INTERVENTIONAL RATIONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective rationalizations improve the explainability of neural networks by selecting a subsequence of the input (i.e., rationales) to explain the prediction results. Although existing methods have achieved promising results, they still suffer from adopting the spurious correlations in data (aka., shortcuts) to compose rationales and make predictions. Inspired by the causal theory, in this paper, we develop an interventional rationalization (Inter-RAT) to discover the causal rationales. Specifically, we first analyse the causalities among the input, rationales and results with a structural causal model. Then, we discover spurious correlations between the input and rationales, and between rationales and results, respectively, by identifying the confounder in the causalities. Next, based on the backdoor adjustment, we propose a causal intervention method to remove the spurious correlations in input and rationales. Further, we discuss reasons why spurious correlations between the selected rationales and results exist by analysing the limitations of the sparsity constraint in the rationalization, and employ the causal intervention method to remove these correlations. Extensive experimental results on three real-world datasets clearly validate the effectiveness of our proposed method.

1 INTRODUCTION

The remarkable success of deep neural networks (DNNs) in natural language understanding tasks has prompted the interest in how to explain the results of DNNs. Among them, the selective rationalization task Lei et al. (2016); Yu et al. (2019; 2021) has received increasing attention, answering the question "What feature has a significant impact on the prediction results of the model?". Specifically, the goal of selective rationalization is to extract a small subset of the input (i.e., rationale) to support and explain the prediction results when yielding them. Existing methods often generate rationales with a conventional framework consisting of a *selector* (aka., *rationale generator*) and a *predictor* Lei et al. (2016). As shown in Figure 1, giving the input X, the *selector* and the *predictor* generate rationales R and prediction results Y cooperatively (i.e., $P(Y|X) = P(Y|R)P(R|X)$). Among them, the *selector* ($P(R|X)$) first extracts a subsequence of the input. Then, the *predictor* ($P(Y|R)$) yields results based only on the selected tokens, and the selected subsequence is defined as the rationale.

Despite the appeal of the rationalization methods, the current implementation is prone to exploit spurious correlations (aka., shortcuts) between the input and labels to yield the prediction results and select the rationales Chang et al. (2020); Wu et al. (2022). We illustrate this problem with an example of the charge prediction[1]. Considering Figure 1, although this case is corresponding to the *Manslaughter*, a DNNs model readily predicts the charge as *Intentional homicide*. Specifically, as *Intentional homicide* occurs more frequently than *Manslaughter*[2] and is often accompanied by tokens denoting violence and death, DNNs do not need to learn the real correlations between the case facts and the charge to yield the result. Instead, it is much easier to exploit spurious correlations in data to achieve high accuracy (i.e., predicting the charge as *Intentional homicide* directly when identifying the tokens about violence and death.). As a result, when facing the cases such as the example in Figure 1, the effectiveness of such DNNs tends to degrade (e.g., the underlined tokens in Figure 1 denoting the offence is negligent will be ignored in rationales extraction and the charge will be misjudged.). Therefore, these types DNNs depending on spurious correlation in data fail to reveal truly critical subsequence for predicting labels.

[1]Charge prediction: predicting the charge such as *Robbery* and *Theft* based on the case fact. Detailed definition of charge prediction is described in section 4.3.

[2]`https://wenshu.court.gov.cn`

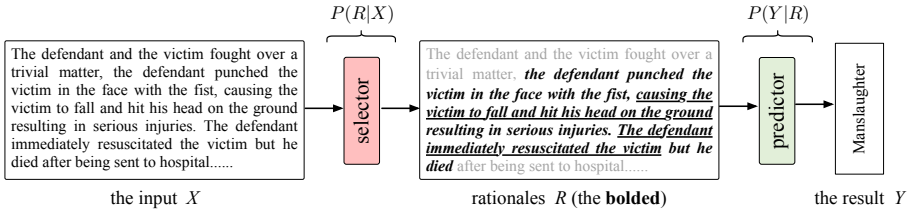

Figure 1: Conventional framework of rationalization presented in this paper. In the charge prediction, the input X represents the case fact and the result Y denotes the charge.

To solve that, Chang et al. (2020) propose an environment-invariant method (INVRAT) to discover the causal rationales. They argue that the causal rationales should remain stable as the environment shifts, while the spurious correlation between input and labels vary. Although this method performs well in selecting rationales, since the environment in rationalization is hard to observe and obtain, we argue that this "*causal pattern*" can be further explored to improve the rationalization.

Along this research line, in this paper, we propose an interventional rationalization (Inter-RAT) method which removes the spurious correlation by the causal intervention Glymour et al. (2016). Specifically, motivated by the causal inference theory, we first formulate the causal relationships among X, R and Y in a Structural Causal Model (SCM) Pearl et al. (2000); Glymour et al. (2016) as shown in Figure 2(a). Then, we identify the confounder C in this SCM, which opens two backdoor paths $X \leftarrow C \rightarrow R$ and $R \leftarrow C \rightarrow Y$, making X and R, R and Y spuriously correlated. Next, we address the above correlations, respectively. For spurious correlations between X and R, we assume the confounder is observed and intervene the X (i.e., calculating $P(R|do(X))$ instead of $P(R|X)$) to block the backdoor path and remove the spurious correlations based on the backdoor adjustment Glymour et al. (2016). Among them, the *do*-operation denotes the pursuit of real causality from X to R. For spurious correlations in R and Y, since by the definition of R (rationales are the only basis for yields prediction results), we argue that there should be no spurious correlations between R and Y. However, in practice, we discover the sparsity constraint commonly defined in rationalization Lei et al. (2016); Cao et al. (2020); Chang et al. (2020); Yu et al. (2019), ensuring the *selector* to extract short rationales, results in the spurious correlations between R and Y. Therefore, we further analyse this discovery and employ the causal intervention to remove these correlations.

Our experiments are conducted on a multi-aspect sentiment analysis dataset BeerAdvocate McAuley et al., a movie reviews prediction dataset MovieReview Zaidan & Eisner (2008) and a legal judgment prediction dataset CAIL Xiao et al. (2018). The experimental results validate the effectiveness of removing spurious correlation with causal interventions, where our proposed approach gains an average improvement of 8.6 token F1-score[3] over the INVRAT baseline on BeerAdvocate, 7.4 token F1-score on MovieReview, and 4.3 F1-score on CAIL.

2 The Conventional Framework of Rationalization

This section formally defines the problem of rationalization, and then presents the details about the conventional rationalization framework consisting of the *selector* and *predictor*, where these two components are trained cooperatively to generate rationales and yield the prediction results.

2.1 Problem Formulation

Considering a text classification task, only the text input $X = \{x_1, x_2, \ldots, x_n\}$, where x_i represents the i-th token, and the discrete ground truth Y are observed during training, while the rationale R is unavailable. The goal of selective rationalization is first adopting the *selector* to learn a binary mask variable $M = \{m_1, m_2, \ldots, m_n\}$, where $m_j \in \{0, 1\}$, and further select a subsequence of input $R = M \odot X = \{m_1 \cdot x_1, m_2 \cdot x_2, \ldots, m_n \cdot x_n\}$, and then employing the *predictor* to re-recode the mask input R to yield the results. Finally, the whole process of rationalization is defined as:

$$P(Y|X) = \underbrace{P(Y|R)}_{predictor}\,\underbrace{P(R|X)}_{selector}. \tag{1}$$

[3]The definition of token F1-score can be found in section 4.1.1.

2.2 SELECTOR

The *selector* divides the process of generating rationales into three steps. First, the *selector* samples each binary value m_j from the probability distribution $P(\widetilde{M}|X) = \{p_1, p_2, \ldots, p_n\}$, where p_j represents the probability of selecting each x_j as the part of the rationale. Specifically, p_j is calculated as $p_j = P(\widetilde{m}_j|x_j) = \text{softmax}(W_e f_e(x_j))$, where the encoder $f_e(\cdot)$ encodes the token x_j into a d-dimensional vector and $W_e \in \mathbb{R}^{2\times d}$. Then, to ensure the sampling operation is differentiable, several reparameterization tricks have been proposed such as policy gradient Lei et al. (2016) and HardKuma Bastings et al. (2019). In this paper, we adopt the Gumbel-softmax method Jang et al. (2017) to achieve this goal:

$$m_j = \frac{\exp\left(\left(\log\left(p_j\right) + g_j\right)/\tau\right)}{\sum_t \exp\left(\left(\log\left(p_t\right) + g_t\right)/\tau\right)}, \tag{2}$$

where τ is a temperature hyperparameter, $g_j = -\log\left(-\log\left(u_j\right)\right)$ and u_j is random sampled from the uniform distribution $U(0, 1)$. Finally, the rationale can be selected as $R = M \odot X = \{m_1 \cdot x_1, m_2 \cdot x_2, \ldots, m_n \cdot x_n\}$. Therefore, we conclude that the probability of generating rationales $P(R|X)$ is calculated as: $P(R|X) = P(M \odot X|X) \equiv P(\widetilde{M}|X)$.

2.3 PREDICTOR

Based on selected rationale tokens R, the *predictor* outputs the prediction results (i.e., calculating $P(Y|R) = P(Y|M \odot X)$), and then R can be seen as an explanation of Y. Specifically, after obtaining R from the *selector*, we adopt the neural network $f_p(\cdot)$ to re-encode the rationale into d-dimensional continuous hidden states to yield results. The objective of the *predictor* is defined as:

$$\mathcal{L}_{task} = \mathbb{E}_{\substack{X,Y\sim\mathcal{D}_{tr} \\ M\sim P(\widetilde{M}|X)}} \left[\ell\left(Y, W_p f_p\left(M \odot X\right)\right)\right], \tag{3}$$

where $\mathcal{D}_{tr}$ denotes the training set, $\ell(\cdot)$ represents the cross-entropy loss function, $W_p \in \mathbb{R}^{N\times d}$ is the trained parameter and N is the number of labels (e.g., $N = 2$ in the binary classification).

2.4 SPARSITY AND CONTINUITY CONSTRAINTS

Since an ideal rationale should be a short and coherent part of original inputs, we add the sparsity and continuity constraints Lei et al. (2016); Chang et al. (2020):

$$\mathcal{L}_{re} = \lambda_1 \left|\alpha - \frac{1}{n}\sum_{j=1}^{n} m_j\right| + \lambda_2 \sum_{j=2}^{n} \left|m_j - m_{j-1}\right|, \tag{4}$$

where the first term encourages the model to select short rationales and α is a predefined sparsity level at the scale of [0, 1], and the second term ensures the coherence of selected tokens. Finally, the overall objective of the rationalization is defined as: $\mathcal{L} = \mathcal{L}_{task} + \mathcal{L}_{re}$.

3 INTERVENTIONAL RATIONALIZATION

In this section, we first reveal how the confounder C causes spurious correlations in rationalization with a causal graph. Then, we remove these correlations by using a causal intervention method.

3.1 STRUCTURAL CAUSAL MODEL

As shown in Figure 2(a), we formulate the causalities among the text input X, rationale R, ground-truth label Y and the confounder C with a Structural Causal Model (SCM) Pearl et al. (2000); Glymour et al. (2016), where the link between two variables represents a causal relationship.

In the following, we introduce the causal graph with these variables at a high-level:

$C \rightarrow X$. The confounder C in rationalization can be seen as the context prior, determining which tokens can be "put" into the text input X. For example, in Figure 1, the context prior decides where the tokens denoting violence and manslaughter appear, and also decides where other tokens that are meaningless appear. In practice, the confounder is commonly partially observed (e.g., in the text classification, we consider the entire label set as the partially observed confounder. See section 3.2 for details.).

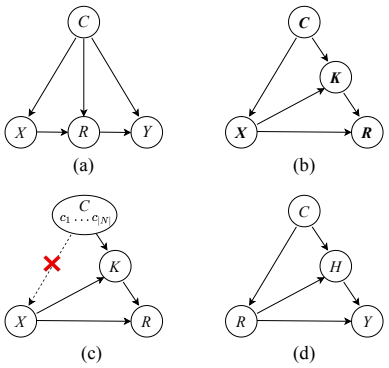

Figure 2: Structural Causal Model for Rationalization.

$X \rightarrow R \leftarrow C$. Besides the *selector* extracts a subsequence of X as the rationale R, making $X \rightarrow R$ holds, R is also affected by the context prior C. Figure 2(b) offers a fine-grained causal relationship between R and C with a mediator K. Specifically, in $X \rightarrow K \leftarrow C$, K denotes the context-specific representation which is a weighted representation of the prior knowledge associated with X in C. In $X \rightarrow R \leftarrow K$, R is affected by the context prior C through K indirectly. For example, in Figure 1, the underlined tokens denoting negligent are ignored in rationalization, since the context prior in C misleads the model to focus on the violence and death feature in X by the mediator K. Detailed examples about explanations for this SCM can be found in Appendix A.1.

$C \rightarrow Y \leftarrow R$. As the *predictor* yields the result based on the rationale, $R \rightarrow Y$ holds. Meanwhile, in the ideal situation, since the rationale is defined as a subsequence of X sufficient to predict the Y, there should be no direct causal relationship between C and Y. However, in practice, rationales R are commonly extracted with shortcut tokens (we will introduce it later in section 3.3), making $C \rightarrow Y$ exists. Figure 2(d) describes a fine-grained scm between Y and C with a mediator H, where H is the context-specific representation of R by using the context prior C. More detailed descriptions of this SCM are presented in Appendix A.2.

From the graph, we find that X and R, R and Y are confounded by the context prior C with two backdoor paths $X \leftarrow C \rightarrow R$ (or $X \leftarrow C \rightarrow K \rightarrow R$ for elaboration) and $R \leftarrow C \rightarrow Y$ (or $R \leftarrow C \rightarrow H \rightarrow Y$). The above backdoor paths result in spurious correlations among the text input X, rationale R, and label Y. Based on this, we propose a causal intervention method to remove the confounding effect by cutting off the link $C \rightarrow X$ and $C \rightarrow R$, respectively.

3.2 Causal Intervention via Backdoor Adjustment

To pursue the real causality from X to R (or R to Y), we adopt the causal intervention $P(R|do(X))$ instead of $P(R|X)$ (or $P(Y|do(R))$ instead of $P(Y|R)$) to remove the effects of confounder C. Next, we introduce the causal intervention method by taking $P(R|do(X))$ as an example, and $P(Y|do(R))$ is similar. Specifically, since adopting the randomized controlled trial to intervene X is impossible, which requires the control over causal features, we apply the backdoor adjustment Glymour et al. (2016) to achieve $P(R|do(X))$ by cutting off $C \rightarrow X$ (Figure 2(c)):

$$P(R|do(X)) = \sum_{i=1}^{|C|} \left[P\left(R|X, K = g_s\left(X, c_i\right)\right) P\left(c_i\right) \right], \tag{5}$$

where the confounder C is stratified into pieces $C = \left\{c_1, c_2, \ldots, c_{|C|}\right\}$, $P\left(c_i\right)$ denotes the prior distribution of c_i, which is calculated before training, and $g_s(\cdot)$ is a function achieving $X \rightarrow K \leftarrow C$. However, the confounder C is commonly hard to observe. Fortunately, based on the existing researches Wang et al. (2020); D'Amour (2019), we can consider the entire label set as the partially observed children of the unobserved confounder. Therefore, we approximate it by designing a dictionary $D_c = \left\{c_1, c_2, \ldots, c_{|N|}\right\}$ as an $N \times d$ matrix, where N represents the number of labels and d is the hidden feature dimension. As described in section 2.2, we conclude $P(R|X) \equiv P(\widetilde{M}|X)$. Therefore, we can achieve $P(R|do(X)) \equiv P(\widetilde{M}|do(X))$. Specifically, to calculate the probability of each token x_j selected as the rationale, the implementation is defined as:

$$P(\widetilde{m}_j|do(X)) = \sum_{i=1}^{|N|} \left[P\left(\widetilde{m}_j|f_s(x_j, k_i)\right) P\left(c_i\right) \right] = \sum_{i=1}^{|N|} \left[\text{softmax}(f_s(x_j, k_i)) P\left(c_i\right) \right]. \tag{6}$$

Among them, $f_s(\cdot)$ is the function achieving $X \rightarrow R \leftarrow K$, $k_i \in K$ is defined as the content-specific representation by using the context prior c_i, we express it as $k_i = g_s(x_j, c_i) = \lambda_i c_i$, where $\lambda_i \in \lambda$. $\lambda \in \mathbb{R}^N$ is the set of the normalized similarity between x_j and each c_i in the confounder set C (i.e., $\lambda = \text{softmax}(f_e(x_j)D_c^T)$). Besides, since Eq (6) requires sampling of C and this sampling is expensive, we try to find an alternative function that would be easy to compute to approximate it. Empirically, based on the results in Xu et al. (2015); Wang et al. (2020); Yue et al. (2020), we can adopt the NWGM approximation to move the outer sum into the softmax (i.e., $P(\widetilde{m}_j|do(X)) \approx \text{softmax}(\sum_{i=1}^{|N|} f_s(x_j, k_i)P(c_i))$). In this paper, we adopt the linear model $f_s(x_j, k_i) = W_1 f_e(x_j) + W_2 k_i = W_1 f_e(x_j) + W_2 \lambda_i c_i$ to fuse the information of the input X and the confounder C. Then, the final implementation of the intervention is formulated as:

$$P(\widetilde{m}_j|do(X)) \approx \text{softmax}(W_1 f_e(x_j) + W_2 \sum_{i=1}^{|N|} \lambda_i c_i P(c_i)). \tag{7}$$

3.3 Limitations on the Predefined Sparsity α in Rationalization

In this section, we discuss why $C \rightarrow Y$ in Figure 2(a) holds in detail. Since rationales are defined as the subsequence of inputs, being sufficient to yield results, $C \rightarrow Y$ should not exist. However, unfortunately, in practical implementation, the sparsity constraint (denoted by α-constraint) in the first term of Eq (4) may result in spurious correlations between the extracted rationale and the predicted result. Specifically, the α-constraint encourages the *selector* to extract α of tokens from the original text input. When the predefined number of extracted tokens is greater than the length of the practical rationale, a few tokens corresponding to shortcuts of Y may still be selected. For example, as α converges to 1, all tokens in the input will be extracted, including the rationales tokens and shortcuts tokens (more examples are shown in Appendix B.1). Then, the shortcuts tokens will hurt the prediction performance. To alleviate this situation, we first construct a fine-grained causal graph (Figure 2(d)) between the selected rationale R and the prediction results Y. Among them, R represents the rationale generated by α-constraint, H denotes the context-specific representation of R based on the context prior C. As mentioned before, from the graph, we find that as there exists a backdoor path $R \leftarrow C \rightarrow H \rightarrow Y$, R and Y are confounded. Then, based on the above observation, the *predictor* adopts the causal intervention methods described in section 3.2 (i.e., calculating $P(Y|do(R)) \approx \text{softmax}(\sum_{i=1}^{|N|} f_r(R, h_i)P(c_i))$ to remove the spurious correlations and further yield prediction results, where $f_r(\cdot)$ is the function to obtain $R \rightarrow Y \leftarrow H$, and $h_i = g_r(R, c_i)$ represents the process of $R \rightarrow H \leftarrow C$. Detailed descriptions of the graph at a high-level and the derivation are shown in Appendix A.2. Besides, although many rationalizers Jain et al. (2020); Paranjape et al. (2020) do not use α-constraint, we believe their constraint on selecting short rationales can be considered as a variant of α-constraint, as detailed in our Appendix B.2. Then, our intervention method will still be effective on these methods.

4 Experiments

In this section, we validate the effectiveness of our method on three real-world tasks including the beer reviews sentiment analysis, movies reviews prediction and the legal judgment prediction.

4.1 Beer Reviews Sentiment Analysis

Beer Reviews Sentiment Analysis is formulated as a sentiment prediction task, predicting the ratings (at the scale of [0, 1]) for the multiple aspects of beer reviews (e.g., appearance, aroma and palate). We use the BeerAdvocate McAuley et al. as our dataset, which is commonly used in the field of rationalization. As there is high sentiment correlation in different aspects in the same beer review Lei et al. (2016), which may confuse the model training, several researches Lei et al. (2016); Bastings et al. (2019) adopt the de-correlated sub-datasets (i.e., a part of BeerAdvocate) in the training stage. However, a high correlated dataset is more conducive to validating our Inter-RAT which is designed to remove the spurious correlations in data. Although Chang et al. (2020) also conduct a correlated sub-dataset, the data split and processing are not available. Therefore, for a fair comparison, different from the previous study which makes experiments on the sub-dataset, we train and validate models on the original BeerAdvocate containing more than 220,000 beer reviews. Besides, following the setup of Chang et al. (2020), we consider the beer review prediction as a binary classification where the ratings ≤ 0.4 as *negative* and ≥ 0.6 as *positive*. Then, the processed BeerAdvocate is a non-balanced dataset. For example, the label distribution in the appearance is *positive*:*negative* $\approx$ 20:1. For testing, we take manually annotated rationales as our test set, detailed statistics are shown in Appendix C.1.

Table 1: Precision, Recall and F1 of selected rationales for the three aspects. Among them, α is the predefined sparsity level.

Methods	α	Appearance			Aroma			Palate		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
RNP	0.1	32.4±0.5	18.6±0.3	23.6±0.4	44.8±0.4	32.4±0.7	37.6±0.5	24.6±0.5	23.5±0.5	24.0±0.5
HardKuma	0.1	53.6±0.1	28.7±0.1	37.4±0.1	29.3±1.4	25.9±3.8	27.3±2.1	7.7±0.1	6.0±0.1	6.8±0.1
INVRAT	0.1	42.6±0.7	31.5±0.6	36.2±0.6	41.2±0.3	39.1±2.8	40.1±1.6	**34.9±1.5**	45.6±0.2	39.5±1.0
Inter-RAT	0.1	**66.0±0.4**	**46.5±0.8**	**54.6±0.7**	**55.4±0.9**	**47.5±0.6**	**51.1±0.8**	34.6±0.8	**48.2±0.4**	**40.2±0.5**
RNP	0.2	39.4±0.4	44.9±0.1	42.0±0.2	37.5±0.1	51.9±0.7	43.5±0.3	21.6±0.4	38.9±0.5	27.8±0.4
HardKuma	0.2	**64.9±0.9**	69.2±1.0	67.0±0.8	37.0±1.3	55.8±1.9	44.5±1.5	14.6±0.3	22.3±0.8	17.7±0.4
INVRAT	0.2	58.9±0.4	67.2±2.3	62.8±1.1	29.3±1.0	52.1±0.6	37.5±0.6	24.0±1.3	55.2±2.3	33.5±1.6
Inter-RAT	0.2	62.0±0.5	**76.7±1.7**	**68.6±0.4**	**44.2±0.1**	**65.4±0.2**	**52.8±0.1**	**26.3±0.6**	**59.1±0.8**	**36.4±0.7**
RNP	0.3	24.2±0.4	41.2±0.8	30.5±0.5	27.1±0.3	55.7±0.8	36.4±0.4	15.4±0.4	42.2±0.9	22.6±0.5
HardKuma	0.3	42.1±0.3	82.4±1.4	55.7±0.5	24.6±0.1	57.7±0.6	34.5±0.2	21.7±0.1	49.7±0.4	30.2±0.1
INVRAT	0.3	41.5±0.4	74.8±0.3	53.4±0.3	22.8±1.6	65.1±1.7	33.8±1.8	20.9±1.1	**71.6±0.4**	32.3±1.3
Inter-RAT	0.3	**48.1±0.7**	**82.7±0.5**	**60.8±0.4**	**37.9±0.7**	**72.0±0.1**	**49.6±0.7**	**21.8±0.1**	66.1±0.8	**32.8±0.1**

4.1.1 COMPARISON METHODS AND EXPERIMENTAL SETUP

In this section, we present the comparison methods with Inter-RAT including RNP Lei et al. (2016), HardKuma Bastings et al. (2019) and INVRAT Chang et al. (2020). Among them, RNP is an original rationalization method which generates rationales by yielding the Bernoulli distribution of each token and sampling from it. HardKuma Bastings et al. (2019) has improved RNP by designing a new reparameterized gradient estimation method. INVRAT is our main baseline to directly compare with, which is also a method of removing the spurious correlation in data. The difference with our Inter-RAT is that INVRAT learns environment invariant representations by obtaining multiple environments from the training set.

For training, we use the pre-trained glove embeddings Pennington et al. (2014) with size 100, and implement the encoder in both $f_e(\cdot)$ and $f_p(\cdot)$ as the bidirectional GRU Cho et al. (2014) with hidden size 100. We optimize the objective of rationalization using Adam Kingma & Ba (2014) with mini-batch size of 256 and an initial learning rate of 0.001. Besides, we consider the α in Eq (4) as {0.1, 0.2, 0.3}, respectively. For testing, we report the token precision, recall and F1-score to evaluate the quality of selected rationales. Among them, token precision is defined as the percentage of how many the selected tokens are in annotated rationales, and token recall is the percentage of annotated rationale tokens that are selected by model. The token F1-score is calculated as $\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$.

4.1.2 EXPERIMENTAL RESULTS

To demonstrate the effectiveness of our Inter-RAT, we briefly compare it with RNP, HardKuma and INVRAT in Table 1, where Inter-RAT outperforms the baselines consistently in finding correct rationales. Specifically, Inter-RAT surpasses RNP and HardKuma on all three aspects (i.e, appearance, aroma and palate) by a large margin in most metrics. Besides, although INVRAT has shown helpful in discovering the de-confounded rationales, Inter-RAT still performs better than it, improving 10.5, 14.0 and 1.4 on the average token F1-score across three aspects, and Inter-RAT has a lower variance illustrating our method is more stable than INVRAT. Such observations strongly demonstrate that Inter-RAT can remove the spurious correlation in data to select rationales effectively.

As discussed in section 3.3, we propose the causal intervention method to alleviate the problem, where several tokens corresponding to spurious correlations in data may be selected and further mislead the prediction with an increasing α. Here, we conduct an experiment to validate the effectiveness of the causal intervention. Since there is only about 1,000 beer reviews in the test set, we report the binary classification F1-score[4] with different α in the dev set which contains about 30,000 reviews. As shown in Figure 3, we make experiments on the palate aspect, and Inter-once is a variant of Inter-RAT, which yields the rationales based on $P(R|do(X))$ but predicts the results based on $P(Y|R)$, rather than $P(Y|do(R))$. From the observation, we can conclude that when α is small (i.e., the length of selected rationales is smaller than real rationales), the difference between Inter-RAT and Inter-once is minor. However, as α increases, Inter-RAT steadily improves, while the Inter-once grows slowly and even degrades. The above observation illustrates that our causal intervention method can alleviate the spurious correlations problem between R and Y caused by the α-constraint.

[4]Different from token F1, F1-score is commonly adopted to evaluate the performance of binary classification.

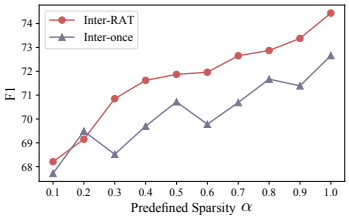

Figure 3: The F1-score on the palate aspect with an increasing α.

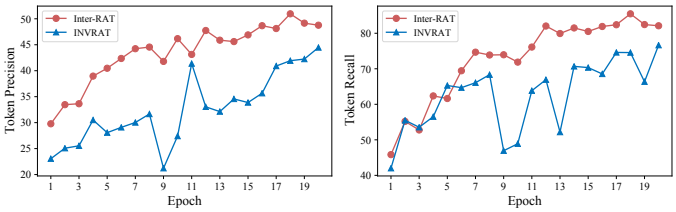

Figure 4: The token precision and recall for rationales on the appearance aspect with $\alpha = 0.3$.

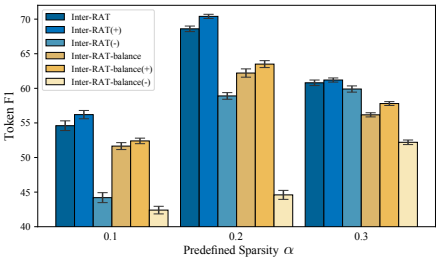

Figure 5: The token F1 for rationales on the appearance aspect with different prior distributions.

As mentioned in section 4.1, we make experiments on the non-balanced dataset, which is different from the previous study Chang et al. (2020; 2019); Huang et al. (2021) adopting the balanced datasets. Therefore, there exists a research question we need to answer:"Does the information of label distributions (or prior distributions) somehow influence Inter-RAT to yield better rationales instead of the causal intervention ?". For instance, as the label distribution in the appearance aspect is *positive*:*negative* $\approx$ 20:1, we conduct experiments on the appearance dataset with $P(c_1) = \frac{20}{21}$ and $P(c_2) = \frac{1}{21}$, where $P(c_i)$ in Eq (5) represents the prior distribution of c_i, c_1 is the *positive* label and c_2 is the *negative* one. The above non-balanced label distribution might be inducing the model to "better rationalize" for the majority class (i.e., the *positive*), further reflecting the improvement of Inter-RAT over the whole dataset. Therefore, we compute token-F1 scores for positive and negative examples separately for a safer evaluation, where we denote the evaluation of Inter-RAT on positive examples as Inter-RAT(+) and on the negative ones as Inter-RAT(-). Figure 5 summarizes the results on the appearance aspect. From the result, we find that the performance of extracting positive rationales is better than extracting the negative, although the difference between the two types results is not significant, and the scores for the negative are still high (better than INVRAT). Therefore, to further validate the effect of label distribution, we add the analysis as follows: we re-run the experiments with $P(c_1) = P(c_2) = \frac{1}{2}$ (i.e., assuming this is a balanced dataset with uniform label distributions) and denote the corresponding model as Inter-RAT-balance. We report experimental results in Figure 5. From the observation, we can find adopting the true prior distribution $P(c_i)$ (Inter-RAT) performs better than the assumed one (Inter-RAT-balance), which demonstrates the prior distribution is critical for the backdoor adjustment method. Besides, it is interesting to see that with a balanced label distribution, the results of the minority label (i.e., *negative*) are worse than using the true label distribution, which suggests that Inter-RAT is not simply "paying more attention" to instances of the majority class.

Besides, comparing with INVRAT, we investigate the model performance by showing the changes in token precision and recall with training epochs. Figure 4 shows the experiments on the appearance aspect with $\alpha = 0.3$. From the observation, we can conclude that Inter-RAT significantly outperforms INVRAT in both precision and recall with lower variance from the training onwards, which proves the effectiveness of our proposed method.

4.2 Movies Reviews Prediction

Besides the beer reviews sentiment analysis task, we also make experiments on another binary classification task (i.e., movie review prediction Zaidan & Eisner (2008)) in the ERASER benchmark DeYoung et al. (2020), which contains token-level human annotations. We follow the same experimental setups in 4.1.1 and report the experimental results with $\alpha = 0.2$ in Table 2. Detailed description of the dataset is shown in Appendix C.1. As shown in the table, Inter-RAT performs better than RNP, HardKuma and INVRAT on the three metrics, which further validates the effectiveness of Inter-RAT. Furthermore, to validate

Table 2: Results on movie reviews, where several results of baselines are quoted from Yu et al. (2021).

Methods	Movie		
	Precision	Recall	F1
RNP	35.6	21.1	24.1
Bert_RNP	–	–	32.2
HardKuma	31.1	28.3	27.0
INVRAT	33.9	24.3	28.3
Inter-RAT	**35.7**$\pm$**0.2**	35.8$\pm$1.7	35.7$\pm$0.8
Bert_Inter-RAT	31.7$\pm$0.1	**43.1**$\pm$**1.1**	**36.5**$\pm$**0.3**

Table 3: LJP results on CAIL. Among them, the underline scores are the state-of-the-art performances in LJP but lacking explainability, and the results in **bold** perform second only to NeurJudge but with an explainable rationale. Results of LJP baselines are quoted from Yue et al. (2021).

Methods	Charges				Law Articles				Terms of Penalty			
	Acc	MP	MR	F1	Acc	MP	MR	F1	Acc	MP	MR	F1
TopJudge	86.5	84.2	78.4	80.2	87.3	85.8	76.3	78.2	38.4	35.7	32.2	31.3
Few-Shot	88.2	87.5	80.6	82.0	88.4	86.8	77.9	79.5	39.6	37.1	30.9	31.6
LADAN	88.3	86.4	80.5	82.1	88.8	85.2	79.5	81.0	38.1	34.0	31.2	30.2
NeurJudge	89.9	87.8	86.8	87.0	90.4	87.2	85.8	86.1	41.7	40.4	37.2	37.3
RNP	85.1±0.2	82.2±0.3	78.1±0.5	79.0±0.5	86.5±0.1	82.1±0.3	77.7±0.7	78.8±0.6	37.1±0.2	30.1±0.3	30.3±0.4	27.7±0.3
HardKuma	86.2±0.2	84.7±1.2	79.0±0.8	80.6±0.4	86.8±0.8	83.3±1.7	77.3±0.8	78.9±0.9	35.8±0.1	34.3±1.3	27.3±0.3	25.8±0.4
INVRAT	85.4±0.2	83.3±0.1	78.7±0.3	80.2±0.3	85.1±0.1	83.1±0.1	76.2±0.1	78.1±0.2	38.5±0.3	34.5±0.7	33.0±0.2	32.0±0.5
Inter-RAT	**89.4±0.2**	**87.5±0.3**	**85.3±0.3**	**85.9±0.3**	**89.5±0.1**	**86.3±0.1**	**83.5±0.3**	**84.6±0.3**	**39.6±0.1**	**36.3±0.4**	**34.3±0.3**	**32.8±0.3**

Inter-RAT is agnostic to the structure of the *selector* and *predictor*, we adopt Bert Devlin et al. to replace bi-GRU in $f_e(\cdot)$ and $f_p(\cdot)$ in both RNP and Inter-RAT, and denote them as Bert_RNP and Bert_Inter-RAT, respectively. From the result, we observe that Bert_Inter-RAT still outperforms Bert_RNP, illustrating the effectiveness of Inter-RAT.

4.3 Legal Judgment Prediction

Since there are only two categories (*positive* and *negative*) in both beer and movie reviews prediction, we further generalize our Inter-RAT to the multi-classification task. Specifically, we focus on the Legal Judgment Prediction (LJP) task, which yields the judgment results such as the charges based on the case fact. We conduct experiments on publicly available datasets of the **C**hinese **AI** and **L**aw challenge (CAIL[5]). CAIL contains criminal cases consisting of the fact description and corresponding charges, law articles, and terms of penalty results. For data processing, referring to Yue et al. (2021), we remove several infrequent and multiple charges cases, and divide the terms into non-overlapping intervals. The detailed statistics of the datasets can be found in Yue et al. (2021). Figure 1 shows an example of LJP, which predicts the charge according to the case fact.

4.3.1 Comparison Methods and Experimental Setup

In addition to comparing RNP Lei et al. (2016), HardKuma Bastings et al. (2019) and INVRAT Chang et al. (2020), we also compare our method with some classical baselines in the LJP task, including TopJudge Zhong et al. (2018), Few-Shot Hu et al. (2018), LADAN Xu et al. (2020) and NeurJudge Yue et al. (2021). All the above baselines are trained by exploiting legal particularities. Among them, NeurJudge is the state-of-the-art model in LJP, which adopts different crime circumstances to yield corresponding results. Meanwhile, it employs a label embedding method to enhance the prediction. We conduct experiments on one of versions of CAIL containing 134,739 cases Yue et al. (2021). For testing, as there are no annotated rationales, we first employ the accuracy (Acc), macro-precision (MP), macro-recall (MR), and macro-F1 (F1) to evaluate the performance of yielding judgment results. Then, we provide a human evaluation for selected rationales in LJP. Detailed description of comparison methods and experimental setups can be found in Appendix C.2.

4.3.2 Experimental Results

To evaluate the performance of our model on LJP, we show the experimental results from two aspects. First, Table 3 shows that our Inter-RAT still performs better than the rationalization methods when generalizing to the multi-classification task. Meanwhile, compared with the LJP approaches (e.g. TopJudge and NeurJudge), even though our model is trained on the three subtasks separately, while these LJP approaches explore the dependencies between tasks and are trained with a multi-task learning framework, our model still achieves promising performance. However, Inter-RAT does not perform better than NeurJudge. A potential reason is that NeurJudge is designed only for LJP, exploiting the legal particularities well (e.g., crime circumstances). In contract, our Inter-RAT is designed for general text classification tasks. Therefore, the performance of Inter-RAT does not surpass NeurJudge. Furthermore, different from the NeurJudge and other LJP baselines, our Inter-RAT can provide an intuitive explanation (i.e., rationales) when yielding the judgment results while

[5] https://github.com/china-ai-law-challenge/CAIL2018

LJP baselines fail to produce them. The above observation provides strong validation of adopting the causal intervention method to remove spurious correlation in data for predicting results. Interestingly, we find there exists a minor difference between Inter-RAT and NeurJudge on yielding the charge and law article. We argue a potential reason is the label embedding method in NeurJudge can be approximated as the causal intervention method. We further discuss it in Appendix D.1 in detail.

Second, as CAIL does not provide annotated rationales like BeerAdvocate, we make a human evaluation to evaluate the performance of selected rationales. Specifically, we sample 100 examples and ask human annotators to evaluate rationales in the charge prediction. Besides, following Sha et al. (2021), we employ three metrics with an interval from 1 (lowest) to 5 (e.g. 2.0 and 3.2) to evaluate rationales, including usefulness (U), completeness (C), and fluency (F). Appendix C.3 describes detailed scoring standards for human annotators. The human evaluation results are shown in Table 4. From the results, we can find Inter-RAT outperforms RNP and INVRAT in all metrics, further demonstrating our causal intervention method can select more sufficient rationales for yielding results.

Table 4: Human evaluation on charge prediction.

Methods	Charge		
	U	C	F
RNP	3.85	3.28	3.34
INVRAT	3.88	3.42	3.41
Inter-RAT	**4.52**	**4.10**	**4.25**

5 Related Work

Rationalization. To improve the explainability of DNNs, the rationalization has attracted increasing attention Lei et al. (2016); Treviso & Martins (2020); Bastings et al. (2019); Chang et al. (2019); Yu et al. (2021). Specifically, Lei et al. (2016) first proposed a rationalization framework which consists of a *selector* and a *predictor*. Following this framework, multiple variants were proposed to improve rationalization. Among them, to replace the Bernoulli sampling distribution in Lei et al. (2016), Bastings et al. (2019) introduced a HardKuma distribution for reparameterized gradient estimates. And Paranjape et al. (2020) studied the Gumbel-softmax trick for reparameterization. Meanwhile, they also adopted the information bottleneck method to manage the trade-off between selecting sparse rationales and yielding accurate results. Additionally, another fundamental direction is adding external components to enhance the original framework. Yu et al. (2019) employed an introspective *selector* which incorporated the prediction results into the selection process. Some researchers Huang et al. (2021); Sha et al. (2021); Cao et al. (2020) proposed an external guider to reduce the difference between the distributions of rationales and input. However, few considered the spurious correlations in data which degraded the rationalization. Among them, Chang et al. (2020) discovered the causal rationales with environment invariant methods by creating different environments. Wu et al. (2022) extracted the rationales from the graph to study the explainability of graph neural networks (GNNs) by the intervention distributions Tian et al. (2006).

Causal Inference. Causal inference Glymour et al. (2016) has been widely explored in various fields, including medicine Richiardi et al. (2013) and politics Keele (2015), which aims to empower models the ability to achieve the causal effect. Recently, several researchesDeng & Zhang (2021); Dong et al. (2020); Yue et al. (2020) introduced causal inference into machine learning with causal intervention to remove the spurious correlations in data. Especially, it has inspired several studies in natural language understanding such as Named Entity Recognition Zhang et al. (2021), Topic modeling Wu et al. (2021), and Relation Extraction Liu et al. (2021). In this paper, we focus on improving the rationalization with causal intervention.

6 Conclusion

In this paper, we proposed a causal intervention method (Inter-RAT) to improve rationalization. To be specific, we first formulated the causalities in rationalization with a structural causal model and revealed how the confounder hurt the performance of selecting rationales with opened backdoor paths. Then, considering the entire label set as the observed confounder set, we introduced a backdoor adjustment method to remove spurious correlations between inputs and rationales, and between rationales and results. Besides, we further discussed the potential bias between selected rationales and predicted results caused by the sparsity constraints, and adopted the above causal intervention method to yield de-confounded prediction results. Experimental results on three real-world datasets have clearly demonstrated the effectiveness of our proposed method.

REFERENCES

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2019.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, (EMNLP), Online, November 16-20*, 2020.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. A game theoretic approach to class-wise selective rationalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. Invariant rationalization. In *Proceedings of the 37th International Conference on Machine Learning, (ICML)*, 2020.

Hanjie Chen and Yangfeng Ji. Learning variational word masks to improve the interpretability of neural text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 4236–4251, 2020.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Xiang Deng and Zhongfei Zhang. Comprehensive knowledge distillation with causal intervention. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020.

Zhang Dong, Zhang Hanwang, Tang Jinhui, Hua Xiansheng, and Sun Qianru. Causal intervention for weakly supervised semantic segmentation. In *NeurIPS*, 2020.

Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3478–3486. PMLR, 2019.

Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 487–498, 2018.

Yongfeng Huang, Yujun Chen, Yulun Du, and Zhilin Yang. Distribution matching for rationalization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI)*, 2021.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR)*. OpenReview. net, 2017.

Luke Keele. The statistics of causal inference: A view from political methodology. *Political Analysis*, 23(3):313–335, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. Element intervention for open relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (ACL/IJCNLP)*, 2021.

Julian J. McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *12th IEEE International Conference on Data Mining, (ICDM), Brussels, Belgium, December 10-13, 2012*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NeurIPS)*, pp. 3111–3119, 2013.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19: 2, 2000.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5):1511–1519, 2013.

Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. Learning from the best: Rationalizing prediction by adversarial information calibration. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Jin Tian, Changsung Kang, and Judea Pearl. A characterization of interventional distributions in semi-markovian causal models. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference (AAAI)*, 2006.

Marcos V. Treviso and André F. T. Martins. The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2020.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10760–10770, 2020.

Xiaobao Wu, Chunping Li, and Yishu Miao. Discovering topics in long-tailed corpora with causal intervention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 175–185, 2021.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat seng Chua. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. Cail2018: A large-scale legal dataset for judgment prediction. In *arXiv preprint arXiv:1807.02478*, 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, (ICML)*, 2015.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3086–3095. Association for Computational Linguistics, 2020.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. Understanding interlocking dynamics of cooperative rationalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NeurIPS*, 2020.

Omar Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 conference on Empirical methods in natural language processing*, pp. 31–40, 2008.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. De-biasing distantly supervised named entity recognition via causal intervention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

A INSTANTIATED STRUCTURE CAUSAL MODEL

A.1 SCM FOR THE SELECTOR

In this section, we describe SCM for the *selector* (Figure 2(b)) in detail with examples:

$C \rightarrow X$. The context prior C determines which tokens can be "put" into the text input X. Among them, the context prior consists of unobserved prior and observed prior (such as the label set). For example, in Figure 1, both observed *Intentional homicide* and *Manslaughter* priors decide where the tokens denoting violence and death appear ; the *Manslaughter* prior determines where the tokens representing manslaughter appear ; the unobserved prior decides where other tokens that are meaningless appear.

$X \rightarrow K \leftarrow C$. K denotes the context-specific representation which is a weighted representation of the prior knowledge associated with X in C. Taking Figure 1 as an example, we assume that the label set consisting of *Intentional homicide*, *Manslaughter*, and *Theft* is the observed prior. Then, we can get the context prior which consists of four parts (i.e., the *Intentional homicide* prior c_1, the *Manslaughter* prior c_2, the *Theft* prior c_3 and the unobserved prior c_4). Next, we calculate the association between X and C, and obtain the corresponding scores, assuming $a_1 = 0.3$, $a_2 = 0.6$, $a_3 = 0.0$, $a_4 = 0.1$, where *Manslaughter* prior c_2 and X are the most relevant, and the *Theft* prior c_3 and X are the least relevant. Finally, we can calculate the K as $a_1c_1 + a_2c_2 + a_3c_3 + a_4c_4 = 0.3c_1 + 0.6c_2 + 0.1c_4$.

$X \rightarrow R \leftarrow K$. As the rationale R is a subsequence of X, $X \rightarrow R$ holds. Besides, $K \rightarrow R$ represents the contextual constitution of the text that affects the composition of rationales. Taking the previous example as an example, since K is calculated as $0.3c_1 + 0.6c_2 + 0.1c_4$, tokens in R will be more inclined with the *Manslaughter* prior c_2.

A.2 SCM FOR THE PREDICTOR

In this section, we describe the detailed causal relationship (Figure 6(a)) between the selected rationales R and the results Y in the *predictor* at a high-level:

$C \rightarrow R$. Based on Eq (5) in section 3.2, we can conclude that the context prior C determines which tokens are corresponding to rationales.

$R \rightarrow H \leftarrow C$. H represents the context-specific representation of R by using the context prior C.

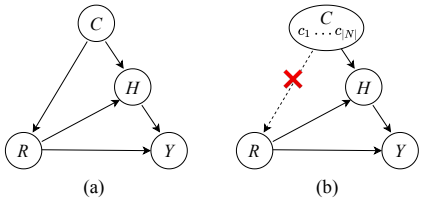

Figure 6: SCM for the *predictor*.

$R \rightarrow Y \leftarrow H$. As rationales R are selected by α-constraint and consist of real rationales and shortcuts tokens, we argue that $R \rightarrow Y$ holds. Besides, the context prior C affects the label Y by the mediator H. The reason is similar to $K \rightarrow R$ in Appendix A.1.

From the graph, we can clearly see the context prior C is the confounder between R and Y, which opens the backdoor path $R \leftarrow C \rightarrow H \rightarrow Y$. Therefore, to remove the spurious correlations between R and Y, we adopt the causal intervention method to calculate $P(Y|do(R))$ by cutting the link $C \rightarrow R$ (Figure 6(b)):

$$
\begin{aligned}
P(Y|do(R)) &= \sum_{i=1}^{|N|} \left[P\left(Y|R, H = g_r\left(R, c_i\right)\right) P\left(c_i\right)\right] \\
&= \sum_{i=1}^{|N|} \left[P\left(Y|f_r(R, h_i)\right) P\left(c_i\right)\right] \\
&= \sum_{i=1}^{|N|} \left[\text{softmax}(f_r(R, h_i)) P\left(c_i\right)\right],
\end{aligned}
\tag{8}
$$

where $f_r(\cdot)$ is the function achieving $R \rightarrow Y \leftarrow H$, $h_i = g_r(R, c_i) = \beta_i c_i$, $\beta_i \in \beta$. $\beta \in \mathbb{R}^N$ is the set of the normalized similarity between R and each c_i in the confounder set C (i.e., $\beta_{=}\text{softmax}(f_p(R)D_c^T)$). Among them, $f_p(\cdot)$ encodes R into a d-dimensional vector, and the dictionary $D_c = \left\{c_1, c_2, \ldots, c_{|N|}\right\}$ is approximated as the observed confounder.

Besides, we also adopt the NWGM approximation to Eq (8) and set $f_r(R, h_i)$ as a linear model (i.e., $f_r(R, h_i) = W_3 f_p(R) + W_4 h_i = W_3 f_p(R) + W_4 \beta_i c_i$). Then, the final objective of intervention is formulated as:

$$\begin{aligned} P(Y|do(R)) &= \sum_{i=1}^{|N|} [\text{softmax}(f_r(R, h_i)) P(c_i)] \\ &\approx \text{softmax}(\sum_{i=1}^{|N|} f_r(R, h_i) P(c_i)) \\ &= \text{softmax}(W_3 f_p(R) + W_4 \sum_{i=1}^{|N|} \beta_i c_i P(c_i)). \end{aligned} \tag{9}$$

B Explanations for α-constraint

B.1 Examples for α-constraint

As mentioned in section 3.3, in the practical, a higher predefined sparsity level α may bring shortcuts tokens which hurt the prediction performance, an extreme example being that all tokens in the text input will be selected. Below, we take a beer review as an example to further illustrate this problem, where this example is adopted to predict scores of the smell aspect.

the original text : *He thinks this beer smells great and tastes terrific .*

rationale : *smells great*

rationale with shortcuts tokens, where α is set to 0.5: *smells great and tastes terrific*. Among them, *"and tastes terrific"* can be considered as shortcuts tokens.

B.2 Discussions on α-constraint

Although several rationalization methods do not set α-constraint to extract rationales, we believe that their methods of constraining the short rationales extraction can be considered a variant of α-constraint, and our intervention method in section 3.3 will still be effective on these methods. Specifically, we argue that these methods should set hyperparameters to encourage the model to select short rationales. However, if the hyperparameters are not set properly, it is possible that more shortcuts tokens will be extracted, making R and Y confounded. For example, for several methods Chen & Ji (2020); Paranjape et al. (2020) adopting the information bottleneck to ensure the model extracts short rationales, there exists a KL divergence between the posterior distribution $P(\widetilde{m}_j|x_j)$ and the prior distribution $r(\widetilde{m}_j)$, where $r(\widetilde{m}_j) = \text{Bernoulli}(\pi)$ for some constant $\pi \in (0, 1)$. For instance, if we set π as 0.1, it means we encourage the model to extract 10% of the input text. Therefore, we consider π as a variant of α proposed by us.

C Setting Details

C.1 Statistics of BeerAdvocate and MovieReview

In this section, we show the detailed statistics of BeerAdvocate and MovieReview in Table 5. Among them, BeerAdvocate contains three aspects, including appearance, aroma and palate. From the Table 5, we can observe that the processed BeerAdvocate is a non-balanced dataset. In the training set, the prior distribution is *positive*:*negative* $\approx$ 20:1 in appearance, *positive*:*negative* $\approx$ 17:3 in aroma, *positive*:*negative* $\approx$ 17:3 in palate. Meanwhile, MovieReview is a balanced dataset with *positive*:*negative* = 1:1.

C.2 Comparison methods and experimental setups for LJP

In addition to comparing RNP Lei et al. (2016), HardKuma Bastings et al. (2019) and INVRAT Chang et al. (2020), we also compare our method with some classical baselines in the LJP task:

Table 5: Detailed statistics of the processed BeerAdvocate and MovieReview.

Dataset	Train		Dev		Test	
	# Pos	# Neg	# Pos	# Neg	# Pos	# Neg
BeerAdvocate (appearance)	202,385	12,897	28,488	1,318	955	14
BeerAdvocate (aroma)	172,299	30,564	24,494	3,396	913	31
BeerAdvocate (palate)	176,038	27,639	24,837	3,203	940	24
MovieReview	800	800	100	100	99	100

- **TopJudge** Zhong et al. (2018) explores the dependencies among the three subtasks in LJP.
- **Few-Shot** Hu et al. (2018) utilizes the charge attributes to identify the confusing charges.
- **LADAN** Xu et al. (2020) learns the distinguished law articles representations for LJP prediction.
- **NeurJudge** Yue et al. (2021) is a circumstance aware approach adopting different crime circumstances to yield corresponding results. Meanwhile, it employs a label embedding method to enhance the prediction.

For training, we adopt the word2vec Mikolov et al. (2013) for word embedding pre-training with size 200, and set the encoder in $f_e(\cdot)$ and $f_p(\cdot)$ as Bi-GRU. Besides, we implement the learning rate of 0.0002 with batch size 256, and take α as 0.2. For evaluating, we employ the accuracy (Acc), macro-precision (MP), macro-recall (MR), and macro-F1 (F1) to evaluate the performance of yielding judgment results.

C.3 Scoring standards for human evaluation

Following Sha et al. (2021), we evaluate the rationales with three metrics: usefulness (U), completeness (C), and fluency (F) in the charge prediction. Among them, each scored from 1 (lowest) to 5. Below, we introduce scoring standards for the above metrics in brief. Detailed standards for human annotators can be found in Sha et al. (2021).

C.3.1 Usefulness

Q: Do you think the selected rationales can be useful for explaining the predicted labels?

- 5: Exactly. Selected rationales are useful for me to get the correct label.
- 4: Highly useful. Although several tokens have no relevance to correct label, most selected tokens are useful to explain the labels.
- 3: Half of them are useful. About half of the tokens are useful for getting labels.
- 2: Almost useless. Almost all of the tokens are useless.
- 1: No Use. The selected rationales are useless for identifying labels.

C.3.2 completeness

Q: Do you think the selected rationales are enough for explaining the predicted labels?

- 5: Exactly. Selected rationales are enough for me to get the correct label.
- 4: Highly complete. Several tokens related to the label are missing.
- 3: Half complete. There are still some important tokens that have not been selected, and they are in nearly the same number as the selected tokens.
- 2: Somewhat complete. The selected tokens are not enough.
- 1: Nonsense. None of the important tokens is selected.

C.3.3 FLUENCY

Q: Do you think the selected rationales are fluent?

- 5: Very fluent.
- 4: Highly fluent.
- 3: Partial fluent.
- 2: Very unfluent.
- 1: Nonsense.

D MORE EXPERIMENTAL RESULTS

D.1 A CAUSAL VIEW ON NEURJUDGE

In this section, from the causal view, we discuss the reason why the difference between Inter-RAT and NeurJudge on the charge and law article prediction is not significant. Here, we explain the observation by taking the charge prediction as an example, and the article prediction is similar. Specifically, NeurJudge adopts a label embedding method to incorporate the semantics of charge into the case fact to yield the corresponding result. We argue that this method can be approximated as the causal intervention method. To illustrate this discovery, we assume Figure 2(d) is the SCM of the charge prediction task, and consider the case fact as R (i.e., α=1) and the charge label set as C. Then the process of label embedding can be formulated as $R \rightarrow H \leftarrow C$ and $R \rightarrow Y \leftarrow H$. The objective of NeurJudge is written as:

$$P(Y|R) = \text{softmax}(\sum_{i=1}^{|N|} f_{neru}(R, h_i)), \tag{10}$$

where $h_i = g_{neru}(R, c_i)$. We can find the difference between the Eq (10) and our causal intervention method is that Eq (10) ignores the prior distribution $P(c_i)$. It is worth noting that although NeurJudge ignores $P(c_i)$, it performs slightly better than our model. A potential reason is that NeurJudge exploits the dependencies among LJP tasks well, while our model is trained on the independent task.

D.2 VISUALIZATION

We provide several visualization cases in CAIL dataset as shown in Figure 7, Figure 8 and Figure 9, which are selected by Inter-RAT and INVRAT, respectively. Among them, annotated rationales are underlined. Inter-RAT and INVRAT rationales are highlighted in **pink** and **green** colors, respectively.

CAIL - Law Article	*Label - Law Article 233*
The defendant was **driving a tricycle without a driver's licence** to deliver walnuts to a certain place. When he drove into the door of the victim's house, **he collapsed the gate pier of the victim's house, causing the victim to be injured** by the collapsed gate pier, **and he died** after being rescued.…	
The defendant was driving a tricycle without a driver's licence to deliver walnuts to a certain place. When he drove into the door of the victim's house, **he collapsed the gate pier of the victim's house, causing the victim to be injured** by the collapsed gate pier, **and he died** after being rescued.…	

Figure 7: Examples of selective rationalization on the law article prediction. Among them, INVRAT fails to predict the law article and consider it as *Law Article 133* with capturing the tokens *driving a tricycle without a driver's licence*. On the contrary, Inter-RAT predicts the law article correctly.

CAIL - Charge *Label - Manslaughter*

The defendant and the victim were both students, after the dormitory relocation, in the new dormitory, the defendant and the victim **for sleeping in the lower bunk bed dispute**, **the defendant picked up a bottle and** forcefully **smashed the victim's head, resulting in** the victim's **head injury**, then the defendant **sent the victim to hospital**. **The victim died** in hospital treatment failed. The forensic medical appraisal showed the injury to the victim's face was a pre-existing injury and the degree of injury was minor. The victim's death was consistent with an acute heart attack where trauma and emotional stress as precipitating factors…

The defendant and the victim were both students, after the dormitory relocation, in the new dormitory, the defendant and the victim **for sleeping in the lower bunk bed dispute**, **the defendant picked up a bottle and forcefully smashed the victim's head**, resulting in the victim's **head injury**, then **the defendant sent the victim to hospital**. **The victim died** in hospital treatment failed. The forensic medical appraisal showed the injury to the victim's face was a pre-existing injury and the degree of **injury was minor**. **The victim's death was consistent with an acute heart attack** where trauma and emotional stress as precipitating factors…

Figure 8: Examples of selective rationalizations on the charge prediction. Although both Inter-RAT and INVRAT predict the charge correctly, Inter-RAT can extract more comprehensive rationales (i.e., *The victim's death was consistent with an acute heart attack*), which support the victim's death was due to the negligence.

CAIL – Term of Penalty *Label - A fixed-term imprisonment of two years*

The trial found that the defendant saw the victim, who was a waiter in the hotel, when he was preparing to go on duty in the lobby of the xx hotel, and he kicked the victim in the buttocks, and a fight broke out between the two. In the course of the fight, **the defendant pulled out the folding fruit knife he was carrying and stabbed the victim** in the inner right thigh. After seeing that the victim‘s leg was bleeding profusely, the victim and his colleague finished sending the victim to the hospital, where it was determined that **the victim suffered minor injuries**. The defendant then **surrendered to the police and made a truthful confession to the crime**.....

The trial found that the defendant saw the victim, who was a waiter in the hotel, when he was preparing to go on duty in the lobby of the xx hotel, and he kicked the victim in the buttocks, and a fight broke out between the two. In the course of the fight, **the defendant pulled out the folding fruit knife he was carrying and stabbed the victim in the inner right thigh**. After seeing that the victim‘s leg was bleeding profusely, the victim and his colleague finished **sending the victim to the hospital**, where it was determined that the victim **suffered minor injuries**. The defendant **then surrendered to the police and made a truthful confession to the crime**.....

Figure 9: Examples of selective rationalization on the term of penalty prediction. Among them, both Inter-RAT and INVRAT predict the term of penalty correctly, but Inter-RAT extracts more plausible rationales (i.e., *sending the victim to the hospital* which is important for sentencing.).