A Causality Inspired Framework for Model Interpretation

Chenwang Wu*
University of Science and Technology of China
Hefei, Anhui, China
wcw1996@mail.ustc.edu.cn

Xiting Wang†
Microsoft Research Asia
Beijing, China
xitwan@microsoft.com

Defu Lian†
University of Science and Technology of China
Hefei, Anhui, China
liandefu@ustc.edu.cn

Xing Xie
Microsoft Research Asia
Beijing, China
xing.xie@microsoft.com

Enhong Chen‡
University of Science and Technology of China
Hefei, Anhui, China
cheneh@ustc.edu.cn

ABSTRACT

This paper introduces a unified causal lens for understanding representative model interpretation methods. We show that their explanation scores align with the concept of average treatment effect in causal inference, which allows us to evaluate their relative strengths and limitations from a unified causal perspective. Based on our observations, we outline the major challenges in applying causal inference to model interpretation, including identifying common causes that can be generalized across instances and ensuring that explanations provide a complete causal explanation of model predictions. We then present CIMI, a Causality-Inspired Model Interpreter, which addresses these challenges. Our experiments show that CIMI provides more faithful and generalizable explanations with improved sampling efficiency, making it particularly suitable for larger pretrained models.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Interpretability, causal inference, machine learning.

ACM Reference Format:
Chenwang Wu, Xiting Wang, Defu Lian, Xing Xie, and Enhong Chen. 2023. A Causality Inspired Framework for Model Interpretation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3580305.3599240

*Work done during an internship at Microsoft Research Asia.
†Corresponding authors.
‡Also affiliated with the State Key Laboratory of Cognitive Intelligence.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599240

1 INTRODUCTION

Although deep learning is widely used in various fields [14, 16, 19], deep models are mostly complex functions that humans cannot understand, which may reduce user trust. For this reason, eXplainable Artificil Intelligence (XAI) has received increasing attention. A fundamental question in XAI is: do explanations reveal important root causes of the model's behavior or merely correlations? The inability to distinguish correlation from causality can result in erroneous explanations for decision-makers [35]. The importance of causality is further highlighted by prominent research in human-computer interaction [41], in which extensive user studies reveal that in XAI, causality increases user trust and helps evaluate the quality of explanations. This result echoes major theories in cognitive science which present that humans build mental models of the world by using causal relationships [34, 46].

XAI provides an ideal environment for causality studies due to its adherence to fundamental causality assumptions, which are usually difficult to verify in other settings. For example, in XAI, we can obtain a set of variables (e.g., input data and model parameters) that construct a complete set of possible causes for model prediction, which ensures the satisfaction of the essential causal sufficiency assumption [36, 42]. In addition, the black-box models to be studied can be easily intervened, allowing the vital do-operator to be performed directly without any further assumptions such as ignorability or exchangeability. In contrast, the inability to perform different do-operators in the same instance is the fundamental problem of causality inference in more general scenarios [36].

Due to its importance and applicability, causality has attracted increasing attention in XAI. Multiple explanation methods [30, 40, 42] utilize causal analysis techniques such as interventions (e.g. input data perturbation), and some have achieved noteworthy success in delivering more trustworthy explanations. Despite this, a formal and unified causal perspective for explainability remains lacking and some key research questions remain challenging to answer, for example:

- **RQ1**: Can the existing explanation methods be framed within a theoretical causal framework? If so, what are the causal models employed, and what distinguishes them from each other?
- **RQ2**: What are the major challenges in leveraging causal inference for model interpretation and what benefits we may achieve by solving these challenges?

- **RQ3**: How can the causal model be improved to overcome these challenges?

In this paper, we aim to bridge the gap between causality and explainability by studying these issues.

We first provide a causal theoretical interpretation for explanation methods including LIME [40], Shapley values [30], and CXplain [42] (RQ1). Our analysis shows that their explanation scores correspond to (average) treatment effect [36] in causal inference to some extent, and they share the same causal graph, with only small differences such as the choices of the treatment (i.e., the perturbed features). This provides a unified view for understanding the precise meaning of their explanations and provides theoretical evidence about their advantages and limitations.

These observations allow us to summarize the core challenge in applying causal inference for model interpretation (RQ2). While it is easy for explanation methods to compute individual causal effects, e.g., understanding how much the model prediction will change when one input feature changes, the core challenge is *how to efficiently discover prominent common causes that can be generalized to different instances from a large number of features and data points.* Addressing this issue requires ensuring that the explanations are **(1) causal sufficiency** for understanding model predictions and can **(2) generalize to different instances**. These become increasingly important when the black-box model grows larger and there are more data points to be explained. In this case, it is vital that the explanations correspond to common causes that can be generalized across many data points, so that we can save users' cognitive efforts.

To solve the above challenges (RQ3), we follow important causal principles, and propose **C**ausality **I**nspired **M**odel **I**nterpreter (CIMI)[1]. Specifically, we first discuss different choices of causal graphs for model interpretation and identify the one that can address the aforementioned challenges. Based on the selected causal graph, we devise training objectives and desirable properties of our neural interpreters following important causal principles. We then show how these training objectives and desirable properties can be achieved through our CIMI framework.

Finally, we conduct extensive experiments on four datasets. The results consistently show that CIMI significantly outperforms baselines on both **causal sufficiency** and **generalizability** metrics on all datasets. Notably, CIMI's **sampling efficiency** is also outstanding, emphasizing that our method is quite timely, because it is more suitable for analyzing large models [57], in which each intervention requires a forward pass through the model. This makes our method particularly suitable for **larger pretrained language models**: its generalizability allows users to save cognitive efforts by checking a fewer number of new inputs, and its sampling efficiency makes it more suitable for analyzing large models, in which each sample (or intervention) requires a forward pass through the model.

2 REVISITING XAI FROM CAUSAL PERSPECTIVE

2.1 Preliminary about Causal Inference

We follow the common terminologies in causal inference [36] to discuss existing and our explanation methods.

[1]The source code of CIMI is available at https://github.com/Daftstone/CIMI.

Causal graph is used to formally depict causal relations. In the graph, each node is a random variable, and each direct edge represents a causal relation, which means that the target node (child) can change in response to the change of the source node (parent).
Do-operator is a mathematical operator for intervention. In general, applying a do-operator $do(E = e)$ on a random variable E means that we set the random variable to value e. For example,

- $P(Y = y|do(E = e))$ is the probability that Y is y when in every instance, E is assigned to e. This is a global intervention that happens to the whole population. In comparison, $P(Y = y|E = e)$ denotes the probability that Y is y on the subpopulation where E is observed to be e.
- $P(Y = y|do(E = e), C = c)$ applies do-operator to the subpopulation where the random variable C has value c.

Treatment effect is an important method to quantify how much causal effect a random variable E has on Y. Suppose that E is a binary value, the average treatment effect T of E on Y is

$$\begin{aligned} T(Y|do(E)) &= \mathbb{E}_c T(Y|do(E), C = c) \\ &= \mathbb{E}_c(Y(do(E = 1), C = c) - Y(do(E = 0), C = c), \end{aligned} \tag{1}$$

where $Y(do(E = e), C = c)$ represents the value of Y when E is set to e and all other causes C is fixed to c.

2.2 Causal Graph of Existing XAI Methods

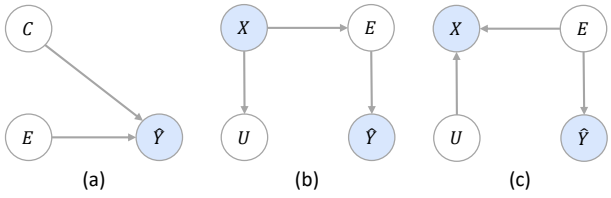

Figure 1: (a) The causal graph for existing methods, in which explanation E is not the sore cause for model prediction $\hat{Y}$; (b) Another causal graph, in which explanations are causally sufficient for prediction, but not generalizable; (c) Our proposed model which explanation E is generalizable and modeled as the only cause for $\hat{Y}$. Observed variables are shaded in blue.

Revisiting existing methods from the causal perspective allows us to show that many well-known perturbation-based methods such as LIME [40], Shapley values [30], and CXPlain [42] actually compute or learn the treatment effect, and that their causal graph corresponds to the one shown in Fig. 1(a). Notably, here we only briefly summarize the commonalities and differences among these XAI methods by presenting the main intuition behind the mathematical analysis, and formal theoretical analysis can be found in Appendix A of our full-version paper[2].

In the causal graph shown in Fig. 1(a), E corresponds to the specific treatment, characterized by one feature (or a set of features) to be perturbed. By $do(E = 1)$, these methods include the feature in the input, while $do(E = 0)$ does the opposite. Then, they obtain the model's outcome $\hat{Y}$ when E is changed and compute the treatment

[2]Since the supplementary material exceeds the space limit of two pages, we put all of the supplementary material into the full version of the paper, and it can be found in https://github.com/Daftstone/CIMI/blob/master/paper.pdf

effect $T(\hat{Y}|do(E)) = \hat{Y}(do(E=1), C=c) - \hat{Y}(do(E=0), C=c)$, where C denotes the context concerning E, or more intuitively, the features that remain unchanged after changing E. The treatment effect then composes (or is equal to) the explanation weight, revealing the extent to which the feature can be considered in the explanation, or "contribution" for each feature in the model prediction. It is worth noting that the do-operator here is directly applied to the data points and collects experiment outcomes, which is different from traditional modeling confounders and converting causal estimand into statistical estimand.

Although all three methods can be summarized using the framework in Fig. 1(a), they differ a little in terms of the following aspects. It is worth emphasizing that we will see how this unified view allows us to easily compare the pros and cons of each work.

- **Intervened features E.** CXPlain and Shapley value only consider one feature as E while LIME uses a set of features as E for testing. Thus, the former two methods cannot measure the causal contribution of a set of features without further extension or assumptions.
- **Context C.** Shapley values consider all subsets of features as possible context, while the other methods take the input instance x as the major context. Accordingly, Shapley values compute the average treatment effect on all contexts (i.e., all possible subsets of features) while others consider individual treatment effects. While individual treatment effects may be computed more efficiently and have a more precise meaning, their ability to generalize to similar inputs may be significantly reduced.
- **Model output $\hat{Y}$.** Most methods track changes in model predictions, while CXPlain observes how input changes the error of the model prediction. Thus, CXPlain may be more useful for debugging, while the others may be more suitable for understanding model behavior.

3 METHODOLOGY

3.1 Causal Graph

Causally insufficiency of explanations in Fig. 1(a). From the previous section, we have seen that existing work adopts the causal graph in Fig. 1(a). The major issue of this framework is that the model prediction $\hat{Y}$ is determined by both the explanation and the context, in other words, the explanation E is not the core cause for $\hat{Y}$. Thus, even if the users have carefully checked the explanations, the problem remains as long as the specific context is a potential cause for the model prediction, thereby the real complete reason for the model prediction cannot be seen.

Solving the causal insufficiency issue. The causal insufficiency of explanations may be addressed by removing context as a causal of the model prediction. Fig. 1(b) and (c) show two possible causal graphs to solve this issue. Here, X denotes the random variable for input instances. E and U are unknown random variables for explanations and non-explanations respectively, where $E = x_e$ means that the explanation for $X = x$ is x_e, and $U = x_u$ means that the non-explanation for $X = x$ is x_u. In both causal graphs, $\hat{Y}$ has the only parent (cause), which is the explanation, making the explanation sufficient to model prediction.

Issue of explanations' generalizability. While both causal graphs allow explanations to be the only cause of model predictions, Fig. 1(b) fails to model the explanation's generalizability: in this causal graph, the explanation may change in arbitrary ways when X changes. Generalizability is very important for model interpretability because it helps to foster human trust and reduce human efforts. Taking a pathological detector as an example, it would be quite disconcerting if entirely different crucial regions of the same patient were detected at different sectional planes. These prominent common causes that can be generalized to various instances help avoid the high cost of repetitive explanation generation and human investigation for similar instances.

Our choice. Considering the above, we choose the causal graph in Fig. 1(c), which resembles the Domain Generalization causal graph [31] and follows its common cause principle to build a shared parent node (in our case E) for two statistically dependent variables (in our case X and $\hat{Y}$). In the causal graph, it is evident that alterations to non-explanatory variable U have no impact on the explanation E or the prediction $\hat{Y}$, only resulting in slight variations in X. This demonstrates the stability of the explanation across different instances of X and its sufficiency as a cause for the model prediction $\hat{Y}$, as E is the only determining factor (parent) for $\hat{Y}$.

3.2 Causality Inspired Problem Formulation

Given the causal graph in Fig. 1(c), we aim to learn unobserved causal factors E and U, where E denotes the generalizable causal explanation for model prediction, and U denotes the non-explanations.

Following the common assumption of existing feature-attribution-based explanations, we assume that E and U could be mapped into the input space of X. More specifically, we assume that E is the set of features in X that influences $\hat{Y}$, while $U = X \backslash E$ is the other features in X that are not included in E. Equivalently, E and U can be represented by learning masks M over X:

- $E = M \odot X$, where $\odot$ is element-wise multiplication, and $M_i = 1$ means that the $i-th$ feature in X is included in the explanation.
- $U = (1 - M) \odot X$, where $M_i = 0$ means the $i-th$ feature in X is included in the non-explanation.

Our **goal** is to learn a function $\boldsymbol{g : X \rightarrow M}$ that inputs an instance $X = x$ and outputs the masks representing the causal factors E and non-causal factor U. Function g is the interpreter in this paper[3].

In our work, we relax $M \in \{0, 1\}$ to $[0, 1]$ (do not discretize the probability vector of g), which not only guarantees the end-to-end training of our neural interpreter but also distinguishes the different contributions of features to the output. We also try to discretize M using deep hashing technique [26], see Section 4.6 for the comparison and discussion.

3.3 Optimization Principles and Modules

It is impractical to directly reconstruct the causal mechanism in the causal graph of Fig. 1(c) since important causal factors are unobservable and ill-defined [31]. However, causal factors in causal graphs need to follow clear principles. We use the following two main principles in causality inference to devise desirable properties.

[3]Although $g : X \rightarrow M$ and the flow in Fig. 1(c) appear to be reversed, this is reasonable because $M = g(x)$ is a normal symmetric equation. Since the direction of flows in our framework does not imply causal direction, defining g is okay as long as $X \rightarrow M$ is a many-to-one (or one-to-one) mapping, which is exactly our cases.

Principle 1. ***Humean's Causality Principle [12]***[4]***:*** *There exists a causal link $x_i \rightarrow \hat{y}$ if the use of all available information results in a more precise prediction of $\hat{y}$ than using information excluding x_i, all causes for $\hat{y}$ are available, and x_i occurs prior to $\hat{y}$.*

Principle 2. ***Independent Causal Mechanisms Principle [38]:*** *The conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

Accordingly, we design three modules to ensure that the extracted explanations (causal factors) satisfy the basic properties required by Principles 1 and 2.

- **Causal Sufficiency Module.** Following Principle 1, we desire to discover E that is causally sufficient for $\hat{Y}$ by ensuring that E contains all the information to predict $\hat{Y}$ and explaining the dependency between X and $\hat{Y}$. Similarly, we also ensure that U is causally insufficient for predicting $\hat{Y}$.
- **Causal Intervention Module.** Following Principle 2, we ensure that U and E are independent by intervening U and guarantee that the learned $g(X) = E$ will not change accordingly. This also allows us to find explanations that can better be generalized to varying cases.
- **Causal Prior Module.** Following Principle 1, we facilitate the learning of explanations by using potential causal hints as inputs to the interpreter and weakly supervise over its output causal masks M. These learning priors enable faster and easier learning.

3.3.1 Causal Sufficiency Module. According to Principle 1, to ensure that E is a sufficient cause of $\hat{Y}$, it is necessary to guarantee that E is the most suitable feature for predicting $\hat{Y} = f(X)$, rather than other features U. In other words, x_e can always predict $f(x)$ through an optimal function f' that maps explanation x_e to $f(x)$, while non-explanation x_u cannot give meaning information for predicting $f(x)$. Accordingly, the causal sufficiency loss can be modeled as follows

$$\mathcal{L}_{s'} = \min_{f'} \mathbb{E}_x(\ell(f(x), f'(x_e)) - \ell(f(x), f'(x_u))), \tag{2}$$

where $\ell(\cdot)$ is the mean squared error loss, $x_e = g(x) \odot x$, $x_u = (1 - g(x)) \odot x$, and x is sampled from the entire model input space.

In practice, finding the optimal f' directly is very difficult due to the vast and sometimes even continuous input space. The interaction between optimizing f' and the interpreter g may also easily lead to unstable training and difficulty in converging to an optimal solution [8]. To address this issue, we approximate the optimal f' by using f, under the assumption that the difference between f' and f is minimum, considering that explanation x_e is in the same space with the origin model inputs X. By setting f' to f, we are actually minimizing each individual treatment effect, which has a precise causal meaning. Besides, since we do not have to optimize f', it may allow us to sample much fewer samples x' to optimize $\mathcal{L}_{s'}$ and learn an interpreter g. In summary, the causal sufficiency loss $\mathcal{L}_s$ rewritten as follows

$$\mathcal{L}_s = \mathbb{E}_x(\ell(f(x), f(x_e)) - \ell(f(x), f(x_u))), \tag{3}$$

where $x_e = g(x) \odot x$, and $x_u = (1 - g(x)) \odot x$.

[4]Although the principle needs a clear occurrence order of variables, we follow it only to determine the relationship between E/U and $\hat{Y}$, which can be satisfied.

3.3.2 Causal Intervention Module. Following Principle 2, we desire U and E to be independent, which makes it possible to find the invariable explanations of neighboring instances and improve the interpreter's generalizability. Despite the lack of true explanations for supervised training, we have the prior knowledge that the learned interpreter g should be invariant to the intervention of U, that is the $do(U)$ does not affect E. Based on this prior knowledge, we design a causal intervention loss to separate explanations.

First, we describe how to intervene on U. Following the common practice [31] in causal inference, we perturb the non-explanation x_u via a linear interpolation between the non-explanation positions of the original instance x and another instance x' sampled randomly from X. The intervention paradigm is shown as follows:

$$x_{int} = \underbrace{g(x) \odot x}_{invariant\ explanation} + \underbrace{(1 - g(x)) \odot ((1 - \lambda) \cdot x + \lambda \cdot x')}_{intervened\ non-explanation}, \tag{4}$$

where $\lambda \sim U(0, \epsilon)$, and ϵ limits the magnitude of perturbation. Furthermore, we can optimize the following causal intervention loss to ensure that U and E are independent.

$$\mathcal{L}_i = \mathbb{E}_x \ell(g(x), g(x_{int})). \tag{5}$$

This loss ensures that the generated explanations do not change before and after intervening in non-explanations. This invariant property guarantees local consistency of explanations. i.e., interpreters should generate consistent explanations with neighboring (or similar) data points. This coincides with the smooth landscape assumption of the loss function of the deep learning model [27], which may help to capture more generalizable features and improve the generalizability of the interpreter.

3.3.3 Causal Prior Module. To facilitate the learning of the interpreter, we 1) inject potential causal hints into the neural network of the interpreter, and 2) design a weakly supervision loss on the output causal masks M.

Interpreter neural network design. A core challenge in XAI is that there lack of prior knowledge about which architecture should be used for the interpreter [42]. When we learn an interpreter with the neural network, it is difficult to decide which neural network structure should be used. If the architecture of g is not as expressive and complex as the black-box model f, then how we can be sure that g has the ability to understand the original black-box f? If g could be more complicated than f, then it is prone to slow training efficiency and overfitting.

Our solution to this problem is inspired by Principle 1, which states that causes (model f) are more effective in predicting the effects (explanation x_e). Hence, we generate the explanation x_e by directly utilizing the parameters of the black-box model f. To achieve this, we use the encoding part of the black-box model f (denoted as f_e) as the encoder in our interpreter model g. The decoder of g is a simple neural network, denoted as ϕ. The ease of learning is supported by information bottleneck theory, which states that information in each layer decreases as we progress through the model [13]. Therefore, the input x contains the most information, while $f_e(x)$ contains less information as the information deemed unnecessary for prediction has been removed. The final prediction and ground-truth explanation use the least amount of information.

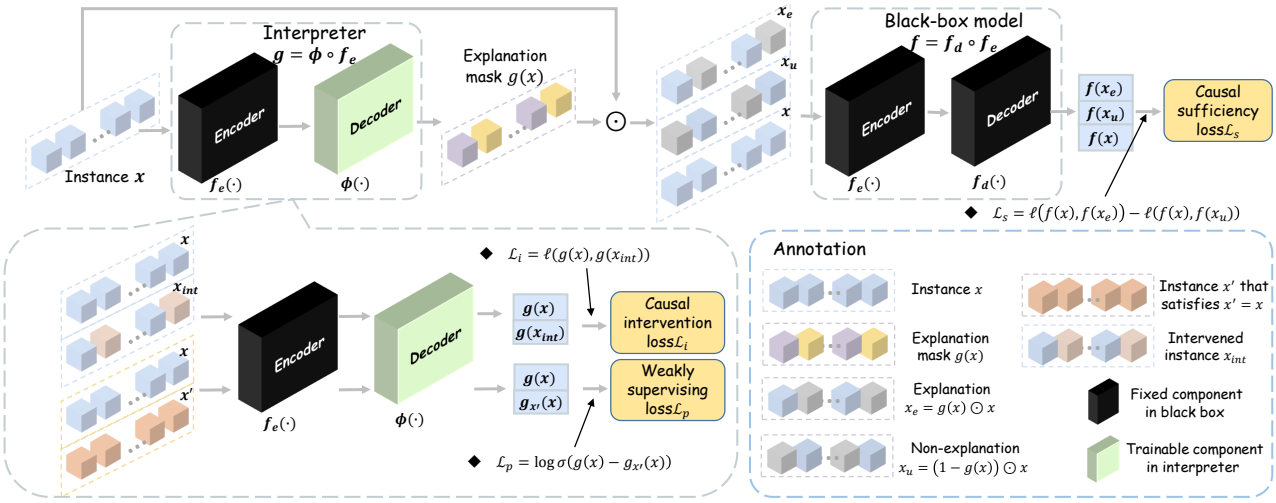

Figure 2: The framework of CIMI. The only trainable component is the decoder ϕ, which is a simple neural network that can be trained with a relatively small number of samples.

Consequently, compared with X, the last embedding layer output $f_e(X)$ is a better indicator to find the explanation.

Based on this observation, we design ϕ so its input concatenates the encoded embedding $f_e(x) \in \mathbb{R}^{|x| \times d}$ and the original instance embedding $v_x \in \mathbb{R}^{|x| \times d}$ along the axis 1, i.e., $[f_e(x); v_x]_1 \in \mathbb{R}^{|x| \times 2d}$, where d is the dimension of embedding, and the operator $[a; b]_i$ denotes the axis i along which matrix a and b will be joined. Therefore, the decoder ϕ maps input $[f_e(x); v_x] \in \mathbb{R}^{|x| \times 2d}$ to $[0, 1]^{|x| \times 1}$, and the i-th dimension of the output represents the probability that the token i is used for explanation. ϕ can be any neural network. In summary, the interpreter g can be reformulated as

$$g(x) = \phi([f_e(x); v_x]_1). \tag{6}$$

By setting the encoder in g as f_e, the architecture of g could be as complicated as f, and such a complex structure helps to fully understand the model's decision-making mechanism. g can also be considered simple, because the parameters of f_e in g are fixed and only the decoder ϕ is learnable, while only requiring a few additional parameters (1-layer LSTM + 2-layers MLP in our paper), avoiding the issues of overfitting and high training cost.

Weakly supervising loss. Without a further regularization loss on the causal factors, there exists a trivial solution (i.e., all explanation masks set to 1) that makes the interpreter collapse. A common regularization in causal discovery is sparsity loss which requires the number of involved causal factors to be small [5]. However, this sparsity loss may fail to adapt to the different requirements of different instances, as the constraints are the same for complicated sentences and simple sentences. Therefore, this poses difficulty in tuning the hyper-parameters for different datasets.

To tackle this issue, we leverage noisy ground-truth labels as a prior for the causal factor E to guide the learning process. Our approach is based on the intuition that the explanation for x should contain more information about x itself than information about another instance x'. Using this, we derive a weakly supervision loss by maximizing the probability that the token in instance x is included in x_e while minimizing the probability that a token not in x (noise) is predicted to be the explanation:

$$\mathcal{L}_p = \mathbb{E}_{x, x', x \neq x'} \log \sigma(g(x) - g_{x'}(x)), \tag{7}$$

where $g_{x'}(x)$ means to map $f_e(x)$ to x' in $g(x)$, refer to Eq. 6, that is, $g_{x'}(x) = \phi([f_e(x); v_{x'}]_1)$. Correspondingly, $g_x(x) = \phi([f_e(x); v_x]_1) = g(x)$, and the subscript in $g_x(x)$ are omitted for simplicity.

This weakly supervising loss prevents the interpreter from overly optimistically predicting all tokens as explanations, which helps alleviate trivial solutions.

3.3.4 Overall Framework and Optimization. **Overall loss function**. Combining the above three modules, the overall optimization objective of CIMI is summarized as follows and the framework is shown in Fig. 2.

$$\min_{\phi} \mathcal{L}_s + \alpha \mathcal{L}_i + \mathcal{L}_p, \tag{8}$$

where α is the trade-off parameter. Notably, the introduction of weakly supervising loss is to avoid the difficulty of tuning regularization parameters, so this term does not require trade-off parameters.

Analysis of the framework. As shown in Fig. 2, the only trainable parameter in our framework is the simple decoder in the interpreter g, which uses a 1-layer LSTM (hidden size is 64) and 2-layers MLP (64×16 and 16×2). This enables us to learn the interpreter efficiently with a small number of forward propagations through f. The validity of our framework can be further verified by considering the information bottleneck theory, which says that during the forward propagation, a neural network gradually focuses on the most important parts in the input by filtering information that is not useful for prediction through the layer [13]. According to this theory, setting the first part of the interpreter to the encoder f_e of the black-box model enables the interpreter to filter a large portion of noisy information that has been filtered by the black-box encoder, thus allowing us to learn the explanations more efficiently and faithfully. A more formal description of the validity of our framework is given in Appendix C.

Table 1: Faithfulness comparison when explaining BERT. * indicate that our method's improvements over the best results of baselines results are statistically significant for $p < 0.001$.**

Method	Clickbait			Hate			Yelp			IMDB		
	DFFOT↓	COMP↑	SUFF↓	DFFOT↓	COMP↑	SUFF↓	DFFOT↓	COMP↑	SUFF↓	DFFOT↓	COMP↑	SUFF↓
Gradient	0.5139	0.3651	0.1308	0.2776	0.4880	**0.1324**	0.4245	0.1497	0.2900	0.3216	0.1390	0.3999
Attention	0.5247	0.3655	0.1213	0.5933	0.2719	0.2718	0.5890	0.0809	0.3935	0.5459	0.0453	0.4496
AXAI	0.5234	0.3641	0.1245	0.4738	0.3210	0.2299	0.5120	0.1115	0.3088	0.4449	0.0891	0.4111
Probing	0.5278	0.3606	0.1249	0.5679	0.3013	0.2462	0.7133	0.0671	0.4392	0.6535	0.0445	0.4531
LIME	0.3994	0.4374	0.0778	0.2800	0.4860	0.1441	0.3346	0.2362	0.3201	0.2777	0.1953	0.4078
KernelSHAP	0.4447	0.4183	0.0725	0.4012	0.3963	0.1897	0.5484	0.0992	0.3488	0.5189	0.0565	0.4297
Rationale	0.5250	0.3651	0.1226	0.5937	0.2719	0.2719	0.5963	0.0838	0.3892	0.5501	0.0420	0.4533
CXPlain	0.4505	0.4092	0.0952	0.3796	0.4414	0.1438	0.4544	0.2287	0.3121	0.2894	**0.3273**	0.4094
Smask	0.5268	0.3561	0.1320	0.6121	0.2722	0.2735	0.5894	0.0839	0.3863	0.5594	0.0446	0.4492
CIMI	**0.3826**	**0.4612**	**0.0416**	**0.2761**	**0.5022**	0.1497	**0.1896**	**0.3100**	**0.2500**	**0.1270**	0.3270	**0.3516**
t-test	***	***	***	***	***		***	***	***	***		***

Relationship with disentanglement. Our method falls into the group of methods that **disentangle the causal effects** of variables [18], while most existing disentanglement methods focus on disentangling the latent representations [53]. The additional causal perspective is essential for ensuring the extraction of common causes to model prediction, which improves both explanation faithfulness and generalizability. To better illustrate the effectiveness, we implement a variant of CIMI that uses the representation disentanglement loss [53], which can be found in Appendix E.7.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets. We use four datasets from the natural language processing domain, including **Clickbait** [2], **Hate** [9], **Yelp** [56], and **IMDB** [32]. See Appendix D.1 for their details.

4.1.2 Black-box Models and Baselines. Although pre-training models are brilliant in many fields, it is difficult to answer what information they encode [25]. For this reason, we choose two well-known pre-training models, BERT [10] and RoBERTa [29], as the black box models to be explained. Notably, the main body shows the experimental results on BERT, and the results of RoBERTa can be found in Appendix E.2 and E.3.

We compare CIMI with **Gradient** [47], **Attention** [3], **LIME** [40], **KernelSHAP** [30], **Rationale** [24], **Probing** [1], **CXPlain** [42], **AXAI** [39], **Smask** [27]. Their details can be found in Appendix D.3. Although there are some works [28, 37] that can be used to extract causal explanations, they often make strict assumptions about the underlying data format, so they cannot be compared fairly, and we put the comparison in Appendix E.1.

4.1.3 Evaluation Metrics. First, we evaluate the causal sufficiency using three faithful metrics (see Appendix D.4 for more details).

Decison Flip-Fraction of Tokens (DFFOT) [43], which measures the minimum fraction of important tokens that need to be erased in order to change the model prediction.

Comprehensiveness (COMP) [11], which measures the faithfulness score by the change in the output probability of the original prediction class after the important tokens are removed.

Sufficiency (SUFF) [11], in contrast to COMP, it only keeps important tokens and compares the changes in output probabilities over the original predicted class. Notably, this metric is not equivalent to the causal sufficiency we focus on.

The number of important tokens is selected from $\{1, 5, 10, 20, 50\}$ and the average performance is taken. Notably, SUFF and COMP have been proven to be more faithful to the model prediction than other metrics [6]. In addition to the above metrics, we use AvgSen to measure the explanation's generalizability.

Average Sensitivity (AvgSen) [52], which measures the average sensitivity of an explanation when the input is perturbed. In our experiments, we replace 5 tokens per instance and calculate the sensitivity of top-10 important tokens.

In addition, we present the generated explanation instances in Appendix E.9 for a more intuitive evaluation.

4.1.4 Parameter Settings. For the two pre-training models used, BERT and RoBERTa, we both add two-layer MLP as decoders for downstream tasks. In all four datasets, the optimization is based on Adam with a learning rate of $1e-5$. The training epoch is 20, and the batch size is 8. For the proposed CIMI, without special instructions, we train 100 epochs on Clickbait and Hate, and 50 epochs on the other two larger datasets to improve efficiency. For the trade-off parameter α, set it to 1, 1, 1, 0.1 on Clickbait, Hate, Yelp, and IMDB respectively. In addition, the perturbation magnitude ϵ in the causal intervention module is set to 0.2.

4.2 Faithfulness Comparison

In this section, we evaluate the causal sufficiency of the explanations using faithfulness metrics. Table 1 summarizes the average results of 10 independent repeated experiments.

First, it can be seen that the proposed method achieves the best or comparable results compared with the baselines on various datasets. In particular, this improvement is more pronounced on more complex datasets (from Clickbait → IMDB). For example, the improvement over the best baselines reaches 119% on IMDB w.r.t. DFFOT metric. Such invaluable property could adapt to the complex trend of black-box models. This gratifying result verifies that CIMI can generate explanations that are more faithful to

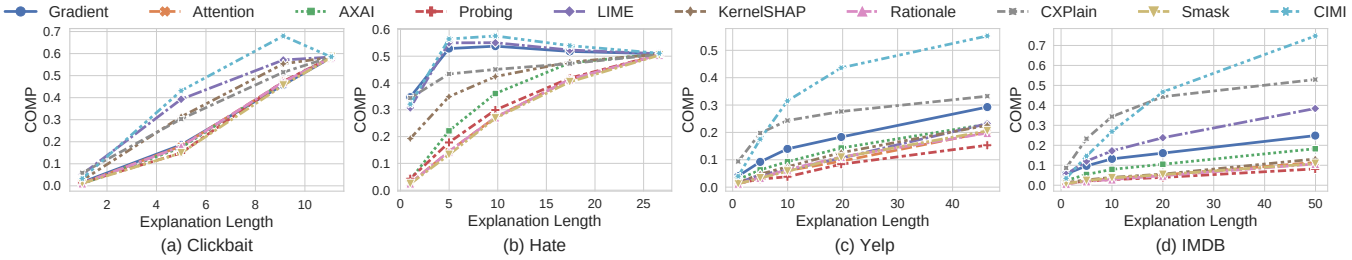

Figure 3: Performance comparison concerning COMP under different length explanations.

Table 2: Generalizability comparison under BERT. IMP indicates the improvement of our method compared to baselines.

Method	Clickbait		Hate		Yelp		IMDB		AVG_IMP(%)
	AvgSen↓	IMP(%)	AvgSen↓	IMP(%)	AvgSen↓	IMP(%)	AvgSen↓	IMP(%)	
Gradient	0.2182	43.18	0.4530	123.68	0.7934	561.19	0.8088	612.60	335.16
Attention	0.2155	41.42	0.5413	167.29	0.8642	620.18	0.9482	735.44	391.09
Lime	0.2036	33.59	0.4689	131.56	0.7880	556.68	0.8555	653.74	343.89
KernelSHAP	0.2022	32.66	0.4989	146.38	0.8280	589.96	0.9180	708.82	369.45
Rationale	0.2163	41.93	0.5440	168.64	0.8650	620.87	0.9497	736.73	392.04
Probing	**0.1460**	-4.23	<u>0.2093</u>	3.34	<u>0.2953</u>	146.08	<u>0.2873</u>	153.14	74.58
CXPlain	0.2101	37.86	0.5066	150.19	0.8135	577.95	0.8315	632.58	349.65
AXAI	0.2146	40.79	0.5127	153.17	0.8221	585.08	0.9103	702.05	370.27
Smask	0.2179	42.98	0.5532	173.20	0.8662	621.80	0.9468	734.16	393.03
CIMI	<u>0.1524</u>		**0.2025**		**0.1200**		**0.1135**		

the model. Second, Gradient has impressive performance in some cases, which indicates that their linear assumption can reflect the model's decision-making process to some extent. Third, among the perturbation-based methods, LIME, KernelSHAP, and CXPlain all show satisfactory performance. Especially LIME based on local linear approximation, which once again verifies the rationality of the first finding, the model linear assumption.

In addition, we also illustrate the performance w.r.t. COMP under different explanation lengths, as shown in Fig. 3 (similar findings can be obtained when concerning SUFF). The experimental results show that regardless of explanation length, CIMI exhibits significant competitiveness. The above results demonstrate the power of causal principle constraints in CIMI for understanding model predictions.

4.3 Generalizability Comparison

We use AvgSen to evaluate the generalizability of the explanations to neighboring (similar) samples. It is undeniable that for AvgSen, some important tokens included in the explanation may be replaced, but the probability is low, especially in Yelp and IMDB which have more tokens. The results are summarized in Table 2. It can be found that the explanations generated by CIMI are the most generalizable. Specifically, in the four datasets, at least 8 of the top-10 important tokens before and after perturbation are consistent, which is impossible for other methods. Besides, as the dataset becomes more complex, our performance remains stable while the baselines decrease significantly. these results demonstrate the outstanding ability of the proposed method to capture invariant generalizable features. Additionally, We conduct generalizability evaluation in attack settings [45, 50], see Appendix E.8 for the comparison.

4.4 Effectiveness of Causal Modules

4.4.1 Effectiveness w.r.t. Faithfulness. In this section, we verify the effectiveness of the proposed three causal modules concerning faithfulness. We define versions that remove causal sufficiency loss, causal intervention loss, weakly supervising loss, and interpreter's encoder f_e as CIMI-s, CIMI-i, CIMI-p, and CIMI-f, respectively. Their sufficient effects are shown in Fig. 4. Overall, removing any module leads to performance degradation, which justifies the design of three modules. Specifically, first, we found that the removal of the causal sufficient module (CIMI-s) has an impact on sufficient performance, which reasonably explains the original intention of our design of this module, ensuring that the explanations E are causally sufficient for the model predictions $\hat{Y}$. Second, the sufficient impact of the causal intervention module (CIMI-i) is marginal, since this module is designed primarily for the explanation's generalizability. Finally, both the weakly supervising loss and the interpreter's encoder design in the causal prior module can assist the model to learn more easily.

4.4.2 Effectiveness w.r.t. Generalizability. In this section, we discuss the generalizable effect of the three causal modules. Keeping the experimental settings consistent with Section 4.3, the results are illustrated in Fig. 5. First, we find that the causal sufficiency module helps improve generalizability against perturbations on Clickbait and Hate, but significantly degrades performance on Yelp and IMDB. We suspect that in the latter two datasets, CIMI-s explanations are not faithful to model prediction (Fig. 4 (c)(d)), then it is difficult to capture the invariant explanations of model decisions from similar instances, resulting in a decrease in generalizability.

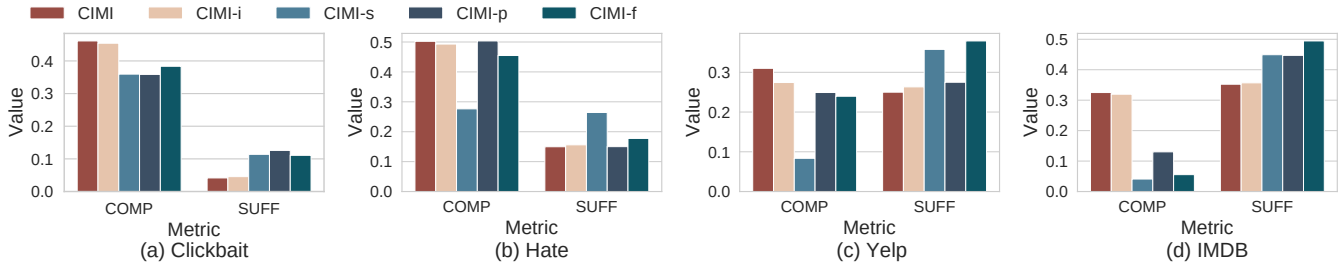

Figure 4: The faithful effect of the causal modules concerning COMP and SUFF.

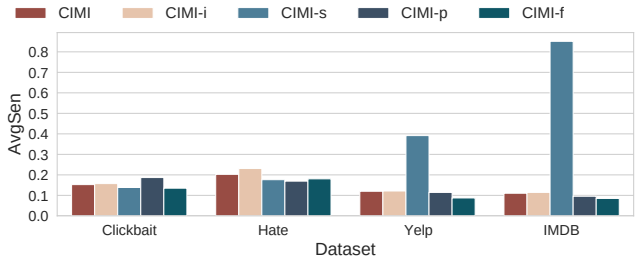

Figure 5: The generalizable effect of the causal modules concerning AvgSen.

Second, only CIMI-i's performance decreases consistently across the four datasets. This is because the causal intervention module aims to make U and E independent to ensure that the explanation can be generalizable to similar instances, that is, generalizability. Removing it resulted in performance degradation justifying the rationality of this module design.

4.5 Sampling Efficiency

Fig. 6 illustrates the performance of various perturbation-based methods under the same forward propagation times to measure the sampling efficiency. First, CXPlain's explanation mechanism makes each sample x perturbed at most $|x|$ times, so it shows high efficiency on small datasets, e,g., Clickbait and Hate. However, it is meaningless to talk about efficiency without explaining quality. Second, LIME performs well on small datasets (e.g., Clickbait), however, as the dataset becomes more complex, more sampling is required to generate high-quality explanations. Rationale's training is unstable and prone to trivial solutions, resulting in insensitivity to the number of sampling. Finally, our method significantly outperforms baselines, especially on Hate, where we only need 3 samplings to outperform baselines with 100 samplings. This benefits from the generalization of the neural network under the constraints of the causal principle, which summarizes the causal laws from a large number of data points and generalizes them to different instances, ultimately improving efficiency.

4.6 Performance Comparison of Different Discretization Strategies

In CIMI, we use Softmax function for differentiable training, in this section, we added two variants of our method that discretize U and E by using Gumbel-Softmax [17] and Deep Hash Learning [4, 26], denoted as CIMI-Gumbel and CIMI-DHL, respectively. As shown in Fig. 7 below, discrete masks help improve the explanation generalizability at the expense of a significant performance decline w.r.t. faithfulness. Specifically, in most cases, Deep Hash Learning contributes to explanations' generalizability (AvgSen), because the change in the mask value domain ($[0, 1] \rightarrow \{0, 1\}$) enables the explanations to be insensitive to noise. However, this discrete mask cannot distinguish the relative importance between features (e.g., when the probabilities of two words being explanations are 0.9 and 0.6, respectively, they are considered indistinguishable explanations after discretization), leading to a significant decline in performance during faithfulness tests that require a correct ordering of features according to their relative importance. We will add these analyses to the paper, which we believe will help readers better understand the motivation for differentiable training.

4.7 Usefulness Evaluation

In addition to allowing us to better understand the model, the explanation can also assist people in debugging the model. Noisy data collection can cause the model to learn wrong correlations during training. To this end, this section analyzes the effectiveness of various explanation methods in removing shortcut features. We use a subset of 20 newsgroups that classify "Christianity" and "Atheism". The reason for choosing this dataset is that there are many shortcut features in its training set, but the test set is clean. For example, 99% of the instances in the training set where the word "posting" appears is in the category of "Atheism".

To test whether an explanation method can help detect shortcuts, we first train a BERT model on the noisy training set. Then, we obtain explanations of different methods and treat a token in the explanation as the potential short-cut if it does not appear in the clean test (more details in Appendix F). We then retrain the classification model after removing shortcuts. The metric for evaluating the quality of shortcuts is based on the retrained model's performance (better classification performance implies that the shortcuts found are more accurate). The result is shown in Fig. 8. First, both explanation methods can effectively remove shortcuts to improve the model performance. Second, the improvement of the proposed CIMI is more obvious, verifying the usefulness on debugging models.

5 RELATED WORK

Existing explainable works can be divided into self-explaining methods and post-hoc methods [33].

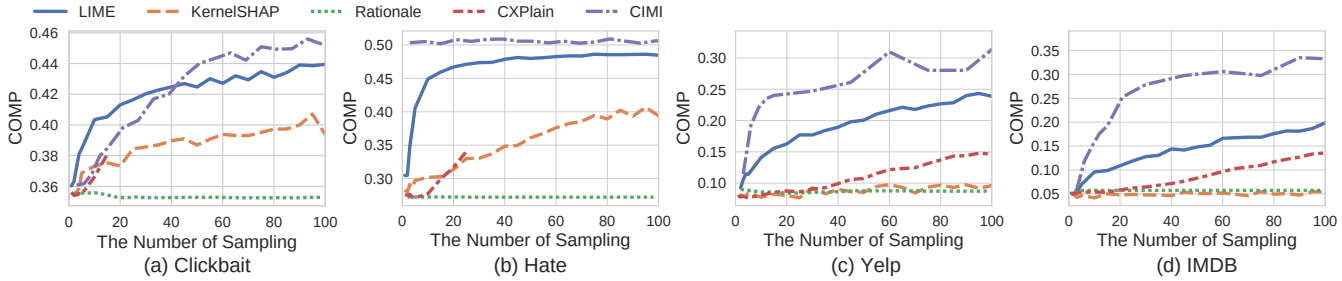

Figure 6: Performance comparison with the different number of sampling (perturbation) w.r.t. SUFF.

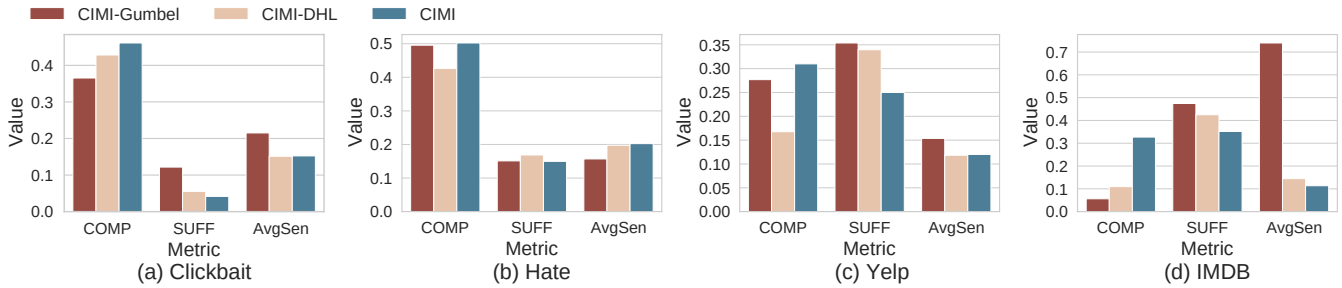

Figure 7: Performance comparison under different discretization strategies.

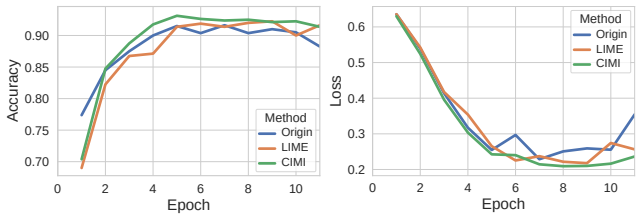

Figure 8: Usefulness evaluation. Classification performance comparison before and after deleting shortcuts.

The self-explaining method focuses on building model architectures that are self-explainable and transparent [51], such as decision tree [22], rule-based models [49], self-attention mechanisms, [3, 55]. In order to provide rules that are easy for humans to understand, they are often too simple to enjoy both interpretability and predictive performance [23]. Recently, methods of integrating add-on modules have received increasing attention [7, 15, 20, 23]. However, the process of generating explanations remains opaque.

Post-hoc interpretation has received more attention as models have gradually evolved into incomprehensible highly nonlinear forms [51]. Gradient-based methods [30, 44, 47, 48, 54] approximate the deep model as a linear and accordingly incorporate the gradient as feature importance. Admittedly, the gradient is only an approximation of the decision sensitivity. Influence function [21] also has been introduced to understand models, which efficiently approximates the impact of perturbations of training data through a second-order optimization strategy. Recently, casual interpretability has attracted increasing attention because it focuses on a fundamental question in XAI: whether existing explanations capture spurious correlations or remain faithful to the underlying causes of model behavior. First, many well-known perturbation-based methods such Shapley values [30], LIME [40], and Smask [27] implicitly use causal inference, and their explanatory scores correspond exactly to the (average) treatment effect [36]. From a causal point of view, the slight difference between them lies only in the number of features selected, the contextual information considered, and the model output. CXPlain [42] explicitly considered the non-informative features should have no effect on model predictions.

6 CONCLUSION

We reinterpreted some classic methods from causal inference and analyzed their pros and cons from this unified view. Then, we revealed the major challenges in leveraging causal inference for interpretation: causal sufficiency and generalizability. Finally, based on a suitable causal graph and important causal principles, we devised training objectives and desirable properties of our neural interpreters and presented an efficient solution, CIMI. Through extensive experiments, we demonstrated the superiority of the proposed method in terms of the explanation's causal sufficiency and generalizability and additionally explored the potential of explanation methods to help debug models.

ACKNOWLEDGMENTS

The work was supported by grants from the National Key R&D Program of China (No. 2021ZD0111801) and the National Natural Science Foundation of China (No. 62022077).

REFERENCES

[1] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR'17*.

[2] Aman Anand. 2020. Clickbait Dataset. (2020). https://www.kaggle.com/datasets/amananandrai/clickbait-dataset.

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of CVPR'17*. 6541–6549.

[4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* (2013).

[5] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. 2020. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems* 33 (2020), 21865–21877.

[6] Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. A Comparative Study of Faithfulness Metrics for Model Interpretability Methods. In *Proceedings of ACL'22*. 5029–5038.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NeurIPS'16* (2016).

[8] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-attentive multi-task learning for explainable recommendation.. In *IJCAI*. 2137–2143.

[9] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 512–515.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *Transactions of the Association for Computational Linguistics* (2020).

[12] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.

[13] Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in nlp. In *International conference on machine learning*. PMLR, 2454–2463.

[14] JB Heaton, Nicholas G Polson, and Jan Hendrik Witte. 2016. Deep learning in finance. *arXiv preprint arXiv:1602.06561* (2016).

[15] Beta-vae Higgins. [n. d.]. Learning basic visual concepts with a constrained variational framework. In *Proceedings of ICLR'17*.

[16] Xu Huang, Defu Lian, Jin Chen, Liu Zheng, Xing Xie, and Enhong Chen. 2023. Cooperative Retriever and Ranker in Deep Recommenders. In *Proceedings of the ACM Web Conference 2023*. 1150–1161.

[17] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[18] Olivier Jeunen, Ciarán Gilligan-Lee, Rishabh Mehrotra, and Mounia Lalmas. 2022. Disentangling causal effects from sets of interventions in the presence of unobserved confounders. *Advances in Neural Information Processing Systems* 35 (2022), 27850–27861.

[19] Yahui Jiang, Meng Yang, Shuhao Wang, Xiangchun Li, and Yan Sun. 2020. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer communications* 40, 4 (2020), 154–166.

[20] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[21] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[22] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of KDD'16*. 1675–1684.

[23] Seungeon Lee, Xiting Wang, Sungwon Han, Xiaoyuan Yi, Xing Xie, and Meeyoung Cha. 2022. Self-explaining deep models with logic rule reasoning. In *Advances in Neural Information Processing Systems*.

[24] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of EMNLP'16*. 107–117.

[25] Zhen Li, Xiting Wang, Weikai Yang, Jing Wu, Zhengyan Zhang, Zhiyuan Liu, Maosong Sun, Hui Zhang, and Shixia Liu. 2022. A unified understanding of deep nlp models for text classification. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2022), 4980–4994.

[26] Defu Lian, Haoyu Wang, Zheng Liu, Jianxun Lian, Enhong Chen, and Xing Xie. 2020. Lightrec: A memory and search-efficient recommender system. In *Proceedings of WWW'20*. 695–705.

[27] Dohun Lim, Hyeonseok Lee, and Sungchan Kim. 2021. Building reliable explanations of unreliable neural networks: locally smoothing perspective of model interpretation. In *Proceedings of CVPR'21*. 6468–6477.

[28] Wanyu Lin, Hao Lan, Hao Wang, and Baochun Li. 2022. Orphicx: A causality-inspired latent variable model for interpreting graph neural networks. In *Proceedings of CVPR'22*. 13729–13738.

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[30] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[31] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality Inspired Representation Learning for Domain Generalization. In *Proceedings of CVPR'22*. 8046–8056.

[32] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL'11*. 142–150.

[33] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *Comput. Surveys* 55, 8 (2022), 1–42.

[34] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2493–2500.

[35] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.

[36] Brady Neal. 2020. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)* (2020).

[37] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, and Mark Davenport. 2020. Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems* 33 (2020), 5453–5467.

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

[39] Arash Rahnama and Andrew Tseng. 2021. An adversarial approach for explaining the predictions of deep neural networks. In *Proceedings of CVPR'21*. 3253–3262.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of KDD'16*. 1135–1144.

[41] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications* 11, 1 (2020), 3923.

[42] Patrick Schwab and Walter Karlen. 2019. Cxplain: Causal explanations for model interpretation under uncertainty. *NeurIPS'19* (2019).

[43] Sofia Serrano and Noah A Smith. 2019. Is Attention Interpretable?. In *Proceedings of ACL'19*. 2931–2951.

[44] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[45] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[46] Steven Sloman. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

[47] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[48] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[49] P.-N. Tan. 2018. *Introduction to data mining*. India.

[50] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, and Enhong Chen. 2023. Influence-Driven Data Poisoning for Robust Recommender Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[51] G Xu, TD Duong, Q Li, S Liu, and X Wang. 2020. Causality Learning: A New Perspective for Interpretable Machine Learning. *IEEE Intelligent Informatics Bulletin* (2020).

[52] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems* 32 (2019).

[53] Linan Yue, Qi Liu, Yichao Du, Yanqing An, Li Wang, and Enhong Chen. 2022. DARE: Disentanglement-Augmented Rationale Extraction. *Advances in Neural Information Processing Systems* 35 (2022), 26603–26617.

[54] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[55] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *Proceedings of CVPR'18*. 8827–8836.

[56] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *NeurIPS'15* (2015).
[57] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]

A SUPPLEMENTARY MATERIAL

The source code of CIMI is available at https://github.com/Daftstone/CIMI, and the full version of the paper (including the main text and all appendices) is available at https://github.com/Daftstone/CIMI/blob/master/paper.pdf.