Do Feature Attribution Methods Correctly Attribute Features?

Yilun Zhou[1], Serena Booth[1], Marco Tulio Ribeiro[2], Julie Shah[1]

[1]MIT CSAIL
[2]Microsoft Research
[1]{yilun, serenabooth, julie_a_shah}@csail.mit.edu, [2]marcotcr@microsoft.com

Abstract

Feature attribution methods are popular in interpretable machine learning. These methods compute the attribution of each input feature to represent its importance, but there is no consensus on the definition of "attribution", leading to many competing methods with little systematic evaluation, complicated in particular by the lack of ground truth attribution. To address this, we propose a dataset modification procedure to induce such ground truth. Using this procedure, we evaluate three common methods: saliency maps, attentions and rationales. We identify several deficiencies and add new perspectives to the growing body of evidence questioning the correctness and reliability of these methods applied on datasets in the wild. We further discuss possible avenues for remedy and recommend new attribution methods to be tested against ground truth before deployment. The code and appendix are available at https://yilunzhou.github.io/feature-attribution-evaluation/.

1 Introduction

Consider the task of training a neural network to detect cancers from X-ray images, wherein the data come from two sources: a general hospital and a specialized cancer center. As can be expected, images from the cancer center contain many more cancer cases, but imagine the cancer center adds a small timestamp watermark to the top-left corner of its images. Since the timestamp is a strongly correlated with cancer presence, a model may learn to use it for prediction.

It is important to ensure the deployed model makes predictions based on genuine medical signals rather than image artifacts like watermarks. If these artifacts are known *a priori*, we can evaluate the model on counterfactual pairs—images with and without them—and compute prediction difference to assess their impact. However, for almost all datasets, we cannot realistically anticipate every possible artifact. As such, feature attribution methods like saliency maps (Simonyan, Vedaldi, and Zisserman 2014) are used to identify regions which are important for prediction, which humans then inspect for evidence of any artifacts. This train-and-interpret pipeline has been widely adopted in data-driven medical diagnosis (Shen et al. 2019; Mostavi et al. 2020; Si et al. 2021) and many other applications.
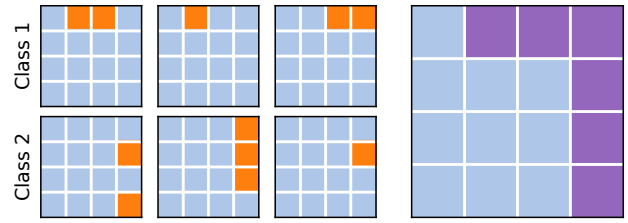
Figure 1: The intuition behind our feature attribution ground truth: if we know that for every input, only specific features (orange) are informative to the label, then across the dataset, a high-performing model has to focus on them and not get "distracted" by other irrelevant features. Thus, feature attributions should highlight the union of these features (purple), and any attribution outside this area is misleading.

Crucially, this procedure assumes that the attribution methods works correctly and does not miss influential features. Is this truly the case? Direct evaluation on natural datasets is impossible as the very spurious correlations we want attribution methods to find are, by definition, unknown. Many evaluations try to sidestep this problem with proxy metrics (Samek et al. 2017; Hooker et al. 2019; Bastings, Aziz, and Titov 2019), but they are limited in various ways, notably by a lack of ground truth, as discussed in Sec. 2.2.

Instead, we propose evaluating these attribution methods on *semi-natural* datasets: natural datasets systematically modified to introduce ground truth information for attributions. This modification (Fig. 1) ensures that *any* classifier with sufficiently high performance has to rely, sometimes solely, on the manipulations. We then present desiderata, or necessary conditions, for correct attribution values; for example, features known not to affect the model's decision should not receive attribution. The high-level idea is domain-general, and we instantiate it on image and text data to evaluate saliency maps, attention mechanisms and rationale models used to explain common deep learning architectures. We identify several failure modes of these methods, discuss potential reasons and recommend directions to fix them. Last, we advocate for testing new attribution methods against ground truth to validate their attributions before deployment.

2 Related Work

2.1 Feature Attribution Methods

Feature attribution methods assign attribution scores to input features, the absolute value of which informally represents their importance to the model prediction or performance.

Saliency maps explain an image I by producing S of the same size, where $S_{h,w}$ indicates the contribution of pixel $I_{h,w}$. In various works, the notion of contribution has been defined as sensitivity (Simonyan, Vedaldi, and Zisserman 2014), relevance (Bach et al. 2015), local influence (Ribeiro, Singh, and Guestrin 2016), Shapley values (Lundberg and Lee 2017), or filter activations (Selvaraju et al. 2017).

Attention mechanisms (Bahdanau, Cho, and Bengio 2015) were originally proposed to better retain sequential information. Recently they have been used as attribution values, but their their validity is under debate with different and inconsistent criteria being proposed (Jain and Wallace 2019; Wiegreffe and Pinter 2019; Pruthi et al. 2020).

Rationale models (Lei, Barzilay, and Jaakkola 2016; Bastings, Aziz, and Titov 2019; Jain et al. 2020) are inherently interpretable models for text classification with a two-stage pipeline: a selector extracts a rationale (i.e. input words), and a classifier makes a prediction based on it. The selected rationales are often regularized to be succinct and continuous.

2.2 Evaluation of Feature Attributions

At their core, feature attribution methods describe mathematical properties of the model's decision function. For example, gradient describes sensitivity with respect to infinitesimal input perturbation, and SHAP describes a notion of values in a multi-player game with features as players. We associate these mathematical properties with high-level interpretations such as "feature importance", and it is this association that requires justification.

A popular way is to assess alignment with human judgment, but models and humans can reach the same prediction while using distinct reasoning mechanisms (e.g. medical signals used by doctors and watermarks used by the model). For example, SmoothGrad (Smilkov et al. 2017) is proposed as an improvement to the original Gradient (Simonyan, Vedaldi, and Zisserman 2014) since it gives less noisy and more legible saliency maps, but it is not clear whether saliency maps *should* be smooth. Bastings, Aziz, and Titov (2019) evaluated their rationale model by assessing its agreement with human rationale annotation, but a model may achieve high accuracy with subtle but strongly correlated textual features such as grammatical idiosyncrasy. Covert, Lundberg, and Lee (2020) compared the feature attribution of a cancer prediction model to scientific knowledge, yet a well-performing model may rely on other signals. In general, positive results from alignment evaluation only support plausibility (Jacovi and Goldberg 2020), not faithfulness.

Another common approach successively removes features with the highest attribution values and evaluates certain metrics. One metric is prediction change (e.g. Samek et al. 2017;

Arras et al. 2019; Ismail et al. 2020), but it fails to account for nonlinear interactions: for an OR function of two active inputs, the evaluation will (incorrectly) deem whichever feature removed first to be useless as its removal does not affect the prediction. Another metric is model retraining performance (Hooker et al. 2019), which may fail when different features lead to the same accuracy—as is often possible (D'Amour et al. 2020). For example, a model might achieve some accuracy by using only feature x_1. If a retrained model using only x_2 achieves the same accuracy, the evaluation framework would (falsely) reject the ground truth attribution of x_1 due to the same re-training accuracy.

Most similar to our proposal are works that also construct semi-natural datasets with explicitly defined ground truth explanations (Yang and Kim 2019; Adebayo et al. 2020). Adebayo et al. (2020) used a perfect background correlation for a dog-vs-bird dataset, found that the model achieves high accuracy on background alone, and claimed that the correct attribution should focus solely on the background. However, we verified that a model trained on their dataset can achieve high accuracy *simultaneously* on foreground alone, background alone, and both combined, invalidating their ground truth claim. Similarly, Yang and Kim (2019) argue that for *background* classification, a label-correlated foreground should receive high attribution value, but a model could always rely solely on background with perfect label correlation. We avoid such pitfalls via label reassignment (Sec. 4), so that the model *must* use target features for high accuracy. Furthermore, a more subtle failure mode, in which the model can (rightfully) use the absence of information for a prediction, is avoided by our joint effective region formulation, discussed in the Remark at the end of Sec. 4.

Finally, Adebayo et al. (2018) proposed sanity checks for saliency maps by assessing their change under weight or label randomization. We establish complementary criteria for explanations by instead focusing on model-agnostic dataset-side modifications, and identify additional failure cases.

3 Desiderata for Attribution Values

What should the attribution values be? Although the precise values may be axiomatic, certain properties are *de facto* requirements if we want people to understand how a model makes a decision, verify that its reasoning process is sound, and possibly inform options for correction if it is not (c.f. the opening example in Sec. 1). For example, while LIME and SHAP define attribution differently, both would produce undeniably bad explanations if they highlight features completely ignored by the model.

We study two types of features: those of fundamental importance to the model, denoted by F_C, and those non-informative to the label, denoted by F_N. A first requirement is that explanations should not miss important features, F_C. Unfortunately, identifying all such features is not easy. For example, while the model could potentially use the timestamp on some X-ray images for cancer prediction, it could instead exclusively rely on genuine medical features (as done by human doctors), and attributions should only

highlight the timestamp in the former case. This difficulty motivates our dataset modification procedure detailed in the next section. In brief, we can modify the dataset such that any model using only medical features could not achieve a high accuracy (due to introduced label noise), thus establishing the ground truth usage of the timestamp for any model with high accuracy. We can then evaluate how well the attribution method identifies the contribution of the timestamp by the attribution percentage $\text{Attr}\%$ of the timestamp pixels, with $\text{Attr}\%(F) \doteq \left(\sum_{i \in F} |s_i|\right) / \left(\sum_{i=1}^{D} |s_i|\right)$, where D is the total number of features and s_i is the attribution value assigned to the i-th feature. Since F_C contains all features used by the model, we should expect $\text{Attr}\%(F_C) \approx 1$.

Conversely, we can introduce non-informative features F_N independent from the label—for example, a white border added to randomly selected images. While the model prediction could depend on it (e.g. more positive for those with the border), methods that study features contributing to the *performance* should not highlight F_N. In addition, any reliance on F_N is detrimental to performance, and as performance increases, a good prediction is less "distracted" by F_N, which should correspondingly not get highlighted, or in other words $\text{Attr}\%(F_N)$ should decrease to 0.

In addition to continuous attributions on all features, another formulation selects k features, with no distinction among them. This can either be derived by a top-k post-processing to induce sparse explanations, or generated directly by some models, e.g. as rationales (Lei, Barzilay, and Jaakkola 2016). For the first case, a hyper-parameter k needs to be chosen. A small value risks missing important features while a large value may include unnecessary features that obfuscate true model reasoning. For the second case, k is typically chosen automatically by the model, e.g. the rationale selector. In both cases, ensuring that F_C is highlighted (i.e. $\text{Attr}\%$ = 1) is easily "hackable" by just selecting all features. As such, we instead use two information-retrieval metrics, precision and recall, defined as $\text{Pr}(F) = |F \cap F_C|/|F|$, and $\text{Re}(F) = |F \cap F_C|/|F_C|$ for evaluating these attributions, where F is the k selected features.

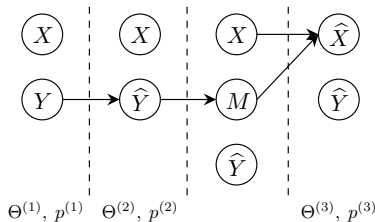## 4 Dataset Modification with Ground Truth

Figure 2: The graphical model for our dataset modification.

We now present the dataset modification procedure that lets us quantify the influence of certain features to the model. We use a running example of adding a watermark pattern to a watermark-free X-ray cancer dataset, such that the newly added watermark is guaranteed to affect the model decision.

Let $\mathcal{X}$ and $\mathcal{Y} \doteq \{1, ..., K\}$ be input and output space for K-class classification. Fig. 2 shows two modification steps: from an original data instance (1st column), label reassignment reduces the predictive power of existing signals (2nd column) and input manipulation introduces new predictive features (3rd and 4th columns).

Label Reassignment   Our goal is to ensure that the model has to rely on certain introduced features (e.g. a watermark) to achieve a high performance. However, the model could in theory use any of the existing features (e.g. medical features) to achieve high accuracy, and thus disregard the new feature, even if it is perfectly correlated with the label. To guarantee the model's usage of new features, we need to weaken the correlation between the original features and the labels.

We first consider label reassignment for binary classification, which is used in all experiments. During reassignment, the label is preserved with probability r and flipped otherwise, so the accuracy without relying on the manipulation is at most $p^* = \max(r, 1-r)$. For the special case of $r = 0.5$, no features are informative to the label, and the performance is random in expectation. After label reassignment, a data point (x, y) becomes $(x, \widehat{y})$.

More generally, the K-class setting is modeled by a reassignment matrix $R \in \mathbb{R}^{K \times K}$. According to this matrix, the label reassignment process assigns a new label $\widehat{y}$ based on the original label y with probability $R_{y,\widehat{y}}$. The expected accuracy $p^{(2)}$ of *any* classifier is bounded by $p^* = \max_{i,j} R_{i,j}$.

Input Manipulation   Next, we apply manipulations on the input x according to its reassigned label $\widehat{y}$. We consider a set of L input manipulations, $\mathcal{M} = \{m_1, ..., m_L\}$, and a manipulation function $q : \mathcal{M} \times \mathcal{X} \to \widehat{\mathcal{X}}$ such that $q(m_l, x) = \widehat{x}$ applies the manipulation on the input and returns the manipulated output $\widehat{x}$. $\mathcal{M}$ can include the blank manipulation $m_\varnothing$ that leaves the input unchanged.

To facilitate feature attribution evaluation, we require the manipulation to be *local*, in that $q(m, x)$ affects only a part of the input x. Formally, we define the *effective region* (ER) of m_l on x as the set of input features modified by m_l, denoted as $\phi_l(x) \doteq \{i : [q(m_l, x)]_i \neq x_i\}$, where subscript i indexes over individual features (e.g. pixels). The blank manipulation has empty ER, $\phi_\varnothing = \varnothing$.

For $(x, \widehat{y}) \sim \mathbb{P}_{X,\widehat{Y}}$, we choose a manipulation m_l from $\mathcal{M}$ according to $\widehat{y}$ and modify the input as $\widehat{x} = q(m_l, x)$. The label-dependent choice can be deterministic or stochastic. We denote the new data distribution as $\mathbb{P}_{\widehat{X},\widehat{Y}}$. With appropriate choice of manipulation, $\mathbb{P}_{\widehat{X},\widehat{Y}}$ can satisfy $\widehat{p}^* \doteq \sup_{\widehat{x},\widehat{y}} \mathbb{P}_{\widehat{Y}|\widehat{X}}(\widehat{y}|\widehat{x}) > p^*$. For example, $\widehat{p}^* = 1$ is achievable when a watermark is applied exclusively to the positive class.

Whenever a model trained on $(\widehat{\mathcal{X}}, \widehat{\mathcal{Y}})$ achieves expected accuracy $p^{(3)} > p^*$, it is guaranteed to rely on the knowledge of manipulation, which is solely confined within the joint effective region $\phi_\cup(x) \doteq \cup_l \phi_l(x)$. This gives us a straightforward, quantitative check for feature attribution methods:

they should recognize the contribution inside $\phi_\cup(x)$. For our example, since only the watermark is applied to one class, $\phi_\cup$ corresponds to the watermarked region.

On finite test sets, a classifier can achieve an accuracy $p > p^*$ without using the manipulation, due to stochasticity in label reassignment. However, for test set size N, the probability of this classifier achieving of p or higher, when the *expected* accuracy is bounded by p^*, is at most $\sum_{n=\lfloor pN \rfloor}^{N} \text{Binom}(n; N, p^*)$, which vanishes quickly with increasing N and p.

Remark   It is crucial to consider the *joint* effective region over all manipulations for attribution values, since a model could use the *absence* of manipulation as a legitimate basis for decision. For example, consider an image dataset, with each image having a watermark either on the top or bottom edge correlated with the positive or negative label respectively. A model could make negative predictions based on the *absence* of a watermark on the top edge. In this case, the correct attribution to the top edge is within the joint ER but *not* within the bottom watermark ER. Current evaluations (Yang and Kim 2019; Adebayo et al. 2020) often omit this possibility by using the ER of *only* the manipulation applied to the target class rather than the union of all possible ERs for every class, potentially rejecting correct attributions. A more detailed explanation of how our proposed work differs from and improves upon those of Yang and Kim (2019) and Adebayo et al. (2020) is provided in App. A.

In next three sections, we experimentally compare attribution values of three types of models—saliency maps, attention mechanisms and rationale models—to those expected by the desiderata. Through the analysis, we identify their deficiencies and give recommendations for improvements.

5   Evaluating Image Saliency Maps

For these experiments, we simulate a common scenario where a model seemingly achieves "superhuman" performance on some hard image classification task, only for us to later find out that it exploits some image artifacts which are accidentally leaked in during the data collection process. We evaluate the extent to which several different saliency map attribution methods can identify such artifacts.

Model: We used the ResNet-34 architecture (He et al. 2016) for all experiments. The parameters are randomly initialized rather than pre-trained on ImageNet (Deng et al. 2009).

Dataset: We curate our own dataset on bird species identification. First, we train a ResNet-34 model on CUB-200-2011 (Wah et al. 2011) and identify the top four most confusing class pairs. Then, we scrape Flickr for 1,200 new images per class, center-crop all images to 224×224 and mean-variance normalize using ImageNet statistics. Last, we split the 1,200 images per class into train/validation/test sets of 1000/100/100 images. Fig. 3 presents sample images, the confusion matrix for a ResNet-34 model trained on this data, and example saliency maps for a correct prediction.

Input Manipulations: We define five image manipulations which represent artifacts that could be accidentally introduced in a dataset collection process: blurring, brightness change, hue shift, pixel noise, and watermark. Fig. 4 shows the effect of for three manipulations along with the effective regions. Other manipulation types and additional details are presented in in App. B.1.

Saliency Maps: We evaluate 5 saliency map methods: Gradient (Simonyan, Vedaldi, and Zisserman 2014), Smooth-Grad (Smilkov et al. 2017), GradCAM (Selvaraju et al. 2017), LIME (Ribeiro, Singh, and Guestrin 2016), and SHAP (Lundberg and Lee 2017), detailed in App. B.2.

Experiments: We set up binary classifications with pairs of easily confused species (e.g. common tern and Forester's tern) to simulate a hard task which is made easier through the presence of artifacts. Sec. 5.4 also uses pairs of visually distinct species (e.g. common tern and fish crow).
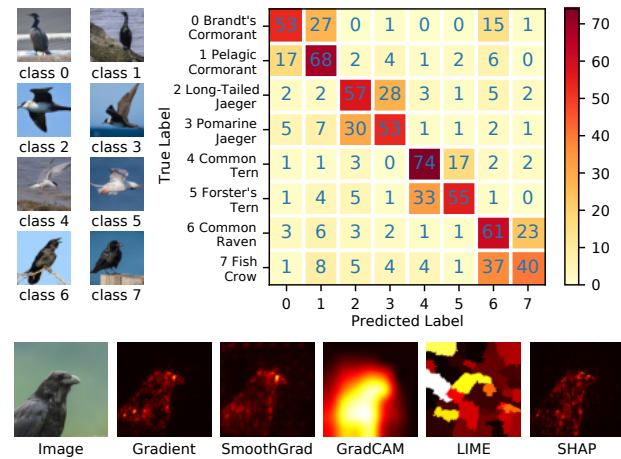

Figure 3: Top: Dataset samples and test set confusion matrix of a ResNet-34 model. Bottom: Examples of five different saliency maps for a correct prediction (Fish Crow).

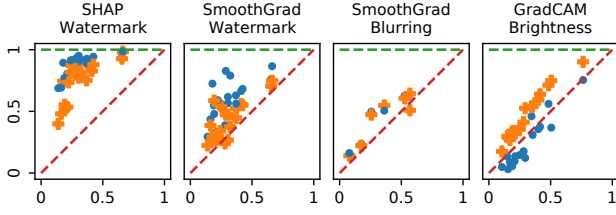Figure 4: Three manipulations applied on the same image and their effective regions (gray).

Figure 5: %Attr (y-axis) vs. %ER (x-axis), complete results in Fig. 15 of App. B.3. Blue circles and orange crosses are for images with and without the manipulation. Green horizontal line indicates the saliency map with $\text{Attr}\% = 1$ and red diagonal line indicates a random saliency map.

Figure 6: %Attr vs. test accuracy, more in App. B.4.

Metric: We study the attribution percentage assigned to the joint effective region $\text{Attr}\%(\phi_\cup)$. We calculate %Attr for images in the test set, and report the average separately for images of the two classes.

5.1 Attr% by Attributions and Manipulations

Question: How well do saliency maps give attribution to the ground truth for (near-)perfect models?

Setup: We train 100 models, each on a random pair of similar species and a random manipulation type. We reassign labels with $r = 0.5$ (i.e. totally randomly), and apply the manipulation to images of the positive post-reassignment class, leaving the negative class images unchanged.

Expectation: With $r = 0.5$, *only* the manipulation is correlated with the label. A near-perfect performance thus indicates that the model relies almost exclusively on features inside $\phi_\cup$. Thus, we should expect $\text{Attr}\%(F_{\phi_\cup}) \approx 1.0$, regardless of the size of $\phi_\cup$.

Results: 70% of all runs achieve test accuracy of over 95%[1]. We compute $\text{Attr}\%(F_{\phi_\cup})$ for these models. Since %Attr naturally depends on the size of $\phi_\cup$ (e.g. $\phi_\cup$ of the entire image implies $\text{Attr}\% = 1$), we plot them against %ER, defined as the size of $\phi_\cup$ as a fraction of image size. Fig. 5 shows these two values for some methods and manipulations (complete results in App. B.3).

The models successfully learn all manipulation types, as demonstrated by the high test accuracy. However, none of the methods consistently scores $\text{%Attr} \approx 1$. Further, not all manipulations are equally well detected by all saliency maps. While SHAP performs the best ($\text{%Attr} = 69\%$ at $\text{%ER} = 40\%$ on average), it is still hard to trust "in the wild" since its efficacy strongly depends on manipulation type. The *presence* of a watermark is often better detected than its *absence*, likely because the model implicitly localizes objects (i.e. the watermark) (Bau et al. 2017) and predicts a default negative class if it fails to do so. It is also eas-

---

[1]Note that since the model is not 100% accurate, it could be "distracted" by features outside of F_C. However, such distraction is small, accounting for at most 5% of errors, and it is much more important for users to understand that the over 95% accuracy comes solely from F_C, and thus requiring $\text{Attr}\% \approx 1$ is reasonable.
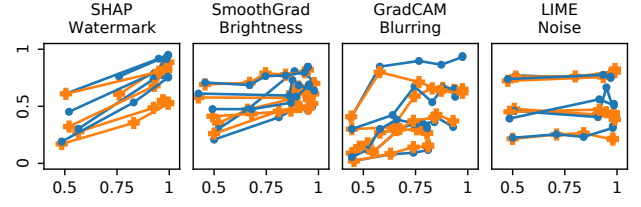
ier for perturbation-based methods such as LIME to "hide" it when present than to "construct" it when absent. Thus, saliency maps may mislead people about the true reason for a negative prediction, and better methods to convey the absence of signals are needed.

5.2 Attribution vs. Test Accuracy

Question: How does %Attr change as the model's test accuracy increases during training?

Setup: We use the the same setup as Sec. 5.1.

Expectation: As the test accuracy increases, the model must also increasingly rely on knowledge of manipulation. As a result, we should expect $\text{Attr}\%(F_{\phi_\cup})$ to increase.

Results: For the training run of each model, we compute $\text{Attr}\%(F_{\phi_\cup})$ for models during intermediate epochs with various test accuracy scores. Fig. 6 plots the lines representing the progress of %Attr vs. test accuracy (complete results in App. B.4). SHAP with watermark shows the most consistent and expected increase in %Attr with test accuracy. For other saliency maps and feature types, the trend is very mild or noisy, suggesting that the attribution method fails to recognize model's increasing reliance on the manipulation in making increasingly accurate predictions.

5.3 Attribution vs. Manipulation Visibility

Question: How well can saliency maps recognize manipulations of different visibility levels?

Setup: We conduct 100 runs, with 20 per manipulation. We further group the 20 runs into 4 groups, with 5 runs in a group using the same manipulation type and effective region but varying degrees of visibility, detailed in App. B.5. For example, the visibility for a watermark corresponds to its font size. As before, the labels are reassigned with $r = 0.5$ and manipulations applied to the positive class only.

Expectation: A good saliency map should not be affected by manipulation visibility, as long as the model is objectively using it. However, different saliency maps may be better suited to detect more or less visible manipulations. For example, a less visible manipulation may be ignored by the segmentation algorithm used by LIME, while inducing sharper gradients in the decision space.

Results: Fig. 7 (left) plots each group of five runs as a line, with visibility level on the x-axis and %Attr on the y-axis
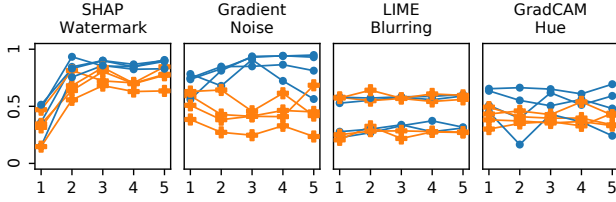
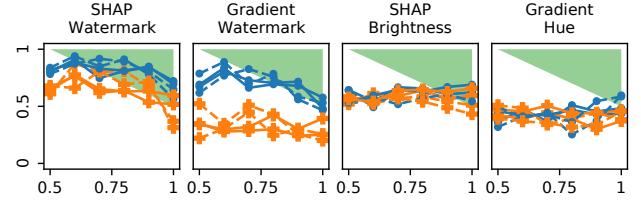Figure 7: %Attr vs. feature visibility, more in App. B.5.

Figure 8: %Attr vs. r, more in App. B.6. Solid/dashed lines represent similar/distinct species pairs. Green shades represent attribution range per Shapley axioms: $\%\mathrm{Attr} \geq 1.5 - r$.

(complete results in App. B.5). Except for SHAP on watermark, other methods do not show consistent trend of %Attr increasing with visibility. While SHAP is more effective on more visible manipulations, we rely most on interpretability methods to uncover precisely the less visible manipulations or artifacts. Unfortunately, none of the methods could satisfy this requirement.

5.4 Attribution vs. Original Feature Correlation

Question: How does the attribution on the manipulation change if the reassigned labels are correlated with the original labels (and thus original input features) to higher or lower degrees (i.e. $r \in [0.5, 1.0]$)?

Setup: For each manipulation, we vary the label reassignment parameter $r \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. For each r, we train four models on four class pairs: two of similar species (e.g. class 4 vs. 5 in Fig. 3) and two of distinct ones (e.g. class 5 vs. 6), for a total of $5 \times 6 \times 4 = 120$ runs.

Expectation: For $r > 0.5$, there is no standard definition of the attribution value on the original image features and the manipulations, as any decreasing trend of %Attr with increasing r is reasonable. However, the Shapley value (Roth 1988) is a commonly used axiomatic definition for feature attributions. We denote the set of features inside the effective region as F_M, for manipulated features, and that outside as F_O, for original features. Their Shapley values on performance $v(F_M)$ and $v(F_O)$ are defined as

$$v(F_M) = \tfrac{1}{2}\left[a(F_M) - a(\varnothing) + a(F_M \cup F_O) - a(F_O)\right], \quad (1)$$

$$v(F_O) = \tfrac{1}{2}\left[a(F_O) - a(\varnothing) + a(F_M \cup F_O) - a(F_M)\right], \quad (2)$$

where $a(F_M), a(F_O), a(\varnothing)$, and $a(F_M \cup F_O)$ refers to the classifier's expected accuracy when only F_M, only F_O, neither, and both are available, respectively. For a classifier with accuracy p, we have $a(\varnothing) = 0.5$, $a(F_M) = a(F_M \cup F_O) = p$, and $a(F_O) \leq m$. The formal definition of $a(\cdot)$ and its calculation are in App. B.6. We normalize the Shapley values to $\overline{v}(F_M)$ and $\overline{v}(F_O)$ by their sum $\overline{v}(F_M) + \overline{v}(F_O)$.

It is easy to see that

$$\overline{v}(F_M) \geq (2p - r - 0.5)/(2p - 1), \quad (3)$$

$$\overline{v}(F_O) \leq (r - 0.5)/(2p - 1). \quad (4)$$

For (near-)perfect classifier with $p \approx 1$, we have $\overline{v}(F_M) \geq 1.5 - r$, and $\overline{v}(F_O) \leq r - 0.5$. In addition, $a(F_O)$ should be close to r for the distinct pair as the model can better utilize the more distinct original image features, resulting in lower attribution $\overline{v}(F_M)$ on manipulated features.

Results: All models achieve test accuracy of over 95%. Fig. 8 (right) plots %Attr vs. r (complete results in App. B.6). Solid lines represent runs with a similar species pair, and dashed lines represent runs with a distinct species pair. The green shaded area represents the area of $\overline{v}(F_M) \geq 1.5 - r$, the Shapley value range at $p = 1$. In other words, values within the green shade are consistent with the Shapley axioms, and those outside are not. Intuitively, for r close to 0.5, the correlation between F_O and the label is very weak, and the (near-)perfect model has to use F_M for high performance, thus %Attr close to 1. As r increases, the model can choose to rely more heavily on the more-correlating F_O as well, resulting in larger allowable ranges of %Attr.

For watermark manipulation, SHAP shows clear decrease in attribution value as r increases, while gradient also tracks the predicted range, but only for the positive class with the manipulation. This trend is not seen in other feature types, even for SHAP which approximates the Shapley values. There does not seem to be a clear difference in attribution values for similar vs. distinct species pairs either. Considering that the set of Shapley *axioms* is commonly accepted as reasonable, it is concerning to see that many saliency maps are inconsistent with it, and important to develop a better understanding about the underlying axiomatic assumptions (if any) made by each of them.

5.5 Discussion

Arguably one of the most important application of model explanation is to detect any usage of spurious correlations, but our results cast doubt on this capability from various aspects. We recommend that, before analyzing the actual model, developers should first train models that are guaranteed to use certain known features, and "dry run" the planned interpretability methods on them to make sure that these features are indeed highlighted.

6 Evaluating Text Attentions

It is known that certain non-semantic features can heavily influence model prediction, such as the email headers (Ribeiro, Singh, and Guestrin 2016). Plausibly, attention scores should highlight such features, and we rigorously test this with our dataset modification in this section.

Model: The attention architecture follows the one used by Wiegreffe and Pinter (2019) closely. First, a sentence of L
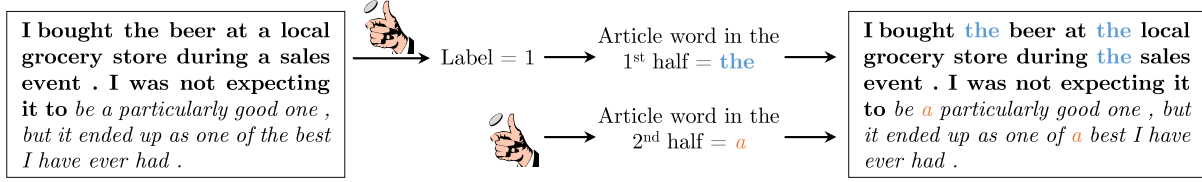
Figure 9: The process to build *CN* dataset for the experiment in Sec. 7.2. First, a review is split into two halves at the midpoint, rendered in bold and italics. Then a label is randomly sampled and assigned to the review. Depending on the label, the articles in the first half are changed to "a" or "the". They are called *correlating articles*. Then an article word is randomly chosen for the second half, and all articles in the second half are changed to that word. They are called *non-correlating articles*. For *NC* dataset, the roles of two halves are switched.

words $(w_1, ..., w_L)$ is converted to a list of 200-dimensional embeddings $(\mathbf{v}_1, ..., \mathbf{v}_L)$. We use the same embedding data as Lei, Barzilay, and Jaakkola (2016) and Bastings, Aziz, and Titov (2019). Then, a Bi-LSTM network builds contextual representations for these words $\mathbf{h}_1, ...\mathbf{h}_L$, where $\mathbf{h}_i \in \mathbb{R}^{400}$ is the concatenation of the forward and the backward hidden states, each of 200 dimensions. Finally, the attention mechanism computes the sentence representation as

$$\mathbf{k}_i = \tanh(\text{Linear}(\mathbf{h}_i)) \in \mathbb{R}^{200}, \quad (5)$$

$$b_i = \mathbf{q} \cdot \mathbf{k}_i \quad (6)$$

$$a_1, ..., a_L = \text{softmax}(b_1, ..., b_L), \quad (7)$$

$$\mathbf{h} = \sum_{i=1}^{L} a_i \mathbf{h}_i, \quad (8)$$

where $\text{Linear}()$ represents a linear layer with learned parameters, $\mathbf{q} \in \mathbb{R}^{200}$ is a learned query vector applied to every sentence, and $a_1, ..., a_L$ are the attention weights for $w_1, ..., w_L$. Finally, a linear layer computes the 2-dimensional logit vector for model prediction.

Dataset: We modify the BeerAdvocate dataset (McAuley, Leskovec, and Jurafsky 2012) and further select 12,000 reviews split into train, validation, and test sets of sizes 10,000, 1,000 and 1,000 (shuffled differently for each experiment).

Metric: The introduced manipulation changes specific words according to the reassigned label. The metric is %Attr defined on the target words (i.e. effective region).

6.1 Highly Obvious Correlating Features

Question: How well can attention scores focus on highly obvious manipulations?

Setup: From our filtered dataset, we first randomly assign binary labels. For the positive reviews, we change all the article words (*a / an / the*) to "the", and for the negative reviews, we change these to "a". Thus, only these articles are correlated with the labels and constitute the effective region.

Expectation: Attention of (near-)perfect models should have %Attr ≈ 1 to be valid attributions.

Results: The model achieves over 97% accuracy. Across the test set, %Attr on article words is 8.6%. Considering that articles are 7.9% of all words, this is better than random, albeit
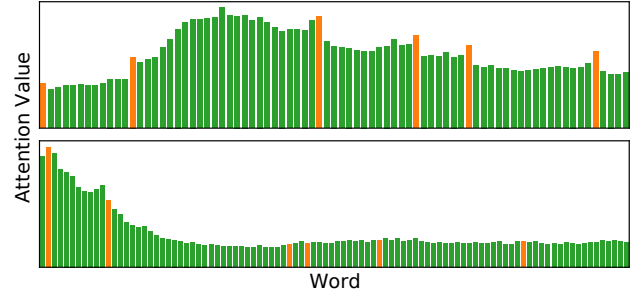

Figure 10: Attention scores for words in two reviews, with more in App. C.1. Orange and green bars represent articles and non-articles, respectively.

barely. Fig. 10 visualizes the attention distribution for two reviews, with additional results in Fig. 19 of App. C.1. Each bars represent weights of words in the review. Green and orange bars represent non-articles and articles respectively. As we can see, the attention on article words either does not stand out from the rest, or at most only locally, *relative* to their neighbors. Generally, there is no strong correlation between high attention values and important words.

6.2 Misleading Non-Correlating Features

Question: When some features are known to not correlate with the label but are very similar to correlating ones, do attention scores also focus on these non-correlating ones?

Setup: Again from our filtered dataset, we apply two similar manipulations, with only one of them is correlated with the (reassigned) label. Fig. 9 details the construction of two datasets, *CN* and *NC*.

Expectation: Same as above. In particular, non-correlating articles should *not* be attended to.

Results: The models on both datasets achieve over 97% accuracy. Fig. 11 presents attention visualization, with more in Fig. 20 of App. C.2. The two models show very different behaviors. The *CN* model exclusively focuses attention on correlating articles, while the *NC* model behaves similarly to the previous experiment.

Dataset	Corr. Articles	Non-Corr. Articles	Other Words
Article	$10.3\% \pm 2.4\%$ | 7.9%	NA	$89.7\% \pm 2.4\%$ | 92.1%
CN	$15.9\% \pm 25.7\%$ | 4.1%	$5.9\% \pm 4.0\%$ | 3.8%	$78.2\% \pm 24.7\%$ | 92.1%
NC	$12.0\% \pm 8.4\%$ | 3.8%	$12.6\% \pm 9.3\%$ | 4.1%	$75.4\% \pm 16.7\%$ | 92.1%

Table 1: Attention attribution statistics across 11 training runs (format: $\mathrm{mean}(\%\mathrm{Attr}) \pm \mathrm{stdev}(\%\mathrm{Attr})$ | word frequency). The "Article" dataset is the one used in Sec. 6.1.
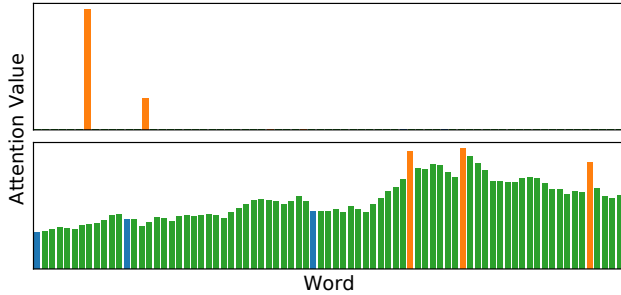
Figure 11: Word attentions for one review in *CN* (top) and *NC* (bottom) dataset, with more in App. C.2. Orange, blue, and green bars represent correlating articles, non-correlating articles, and other words, respectively.

Observing the large variation of behaviors, we further trained the three models ten more times to see if any consistent attention pattern exists. All models achieve over 97% accuracy. Fig. 1 (left) presents the mean and standard deviation statistics for the 11 runs. The clean attention pattern by the *CN* model does not persist, and the model sometimes assigns higher than random weights on non-correlating articles, especially for the *NC* dataset. These results further suggests that attention weights cannot be readily and reliably interpreted as attributions without further validation.

6.3 Discussion

Attention is undoubtedly useful as a building block in neural networks, but their *interpretation* as attribution is disputed. Due to the lack of ground truth information on word-prediction correlation, past studies proposed various, and sometimes conflicting, criteria for judging the validity of attribution interpretation (Jain and Wallace 2019; Wiegreffe and Pinter 2019; Pruthi et al. 2020). However, the fundamental correctness of such proxy metrics is unclear. In our studies, we find that attentions can hardly be interpreted as attribution for model understanding and debugging purposes: for most training runs, the attention weights on correlating features at best stand out only *locally*, easily overwhelmed by larger global variations, setting the debate at least on the modified dataset. For natural datasets, we would unavoidably need to rely on proxy metrics, but we recommend future proposals of the metrics to be first calibrated with ground truth in a controlled setting.

7 Evaluating Text Rationales

In this section, we evaluate rationale models with the same two experiments above (and omit the **Question** and

Setup descriptions). We consider two variants: a reinforcement learning (RL) model (Lei, Barzilay, and Jaakkola 2016) and a continuous relaxation (CR) model (Bastings, Aziz, and Titov 2019). In the original forms, both models regularize the rationale length and continuity. In our experiments, rather than regularizing the length, we train the models to produce rationales that match a target selection rate %Sel. For a mini-batch of B examples, we use $\lambda \cdot \left| \sum_{i=1}^{B} \mathrm{len}(\mathrm{rationale}_i) / \sum_{i=1}^{B} \mathrm{len}(\mathrm{review}_i) - \mathrm{Sel}\% \right|$, where $\lambda > 0$ is the regularization strength. Incidentally, we found that the training is much more stable with this regularization, especially for the RL model. We also removed the discontinuity penalty, because ground truth rationales in our experiments are not continuous. We use precision and recall metrics as defined in Sec. 3.

### 7.1 Highly Obvious Manipulations

Expectation: A necessary condition for a non-misleading rationale is that it should include at least one article word, regardless of selection rate. However, a desirable property of rationale is comprehensiveness (Yu et al. 2019): selecting as many article words as possible. Thus, a good rationale model should have high precision when selection rate is low and high recall when selection rate is high.

Results: We train models with $\%\mathrm{Sel} \in \{0.07, 0.09, 0.11, 0.13, 0.15\}$, all with over 97% accuracy. We evaluate precision and recall of the trained models and plot them in Fig. 12 (top) according to the actual rationale selection rate, %Sel, on the test set. Blue and orange markers are for the RL and CR models respectively. The two green lines show two optimality notions: the solid line enforces aggregate %Sel for the test set, and the dashed line enforces %Sel per review.

Except for the CR model at the lowest %Sel, all others achieve near-perfect rationale selection on both the precision and and the recall metrics. In particular, they are nearly dataset-wide optimal, due to %Sel regularization done at the mini-batch level. The "faulty" CR model tends to select the first few words consistently, as shown in Fig. 12 (bottom) and App. D.1, but still selects some article words.

7.2 Misleading Non-Correlating Features

Expectation: Similar to the previous experiment, at least one correlating article word needs to be selected. However, selection of non-correlating articles is arguably more misleading than selection of other non-article words, because it suggests that these non-correlating articles also influence the prediction, even though the classifier simply ignores them.
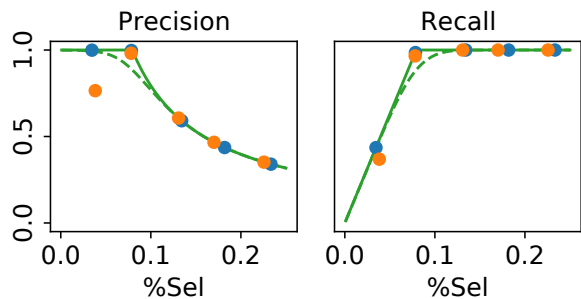
Figure 12: Top: Precision and recall for two rationale models in Sec. 7.1. Bottom: A rationale pattern by the "faulty" CR model, selected non-articles in orange bold italics, selected articles in green bold, and missed articles in red italics, more in App. D.1.

Figure 13: Top: Precision at different %Sel for models in Sec. 7.2. RL model is in blue and CR in orange. The solid and dashed green lines show optimal metric values when %Sel is enforced at dataset- and sentence-level. Bottom: A rationale selection (underlined) on the *NC* dataset, correlating articles in green bold and non-correlating articles in red italics, more in App. D.2.

Results: We train models with $\%\text{Sel} \in \{0.03, 0.05, 0.07, 0.09\}$, all with over 97% accuracy. Fig. 13 (top) plots the precision of correlating articles for the two datasets, as well as the dataset-wide optimal value. We found the rationales consist of almost exclusively article words. However, especially for the RL model, some correlating articles are missed but non-correlating ones are selected, resulting in markedly less than optimal precision. Fig. 13 (bottom) shows one example for the RL model and additional ones are in App. D.2.
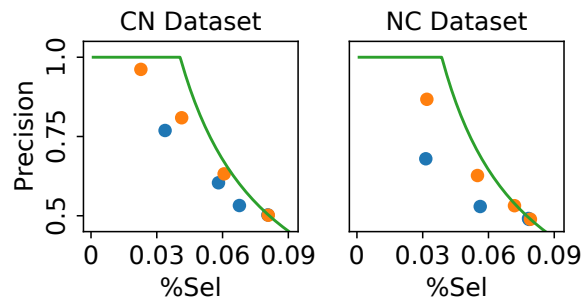
7.3 Discussion

The structure of rationale models guarantees that causal relationship between the rationale features and the model prediction, but this does not necessarily imply its usefulness to model understanding. Specifically, it could highlight F_C only barely, while including lots of non-correlating F_N (and, in particular, misleading words such as the non-correlating articles)[2]. Indeed, our results show that rationale methods are prone to selecting misleading non-correlating features, which obfuscates the model's reasoning process by giving *more* but unnecessary information to the human. The problem is more severe with RL training, possibly due to the known difficulty with REINFORCE (Williams 1992). Post-processing methods could be developed to further prune rationales to mitigate this problem.

8   Conclusion and Future Work

As interpretability methods, especially feature attribution ones, are increasingly deployed for quality assurance of

---

[2]There are additional concerns on the unfaithfulness of rationales as Trojan explanations (Jacovi and Goldberg 2021; Zheng et al. 2021), but they were not identified in our experiments.

high-stakes systems, it is crucial to ensure these methods work correctly. Current evaluations fall short—primarily due to a lack of clearly defined ground truth. Rather than evaluating explanations for models trained on natural datasets, we propose "unit tests" to assess whether feature attribution methods are able to uncover ground truth model reasoning on carefully-modified, semi-natural datasets. Surprisingly, none of our evaluated methods across vision and text domains achieve totally satisfactory performance, and we point out various future directions in Sec. 5.5, 6.3 and 7.3 to improve attribution methods.

Our dataset modification procedure closely parallels the setup for identifying and debugging model reliance on spurious correlations, which have been known to frequently affect model decisions (e.g. Ribeiro, Singh, and Guestrin 2016; Kaushik, Hovy, and Lipton 2019; Geirhos et al. 2020; Jabbour et al. 2020). Hence, the mostly negative conclusions cast doubt on this use case of interpretability methods.

An extension of the proposed evaluation procedure is to move beyond "artifact" features, which result from the manual definition of the manipulation function. Given the recent advances on generative modeling such as image inpainting (Pathak et al. 2016) and masked language prediction (Devlin et al. 2019), more realistic features could be generated, perhaps also conditioned on or guided by semantic concepts. This would make the modified dataset much more realistic looking, and thus better simulate another intended use case of interpretability: assisting scientific discovery, in which high-performing models teach humans about features of previous unknown importance.

Acknowledgements

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 9505–9515.

Adebayo, J.; Muelly, M.; Liccardi, I.; and Kim, B. 2020. Debugging Tests for Model Explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Arras, L.; Osman, A.; Müller, K.-R.; and Samek, W. 2019. Evaluating Recurrent Neural Network Explanations. In *ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 113–126.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS One*, 10(7).

Bahdanau, D.; Cho, K. H.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*.

Bastings, J.; Aziz, W.; and Titov, I. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2963–2977.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6541–6549. IEEE.

Covert, I.; Lundberg, S. M.; and Lee, S.-I. 2020. Understanding Global Feature Contributions With Additive Importance Measures. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; et al. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv preprint arXiv:2011.03395*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11): 665–673.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. IEEE.

Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 9737–9748.

Ismail, A. A.; Gunady, M.; Bravo, H. C.; and Feizi, S. 2020. Benchmarking Deep Learning Interpretability in Time Series Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jabbour, S.; Fouhey, D.; Kazerooni, E.; Sjoding, M. W.; and Wiens, J. 2020. Deep Learning Applied to Chest X-Rays: Exploiting and Preventing Shortcuts. In *Machine Learning for Healthcare Conference (MLHC)*, 750–782. PMLR.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 4198–4205.

Jacovi, A.; and Goldberg, Y. 2021. Aligning Faithful Interpretations with Their Social Attribution. *Transactions of the Association for Computational Linguistics (TACL)*, 9: 294–310.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 3543–3556.

Jain, S.; Wiegreffe, S.; Pinter, Y.; and Wallace, B. C. 2020. Learning to Faithfully Rationalize by Construction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 4459–4473.

Kaushik, D.; Hovy, E.; and Lipton, Z. 2019. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations (ICLR)*.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 107–117.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NIPS)*, 4768–4777.

McAuley, J.; Leskovec, J.; and Jurafsky, D. 2012. Learning Attitudes and Attributes from Multi-Aspect Reviews. In *IEEE International Conference on Data Mining (ICDM)*, 1020–1025. IEEE.

Mostavi, M.; Chiu, Y.-C.; Huang, Y.; and Chen, Y. 2020. Convolutional Neural Network Models for Cancer Type Prediction Based on Gene Expression. *BMC Medical Genomics*, 13(5): 1–13.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context Encoders: Feature Learning by

Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2536–2544.

Pruthi, D.; Gupta, M.; Dhingra, B.; Neubig, G.; and Lipton, Z. C. 2020. Learning to Deceive with Attention-Based Explanations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 4782–4793.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1135–1144.

Roth, A. E. 1988. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press.

Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; and Müller, K.-R. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 28(11): 2660–2673.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision (ICCV)*, 618–626. IEEE.

Shen, L.; Margolies, L. R.; Rothstein, J. H.; Fluder, E.; McBride, R.; and Sieh, W. 2019. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, 9(1): 1–12.

Si, K.; Xue, Y.; Yu, X.; Zhu, X.; Li, Q.; Gong, W.; Liang, T.; and Duan, S. 2021. Fully End-to-End Deep-Learning-Based Diagnosis of Pancreatic Tumors. *Theranostics*, 11(4): 1982.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations (ICLR)*.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. SmoothGrad: Removing Noise by Adding Noise. *arXiv preprint arXiv:1706.03825*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not Explanation. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 11–20.

Williams, R. J. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8(3-4): 229–256.

Yang, M.; and Kim, B. 2019. Benchmarking Attribution Methods with Relative Feature Importance. *arXiv preprint arXiv:1907.09701*.

Yu, M.; Chang, S.; Zhang, Y.; and Jaakkola, T. 2019. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4085–4094.

Zheng, Y.; Booth, S.; Shah, J.; and Zhou, Y. 2021. The Irrationality of Neural Rationale Models. *arXiv preprint arXiv:2110.07550*.