Adversarial Infidelity Learning for Model Interpretation

Jian Liang[1*], Bing Bai[1*], Yuren Cao[1], Kun Bai[1], Fei Wang[2]
[1]Cloud and Smart Industries Group, Tencent, China
[2]Department of Population Health Sciences, Weill Cornell Medicine, USA
{joshualiang,icebai,laurenyrcao,kunbai}@tencent.com
few2001@med.cornell.edu

ABSTRACT

Model interpretation is essential in data mining and knowledge discovery. It can help understand the intrinsic model working mechanism and check if the model has undesired characteristics. A popular way of performing model interpretation is Instance-wise Feature Selection (IFS), which provides an importance score of each feature representing the data samples to explain how the model generates the specific output. In this paper, we propose a Model-agnostic Effective Efficient Direct (MEED) IFS framework for model interpretation, mitigating concerns about sanity, combinatorial shortcuts, model identifiability, and information transmission. Also, we focus on the following setting: using selected features to directly predict the output of the given model, which serves as a primary evaluation metric for model-interpretation methods. Apart from the features, we involve the output of the given model as an additional input to learn an explainer based on more accurate information. To learn the explainer, besides fidelity, we propose an Adversarial Infidelity Learning (AIL) mechanism to boost the explanation learning by screening relatively unimportant features. Through theoretical and experimental analysis, we show that our AIL mechanism can help learn the desired conditional distribution between selected features and targets. Moreover, we extend our framework by integrating efficient interpretation methods as proper priors to provide a warm start. Comprehensive empirical evaluation results are provided by quantitative metrics and human evaluation to demonstrate the effectiveness and superiority of our proposed method. Our code is publicly available online at https://github.com/langlrsw/MEED.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; **Instance-based learning**; *Neural networks*.

KEYWORDS

model interpretation, black-box explanations, infidelity, adversarial learning.

∗ Equal contributions from both authors.

KDD '20, August 23–27, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403071

ACM Reference Format:
Jian Liang, Bing Bai, Yuren Cao, Kun Bai, Fei Wang. 2020. Adversarial Infidelity Learning for Model Interpretation. In *26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3394486.3403071

1 INTRODUCTION

The interpretation of data-driven models explains their input-output relationship, which provides information about whether the models admit some undesired characteristics, and thus can guide people to use, debug, and improve machine learning models. The model interpretation has an increasing demand in many real-applications, including medicine [36], security [8], and criminal justice [22].

Existing research on model interpretation can be categorized into *model-specific methods* and *model-agnostic methods*. Model-specific methods take advantage of the knowledge of the model itself to assist explanations, such as gradient-based methods for neural networks, whereas model-agnostic methods can explain any black-box system. Instance-wise Feature Selection (IFS) is a well known model-agnostic interpretation method. It produces an importance score of each feature for representing a data sample [12], which indicates how much each feature dominates the model's output. For this kind of approach, desired properties for ideal explanations (feature importance scores) are as follows.

- Expressiveness: the number of features with relatively high scores should be small [27].
- Fidelity: the model output should primarily depend on high-score features [7, 9, 13, 15, 19, 27, 30, 39].
- Low sensitivity: feature scores should be robust against adversarial attacks [11, 14, 39, 40].
- Sanity: feature scores should be dependent of the model [3].

Recent research for IFS-based model explanation can be divided into *(local/global) feature attribution methods* [4, 39][1] and *direct model-interpretation (DMI) methods*. Local feature attribution methods provide some sensitivity scores of the model output concerning the changes of the features in the neighborhood. In contrast, global feature attribution methods directly produce the amount of change of the model output given changes of the features. Other than providing the change of the model output, DMI is a more straightforward approach to select features and use a model to approximate the output of the original black-box model [9, 35].

In this paper, we attempt to tackle the DMI problem. *When given a data sample and the model to be explained, what features does the model use primarily to generate the output?* A straightforward approach is to develop a feature attribution network (which we

[1]In this paper, the definitions of global and local explanations follow the description of Ancona *et al.* [4] and Yeh *et al.* [39], and distinct from that of Plumb *et al.* [26].

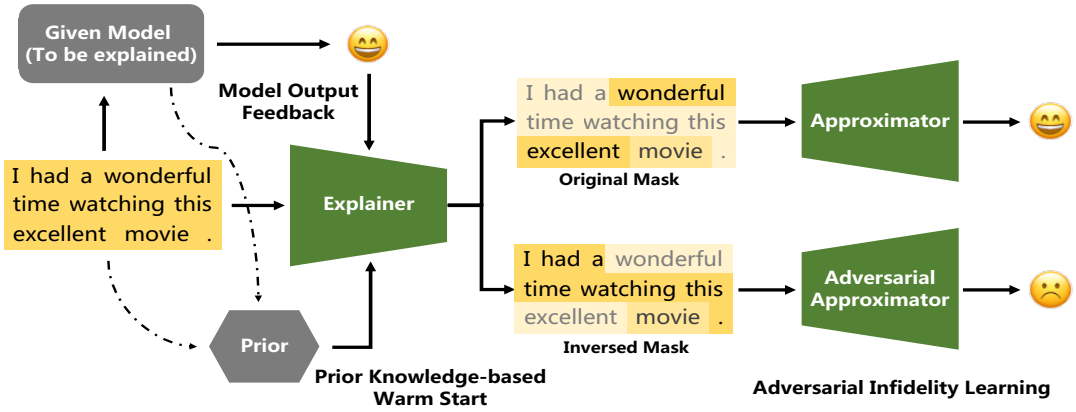

Figure 1: The architecture of our proposed framework. Taking a sentence as an example, we train an explainer to select important words and an approximator to predict the output of the original model. The model output is also used as an input for the explainer. The AIL module trains the explainer to render the adversarial approximator cannot predict the model output well based on the unselected words. As an extension, efficient interpretation methods, *e.g.*, gradient-based methods, can be integrated to provide a warm start. Best view in color.

refer to as the explainer) to produce a soft/hard mask to highlight essential features, and a prediction network (which we refer to as the approximator) to approximate the output of the original model [9]. However, this straightforward approach may cause the effectiveness and efficiency related concerns in the following.

- **Sanity problem** [3]: a mask may be irrelevant to the original model, but only relate to the features of a specific sample. As a consequence, the selected features of the trained explainer may be different with those truly used by the original model, which is not expected in interpreting the model.
- **Combinatorial shortcuts problem**: the entries of the mask may not select good features, but rather act as additional features themselves for better approximation performances [17, 37], because it is a function of all the input features. For example, the explainer could choose to mask out the first half of the features for positive samples, and the second half of the features for negative samples. The approximator can utilize this pattern to predict the target while completely ignore whether good features are selected.
- **Model identifiability problem**: similar approximation performances can be achieved by different groups of feature. It is difficult to decide which group is the best.
- **Information transmission problem** [25]: it is difficult to transmit effective supervised information to the explainer, because the mask is unsupervised.

To address these issues, we propose a Model-agnostic Effective Efficient Direct (MEED) model-interpretation method for the DMI problem. The overall architecture of our proposed framework is presented in Figure 1. The major components include model output feedback, adversarial infidelity learning, and prior knowledge-based warm start, which we describe as follows.

Firstly, we propose to enrich the input information of the explainer to boost the effectiveness and the efficiency of the feature selection process. Existing research treats raw features only as the input to the neural network-based explainer [7, 9, 30]. The absence of the original model to include for the explainer's input may render the mask out of the explainer uncorrelated with the original model and then cause the sanity problem. Nonetheless, it is not trivial to input a whole model into the neural network-based explainer. Therefore, we propose to incorporate the model output as another input signal. Apart from the sanity problem, the model output can provide rich information for the explainers to select essential features, and make the learning process more precise, especially in applications like regression or representation learning. In other words, the information transmission problem can also be mitigated.

Secondly, we propose to exploit the unselected features for mitigating the combinatorial shortcuts and model identifiability problems. Inspired by Hooker *et al.* [15], we attempt to achieve an auxiliary goal that *the unselected features should contain the least useful information*. To achieve this, we propose an Adversarial Infidelity Learning (AIL) mechanism. Specifically, we develop another approximator that learns to approximate the original model output using the unselected features. Then our explainer learns to select features to minimize such approximation accuracy. The learning processes run alternately. Intuitively, the convergence of such an adversarial learning process will render the masks uncorrelated with the model output, and then can mitigate the combinatorial shortcuts problem. On the other hand, this learning process exploits the unselected features, which are often (at least relatively) ignored, to introduce additional supervised information for a certain group of selected features, and then can improve model identifiability. These properties are demonstrated by our theoretical analysis and experimental results.

Finally, we extend our framework to further mitigate the information transmission problem by integrating prior knowledge. Specifically, we integrate explanations provided by efficient interpretation methods as priors to provide a warm start. The constraints of the priors fade out when the number of training epochs grows to learn a more powerful explainer by the end-to-end framework.

We follow Chen [9] to perform a predictive evaluation to see whether the selected features contribute to sufficient approximate accuracy. Comprehensive empirical evaluation results on four real-world benchmark datasets are provided with quantitative evaluation metrics and human-evaluations to demonstrate the effectiveness and superiority of our proposed method. Moreover, we validate our method on a real-world application: teenager/adult classification based on mobile sensor data from 5 million of Tencent users who play the popular *Honor of Kings, a.k.a. Arena of Valor* game.

2 RELATED WORKS

Model interpretation methods based on IFS can be categorized into local/global methods as introduced in the introduction. Local methods includes 1) gradient-based methods, such as Gradient (Grad) [32] and Guided Back-Propagation [34], 2) sampling-based methods, *i.e.*, perform sensitivity analysis by sampling points around the given data sample, such as LIME [27], kernel SHAP [23] and CXPlain [30], and 3) hybrid methods, such as SmoothGrad [33], Squared SmoothGrad [33], VarGrad [2], and INFD [39]. On the other hand, global methods include Gradient × Input [31], Integrated Gradients [35], DeepLIFT [31] and LRP [6], among others. These methods do not directly tackle the DMI problem.

For the DMI problem, being inherently interpretable, tree- [29] and rule-based [5] models have been proposed to approximate the output of a complex black-box model with all features. The models themselves provide explanations, including feature importance. However, they may lack the ability for accurate approximations when the original given model is complex. Recently, L2X [9] and VIBI [7] have been proposed as variational methods to learn a neural network-based approximator based on the selected features. The unselected features are masked out by imputing zeros. VIBI improves L2X to encourage the briefness of the learned explanation by adding a constraint for the feature scores to a global prior (in contrast, our priors are conditioned on each sample). Since L2X and VIBI only input features to their explainers, and they directly select features to approximate the model out, therefore, they both may suffer from the sanity, combinatorial shortcuts, model identifiability, model identifiability, and information transmission problems.

In contrast, our method tackle these problems for DMI by leveraging more comprehensive information from the model output, the proposed adversarial infidelity learning mechanism, and the proposed prior-knowledge integration. We note that Zhu *et al.* [41] proposed an adversarial attention network. However, their objective is to eliminate the difference in extracted features for different learning tasks, which is different from ours.

3 METHODOLOGY

In this section, we present the detailed methodology of our method. First, we define the notations and problem settings of our study.

Consider a dataset $\mathcal{D} = \{\mathbf{x}^i\}_{i=1}^n$ consisting of n independent samples. For the ith sample, $\mathbf{x}^i \in \mathcal{X} \subset \mathbb{R}^d$ is a feature vector with d dimensions, $\mathbf{y}^i = M(\mathbf{x}^i) \in \mathcal{Y} \subset \mathbb{R}^c$ is the output vector of a given data-driven model $M \in \mathcal{M}$ (note that y^i may be different from the true label of the sample). The conditional output distribution $p(\mathbf{y} \mid \mathbf{x})$ is determined by the given model. For classification tasks, c is the number of classes.

We do not assume the true label of each feature vector is available for training or inference. We develop a neural network-based IFS explainer E, which outputs a feature-importance-score vector $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^d$ for a data sample $\mathbf{x}$ and the model M. As discussed by Yeh *et al.* [39], the explainer should be a mapping that $E : \mathcal{X} \times \mathcal{M} \to \mathcal{Z}$. Since it is not trivial to treat an arbitrary model in $\mathcal{M}$ as an input to a neural network, we compromise by involving the model output as an alternative such that $E : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. We select top-$k$ features according to $\boldsymbol{z}$, where k is a user-defined parameter. The indices of k selected features are denoted by $S \subset \{1, \ldots, d\}$. For a feature vector $\mathbf{x}$, the selected features are denoted by $\mathbf{x}_S$, whereas the unselected features are denoted by $\mathbf{x}_{\bar{S}}$. Throughout the paper, we denote $[k']$ as the index set $\{1, 2, \ldots, k'\}$ for some integer k'.

The goal of our learning approach is to train a neural network-based explainer E over the dataset $\mathcal{D}$ and then generalize it to a testing set to see whether the selected features contribute to sufficient approximate accuracy. The quantitative evaluations of the explainer are described in Section 4.0.2.

3.1 Our Framework

The architecture of our framework is illustrated in Fig. 1. We explain a given model by providing IFS for each specific data sample. The IFS is embodied as a feature attribution mask provided by a learned explainer with the features and the model output of the data sample as inputs. We train an approximator to use selected/masked features to approximate the model output. We also train an adversarial approximator to use unselected features to approximate the model output, and then train the explainer to select features to undermine the approximation, which is referred to as the AIL mechanism. As an extention, integrating efficient model-interpretation methods is also introduced to provide a warm start.

3.2 Adversarial Infidelity Learning

As discussed in the introduction, a straightforward approach to optimize the selection indices S is directly maximizing the mutual information $I(\mathbf{x}_S; \mathbf{y})$ between selected features $\mathbf{x}_S$, and the model output $\mathbf{y}$ [7, 9]. To tackle the combinatorial shortcuts and model identifiability problems, we propose an auxiliary objective: minimizing the mutual information $I(\mathbf{x}_{\bar{S}}; \mathbf{y})$ between unselected features $\mathbf{x}_{\bar{S}}$ and the model output $\mathbf{y}$. Because compared with the selected features, the unselected features should contain less useful information. Therefore, the basic optimization problem for S is:

$$\max_S \; I(\mathbf{x}_S; \mathbf{y}) - I(\mathbf{x}_{\bar{S}}; \mathbf{y}) \;\; \text{s.t. } S \sim E(\mathbf{x}, \mathbf{y}). \tag{1}$$

We can be guided by the Theorem 1 to optimize the explainer.

Theorem 1. *Define*

$$S^* = \operatorname*{argmax}_S \mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_S) - \log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}})], \tag{2}$$

where the expectation is over $p(\mathbf{y} \mid \mathbf{x})$. Then S^* is a global optimum of Problem (1)*. Conversely, any global optimum of Problem* (1) *degenerates to S^* almost surely over $p(\mathbf{x}, \mathbf{y})$.*

The proof is deferred to Appendix A.1. Problem (1) and Theorem 1 show that the auxiliary objective exploits $\mathbf{x}_{\bar{S}}$ to involve additional supervised information, and then improves model identifiability.

According to Theorem 1, we deveop an approximator $A_s : \mathcal{X} \to \mathcal{Y}$ to learn the conditional distribution $p(\mathbf{y} \mid \mathbf{x}_S)$. We achieve this by optimizing a variational mapping: $\mathbf{x}_S \to q_s(\mathbf{y} \mid \mathbf{x}_S)$ to let $q_s(\mathbf{y} \mid \mathbf{x}_S)$ approximate $p(\mathbf{y} \mid \mathbf{x}_S)$. We define $q_s(\mathbf{y} \mid \mathbf{x}_S) \propto \exp(-\ell_s(\mathbf{y}, A_s(\tilde{\mathbf{x}}_S)))$, where ℓ_s denotes the loss function corresponding to the conditional distribution $p(\mathbf{y} \mid \mathbf{x}_S)$ (*e.g.,* mean square error for Gaussian distribution, and categorical cross entropy for categorical distribution), and $\tilde{\mathbf{x}}_S \in \mathcal{X}$ which is defined as: $(\tilde{\mathbf{x}}_S)_j = x_j$ if $j \in S$ and $(\tilde{\mathbf{x}}_S)_j = 0$ otherwise. We let $q_m(\mathbf{y} \mid \tilde{\mathbf{x}}_S)$ denote the output distribution of $M(\tilde{\mathbf{x}}_S)$. We approximate $\mathbf{y}$ by A_s instead of M, because as discussed by Hooker *et al.* [15], $p(\mathbf{y} \mid \mathbf{x}_S) \neq q_m(\mathbf{y} \mid \tilde{\mathbf{x}}_S)$, then $A_s(\tilde{\mathbf{x}}_S)$ may approximate more accurate than $M(\tilde{\mathbf{x}}_S)$ does.

Similarly, we develop another approximator $A_u : \mathcal{X} \to \mathcal{Y}$ to learn $q_u(\mathbf{y} \mid \mathbf{x}_{\bar{S}}) \propto \exp(-\ell_u(\mathbf{y}, A_u(\tilde{\mathbf{x}}_{\bar{S}})))$ to approximate $p(\mathbf{y} \mid \mathbf{x}_{\bar{S}})$. Then we can show that Problem (1) can be relaxed by maximizing variational lower bounds and alternately optimizing:

$$\max_{A_s, A_u} \ \mathbb{E}[\log q_s(\mathbf{y} \mid \mathbf{x}_S) + \log q_u(\mathbf{y} \mid \mathbf{x}_{\bar{S}})] \ \text{ s.t. } S \sim E(\mathbf{x}, \mathbf{y}), \tag{3}$$

$$\max_{E} \ \mathbb{E}[\log q_s(\mathbf{y} \mid \mathbf{x}_S) - \log q_u(\mathbf{y} \mid \mathbf{x}_{\bar{S}})] \ \text{ s.t. } S \sim E(\mathbf{x}, \mathbf{y}). \tag{4}$$

First, Problem (3) is optimized to learn $q_s(\mathbf{y} \mid \mathbf{x}_S)$ and $q_u(\mathbf{y} \mid \mathbf{x}_{\bar{S}})$ to approximate $p(\mathbf{y} \mid \mathbf{x}_S)$ and $p(\mathbf{y} \mid \mathbf{x}_{\bar{S}})$, respectively. Then Problem (4) is optimized to learn the explainer E to find good explanations according to Theorem 1. Since 1) $q_u(\mathbf{y} \mid \mathbf{x}_{\bar{S}})$ is maximized by optimizing A_u and then minimized by optimizing E, which is an *adversarial learning* process, and 2) minimizing $q_u(\mathbf{y} \mid \mathbf{x}_{\bar{S}})$ represents *infidelity, i.e.,* undermining performance to approximate $\mathbf{y}$ (by excluding selected features), the alternate optimization process can be regarded as an adversarial infidelity learning mechanism.

Since optimizing Problems (3) and (4) for all possible S requires $\binom{d}{k}$ times of computation for the objectives, we follow L2X [9] to apply the Gumbel-softmax trick to approximately sample a k-hot vector. Specifically, let $\boldsymbol{z} = E(\mathbf{x}, \mathbf{y})$ for a pair of inputs $(\mathbf{x}, \mathbf{y})$, where for $j \in [d]$, $z_j \geq 0$ and $\sum_j z_j = 1$. Then we define the sampled vector $\mathbf{v} \in \mathcal{V} \subset \mathbb{R}^d$, where for a predefined $\tau > 0$,

$$\begin{aligned} v_j &= \max_{l \in [k]} \frac{\exp((\log z_j + \xi_j^l)/\tau)}{\sum_{j'=1}^d \exp((\log z_{j'} + \xi_{j'}^l)/\tau)}, \ j \in [d], \\ \xi_j^l &= -\log(-\log u_j^l), \ u_j^l \sim \text{Uniform(0,1)}, \ j \in [d], l \in [k]. \end{aligned} \tag{5}$$

Denote the above random mapping by $G : \mathcal{Z} \to \mathcal{V}$, approximate $\tilde{\mathbf{x}}_S \approx \mathbf{x} \odot G(E(\mathbf{x}, \mathbf{y}))$ and $\tilde{\mathbf{x}}_{\bar{S}} \approx \mathbf{x} \odot (\mathbf{1}^d - G(E(\mathbf{x}, \mathbf{y})))$, where $\mathbf{1}^d \in \mathbb{R}^d$ with all elements being 1, and $\odot$ denotes element-wise product. Define the losses for selected and unselected features, respectively,

$$\begin{aligned} \mathcal{L}_s &= \frac{1}{n} \sum_{i=1}^n \ell_s(\mathbf{y}^i, A_s(\mathbf{x}^i \odot G(E(\mathbf{x}^i, \mathbf{y}^i)))) \\ \mathcal{L}_u &= \frac{1}{n} \sum_{i=1}^n \ell_u(\mathbf{y}^i, A_u(\mathbf{x}^i \odot (\mathbf{1}^d - G(E(\mathbf{x}^i, \mathbf{y}^i))))). \end{aligned} \tag{6}$$

Then we can relax Problems (3) and (4) as

$$\min_{A_s, A_u} \ \mathcal{L}_s + \mathcal{L}_u, \tag{7}$$

$$\min_{E} \ \mathcal{L}_s - \mathcal{L}_u. \tag{8}$$

For inference, one can select the top-k features of $E(\mathbf{x}^t, \mathbf{y}^t)$ for a testing sample $\mathbf{x}^t$, where $\mathbf{y}^t = M(\mathbf{x}^t)$.

3.3 Theoretical Analysis

In this section, we first derive variational lower bounds to show the connection between the goal in Eq. (1) and the realization in Eq. (3) and (4). The derivation also shows that the approximator is possible to be superior to the given model to predict the model output using masked features. We also show that our AIL mechanism can mitigate the combinatorial shortcuts problem.

The variational lower bounds are as follows and derived in Appendix A.2. First, for selected features, we have for any $q_s(\mathbf{y} \mid \mathbf{x}_S)$:

$$\begin{aligned} &I(\mathbf{x}_S; \mathbf{y}) = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}|\mathbf{x}} \mathbb{E}_{S|\mathbf{x},\mathbf{y}} \mathbb{E}_{\mathbf{y}|\mathbf{x}_S}[\log p(\mathbf{y} \mid \mathbf{x}_S)] + \text{Const}, \\ &\mathbb{E}_{\mathbf{y}|\mathbf{x}_S}[\log p(\mathbf{y} \mid \mathbf{x}_S)] \geq \mathbb{E}_{\mathbf{y}|\mathbf{x}_S}[\log q_s(\mathbf{y} \mid \mathbf{x}_S)], \end{aligned} \tag{9}$$

where the equality holds if and only if $q_s(\mathbf{y} \mid \mathbf{x}_S) = p(\mathbf{y} \mid \mathbf{x}_S)$. Therefore, if $M(\tilde{\mathbf{x}}_S)$'s output distribution $q_m(\mathbf{y} \mid \tilde{\mathbf{x}}_S) \neq p(\mathbf{y} \mid \mathbf{x}_S)$, it is possible that $\mathbb{E}_{\mathbf{y}|\mathbf{x}_S}[\log q_s(\mathbf{y} \mid \mathbf{x}_S)] > \mathbb{E}_{\mathbf{y}|\mathbf{x}_S}[\log q_m(\mathbf{y} \mid \tilde{\mathbf{x}}_S)]$, *i.e.,* $A_s(\tilde{\mathbf{x}}_S)$ can be more accurate than $M(\tilde{\mathbf{x}}_S)$ to estimate $\mathbf{y}$.

Similarly, for unselected features, we have for any $q_u(\mathbf{y} \mid \mathbf{x}_{\bar{S}})$:

$$\begin{aligned} &I(\mathbf{x}_{\bar{S}}; \mathbf{y}) = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}|\mathbf{x}} \mathbb{E}_{S|\mathbf{x},\mathbf{y}} \mathbb{E}_{\mathbf{y}|\mathbf{x}_{\bar{S}}}[\log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}})] + \text{Const}. \\ &\mathbb{E}_{\mathbf{y}|\mathbf{x}_{\bar{S}}}[\log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}})] \geq \mathbb{E}_{\mathbf{y}|\mathbf{x}_{\bar{S}}}[\log q_u(\mathbf{y} \mid \mathbf{x}_{\bar{S}})]. \end{aligned} \tag{10}$$

On the other hand, since A_s actually receives the selected features $\mathbf{x}_S$ and the feature-attribution mask $\mathbf{v}$ as inputs, where $\mathbf{v} = G(E(\mathbf{x}, \mathbf{y}))$, what A_s actually learns is the conditional distribution $p(\mathbf{y} \mid \mathbf{x}_S, \mathbf{v})$. Through the straightforward learning mentioned in the introduction, *i.e.,* removing $\mathcal{L}_u$ in Eq. (7) and (8), it could cause the combinatorial shortcuts problem for A_s to learn $p(\mathbf{y} \mid \mathbf{v})$ only, resulting in the feature selection process $\mathbf{x}_S$ meaningless. Fortunately, Theorem 2 shows that our AIL can help to avoid this problem by encouraging the independence between $\mathbf{v}$ and $\mathbf{y}$, then it will be hard for A_s to approximate $\mathbf{y}$ solely by $\mathbf{v}$. Thus, A_s will have to select useful features from $\mathbf{x}$. The proof can be found in Appendix A.3.

Theorem 2. *For the optimized problem in Eq.* (7) *and* (8)*, the independence between* $\mathbf{v}$ *and* $\mathbf{y}$ *is encouraged.*

3.4 Extension Considering Prior Knowledge

As described in Section 3.2, the feature attribution layer (the output layer of the explainer) is in the middle of networks. Since it is optimized by an end-to-end learning process, the information transmission is inefficient. Therefore, we propose to involve efficient interpretations as good priors of feature attribution.

Let $\mathbf{r} \in \mathbb{R}^d$ be a feature-importance-score vector generated by another interpretation method H for a sample $\mathbf{x}$ and the model M. Assume for $j \in [d]$, $r_j \geq 0$ and $\sum_j r_j = 1$, which can be easily achieved through a softmax operation.

Given $\boldsymbol{z} = E(\mathbf{x}, M(\mathbf{x}))$ for a sample $\mathbf{x}$, we can regard z_j as $p(\delta = j \mid \mathbf{x}, M, E)$ for $j \in [d]$, where $\delta \in [d]$ denotes whether the jth feature should be selected. Similarly, we can regard r_j as $p(\delta = j \mid \mathbf{x}, M, H)$. Then assuming conditional independence between the interpretation models E and H given δ, $\mathbf{x}$ and M, we can obtain

$$p(\delta = j \mid \mathbf{x}, M, E, H) = \frac{p(\delta = j \mid \mathbf{x}, M, E) p(\delta = j \mid \mathbf{x}, M, H)}{\sum_{j'=1}^d p(\delta = j' \mid \mathbf{x}, M, E) p(\delta = j' \mid \mathbf{x}, M, H)}. \tag{11}$$

The derivation details can be found in Appendix A.4.

Nonetheless, as we expect that the end-to-end learning process can generate better explanations, the prior explanation should decrease its influence when the number of epochs $m \in \mathbb{Z}_+$ increases. Therefore, we define

$$\tilde{z} := \frac{[p(\delta = j \mid \mathbf{x}, M, E)^m p(\delta = j \mid \mathbf{x}, M, H)]^{1/(m+1)}}{\sum_{j'=1}^{d} [p(\delta = j' \mid \mathbf{x}, M, E)^m p(\delta = j' \mid \mathbf{x}, M, H)]^{1/(m+1)}}. \quad (12)$$

For Eq. (5), we replace z with the re-estimated $\tilde{z}$ defined in Eq. (12).

In addition, we add a constraint for the explanations, learning $z = E(\mathbf{x}, M(\mathbf{x}))$ to be close to $\tilde{z}$ for a loss $\ell_e(\cdot, \cdot)$:

$$\mathcal{L}_e = \frac{1}{n} \sum_{i=1}^{n} \frac{\ell_e(\tilde{z}^i, z^i)}{m+1}. \quad (13)$$

The constraint will fade out when the number of epochs m becomes large, and thus only contributes for a warm start.

4 EXPERIMENTS

We conduct comprehensive evaluations on five datasets:

- IMDB sentiment analysis dataset [24],
- MNIST dataset [20] to classify `3` and `8`,
- Fashion-MNIST dataset [38] to classify `Pullover` and `Coat`,
- ImageNet dataset [10] to classify `Gorilla` and `Zebra`,
- our established mobile sensor dataset from Tencent *Honor of Kings* game for teenager recognition, which we refer to as Tencent Gaming Dataset (TGD) in this paper.

The detailed re-organization process for each data will be introduced in the following sections.

4.0.1 Methods for Comparison. We compare our method (Ours) with the state-of-the-art model-agnostic baselines: LIME [27], kernel SHAP (SHAP) [23], CXPlain (CXP) [30], INFD [39], L2X [9] and VIBI [7]. We also compare model-specific baselines: Gradient (Grad) [32] and Gradient $\times$ Input (GI) [31].

4.0.2 Evaluation Metrics. We follow Chen [9] to perform a predictive evaluation for the fidelity of both the selected and the unselected features. For the Fidelity of the Selected features, given an explanation, *e.g.,* selected features $\mathbf{x}_S$, from an arbitrary IFS interpretation method, we evaluate whether the given model M truly use the selected features primarily to generate the very output $\mathbf{y} = M(\mathbf{x})$. To answer this, we need to approximate $\mathbf{y}$ based on selected features $\mathbf{x}_S$. Thus, we evaluate by the consistency between $\mathbf{y}$ and $M(\tilde{\mathbf{x}}_S)$ (recall that $\tilde{\mathbf{x}}_S$ is $\mathbf{x}$ with unselected features imputed by zeros), denoted by FS-M(%). However, M is trained on all features of $\mathbf{x}$, not on $\tilde{\mathbf{x}}_S$ [15]. Therefore, we additionally propose to evaluate the consistency between $\mathbf{y}$ and $A'_S(\tilde{\mathbf{x}}_S)$ as a reference, denoted by FS-A(%), where A'_S is an trained approximator on $\mathcal{D}$ to learn the mapping $\tilde{\mathbf{x}}_S \to \mathbf{y}$. High FS-M or FS-A result suggests high importance of the selected features. Similarly, for the Fidelity of the Unselected features, we evaluate the consistency between $\mathbf{y}$ and $M(\tilde{\mathbf{x}}_{\bar{S}})$, denoted by FU-M(%); and the consistency between $\mathbf{y}$ and $A'_u(\tilde{\mathbf{x}}_{\bar{S}})$, denoted by FU-A(%), where A'_u is an trained approximator on $\mathcal{D}$ to learn the mapping $\tilde{\mathbf{x}}_{\bar{S}} \to \mathbf{y}$. Low FU-M or FU-A result suggests high importance of the selected features. Note that low FS-A or high FU-A is possible because the number of selected features are usually small. Nonetheless, simultaneous results of high FS-A and low FU-A suggest good selected features.

For human evaluation, we also follow Chen [9] to evaluate Fidelities of Selected features denoted by FS-H(%), *i.e.,* whether the predictions made by a human using selected features are consistent with those made by the given model using all the features. We adopt this metric to evaluate whether human can understand how the given model makes decisions. Note that sometimes human may not understand or be satisfied with features selected by the given model. After all, we are explaining the given model, not human.

For the evaluation metric for the fidelities, we report top-1 accuracy (ACC@1), since the five tasks are all binary classification. Specifically, the model outputs are transformed to categorical variables to compute accuracy. We adopt binary masks to select features, *i.e.,* top k values of $z = E(\mathbf{x}, \mathbf{y})$ are set to 1, others are set to 0, and then we treat $\mathbf{x} \odot z$ as the selected features. On the other hand, we evaluate the influence of adversarial examples on the feature importance scores by the sensitivity score, SEN(%), proposed by Yeh *et al.* [39]. We also report the average explanation Time (by second) Per Sample (TPS) on a single NVidia Tesla m40 GPU.

4.0.3 Implementation Details. In Eq. (6), ℓ_s adopts cross-entropy. ℓ_u adopts cross-entropy for IMDB, MNIST, and TGD, and adopts Wasserstein distance [21] for Fashion-MNIST and ImageNet. The weights for $\mathcal{L}_u$ in Eq. (7) and (8) are 1e-3 for MNIST and 1 for all other datasets. We adopt the GI method to provide the prior explanations. ℓ_e in Eq. (13) is mean absolute error with the weight to be 1e-3 for ImageNet and 0 for others. $\tau = 0.5$ for all the datasets. We constrain the model-capacity of each method to be the same to acquire fair comparisons. For each dataset, we split half test data for validation. For each method on each dataset, we repeat 20 times and report the averaged results. Our implementation uses Keras with Tensorflow [1] backends. We list all other details in Appendix B.

4.1 IMDB

The IMDB [24] is a sentiment analysis dataset which consists of 50,000 movie reviews labeled by positive/negative sentiments. Half reviews are for training, and the other half is for testing. We split half of the testing reviews for validation. For the given model, we follow Chen *et al.* [9] to train a 1D convolutional neural network (CNN) for binary sentiment classification, and achieve the test accuracy of 88.60%. We develop our approximator with the same architecture as the given model. And we develop our explainer with the 1D CNN used by L2X [9] with a global and a local component. For a fair comparison, each method selects top-10 important words as an explanation for a review.

As shown in Table 1, our method significantly outperforms state-of-the-art baseline methods. Especially, our FS-M score shows nearly optimal fidelity, which is objectively validated by the original given model. Given that our FU-A score is similar to those of baselines, which shows that our selected features are indeed important, we demonstrate that the effectiveness and superiority of our method are significant. We present some examples of selected words of our method and the state-of-the-art baseline VIBI in Fig. 2. As shown in Fig. 2, our method not only selects more accurate keywords, but also provides more interpretable word combinations, such as “I enjoyed it”, “fabulous dancer”, “extremely bad movie”, and “excellent thriller”. Even though “I”, “it”, “dancer”, “movie”, and “thriller” are not apparent whether they are positive or negative

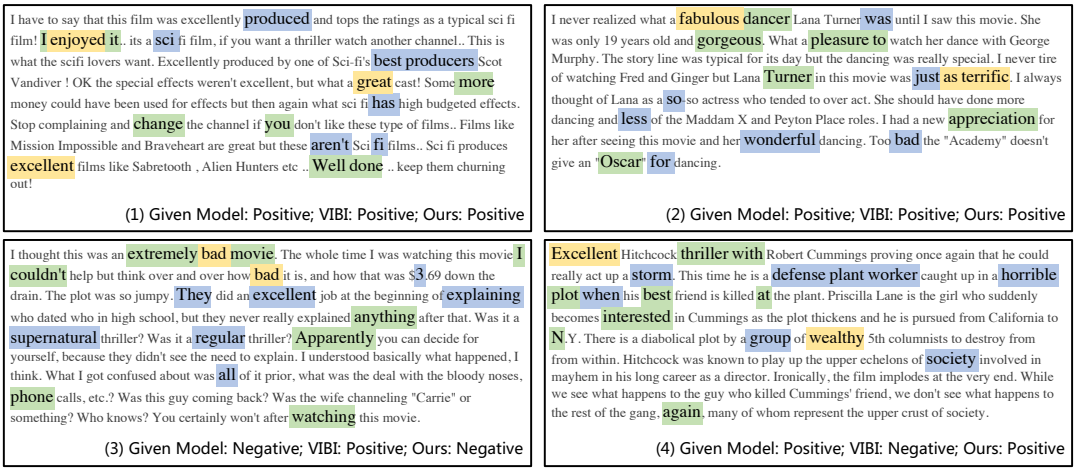

Figure 2: Examples of explanations on the IMDB dataset. The labels predicted by the original given model using all the words, the words selected by the state-of-the-art baseline VIBI, and the words selected by our method are shown, respectively, at the bottom of each panel. Keywords picked by VIBI, our method, and both methods are highlighted in blue, green, and yellow, respectively. In (1) and (2), the given model output consistent predictions using the selected words by both VIBI and ours, whereas (3) and (4), the prediction using full words is inconsistent with that using VIBI's selected words but is still consistent with that using our selected words. Best view in colors.

Table 1: Results on the IMDB dataset. † denotes the method uses additional information.

Method	FS-M	FU-M	FS-A	FU-A	FS-H	TPS
Grad	85.58	87.58	86.18	85.79	73.58	5e-5
GI	87.31	86.25	87.88	83.86	78.23	5e-5
LIME	89.75	82.13	88.53	82.96	83.98	3e-2
SHAP	50.17	99.16	50.24	99.50	53.22	4e-2
†CXP	90.60	80.01	90.70	83.04	84.97	1e-4
INFD	40.50	99.80	64.50	96.70	46.27	3e-0
L2X	89.23	82.90	89.05	83.81	83.49	1e-4
VIBI	90.79	80.36	90.12	82.57	84.33	1e-4
Ours	98.48	59.05	98.70	81.83	92.98	1e-4

words. Especially in Fig. 2 (2), our method picks the word "Oscar", which is not explicit positive, but its underlying logic suggests positive sentiment. These inspiring examples support the significant superiority of our method.

4.1.1 Human Evaluation. We also evaluate with the help of humans to quantify how interpretable are those selected words. We randomly select 500 reviews in the testing set for this human evaluation. We invite 20 Tencent employees to infer the sentiment of a review given only the selected words. The explanations provided by different interpretation methods are randomly mixed before sent to these employees. The final prediction of each review is averaged over multiple human annotations. For the explanations that are difficult to infer the sentiments, we ask the employees to provide random guesses. Finally, as shown by the FS-H scores in Table 1, our method significantly outperforms baseline methods as well.

Table 2: Results of the ablation study.

Method	FS-M	FU-M	FS-A	FU-A
Ours	98.48	59.05	98.70	81.83
w/o Output	92.47	54.28	91.99	79.82
w/o AIL	78.31	97.62	99.33	94.42
w/o Prior	98.38	60.22	98.97	81.07

4.1.2 Ablation Study. We evaluate three variants of our method by ablating our three components, *i.e.,* the model output feedback (Output), AIL, and prior knowledge-based warm start (Prior). In Table 2, we show the effectiveness of both the model output feedback and AIL. It is worthy of mentioning that, although the warm start strategy does not improve the final scores, it boosts the convergence rate at the start of optimization, as shown in Fig. 3.

4.1.3 Sanity Check. We perform the model and data randomization tests suggested by Adebayo *et al.* [3]. We evaluate the sanity by the cosine correlation between binary masks, *i.e.,* the original mask, and the other one resulted from randomization. The sanity scores for model and data randomization tests are 9.39% and 10.25%, respectively, which shows that our explanations are dependent on models and then are valid.

4.2 MNIST

We select the data of 3 and 8 of the MNIST dataset [20] for binary classification with 11, 982 training and 1984 testing images. We train a 2D CNN with two convolutional layers for the classification and achieve the test accuracy of 99.89%. We develop our approximators are by the same architecture as the given model, whereas we develop

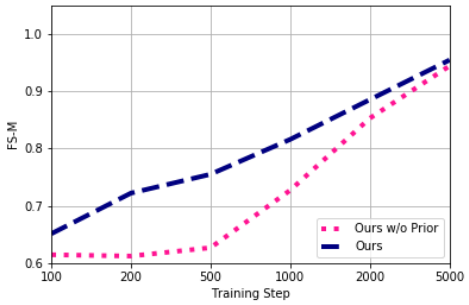

Figure 3: The FS-M scores of our method with and without the prior knowledge-based warm start.

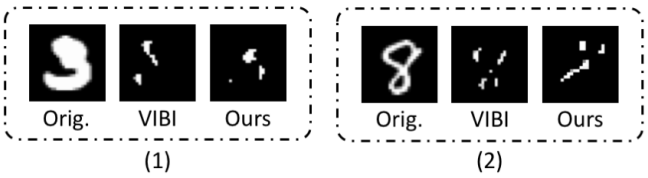

Figure 4: Examples of explanations on the MNIST dataset. In each panel, an original image and the masked images by VIBI and our method are presented from left to right. The labels inferred by the original model using the full images are 3 and 8 for panels (1) and (2), respectively.

Table 3: Results on the MNIST dataset.

Method	FS-M	FU-M	FS-A	FU-A	SEN	FS-H	TPS
Grad	98.19	68.75	99.55	99.55	139	94.37	5e-5
GI	99.45	67.64	99.55	99.24	120	94.58	5e-5
LIME	80.37	99.75	82.46	99.80	62.6	70.29	3e-2
SHAP	92.74	90.83	98.87	99.75	87.2	87.75	4e-2
†CXP	99.40	64.77	99.70	99.24	91.8	94.57	1e-4
INFD	89.62	96.62	99.95	99.95	101	84.33	1e-0
L2X	91.38	91.18	98.54	99.65	6.90	86.58	1e-4
VIBI	98.29	86.29	99.29	99.65	6.35	92.17	1e-4
Ours	99.04	74.70	99.80	99.70	6.11	94.46	1e-4

our explainer by a 2D CNN with three convolutional layers only. Each method selects top-25 pixels as an explanation for an image.

As shown in Table 3, our method still outperforms state-of-the-art model-agnostic baseline methods except for the CXPlain method, which uses the additional true label for each sample and is highly sensitive to adversarial examples. The Grad and GI methods are model-specific and not robust when facing challenging data, *e.g.*, see Tables 1, 4, and 5. Compared with the next-best model-agnostic baseline VIBI, the strength of our method determined by the FU-M score is to select the necessary features of a sample for recognition. We show some examples of selected pixels of our method and VIBI in Fig. 4. As shown in Fig. 4 (1), the VIBI masked image is closer to 8 than 3, whereas our masked image is more similar to 3. In

Figure 5: Examples of explanations on the Fashion-MNIST dataset. The first, second and third lines list the original images, the images masked by VIBI, the images masked by our method, respectively.

Table 4: Results on the Fashion-MNIST dataset.

Method	FS-M	FU-M	FS-A	FU-A	SEN	TPS
Grad	58.60	76.20	93.70	95.45	2528	5e-5
GI	62.15	66.25	93.35	94.45	2662	5e-5
LIME	75.63	94.30	73.60	97.35	61.03	3e-2
SHAP	63.29	55.78	93.97	95.58	84.59	4e-2
†CXP	59.65	16.50	94.85	95.10	107	1e-4
INFD	100.0	45.80	100.0	100.0	87.37	5e-1
L2X	77.30	87.30	89.85	96.00	1.76	1e-4
VIBI	84.10	70.85	91.90	94.40	17.36	1e-4
Ours	97.80	66.65	99.80	99.65	0.70	1e-4

Fig. 4 (2), though the VIBI masked image is similar to 8, but it is also close to a 3. In contrast, our masked image can never be 3. Since we are interpreting the recognition logic of model rather than that of humans, it is important to select features in favor of the machine logic, *e.g.*, considering both possibility and impossibility.

4.2.1 Human Evaluation. We randomly select 500 images in the testing set for this human evaluation. We invite 15 Tencent employees who are experts in machine learning to perform the same binary classification given only the masked images. Specifically, we ask the subjects to provide two scores in the range of $[0, 1]$ for each image. Each score is for the possibility of each class (3 or 8). We perform ℓ_1 normalization for the scores as the final prediction probabilities. Other settings and procedures are similar to Section 4.1.1. Finally, as shown by the FS-H scores in Table 3, our method significantly outperforms baseline methods as well.

4.3 Fashion-MNIST

The Fashion-MNIST dataset [38] is a dataset of Zalando's article images, which consists of 28×28 images of 10 classes. We select the data of `Pullover` and `Shirt` for binary classification dataset with 12,000 training and 2000 testing images. We train a 2D CNN with the same architecture as that for MNIST for the classification and achieve the test accuracy of 92.20%. The architectures of our approximators and explainer are also the same as those for MNIST. Each method selects top-64 pixels as an explanation for an image.

As shown in Table 4, our method outperforms state-of-the-art baseline methods except for the INFD method. Since INFD adds

Figure 6: Examples of explanations on the ImageNet dataset. In each panel, an original image and the masked images by VIBI and our method are presented from left to right. Best view in colors.

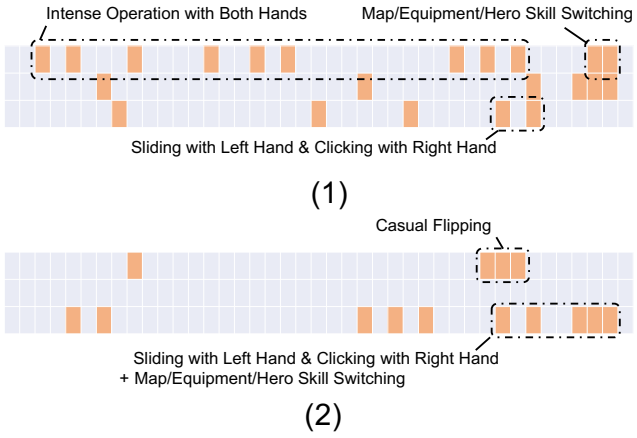

Figure 7: Examples of explanations on the TGD dataset (Honor of Kings). Each panel shows the mask for each time-series data sample with the time dimension of 3. (1) and (2) show samples of a teenager and an adult, respectively. Selected features are in orange. Best view in colors.

perturbation to each feature and directly performs regression, it is suitable for well-aligned data, *e.g.,* the Fashion-MNIST dataset. However, as shown in Tables 1 and 3, INFD's performances are disappointing for data that are not well-aligned. Therefore, our method is more robust. Moreover, INFD is extremely time-consuming to be applied to practical applications and is sensitive to adversarial examples. Compared with the next-best baseline VIBI, we show some examples of selected pixels in Fig. 5. As shown in Fig. 5, VIBI primarily focuses on the contours, whereas our method focuses on relatively fixed local regions. Since the data are well-aligned, the explanations provided by our method are more consistent with the machine logic of the original model.

4.4 ImageNet

We select the data of `Gorilla` and `Zebra` from ImageNet [10] for binary classification. We adopt the MobileNet [16] and train only the top layer for the classification and achieve the test accuracy of 100%. We develop our approximators with the same architecture and adopt the U-Net [28] for our explainer. Each method selects top-10% pixels as an explanation for an image. As shown in Table 5, our method outperforms state-of-the-art baselines. We exhibit some

Table 5: Results on the ImageNet dataset.

Method	FS-M	FU-M	FS-A	FU-A	SEN	TPS
Grad	55	91	66	99	1e+6	2e-3
GI	56	95	74	89	1e+6	2e-3
LIME	77	98	72	96	76.1	1e-0
SHAP	66	96	75	90	89.4	2e-0
†CXP	77	96	61	95	41.38	5e-3
INFD	-	-	-	-	-	4e+3
L2X	78	99	75	96	41.32	5e-3
VIBI	78	98	78	96	40.10	5e-3
Ours	83	98	90	93	39.91	5e-3

examples of selected pixels in Fig. 6 and compare them with the best baseline VIBI. As shown in Fig. 6 (1), our selected pixels are more concentrated on the label-related regions, which demonstrates that our method can improve the model identifiability. Fig. 6 (2) shows that our method can better avoid irrelevant regions, *e.g.,* the ground and the back of an ostrich.

4.5 TGD

Last, we apply our method to the Tencent Gaming Dataset (TGD), which consists of 100 million samples from 5 million gamers. Each sample is a 3×643 time-series data with the time dimension and feature dimension of 3 and 643, respectively. We extract the features from inertia sensors and touch information of mobile phones, in both time and frequency domains, and categorize in 41 groups. Each feature vector of a sample corresponds to a 2-second operation during the game. Three vectors are ordered by time, but not necessarily continuous in time. The learning task is the teenage gamer (age $\leq$ 17) recognition. The original model is a stacked LSTM with an accuracy of 90.16%. The approximator uses the same structure, and the explainer is also a stacked LSTM. Our method achieves the FS-M, FU-M, FS-A, FU-A, and SEN scores of 95.68%, 82.24%, 95.33%, 82.37%, and 0.18%, respectively, selecting only 10% of features. We show examples of selected features in Fig. 7. In Fig. 7 (1), the teenage gamer performs a complex operation excitedly at the start but performs a monotonous/regular operation at the end. Whereas, in Fig. 7 (2), the adult gamer starts with casual flipping of the mobile phone, and ends with a complex/skilled operation.

5 CONCLUSION

In this paper, we investigate the model interpretation problem in the favored direction of Instance-wise Feature Selection (IFS). We propose a Model-agnostic Effective Efficient Direct (MEED) IFS framework for model interpretation. Specifically, we consider the model output feedback as an additional input to learn an explainer to mitigate the sanity and information transmission problems. Furthermore, we propose an adversarial infidelity learning (AIL) mechanism to screen relative unimportant features for mitigating the combinatorial shortcuts and the model identifiability problems. Our theoretical analyses show that AIL can mitigate the model identifiability and combinatorial shortcuts problems. Our experimental results reveal that AIL can mitigate the model identifiability problem and learn more necessary features. Moreover, our extension

to integrate efficient interpretation methods as proper priors has been shown to provide a warm start and mitigate the information transmission problem. Comprehensive empirical evaluation results provided by quantitative metrics and human evaluation demonstrate the effectiveness, superiority, and robustness of our method.

ACKNOWLEDGMENTS

Jian Liang, Bing Bai, Yuren Cao, and Kun Bai would like to acknowledge the support from the TuringShield team of Tencent. Fei Wang would like to acknowledge the support from AWS Machine Learning for Research Award and Google Faculty Research Award.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning.. In *OSDI*, Vol. 16. 265–283.

[2] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. 2018. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307* (2018).

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*. 9505–9515.

[4] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. A unified view of gradient-based attribution methods for deep neural networks. In *NIPS 2017-Workshop on Interpreting, Explaining and Visualizing Deep Learning*. ETH Zurich.

[5] Robert Andrews, Joachim Diederich, and Alan B Tickle. 1995. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems* 8, 6 (1995), 373–389.

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[7] Seojin Bang, Pengtao Xie, Wei Wu, and Eric Xing. 2019. Explaining a black-box using Deep Variational Information Bottleneck Approach. *arXiv preprint arXiv:1902.06918* (2019).

[8] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E Smith, and Subbarao Kambhampati. 2019. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 29. 86–96.

[9] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *International Conference on Machine Learning*. 883–892.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[11] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 13567–13578. http://papers.nips.cc/paper/9511-explanations-can-be-manipulated-and-geometry-is-to-blame.pdf

[12] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[13] Satoshi Hara, Koichi Ikeno, Tasuku Soma, and Takanori Maehara. 2019. Feature Attribution As Feature Selection. https://openreview.net/forum?id=H1lS8oA5YQ

[14] Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling Neural Network Interpretations via Adversarial Model Manipulation. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 2921–2932. http://papers.nips.cc/paper/8558-fooling-neural-network-interpretations-via-adversarial-model-manipulation.pdf

[15] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*. 9734–9745.

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[17] Sarthak Jain and Byron C Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3543–3556.

[18] Alexia Jolicoeur-Martineau. 2018. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734* (2018).

[19] Ashkan Khakzar, Soroosh Baselizadeh, Saurabh Khanduja, Seong Tae Kim, and Nassir Navab. 2019. Explaining Neural Networks via Perturbing Important Learned Features. *arXiv preprint arXiv:1911.11081* (2019).

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[21] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10285–10295.

[22] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[23] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[24] Andrew L 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 the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics, 142–150.

[25] Seongsik Park, Seijoon Kim, Hyeokjun Choe, and Sungroh Yoon. 2019. Fast and efficient information transmission with burst spikes in deep spiking neural networks. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[26] Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. 2018. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*. 2515–2524.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. ACM, 1135–1144.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[29] Patrick Schwab and Helmut Hlavacs. 2015. Capturing the essence: Towards the automated generation of transparent behavior models. In *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*.

[30] Patrick Schwab and Walter Karlen. 2019. CXPlain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*. 10220–10230.

[31] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3145–3153.

[32] 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).

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

[34] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR (workshop track)*.

[35] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3319–3328.

[36] Fei Wang, Rainu Kaushal, and Dhruv Khullar. 2019. Should Health Care Demand Interpretable Artificial Intelligence or Accept âĂIJBlack BoxâĂİ Medicine? *Annals of Internal Medicine* (2019).

[37] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 11–20.

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[39] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (In) fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems*. 10965–10976.

[40] Xinyang Zhang, Ningfei Wang, Shouling Ji, Hua Shen, and Ting Wang. 2018. Interpretable Deep Learning under Fire. *arXiv preprint arXiv:1812.00891* (2018).

[41] Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial Attention Modeling for Multi-dimensional Emotion Regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 471–480.

A PROOFS AND DERIVATIONS

A.1 Proof of Theorem 1

Proof. The proof follows that of Theorem 1 of Chen *et al.* [9].

(1) Forward direction: Given the definition of S^*, we have for any pair $(\mathbf{x}, \mathbf{y})$, and any explainer $E : S \mid \mathbf{x}, \mathbf{y}$,

$$\mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_S) - \log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}})] \leq \mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_{S^*}) - \log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}^*})] \tag{14}$$

In the case when S^* is a set instead of a singleton, we identify S^* with any distribution that assigns arbitrary probability to each elements in S^*, and with zero probability outside S^*. With abuse of notation, $S^*(\mathbf{x}, \mathbf{y})$ indicates both the set function that maps every pair $(\mathbf{x}, \mathbf{y})$ to a set S^* and any real-valued function that maps $(\mathbf{x}, \mathbf{y})$ to an element in S^*. Taking expectation over the distribution of $(\mathbf{x}, \mathbf{y})$, and adding $\mathbb{E}[\log p(\mathbf{y})]$ at both sides, we have

$$I(\mathbf{x}_S; \mathbf{y}) - I(\mathbf{x}_{\bar{S}}; \mathbf{y}) \leq I(\mathbf{x}_{S^*}; \mathbf{y}) - I(\mathbf{x}_{\bar{S}^*}; \mathbf{y}) \tag{15}$$

for any explainer $E : S \mid \mathbf{x}, \mathbf{y}$.

(2) Reverse direction: The reverse direction is proved by contradiction. Since the optimal explanation $S \mid \mathbf{x}, \mathbf{y}$ satisfies

$$I(\mathbf{x}_{S'}; \mathbf{y}) - I(\mathbf{x}_{\bar{S}'}; \mathbf{y}) \leq I(\mathbf{x}_S; \mathbf{y}) - I(\mathbf{x}_{\bar{S}}; \mathbf{y}) \tag{16}$$

for any other $S' \mid \mathbf{x}, \mathbf{y}$, assume the optimal explanation $S \mid \mathbf{x}, \mathbf{y}$ is such that there exists a set $\mathcal{S}$ of nonzero probability, over which $S \mid \mathbf{x}, \mathbf{y}$ does not degenerates to an element in S^*. Concretely, we define $\mathcal{S}$ as

$$\mathcal{S} = \{\mathbf{x}, \mathbf{y} : p(S \notin S^* \mid \mathbf{x}, \mathbf{y}) > 0\}. \tag{17}$$

For any $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}$, we have

$$\mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_S) - \log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}})] < \mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_{S^*}) - \log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}^*})], \tag{18}$$

where $S^*(\mathbf{x}, \mathbf{y})$ is a deterministic function in the set of distributions that assign arbitrary probability to each elements in S^*, and with zero probability outside S^*. Outside $\mathcal{S}$, we always have

$$\mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_S) - \log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}})] \leq \mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_{S^*}) - \log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}^*})] \tag{19}$$

from the definition of S^*. As $\mathcal{S}$ is of nonzero size over $p(\mathbf{x}, \mathbf{y})$, combining Eq. (18) and Eq. (19), taking expectation with respect to $p(\mathbf{x}, \mathbf{y})$ and adding $\mathbb{E}[\log p(\mathbf{y})]$ at both sides, we have

$$I(\mathbf{x}_S; \mathbf{y}) - I(\mathbf{x}_{\bar{S}}; \mathbf{y}) < I(\mathbf{x}_{S^*}; \mathbf{y}) - I(\mathbf{x}_{\bar{S}^*}; \mathbf{y}) \tag{20}$$

which is a contradiction to Eq. (16). □

A.2 Derivations for The Variational Lower Bounds

Proof. First, for selected features, we have:

$$\begin{aligned} I(\mathbf{x}_S; \mathbf{y}) &= \mathbb{E}\left[\log \frac{p(\mathbf{x}_S, \mathbf{y})}{p(\mathbf{x}_S)p(\mathbf{y})}\right] = \mathbb{E}\left[\log \frac{p(\mathbf{y} \mid \mathbf{x}_S)}{p(\mathbf{y})}\right] \\ &= \mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_S)] + \text{Const.} \\ &= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}|\mathbf{x}}\mathbb{E}_{S|\mathbf{x},\mathbf{y}}\mathbb{E}_{\mathbf{y}|\mathbf{x}_S}[\log p(\mathbf{y} \mid \mathbf{x}_S)] + \text{Const.} \end{aligned} \tag{21}$$

For any $q_s(\mathbf{y} \mid \mathbf{x}_S)$, we obtain the lower bound by applying the Jensen's inequality:

$$\begin{aligned} \mathbb{E}_{\mathbf{y}|\mathbf{x}_S}[\log p(\mathbf{y} \mid \mathbf{x}_S)] &\geq \int [\log q_s(\mathbf{y} \mid \mathbf{x}_S)] dp(\mathbf{y} \mid \mathbf{x}_S) \\ &= \mathbb{E}_{\mathbf{y}|\mathbf{x}_S}[\log q_s(\mathbf{y} \mid \mathbf{x}_S)]. \end{aligned} \tag{22}$$

It is similar for unselected features. □

A.3 Proof of Theorem 2

Proof. For the problem in Eq. (7) and (8), A_u learns $p(\mathbf{y} \mid \mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})$, where $\mathbf{1}$ is a vector with all the elements being 1.

Therefore, our AIL mechanism learns $\mathbf{v}$ to minimize $p(\mathbf{y} \mid \mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})$. Since

$$\begin{aligned} &I((\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v}); \mathbf{y}) \\ &= \mathbb{E}\left[\log \frac{p(\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v}, \mathbf{y})}{p(\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})p(\mathbf{y})}\right] \\ &= \mathbb{E}\left[\log \frac{p(\mathbf{y} \mid \mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})}{p(\mathbf{y})}\right] \\ &= \mathbb{E}[\log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})] + \text{Const.} \\ &= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}|\mathbf{x}}\mathbb{E}_{\mathbf{v}|\mathbf{x},\mathbf{y}}\mathbb{E}_{\mathbf{y}|\mathbf{x}_{\bar{S}},\mathbf{1}-\mathbf{v}}[\log p(\mathbf{y} \mid \mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})] \\ &\quad + \text{Const.}, \end{aligned} \tag{23}$$

the mutual information $I((\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v}); \mathbf{y})$ is minimized. By the property of mutual information, we assume that the minimization of $I((\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v}); \mathbf{y})$ encourages the independence between $(\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})$ and $\mathbf{y}$, which leads to

$$p(\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v}, \mathbf{y}) = p(\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})p(\mathbf{y}) \tag{24}$$

Thus, by marginalizing $\mathbf{x}_{\bar{S}}$ at both sides, we have

$$\begin{aligned} \int_{\mathbf{x}_{\bar{S}}} p(\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v}, \mathbf{y}) &= \int_{\mathbf{x}_{\bar{S}}} p(\mathbf{x}_{\bar{S}}, \mathbf{1} - \mathbf{v})p(\mathbf{y}) \\ &\Rightarrow p(\mathbf{1} - \mathbf{v}, \mathbf{y}) = p(\mathbf{1} - \mathbf{v})p(\mathbf{y}). \end{aligned} \tag{25}$$

Therefore, the independence between $\mathbf{1} - \mathbf{v}$ and $\mathbf{y}$ is encouraged as well. Since $\mathbf{1} - \mathbf{v}$ and $\mathbf{v}$ are deterministic between each other, then for any set $\mathcal{S}_1$ such that $\mathbf{v} \in \mathcal{S}_1$, there exists a fixed set $\mathcal{S}_2$ such that $\mathbf{1} - \mathbf{v} \in \mathcal{S}_2$, and vice versa. Thus we have for any set $\mathcal{S}_1$,

$$\begin{aligned} p(\mathbf{v} \in \mathcal{S}_1, \mathbf{y}) &= p(\mathbf{1} - \mathbf{v} \in \mathcal{S}_2, \mathbf{y}) \\ &= p(\mathbf{1} - \mathbf{v} \in \mathcal{S}_2)p(\mathbf{y}) = p(\mathbf{v} \in \mathcal{S}_1)p(\mathbf{y}). \end{aligned} \tag{26}$$

Thus, the independence between $\mathbf{v}$ and $\mathbf{y}$ is also encouraged. □

A.4 Derivation for Eq. (11)

Proof.

$$\begin{aligned} &p(\delta = j \mid \mathbf{x}, M, E, H) \\ &= \frac{p(\delta = j, E, H \mid \mathbf{x}, M)}{p(E, H \mid \mathbf{x}, M)} \\ &= \frac{p(\delta = j \mid \mathbf{x}, M)p(E, H \mid \delta = j, \mathbf{x}, M)}{p(E, H \mid \mathbf{x}, M)} \\ &= \frac{p(\delta = j \mid \mathbf{x}, M)p(E \mid \delta = j, \mathbf{x}, M)p(H \mid \delta = j, \mathbf{x}, M)}{p(E, H \mid \mathbf{x}, M)} \\ &= \frac{p(\delta = j \mid \mathbf{x}, M)\frac{p(E, \delta=j|\mathbf{x}, M)}{p(\delta=j|\mathbf{x}, M)}\frac{p(H, \delta=j|\mathbf{x}, M)}{p(\delta=j|\mathbf{x}, M)}}{p(E, H \mid \mathbf{x}, M)} \\ &= \frac{p(\delta = j \mid \mathbf{x}, M)\frac{p(E|\mathbf{x}, M)p(\delta=j|\mathbf{x}, M, E)}{p(\delta=j|\mathbf{x}, M)}\frac{p(H|\mathbf{x}, M)p(\delta=j|\mathbf{x}, M, H)}{p(\delta=j|\mathbf{x}, M)}}{p(E, H \mid \mathbf{x}, M)} \\ &= C(E, H \mid \mathbf{x}, M)\frac{p(\delta = j \mid \mathbf{x}, M, E)p(\delta = j \mid \mathbf{x}, M, H)}{p(\delta = j \mid \mathbf{x}, M)}, \end{aligned} \tag{27}$$

where

$$C(E, H \mid \mathbf{x}, M) = \frac{p(E \mid \mathbf{x}, M)p(H \mid \mathbf{x}, M)}{p(E, H \mid \mathbf{x}, M)}. \tag{28}$$

The third equation is from the assumption of conditional independence between the interpretation models E and H given $\delta, \mathbf{x}$ and M. Assuming $p(\delta = j \mid \mathbf{x}, M) = \frac{1}{d}$ for all $j \in [d]$, because we have no knowledge of the explainer, then we have

$$\begin{aligned} p(\delta = j \mid \mathbf{x}, M, E, H) &= \frac{p(\delta = j \mid \mathbf{x}, M, E, H)}{\sum_{j'=1}^{d} p(\delta = j' \mid \mathbf{x}, M, E, H)} \\ &= \frac{p(\delta = j \mid \mathbf{x}, M, E)p(\delta = j \mid \mathbf{x}, M, H)}{\sum_{j'=1}^{d} p(\delta = j' \mid \mathbf{x}, M, E)p(\delta = j' \mid \mathbf{x}, M, H)}. \end{aligned} \tag{29}$$

□

B IMPLEMENTATION DETAILS

B.1 Details for Our Method

For ℓ_u with cross entropy loss, in Eq. (8) we still minimize $\mathcal{L}_u$, and just replace the target $\mathbf{y}$ for ℓ_u by $\mathbf{1} - \mathbf{y}$, following the suggestion of the Relativistic GAN *et al.* [18]. For ℓ_u with the Sliced Wasserstein distance [21], the number of random vectors is 128 for Fashion-MNIST and 256 for ImageNet.

For optimizers, we use RMSprop for IMDB and Adadelta for image datasets, with the default hyperparameters. The learning rates are fixed. For TGD, we use Adam with learning rate of $2e-4$, $\beta_1 = 0.5, \beta_2 = 0.999$, with the learning-rate decay of 30% for every 50,000 steps. The batch size is 32 for IMDB and ImageNet, 128 for MNIST and Fashion-MNIST, and 1024 for TGD. The hyper-parameter tuning set for both $\mathcal{L}_u$ and $\mathcal{L}_e$ is $\{0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$.

For IMDB, we adopt the structure of the original model of L2X [9] for the same structure for our original model and approximators. We adopt adopt the structure of the explainer of L2X [9] for the structure for our explainer, with a global and a local component. For MNIST and Fashion-MNIST, the structure of our original model and approximators is shown in Table 6. Whereas the structure for our explainer is shown in Table 7. For ImageNet, we adopt the MobileNet module in the package of keras.applications.mobilenet without the top layer as our explainer. The model parameters pre-trained on ImageNet are fixed. We only stack a global max-pooling layer and learn a full-connected top layer. We adopt the preprocess_input function in the keras.applications.mobilenet package for image pre-processing. We perform a max-pooling of kernel size and stride of 4, before generate the feature important scores, and perform an up-sampling with kernel size and stride of 4 when masking pixels. We adopt the adopt the U-Net [28] for our explainer, whose structure is complex and then omitted due to space limitations. Similarly, the structures of our modules for TGD are also omitted. Readers may refer to the publicly-available code for more implementation details.

For IMDB, the model output $\mathbf{y}$ is input to a MLP with three hidden layers with 100 neurons and the ReLu activation, before being concatenated to the global component of the explainer. For image datasets, the model output $\mathbf{y}$ is linearly mapped to the same shape with the first channel of an image and is concatenated to the raw image as an additional channel. For TGD, the model output $\mathbf{y}$ is linearly mapped to the same shape as an raw data sample and concatenated to the data sample in the feature dimension.

Table 6: The CNN structure for our original model and approximators for MNIST and Fashion-MNIST.

Layer	# Filters	Kernel Size	Stride	# Padding	Activation
Convolution	32	3	3	default	ReLu
Convolution	64	3	3	default	ReLu
Max-pooling	-	2	2	default	-
Dropout(0.25)	-	-	-	-	-
Flatten	-	-	-	-	-
Fully-Connected	128	-	-	-	ReLu
Dropout(0.5)	-	-	-	-	-
Fully-Connected	2	-	-	-	Softmax

Table 7: The CNN structure for our explainer for MNIST and Fashion-MNIST.

Layer	# Filters	Kernel Size	Stride	# Padding	Activation
Convolution	32	3	3	same	ReLu
Convolution	64	3	3	same	ReLu
Convolution	1	3	3	same	Linear

B.2 Details for Baseline Methods

For Grad, we compute the gradient of the selected class with respect to the input feature and uses the absolute values as importance scores. We perform summation operations to form the importance scores with proper shapes. For GI, the gradient is multiplied by the input feature before calculate the absolute value. For INFD, we select the Noisy Baseline method for consistent comparisons, since its another method Square is only suitable for image datasets. The structures of explainers are the same for CXP, L2X, VIBI, and Ours. The structures of original models and approximators are the same for L2X, VIBI, and Ours.

The hyper-parameters of each method are tuned according the strategy mentioned in their respective papers.

On ImageNet, for all the baseline methods, we perform a max-pooling of kernel size and stride of 4 for the feature important scores, and perform an up-sampling with kernel size and stride of 4 when masking pixels.