multi-task classifier to obtain the attribute values and through the black-box classifier to obtain the target label. We seek to explain the behavior of the target output Y using the attributes A on this dataset by performing the following experiments:

- **Linear baseline**, we first consider an interpretable linear approximation of target label behavior w.r.t. the attributes as the underlying black-box model. We use this approximation as ground truth and assess the validity of necessity and sufficiency scores in capturing this ground truth linear behavior.

- **Black-box explanations**, we consider a complex black-box classifier built on a pretrained VGG backbone and generate sufficiency and necessity scores to explain it. In conjunction with generated counterfactual images, we use these scores to analyze how increasing and decreasing the attributes affects the classification into attractive and not-attractive labels by the classifier.

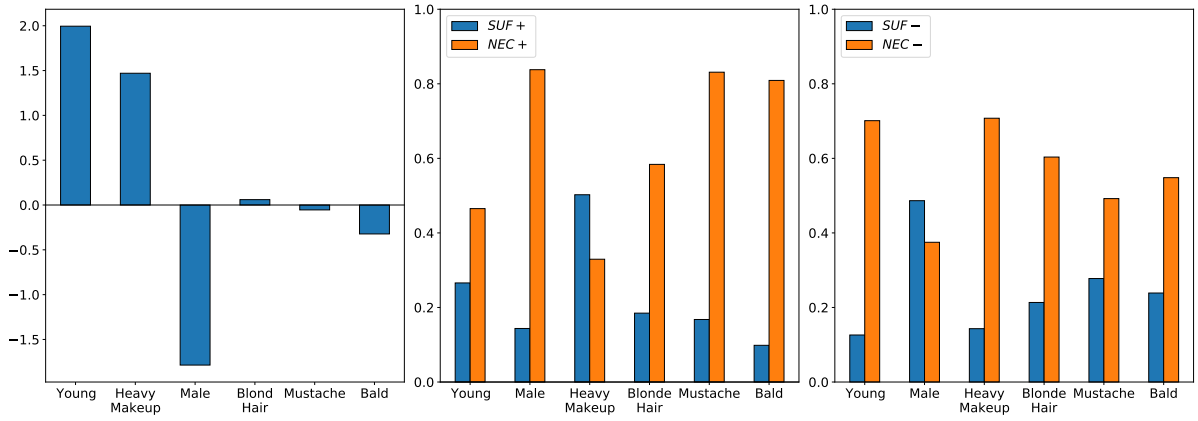

Figure 2.5. Left: The coefficients of the known black-box logistic regressor. **Center**: The sufficiency and necessity scores explaining the logistic regressor behavior when attributes are increased and in **Right** when decreased.

2.4.1 Linear baseline

Our pipeline of generating counterfactual images and explanation scores is agnostic to the type of black-box model being explained. This implies that it is independent of the type of machine learning or deep learning model used. However, one way to test the quality of our explanations is by generating explanations for a case where we actually have access to the decision making rationale of the underlying black-box model. To this end, we choose a logistic regression classifier as the model whose decision we seek to explain. This classifier takes real values corresponding to the feature attributes obtained from the multi-task classifier, and predicts the target label of attractiveness using only these values. The coefficients of the logistic regressor corresponding to the different features gives us an indication of how the model is making its decisions. We compare this to sufficiency and necessity scores generated by our method which seeks to explain this logistic regression model.

We observe from Fig. 2.5 that indeed, features that have negative attributions such as those for male, mustache and bald, in the logistic regressor model carry a large value of necessity when we move in directions of their increase. This high necessity score indicates that leaving them unset as compared to increasing them is most important to allow attractive individuals keep

their attractiveness score high. Similarly, for people classified as unattractive, the features that had high positive attributions in the logistic regressor such as Young and Heavy Makeup, carry the highest values of sufficiency when we increase their values. This is indicative of the fact that in order to flip the outcome from not-attractive to attractive, increasing heavy makeup and youngness are the two most important factors. We can interpret the SUF/NEC^{-} scores in a similar way. This kind of contrastive and counterfactual analysis is not possible through simple coefficients obtained from the logistic regressor. This makes it important to use these notions of sufficiency and necessity over standard co-efficient based attributions that have been observed to have multiple shortcomings due to their overly simplistic nature. These include issues like their dependence on feature pre-processing methods, as well as instability due to different feature selection[59].

2.4.2 Black-box Explanations

We use our pipeline to generate global explanations for the black-box attractiveness classifier. Here, the black-box classifier is built on a pretrained VGG backbone. Our explanations are two-fold. First, we provide sufficiency and necessity scores on a population level for the 6 different attributes. In addition to this, we also provide users with the counterfactual images that were generated by our shift predictor during the computation of the scores. The ability to have both feature attributions, as well as the images that led to the computation of those attributions allows the user to understand model behavior at a deeper level.

Fig. 2.4 contains a set of representative images. We can see how the original image changes in the direction of decrease and increase of the explanatory attributes. Fig. 2.6 shows the overall sufficiency and necessity scores of attributes in both positive and negative directions of increase. This gives us a detailed analysis of how the features are affecting the classifier output. For instance, a high SUF^{+} value of the attributes Heavy Makeup, Blond Hair, and Young implies that moving in the direction of increase of these attributes is most sufficient to flip an outcome of unattractiveness to attractive. Similarly, a high SUF^{-} value of the attributes Male, and Bald

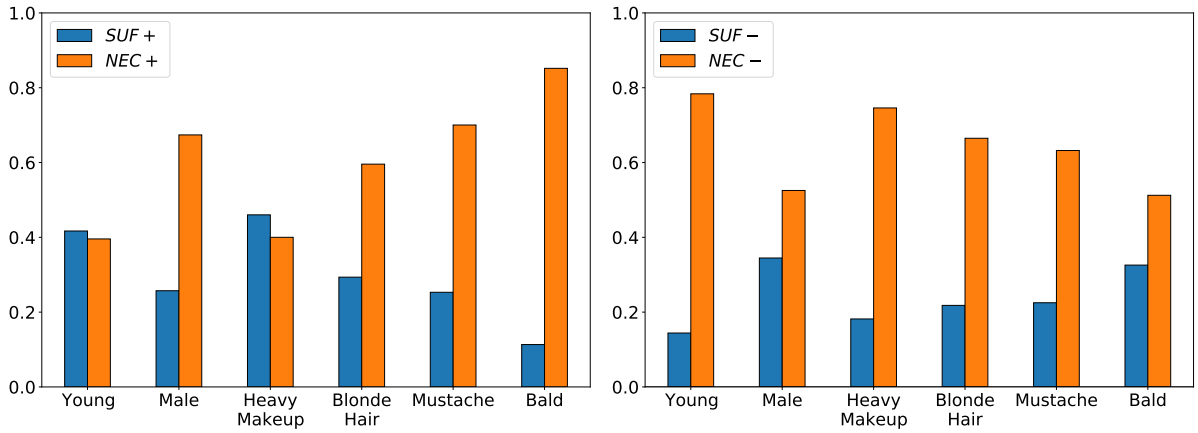

Figure 2.6. Sufficiency and necessity scores as global explanations.

reflects that when moving in the direction of decrease of these attributes, we are most likely to be able to flip our outcome from not attractive to attractive. The necessity scores inform us about the attributes that are most important to be left "unset" in their default state as compared to increasing or decreasing them, in order for a person classified as attractive to maintain that classification. A high NEC^{+} score of Baldness, Moustache, and Male is indicative of the fact that one should avoid increasing these attributes if they wish to remain classified as attractive by the classifier. Similarly, the high NEC^{-} scores of Young, Heavy Makeup, and Blond Hair are indicative of the fact that one should avoid decreasing these attributes if they wish to remain classified as attractive by the classifier. With the generated counterfactuals images as evidence, the necessity and sufficiency scores provide a holistic understanding of the black-box classifier to the end-user.

2.5 Summary

In this work, we proposed an end-to-end pipeline that generated counterfactuals from a pretrained generative model and used that to help compute probabilistic causal counterfactual scores. These scores, along with the generated images, served as explanations for any underlying black-box image classifier. Our work also highlighted the need and advantages of these contrastive explanations over simple feature attributions. However, one of the drawbacks of our

current method is that it does not effectively disentangle the effects between attributes. We would want to improve on that aspect by learning a structural causal model that can model the effects which attributes have on one another as well. This would also allow us to extend our analysis to compute the direct and indirect effects [60] of attributes on the target label. Furthermore, we would like to apply this pipeline to detect and mitigate bias in image classification systems.

Chapter 2, in full, is a reprint of the material as it appears Alipour, K., Lahiri, A., Adeli, E., Salimi, B., and Pazzani, M. (2022). Explaining Image Classifiers Using Contrastive Counterfactuals in Generative Latent Spaces. arXiv preprint arXiv:2206.05257.

Chapter 3

Combining Counterfactuals With Shapley Values To Explain Image Models

In this chapter, we present our proposed method that combines the game theoretic notion of shapley values with counterfactual image generation to help understand model decision making. This contrastive approach can be used by stakeholders such as model auditors to investigate blackbox model behaviour.

3.1 Motivation & Overview

Understanding the decision-making rationale behind complicated black-box models has emerged as one of the most critical tasks in the greater overarching goal of making AI systems transparent and trustworthy. Amidst several proposed methods, the well-studied concept of Shapley values [61] from Game Theory literature has emerged as a principled framework to obtain feature attributions as explanations. By virtue of their strong axiomatic guarantees, they lend themselves favorably to the task of distributing model outputs fairly among the different input features. However, one of the major issues in using Shapley values for model explanation is the computation time [62]. It grows exponentially with the number of features involved. This problem is particularly exacerbated in end-to-end models that operate directly on images since, usually, the features in images are defined on a pixel level. This results in thousands of features. In addition, the individual pixel values are not interpretable to humans, and therefore

any attribution attached to individual pixels does not help with high-level model understanding. It has also been shown that humans understand and explain things by comparison and contrast [20]. On their own, feature attribution numbers obtained by Shapley values do not allow for this contrastive notion and lack a reference or visual aid to compare to. This has been an essential barrier for generating explanations based on Shapley values [63].

To close the gap, we propose an approach that incorporates high-level interpretable features and employs generative models to produce counterfactual images corresponding to specific changes in the interpretable features. These counterfactuals are then used to compute Shapley values which explain the difference in prediction scores between the original and counterfactual image. This process enables us to understand model behavior using the contrastive explanations (w.r.t. Shaply values) provided for arbitrary input images.

3.2 Related Work

Shapley values have been widely used in the context of Explainable AI (XAI) to provide feature attributions as explanations [64, 65]. However, since computing Shapley values is intractable [62], various approximations have been used [66]. For models that work on images as inputs, popular techniques include aggregating neighboring pixels to form sub-pixels [67] to be more interpretabile. Other works require gradients to obtain pixel-level explanations and try to compute them efficiently [68]. We propose a model-agnostic approach that generates explanations in terms of a limited number of high-level human interpretable features.

Significant work has gone in to producing contrastive explanations [24, 10]. These have primarily been shown to work on structured tabular data. Counterfactual image generation has also been an active space of research [51, 52]. Often they require re-training a generative model [69], which is expensive or are specific to the choice of generative model [56]. We provide a minimal and scalable training process by utilizing only pretrained GANs for generating counterfactual images. Recent work by [70] comes closest to our work and seeks to use Shapley

values to explain models trained on high dimensional data. They use classical computer vision techniques to generate interpretable features. In contrast, we use generative models to produce counterfactual images to compute Shapley values efficiently while also providing contrastive explanations.

3.3 Preliminaries

In this section, we briefly discuss the concepts that are essential to understand our framework such as Generative Models, and Shapley Values.

3.3.1 Contrastive Counterfactuals & Generative Models

Contrastive counterfactuals have been the building blocks of ideas in philosophy and cognition that guide people's understanding and dictate how we explain things to one another [20] and have been argued to be central to explainable AI [21]. We are specifically interested in the implementation of this framework in the image classification problem in order to allow us to generate counterfactual images at will such that they increase or decrease the presence of a set of interpretable feature attributes. Generative models are vastly popular in different fields of AI, and their recent advancements in creating realistic images have made them a viable approach to produce a latent representation of an image dataset. In our experiments, we utilize StyleGAN2 [48] as a state-of-the-art generative model which can be used to generate high resolution and realistic images in different domains. StyleGAN feeds the latent variable into a mapping network that transforms it into an intermediate latent variable. Aside from its ability to produce styles, this transformation also provides the intermediary latent space as a more regulated domain to learn and traverse through interpretable attributes. We use the manipulation of this latent space to provide us with realistic counterfactual images according to our specifications.

3.3.2 Shapley Values

Shapley value is a concept from Game Theory that provides a unique solution to fairly allocating the total payout from a game to its individual players. A coalition is a set of players playing the game. A grand coalition contains all the players, while an empty coalition contains none. A value function $v(S)$ provides the scores obtained by the coalition S on playing the game. The score obtained by playing the game is called the payoff.

In the context of XAI, Shapley values have been used as a means to obtain feature contribution. The input features are the players in this game of obtaining predictions from the model. The prediction is then fairly distributed among the input features. For a coalition set S, the following formula provides the Shapley value ϕ for feature i:

$$\phi_i(v) = \sum_{S \subset N \setminus i} \frac{|S|!(n-|S|-1)!}{n!}(v(S \cup i) - v(S)) \tag{3.1}$$

where $n = |N|$ is the total number of input features. This is the weighted sum of marginal contribution of feature i across all possible coalitions that do not contain the feature i. Shapley values satisfy desirable axioms that make them make a good choice for generating feature contributions. We describe 3 of those axioms now-

Null : The null axioms states that if a feature that does not change the output when added to any coalition, it will get a Shapley value of zero.

Efficiency : The efficiency property states that the sum of Shapley values is equal to the difference between the prediction obtained by the grand coalition and empty coalition.

Symmetry: If two features behave in the same way across all possible coalitions, then their Shapley values will end up being the same.

The grand coalition is usually the entire instance to be explained, while the null coalition is an average instance or an instance composed entirely of default replacements for missing values for each feature(since they are all out of the coalition in the null coalition). The Shapley values

are then used to decompose the difference between an instance's prediction and the average prediction of the model to the individual features of the instance through the efficiency axiom. Since we iterate over all possible subsets of features that do not contain the *i*th feature, computing Shapley values in exponential in the number of features. This is particularly problematic in case of images, where we have large numbers of pixels as input features. For partial coalitions, we need to compute the model output given only the members in the coalition, which results in partially formed instances. This is usually done by either marginalizing [65] or conditioning [64]. However, both of these techniques are known to have issues [63]. Further, this vanilla Shapley value-based decomposition explains away the difference between the prediction of the input instance and the model prediction on an "average" instance. This average model prediction does not always correspond to a sensible input [66]. It has also been shown that standard feature contributions in isolation do not help humans reason as well as contrastive explanations do [20]. We aim to overcome these issues through our modified contrastive formulation of Shapley values using high-level interpretable input features as players.

3.3.3 Generating Counterfactual Images

We use a Shift Predictor model to obtain counterfactual images according to our needs. A shift predictor model is an MLP model that takes latent variable of an image from a generative model G and generates the latent variable for its counterfactual based on the attributes produced by a classifier. For a generative model $G : \mathscr{R}^d \rightarrow \mathscr{R}^n$ having a latent space with dimension d and a classifier $C : \mathscr{R}^n \rightarrow \mathscr{R}^m$ that predicts m attributes, we define our shift predictor as $M(\mathbf{z}, \hat{\mathbf{y}}) : \mathscr{R}^d \times \mathscr{R}^m \rightarrow \mathscr{R}^d$, where $\mathbf{z} \in \mathscr{R}^d$ is the latent variable for the input image and $\hat{\mathbf{y}}$ denotes the attributes for the intended counterfactuals. During training, shift predictor learns the directions in the latent space of G that correspond to changes in the attributes predicted by the classifier. Without need for any manual labeling, the training procedure only requires the latent variables of images from G to input the shift predictor and supervise it with the labels generated by the classifier.

During training, shift predictor learns to produce a counterfactual latent variable that satisfies any combination of attributes defined by $\hat{\mathbf{y}}$. In other words, if the classifier predicts a set of attributes $\mathscr{A} = \{A_1, A_2, ..., A_m\}$, shift predictor provides a counterfactual latent variable compatible with any selected subset of attributes $\bar{\mathscr{A}}$:

$$\hat{\mathbf{z}} = M(\mathbf{z}, \{A_i = \hat{a}_i \mid A_i \in \bar{\mathscr{A}}\}). \tag{3.2}$$

Under proper training, the shift predictor is an approximation of latent variable distribution conditioned by the subset of attributes $\bar{\mathscr{A}}$:

$$\hat{\mathbf{z}} \sim P(\mathbf{z} | \{A_i = \hat{a}_i \mid A_i \in \bar{\mathscr{A}}\}),\ \bar{\mathscr{A}} \subseteq \mathscr{A}. \tag{3.3}$$

The loss function in the training process pursues two objectives: 1) minimizing the error in prediction of attributes $\bar{\mathscr{A}}$ for the counterfactual image, 2) assuring a level of faithfulness of the generated counterfactual to the original image. The attribute loss $\mathscr{L}_a$ is defined as a cross entropy between the conditioned attributes and the attributes predicted by the classifier. During training, the conditioned attributes $\bar{\mathscr{A}}$ are distinguished from unset attributes so the loss will be only calculated for them. On the other hand, the faithfulness loss $\mathscr{L}_f$ is calculated as the normal distance between the original latent variables and their counterfactuals. The overall loss in the training process is defined as a combination of these two losses with a faithfulness factor γ which defines a balance between attribute accuracy of counterfactuals and their faithfulness to the original input:

$$\mathscr{L} = \mathscr{L}_a + \gamma \mathscr{L}_f = \sum_{A_i \in \bar{\mathscr{A}}} -\hat{\mathbf{y}}_i log(\mathbf{y}_i) + \gamma ||\hat{\mathbf{z}} - \mathbf{z}|| \tag{3.4}$$

3.3.4 Contrastive Explanations Using Shapley Values

The efficiency property of Shapley values allows us to decompose the score difference between the game played by the grand coalition and the empty coalition among all the players.

We define players in terms of interpretable attributes of the images. In images of human faces, these attributes could be attributes like hair, makeup, etc. We work in this interpretable space instead of pixel-level features. We define the empty coalition as the set of all zeros for these attributes. It corresponds to their natural or default values that they take on in the image. However, we can define the grand coalition by either increasing the attributes or decreasing them. We use the shift predictor described above to obtain images corresponding to this increase or decrease in the specific attribute of the image. For instance, we can define the grand coalition to be +1 for all interpretable features, which would correspond to an image that has been forced to increase the presence of all attributes in itself. This provides us with 2 images to compare - the original image corresponding to the empty coalition and the counterfactual image with a set of increased or reduced attributes($+1\uparrow, -1\downarrow$). We use Shapley values to decompose the difference in prediction obtained for these two images into their interpretable features. Our shift predictor can take any arbitrary vector specifying the direction of change of attributes and return a counterfactual image. This overcomes issues found in other traditional Shapley value-based frameworks of missing data for members out of coalition. For partial coalitions, we set the values for members out of the coalition to be 0. This instructs the shift predictor to not make any changes to that attribute. Hence, we directly compute value functions for any coalition using this pipeline. In this setup, we gain a contrastive notion and no longer compare to an "average" instance. Instead, we have two specific images- the original and its counterfactual, and we explain the difference in their predictions Shapley values.

3.4 Experiments

We run experiments to explain a classifier that is trained on face images. We have annotations for the set of interpretable attributes A that we will use to explain the model's behavior. We train a multi-task classifier built on top of a pretrained VGG cite backbone to predict interpretable attributes A for new unseen images. The CelebA dataset[11] is used as

Figure 3.1. Examples of original images and their corresponding counterfactual images.

the training set and provides 39 binary attributes, including attractiveness which we use as the target output Y for any black-box classifier of choice. As the set of interpretable explanatory attributes(A), we choose five other labels: blonde hair, heavy makeup, baldness, youngness, and maleness. We model attractiveness as the positive class $(y = 1.0)$ and unattractiveness as the negative class $(\hat{y} = 0.0)$ for the classifier to predict. The black-box classifier is built on a pretrained VGG backbone. We train our shift predictor model, as described earlier, with a faithfulness γ value of 0.09. We pass the images through the multi-task classifier to obtain the attribute values and through the black-box classifier to obtain the target label.

We sample images randomly from the StyleGan2 Generative Model and pass them through the black-box model. For every instance, we generate a counterfactual image through the shift predictor by defining our grand coalition in terms of increase or decrease of feature attributes. We obtain Shapley values-based contributions that explain away the difference in prediction between the original image(empty coalition) and the counterfactual image(grand coalition). Both

Table 3.1. Shapley Value based contributions explaining the difference in predictions between each pair of original and counterfactual images from Figure 3.1. Each element is of the form $(\uparrow / \downarrow, attribution)$.

IMAGE	YOUNG	HEAVY MAKEUP	BLOND HAIR	BALD	MALE
1	$\downarrow$ -0.28	$\downarrow$ -0.02	$\downarrow$ -0.03	$\uparrow$ -0.34	$\uparrow$ 0.07
2	$\downarrow$ -0.20	$\downarrow$ -0.07	$\downarrow$ -0.03	$\uparrow$ -0.23	$\downarrow$ -0.04
3	$\uparrow$ 0.23	$\uparrow$ 0.37	$\uparrow$ 0.15	$\downarrow$ 0.10	$\uparrow$ -0.05
4	$\uparrow$ 0.16	$\uparrow$ 0.18	$\uparrow$ 0.13	$\downarrow$ 0.23	$\downarrow$ 0.09

the original and contrastive images along with the Shapley values of the interpretable features are given as explanations. We report the Shapley value-based attributions for images in Figure 3.1 in Table 3.1. The original images have their default attributes, while counterfactual images are generated by modifying the interpretable images as shown by $\uparrow$ and $\downarrow$ signs for the increase and decrease of attributes, respectively, as listed in Table 3.1. We can observe that for the original image O_1, we generate the counterfactual image C_i by increasing maleness and baldness while decreasing youngness, heavy makeup, and blond hair. When this is done, the attractiveness score drops from 0.73 to 0.12. This drop is mainly due to increasing baldness which accounts for 0.34 of the 0.61 difference in attractiveness, while decreasing youngness(increasing age) contributed to 0.28 of the total drop. Similarly, we can look at other pairs of original images and counterfactuals, in conjunction with the assigned Shapley attributions to understand model behaviour.

3.5 Summary

As next steps, we want to make the shift predictor causal by incorporating causal graphs and also work towards a computationally efficient algorithm to compute these explanations.

Chapter 3, in full, is a reprint of the material as it appears in Lahiri, A., Alipour, K., Adeli, E., and Salimi, B. (2022). Combining counterfactuals with shapley values to explain image models. arXiv preprint arXiv:2206.07087.The dissertation/thesis author was the primary

investigator and author of this paper.

Chapter 4

Contrastive and Causal Shapley Values For Model Explanations

In this chapter, we introduce a framework that integrates game theoretic shapley values with the causal notions of direct, indirect and total effects. This enables us to obtain a more detailed understanding of model behaviour by enabling us to capture direct functional effects of a feature, indirect effects by virtue of affecting other influential features, and total effects. It also establishes a generalised standard contrastive setup for shapley values that can be coupled with any arbitrary value function to enable comparsion and contrast.

4.1 Motivation & Overview

Shapley values have emerged as a popular and principled method for generating post-hoc model agnostic explanations. These values possess unique and desirable axiomatic properties that make them ideal for representing feature attributions. However, recent works have shown that popular methods to estimate Shapley values can lead to counter-intuitive explanations. Furthermore, these explanations often lack a human-centric perspective and fail to offer actionable insights.

To overcome these limitations, we introduce a contrastive notion of Shapley values, thereby useful to assist human understanding through comparison and contrast. Our approach provides feature attributions that explain the disparities in predictions obtained for arbitrary pairs

of data instances, allowing for meaningful comparisons between instances. We leverage our contrastive Shapley framework and integrate the widely studied notions of causal direct, indirect and total effects with Shapley values. This integration enables us to provide more robust and specific information about the influence of features which captures different kinds of causal dependencies and comes with axiomatic guarantees. For instance, we can determine whether a feature directly affects the outcome through functional dependencies in the output prediction function or if it indirectly impacts the final prediction through interactions with other features. We present specific examples and real world settings where the combination of contrastive and causal Shapley values offers more intuitive, detailed, and informative explanations than existing methods.

The rise of Machine Learning(ML) has resulted in the widespread use of complex models across various critical sectors, underscoring the importance of comprehending how these models utilize input features to make predictions [71]. The Shapley value, originally developed in game theory and later adapted for explaining black-box ML models, plays a significant role in this context [61, 72]. Shapley value-based methods are renowned for their equitable distribution of "total payoff" among input features, reflecting the principle of distributive justice [73]. These methods offer valuable insights into the decision-making processes of models by uncovering intricate feature interactions and dependencies. However, traditional Shapley value techniques often face challenges in incorporating both contrastivity and causality, which are essential elements for creating explanations that align with human comprehension [21].

The integration of contrastive reasoning into Shapley value-based methods is crucial for bridging the gap between transparency and explanations that resonate with human cognition. Disciplines such as sociology, philosophy, and psychology have highlighted the natural inclination of humans towards contrastive analysis, where alternative scenarios are explored to gain a deeper understanding of the environment [20, 8, 57, 74, 75, 9]. In the field of Explainable AI (XAI), users often seek explanations in a comparative context, such as understanding why a ML model approved a loan for one individual but rejected another with similar financial profiles.

Contrastive explanations enable individuals to derive actionable insights and explore recourse strategies, empowering them to understand their situation and potentially bring about changes [30, 76, 3]. The integration of contrastive analysis into Shapley value-based methods goes beyond individual instances and incorporates comparative assessments, considering both the contributions of input features to predictions and their relative importance in comparison to alternative feature sets. By doing so, these methods provide comprehensive, contextually relevant explanations in XAI, aligning with human cognitive processes and enhancing our understanding of complex AI models.

Furthermore, conventional Shapley value-based methods in XAI often fall short by overlooking the causal relationships among input features. This oversight can result in explanations that contradict the inherent causal structures present in the data. For example, in a credit scoring model, features like "credit history" and 'income" may have causal interactions that influence the "credit score" outcome. An ideal explanation from this model should respect these causal relationships by providing insights into how changes in "income" can impact "credit history" and subsequently affect the "credit score". However, the causal dynamics among features within a model are often intricate and complex. To accurately capture these interactions, it is essential to differentiate between the direct and indirect influences of a feature on the outcome [57, 77]. The direct influence refers to the feature's independent effect on the outcome, while the indirect influence is mediated through its interactions with other variables. In the context of the credit scoring model, "income" may have a direct impact on "credit score", while its indirect effect may be channeled through "credit history". While some recent attempts[78],[79] have been made to incorporate causal reasoning into Shapley value explanations, these methods predominantly remain non-contrastive, focusing on the causal influences on a single instance rather than comparing outcomes across instances . This lack of contrastive analysis limits the scope of such explanations, as they do not address questions like "Why was this outcome predicted for A and not for B?". It also prevents these methods from directly aligning with Pearl's notions of direct and indirect effects [77]. This integration of causality and contrast into a unified explanation

framework remains a significant challenge.

The main contributions of our paper are as follows. We propose a novel contrastive formulation that computes attributions by capturing differences between user-defined instance pairs, as opposed to differences relative to an average baseline fixed instance. This transition allows for more intuitive, human-aligned exploration of the model's predictions. Critically, our method retains the unique properties of Shapley values—such as their unbiased distribution of attribution and accounting for cooperative interactions among features—while seamlessly integrating well-established notions of direct, indirect, and total effects from causal inference literature. The resulting attributions enable the disentanglement of direct functional dependencies from indirect influences, contributing to a deeper understanding of the model's decision-making process. It enables us obtain a set of three shapley based attributions that provide a holistic view about the feature's influence on model output. We also provide a practical method for estimating these values in real-world datasets, requiring only a causal graph and observational data from the user. Therefore, our framework lays the groundwork for more interpretable, actionable, and human-centric explanations in ML.

In the related work section, we describe the prior work on Shapley value and how it relates to our formulation. Section 4.2 first formalizes the notion of contrastive Shapley value and then describes different value functions to capture causal dependencies to evaluate direct, indirect and total effect of an attribute on the outcome. We use our method over real and synthetic data to generate feature attributions and demonstrate its effectiveness of capture causal dependencies in Section 4.3.

4.2 Contrastive & Causal Shapley

In this section, we present Contrastive Counterfactual Shapley values and compare them with traditional methods for feature attribution. Let $f(\cdot)$ be an arbitrarily chosen decision-making or predictive algorithm trained on features $\mathbf{X}$ that generates outcomes for individuals with

Table 4.1. $\mathbf{x}_t$ and $\mathbf{x}_r$ instances for the example on Job Approval.

	Years of Experience	Age	Advanced Degree	Interview	Job Approval
$\mathbf{x}_t$	1	1	1	0	1
$\mathbf{x}_t$	0	0	0	0	0

Table 4.2. Contrastive Shapley Values for feature attributions of scenario in Figure 4.3

	Years of Experience	Age	Advanced Degree	Interview
Direct	0	0	1	0
Indirect	$1/6$	$1/6$	$-1/3$	0
Total	$1/6$	$1/6$	$2/3$	0

feature vectors $\mathbf{x}$. Our objective is to provide an explanation for a specific prediction $f(\mathbf{x}_r)$ for an instance $\mathbf{x}_r$ in contrast to another individual prediction $f(\mathbf{x}_t)$ for an instance $\mathbf{x}_t$, while capturing the underlying causal relationships between features. Furthermore, we aim to generate fine-grained explanations that rank features based on their direct, indirect, and total effects on the outcome. By doing so, we can contrast the importance of each feature between the two predictions and gain a more in-depth understanding of the factors influencing the decisions made by the algorithm. Before formalizing the notion of contrastive Shapley value, we demonstrate the importance of calculating attribution scores with the following example.

Example 4.2.1 *Consider a simplified scenario where an ML model is trained to predict Job Approval (J) for an individual, using four inputs: Years of Experience (E), Age (A), possession of an Advanced Degree (D), and the result of the Interview Process (I), all of which can adopt binary values. Within this model, there are clear functional relationships indicating that Job Approval (J) is directly influenced by the Interview Process (I) and the presence of an Advanced Degree (D). For simplicity, suppose these dependencies can be captured by the following equations and has causal graph as shown in Figure 4.3:*

$$D = E \wedge A \qquad (4.1) \qquad\qquad J = I \vee D \qquad (4.2)$$

Let's examine two instances, $\mathbf{x}_t$ and $\mathbf{x}_r$, as depicted in Table 4.1. These two individuals contrast in all features except for I. The question at hand is: "What did $\mathbf{x}_t$ do differently from $\mathbf{x}_r$ that enabled $\mathbf{x}_t$ to secure the job?" As we explain the contribution of different features towards $\mathbf{x}_t$

relative to $\mathbf{x}_r$, we want the explanation framework to draw the following conclusions:

1. *The Interview Process I was identical for both individuals, therefore it has no marginal contribution to the $\mathbf{x}_t$'s outcome.*

2. *Age (A) and Years of Experience (E) did not directly influence job approval, but they did have an indirect effect. This kind of attribution is often crucial in contexts of justice and fairness where the goal is to determine whether an ML model utilized sensitive attributes directly or indirectly. The direct impact of an advanced degree (D) is* 1*, as it is the sole factor directly influencing the prediction outcome in this case. This information is vital for generating recourse for $\mathbf{x}_r$, potentially aiding them in reversing the decision in their favor.*

This toy example describes explanations that are useful to illustrate the key factors that helped $\mathbf{x}_t$ achieve a favorable outcome as compared to $\mathbf{x}_r$. In practice, the explanations are non-binary and the user would be interested in the relative importance of each of the features.

In this work, we leverage the Shapley value framework and extend it to generate contrastive and causal attribution scores to generate explanations consistent with the example described above.

Shapley value. Shapley value is a concept from Game Theory that provides a unique solution to fairly allocating the total payout from a game to its individual players. A coalition is a set of players playing the game. A grand coalition contains all the players, while an empty coalition contains none. A value function $v(S)$ provides the scores obtained by the coalition S on playing the game. The score obtained by playing the game is called the payoff.

To explain the output of an ML model, Shapley values are used to calculate the contribution of different features towards the outcome. The features $\mathbf{X}$ are modelled as the players in the game of obtaining predictions from the model where the model output is then fairly distributed among the input features. Using the Shapley value framework, the contribution ϕ_i for feature i is

calculated as follows:

$$\phi_i(v) = \sum_{S \subset X \setminus i} \frac{|S|!(n-|S|-1)!}{n!}(v(S \cup i) - v(S)) \tag{4.3}$$

where $n = |\mathbf{X}|$ is the total number of input features and v denotes the value function for a coalition set S. Intuitively, the feature importance score is the weighted sum of marginal contribution of feature i across all possible coalitions that do not contain the feature i. This formulation guarantees that for any value function v, attribution scores of all features add up to the value of grand coalition, i.e. $v(S = \mathbf{X})$. This property is often known as the efficiency property, which is one of the requirements of our scoring mechanism (as described in Example 4.2.1).

Contrastive Shapley value. We now extend the Shapley value framework to propose a contrastive feature attribution score by defining a contrastive value function v_{ct} for a target data point $\mathbf{x}_t$ with respect to a reference data point $\mathbf{x}_r$ as follows.

$$v_{ct}(S, \mathbf{x}_t, \mathbf{x}_r) = v(S \leftarrow \mathbf{x}_t) - v(S \leftarrow \mathbf{x}_r) \tag{4.4}$$

where $v(S \leftarrow \mathbf{x})$ denotes the value function where features in S are assigned the values according to $\mathbf{x}$. We show that calculating feature attribution scores with v_{ct} satisfy the following axiom.

Axiom 4.2.1 (Contrastive Efficiency) *Consider two datapoints $\mathbf{x}_r$ and $\mathbf{x}_t$ and a predictive algorithm $f(\cdot)$. The attribution scores of all features calculated using contrastive value function add up to the difference between the classifier predictions for these inputs,*

$$\sum_{i \in \mathbf{X}} \phi_i(v_{ct}) = f(\mathbf{x}_t) - f(\mathbf{x}_r)$$

Proof 1 *Using the efficiency axiom of Shapley value,*

$$\sum_{i \in \mathbf{X}} \phi_i(v_{ct}) \quad = \quad v_{ct}(\mathbf{X}, \mathbf{x}_t, \mathbf{x}_r) = v(\mathbf{X} \leftarrow \mathbf{x}_t) - v(\mathbf{X} \leftarrow \mathbf{x}_r) = f(\mathbf{x}_t) - f(\mathbf{x}_r)$$

In the upcoming sections, our focus revolves around exploring and refining contrastive Shapley scores using a range of formulations. Our primary objective is to effectively capture the diverse causal effects exhibited by features by coming up with appropriate value functions.

4.2.1 Contrastive Direct Shapley

Direct influence of a feature on the model outcome is crucial to understand if the classifier uses the feature for its prediction or not. The contrastive direct influence of a feature X_i captures the marginal change in prediction outcome when the attribute X_i is changed to $\mathbf{x}_t^i$ instead of $\mathbf{x}_r^i$ but all indirect effects do not change. This means that the direct attribution score of a feature i should be 0 if i is not used by the prediction algorithm (or has no direct impact on the outcome). We refer to this property as the functional irrelevance axiom and define it formally as follows.

Axiom 4.2.2 (Functional Irrelevance) *Consider a prediction algorithm $f(\cdot)$. If $f(\mathbf{x}) = f(\mathbf{x}')$ for all $\mathbf{x}$ and $\mathbf{x}'$ such that $\mathbf{x}_j = \mathbf{x}'_j, \forall j \neq i$ then the direct Shapley score of feature X_i denoted as $\phi_i^{de}(\mathbf{x}_t, \mathbf{x}_r)$ is zero.*

Functional Irrelevance (Axiom 4.2.2) is equivalent to the null axiom for traditional Shapley value. Since Shapley value based attribution guarantees fair allocation, we can capture the degree of influence of a feature by defining an appropriate value function. We use Pearl's notion of direct and indirect effects, as described in [77], to propose corresponding value functions.

Pearl defines the natural direct effect of a feature X on Y as the change that occurs when altering the value of X from x to x' within a specific environment $U = u$. This change is quantified as the difference in Y when all other variables $\mathbf{Z}$ are set according to $X \leftarrow x'$

(symbolized as $Z_{X\leftarrow x'}(u)$), while maintaining the assignment of x to X. This is formally expressed as $Y_{X\leftarrow x', Z_{X\leftarrow x}(u)}(u) - Y_{X\leftarrow x}(u)$. We broaden this concept to include a coalition set $\mathbf{S}$, examining the direct effect of transitioning the values in $\mathbf{S}$ from $\mathbf{s}$ to $\mathbf{s}'$ within the context $U = u$. In this scenario, the set $\mathbf{Z}$ represents features outside the coalition, i.e., $\mathbf{S}' = \mathbf{X} \setminus \mathbf{S}$. The direct-effect contrastive value function calculates the expected direct effect of modifying features in the coalition from $\mathbf{x}_r$ to $\mathbf{x}_t$. This is formally defined as follows:

$$v^d(S, \mathbf{x}_t, \mathbf{x}_r) = E_u[f(S \leftarrow \mathbf{x}_t, S'_{\mathbf{x}_r}(u))] - E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_r}(u))]$$

Here $S'_{\mathbf{x}_r}(u)$ is shorthand for $S'_{S\leftarrow \mathbf{x}_r}(u)$ which represents setting the features in S' to be equal to values they would have attained had features in S been equal to their corresponding values in $\mathbf{x}_r$. Intuitively, the direct effect measures the change in $f(\cdot)$ when all variables in S' are forced to be according to $\mathbf{x}_r$ and S is changed from $\mathbf{x}_r$ to $\mathbf{x}_t$. Using this notion, the direct effect contrastive Shapley value of a feature i is defined as

$$\phi_i^{de}(\mathbf{x}_t, \mathbf{x}_r) = \sum_{S \subset X \setminus i} \frac{|S|!(n-|S|-1)!}{n!} \Big(E_u[f(S \cup \{i\} \leftarrow \mathbf{x}_t, S''_{\mathbf{x}_r}(u))] - E_u[f(S \cup \{i\} \leftarrow \mathbf{x}_r, S''_{\mathbf{x}_r}(u))] \\ - \big(E_u[f(S \leftarrow \mathbf{x}_t, S'_{\mathbf{x}_r}(u))] - E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_r}(u))] \big) \Big)$$

where $S'' = S' \setminus \{i\}$.

Using the efficiency axiom, we get the following result.

Proposition 4.2.1 *Consider two datapoints $\mathbf{x}_r$ and $\mathbf{x}_t$ and a predictive algorithm $f(\cdot)$. The direct effect contrastive Shapley scores of all features adds up to the difference in predictions between $\mathbf{x}_t$ and $\mathbf{x}_r$: $(f(\mathbf{x}_t) - f(\mathbf{x}_r))$.*

Further, we can show that if $\mathbf{x}_r$ and $\mathbf{x}_t$ data points have the same value for feature i, i.e., $\mathbf{x}_t^i = \mathbf{x}_r^i$, then direct effect $\phi_i^{de}(\mathbf{x}_t, \mathbf{x}_r) = 0$. Intuitively, this holds because setting the feature i to that same value in the final function through its inclusion in the coalition will have the same marginal effect in both the cases where members in-coalition are set according to $\mathbf{x}_t$ and when

set according to $\mathbf{x}_r$. Therefore, they will cancel out in each coalition and will result in a $\phi_i^{de} = 0$.

Example 4.2.2 *In our example of Job Approval, we find that the ϕ^{de} value for the Advanced Degree feature is 1, indicating a direct effect on the Job Approval decision. On the other hand, the ϕ^{de} values for the remaining variables are all 0, suggesting no direct influence on the outcome. This aligns with our expectations, as the causal graph indicates that only the Advanced Degree and Interview Process variables have a direct impact on the Job Approval decision. However, since both instances in this comparison have the same Interview Process value, the Interview Process variable cannot exert a direct effect on the outcome. With these attribution values, an end user can discern that the favorable outcome for the target instance is solely attributed to the Advanced Degree feature, which has a direct effect of 100% on the Job Approval decision.*

4.2.2 Contrastive Indirect Shapley

In this part, we conceptualize the indirect effect value function as a means to calculate the indirect Shapley value of a feature. This function quantifies the cumulative impact that a feature i exerts on the prediction result via its descendants, with the exception of the target outcome. We begin by establishing a key property that is vital for capturing indirect effects.

Axiom 4.2.3 (Causal Irrelevance) *In a causal graph G, if either of the following conditions hold, then the indirect effect of a feature i is* 0*.*

1. *feature i has no ancestors and $f(\cdot)$ does not use any of the descendants of i for prediction, then $\phi_i^{ie} = 0$, i.e. $f(\mathbf{x}) = f(\mathbf{x}'), \forall \mathbf{x}, \mathbf{x}'$ where $\mathbf{x}_j = \mathbf{x}'_j, \forall j \neq \text{DESCENDANTS}_G(i)$.*

2. *feature i has no descendants and $f(\cdot)$ does not use i for prediction, i.e. $f(\mathbf{x}) = f(\mathbf{x}'), \forall \mathbf{x}, \mathbf{x}'$ where $\mathbf{x}_j = \mathbf{x}'_j, \forall j \neq i$.*

This axiom is similar to the null axiom, which characterizes the scenarios where the indirect effect of a feature i is absent. We now use Pearl's notion of indirect effect to define

indirect effect contrastive value function as

$$v^i(S, \mathbf{x}_t, \mathbf{x}_r) = E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_t}(u))] - E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_r}(u))]$$

Intuitively, the value function captures the effect of changing out-of-coalition features S' from $\mathbf{x}_r$ to $\mathbf{x}_t$ while fixing in-coalition features to that of $\mathbf{x}_r$. Note that $v^i(\mathbf{X}, \mathbf{x}_t, \mathbf{x}_r) = 0$. Using this value function, indirect effect shapley value of feature i is defined as

$$\phi_i^{ie}(\mathbf{x}_t, \mathbf{x}_r) = \sum_{S \subset X \setminus i} \frac{|S|!(n-|S|-1)!}{n!} \Big(E_u[f(S \cup \{i\} \leftarrow \mathbf{x}_r, S''_{\mathbf{x}_t}(u))] - E_u[f(S \cup \{i\} \leftarrow \mathbf{x}_r, S''_{\mathbf{x}_r}(u))] \\ - \big(E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_t}(u))] - E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_r}(u))] \big) \Big)$$

where $S'' = S' \setminus \{i\}$. Using the efficiency axiom, we get the following result.

Proposition 4.2.2 *Consider two datapoints $\mathbf{x}_r$ and $\mathbf{x}_t$ and a predictive algorithm $f(\cdot)$. The indirect effect contrastive shapley scores of all features add up to zero.*

Example 4.2.3 *In our Job Approval example scenario, Years of Experience (E) and Age (A) receive positive ϕ^{ie} values of $\frac{1}{6}$ (Table **??**). These positive values indicate the indirect influence of (E) and (A) on the Job Approval decision when comparing the target instance to the baseline instance. Specifically, when (E) and (A) are set to 1 in the target instance and 0 in the baseline, they contribute positively to the outcome. The Advanced Degree (D) feature obtains a negative ϕ^{ie} value of $\frac{-1}{3}$. This negative value signifies that the indirect influences of other features pass through (D) and are not attributed to it directly. The Interview Process feature has an ϕ^{ie} value of 0, indicating that it does not have any indirect influence on the Job Approval decision. The estimation of ϕ^{ie} values provides valuable insights to end-users, confirming that both (E) and (A) indirectly affect the model's outcome of Job Approval (J) through their impact on (D).*

4.2.3 Contrastive Total Shapley

We now assess the overall causal impact of a feature on the model outcome. To achieve this, we consider the total causal effect-based attribution, which adheres to the following axiom.

Axiom 4.2.4 *If a feature i has no causal effect on the prediction outcome $f(\cdot)$ (has zero direct and indirect attribution), then ϕ_i^{te} should be zero.*

We define total effect contrastive value function by calculating the expected difference in the model outcome for the two inputs, $\mathbf{x}_t$ and $\mathbf{x}_r$.

$$v^t(S,\mathbf{x}_t,\mathbf{x}_r) = E_u[f(S \leftarrow \mathbf{x}_t, S'_{\mathbf{x}_t}(u))] - E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_r}(u))]$$

$$\phi_i^{te}(\mathbf{x}_t,\mathbf{x}_r) = \sum_{S \subset X \setminus i} \frac{|S|!(n-|S|-1)!}{n!} \Big(E_u[f(S \cup \{i\} \leftarrow \mathbf{x}_t, S''_{\mathbf{x}_t}(u))] - E_u[f(S \cup \{i\} \leftarrow \mathbf{x}_r, S''_{\mathbf{x}_r}(u))] - \big(E_u[f(S \leftarrow \mathbf{x}_t, S'_{\mathbf{x}_t}(u))] - E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_r}(u))] \big) \Big)$$

where $S'' = S' \setminus \{i\}$.

By employing the total effect contrastive value function, we can generate Shapley values that are causally consistent for the various features. In the following, we demonstrate the properties exhibited by these feature attributions.

Proposition 4.2.3 *Total effect contrastive value of $\mathbf{x}_t$ with respect to $\mathbf{x}_r$ is equivalent to the different between the direct effect of $\mathbf{x}_t$ minus the indirect effect of the $\mathbf{x}_r$ relative to the target.*
$v^t(S,\mathbf{x}_t,\mathbf{x}_r) = v^d(S,\mathbf{x}_t,\mathbf{x}_r) - v^i(S,\mathbf{x}_r,\mathbf{x}_t)$

Proof 2 *Using the definition of direct and indirect value functions,*

$$v^d(S,\mathbf{x}_t,\mathbf{x}_r) = E_u[f(S \leftarrow \mathbf{x}_t, S'_{\mathbf{x}_r}(u))] - E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_r}(u))]$$

$$v^i(S,\mathbf{x}_r,\mathbf{x}_t) = E_u[f(S \leftarrow \mathbf{x}_t, S'_{\mathbf{x}_r}(u))] - E_u[f(S \leftarrow \mathbf{x}_t, S'_{\mathbf{x}_t}(u))]$$

Therefore,

$$v^d(S,\mathbf{x}_t,\mathbf{x}_r) - v^i(S,\mathbf{x}_r,\mathbf{x}_t) = E_u[f(S \leftarrow \mathbf{x}_t, S'_{\mathbf{x}_t}(u))] - E_u[f(S \leftarrow \mathbf{x}_r, S'_{\mathbf{x}_r}(u))] \quad (4.5)$$

$$= v^t(S,\mathbf{x}_t,\mathbf{x}_r) \quad (4.6)$$

This result allows us to obtain a similar result for the shapley values $\phi_i^{ie}(v)$,$\phi_i^{de}(v)$,$\phi_i^{te}(v)$.

Proposition 4.2.4 $\phi_i^{te}(\mathbf{x}_t,\mathbf{x}_r) = \phi_i^{de}(\mathbf{x}_t,\mathbf{x}_r) - \phi_i^{ie}(\mathbf{x}_r,\mathbf{x}_t)$

Proof 3 *Using the total shapley value formulation and the previous result,*

$$\phi_i^{te}(\mathbf{x}_t,\mathbf{x}_r) = \sum_{S \subset X \setminus i} \frac{|S|!(n-|S|-1)!}{n!} \qquad \Big(v^t(S \cup \{i\},\mathbf{x}_t,\mathbf{x}_r) - v^t(S,\mathbf{x}_t,\mathbf{x}_r)\Big)$$

Representing $\frac{|S|!(n-|S|-1)!}{n!}$ *as Coalition Weight,* $W(S,n)$ *,*

$$\begin{aligned}
\phi_i^{te}(\mathbf{x}_t,\mathbf{x}_r) &= \sum_{S \subset X \setminus i} W(S,n)\Big(v^d(S \cup \{i\},\mathbf{x}_t,\mathbf{x}_r) - v^i(S \cup \{i\},\mathbf{x}_r,\mathbf{x}_t) - v^d(S,\mathbf{x}_t,\mathbf{x}_r) + v^i(S,\mathbf{x}_r,\mathbf{x}_t)\Big) \\
&= \sum_{S \subset X \setminus i} W(S,n)\Big(v^d(S \cup \{i\},\mathbf{x}_t,\mathbf{x}_r) - v^d(S,\mathbf{x}_t,\mathbf{x}_r) - (v^i(S \cup \{i\},\mathbf{x}_r,\mathbf{x}_t) - v^i(S,\mathbf{x}_r,\mathbf{x}_t))\Big) \\
&= \sum_{S \subset X \setminus i} W(S,n)(v^d(S \cup \{i\},\mathbf{x}_t,\mathbf{x}_r) - v^d(S,\mathbf{x}_t,\mathbf{x}_r)) - \sum_{S \subset X \setminus i} W(S,n)(v^i(S \cup \{i\},\mathbf{x}_r,\mathbf{x}_t) \\
&- v^i(S,\mathbf{x}_r,\mathbf{x}_t))
\end{aligned}$$

Therefore,

$$\phi_i^{te}(\mathbf{x}_t,\mathbf{x}_r) = \phi_i^{de}(\mathbf{x}_t,\mathbf{x}_r) - \phi_i^{ie}(\mathbf{x}_r,\mathbf{x}_t) \quad (4.7)$$

Example 4.2.4 *In the Job Approval case, we obtain the following* Contrastive Total Shapley *values that represent the comprehensive influence of each feature. The Advanced Degree (A) receives the highest attribution of* $\frac{2}{3}$*, while both Years Of Experience (E) and Age (A) are assigned a value of* $\frac{1}{6}$*. This aligns with our understanding, given that the Advanced Degree directly impacts Job Approval and intuitively exerts a combined effect of its ancestors, (E) and (A). Conversely, the Interview Process (I) possesses a Contrastive Total Shapley value of 0, as it held the same*

value (0) in both $\mathbf{x}_r$ *and* $\mathbf{x}_t$.

4.2.4 Estimation

To compute contrastive Shapley values, estimating nested counterfactuals using observational data is crucial for accurately assessing the direct and indirect effects of features. While the identifiability of direct and indirect effects from observational data has been established [77, 80], estimating nested counterfactuals for high-dimensional mediating variables in real data can be challenging. In this paper, we address this challenge by leveraging the underlying structural equations with an additive noise model assumption. We employ Pearl's three-step procedure for counterfactual estimation: Abduction, Action, and Prediction [81, 57]. This approach allows us to estimate both counterfactuals and nested counterfactuals and compute contrastive Shapley values.

4.3 Experiments

In this section, we evaluate the effectiveness of our framework to generate useful attributions for both synthetic and real-world datasets. We discuss how direct, indirect and total contributions of a feature help to uncover observations about the behavior of the ML model, which were not possible earlier.

4.3.1 Real Data

In this experiment, we utilize the Adult income dataset from the UCI Repository [82] and adopt the same pre-processing steps as performed in the DiCE library [2]. We train a Random Forest Classifier and employ our method to generate explanations for a randomly selected subset of four inputs.

We adopt the causal graph proposed in [83] and estimate the underlying structural equations using additive noise models on the original dataset. Figure 4.2 illustrates the feature explanations generated by our method for four pairs of input instances. Each score provides

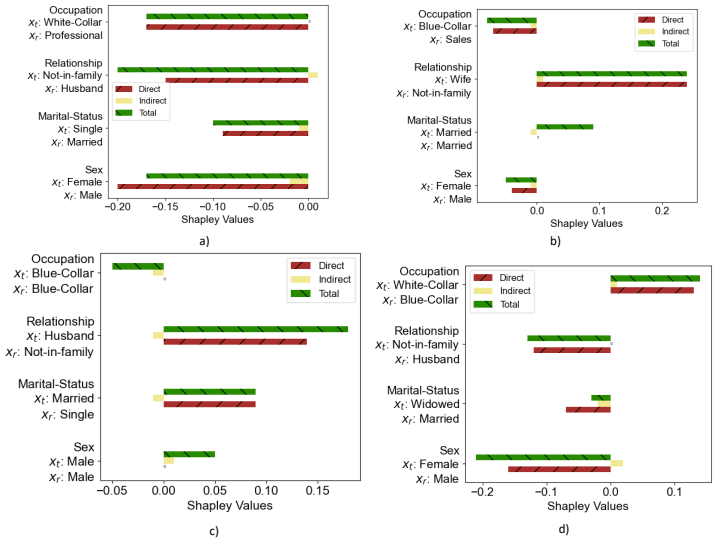

Figure 4.1. Comparison of ϕ^{de}, ϕ^{te} with ground truth scores estimated using the structural causal model for two different input pairs.

insights into how feature i contributes directly|indirectly|in total to the prediction of $\mathbf{x}_t$ in comparison to $\mathbf{x}_r$.

Figure 4.2 shows the negative influence of $\mathbf{x}_t$'s `gender = Female` and `Marital Status = Single`. This observation is consistent with prior XAI studies on this dataset, which have shown a strong influence of `Marital Status` on the outcome [84, 85]. This shows that changing marital status to `Married` could help the $\mathbf{x}_t$ individual improve the chances of outcome by around 10%. We observe a similar scenario in Figure 4.2 (b) where `gender = Female` has a negative impact but `Relationship = Wife` has a positive influence as compared to `Relationship = Not-in-family`. While comparing Figure 4.2 (a) and (d), we observe that `White-Collar` has a positive influence as compared to the $\mathbf{x}_r$ individual in (d) but a negative influence in (a). This shows that the same feature value can have positive or negative impact on the model outcome, and generating contrastive explanations can help to discover such behaviors. This analysis demonstrated the advantages of evaluating direct, indirect and total contribution of different features in contrast to a reference individual.

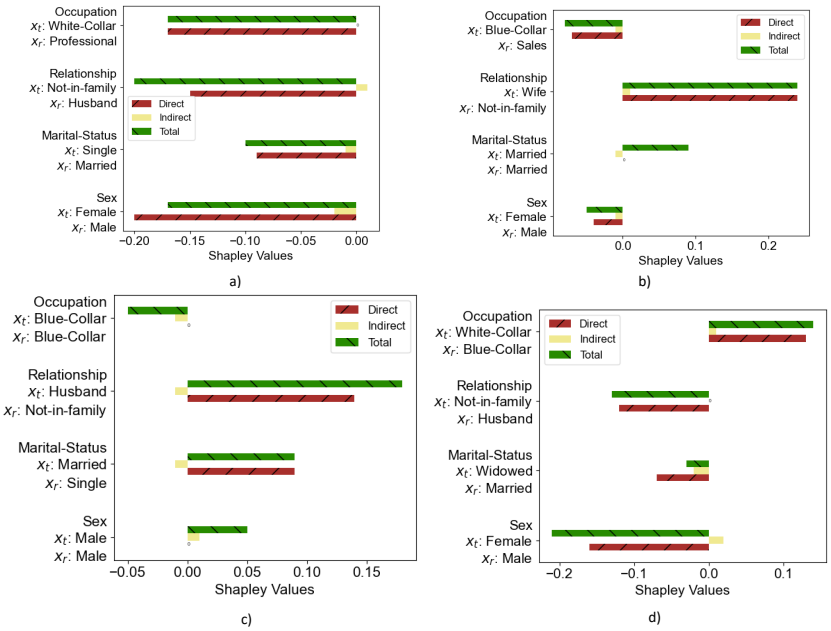

Figure 4.2. Direct, indirect and total attribution scores for four different pairs of data points from the Adult dataset.

4.3.2 Synthetic Experiment

We consider the causal graph considered in Example 4.2.1 (graph shown in Figure4.3) with four features (E, A, D, I) used to predict job approval (J). We generate the data using the equations.

$$E \sim \mathcal{N}(0,1), \quad A \sim \mathcal{N}(0,1), \quad D = 2*A + 3*D + \mathcal{N}(0,1), \quad I \sim \mathcal{N}(0,1), \quad J = D + 5*I$$

We consider two different pairs of reference and target data points (as shown in Figure 4.1) to compare the Shapley scores for the four features. Figure 4.1 shows that the attribution scores for all features are consistent with the ground truth values. This evaluation demonstrates the effectiveness of our method to evaluate the different Shapley values.

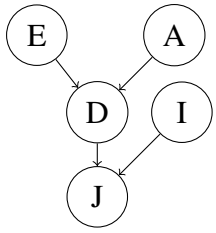

Figure 4.3. Causal model representing the effects in the example.

4.4 Related Work & Summary

In recent years, Shapley value-based explanations have gained significant attention in the field of explainable artificial intelligence (XAI) due to their strong theoretical foundations and game-theoretic principles. Lundberg and Lee [64] unified various explainability methods and introduced SHAP, which provides an approximation to Shapley values. Their approach incorporates conditional expectations over out-of-coalition features. However, [86] highlighted certain cases where this methodology produces incorrect explanations, suggesting the use of marginal expectations instead. They also proposed an interventional causal interpretation of Shapley values. Heskes et al. [78] integrated interventions into the value function, replacing conditional expectations with interventions. Another approach by Frye et al. [79] aimed to incorporate causal knowledge and structure by discarding permutations that do not adhere to the causal ordering, although this modification sacrifices the symmetry property of Shapley values. Heskes et al. [78] put forth definitions for direct and indirect effects in the framework of Shapley values. Nonetheless, these definitions depart from the well-established concepts of natural direct and indirect effects in causality and has a different interpretation. Importantly, these Shapley value variants are not inherently contrastive and are applied to individual points at a time. Moreover, Heskes et al. [78] defined direct and indirect effects with respect to an ”unknown” instance, preventing direct comparison with user-provided instances. Instead, the reference point defaults to the counterfactual situation when the value is unknown.

Contrastive explanations have also been the focus of recent research due to their intuitive appeal. However, these methods typically approach contrastiveness through a lens of sufficiency and necessity rather than in relation to another instance. For example, [87] generate contrastive explanations by identifying features that must minimally be present for a specific outcome and absent for its non-occurrence. It’s noteworthy that these are contrastive with respect to the target instance’s outcomes and not input features. This concept of contrastiveness is also used in [10], where probabilistic versions of sufficiency and necessity are used to provide explanations. [66]

discuss the selection of an arbitrary baseline data point in the context of Shapley values but do not explore the contrastive or causal implications of this approach. Our work aligns with the notion of contrastiveness as defined in [88], which posits that people seeking contrastive explanations are comparing two cases, asking ”Why P rather than Q?”. We aim to encapsulate this concept in our proposed method. Counterfactuals also present a form of contrast, but this is done in terms of the outcome label, not the features [2], [89]. Moreover, these methods are either non-causal [89] or apply causal constraints post-hoc after generating the counterfactual [2].

Chapter 4, in part is currently being prepared for submission for publication of the material. Lahiri, Aditya; Galhotra, Sainyam; Shanmugam, Karthik;Salimi,Babak. The dissertation/thesis author was the primary investigator and author of this material.

Chapter 5

Future Work

In this thesis, we look at approaches to understand black-box model decision making. Our methods encompassed vision based models as well as models trained on tabular data. We integrated game theoretic notions with causal effects, probabilistic scores as well as focussed on the importance and utility of contrastive approaches aided by conterfactuals. As future work, we can look to make the computation of these shapley value based scores more efficient. For certain value functions, perhaps we could gain a speed up on the exponential time taken to compute these attributions. We could also optimize these for certain causal structures. Furthermore, we could look at the application of these frameworks on other modalities such as text.

Bibliography

[1] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *International Conference on Machine Learning*, pp. 2376–2384, PMLR, 2019.

[2] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617, 2020.

[3] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[4] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: a review," *arXiv preprint arXiv:2010.10596*, 2020.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[6] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *International Conference on Machine Learning*, pp. 2484–2493, PMLR, 2019.

[7] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[8] T. Gerstenberg, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum, "How, whether, why: Causal judgments as counterfactual contrasts.," in *CogSci*, 2015.

[9] A. Morton, "Contrastive knowledge," *Contrastivism in philosophy*, pp. 101–115, 2013.

[10] S. Galhotra, R. Pradhan, and B. Salimi, "Explaining black-box algorithms using probabilistic contrastive counterfactuals," in *Proceedings of the 2021 International Conference on Management of Data*, pp. 577–590, 2021.

[11] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[12] N. Sani, D. Malinsky, and I. Shpitser, "Explaining the behavior of black-box prediction algorithms with causal learning," *arXiv preprint arXiv:2006.02482*, 2020.

[13] N. Pawlowski, D. C. Castro, and B. Glocker, "Deep structural causal models for tractable counterfactual inference," *arXiv preprint arXiv:2006.06485*, 2020.

[14] Á. Parafita and J. Vitrià, "Explaining visual models by causal attribution," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4167–4175, IEEE, 2019.

[15] S. Dash, V. N. Balasubramanian, and A. Sharma, "Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals," *arXiv preprint arXiv:2009.08270*, 2020.

[16] M. T. Bahadori and D. Heckerman, "Debiasing concept-based explanations with causal analysis," in *International Conference on Learning Representations*, 2020.

[17] A. Khademi and V. Honavar, "A causal lens for peeking into black box predictive models: Predictive model interpretation via causal attribution," *arXiv preprint arXiv:2008.00357*, 2020.

[18] M. N. Zaeem and M. Komeili, "Cause and effect: Concept-based explanation of neural networks," *arXiv preprint arXiv:2105.07033*, 2021.

[19] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards automatic concept-based explanations," *arXiv preprint arXiv:1902.03129*, 2019.

[20] M. M. De Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI Fall Symposium Series*, 2017.

[21] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[22] C. Russell, M. J. Kusner, J. Loftus, and R. Silva, "When worlds collide: integrating different counterfactual assumptions in fairness," in *Advances in Neural Information Processing Systems*, pp. 6414–6423, 2017.

[23] S. Greenland and J. M. Robins, "Epidemiology, justice, and the probability of causation," *Jurimetrics*, vol. 40, p. 321, 1999.

[24] R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma, "Towards unifying feature attribution and counterfactual explanations: Different means to the same end," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 652–663, 2021.

[25] D. S. Watson, L. Gultchin, A. Taly, and L. Floridi, "Local explanations via necessity and sufficiency: unifying theory and practice," in *Uncertainty in Artificial Intelligence*, pp. 1382–1392, PMLR, 2021.

[26] L. Bertossi, J. Li, M. Schleich, D. Suciu, and Z. Vagena, "Causality-based explanation of classification outcomes," *arXiv preprint arXiv:2003.06868*, 2020.

[27] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *International Conference on Artificial Intelligence and Statistics*, pp. 895–905, PMLR, 2020.

[28] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Inverse classification for comparison-based interpretability in machine learning," *arXiv preprint arXiv:1712.08443*, 2017.

[29] D. Mahajan, C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers," *arXiv preprint arXiv:1912.03277*, 2019.

[30] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19, 2019.

[31] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, vol. 9, pp. 11974–12001, 2021.

[32] M. O'Shaughnessy, G. Canal, M. Connor, M. Davenport, and C. Rozell, "Generative causal explanations of black-box classifiers," *arXiv preprint arXiv:2006.13913*, 2020.

[33] S. An, J.-J. Jeon, and H. Choi, "Exon: Explainable encoder network," *arXiv preprint arXiv:2105.10867*, 2021.

[34] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining classifiers with causal concept effect (cace)," *arXiv preprint arXiv:1907.07165*, 2019.

[35] D. C. Castro, J. Tan, B. Kainz, E. Konukoglu, and B. Glocker, "Morpho-mnist: quantitative assessment and diagnostics for representation learning," *Journal of Machine Learning Research*, vol. 20, no. 178, pp. 1–29, 2019.

[36] J. J. Thiagarajan, V. Narayanaswamy, R. Anirudh, P.-T. Bremer, and A. Spanias, "Accurate and robust feature importance estimation under distribution shifts," *arXiv preprint arXiv:2009.14454*, 2020.

[37] P. Schwab and W. Karlen, "Cxplain: Causal explanations for model interpretation under uncertainty," *arXiv preprint arXiv:1910.12336*, 2019.

[38] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

[39] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9841–9850, 2020.

[40] K. Liu, G. Cao, F. Zhou, B. Liu, J. Duan, and G. Qiu, "Towards disentangling latent space for unsupervised semantic face editing," *IEEE Transactions on Image Processing*, 2022.

[41] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or, "Face identity disentanglement via latent space mapping," *arXiv preprint arXiv:2005.07728*, 2020.

[42] Y.-D. Lu, H.-Y. Lee, H.-Y. Tseng, and M.-H. Yang, "Unsupervised discovery of disentangled manifolds in gans," *arXiv preprint arXiv:2011.11842*, 2020.

[43] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," in *International conference on machine learning*, pp. 9786–9796, PMLR, 2020.

[44] C. Tzelepis, G. Tzimiropoulos, and I. Patras, "Warpedganspace: Finding non-linear rbf paths in gan latent space," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6393–6402, 2021.

[45] H. Yang, L. Chai, Q. Wen, S. Zhao, Z. Sun, and S. He, "Discovering interpretable latent space directions of gans beyond binary attributes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12177–12185, 2021.

[46] O. K. Yüksel, E. Simsar, E. G. Er, and P. Yanardag, "Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14263–14272, 2021.

[47] I. Gat, G. Lorberbom, I. Schwartz, and T. Hazan, "Latent space explanation by intervention," *arXiv preprint arXiv:2112.04895*, 2021.

[48] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

[49] Z. Li and C. Xu, "Discover the unknown biased attribute of an image classifier," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14970–14979, 2021.

[50] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," *arXiv preprint arXiv:1907.03077*, 2019.

[51] D. Nemirovsky, N. Thiebaut, Y. Xu, and A. Gupta, "Countergan: Generating realistic counterfactuals with residual generative adversarial nets," *arXiv preprint arXiv:2009.05199*, 2020.

[52] P. Samangouei, A. Saeedi, L. Nakagawa, and N. Silberman, "Explaingan: Model explanation via decision boundary crossing transformations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 666–681, 2018.

[53] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, and I. Mosseri, "Explaining in style: Training a gan to explain a classifier in stylespace," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 693–702, 2021.

[54] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

[55] W. Nie, A. Vahdat, and A. Anandkumar, "Controllable and compositional generation with latent-space energy-based models," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[56] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 3, pp. 1–21, 2021.

[57] J. Pearl, *Causality*. Cambridge university press, 2009.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[59] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[60] J. Pearl, "Direct and indirect effects," *arXiv preprint arXiv:1301.2300*, 2013.

[61] E. Winter, "The shapley value," *Handbook of game theory with economic applications*, vol. 3, pp. 2025–2054, 2002.

[62] G. Van den Broeck, A. Lykov, M. Schleich, and D. Suciu, "On the tractability of shap explanations," in *Proceedings of the 35th Conference on Artificial Intelligence (AAAI)*, 2021.

[63] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," in *International Conference on Machine Learning*, pp. 5491–5500, PMLR, 2020.

[64] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[65] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE symposium on security and privacy (SP)*, pp. 598–617, IEEE, 2016.

[66] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *International conference on machine learning*, pp. 9269–9278, PMLR, 2020.

[67] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[68] X. Li, Y. Zhou, N. C. Dvornek, Y. Gu, P. Ventola, and J. S. Duncan, "Efficient shapley explanation for features importance estimation under uncertainty," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 792–801, Springer, 2020.

[69] S. Dash, V. N. Balasubramanian, and A. Sharma, "Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 915–924, 2022.

[70] D. de Mijolla, C. Frye, M. Kunesch, J. Mansir, and I. Feige, "Human-interpretable model explainability on high-dimensional data," *arXiv preprint arXiv:2010.07384*, 2020.

[71] H. Lakkaraju, J. Adebayo, and S. Singh, "Explaining machine learning predictions: State-of-the-art, challenges, and opportunities," *NeurIPS Tutorial*, 2020.

[72] D. Vale, A. El-Sharif, and M. Ali, "Explainable artificial intelligence (xai) post-hoc explainability methods: Risks and limitations in non-discrimination law," *AI and Ethics*, pp. 1–12, 2022.

[73] J. Konow, "Which is the fairest one of all? a positive analysis of justice theories," *Journal of economic literature*, vol. 41, no. 4, pp. 1188–1239, 2003.

[74] P. Lipton, "Contrastive explanation," *Royal Institute of Philosophy Supplement*, vol. 27, pp. 247–266, 1990.

[75] E. Grynaviski, "Contrasts, counterfactuals, and causes," *European Journal of International Relations*, vol. 19, no. 4, pp. 823–846, 2013.

[76] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera, "Algorithmic recourse under imperfect causal knowledge: a probabilistic approach," *arXiv preprint arXiv:2006.06831*, 2020.

[77] J. Pearl, "Direct and indirect effects," in *Probabilistic and causal inference: The works of Judea Pearl*, pp. 373–392, 2022.

[78] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen, "Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models," *Advances in neural information processing systems*, vol. 33, pp. 4778–4789, 2020.

[79] C. Frye, C. Rowat, and I. Feige, "Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability," *arXiv preprint arXiv:1910.06358*, 2019.

[80] J. Correa, S. Lee, and E. Bareinboim, "Nested counterfactual identification from arbitrary surrogate experiments," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6856–6867, 2021.

[81] M. Glymour, J. Pearl, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[82] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[83] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in data release," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1335–1344, 2017.

[84] B. Salimi, J. Gehrke, and D. Suciu, "Bias in olap queries: Detection, explanation, and removal," in *Proceedings of the 2018 International Conference on Management of Data*, pp. 1021–1035, 2018.

[85] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin, "Fairtest: Discovering unwarranted associations in data-driven applications," in *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 401–416, IEEE, 2017.

[86] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable ai: A causal problem," in *International Conference on artificial intelligence and statistics*, pp. 2907–2916, PMLR, 2020.

[87] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *Advances in neural information processing systems*, vol. 31, 2018.

[88] T. Miller, "Contrastive explanation: A structural-model approach," *The Knowledge Engineering Review*, vol. 36, p. e14, 2021.

[89] A. Van Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pp. 650–665, Springer, 2021.