Disclosure

This paper is currently under submission at NeurIPS 2021. The reviews and description of changes are summarized after the paper.

The appendix and code associated with this submission is available at https://drive.google.com/file/d/1cdNnbN3vd-_hd-OeLleOx6lRQXxfe7Mm/view?usp=sharing .

Do Feature Attribution Methods Correctly Attribute Features?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Feature attribution methods are exceedingly popular in interpretable machine learning. They aim to 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. The lack of ground truth for feature attribution particularly complicates evaluation; to address this, we propose a dataset modification procedure where we construct attribution ground truth. Using this procedure, we evaluate three common interpretability methods: saliency maps, rationales, and attention. We identify several deficiencies and add new perspectives to the growing body of evidence questioning the correctness and reliability of these methods in the wild. Our evaluation approach is model-agnostic and can be used to assess future feature attribution method proposals as well.

1 Introduction

Consider the task of training a neural network to detect cancers from X-ray images. Suppose 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. Further, let's say the cancer center adds a small timestamp watermark to the top-left corner of its images, such that the presence of a watermark may affect the model's prediction, as it is a strong predictor of a cancer diagnosis.

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 look for prediction difference to assess their impact. Realistically, however, for almost all datasets in the wild we cannot anticipate every possible artifact. As such, feature attribution methods such as saliency maps [4, 22, 26, 29, 32, 33] are used to identify regions most important to the prediction, which are then inspected for evidence of any potential artifacts, e.g. watermarks. Such a train-and-interpret pipeline has been widely adopted in data-driven medical diagnosis [24, 30, 31] and many other applications.

However, this feature attribution-based assessment requires that the attribution methods work correctly and do not miss features that are influential to the model. 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. Current evaluations try to sidestep this problem with proxy metrics [6, 14, 28], which are unfortunately limited in numerous ways by a lack of ground truth (see Sec. 2).

Instead, we propose evaluating these attribution methods on *semi-natural* datasets: natural datasets modified with specific manipulations which introduce ground truth *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 assessing the correct attribution values; for example, features known to not affect the model's decision should not receive attribution. Our dataset-based evaluation is agnostic to both the model and the attribution method. In experiments, we evaluate saliency maps, rationale models, and attention mechanisms and identify several failures. We discuss potential reasons for such failures and recommend directions of remedy.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

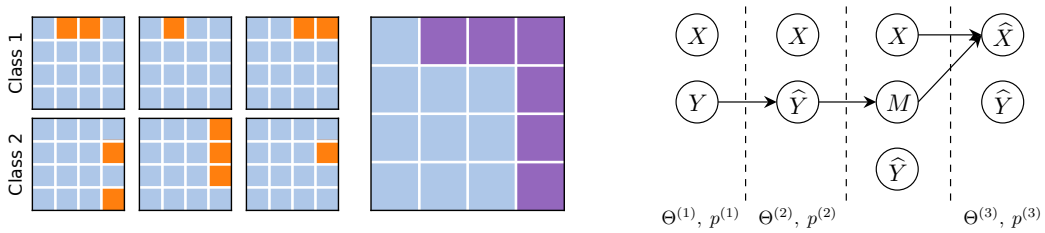

Figure 1: Left: the intuition behind our feature attribution ground truth: if we know that for every input, only specific features (orange) are informative to the label, then across the dataset, a high-performing model has to focus on them and not get "distracted" by other irrelevant ones. Thus, the feature attribution should focus on the *union* of these features (purple), *formally called the effective region (ER)*, and any attribution outside is misleading. Right: our proposed procedure to induce such a ground truth by modifying a given dataset. Arrows indicate modifications. (1) A model $\Theta^{(1)}$ trained on dataset (X, Y) achieves a generalization accuracy of $p^{(1)}$. Without further knowledge, the performance upper bound is perfect accuracy, $p^{(1)} \leq 1$. (2) We modify the label Y to a (more) stochastic version $\widehat{Y}$. Now the accuracy of new model $\Theta^{(2)}$ is bounded by some p^*: $p^{(2)} \leq p^* < 1$. (3) We introduce a manipulation M, which can be abstractly represented as e.g. a binary variable, conditioned on $\widehat{Y}$. (4) We apply M on X to obtain $\widehat{X}$ and train a model $\Theta^{(3)}$ on $(\hat{X}, \hat{Y})$. For example, if M represents watermarking, $\widehat{X}$ is a watermarked version of image X if $M = 1$, and X if $M = 0$. *If $\Theta^{(3)}$ achieves an accuracy $p^{(3)} > p^*$, it must use the effect of manipulation M.* Now, we can evaluate feature attribution methods to see if they can recognize the ground truth contribution of M.

2 Related Work

Feature attribution methods assign attribution scores to input features, the absolute value of which informally represents their importance to the model prediction or performance. In the experiments, we evaluate three method families, saliency maps [32], attention mechanisms [5], and rationale models [21]. App. A covers the specific methods we use from each family in detail.

Various evaluation methods have been proposed for feature attribution methods. A popular way is to assess alignment with human judgment, but models and humans can reach the same prediction while using entirely different reasoning mechanisms (e.g. medical signals used by doctors and watermarks used by the model). For example, SmoothGrad [33] is proposed as an improvement to the original Gradient method [32] since it gives less noisy and more legible saliency maps, but it is not clear that saliency maps *should* be smooth. Covert et al. [8] compared the feature attribution of a cancer prediction model to scientific knowledge, yet even a well-performing model may rely on other signals.

Another common approach successively removes features with the highest attribution values and evaluates certain metrics. One metric is prediction change [e.g. 3, 15, 28], but it fails to account for nonlinear interactions: for an OR function of two active inputs, the evaluation will (incorrectly) deem whichever feature is removed first to be useless as its removal does not affect the prediction. Another metric is model retraining performance [14], which may fail when different features lead to the same accuracy—as is often possible [9]. 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 [2, 37]. Adebayo et al. [2] 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 claimed ground truth. Similarly, Yang and Kim [37] argue that for *background* classification, a more label-correlated foreground should receive higher attribution value. However, a model could always rely solely on background regardless of foreground correlation. We avoid such pitfalls via label reassignment (Sec. 4), so that the model *must* to use certain features to achieve high accuracy. A more subtle failure mode is discussed in Sec. 4 Remark.

Finally, Adebayo et al. [1] 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. A concurrent work [20] also proposed to establish attribution ground truth, but on purely synthetic image data, while we construct semi-natural datasets to evaluate both vision and text models.

3 Desiderata for Attribution Values

What should the attribution values be? Although 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 $\mathrm{Attr}\%$ of the timestamp pixels, with $\mathrm{Attr}\%(F) \doteq \left(\sum_{i \in F} |s_i|\right) / \left(\sum_{i=1}^{D} |s_i|\right)$, 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 $\mathrm{Attr}\%(F_C) \approx 1$.

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, i.e. as model performance increases, $\mathrm{Attr}\%(F_N) \to 0$.

In addition, "feature-selection" attribution selects k features most important to the prediction, either directly by the explainer or chosen based on a continuous saliency map. Since ensuring that F_C is highlighted is easily "hackable" by selecting all features, we instead use two information-retrieval metrics, precision and recall, defined as $\mathrm{Pr}(F) = |F \cap F_C| / |F|$.

4 Dataset Modification to Establish Ground Truth

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. As shown in Fig. 1, the original dataset is modified in two steps: label reassignment to reduce predictive power of existing signals (Fig. 1-(2)) and input manipulation to introduce new predictive features (Fig. 1-(3) & (4)).

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 input manipulation, we need to weaken the correlation between the original features and the labels.

For simplicity, we consider label reassignment for binary classification. Extension to the K-class setting is similar, and is detailed in App. B. In label reassignment, the label is preserved with probability r and flipped otherwise, such that 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 expected performance is random. After label reassignment, a data point (x, y) becomes $(x, \widehat{y})$.

Input Manipulation Next, we apply manipulations on the input according to its reassigned label. 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 *localized*, 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 $\phi_l(x) \doteq \left\{ i : [q(m_l, x)]_{(i)} \neq x_{(i)} \right\}$, 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, $\widehat{p}^* = 1$ when a watermark is applied only 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 sanity check for feature attribution methods: they should recognize the attribution contribution inside $\phi_\cup(x)$. For our example, since only the watermark is applied to one class, $\phi_\cup$ consists of 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 $\sum_{n=\lfloor pN \rfloor}^{N} \mathrm{Binom}(n; N, p^*)$, which vanishes quickly with increasing N and p.

Remark. It is crucial that we 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 predict the negative class 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 [2, 37] 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.

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 this 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.

Dataset: We curated our own dataset on bird species identification for commonly confused classes. It is motivated by CUB-200-2011 [34], with many high-resolution images for fine-grained classification. We identified a ResNet-34 model's [13] four most mistaken class pairs in CUB-200-2011 and scraped Flickr for 1,200 new images for each of these classes. We center-cropped all images to 224×224 and mean-variance normalized using ImageNet statistics [10]. We split the 1,200 images per class into train/validation/test sets of 1000/100/100 images. Fig. 2 presents sample images, the confusion matrix for a ResNet-34 model trained on this data, and example saliency maps for a correct prediction.

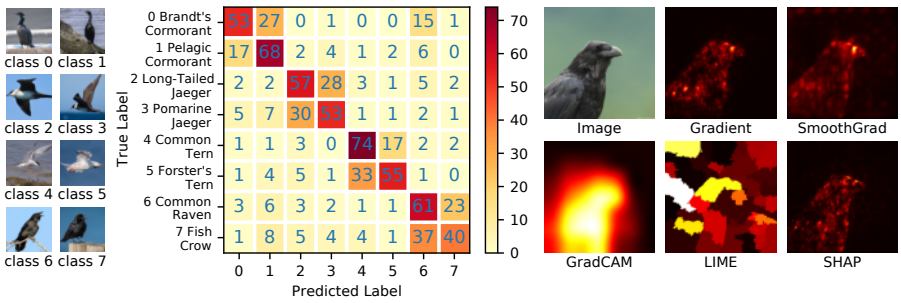

Figure 2: Left: Dataset samples. Middle: Test set confusion matrix of a ResNet-34 model. Right: Examples of five different saliency maps for a correct prediction (Fish Crow).

Input Manipulations: We select five image manipulations which represent artifacts that could be accidentally introduced in a dataset collection process: blurring, brightness change, hue change, pixel noise, and watermark. Details and examples of each manipulation are in App. C.1.

Saliency Maps: We evaluate 5 standard saliency map methods: Gradient [32], SmoothGrad [33], GradCAM [29], LIME [26], and SHAP [22]. These methods are described in detail in App. C.2.

Experiments: We choose 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. All experiments are binary classifications. Sec. 5.4 also uses pairs of distinct species (e.g. common tern and fish crow).

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

5.1 Attr% for Various Attribution Methods and Manipulations

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

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

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

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

The models successfully learned all manipulation types, as demonstrated by the high test accuracies; however, none of the methods consistently scores $\%\text{Attr} \approx 1$, as expected. Further, these manipulations are not all 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 consistently better detected than its *absence*, likely because the model implicitly localizes objects (i.e. the watermark) [7] and predicts a default negative class if it fails to do so. It is also easier 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 absence are needed.

5.2 Attribution vs. Test Accuracy

Question: How does $\%\text{Attr}$ 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 rely more and more on knowledge of manipulation. As a result, we should expect $\text{Attr}\%(F_{\phi_\cup})$ to also 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. 3 (right) plots the lines representing the progress of $\%\text{Attr}$ vs. test accuracy (complete results in App. C.6). SHAP with watermark shows the most consistent and expected increase in $\%\text{Attr}$ with test accuracy. For other saliency maps and feature

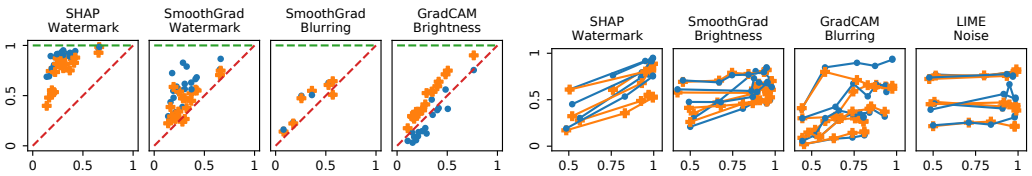

Figure 3: Left: $\%\text{Attr}$ (y-axis) vs. $\%\text{ER}$ (x-axis), complete results in Fig. 10 of App. C.3. Right: $\%\text{Attr}$ (y-axis) vs. test accuracy (x-axis), complete results in Fig. 13 of App. C.6. Blue circles and orange crosses are for images with and without the manipulation. Green horizontal line indicates the saliency map with $\text{Attr}\% = 1$ and red diagonal line indicates a random saliency map.

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. C.4. For example, the visibility for a watermark refers to its font size. As before, the labels are reassigned with $r = 0.5$ and manipulations applied to the positive class.

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 larger gradients in the decision space. From a practical perspective, it is more important for them to identify less visible manipulations, since the others are more likely to be detected by human eyes as well during dataset visualization.

Results: Fig. 4 (left) plots each group of five runs as a line, with visibility level on the x-axis and $\%\mathrm{Attr}$ on the y-axis (complete results in App. C.4). Except for SHAP on watermark, other methods do not show consistent trend of $\%\mathrm{Attr}$ 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 label reassignment is correlated with the 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. 2) and two of distinct ones (e.g. class 5 vs. 6). These variations total $5 \times 6 \times 4 = 120$ 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 $\%\mathrm{Attr}$ with increasing r is reasonable. However, the Shapley value [27] 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. For (near-)perfect classifier with $p \approx 1$, we have their Shapley values (normalized to have a sum of 1) satisfy $\bar{v}(F_M) \geq 1.5 - r$, and $\bar{v}(F_O) \leq r - 0.5$, as derived in App. C.5. In addition, for distinct class pairs, the model can better utilize the more distinct original image features, resulting in lower attribution $\bar{v}(F_M)$ on manipulated features.

Results: All models achieve test accuracy of over 95%. Fig. 4 (right) plots $\%\mathrm{Attr}$ vs. r (complete results in App. C.5). 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 $\bar{v}(F_M) \geq 1.5 - r$, 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 theoretical Shapley values should mostly concentrate on F_M for the (near-)perfect model, with $\%\mathrm{Attr}$ close to 1. As r increases, the model can choose to rely more heavily on F_O for its prediction, resulting in larger allowable ranges of $\%\mathrm{Attr}$.

For watermark manipulation, SHAP shows clear decreases in attribution value as r increases, while gradient also tracks the predicted range, but only for the class with the manipulation. This trend is not

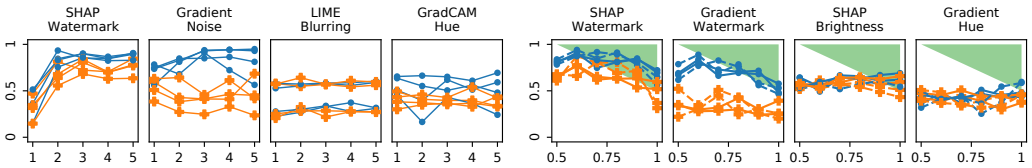

Figure 4: Left: $\%\mathrm{Attr}$ vs. feature visibility , complete results in Fig. 11 of App. C.4. Right: $\%\mathrm{Attr}$ vs. r, complete results in Fig. 12 of App. C.5. Blue/orange colors indicate feature presence/absence. Solid/dashed lines represent similar/distinct species pairs. Green shades represent $\%\mathrm{Attr} \geq 1.5 - r$.

seen in other feature types, even if computed by SHAP which uses approximation, and shows relative insensitivity with respect to r. In addition, there does not seem to be a clear difference in attribution values for similar vs. distinct species pairs. Since the Shapley value *axiom* is commonly accepted as generally reasonable, it is concerning that many saliency maps are inconsistent with it, and important to better understand the underlying axiomatic assumptions (if any) made by each saliency map.

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. Therefore, 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 Attention

It is known that certain non-semantic features can heavily influence model prediction (e.g. email header in the Newsgroup dataset [26]). We study whether the word attention score can saliently highlight any potential usage of such non-semantic features. Details of the implemented dot-product attention mechanism are presented in App. D.1.

Dataset: We use the BeerAdvocate dataset [23] as the basis for modification and further select 12,000 reviews split into 10,000/1,000/1,000 train/validation/test sets, without the label information (which is set differently for each experiment).

Metric: The introduced manipulation changes specific words according to the (reassigned) label. The metric is $\%\mathrm{Attr}$ defined on the set of 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, the articles are the only words correlated with the labels and constitute the effective region.

Expectation: Attention of (near-)perfect models should have $\%\mathrm{Attr} \approx 1$ to be valid attributions.

Results: The model achieves over 97% accuracy. Across the test set, $\%\mathrm{Attr}$ on article words is 8.6%. Considering that articles are 7.9% of all words, this is better than random, albeit barely. Fig. 5 visualizes the attention distribution for two reviews, with additional results in Fig. 14 of App. D.2. Each bars represent weights of words in the review. Green bars represent non-articles and orange bars represent articles. The attention on article words either does not stand out from the rest, or only stands out *relative* to their neighbors. Generally, it does not consistently highlight important features.

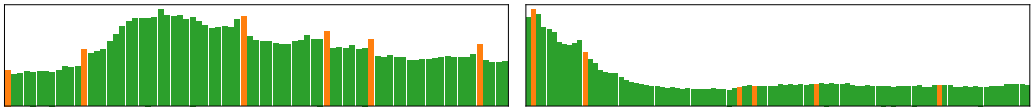

Figure 5: Attention values for each token in two reviews. Orange/green bars represent articles/non-articles. More in Fig. 14 of App. D.2.

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 applied two similar manipulations, with only one of them is correlated with the (reassigned) label. Fig. 6 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. 7 presents attention visualization, with more in Fig. 15 of App. D.3. 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.

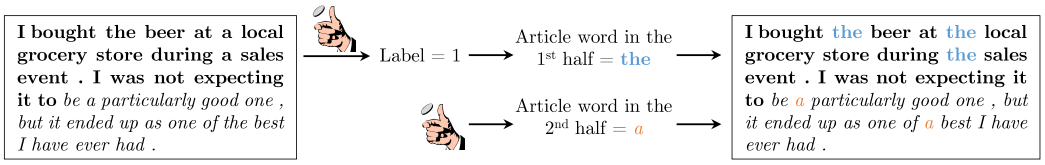

Figure 6: The process to build *CN* dataset for the experiment in Sec. 7.2. First, a review is split into two halves at the midpoint, shown in **bold** and *italics*. Then a label is randomly sampled and assigned to the review. Depending on the label, the articles in the first half are changed to "a" 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.

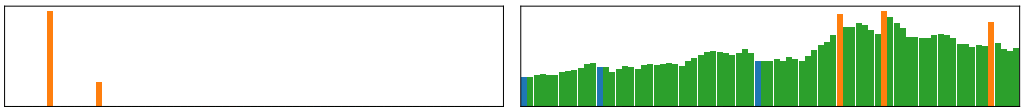

Figure 7: Attention scores for one review in *CN* (left) and *NC* (right) dataset. Orange/blue/green bars represent corr. articles/non-corr. articles/other words. More in Fig. 15 of App. D.3.

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. Tab. 1 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.

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 over 11 training runs, in the format of "$\text{mean}(\%\text{Attr}) \pm \text{stdev}(\%\text{Attr}) \mid \text{word frequency}$". The "Article" dataset is the one used in Sec. 6.1.

6.3 Discussion

Attention is undoubtedly useful as a building block in neural networks, but their *interpretation* as attribution is disputed. Past studies proposed various, and sometimes conflicting, criteria for such interpretation [17, 25, 35], but their correctness is unclear. In our studies, the answer is mostly negative: for most training runs, the attention weights on correlating features at best only stand out *locally*, easily overwhelmed by larger global variations. Therefore, we recommend that future proposals should first be calibrated with ground truth in a controlled setting.

7 Evaluating Text Rationales

In this section, we conducted the same two experiments above (and omit the **Question** and **Setup** descriptions), but for two rationale models, a reinforcement learning (RL) model [21] and a continuous relaxation (CR) model [6]. 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 $\%\text{Sel}$. For a mini-batch of B examples, we use $\lambda \cdot \left| \sum_{i=1}^{B} \text{len}(\text{rationale}_i) / \sum_{i=1}^{B} \text{len}(\text{review}_i) - \text{Sel}\% \right|$, where $\lambda > 0$ is the regularization strength. Incidentally, we also found that the training is much more stable with this regularization, especially for the RL model. Additionally, we 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 [38]: 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 trained models with $\%\text{Sel} \in \{0.07, 0.09, 0.11, 0.13, 0.15\}$, all with over 97% accuracy. We evaluate precision and recall of the trained models and plot them in Fig. 8 (left) according to the actual rationale selection rate, $\%\text{Sel}$, 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 $\%\text{Sel}$ for the test set, and the dashed line enforces $\%\text{Sel}$ individually per review.

Except for the CR model at the lowest $\%\text{Sel}$, 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 $\%\text{Sel}$ regularization done at the mini-batch level. The "faulty" CR model tends to select the first few words regardless, as shown in Fig. 16 of App. E.1, but still selects some article words.

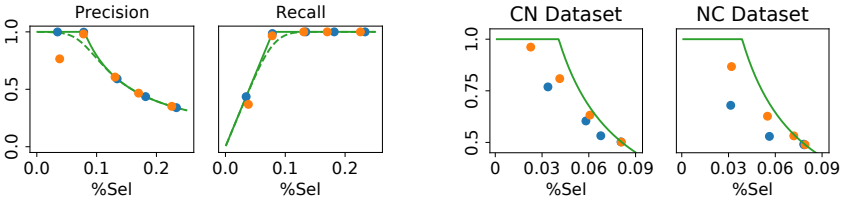

Figure 8: Left: Precision and recall for two rationale models in Sec. 7.1. Right: Precision at different $\%\text{Sel}$ for models in Sec. 7.2. RL model is in blue and CR in orange. The solid and dashed green lines show optimal metric values when $\%\text{Sel}$ is enforced at dataset- and sentence-level.

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 learns to ignore them.

Results: We trained models with $\%\text{Sel} \in \{0.03, 0.05, 0.07, 0.09\}$, all with over 97% accuracy. Fig. 8 (right) 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 (Fig. 17 of App. E.2). However, especially for the RL model, some correlating articles are missed while at the same time non-correlating ones are selected, resulting in markedly less than optimal precision.

7.3 Discussion

The structure of rationale models guarantees that the classifier objectively relies on the rationale features for prediction, but this does not mean that the rationale is contains all the correlating features F_C. 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). 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 [36]. Post-processing methods could be developed to further prune rationales to mitigate this problem.

8 Conclusion

As feature attribution methods are increasingly deployed to ensure the correctness of 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 information about how models trained on carefully-modified, semi-natural datasets make decisions. 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 for improving attribution methods. While our ground truth consists of known dataset artifacts rather than "natural" features, similar artifacts have been found in popular image and text datasets, and have been known to affect model decisions [e.g. 12, 16, 19, 26]. Moreover, any limitations of attribution methods should be carefully studied, and practitioners should be informed accordingly.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in neural information processing systems*, volume 31, pages 9505–9515, 2018.

[2] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

[3] Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, 2019.

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

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, 2019.

[7] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[8] Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions through additive importance measures. *arXiv preprint arXiv:2004.00668*, 2020.

[9] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

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

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018.

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *arXiv preprint arXiv:1806.10758*, 2018.

[15] Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *arXiv preprint arXiv:2010.13924*, 2020.

[16] Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pages 750–782. PMLR, 2020.

[17] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[18] Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*, 2020.

[19] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

[20] Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Sanity simulations for saliency methods. *arXiv preprint arXiv:2105.06506*, 2021.

[21] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, 2016.

[22] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[23] Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE, 2012.

[24] Milad Mostavi, Yu-Chiao Chiu, Yufei Huang, and Yidong Chen. Convolutional neural network models for cancer type prediction based on gene expression. *BMC medical genomics*, 13:1–13, 2020.

[25] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, 2020.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, pages 1135–1144, 2016.

[27] Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[28] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017. doi: 10.1109/TNNLS.2016.2599820.

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[30] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12, 2019.

[31] Ke Si, Ying Xue, Xiazhen Yu, Xinpei Zhu, Qinghai Li, Wei Gong, Tingbo Liang, and Shumin Duan. Fully end-to-end deep-learning-based diagnosis of pancreatic tumors. *Theranostics*, 11 (4):1982, 2021.

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

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

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

[35] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

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

[37] Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*, 2019.

[38] Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4085–4094, 2019.

[39] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Broader Impact

Automated models need to be held to high standard to ensure that predictions are really "right for the right reason". Since a full theoretical understanding of neural models is currently out-of-reach, interpretability methods are often tasked for such purposes. Our work provides a set of computational criteria to identify flaws in these methods. However, it is important to note that similar to other works [1], the criteria are necessary, but not sufficient, conditions. In other words, failing our criteria indicate issues with the tested method, while passing them does not automatically guarantee that they are perfectly working. Therefore, we consider our work as a step toward comprehensively evaluating interpretability methods, rather than the final definitive solution.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] in the section above.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] as supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] in the code, and using the default ResNet-34 setting.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] did not repeat model training.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] training time is not a concern as the dataset is pretty small. A training run can typically finish in 10 minutes on GeForce GTX 1080.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] but the majority are under Creative Commons or Public Domain license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] only public images on Flickr are downloaded.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] no personally identifiable information or offensive content is contained.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Do Feature Attribution Methods Correctly Attribute Features? (/pdf?id=k_wRceOjE9f)

Anonymous

21 May 2021 (modified: 02 Jun 2021) NeurIPS 2021 Conference Blind Submission Readers: Conference, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Reviewers, Paper3636 Authors

Keywords: interpretability, explainability, feature attribution, evaluation of interpretability

TL;DR: We "unit test" several popular feature attribution algorithms for CV and NLP models against known highly important features to see if they can identify them; (un)surprisingly, they mostly can't.

Abstract: Feature attribution methods are exceedingly popular in interpretable machine learning. They aim to 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. The lack of ground truth for feature attribution particularly complicates evaluation; to address this, we propose a dataset modification procedure where we construct attribution ground truth. Using this procedure, we evaluate three common interpretability methods: saliency maps, rationales, and attention. We identify several deficiencies and add new perspectives to the growing body of evidence questioning the correctness and reliability of these methods in the wild. Our evaluation approach is model-agnostic and can be used to assess future feature attribution method proposals as well.

Supplementary Material: zip (/attachment?id=k_wRceOjE9f&name=supplementary_material)

Submission History: Hidden as part of the resubmission bias experiment

Revealed to AUTHOR INFORMATION REDACTED

Add **Official Comment** **Withdraw** **Author Survey**

Author Discussion **Committee Discussion** **All**

Reply Type: all Author: everybody Visible To: all readers **15 Replies**

Hidden From: nobody

[−] A quick note to AC

NeurIPS 2021 Conference Paper3636 Authors *AUTHOR INFORMATION REDACTED*

Official Comment by Paper3636 Authors 20 Aug 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Authors

Comment:

(This note is directed at AC and not visible to the reviewers)

Hello AC,

Thanks for organizing the review effort for our paper. We are glad to see that the reviewers gave us several valuable suggestions. However, we are a bit concerned about the review by reviewer MVVg, who, despite giving a confidence rating of 5, we believe to have several misunderstanding of our technical proposal, specifically in point 1a, 1b, 5 and 7 raised in the review. We responded to each point in the review and hope that it could clarify things up. We would appreciate if you could follow up with the reviewer during committee discussion to ensure that any misunderstanding is addressed. And we are definitely happy to answer any further questions from this (and other) reviewer.

Best, Paper Authors

Add **Official Comment**

[−] Ethics Review of Paper3636 by Ethics Reviewer pBgP

NeurIPS 2021 Conference Paper3636 Ethics Reviewer pBgP

Ethics Review by Paper3636 Ethics Reviewer pBgP 11 Aug 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Reviewers Submitted, Paper3636 Ethics Reviewers Submitted, Paper3636 Authors

Ethical Issues: No

Ethics Review:

The reviewer e7P6 mentions that the limitations of the method are not discussed. However, best to my understanding, such limitations are not related to ethical dimensions. Therefore, I did not find any ethical issues with the paper.

Issues Acknowledged: Yes

Issues Acknowledged Description:

The limitation of the technique (about natural modification) has been mentioned in Line 353 and such is clarified by the authors.

Recommendation:

Authors can highlight the limitations much before the conclusion which will be also useful to scope the work.

Add **Official Comment**

[−] Ethics Review of Paper3636 by Ethics Reviewer mz8x

NeurIPS 2021 Conference Paper3636 Ethics Reviewer mz8x

Ethics Review by Paper3636 Ethics Reviewer mz8x 10 Aug 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Reviewers Submitted, Paper3636 Ethics Reviewers Submitted, Paper3636 Authors

Ethical Issues: No

Ethics Review:

No apparent ethical concerns identified

Issues Acknowledged: Yes

Issues Acknowledged Description:

N/A

Recommendation:

N/A

Add **Official Comment**

[−]

Official Review of Paper3636 by Reviewer wStr

NeurIPS 2021 Conference Paper3636 Reviewer wStr

Official Review by Paper3636 Reviewer wStr 25 Jul 2021 (modified: 03 Sept 2021)

Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Summary:

The paper adds to the growing body of works on evaluating model explanations. The main contribution of the paper is to propose a test to generate "ground truth" attributions. This procedure operates by first flipping the labels in the training data (with some probability) to essentially break the feature-to-class mapping, and then injecting some class conditional artifacts into specific areas of the inputs. Since these artifacts are class-conditional, these are the only discriminators between different classes. A good explainability method should attribute these (and only these) areas of the input. The paper then carries out an evaluation, concluding that SHAP performs the best among the feature attribution methods and attention, as expected, does not perform so well.

Main Review:

The algorithm proposed in the paper is quite intuitive and the reasoning behind it has been known to the interpretable ML community (as pointed out in line 355). Nonetheless, the paper does a good job of formalizing it. The paper is easy to follow. Overall, it seems like a nice contribution to the growing literature on evaluating explainability methods. There are still some places (mostly experiments) where the paper can be improved. See the detailed suggestions below:

1. One of the biggest points that the paper should address is the fact that the feature attributions are not additive for many explainability methods used in the experiments (e.g., SHAP follows the additivity property while LIME does not). However, the metric of Attr% seems to assume additivity. How does this impact the results? Is this the reason why the gradient based methods seem to do worse than SHAP (Fig 3)? Perhaps a "softer" metric here would help? For instance, instead of using Attr%, one could count how many of the pixels contained inside $\phi_U(x)$ are within top-X% pixels of the attributions method?
2. Related to the previous point, it would help to evaluate methods like IntegratedGradients (IG) and Layerwise Relevance Propagation (LRP) that follow this additivity property. Also, since Gradient and SmoothGrad are quite similar (minus the averaging over multiple perturbed images), it might help to drop Gradient altogether and use the space for IG and LRP.
3. Since LIME uses super pixels, how does the paper ensure that the $\phi_U(x)$ is included in one super pixel? If it spans multiple super pixels, how does that impact the Attr% metric?
4. Regarding line 177, what was the observed test set accuracy of these models?
5. I am not sure about the statement in line 183. Sure when the full image is considered, Attr% will be 100 even for bad interpretability methods. However, even when the size of ϕ_U is quite small, say 10% of the image, we would still expect good interpretability methods to have Attr% = 100. So the %ER and the diagonal lines in Figure 3 seems to be somewhat misleading. Also, it is quite surprising that the performance of gradient based methods is close to random in Figure 3. Again, is this just an artifact of the lack of additivity property?
6. Regarding "their correctness is unclear" statement in line 298, I am not sure what is meant here. Could you please clarify?

POST REBUTTAL COMMENTS

I thank the authors for their response. After reading the comments, I decided to keep my current score. I would highly encourage the authors to add the top-X% metric, look into using methods like IG which have nice theoretical properties and hence are worth comparing. Finally, I am not sure if assigning the attribution of a superpixel to all the pixels inside it is the right choice. This way, pixels contained in superpixels with large area might get unfair disadvantage (a very large superpixel will get a large attribution just due to its size). So normalizing by size might help.

Nonetheless, I think the paper formalizes a very intuitive test which "makes a lot of sense" and provides a nice complement to evaluating models based on their theoretical properties.

Limitations And Societal Impact:
I do not see any immediate negative impact.

Needs Ethics Review: No

Time Spent Reviewing: 6

Rating: 6: Marginally above the acceptance threshold

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Add **Official Comment** **Review Feedback**

[−] **Author Response**

NeurIPS 2021 Conference Paper3636 Authors *AUTHOR INFORMATION REDACTED*

Official Comment by Paper3636 Authors 10 Aug 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Comment:

Thank you for your comments. We address them by the point, but we could not post an updated draft per the conference submission policy.

1. We would like to clarify that the motivation for the Attr% definition is, and should be, independent from the explainability method used. Instead, it should arise out of the intended use of feature attribution methods -- they are used to suggest “important” input regions, so that we could check if they are making the correct decision for the correct reason (rather than focusing on spurious correlation, for example). Thus, an additive attribution inside the ground-truth region is a reasonable definition to evaluate the accuracy (i.e. concentration and sharpness) of the “highlighting” behavior. If the users prefer to use the Top-X% metric (which does have a hyperparameter dependency), this is also fine. It’s just that they do a “dry run” on this metric using our dataset modification method before trusting whatever explainer to be correct. We will add the discussion to the revision and note alternatives to the Attr% metric.

   As for poor performance of gradient methods, we believe that they are too “local” (infinitesimally local to be exact), and neural nets are known to have quirkiness in the decision surface as evidenced by adversarial examples.
2. We studied the gradient method because it is arguably the first saliency map proposed (back in 2013) and still widely used, and SmoothGrad because it is widely considered as an improved version of gradient. We will perform some evaluation with IG and LRP, and also focus on open-source our code as a high-quality extensible toolbox so that the general method could be useful for any explanation method as well.
3. We follow the default LIME implementation by the original authors (https://github.com/marcotcr/lime (https://github.com/marcotcr/lime)), and all pixels in the super-pixel receive the same attribution (which is the regression coefficient). Apparently, a high-quality LIME explanation depends on the correct segmentation result in the first place, which is implemented by the quick shift algorithm by default and indeed fails to work

correctly on many spurious correlation types. We consider this to be a drawback for LIME, especially for detecting spurious correlations in our demonstrated use-cases. Some additional discussions on LIME are in Line 192.

4. They are almost perfect, with all of the used ones achieving over 95% accuracy (Line 182).
5. The green horizontal line (very top in the figure) represents the desirable behavior of the ideal saliency map. The red diagonal line is for the random saliency map (so yes, some methods are worse than random), as described in Fig 3 caption. We discussed the issue with gradient in Point 1 above.
6. The paper is motivated by the observation that a direct evaluation of explanation in general would result in a circular argument, and thus evaluations are often done with indirect metrics, such as prediction change under feature removal, removal-and-retraining (ROAR), agreement with human judgement, etc., which are flawed in various ways, as discussed in Related Work. Similarly, various indirect evaluations of attention have been proposed, such as the adversarial construction proposed in the "Attention is Not Explanation" paper. However, fundamentally, they do not evaluate the most basic question of "do attentions truly highlight important words in text?" Thus, combined with seemingly contradictory evidence in "Attention is Not Not Explanation", that's why we say that the correctness of these evaluation methods is unclear, and propose to directly evaluate with ground truth information, and indeed highlight that the feature-attribution interpretation of attention is often problematic.

Thanks again for your comments and please let us know if you have additional questions.

Add **Official Comment**

[−] # Official Review of Paper3636 by Reviewer e7P6

NeurIPS 2021 Conference Paper3636 Reviewer e7P6

Official Review by Paper3636 Reviewer e7P6 19 Jul 2021 (modified: 17 Aug 2021)

Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Summary:

This paper provides evaluation metrics for evaluating attribution maps for image and text classifier, and also generated rationale. The idea is to artificially inject spurious correlation into the input and test whether explanation methods are able to select these correlating features. Label noise was added to the original labels to make sure models rely on the artificially-injected correlating features. The authors extensively studied image and text classifiers and text-rationale generators. Five attribution methods in image classification were tested, Attention was used as an explanation method in NLP experiments. The main conclusion stated is most explanation methods do not pass the sanity check (%Attr is not near 1).

Main Review:

++ Originality

- This paper in principle is very similar to BIM (later re-named into BAM) by Yang and Kim 2019 (cite [36]). It is reasonable to call this an extension of BIM (e.g. BIM v2); however, the authors claim this to be an independent, new benchmark. Little to no reasons were provided.
- A lot of papers have tested their methods on synthetic data to confirm the explanations select the (expected) correlated features. However, a key difference between this work and others is the consideration of effective region (ER), which appears to be well-thought-out!

++ Quality

- L182: If models achieve 95% test-set accuracy (not 100%). How do you guarantee what features it uses for all the cases? (i.e. how can we expect %Attr = 1).
- Sec. 5.3 has no motivation provided. To humans, one may expect %Attr increases with visibility. However, this need not be the case with CNNs as we know human and computer vision differs significantly (e.g. in perceiving adversarial examples).
- All the main details of Sec. 5.3 are in the appendix.

- I like the authors' intention of performing thorough experiments.

++ Clarity

- The authors want to discuss other tests that evaluate the correctness and reliability of attribution methods e.g. [1]. For example, LIME (and SHAP) has the reproducibility issue (due to the random sampling), which may cause %Attr to never reach exactly 1.0.
- What are the motivations for the five image manipulation techniques?
- The paper is packed and a lot of key details are in the Appendix, which is not required to read by the reviewers.

[1] SAM: The Sensitivity of Attribution Methods to Hyperparameters. CVPR 2020

++ Significance

- "Attention is not an explanation" has been found to be the case in many recent studies. Especially, in Transformers where there are multiple attention heads, each serving a different job. Therefore, the result that Attention is not performing well based on this dataset/metric is not too surprising or the most useful in assessing the proposed metrics.
- To be absolutely convincing, authors need to train models to 100% accuracy on the test set, and using the current data-generation to ensure models must rely on the artificially-injected features (a.k.a. "image-manipulation" techniques). Because besides groundtruth explanations, you need groundtruth networks.
- I do not see the NLP experiments using Attention + text classifier to add any new insights (given the existing image classifier experiments).
- While the text injection image-manipulation is well-motivated, the choice of the rest of the artificial features was arbitrary and do not provide any specific insights.
- In overall, this paper is very similar to BIM/BAM [36]. Therefore, the added value to the community is limited.

===== After Rebuttal ===== I'm satisfied with the author's responses. I think this work is an improved version of BAM and I like the work's thoroughness. So I tend to accept this paper. Rating changed to 7 (from a 5).

A main issue with the paper is that the paper is too dense and many key details are in the Appendix. So actually, the most useful version of this paper may be in an arXiv or a journal format (many pages).

Limitations And Societal Impact:
No limitations about this work were discussed. Societal impact: N/A

Ethical Concerns:
N/A

Needs Ethics Review: Yes
Time Spent Reviewing: 2.5
Rating: 7: Good paper, accept
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Add **Official Comment** **Review Feedback**

[−] **Thank you for your review update!**

NeurIPS 2021 Conference Paper3636 Authors *AUTHOR INFORMATION REDACTED*
Official Comment by Paper3636 Authors 18 Aug 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Comment:

We are glad to see that our response clarifies things up and thank you for updating your review and score. We understand your concern that many details of the paper are left to the appendix. Indeed, it is hard to fully fit a rigorous and comprehensive evaluation framework for general feature attribution methods within 9 pages. We made the intentional trade-off of focusing on the methodological rigor, which, in our opinion, is the biggest issue with existing more ad hoc proposals on individual scenarios. Since both the image and the text experiments can be seen as "instantiations" of our framework, we chose to leave the details to appendix. Nevertheless, if the paper is accepted, we would put review clarifications and the more important experimental details in the additional page allowed by the camera-ready version.

More importantly, since ML interpretability is such a fast moving field, and many new explainers are still evaluted with less-than-optimal and potentially flawed metrics such as human annotation agreement, it is crucial for the community to be aware of their limitations, discussed in the Related Work, and establish the importance of rigorous and objective evaluation. Thus, the NeurIPS conference, with its fast turn-around and wide publicity, is the ideal place for the dissemination of ideas. We plan to extend this work to include more explainers, more domains, and an easy-to-use package for off-the-shelf evaluation, and post updates on arxiv. If we feel that additional technical contribution is significant, we will also submit the extended version to a journal.

Add **Official Comment**

[−] **similarity and differences w.r.t. [2,36]**

NeurIPS 2021 Conference Paper3636 Reviewer e7P6

Official Comment by Paper3636 Reviewer e7P6 03 Sept 2021

Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Comment:

Thanks author for your reply!

Could you also elaborate more on both the **similarities** and **differences** w.r.t. [2] and [36]? I think this is an important point that might not be discussed sufficiently in the Related Work. Therefore, it could hinder readers and here, reviewers, from more fully appreciating this work.

L58-69 talks about 2 subtle points, one w.r.t. [2] and one w.r.t.[36]. I wonder (1) what are the contributions in this submission that were not in [2] and [36]; and (2) what are the similarities in findings/evaluation-method between this submission and those two?

Add **Official Comment**

[−] **Re: similarity and differences w.r.t. [2,36]**

NeurIPS 2021 Conference Paper3636 Authors *AUTHOR INFORMATION REDACTED*

Official Comment by Paper3636 Authors 03 Sept 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Comment:

Thank you for your question. Below are some more detailed explanations that we are looking to include in the appendix or the extra page allowed by the final version.

In terms of similarity, the high-level idea is similar: for a natural dataset, we do not know individual feature contribution, but generally highly correlated and easily discriminative features should have high contribution, and we realize this notion by injecting such features directly. All three works can be seen as operationalizations of this idea.

However, this high-level idea needs two caveats, which set our paper aparts from [2] and [36]. First, features that we think to be "easily discriminative" may not be be considered as such by neural nets. For example, (Geirhos et al.) showed that they have particular inclination toward textures. Thus, we can't really trust them to actually pick up and use our introduced features, *unless* that we are assured that focusing on other features cannot achieve the performance that the model is achieving now. In this aspect, we concretely demonstrated that network trained in [2] *can* achieve good performance when using the "other features" exclusively. In [36], the authors demonstrated this principle, but used out-of-distribution data, so it is not clear whether the failure is due to achieve good performnace is really due to the network indeed ignoring the "other features" or due to instability of out-of-distribution extrapolation (as discussed in our original response).

The second caveat is that the "other features" cannot have any information on the injected features. For example, in Fig. 4 (Page 6) of https://arxiv.org/pdf/2011.05429.pdf (https://arxiv.org/pdf/2011.05429.pdf) (the extended version of [2]), the saliency map that perfectly crops out the foreground dog shape could imply that the network actually uses the contour of the dog, which is inconsistent with their target conclusion that "foreground doesn't matter in highly correlated background dataset". In other words, background with a dog contour cropped out is very informative to the fact that the foreground is a dog. The similar cropping procedure is used in [36] as well. Another loophole is discussed in the Remark of the paper, where the lack of one feature at one place could imply the presence of another feature at another place. By comparison, in our work, we define the joint effective region such that all injected features are contained within, and then we are assured that the features outside of the region absolutely cannot contribute to the high performance, and then evaluate Attr% for that region.

Finally, compared to [2] and [36], our framing is more general. We start with the goal of evaluating feature attribution in general, and proposes our domain-agnostic framework in Fig. 1 and Sec. 3 and 4. Thus, it would be straightforward to instantiate our idea to other domains such as graph, speech, or time-series data. By comparison, both [2] and [36] are focused on the foreground/background patch setting, and introduces necessary concepts under this context. In addition, we also proposed ways to evaluate feature-selection style attributions, more common for text models, as discussed in Line 99.

Overall, we believe that our work adds value to the community as a more rigorous and general framework than existing ones, and such rigor is crucially needed as we demonstrated in particular that the formulation in [2] does not actually support its conclusion (Line 63).

(Geirhos et al.): ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, ICLR 2019, https://arxiv.org/abs/1811.12231 (https://arxiv.org/abs/1811.12231)

Add **Official Comment**

[−] **thanks; suggestion to add this reply to final paper**

NeurIPS 2021 Conference Paper3636 Reviewer e7P6

Official Comment by Paper3636 Reviewer e7P6 03 Sept 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Comment:

Thanks authors for your reply!

I'd love to see your reply/discussion added to the final paper.

Add **Official Comment**

View 1 More Reply →

[−] **Author Response**

NeurIPS 2021 Conference Paper3636 Authors *AUTHOR INFORMATION REDACTED*

Official Comment by Paper3636 Authors 10 Aug 2021 (modified: 10 Aug 2021)

Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Comment:

Thank you for your comments. We address them by the point, but we could not post an updated draft per the conference submission policy.

Originality:

Indeed, similar sanity checks, from the original paper by Adebayo et al. to more recent ones including BAM and cite [2, 19] in the paper, have been proposed. We do not claim the use of ground truth information as the first of its kind (c.f. last paragraph of Related Work), but many of them, including [2] and BAM are flawed in one way or another, as explained in Line 62-64, as well as Line 141. Specifically, for the BAM work mentioned by the reviewer, there are two issues, both of which are addressed in our benchmark. First, despite the foreground (i.e. object) injection, the label is still perfectly correlated with the background (i.e. scene). Thus, it is unclear whether the foreground should have high attribution values, since the model could rely exclusively on the background. We address this issue via our label reassignment (to either a completely random value or a less correlated value in Sec. 5.4), to ensure that the model could not use the original (i.e. background in their case) features. Note that while the BAM paper demonstrates low model accuracy on $X_{\varnothing,s}$ for weakly correlated features (i.e. first couple columns) on the bottom table in Fig 2, this dataset itself is out-of-distribution (OOD), as all the training images are in the form of a combined foreground-background rendering, while the dataset contains only the background. Thus, the metric is complicated by OOD extrapolation behavior of the model, which is hard to characterize, and argued against being used in the ROAR paper (cite [13]). By comparison, in our model, no OOD data are used and the models are fairly evaluated.

Second, assuming that the model does use foreground information, BAM only evaluates attribution inside the specific foreground patch in an image. However, in the Remark of page 4, we give a counterexample in which a model could use the absence of a feature as the reason for decision, which would result in a false negative by the BAM evaluation procedure for the saliency map that correctly highlights this reasoning, which motivates our definition of the joint effective region (ER).

These tests are not designed for sufficiency (i.e. explanations that pass their assessment are not guaranteed to be correct). Given this, and the fact that they are not necessary either (i.e. they can falsely mark correct explanations as incorrect), their utilities as neither necessary nor sufficient tests are concerning. To address the above problems, in Sec. 3, we start from first principles of how explanations are, or should be, used in practice, and construct our dataset modification and corresponding evaluation metrics rigorously from these principles, including the formal notion of effective region (ER), as the reviewer has pointed out. Therefore, we do not consider our work as an extension of BAM, but an advocate for more rigorously evaluating attributions.

Clarity:

1. For what it's worth, the majority of the test runs in the >=95% category do achieve 100% test accuracy. However, even then, it could be argued that the true accuracy may not be 100%, maybe 99% or 98% and we just get lucky on the test set. Furthermore, if we want to know if a feature is used or not *at all*, there is a very simple way: just perturing that feature value and see if the predicted probability or the logit changes. If so, then arguably, the model "uses" that feature.

   However, practically we are not interested in every bit of influence by every feature. Maybe a feature very infrequently and mildly affects the decision, while another feature is absolutely important. Do we want the feature attribution to assign equal values on both, or do we want it to highlight the latter? From a practical perspective in terms of helping people understand the model by e.g. recognizing potential spurious correlation, we would argue that the latter is far more desirable. Specifically, in the experiment, the >=95%

accuracy is solely due to the injected manipulation, while the original features could be responsible for at most the remaining <=5% errors. Suppose that the task is cancer detection. A 95% accurate model would be well-trusted by human doctors, yet it is absolutely important that people understand that the predominantly influential feature is the watermark, despite potentially minor influence by other features. Thus, even if we would not expect Attr% being 1, it is reasonable to require it to be very high. We will elaborate and tone down the requirement in the text.

2. We agree with the human vs. CNN difference. Furthermore, we actually know that the CNN model successfully detects the manipulations since the accuracy is very high. Thus, the only question is whether the *explainer* could identify that the model detects and uses the manipulation effects. From a practical perspective, obviously, we would want the explainer to succeed for all visibility levels, but arguably, it is more important for them to identify less visible manipulations, since the more visible ones are more likely to be detected by human eyes as well during dataset visualization. Unfortunately, we found that no method works consistently better at low-visibility level than high-visibility level, which further increases the doubt of whether these methods could really help us identify subtle but important features.
3. Indeed, we are quite limited by space since we want to demonstrate our approach as a general evaluation method applicable to both image and text domains, unlike many existing works that mostly focus on image. We would appreciate the reviewer's input on restructuring, or if they believe that the entire 5.3 could be delegated to the appendix as a supplementary study.

Clarity:

1. Indeed, and we believe that such an issue should definitely be known to users of these methods. In practice, we try to decrease the variance as much as possible by using a large number of samples in LIME and the entire test set as the background distribution in SHAP. With these hyperparameters, we can confirm that they *seem* to generally work well on the original training set without the injected manipulation (Fig 2).
2. The common theme is they represent common but unsuspecting effects.
   - Peripheral blurring could be due to either camera in motion or an intentional, artistic post-processing effect to highlight the main subject of the image.
   - Brightness change could be due to times of the day, or the use of artificial light to illuminate the subject.
   - Hue change could be due to errors in conversion of different color space encodings (https://en.wikipedia.org/wiki/RGB_color_space#/media/File:CIE1931xy_gamut_comparison.svg (https://en.wikipedia.org/wiki/RGB_color_space#/media/File:CIE1931xy_gamut_comparison.svg)), which may result in color loss or distortion.
   - Pixel noise could be due to lossy compression or data loss during transmission.
   - Watermark is a commonly employed technique to attribute the author/organization of the image.

   Normally, none of them should be expected to correlate with the label. However, especially with image scraping on the web and crowdsourced dataset construction, it is possible that some spurious correlations leak into the final dataset.
3. We thank the reviewer for taking the time to read the appendix. Indeed, we found ourselves quite limited by space, but we hope that Sec. 3 and 4 provides sufficient and clear expositions of our core technical ideas, and we would appreciate any feedback on presentation improvement.

Significance:

1. While this paper argues the invalidity of attentions as explanations, another paper “attention is not not explanation” provides reasons why they could still be explanations. However, fundamentally, all evidence is indirect, in that they propose properties that are believed to be correlated with explanation validity, but are never tested in a controlled setting. We argue that such a controlled setting is necessary, due to various flaws in many indirect methods (e.g. prediction change or retraining accuracy under feature removal, agreement with human judgement) in Line 44 - 58.

   For more discussions, we refer the reviewer to the first point of our response to reviewer wStr above.
2. The accuracy problem is addressed in the first point of the “Quality” section response.
3. As discussed earlier, most existing efforts focus on the vision domain. However, text explanations also suffer from similar potential problems (both pointed out by Jain & Wallace, and due to similar mechanics of the explainer such as the use of gradient information), yet there is little systematic study. Thus, we would like to

pose our paper as a general method to evaluate feature attribution methods to maximally benefit the community. We focus on attention as it has generated particularly interesting discussions from the two papers arguing whether attention “is not”, or “is not not” explanation.

Furthermore, a type of saliency map is computed by a “feature selection” style procedure, such rationale models as discussed in Line 99, which requires different criteria for the reasons explained there.

4. We have addressed this above, and will add them in the revision.
5. We elaborated on key differences in the first point of our response.

Limitations and Societal Impact:

We did discuss the limitation of our method, in that it is best suited for evaluating the explainer's capability to detect artifacts due to our ground truth construction procedure, in Line 353. We will focus future work on how to make the artifact injection look “natural”, potentially using generative models.

Thanks again for your comments and please let us know if you have additional questions.

Add **Official Comment**

[−] # Official Review of Paper3636 by Reviewer MVVg

NeurIPS 2021 Conference Paper3636 Reviewer MVVg

Official Review by Paper3636 Reviewer MVVg 14 Jul 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Summary:
This paper proposes a dataset modification, including label reassignment and input manipulation, to construct the attribution ground truth to facilitate the evaluation of attribution methods. This paper conducts a series of experiments to reveal the weakness of several attribution methods.

Main Review:
[Strengths]

1. This paper focuses on an important issue, i.e., the evaluation of the attribution methods.
2. Authors apply the proposed evaluation metric on various experiments, including both the NLP task and the CV task.

[Weaknesses]

1. Issues in formulation.

a) For two important variables F_C and F_N (sets of informative features and non-informative features) used in the evaluation metric (in Lines 80-81), theoretically, these two variables can be relatively accurately determined when we set r=0.5 (see Line 115). However, for a more general case 0<r<1, F_C and F_N cannot be faithfully determined, because informative features include both the manmade feature (after dataset manipulation) and some features in the original dataset. This is a theoretic problem.

b) Authors then consider a special case when r=0.5. In this case, the manipulation is equivalent to a backdoor attack [cite 1]. Therefore, authors are suggested to clarify the relationship between the proposed method and the backdoor attack, and cite the related papers. Besides, the model trained on the manipulated dataset, rather than normal datasets. Thus, there is no mechanism to ensure that the evaluation based on the manipulated dataset reflects the behavior of attribution mehtods on models trained on normal datasets. [cite 1] T. Gu, K. Liu, B. Dolan-Gavitt and S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," in IEEE Access, vol. 7, pp. 47230-47244, 2019, doi: 10.1109/ACCESS.2019.2909068.

c) In line 180, the authors claim that when r=0.5, “a perfect performance thus indicates that the model relies exclusively on features inside $\phi_{\cup}$.” Based on this, the authors construct the attribution ground truth. However, can the authors provide theoretical proof for this claim? I think this may be a big problem, because a faithful evaluation metric should not be defined under the assumption of "a perfect performance." In real experiments, a model cannot achieve a perfect

performance. For example, in line 182, the authors select models with test accuracy over 95%, but we still cannot prevent the model from using features outside the effective region for classification. In this case, the ground truth cannot be faithfully determined.

2. The authors only evaluate 5 attribution methods for image saliency maps in this paper, which are quite limited for a study focusing on the evaluation. The authors are suggested to conduct a more comprehensive evaluation on more attribution methods.
3. Inequalities in Line 232 for the Shapley value may be incorrect. In the proof of Appendix C.5, when r is not equal to 0.5, theoretically, the model does not exclusively use the manmade feature after the manipulation, but also uses some features in the original dataset. This means that F_O also contains some features in the original dataset for classification. Therefore, the equality $a(F_M) = a(F_M \cup F_O)$ does not hold. Please explain this issue.
4. The construction of CN and NC dataset in Figure 6 is ad-hoc. Is it a standard experimental setting to split the sentences into two parts, and change articles of two parts differently?
5. In Lines 224-234, authors use the Shapley value to design the evaluation metric. Is the Shapley value considered as the ground-truth explanation here? If so, why do not directly use the Shapley value for evaluation?
6. In Figure 4, the result shows that the SHAP has a better performance than other attribution methods. However, SHAP is designed based on the Shapley value, and the evaluation metric is also designed using the Shapley value. Does the proposed metric have a partiality towards SHAP?
7. This paper only evaluates the correctness and reliability of several existing attribution methods, but does not discuss how to improve feature attribution methods based on the evaluation result.

Limitations And Societal Impact:
Yes

Needs Ethics Review: No

Time Spent Reviewing: 1.5

Rating: 3: Clear rejection

Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Add **Official Comment** **Review Feedback**

[−] **Author Response**

NeurIPS 2021 Conference Paper3636 Authors *AUTHOR INFORMATION REDACTED*

Official Comment by Paper3636 Authors 10 Aug 2021 Program Chairs, Paper3636 Senior Area Chairs, Paper3636 Area Chairs, Paper3636 Ethics Reviewers, Paper3636 Reviewers Submitted, Paper3636 Authors

Comment:

Thank you for your comments. We address them by the point, but we could not post an updated draft per the conference submission policy.

1. (a) Indeed, there is no general consensus on what the attribution value should be for $r \neq 0.5$. However, we did consider (and as the reviewer also pointed out later in the review) the case of r > 0.5, using Shapley value as the “ground truth”. We refer to discussions later in the response for more detail.
   (b) The backdoor attack is a form of adversarial attack that triggers wildly different behavior for out-of-distribution data (i.e. those containing backdoor). In our work, we focus on the interpretability side and nothing is out of distribution: the model is trained on modified data and also tested on similarly modified inputs as well. Thus, we should expect (and indeed see) that the model works well, yet interpretability methods mostly fail to detect the true reason for the high performance.

On natural datasets without ways to intervene, evaluating interpretability methods on “ground truth explanation” is by definition a circular argument (Line 26), and earlier attempts (cite [2, 36] in the paper) to evaluate against “ground truth” also need to modify the dataset in one way or another. However, they are all flawed in certain ways, as discussed in both the Related Work and our response to the Originality comment by Reviewer e7P6 above). Finally, as we discussed in conclusion (Line 353), the setups in our experiment do represent possible accidental dataset corruption which could lead to artificially high test accuracy, and it is absolutely critical that the explanation methods we trust could correctly identify the true reason.

(c) A similar concern is also raised by Reviewer e7P6 in the first comment about Quality, so we give the same response below. For what it’s worth, the majority of the test runs in the >=95% category do achieve 100% test accuracy. However, even then, it could be argued that the true accuracy may not be 100%, maybe 99% or 98% and we just get lucky on the test set. Furthermore, if we want to know if a feature is used or not *at all*, there is a very simple way: just perturing that feature value and see if the predicted probability or the logit changes. If so, then arguably, the model “uses” that feature.

However, practically we are not interested in every bit of influence by every feature. Maybe a feature very infrequently and mildly affects the decision, while another feature is absolutely important. Do we want the feature attribution to assign equal values on both, or do we want it to highlight the latter? From a practical perspective in terms of helping people understand the model by e.g. recognizing potential spurious correlation, we would argue that the latter is far more desirable. Specifically, in the experiment, the >=95% accuracy is solely due to the injected manipulation, while the original features could be responsible for at most the remaining <=5% errors. Suppose that the task is cancer detection. A 95% accurate model would be well-trusted by human doctors, yet it is absolutely important that people understand that the predominantly influential feature is the watermark, despite potentially minor influence by other features. Thus, even if we would not expect Attr% being 1, it is reasonable to require it to be very high. We will elaborate and tone down the requirement in the text.

2. The main goal of this paper is to provide a rigorous framework for evaluating feature attribution methods, rather than a critique of existing methods. Thus, we strive to demonstrate the wide applicability of the framework, including vision and text domains, and soft and hard (i.e. feature-selection-style) attribution. With limited space allowed by the page limit, we believe that we achieved a good coverage of breadth of application domains and breadth of explanation methods for each domain, considering that all the methods we evaluated are very popularly used. If the reviewer has a specific method suggestion, we'd gladly incorporate it.
3. Thanks for pointing this out. This is definitely a subtle point and highlights a (potentially undesirable) particularity of Shapley value definition. The definition is for $a(F_M)$ is the expected performance of the model using only the information of F_M. The computational definition of “using only the information of F_M” is that, suppose the model can distinguish between images with different F_M value, but cannot distinguish images with the same F_M value (since it cannot use features beyond F_M), how good the model can be. To calculate this number, we can “group” images in the distribution according to their F_M value, run the model on images, and shuffle the predicted labels within each group. The final shuffling part essentially makes the model unable to distinguish images with the same F_M (i.e. within each group).

   However, the set-up is that the labels are deterministically correlated with F_M. So a perfect model should predict this label, regardless of F_O values on all images within the group, and the label shuffling does not hurt the performance. Thus, we do have $a(F_M) = a(F_M \cup F_O)$. This is also why F_M and F_O are not symmetric. Again, we do not argue whether this is reasonable or not, but simply that this is how Shapley value calculation should be carried out mechanically.

   At the end of the day, whether we want to accept Shapley value as the definition, in light of this subtlety, is determined by the user of the explanation. We took the affirmative answer here as it is the most common axiomatic definition (as explained in more detail in Point 5 below).
4. We are not aware of such a technique being used, but it does inject an interesting and challenging correlation pattern into the dataset, and indeed reveals limitations of both attention and rationale models (Table 1 and Figure 8 right). Thus, we would love for it to be more widely adopted as a standard benchmark for papers proposing new or improved explanation methods.

5. Compared to the r=0.5 case, arguably the definition for attribution value for r>0.5 is less unanimous and more up-to-interpretation. However, we believe that Shapley value is by far the most agreed upon metric for general feature attribution, as evidenced by both the extreme popularity of the original SHAP paper and numerous proposals to improve it (such as L-Shapley and C-Shapley methods). Therefore, we also used Shapley values. For the question of "why do not directly use the Shapley value for evaluation", we do not fully understand, but to clarify, we took the Shapley values as the ground truth (which is known in our case by explicit computation but is generally intractable to evaluate), and see if current explanation methods (such as SHAP, which approximates Shapley value calculation) produces results that agree with the Shapley value "ground truth".
6. Yes, SHAP is intended to approximate Shapley value. As discussed above, whether Shapley value should be the interpretation is ultimately up to the user. But if we take it as so, which is reasonable due to its popularity and several axiomatically nice features (details in the original SHAP paper), then indeed we should expect SHAP to track the Shapley values. Conversely, we showed that other methods don't track Shapley values that nicely. Do they compute some other axiomatic definitions that have not been discovered, or do they just produce low quality metrics in general? This question needs to be answered before we should trust the other methods, and is a future direction.
7. While directions for improvement are beyond the main scope of the paper, we did provide some initial thoughts in the paper. For example, filter-based CNN methods such as GradCAM may need to better account for "non-object-like" image patches, and perturbation-based methods such as LIME may benefit from ways to construct certain absent artifacts in addition to the current mechanism of purely suppressing present ones, as discussed around Line 191. For rationale models, an additional pruning step may benefit the ease of interpretability, as discussed in Line 344. Attention models are an intrinsic component of the model and more rigid, so we are not sure how to improve it (and several papers argue for not using it as interpretation at all). But we argue for better ground-truth calibration if people still want to use it, around Line 300.

Thanks again for your comments and please let us know if you have additional questions.

Add **Official Comment**

[−] **Your Predicted Probability of Acceptance**

AUTHOR INFORMATION REDACTED

Add **Official Comment**

About OpenReview (/about)
Hosting a Venue (/group?id=OpenReview.net/Support)
All Venues (/venues)
Sponsors (/sponsors)
Frequently Asked Questions (/faq)
Contact (/contact)
Feedback
Terms of Service (/legal/terms)
Privacy Policy (/legal/privacy)

OpenReview (/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through Code for Science & Society (https://codeforscience.org/). We gratefully acknowledge the support of the OpenReview Sponsors (/sponsors).

Description of Changes

We made several changes to both the main text and the supplementary material in response to the review feedback.

1. We toned down the requirement that Attr% needs to be 1 for experiment in 5.1, since the non-correlating features can still contribute to the prediction as evidenced by a non-perfect accuracy. However, they should still be very close to 1, since the contribution of non-correlating features is minimal. This point is discussed in detail in our response to two of the reviewers.
2. We added discussion to top-k selection, which is actually a special case of the feature-selection style attribution, in the last paragraph of Section 3, as requested by one reviewer.
3. We added extensive discussion on the similarities and differences to [2, 36] in Appendix A.1, as requested by one reviewer.
4. We added discussions on how our introduced image manipulations could realistically occur in real world in Appendix C.1, as requested by one reviewer.

We would be happy for this paper to be considered for a journal submission.