Invariant Rationalization

Shiyu Chang [*1] **Yang Zhang** [*1] **Mo Yu** [*2] **Tommi S. Jaakkola** [3]

Abstract

Selective rationalization improves neural network interpretability by identifying a small subset of input features — the rationale — that best explains or supports the prediction. A typical rationalization criterion, *i.e.* maximum mutual information (MMI), finds the rationale that maximizes the prediction performance based only on the rationale. However, MMI can be problematic because it picks up spurious correlations between the input features and the output. Instead, we introduce a game-theoretic invariant rationalization criterion where the rationales are constrained to enable the same predictor to be optimal across different environments. We show both theoretically and empirically that the proposed rationales can rule out spurious correlations and generalize better to different test scenarios. The resulting explanations also align better with human judgments. Our implementations are publicly available at https://github.com/code-terminator/invariant_rationalization.

1. Introduction

A number of selective rationalization techniques (Lei et al., 2016; Li et al., 2016b; Chen et al., 2018a;b; Yu et al., 2018; Carton et al., 2018; Bastings et al., 2019; Yu et al., 2019; Chang et al., 2019) have been proposed to explain the predictions of complex neural models. The key idea driving these methods is to find a small subset of the input features – *rationale* – that suffices on its own to yield the same outcome. In practice, rationales that remove much of the spurious content from the input, *e.g.*, text, could be used and examined as justifications for model's predictions.

The most commonly-used criterion for rationales is the maximum mutual information (MMI) criterion. In the context of natural language processing (NLP), it defines rationale as the subset of input text that maximizes the mutual information between the subset and the model output, subject to the constraint that the selected subset remains within a prescribed length. Specifically, if we denote the random variables corresponding to input as $\boldsymbol{X}$, rationales as $\boldsymbol{Z}$ and the model output as Y, then the MMI criterion finds the explanation $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{X})$ that yields the highest prediction accuracy of Y.

MMI criterion can nevertheless lead to undesirable results in practice. It is prone to highlighting spurious correlations between the input features and the output as valid explanations. While such correlations represent statistical relations present in the training data, and thus incorporated into the neural model, the impact of such features on the true outcome (as opposed to model's predictions) can change at test time. In other words, MMI may select features that do not explain the underlying relationship between the inputs and outputs even though they may still be faithfully reporting the model's behavior. We seek to modify the rationalization criterion to better tailor it to find causal features.

As an example, consider figure 1 that shows a beverage review which covers four aspects of beer: *appearance*, *smell*, *palate*, and *overall*. The reviewers also assigned a score to each of these aspects. Suppose we want to find an explanation supporting a positive score to *smell*. The correct explanation should be the portion of the review that actually discusses smell, as highlighted in green. However, reviews for other aspects such as *palate* (highlighted in red) may co-vary with *smell* score since, as senses, smell and palate are related. The overall statement as highlighted in blue would typically also clearly correlate with any individual aspect score, including *smell*. Taken together, sentences highlighted in green, red and blue would all be highly correlated with the positive score for *smell*. As a result, MMI may select any one of them (or some combination) as the rationale, depending on precise statistics in the training data. Only the green sentence constitutes an adequate explanation.

Our goal is to design a rationalization criterion that approximates finding causal features. While assessing causality is challenging, we can approximate the task by searching instead features that are in some sense *invariant*. This notion was recently introduced in the context of invariant risk

[*]Equal contribution [1]MIT-IBM Watson AI Lab [2]IBM Research [3]CSAIL MIT. Correspondence to: Shiyu Chang <shiyu.chang@ibm.com>, Yang Zhang <yang.zhang2@ibm.com>, Mo Yu <yum@us.ibm.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Beer - Smell Label - Positive

375ml corked and caged bottle with bottled on date november 30 2005 , poured into snifter at brouwer 's , reviewed on 5/15/11 . aroma : pours a clear golden color with orange hues and a whitish head that leaves some lacing around glass . smell : **lots of barnyaardy funk with tons of earthy aromas , grass and some lemon peel** . palate : **similar to the aroma , lots of funk , lactic sourness , really earthy with citrus notes and oak . many layers of intriguing earthy complexities** . overall : **very funky and earthy gueuze , nice and crisp with good drinkability** .

Figure 1. An example beer review and possible rationales explaining why the score on the smell aspect is positive. **Green highlights** the review on the smell aspect, which is the true explanation. **Red highlights** the review on the taste aspect, which has a high correlation with the smell. **Blue highlights** the overall review, which summarizes all the aspects, including smell. All three sentences have high predictive powers of the smell score, but only the green sentence is the desired explanation.

minimization (IRM) (Arjovsky et al., 2019). The main idea is to highlight spurious (non-causal) variation by dividing the data into different environments. The same predictor, if based on causal features, should remain optimal in each environment separately.

In this paper, we propose invariant rationalization (INVRAT), a novel rationalization scheme that incorporates the invariance constraint. We extend the IRM principle to neural predictions by resorting to a game-theoretic framework to impose invariance. Specifically, the proposed framework consists of three modules: a rationale generator, an environment-agnostic predictor as well as an environment-aware predictor. The rationale generator generates rationales $\boldsymbol{Z}$ from the input $\boldsymbol{X}$, and both predictors try to predict Y from $\boldsymbol{Z}$. The only difference between the two predictors is that the environment-aware predictor also has access to which environment each training data point is drawn from. The goal of the rationale generator is to restrict the rationales in a manner that closes the performance gap between the two predictors while still maximizing the prediction accuracy of the environment-agnostic predictor.

We show theoretically that INVRAT can solve the invariant rationalization problem, and that the invariant rationales generalize well to unknown test environments in a well-defined minimax sense. We evaluate INVRAT on multiple datasets with false correlations. The results show that INVRAT does significantly better in removing false correlations and finding explanations that better align with human judgments. Both data and code will become publicly available.

2. Preliminaries: MMI and Its Limitation

In this section, we will formally review the MMI criterion and analyze its limitation using a probabilistic model. Throughout the paper, upper-cased letters, X and $\boldsymbol{X}$, denote random scalars and vectors respectively; lower-cased letters, x and $\boldsymbol{x}$, denote deterministic scalars and vectors respectively; $H(\boldsymbol{X})$ denotes the Shannon entropy of $\boldsymbol{X}$; $H(Y|\boldsymbol{X})$ denotes the entropy of Y conditional on $\boldsymbol{X}$; $I(Y;\boldsymbol{X})$ denotes the mutual information. Without causing ambiguities, we use $p_{\boldsymbol{X}}(\cdot)$ and $p(\boldsymbol{X})$ interchangeably to denote the probabilistic mass function of $\boldsymbol{X}$.

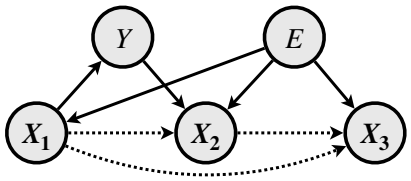

Figure 2. A probabilistic model illustrating different parts of an input that have different probabilistic relationships with the model output Y. A sentence $\boldsymbol{X}$ can be divided into three variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$. All $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ can be highly correlated with Y, but only $\boldsymbol{X}_1$ is regarded as a plausible explanation.

2.1. Maximum Mutual Information Criterion

The MMI objective can be formulated as follows. Given the input-output pairs $(\boldsymbol{X}, Y)$, MMI aims to find a rationale $\boldsymbol{Z}$, which is a masked version of $\boldsymbol{X}$, such that it maximizes the mutual information between $\boldsymbol{Z}$ and Y. Formally,

$$\max_{\boldsymbol{m}\in\mathcal{S}} I(Y;\boldsymbol{Z}) \quad \text{s.t. } \boldsymbol{Z} = \boldsymbol{m}\odot\boldsymbol{X}, \tag{1}$$

where $\boldsymbol{m}$ is a binary mask and $\mathcal{S}$ denotes a subset of $\{0,1\}^N$ with a sparsity and a continuity constraints. N is the total length in $\boldsymbol{X}$. We leave the exact mathematical form of the constraint set $\mathcal{S}$ abstract here, and it will be formally introduced in section 3.5. $\odot$ denotes the element-wise multiplication of two vectors or matrices. Since the mutual information measures the predictive power of $\boldsymbol{Z}$ on Y, MMI essentially tries to find a subset of input features that can best predict the output Y.

2.2. MMI Limitations

The biggest problem of MMI is that it is prone to picking up spurious probabilistic correlations, rather than finding the causal explanation. To demonstrate why this is the case, consider a probabilistic graph in figure 2, where $\boldsymbol{X}$ is divided into three variables, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$, which represents the three typical relationship with Y: $\boldsymbol{X}_1$ influences Y; $\boldsymbol{X}_2$

is influenced by Y; $\boldsymbol{X}_3$ has no direction connections with Y. The dashed arrows represent some additional probabilistic dependencies among $\boldsymbol{X}$. For now, we ignore E.

As observed from the graph, $\boldsymbol{X}_1$ serves as the valid explanation of Y, because it is the true cause of Y. Neither $\boldsymbol{X}_2$ nor $\boldsymbol{X}_3$ are valid explanations. However, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ can all be highly predicative of Y, so the MMI criterion may select any of the three features as the rationale. Concretely, consider the following toy example with all binary variables. Assume $p_{\boldsymbol{X}_1}(1) = 0.5$, and

$$p_{Y|\boldsymbol{X}_1}(1|1) = p_{Y|\boldsymbol{X}_1}(0|0) = 0.9, \tag{2}$$

which makes $\boldsymbol{X}_1$ a good predictor of Y. Next, define the conditional prior of $\boldsymbol{X}_2$ as

$$p_{\boldsymbol{X}_2|Y}(1|1) = p_{\boldsymbol{X}_2|Y}(0|0) = 0.9.$$

According to the Bayes rule,

$$p_{Y|\boldsymbol{X}_2}(1|1) = p_{Y|\boldsymbol{X}_2}(0|0) = 0.9, \tag{3}$$

which makes $\boldsymbol{X}_2$ also a good predictor of Y. Finally, assume the conditional prior of $\boldsymbol{X}_3$ is

$$p_{\boldsymbol{X}_3|\boldsymbol{X}_1,\boldsymbol{X}_2}(1|1,1) = p_{\boldsymbol{X}_3|\boldsymbol{X}_1,\boldsymbol{X}_2}(0|0,0) = 1, \text{ and}$$
$$p_{\boldsymbol{X}_3|\boldsymbol{X}_1,\boldsymbol{X}_2}(1|0,1) = p_{\boldsymbol{X}_3|\boldsymbol{X}_1,\boldsymbol{X}_2}(1|1,0) = 0.5.$$

It can be computed that

$$p_{Y|\boldsymbol{X}_3}(1|1) = p_{Y|\boldsymbol{X}_3}(0|0) = 0.9. \tag{4}$$

In short, according to equations (2), (3) and (4), we have constructed a set of priors such that the predictive power of $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ is *exactly the same*. As a result, there is no reason for MMI to favor $\boldsymbol{X}_1$ over the others.

In fact, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ correspond to the three highlighted sentences in figure 1. $\boldsymbol{X}_1$ corresponds to the smell review (green sentence), because it represents the true explanation that influences the output decision. $\boldsymbol{X}_2$ corresponds to the overall review (blue sentence), because the overall summary of the beer inversely influenced by the smell score. Finally, $\boldsymbol{X}_3$ corresponds to the palate review (red sentence), because the palate review does not have a direct relationship with the smell score. However, $\boldsymbol{X}_3$ may still be highly predicative of Y because it can be strongly correlated with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Therefore, we need to explore a novel rationalization scheme that can distinguish $\boldsymbol{X}_1$ from the rest.

3. Adversarial Invariant Rationalization

In this section, we propose invariant rationalization, a rationalization criterion that can exclude rationales with spurious correlations, utilizing the extra information provided by an environment variable. We will introduce INVRAT, a game-theoretic approach to solving the invariant rationalization problem. We will then theoretically analyze the convergence property and the generalizability of invariant rationales.

3.1. Invariant Rationalization

Without further information, distinguishing $\boldsymbol{X}_1$ from $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ is a challenging task. However, this challenge can be resolved if we also have access to an extra piece of information: the environment. As shown in figure 2, an environment is defined as an instance of the variable E that impacts the prior distribution of $\boldsymbol{X}$ (Arjovsky et al., 2019). On the other hand, we make the same assumption as in IRM that the $p(Y|\boldsymbol{X}_1)$ remains the same across the environments (hence there is no edge pointing from E to Y in figure 2), because $\boldsymbol{X}_1$ is the true cause of Y. A general guidance on how to choose the environments is presented in appendix A. As we will show soon, $p(Y|\boldsymbol{X}_2)$ and $p(Y|\boldsymbol{X}_3)$ will *not* remain the same across the environments, which distinguishes $\boldsymbol{X}_1$ from $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$.

Back to the binary toy example in section 2.2, suppose there are two environments, e_1 and e_2. In environment e_1, all the prior distributions are exactly the same as in section 2.2. In environment e_2, the priors are almost the same, except for the prior of $\boldsymbol{X}_1$. For notation ease, define $q_{\boldsymbol{X}}(\cdot)$ as the probabilities under environment e_2, *i.e.* $p_{\boldsymbol{X}|E}(\cdot|e_2)$. Then, we assume that

$$q_{\boldsymbol{X}_1}(1) = 0.6.$$

It turns out that such a small difference suffices to expose $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$. In this environment, $q(Y|\boldsymbol{X}_1)$ is the same as in equation (2) as assumed. However, it can be computed that

$$q_{Y|\boldsymbol{X}_2}(1|1) \approx 0.926, \quad q_{Y|\boldsymbol{X}_2}(0|0) \approx 0.867,$$
$$q_{Y|\boldsymbol{X}_3}(1|1) \approx 0.912, \quad q_{Y|\boldsymbol{X}_3}(0|0) \approx 0.883,$$

which are different from equations (3) and (4). Notice that we have not yet assumed any changes in the priors of $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$, which will introduce further differences. The fundamental cause of such differences is that Y is independent of E *only when* conditioned on $\boldsymbol{X}_1$, so $p_{Y|\boldsymbol{X}_1}(\cdot|\cdot)$ would not change with E. We call this property *invariance*. However, the conditional independence does not hold for $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$.

Therefore, given that we have access to multiple environments during training, *i.e.* multiple instances of E, we propose the invariant rationalization objective as follows:

$$\max_{\boldsymbol{m}\in\mathcal{S}} I(Y;\boldsymbol{Z}) \quad \text{s.t. } \boldsymbol{Z} = \boldsymbol{m}\odot\boldsymbol{X}, \quad Y \perp E \mid \boldsymbol{Z}, \tag{5}$$

where $\perp$ denotes probabilistic independence. The only difference between equations (1) and (5) is that the latter has the invariance constraint, which is used to screen out $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$. In practice, finding an eligible environment is feasible. In the beer review example in figure 1, a possible choice of environment is the brand of beer, because different beer brands have different prior distributions of the review in each aspect – some brands are better at the appearance, others better at the palate. Such variations in priors suffice to expose the non-invariance of the palate review or the overall review in terms of predicting the smell score.

3.2. The INVRAT Framework

The constrained optimization in equation (5) is hard to solve in its original form. INVRAT introduces a game-theoretic framework, which can approximately solve this problem. Notice that the invariance constraint can be converted to a constraint on entropy, *i.e.*,

$$Y \perp E \mid \boldsymbol{Z} \Leftrightarrow H(Y|\boldsymbol{Z}, E) = H(Y|\boldsymbol{Z}), \tag{6}$$

which means if $\boldsymbol{Z}$ is invariant, E cannot provide extra information beyond $\boldsymbol{Z}$ to predict Y. Guided by this perspective, INVRAT consists of three players, as shown in figure 3:

- an environment-agnostic/-independent predictor $f_i(\boldsymbol{Z})$;
- an environment-aware predictor $f_e(\boldsymbol{Z}, E)$; and
- a rationale generator, $g(\boldsymbol{X})$.

The goal of the environment-agnostic and environment-aware predictors is to predict Y from the rationale $\boldsymbol{Z}$. The only difference between them is that the latter has access to E as another input feature but the former does not. Formally, denote $\mathcal{L}(Y; f)$ as the cross-entropy loss on a single instance. Then the learning objective of these two predictors can be written as follows.

$$\mathcal{L}_i^* = \min_{f_i(\cdot)} \mathbb{E}[\mathcal{L}(Y; f_i(\boldsymbol{Z}))],\ \mathcal{L}_e^* = \min_{f_e(\cdot,\cdot)} \mathbb{E}[\mathcal{L}(Y; f_e(\boldsymbol{Z}, E))], \tag{7}$$

where $\boldsymbol{Z} = g(\boldsymbol{X})$. The rationale generator generates $\boldsymbol{Z}$ by masking $\boldsymbol{X}$. The goal of the rationale generator is also to minimize the invariance prediction loss $\mathcal{L}_i^*$. However, there is an additional goal to make the gap between $\mathcal{L}_i^*$ and $\mathcal{L}_e^*$ small. Formally, the objective of the generator is as follows:

$$\min_{g(\cdot)} \mathcal{L}_i^* + \lambda h(\mathcal{L}_i^* - \mathcal{L}_e^*), \tag{8}$$

where $h(t)$ a convex function that is monotonically increasing in t when $t < 0$, and *strictly* monotonically increasing in t when $t \geq 0$, *e.g.*, $h(t) = t$ and $h(t) = \text{ReLU}(t)$.

3.3. Convergence Properties

This section justifies that equations (7) and (8) can solve equation (5) in its Lagrangian form. If the representation power of $f_i(\cdot)$ and $f_e(\cdot, \cdot)$ is sufficient, the cross-entropy loss can achieve its entropy lower bound, *i.e.*,

$$\mathcal{L}_i^* = H(Y|\boldsymbol{Z}), \quad \mathcal{L}_e^* = H(Y|\boldsymbol{Z}, E).$$

Notice that the environment-aware loss should be no greater than the environment-agnostic loss, because of the availability of more information, *i.e.*, $H(Y|\boldsymbol{Z}) \geq H(Y|\boldsymbol{Z}, E)$. Therefore, the invariance constraint in equation (6) can be rewritten as an inequality constraint:

$$H(Y|\boldsymbol{Z}) = H(Y|\boldsymbol{Z}, E) \Leftrightarrow H(Y|\boldsymbol{Z}) \leq H(Y|\boldsymbol{Z}, E). \tag{9}$$

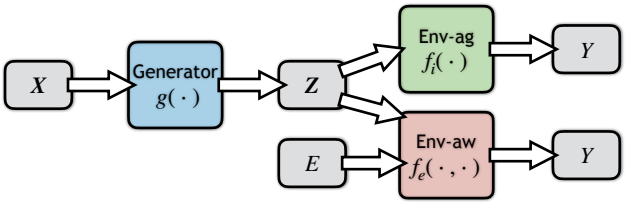

Figure 3. The INVRAT framework with three players: the rationale generator, environment-agnostic and -aware predictors.

Finally, notice that $I(Y; \boldsymbol{Z}) = H(Y) - H(Y|\boldsymbol{Z})$. Thus the objective in equation (8) can be regarded as the Lagrange form of equation (5), with the constraint rewritten as an inequality constraint

$$h(H(Y|\boldsymbol{Z}) - H(Y|\boldsymbol{Z}, E)) \leq h(0). \tag{10}$$

According to the KKT conditions, $\lambda > 0$ when equation (10) is binding. Moreover, the objectives in equations (7) and (8) can be rewritten as a minimax game

$$\min_{g(\cdot), f_i(\cdot)} \max_{f_e(\cdot,\cdot)} \mathcal{L}_i(g, f_i) + \lambda h(\mathcal{L}_i(g, f_i) - \mathcal{L}_e(g, f_e)), \tag{11}$$

where

$$\mathcal{L}_i(g, f_i) = \mathbb{E}[\mathcal{L}(Y; f_i(\boldsymbol{Z}))],\ \ \mathcal{L}_e(g, f_e) = \mathbb{E}[\mathcal{L}(Y; f_e(\boldsymbol{Z}, E))].$$

Therefore, the generator plays a co-operative game with the environment-agnostic predictor, and an adversarial game with the environment-aware predictor. The optimization can be performed using alternate gradient descent/ascent.

3.4. Invariance and Generalizability

In our previous discussions, we have justified the invariant rationales in the sense that it can uncover consistent and causal explanations and leave out spurious statistical correlations. In this section, we further justify invariant rationale in terms of generalizability. We consider two sets of environments, a set of training environments $\{e_t\}$ and a test environment e_a. Only the training environments are accessible during training. The prior distributions in the test environment are completely unknown. The question we want to ask is: does keeping the invariant rationales and dropping the non-invariant rationales improve the generalizability in the unknown test environment?

Assume that 1) the training data are sufficient, 2) the predictor is environment-agnostic, 3) the predictor has sufficient representation power, and 4) the training converges to the global optimum. Under these assumptions, any predictor is able to replicate the training set distribution (with all the training environments mixed) $p(Y|\boldsymbol{Z}, E \in \{e_t\})$, which is optimal under the cross-entropy training objective. In the test environment e_a, the cross-entropy loss of this predictor

is given by

$$\mathcal{L}^*_{\text{test}}(\boldsymbol{Z}) = H(p(Y|\boldsymbol{Z}, e_a); p(Y|\boldsymbol{Z}, \{e_t\})).$$

where $p(Y|\boldsymbol{Z}, \{e_t\})$ is short for $p(Y|\boldsymbol{Z}, E \in \{e_t\})$. $\mathcal{L}^*_{\text{test}}(\boldsymbol{Z})$ cannot be evaluated because the prior distribution in the test environment is unknown. Instead, we consider the worst scenario. For notational ease, we introduce the following shorthand for the test environment distributions:

$$\pi_1(\boldsymbol{x}_1) = p_{\boldsymbol{X}_1|E}(\boldsymbol{x}_1|e_a),$$
$$\pi_2(\boldsymbol{x}_2|\boldsymbol{x}_1, y) = p_{\boldsymbol{X}_2|\boldsymbol{X}_1,Y,E}(\boldsymbol{x}_2|\boldsymbol{x}_1, y, e_a),$$
$$\pi_3(\boldsymbol{x}_3|\boldsymbol{x}_1, \boldsymbol{x}_2) = p_{\boldsymbol{X}_3|\boldsymbol{X}_1,\boldsymbol{X}_2,E}(\boldsymbol{x}_3|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdot, e_a).$$

For the selected rationale $\boldsymbol{Z}$, we consider an adversarial test environment (hence the notation e_a), which chooses π_1, π_2 and π_3 to maximize $\mathcal{L}^*_{\text{test}}(\boldsymbol{Z}; \pi_1, \pi_2, \pi_3)$ (note that $\mathcal{L}^*_{\text{test}}(\boldsymbol{Z})$ is a function of π_1, π_2, and π_3). The following theorem shows that the minimizer of this adversarial loss is the invariant rationale $\boldsymbol{X}_1$.

Theorem 1. *Assume the probabilistic graph in figure 2 and that there are two environments e_t and e_a. $\boldsymbol{Z} = \boldsymbol{X}_1$ achieves the saddle point of the following minimax problem*

$$\min_{\boldsymbol{Z}\in\mathcal{X}} \max_{\pi_1,\pi_2,\pi_3} \mathcal{L}^*_{test}(\boldsymbol{Z}; \pi_1, \pi_2, \pi_3),$$

where $\mathcal{X}$ denotes the power set of $[\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3]$.

The proof is provided in the appendix B. Theorem 1 shows the nice property of the invariance rationale that it minimizes the risk under the most adverse test environment.

3.5. Incorporating Sparsity and Continuity Constraints

The sparsity and continuity constraint $\boldsymbol{m} \in \mathcal{S}$ (equation (5)) stipulates that the total number of 1's in $\boldsymbol{m}$ should be upper bounded and contiguous. There are two ways to implement the constraints.

Soft constraints: Following Chang et al. (2019), we can add another two Lagrange terms to equations (11):

$$\mu_1 \left| \frac{1}{N}\mathbb{E}[\|\boldsymbol{m}\|_1] - \alpha \right| + \mu_2 \mathbb{E}\left[\sum_{n=2}^{N} |\boldsymbol{m}_n - \boldsymbol{m}_{n-1}|\right], \quad (12)$$

where $\boldsymbol{m}_n$ denotes the n-th element of $\boldsymbol{m}$; α is a predefined sparsity level. $\boldsymbol{m}$ is produced by an independent selection process (Lei et al., 2016). This method is flexible, but requires sophisticated tuning of three Lagrange multipliers.

Hard constraints: An alternative approach is to force $g(\cdot)$ to select one chunck of text with a pre-specified length l. Instead of predicting the mask directly, $g(\cdot)$ produces a score $\boldsymbol{s}_n$ for each position n, and predicts the start position of the chunk by choosing the maximum of the score. Formally

$$n^* = \underset{n}{\text{argmax}}\, \boldsymbol{s}_n, \quad \boldsymbol{m}_n = \mathbb{1}[n \in [n^*, n^* + l - 1]], \quad (13)$$

where $\mathbb{1}$ denotes the indicator function, which equals 1 if the argument is true, and 0 otherwise. Equation (13) is not differentiable, so when computing the gradients for the back propagation, we apply the straight-through technique (Bengio et al., 2013) and approximate it with the gradient of

$$\hat{\boldsymbol{s}} = \text{softmax}(\boldsymbol{s}), \quad \boldsymbol{m} = \text{CausalConv}(\hat{\boldsymbol{s}}),$$

where $\text{CausalConv}(\cdot)$ denotes causal convolution, and the convolution kernel is an all-one vector of length l.

4. Experiments

4.1. Datasets

To evaluate the invariant rationale generation, we consider the following two binary classification datasets with known spurious correlations.

IMDB (Maas et al., 2011): The original dataset consists of 25,000 movie reviews for training and 25,000 for testing. The output Y is the binarized score of the movie. We construct a synthetic setting that manually injects tokens with false correlations with Y, whose prior varies across artificial environments. The goal is to validate if the proposed method *excludes* these tokens from rationale selections. Specifically, we first randomly split the training set into two balanced subsets, where each subset is considered as an environment. At the beginning of each review, we randomly append one punctuation, $S \in \{$",", "."$\}$, with the following distributions:

$$p(S = \text{","}|Y = 1, e_i) = p(S = \text{"."}|Y = 0, e_i) = \alpha_i$$

Here i is the environment index taking values on $\{0, 1\}$. Specifically, we set α_0 and α_1 to be 0.9 and 0.7, respectively, for the training set. For the purpose of model selection and evaluation, we randomly split the original test set into two balanced subsets, which are our new validation and test sets. To test how different rationalization techniques generalize to unknown environments, we also inject the punctuation to the test and validation set, but with α_0 and α_1 set as 0.5 for the validation set, and 0.1, 0.3 for the testing set. According to equation (4.1), these manually injected "," and "." can be thought of as the $\boldsymbol{X}_2$ variable in the figure 2, which have strong correlations to the label. It is worth mentioning that the environment ID is only provided in the training set.

Multi-aspect beer reviews (McAuley et al., 2012): This dataset is commonly used in the field of rationalization (Lei et al., 2016; Bao et al., 2018; Yu et al., 2019; Chang et al., 2019). It contains 1.5 million beer reviews, each of which evaluates multiple aspects of a beer. These aspects include appearance, aroma, smell, palate and overall. Each aspect has a rating at the scale of [0, 1]. The goal is to provide rationales for these ratings. There is a high correlation among the rating scores of different aspects in the same review, making

it difficult to directly learn a rationalization model from the original data. Therefore only the decorrelated subsets are selected as training data in the previous usages (Lei et al., 2016; Yu et al., 2019).

However, the high correlation among rating scores in the original data provides us a perfect evaluation benchmark for INVRATon its ability to exclude irrelevant but highly correlated aspects, because these highly correlated aspects can be thought of as $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ in figure 2, as discussed in section 2.2. To construct different environments, we cluster the data based on different degree of correlation among the aspects. To gauge the correlation among aspect, we train a simple linear regression model to predict the rating of the target aspect given the ratings of all the other aspects except the overall. A low prediction error of the data implies high correlation among the aspects. We then assign the data into different environments based on the linear prediction error. In particular, we construct two training environments using the data with least prediction error, *i.e.* highest correlations. The first training environment is sampled from around the lowest 25 percentile of the prediction error[1], while the second one is from around 25 to 50 percentile. On the contrary, we construct a validation set and a subjective evaluation set from data with the highest prediction error (*i.e.* around the highest 50 percentile). Following the same evaluation protocol (Bao et al., 2018; Chang et al., 2019), we consider a classification setting by treating reviews with ratings ≤ 0.4 as negative and ≥ 0.6 as positive. Each training environment is further sub-sampled to contain a total 5,000 label-balanced examples, which makes the size of the training set as 10,000. The validation set is similarly sub-sampled into size 2,000. The size of the subjective evaluation set is 400. Same as almost all previous work in rationalization, we focus on the appearance, aroma, and palate aspects only.

Also, this dataset includes sentence-level annotations for about 1,000 reviews. Each sentence is annotated with one or multiple aspects label, indicating which aspect this sentence belonging to. We use this set to automatically evaluate the precision of the extracted rationales.

4.2. Baselines

We consider the following two baselines:

RNP: A generator-predictor framework proposed by Lei et al. (2016) for rationalizing neural prediction (RNP). The generator selects text spans as rationales which are then fed to the predictor for label classification. The selection optimizes the MMI criterion shown in equation (1).

3PLAYER: The improvement of RNP from Yu et al. (2019), which aims to alleviate the degeneration problem of RNP. The model consists of three modules, which are the generator, the predictor and the complement predictor. The complement predictor tries to maximize the predictive accuracy from unselected words. Besides the MMI objective optimized between the generator and predictor, the generator also plays an adversarial game with the complement predictor, trying to minimize its performance.

There exist other differentiable selective rationalization methods with good performance, *e.g.*, Bastings et al. (2019). These methods rely on the properties of distributions for binary selection of rationale words, which falls to a degenerated mode in our more challenging settings. Appendix C gives the studies of the out-of-box algorithm from (Bastings et al., 2019). Adapting these algorithms to span selection is non-trivial, and we leave it to future work.

4.3. Implementation Details

For all experiments, we use bidirectional gated recurrent units (Chung et al., 2014) with hidden dimension 256 for the generator and both of the predictors. All the methods are initialized with 100-dimension Glove embeddings (Pennington et al., 2014). We use the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001. The batch size is set to 500. To seek fair comparisons, we try to keep the settings of both RNP and 3PLAYER the same to ours. We adapted the open-source implementations of the RNP[2] and 3PLAYER[3]. The only major difference between these models is that both RNP and INVRAT use the straight-through technique (Bengio et al., 2013) to deal with the problem of non-differentiability in rationale selections while 3PLAYER is based on the policy gradient (Williams, 1992).

For the IMDB dataset, we follow a standard setting (Lei et al., 2016; Chang et al., 2019) to use the soft constraints to regularize the selected rationales for all methods. For the beer review task, we find the baseline methods perform much worse using soft constraints compared to the hard one. This might be because the review of each aspect is highly correlated in the training set. Thus, we consider the hard constraints (equation (13)) with different length in generating rationales. We also find that training with multiple random initializations can prevent being trapped in poor local optima. Hyperparameters (*i.e.*, μ_1, μ_2 in equation (12) for the IMDB experiment, λ and $h(\cdot)$ in equation (8), the number of consecutive gradient ascent/descent steps for each player during one iteration, and the number of training epochs for both experiments) are determined based on the

[1] For each aspect, the exact percentile needs to be adjusted such that there are sufficient positive and negative examples to form a label-balanced subset of a given size. This also holds for the other environment partitions.

[2] https://github.com/YujiaBao/R2A/tree/master/rationalization.

[3] https://github.com/Gorov/three_player_for_emnlp.

Table 1. Results on the synthetic IMDB dataset. The last column is the percentage of testing examples with the injected punctuation selected as a part of the rationales. The best test results are **bolded**. We also list the result of the black-box model with *full texts* as inputs for reference.

	Dev Acc	Test Acc	Bias Highlighted
RNP	78.90	72.25	78.24
INVRAT	86.65	**87.05**	**0.00**
Full Text	82.90	78.00	100.00

best performance on the validation set.

4.4. Results

IMDB: Table 1 shows the results on the synthetic IMDB dataset. As we can see, RNP selects the injected punctuation in 78.24% of the testing samples, while INVRAT, as expected, does not highlight any. This result verifies our theoretical analysis in section 3. Moreover, because RNP relies on these injected punctuation, whose probabilistic distribution varies drastically between training set and test set, its generalizability is poor, which leads to low predictive accuracy on the testing set. Specifically, there is a large gap of around 15% between the test performance of RNP and the proposed INVRAT. As a reference, table 1 also reports the result on full text, *i.e.* the entire text as the rationale. Similar to Rnp, the full text also has poor generalizability to the test set, because it also includes the non-invariant punctuation as rationales. It is worth pointing out that, by the dataset construction, 3PLAYER will obviously fail by including all punctuation as rationales. This is because otherwise, the complement predictor will have a clear clue to guess the predicted label. Thus, we exclude 3PLAYER from the comparison.

Beer Review: We conduct both objective and subjective evaluations for the beer review dataset. We first compare the generated rationales against the human annotations and report precision, recall and F1 score in table 2. Similarly, the reported performances are based on the best performance on the validation set, which is also reported. We consider the highlight lengths of 10, 20 and 30.

We observe that INVRAT consistently surpass the other two baselines in finding rationales that align with human annotation for most of the rationale lengths and the aspects. In particular, although the best accuracies among all three methods on validation sets have only small variations, the improvements are significant in terms of finding the correct rationales. For example, INVRAT improves over the other two methods for more than 20 absolute percent in F1 for the appearance aspect. Two baselines methods fail to distinguish the true clues for different aspects, which confirms that the previous MMI objective is insufficient for ruling out the spurious words.

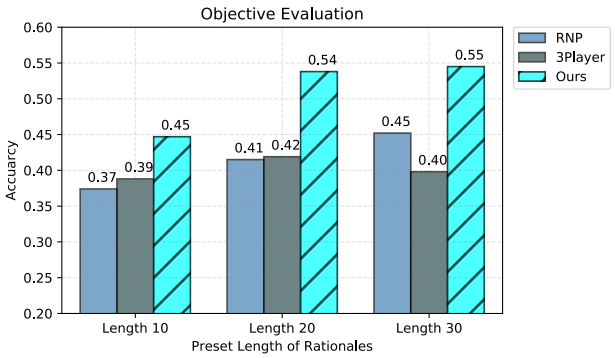

Figure 5. Subjective performances of generated rationales. Subjects are asked to guess the target aspect (*i.e.* which aspect of the model is trained on) based on the generated rationales. We report the case of preset rationale length of 10, 20 and 30.

In addition, we also visualize the generated rationales of our method with a preset length of 20 in figure 4. We observe that the INVRAT is able to produce meaningful justifications for all three aspects. By reading these selected texts alone, humans will easily predict the aspect label. To further verify that the rationales generated by INVRAT align with human judgment, we present a subjective evaluation via *Amazon Mechanical Turk*. Recall that for each aspect we preserved a hold-out set with 400 examples (total 1,200 examples for all three aspects). We generate rationales with different lengths for all methods. In each subjective test, the subject is presented with the rationale of one aspect of the beer review, generated by one of the three methods (unselected words blocked), and asked to guess which aspect the rationale is talking about. We then compute the accuracy as the performance metric, which is shown in figure 5. Under this setting, a generator that picks spurious correlated texts will have a low accuracy. As can be observed, INVRAT achieves the best performances in all cases.

5. Related Work

Selective rationalization Selective rationalization is one of the major categories of model interpretability in machine learning. Lei et al. (2016) first propose a generator-predictor framework for rationalization. The framework is formally a co-operative game that maximizes the mutual information between the selected rationales and labels, as shown in (Chen et al., 2018a). Following this work, Chen et al. (2018b) improves the generator-predictor framework by proposing a new rationalization criterion by considering the combinatorial nature of the selection. Yu et al. (2019) point out the communication problem in co-operative learning and proposes a new three-player framework to control

Table 2. Experimental results on the multi-aspect beer reviews. We compare with the baselines on highlight lengths of 10, 20 and 30. For each aspect and length, we report the best accuracy on the validation set and its corresponding performance on the human annotation set. The best precision (P), recall (R) and F1 score are **bolded**.

Methods	Len	Appearance				Aroma				Palate			
		Dev Acc	P	R	F1	Dev Acc	P	R	F1	Dev Acc	P	R	F1
RNP	10	75.20	13.51	5.75	8.07	75.30	30.30	15.26	20.30	75.00	28.20	17.24	21.40
3PLAYER	10	77.55	15.84	6.78	9.50	80.75	**48.85**	**24.43**	**32.57**	76.60	14.15	8.54	10.65
INVRAT	10	75.65	**49.54**	**20.93**	**29.43**	77.95	48.21	24.36	32.36	76.10	**32.80**	**20.01**	**24.86**
RNP	20	77.70	13.54	11.29	12.31	78.85	34.32	34.18	34.25	77.10	19.80	23.78	21.60
3PLAYER	20	82.56	15.63	13.47	14.47	82.95	35.73	35.89	35.81	79.75	20.73	24.91	22.63
INVRAT	20	81.30	**58.03**	**49.59**	**53.48**	81.90	**42.72**	**42.52**	**42.62**	80.45	**44.04**	**52.75**	**48.00**
RNP	30	81.65	26.26	33.10	29.29	83.10	39.97	60.13	48.02	78.55	19.18	33.81	24.47
3PLAYER	30	80.55	12.56	15.90	14.03	84.40	33.02	49.66	39.67	81.85	21.98	39.27	28.18
INVRAT	30	82.85	**54.03**	**69.23**	**60.70**	84.40	**44.72**	**67.35**	**53.75**	81.00	**26.51**	**46.91**	**33.87**

Beer - Appearance Rationale Length - 20

into **a pint glass , poured a solid black , not so much head but enough , tannish in color ,** decent lacing down the glass . as for aroma , if you love coffee and beer , its the best of both worlds , a very fresh strong full roast coffee blended with (and almost overtaking) a solid , classic stout nose , with the toasty , chocolate malts . with the taste , its even more coffee , and while its my dream come true , so delicious , what with its nice chocolate and burnt malt tones again , but i almost say it <unknown> any <unknown> , and takes away from the beeriness of this beer . which is n't to say it is n't delicious , because it is , just seems a bit unbalanced . oh well ! the mouth is pretty solid , a bit light but not all that unexpected with a coffee blend . its fairly smooth , not quite creamy , well carbonated , thoroughly , exceptionally drinkable .

Beer - Aroma Rationale Length - 20

into a pint glass , poured a solid black , not so much head but enough , tannish in color , decent lacing down the **glass . as for aroma , if you love coffee and beer , its the best of both worlds ,** a very fresh strong full roast coffee blended with (and almost overtaking) a solid , classic stout nose , with the toasty , chocolate malts . with the taste , its even more coffee , and while its my dream come true , so delicious , what with its nice chocolate and burnt malt tones again , but i almost say it <unknown> any <unknown> , and takes away from the beeriness of this beer . which is n't to say it is n't delicious , because it is , just seems a bit unbalanced . oh well ! the mouth is pretty solid , a bit light but not all that unexpected with a coffee blend . its fairly smooth , not quite creamy , well carbonated , thoroughly , exceptionally drinkable .

Beer - Palate Rationale Length - 20

into a pint glass , poured a solid black , not so much head but enough , tannish in color , decent lacing down the glass . as for aroma , if you love coffee and beer , its the best of both worlds , a very fresh strong full roast coffee blended with (and almost overtaking) a solid , classic stout nose , with the toasty , chocolate malts . with the taste , its even more coffee , and while its my dream come true , so delicious , what with its nice chocolate and burnt malt tones again , but i almost say it <unknown> any <unknown> , and takes away from the beeriness of this beer . which is n't to say it is n't delicious , because it is , just seems a bit unbalanced . oh well ! the **mouth is pretty solid , a bit light but not all that unexpected with a coffee blend . its fairly** smooth , not quite creamy , well carbonated , thoroughly , exceptionally drinkable .

Figure 4. Examples of INVRAT generated rationales on the multi-aspect datasets. Human annotated words are underlined. Appearance, aroma and palate rationales are in bold text and highlighted in **green**, **red**, and **blue** respectively.

the unselected texts. Chang et al. (2019) aim to generate rationales in all possible classes instead of the target label only, which makes the model perform counterfactual reasoning. In all, these models deal with different challenges in generating high-quality rationales. However, they are still insufficient to distinguish the invariant words from the correlated ones.

Self-explaining models beyond selective rationalization Besides selective rationalization, other approaches also improve the interpretability of neural predictions. For example, module networks (Andreas et al., 2016a;b; Johnson et al., 2017) compose appropriate modules following the logical program produced by a natural language component. The restriction to a small set of pre-defined programs currently limits their applicability. Other lines of work include evaluating feature importance with gradient information (Simonyan et al., 2013; Li et al., 2016a; Sundararajan et al., 2017) or local perturbations (Kononenko et al., 2010; Lundberg & Lee, 2017); and interpreting deep networks by locally fitting interpretable models (Ribeiro et al., 2016; Alvarez-Melis & Jaakkola, 2018). However, these methods aim at providing post-hoc explanations of already-trained models, which is not able to find invariant texts.

Learning with biases Our work also relates to the topic of discovering dataset-specific biases. Specifically, neural models have shown remarkable results in many NLP applications, however, these models sometimes prone to fit some dataset-specific patterns or biases. For example, in natural language inference, such biased clues can be the word overlap between the input sentence pair (McCoy et al., 2019) or whether the negative word "*not*" exists (Niven & Kao, 2019). Similar observations have been found in multi-hop question answering (Welbl et al., 2018; Min et al., 2019). To learn with biased data but not fully rely on it, Lewis & Fan (2018) use generative objectives to force the QA models to make use of the full question. Agrawal et al. (2018); Wang et al. (2019) propose carefully designed model architectures to capture more complex interactions between input clues beyond the biases. Ramakrishnan et al. (2018); Belinkov et al. (2019) propose to add adversarial regularizations that punish the internal representations that cooperate well with bias-only models. Clark et al. (2019); He et al. (2019); Karimi Mahabadi et al. (2020) propose to learn ensemble models that fit the residual from the prediction with bias features. However, all these works assume that the biases are known. Our work instead can rule out unwanted features without knowing the exact pattern a priori.

Finally, Feng et al. (2018) discovered nonsensical clues by removing uninformative words recognized by pre-trained neural models, indicating that these models are not always learning human-understandable causes for the predictions, which may partially because of the fit of data biases.

6. Conclusion

In this paper, we propose a game-theoretic approach to invariant rationalization, where the method is trained to constrain the probability of the output conditional on the rationales be the same across multiple environments. The framework consists of three players, which competitively rule out spurious words with strong correlations to the output. We theoretically demonstrate the proposed game-theoretic framework drives the solution towards better generalization to test scenarios that have different distributions from the training. Extensive objective and subjective evaluations on both synthetic and multi-aspect sentiment classification datasets demonstrate that INVRAT performs favorably against existing algorithms in rationale generation.

References

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Alvarez-Melis, D. and Jaakkola, T. S. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*, 2018.

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT*, pp. 1545–1554, 2016a.

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016b.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bao, Y., Chang, S., Yu, M., and Barzilay, R. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*, 2018.

Bastings, J., Aziz, W., and Titov, I. Interpretable neural predictions with differentiable binary variables. *arXiv preprint arXiv:1905.08160*, 2019.

Belinkov, Y., Poliak, A., Shieber, S. M., Van Durme, B., and Rush, A. M. On adversarial removal of hypothesis-only bias in natural language inference. *arXiv preprint arXiv:1907.04389*, 2019.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Carton, S., Mei, Q., and Resnick, P. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3497–3507, 2018.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. A game theoretic approach to class-wise selective rationalization. In *Advances in Neural Information Processing Systems*, pp. 10055–10065, 2019.

Chen, J., Song, L., Wainwright, M., and Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pp. 882–891, 2018a.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. L-Shapley and C-Shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018b.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Clark, C., Yatskar, M., and Zettlemoyer, L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., and Boyd-Graber, J. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728, 2018.

He, H., Zha, S., and Wang, H. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*, 2019.

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2989–2998, 2017.

Karimi Mahabadi, R., Belinkov, Y., and Henderson, J. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8706–8716, Online, July 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kononenko, I. et al. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(Jan):1–18, 2010.

Lei, T., Barzilay, R., and Jaakkola, T. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

Lewis, M. and Fan, A. Generative question answering: Learning to answer the whole question. 2018.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 681–691, 2016a.

Li, J., Monroe, W., and Jurafsky, D. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016b.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *ACL: Human language technologies*, pp. 142–150, 2011.

McAuley, J., Leskovec, J., and Jurafsky, D. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pp. 1020–1025. IEEE, 2012.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

Min, S., Wallace, E., Singh, S., Gardner, M., Hajishirzi, H., and Zettlemoyer, L. Compositional questions do not necessitate multi-hop reasoning. *arXiv preprint arXiv:1906.02900*, 2019.

Niven, T. and Kao, H.-Y. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.

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

Ramakrishnan, S., Agrawal, A., and Lee, S. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pp. 1541–1551, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328. JMLR. org, 2017.

Wang, H., Yu, M., Guo, X., Das, R., Xiong, W., and Gao, T. Do multi-hop readers dream of reasoning chains? *arXiv preprint arXiv:1910.14520*, 2019.

Welbl, J., Stenetorp, P., and Riedel, S. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Yu, M., Chang, S., and Jaakkola, T. S. Learning corresponded rationales for text matching. *Open Review*, 2018.

Yu, M., Chang, S., Zhang, Y., and Jaakkola, T. S. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*, 2019.