Evaluation of Attribution Explanations without Ground Truth

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper proposes a metric to evaluate the objectiveness of explanation methods of neural networks, *i.e.,* the accuracy of the estimated importance/attribution/saliency values of input variables. This is crucial for the development of explainable AI, but it also presents significant challenges. The core challenge is that people usually cannot obtain the ground-truth value of the pixel-wise attribution of an input. Thus, we design a metric to evaluate the objectiveness of the attribution map without ground truth. Our metric is broadly used to evaluate eight benchmark methods of attribution explanations, which provides new insights into attribution methods. *We will release the code when the paper is accepted.*

1 Introduction

Nowadays, many methods have been proposed to explain feature representations of a deep neural network (DNN) in a post-hoc manner. In this research, we limit our attention to existing methods of estimating the *importance/attribution/saliency* of input variables or neural activations in an intermediate layer to the network output [37, 30, 33, 8, 10], which present the mainstream in the field of explainable AI. To avoid ambiguity, importance/saliency/attribution maps are all termed *attribution maps* in this paper.

In this paper, we aim to evaluate the objectiveness of eight existing attribution methods, instead of proposing a new attribution method to explain the DNN. Previous studies of estimating attribution maps mainly explained pixel-wise importance encoded by the DNN. In comparison, our research evaluates whether their explanations objectively reflect the true attribution in the DNN, instead of simply generating seemingly reasonable attribution maps from the perspective of human users. Specifically, we aim to fairly evaluate the objectiveness of attribution maps. *E.g.* if the attribution value of a pixel is twice that of another pixel, then the first pixel is supposed to contribute exactly twice numerical values to the prediction *w.r.t.* the second pixel. In this case, the attribution value can be considered objective.

The motivation of evaluating the objectiveness of attribution methods is that features encoded by the DNN are usually far away from the logic of human cognition. In fact, people cannot even obtain the true logic used for inference by the DNN to evaluate whether the explanation objectively reflects true attributions of input variables. This problem can be analyzed from two perspectives.

First, there is a large gap between visually reasonable attributions and ground-truth explanations. As Figure 1 shows, although most attribution results seem reasonable to some extent, explanations for the same input image are quite different. To this end, several previous studies [14, 44] evaluated whether the attribution map looks reasonable to human users in a qualitative manner. However, there is no theory to prove the intuitively correct attribution map really reflects the truth in the DNN. Thus, in this paper, we try to develop a metric to evaluate the objectiveness of attribution methods.

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

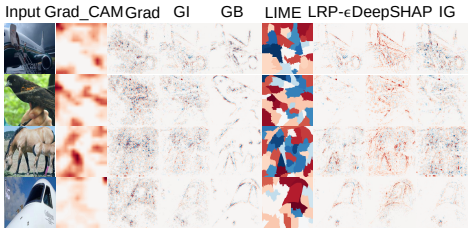

Figure 1: Examples of attribution maps generated by different methods. These attribution methods provide quite different attribution maps for the same image. Furthermore, in various experiments, we show that explanations should not be evaluated using ground truth labeled according to human cognition. Supplementary materials provide more attribution maps.

Table 1: Review of attribution methods

Method	What to explain	Limitations
Grad-CAM [35]	Attribution distribution at an intermediate layer	Usually explaining features at high layers
Grad [38]	Pixel-wise attribution	–
GI [37]	Pixel-wise attribution	–
IG [40]	Pixel-wise attribution	–
GB [39]	Pixel-wise attribution	Requirement to use ReLU as non-linear layers
Shapley Value [36]	Pixel-wise attribution	NP-hard problem
DeepSHAP [30]	Pixel-wise attribution	Only be applied to certain architectures, which are designed with specific backward rules.
LIME [33]	Super-pixel level attribution	Analysis at the super-pixel level, rather than the pixel level
LRP-ϵ [8]	Pixel-wise attribution	Only be applied to certain architectures, whose relevance propagation rules of all layers should be defined.

Second, we cannot construct a dataset for a certain task with intuitive ground-truth attributions according to human cognition. It is because the DNN usually does not only learn the knowledge that it is supposed to learn. Section 4 demonstrates that previous studies, which tried to construct specific datasets with ground-truth attributions, could not ensure the trustworthiness of the evaluation.

Existing evaluation metrics for the objectiveness of explanations also have certain flaws, which we will discuss in detail in the second paragraph of Section 2.

How to evaluate the objectiveness of the explanation? As mentioned above, people cannot obtain ground-truth explanations, which objectively reflect the true logic of the DNN instead of fitting human cognition. Therefore, defining theoretical rules to examine the objectiveness of explanations is still an open problem. To this end, some studies defined [36, 15] game-theoretic axioms for attributions to identify better attribution methods. For example, Dr. Shapley [36] proposed *linearity, dummy, symmetry, and efficiency axioms* for objective attributions. Besides, Deng et al. [15] proved the *interaction distribution axiom*, which uses interactions between input variables to explain the rationale of Shapley values. Although we admit that the above five axioms are not the only requirements for objective attributions, we believe these five axioms are good enough to evaluate the objectiveness of attribution methods. Please see Section 3 for more discussions.

Designing the evaluation metric. Inspired by the above analysis, we aim to design a metric to evaluate the objectiveness of attribution explanations, according to all the above five axioms. We will also discuss how to use these axioms to ensure the solid foundation of this metric. To this end, the aforementioned five axioms seem to be a good evaluation standard, but the evaluation based on attributions (*i.e.,* the Shapley value) satisfying the five axioms has two dramatic problems. First, the computational cost of accurately evaluating the objectiveness of attributions is NP-hard, which is unaffordable. Second, a seemingly plausible solution is to use approximated Shapley value [12, 30, 5] for the evaluation. However, the approximation error is usually even larger than the gap between other explanations and accurate Shapley values, which makes this evaluation metric unreliable.

To address these two problems, we design a new metric to evaluate the objectiveness of explanation methods, which evaluates the bias of attribution maps, instead of directly evaluating the accuracy of pixel-wise attributions. This metric simultaneously solves two challenges, *i.e.,* reducing the computational cost, and ensuring high accuracy. More crucially, this metric can be derived from the aforementioned axioms (see supplementary materials). **Besides, we find that this evaluation has no partiality to Shapley-value-based explanations** (*e.g.* DeepSHAP [30]) due to the gap between the Shapley-value-based explanation and the accurate Shapley value (see Figure 8).

In this paper, we used our metric to evaluate eight widely-used attribution methods in Table 1. We conducted experiments to evaluate these methods on image datasets [26, 17] and tabular datasets [42].

Evaluation on pixels or super-pixels? How to determine the basic unit (*e.g.* the elementary concept) of the explanation is another open problem without a trustworthy solution. Therefore, to simplify the story, this paper only focuses on the objectiveness of the pixel-wise attribution.

Contributions of our paper can be concluded as follows. (1) We propose a standard metric based on game theory to evaluate the objectiveness of attribution methods without knowing ground-truth explanations. (2) The metric of the pixel-wise bias of the attribution map can be estimated efficiently and accurately, which avoids falling into the computational bottleneck of using accurate pixel-wise attributions for evaluation. (3) Because our metric does not need any annotations of ground-truth explanations, our metric can be broadly applied to different DNNs trained on different datasets.

2 Related Work

For the evaluation of attribution/importance/saliency explanation methods, the qualitative analysis of explanation results via subject judgement [14, 44, 31] is a classical perspective for the evaluation. However, we would like to limit our discussions to the quantitative evaluation metric for explanation methods.

Yang and Kim [45] and Oramas et al. [32] built a specific dataset to generate intuitive ground-truth explanations for evaluation. However, in Section 4, we find that we cannot assume the DNN not to use noises for classification, which hurts the trustworthiness of the evaluation. Another kind of classical evaluation metrics [34, 23, 4, 18, 25, 43, 16, 22, 6, 31] was sequentially removing pixels from pixels with the lowest attributions to those with the highest attributions (or from the highest attributions to the lowest attributions). They used the decreasing speed of the DNN's performance to evaluate the quality of explanation results. If people masked pixels with the lowest attribution values first, then a slow decrease of the network output score indicated a high quality of the attribution map. However, previous studies [36] have pointed out that the method of sequentially removing pixels could not objectively reflect the true importance of pixels. Section 4 has discussed this issue both theoretically and experimentally.

Besides, some studies evaluated explanation results from other perspectives. Arras et al. [7] and Vu et al. [41] evaluated attribution maps from the perspective of adversarial attacks by adding random noise to the input. Some studies [2, 19, 3] evaluated the robustness of explanation methods *w.r.t.* the perturbation. Adebayo et al. [2] randomized layers of a DNN from the top to the bottom, and visualized the change of attribution maps. Bhatt et al. [9] proposed three desirable properties of the explanation methods, including low sensitivity, high faithfulness, and low complexity to evaluate explanation methods. Warnecke et al. [43] used multiple perspectives to evaluate explanations. Yeh et al. [46] evaluated explanation methods using (in)fidelity and sensitivity. Supplementary materials summarize various evaluation perspectives of previous studies.

Unlike previous studies, in this paper, we aim to evaluate the objectiveness of attribution estimation without annotations of ground-truth attributions. We believe that the objectiveness is the most crucial issue for attribution methods. Please see supplementary materials for more discussions.

3 Evaluation metrics of explanations

Axioms for objectiveness. Letting the attribution method objectively explain a DNN is a key issue for attribution methods, but it is still an open problem. In recent years, many studies have attempted to explore the trustworthiness of an explanation result. A classical way is to define various axioms to constrain attribution maps. Dr. Shapley [36] proposed *linearity, dummy, symmetry, and efficiency axioms*, and considered that trustworthy attributions should satisfy such axioms (see supplementary materials for details). Furthermore, Deng et al. [15] further proposed the *interaction distribution axiom* as shown in Theorem 2, and used interactions between input variables to explain the rationale of different attribution methods.

The composition of the five axioms has been widely considered as a typical requirement for the objective explanation of DNNs [36, 5, 30], although we admit that the five axioms are not the only requirements for the evaluation of the objectiveness of explanations, and we welcome further axioms. To this end, Shapley [36] has proved that the Shapley value was the unique solution that satisfied *linearity, dummy, symmetry, and efficiency axioms* (see Theorem 1). Deng et al. [15] proved that the Shapley value also satisfies the interaction distribution axiom (see Theorem 2).

However, we cannot directly use the five axioms for evaluation due to the impractical computational cost. For example, the Shapley value is hard to be accurately computed, which will be discussed in

detail later. *Thus, we propose a relatively accurate metric satisfying above axioms without a high computational cost to evaluate the objectiveness of attributions.*

Details about axioms and the Shapley value. The inference of a DNN can be regarded as a game with n players $\Omega = \{1, 2, \cdots, n\}$. Each player i is an input variable (*e.g.* a pixel/local patch/super-pixel). The DNN does not use each input variable (player) individually. Instead, a set of input variables (players) $T \subseteq \Omega$ may cooperate with each other and form a certain inference pattern (a coalition) to affect the network output (*i.e.,* pursuing a high reward). In this case, the DNN is taken as the game F, $F : 2^{\Omega} \to \mathbb{R}$. $F(T)$ represents the output of the DNN when only input variables in T are present, and all other variables in $\Omega \setminus T$ are masked. Specifically, $F(T)$ is computed on the sample, in which variables in $\Omega \setminus T$ are replaced with their baseline values. The baseline value is computed as the average pixel value over all images [5]. $F(\Omega) - F(\emptyset)$ denotes the total reward of all players. An attribution method tries to allocate the network output (the total reward) to each input variable $i \in \Omega$ (each player) as its attribution $A_i \in \mathbb{R}$. The Shapley value of the player i can be computed as follows.

$$A_i^* = \sum\nolimits_{T \subseteq \Omega \setminus \{i\}} \frac{|T|!(|\Omega|-|T|-1)!}{|T|!} \left[F(T \cup \{i\}) - F(T)\right] = \mathbb{E}_r \left[\mathbb{E}_{|T|=r, T \not\ni i} \left[F(T \cup \{i\}) - F(T)\right]\right], \quad (1)$$

where $|\Omega|$ denotes the total number of input variables (players) in an input sample, and r denotes the number of input variables (players) in a sampled subset T.

Theorem 1 *Shapley [36] proved that the Shapley value was the unique unbiased metric which satisfies the linearity axiom, the dummy axiom, the symmetry axiom, and the efficiency axiom.*

Theorem 2 *The Shapley value satisfies the interaction distribution axiom, i.e., $A_i = \sum_{T \subseteq \Omega: i \in T} D(T)/|T|$, where $D(T)$ represents the Harsanyi dividend [21] of the set of players T. The Harsanyi dividend $D(T)$ measures the numerical utility created by the interaction patterns among exactly all players in T. In this way, the Shapley value uniformly allocates each interaction pattern $D(T)$ to all players in this pattern. In comparison, many other attribution methods, such as the LRP-ϵ [8], the GI [37], the Grad-CAM [35], do not satisfy the interaction distributive axiom [15].*

Bias of the attribution map. The Shapley value is not a practical metric for evaluation because of its NP-hard computational cost. Although previous studies [30, 13, 12, 5] proposed various methods to approximate the Shapley value with a relatively low computational cost, Aas et al. [1] showed that accurate Shapley values were dramatically different from the approximated Shapley value. In particular, Figure 6 shows that the gap between the approximated Shapley value and the accurate Shapley value is sometimes even larger than the gap between other explanations and the accurate Shapley value.

Therefore, instead of directly computing Shapley values for evaluation, we propose a new metric to evaluate the objectiveness of the attribution map. We use the bias of the attribution map, which reflects the difference between the explanation result and the unbiased estimation of the attribution map, as the metric to evaluate the objectiveness of the attribution method. Because the attribution bias is derived on the basis of the Shapley value, it potentially reflects the aforementioned five axioms. In addition, unlike Shapley values, the proposed metric is supposed to be efficiently computed with much higher accuracy than the approximated Shapley value.

More crucially, Figure 6 shows that the Shapley-value-based attribution methods are not favored by our evaluation metric. This verifies the fairness of this metric.

To compute the attribution bias, we are given an input sample $I \in \mathbf{I}$. Let us consider the DNN F with a single scalar output $y = F(I)$. For DNNs with multiple outputs, existing methods [37, 30, 33, 8, 10] usually explain each individual output dimension independently. Although our study is not limited to visual data, let us use image classification as an example to introduce the algorithm. Let $\{a_i\}$ denote the pixel-wise attribution map estimated by a specific attribution method. We aim to evaluate the bias of $\{a_i\}$. People in game theory usually hope the network output equals the sum of pixel-wise attribution values [36, 5, 30] in order to let the attributions act like causal factors for the model output, to some extent. However, not all explanation methods satisfy this property. To ensure fair comparisons, we use λ to normalize attributions yielded by an attribution method, so as to make attributions fit this property without loss of generality. Nevertheless, λ can be **eliminated** in Equation (3) during the evaluation process, without affecting the evaluation result.

$$y = b + \sum\nolimits_{i \in \Omega} A_i, \qquad \text{s.t.} \quad A_i = \lambda a_i \quad (2)$$

b denotes an anchor value for y; i denotes the index of each pixel in the input image; Ω denotes the set of all pixels in the image. Since many attribution methods [35, 38] mainly compute relative values of attributions $\{a_i\}$, instead of a strict attribution map $\{A_i\}$, we use λ to bridge $\{A_i\}$ and $\{a_i\}$, which is independent with the index i. The aforementioned Shapley values $\{A_i^*\}$ can be considered as the ground truth of $\{A_i\}$ [30]. The estimated attribution of each pixel can be assumed to follow a Gaussian distribution $A_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ [12]. Attribution distributions of different pixels can be further assumed to share a unified variance, *i.e.,* $\sigma_1^2 \approx \sigma_2^2 \approx ... \approx \sigma_n^2$.

Evaluating the attribution distribution $\{a_i\}$ has two steps. *First, we sample pixels whose attributions are more likely to have large deviations.* To this end, we sample the set of pixels S with top-ranked high (or low) attributions. These pixels are more likely to be significantly biased towards high (or low) attribution values. Meanwhile, from another perspective, considering the Gaussian distribution of $\{A_i\}$, the distribution of the sampled attribution values is close to the Gumbel distribution [20].

Second, we evaluate the bias of the sampled attributions. Although the Shapley value can be considered as a standard formulation of the pixel-wise attribution, it usually cannot be accurately computed because of its high computational cost. In order to evaluate the sampled attribution values without significantly increasing the computational cost, we applied the Shapley value approximated by the sampling method [12], denoted by A_i^{shap}. The approximation equation is given by the second equation of Equation (1). Just like the target attribution distribution A_i, A_i^{shap} is assumed to follow a Gaussian distribution $\mathcal{N}(A_i^*, (\sigma^{\text{shap}})^2)$. A_i^{shap} is an unbiased approximation of the true Shapley value A_i^*, Thus, the average value over different pixels in S satisfies $\frac{\sum_{i\in S} A_i^{\text{shap}}}{|S|} \sim \mathcal{N}(\frac{\sum_{i\in S} A_i^*}{|S|}, \frac{(\sigma^{\text{shap}})^2}{|S|})$.

We use $\beta = \frac{\sum_{i\in S} A_i^{\text{shap}}}{|S|\|A^{\text{shap}}\|}$ as the anchor value of our metric. The difference between the average value of highest (or lowest) attribution values in S and the anchor value, $\left|\frac{\sum_{i\in S} A_i}{|S|\|A\|} - \beta\right|$, can reflect the bias of the attribution map, as follows.

$$M_{\text{pixel}} = \mathbb{E}_I\left[\left|\frac{\sum_{i\in S} A_i}{|S|\|A\|} - \frac{\sum_{i\in S} A_i^{\text{shap}}}{|S|\|A^{\text{shap}}\|}\right|\right] = \mathbb{E}_I\left[\frac{1}{|S|}\left|\frac{\sum_{i\in S} a_i}{\|a\|} - \frac{\sum_{i\in S} A_i^{\text{shap}}}{\|A^{\text{shap}}\|}\right|\right] \quad \text{s.t. } \forall i\in S, j\in\Omega\backslash S, a_i \geq a_j \text{ or } \forall i\in S, j\in\Omega\backslash S, a_i \leq a_j, \tag{3}$$

where Ω is the set of all pixels in an image. $\|A^{\text{shap}}\|$ and $\|a\|$ are used for normalization. A small value of M_{pixel} indicates the low bias of the attribution map.

Can the above metric be used to estimate attributions or have partiality to Shapley-value-based methods? The proposed metric is designed to evaluate attribution methods, rather than approximate pixel-wise attributions. Please see supplementary materials for discussions on why not use the proposed metric for explanations. Besides, although the metric has the same theoretical foundation as the Shapley value, our metric has no partiality to the DeepSHAP. For example, experimental results showed that LRP-ϵ outperformed DeepSHAP.

In addition, the proposed metric can also be used to evaluate the attribution of neural activations in the intermediate layer, such as those generated by Grad-CAM. In this case, we regard the target intermediate-layer feature as the input to compute A_i^{shap}, so as to implement the evaluation.

Compared with the estimated Shapley value, the proposed metric can be more accurately measured with much less sampling of testing samples (see Exp.1 in Section 5). Computing the accurate Shapley value is an NP-hard problem with the computational cost $\mathcal{O}(N \cdot 2^N)$, according to Equation (1). In comparison, the computational cost of the evaluation based on our metric is $\mathcal{O}(mN)$, where N denotes the number of pixels in the image, and m is the sampling number of each subset T in Equation (1). Besides, Figure 4 shows that the estimated metric is accurate enough for evaluation.

In comparison, the computational cost of approximating the Shapley value A_i^{shap} for evaluation is $\mathcal{O}(m^{\text{shap}}N)$, where m^{shap} is the number of sampling to approximate the Shapley value. Theoretically, we must set $m^{\text{shap}} = |S|m$, if we want the evaluation based on $|A_i - A_i^{\text{shap}}|$ to achieve the same accuracy as the proposed metric, *i.e.,* its computational cost is $|S|$ times as our metric. **Thus, our metric is much less biased than the approximated Shapley value calculated with the same computational cost.** This conclusion is verified in experiments in Figure 4 and is theoretically proved in supplementary materials.

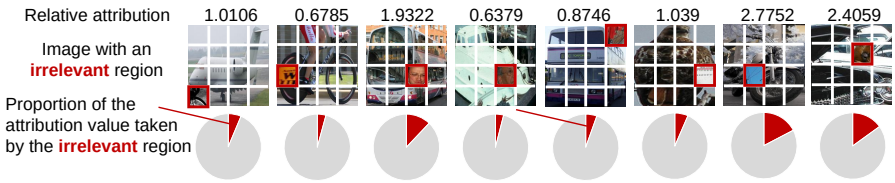

Figure 2: Disproof of the assumption that regions irrelevant to the classification had no contribution to the classification. Each image was divided into 4×4 regions. The red box indicated the region irrelevant to the target class, which was pasted into the image. The Shapley value of the irrelevant region was not zero, which demonstrated that it was untrustworthy to create the ground-truth explanation using irrelevant objects. Here, Shapley values of irrelevant regions were accurately computed via brute-force enumeration without approximation. The relative attribution of the irrelevant region was reported on the top of the image. We found that DNNs usually model features of irrelevant regions. Thus, the assumption was not convincing enough for evaluation.

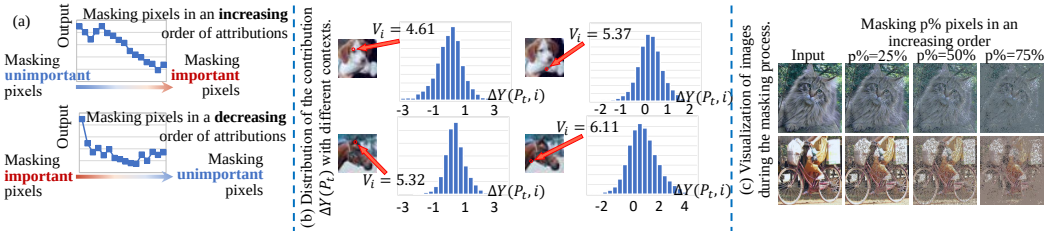

Figure 3: (a) Schematics of the evaluation based on ranked attributions. (b) Evaluating the trustworthiness of the evaluation method based on the ranked attributions. We randomly select the pixel i in the input image, and compute $\Delta Y(P_t, i)$ and $V(P_t)$ over different contexts P_t. We find that the contribution of the pixel to the output is unstable *w.r.t.* different contexts. Thus, this evaluation method is not convincing. (c) The masked pixels distribute over the whole image when we evaluate a specific attribution map. Therefore, the evaluation based on the removal of ranked pixels is conducted under specific out-of-distribution contexts, which cannot reflect the true importance of the pixel. In comparison, our evaluation metric is computed considering all contexts.

4 Problems of previous evaluation metrics

Can we create ground-truth explanations according to human cognition for evaluation? Several previous studies [45, 11] tried to obtain intuitive ground-truth explanations by creating synthetic datasets. These intuitive ground-truth explanations were used to evaluate the attribution map. Specifically, they pasted an irrelevant object *w.r.t.* the task into the image. Pixels of the irrelevant object were assumed to be assigned with zero attribution to the prediction.

However, we found that this assumption is not convincing. A common sense is that the DNN does not make inferences in the same way as people [24], so we conducted experiments to examine the above assumption. We followed the same settings to build up a synthetic dataset by modifying images in the Pascal VOC 2012 dataset [17]. We constructed images containing both relevant regions and irrelevant regions. We divided each image into 4×4 regions. Then, we randomly replaced a region with an image patch of 56×56 pixels cut from an image in a different class. In this way, the replaced region was regarded irrelevant to the ground-truth class, and other regions were regarded as relevant to the ground-truth class. Then, we trained an AlexNet [27] on the dataset, and computed the accurate Shapley value of the irrelevant region in the image. In this case, we only needed to compute sixteen Shapley values A^*_i, $i = 1, 2, \cdots, 16$, for the sixteen image regions. According to Equation (1), the cost of computing accurate Shapley values for as few as sixteen regions was still affordable, unlike computing pixel-wise Shapley values. Accurate Shapley values for all image regions could be computed within ten minutes without any approximation.

As Figure 2 shows, the Shapley value of the irrelevant region was **not** zero. We also showed the relative attribution of the irrelevant region $\hat{k}$, $\frac{|A^*_{\hat{k}}|}{\mathbb{E}_{i\in\Omega}[|A^*_i|]}$, on the top of images, where Ω denotes the set of all regions in the image. The average relative attribution over all images was 1.0336. Therefore, it was inappropriate to assume that the irrelevant region has no attribution to the classification.

Trustworthiness of using the ranked attributions for evaluation. Several previous studies [34, 4, 23, 25, 43] evaluated the ranking order of pixel-wise attribution values. As Figure 3 (a) shows,

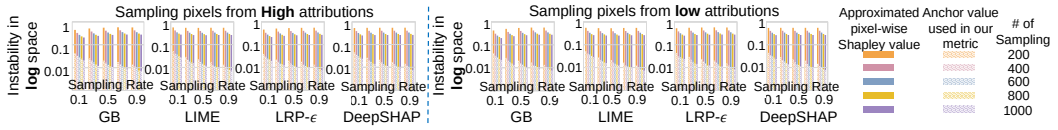

Figure 4: Instability of the approximated pixel-wise Shapley value and the instability of the anchor value $\beta = (\sum_{i \in S} A_i^{\text{shap}})/(|S| \|A^{\text{shap}}\|)$ of our bias metric. S denotes the set of pixels sampled according to attribution maps from different explanation methods. Results show that the anchor value of our metric exhibited a much lower instability than the approximated pixel-wise Shapley value.

for each time step t, they masked the pixel with the t-th lowest (or highest) attribution value in a sequential manner. In this way, if the attribution value was estimated more accurately, then the performance of the DNN was supposed to decrease more slowly (or more quickly).

However, many studies [33, 13] found that the numerical contribution of a pixel to the performance significantly depended on the context of other pixels. In other words, the influence of masking a pixel on the DNN's output significantly depended on the order when the pixel was masked, *i.e.,* if the pixel was masked earlier or later, then the pixel would have a different context. Each specific context usually made the pixel have a different influence on the DNN's output.

In this way, we designed a new experiment to verify that a pixel's marginal attribution strongly depended on the order (or more precisely the context) when it was masked. If so, the evaluation based on masking top-ranked pixels would not faithfully reflect true attributions of pixels. We trained a LeNet [29] on the CIFAR-10 dataset [26]. Given a pixel i in the input image, we computed its marginal attribution when this pixel was the $(t+1)$-th masked pixel, as $\Delta Y(P_t, i) = Y_{\text{masking } P_t} - Y_{\text{masking } P_t \& i}$. P_t represents a set of t pixels that had been masked before the t-th step. $Y_{\text{masking } P_t}$ denotes the network output when we fed the image with pixels in P_t being masked into the DNN. Similarly, $Y_{\text{masking } P_t \& i}$ corresponds to the network output when we further masked an additional pixel i. Then, $V_i = \mathbb{E}_t \mathbb{E}_{P_t : \|P_t\| = t}[\max(|\frac{\Delta Y(P_t, i)}{\Delta \bar{Y}(i)}|, |\frac{\Delta \bar{Y}(i)}{\Delta Y(P_t, i)}|)]$ measures the instability of $\Delta Y(P_t, i)$, when people masked the pixel i at different orders with different contexts, where $\Delta \bar{Y}(i) = \mathbb{E}_{t'} \mathbb{E}_{P'_{t'} : \|P'_{t'}\| = t'}[\Delta Y(P'_{t'}, i)]$. Figure 3 (b) reports the distribution of $\Delta Y(P_t, i)$ over different orders and V_i values for specific pixels. A large value of V_i indicates that the marginal attribution $\Delta Y(P_t, i)$ was unstable. We found that the marginal attribution of a pixel was significantly affected by the order of masking. In particular, for some pixels, the marginal attribution $\Delta Y(P_t, i)$ was sometimes more than five times of the average marginal attribution of the same pixel over different orders. Moreover, as Figure 3 (c) shows, the masked pixels distributed over the whole image, and formed an irregular context without an explainable meaning. The evaluation under irregular unexplainable contexts may not reflect the true importance of the pixel. Thus, it still has technical flaws to evaluate attribution methods based on the ranking order of attributions.

5 Experiments

To evaluate attribution methods, we conducted experiments on both visual data and tabular data for evaluation, which included the Pascal VOC 2012 [17] dataset, the CIFAR-10 [26] dataset, and the TV news channel commercial detection dataset (a tabular dataset) [42]. In experiments, we evaluated the following explanation methods, including the Gradient (Grad) [38], Gradient$\times$Input (GI) [37], integrated gradient (IG) [40], guided back-propagation (GB) [39], layer-wise relevance propagation (LRP) [8], DeepSHAP [37], LIME [33], and grad-CAM [35]. Figure 1 shows attribution maps yielded by these methods. Supplementary materials revisit these attribution methods.

Implementation Details: To approximate the Shapley value for each pixel, we sampled the set T in Equation (1) for 1000 times for each pixel of images in the CIFAR-10 dataset, and sampled T for 100 times for each pixel of images in the Pascal VOC 2012 dataset. We sampled the top-10%, 30%, 50%, 70%, 90% pixels with the highest/lowest values. In this way, the proposed bias metric can consider all pixels for evaluation. Besides, LRP-ϵ was not used on residual networks, because the relevance propagation rules of some structures in ResNet were not defined, to the best of our knowledge.

Exp. 1: Stability of the proposed bias metric. People usually used a sampling-based method [12] to approximate the Shapley value, whose computation is originally NP-hard. This experiment compared the convergence speed between the approximated Shapley value and the convergence speed

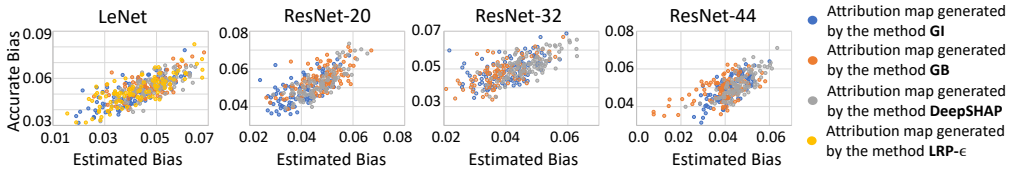

Figure 5: The positive relationship between the bias metric and the accurate pixel-wise bias (with an NP-hard computational cost). Each point corresponded to an attribution map generated by an explanation method. The bias metric was positively correlated with the accurate pixel-wise bias, which verified the objectiveness of the proposed bias metric.

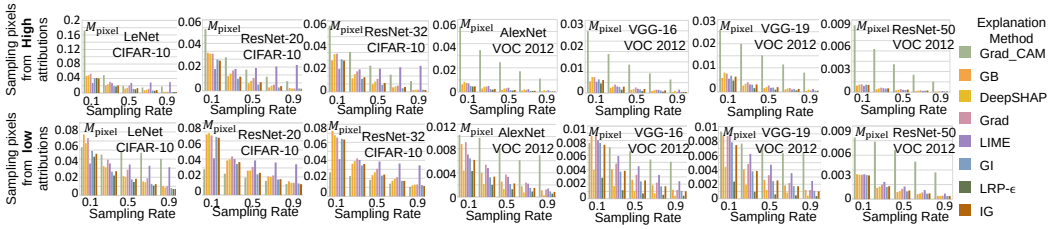

Figure 6: Bias of the attribution map at the pixel level. LRP-ϵ provided attribution maps with the least bias on LeNet, AlexNet, and VGG-16/19. GI and GB outperformed other attribution methods on ResNets. Due to the limitation of the page number, we provided results on the ResNet-44/56/50 and explicit result numbers in supplementary materials.

of the anchor value of the bias metric *w.r.t.* an increasing sampling times. Given a trained DNN and an image $I \in \mathbf{I}$, we computed the approximated Shapley value multiple times. We normalized the approximated Shapley value as $\alpha = A^{\text{shap}}/\|A^{\text{shap}}\|$, just like the normalization in Equation (3). We computed $\beta = (\sum_{i\in S} A_i^{\text{shap}})/(|S|\|A^{\text{shap}}\|)$ multiple times, which was the anchor value for the proposed bias metric in Equation (3). The instability of the approximated Shapley value was computed as $\text{Instability}_{\text{shap}} = \mathbb{E}_{I\in\mathbf{I}}\left[\frac{\mathbb{E}_{u,v:u\neq v}|\alpha_{(u)}-\alpha_{(v)}|}{\mathbb{E}_w[|\alpha_{(w)}|]}\right]$, and the instability of the anchor value of the bias metric was computed as $\text{Instability}_{\text{ours}} = \mathbb{E}_{I\in\mathbf{I}}\left[\frac{\mathbb{E}_{u,v:u\neq v}|\beta_{(u)}-\beta_{(v)}|}{\mathbb{E}_w[|\beta_{(w)}|]}\right]$. $\alpha_{(u)}$ and $\beta_{(u)}$ denoted the u-th computation result of the approximated Shapley value and the u-th result of the anchor value of the bias metric, respectively.

In this experiment, we used the LeNet trained on the CIFAR-10 dataset. We explored the change of the instability along with the increase of the sampling number and the increase of the pixel number. As Figure 4 shows, the instability of our anchor value is much lower than the instability of the approximated pixel-wised Shapley value. This result demonstrated that the anchor value of the bias metric converged much faster, and thereby was more reliable than the approximated pixel-wise Shapley value.

Exp. 2: Trustworthiness of the proposed bias metric. It was a challenge to evaluate the trustworthiness of the proposed bias metric. We needed to compare the proposed bias metric with the accurate pixel-wise bias, whose computation needed accurately computed Shapley values. As aforementioned, accurately computing the Shapley value was NP-hard and unaffordable. Thus, we only computed relatively accurate Shapley values just for a few pixels using the method in [12] with a huge number of sampling. In this case, we could consider that the estimated Shapley values were accurate enough to verify the trustworthiness of the proposed bias metric. In this experiment, we used LeNet and ResNet-20/32/44 trained using the CIFAR-10 [26] dataset.

The NP-hard computational cost of calculating the accurate pixel-wise bias significantly limited the applicability of the accurate bias. To this end, we only sampled 10% pixels with the highest attributions to evaluate the proposed bias metric. Figure 5 compares the proposed bias metric with the accurate pixel-wise bias. Each point corresponded to a specific evaluation result. We found that the bias metric was roughly positively correlated with the accurate bias among sampled pixels, which verified the objectiveness of the bias metric.

Exp. 3: Evaluating the bias of the attribution maps. Figure 6 and Figure 7 show bias values of attribution maps, which were generated by different attribution methods on images and tabular data, respectively. GI and GB provided the least biased attribution maps for ResNets at the pixel level. For AlexNet, VGG-16/19, and LeNet, LRP-ϵ outperformed other methods. Besides, we found that

Figure 7: Bias of attribution maps estimated on the TV news channel commercial detection dataset [42]. IG and DeepSHAP outperformed other attribution methods.

Figure 8: Difference between attribution maps generated by DeepSHAP and those of accurately-computed Shapley values.

the performance of LIME was volatile, *i.e.,* in some cases, LIME performed quite well, but in some other cases, LIME performed the worst. This phenomenon was obvious in the CIFAR-10 dataset. It was because LIME calculated attribution maps for super-pixels. The number of super-pixels in an image from the CIFAR-10 dataset was limited. In this way, many pixels within a single super-pixel shared the same attribution value in results of LIME, which made it hard to sample these pixels with significantly biased attribution values. According to Figure 7, we found that IG and DeepSHAP usually achieved the least biased attribution maps.

In particular, we also found that our metric had no partiality to Shapley-value-based methods, such as DeepSHAP. Figure 8 shows the clear difference between the DeepSHAP's attribution map and relatively accurate Shapley values. We computed such relatively accurate Shapley values by sampling 20,000 testing samples from a single input image, to ensure its accuracy. It shows that the DeepSHAP's attribution map was clearly different from relatively accurate Shapley values. Furthermore, Aas et al. [1] also showed that accurate Shapley values computed with an NP-hard cost were usually dramatically different from Shapley values approximated by practical methods.

Discussions on fitness with human intuition. Unlike previous studies pursuing an explanation towards human intuition, our study evaluated whether these attribution methods objectively reflect true attributions of input variables. Nevertheless, objective attributions selected by our evaluation metrics usually also fit human intuition, which has been discussed in supplementary materials.

Seemingly contradictory with previous metrics, but actually not. We noticed that our evaluation results seemed to conflict with [2] and ROAR [23], but actually not. Adebayo et al. [2] criticized the GB for being insensitive to the randomization of DNNs. The evaluation metric ROAR [23] showed that the removal of important pixels based on GB did not affect the performance of the DNN significantly. Theoretically, the evaluation in [2] naturally welcomed smoothed attributions (*e.g.* Grad-CAM), and criticized edge-like attributions (*e.g.* GB). However, this phenomenon also showed that edge-like pixels had large impacts on network features, which actually reflected the true phenomenon of signal processing in DNNs. Besides, ROAR might have a partiality to Grad-CAM. It was because removing pixels from a smooth surface/region was more likely to bring in additional noisy features than removing pixels from edges. To this end, GB usually assigned edges with large attributions, which caused the performance of the DNN was not significantly affected. In comparison, the proposed bias metric in this paper could measure the objectiveness of explanation methods more accurately than [2, 23]. Please see supplementary materials for more discussions.

6 Conclusion and Discussions

In this paper, we propose a metric, *i.e.,* the bias of attribution maps at the pixel level, to evaluate the objectiveness of attribution methods. The proposed evaluation metric is computed without a need for ground-truth explanations. Our metric can be applied to widely-used explanation methods *w.r.t.* different DNNs learned using different datasets. We also verified the stability and trustworthiness of our metric in various experiments. Experimental results showed that our metric could effectively evaluate the bias of attribution maps, and attribution maps from LRP-ϵ, GI, and GB exhibited lower bias than other attribution methods. Kumar et al. [28] pointed out two technical flaws of the Shapley value. First, the Shapley value could not totally avoid the OOD problem with the setting of baseline values. Second, the selection of players (*i.e.,* the partition of input variables) also affected the Shapley value. They are both open problems for Shapley values. However, the baseline value and the partition of variables are two problems orthogonal to the evaluation of the attributions in the scenario where the partition of input variables and their baseline values have been given. These problems do not hurt the soundness of our study. Please see supplementary materials for more discussions.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv:1903.10464*, 2019.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *In NIPS*, 2018.

[3] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. *In arXiv:1806.07538*, 2018.

[4] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR*, 2018.

[5] Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley values approximation. *arXiv:1903.10992*, 2019.

[6] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "what is relevant in a text document?": An interpretable machine learning approach. *PLoS ONE*, 12:E0181142, 08 2017. doi: 10.1371/journal.pone.0181142.

[7] Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. *In arXiv:1904.11829*, 2019.

[8] Sebastian Bach, Alexander Binder, Gregoire Montavon, Frederick Klauschen, Klausrobert Muller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7), 2015.

[9] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. *IJCAI*, 2020.

[10] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. *In International Conference on Artificial Neural Networks (ICANN)*, 2016.

[11] Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. Can i trust the explainer? verifying post-hoc explanatory methods. *arXiv: 1910.02065*, 2019.

[12] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *In Computers & Operations Research*, 36(5):1726–1730, 2009.

[13] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018.

[14] Xiaocong Cui, Jung Min Lee, and J Hsieh. An integrative 3c evaluation framework for explainable artificial intelligence. *In The annual Americas Conference on Information Systems (AMCIS)*, 2019.

[15] Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, and Xia Hu. A general taylor framework for unifying and revisiting attribution methods. *arXiv preprint arXiv:2105.13841*, 2021.

[16] Jay DeYoung, Sarthak Jain, Nazneen Rajani, Eric P. Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *ArXiv*, abs/1911.03429, 2020.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *In International Journal of Computer Vision*, 88(2):303–338, June 2010.

[18] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *In arXiv:1704.03296v1*, 2017.

[19] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *In AAAI*, 2019.

[20] Emil Julius Gumbel. Les valeurs extrêmes des distributions statistiques. In *Annales de l'institut Henri Poincaré*, volume 5, pages 115–158, 1935.

[21] John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.

[22] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *NeurIPS*, 2021.

[23] Sara Hooker, Dumitru Erhan, Pieterjan Kindermans, and Been Kim. Evaluating feature importance estimates. In *NIPS*, 2019.

[24] Georgin Jacob, RT Pramod, Harish Katti, and SP Arun. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature communications*, 12 (1):1872, March 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22078-3.

[25] Pieterjan Kindermans, Kristof T Schutt, Maximilian Alber, Klausrobert Muller, Dumitru Erhan, Been Kim, and Sven Dahne. Learning how to explain neural networks: Patternnet and patternattribution. In *ICLR*, 2018.

[26] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *In Computer Science Department, University of Toronto, Tech. Rep*, 1, 2009.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *In NIPS*, 2012.

[28] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler. Problems with shapley-value-based explanations as feature importance measures. In *ICML 2020*, 2020.

[29] Yann LeCun, Lèon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *In Proceedings of the IEEE*, 1998.

[30] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *In NIPS*, 2017.

[31] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, 2018.

[32] Jose Oramas, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. *ICLR*, 2019.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. *In KDD*, 2016.

[34] Wojciech Samek, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klausrobert Muller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks*, 28(11):2660–2673, 2017.

[35] 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 ICCV*, 2017.

[36] Lloyd S Shapley. A value for n-person games. *In Contributions to the Theory of Games*, 2(28): 307–317, 1953.

[37] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *In arXiv:1605.01713*, 2016.

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

[39] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *In arXiv:1412.6806*, 2014.

[40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[41] Minh N Vu, Truc D Nguyen, NhatHai Phan, Ralucca Gera, and My T Thai. Evaluating explainers via perturbation. *In arXiv:1906.02032*, 2019.

[42] Apoorv Vyas, Raghvendra Kannao, Vineet Bhargava, and Prithwijit Guha. Commercial block detection in broadcast news videos. In *ICVGIP '14*, 2014.

[43] Alexander Warnecke, Daniel Arp, Christian Wressnegger, and Konrad Rieck. Evaluating explanation methods for deep learning in security. *IEEE European Symposium on Security and Privacy*, 2020.

[44] Fan Yang, Mengnan Du, and Xia Hu. Evaluating explanation without ground truth in interpretable machine learning. *In arxiv: 1907.06831*, 2019.

[45] Mengjiao Yang and Been Kim. Bim: Towards quantitative evaluation of interpretability methods with ground truth. *In arXiv:1907.09701*, 2019.

[46] Chihkuan Yeh, Chengyu Hsieh, Arun Sai Suggala, David I Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, pages 10965–10976, 2019.

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] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [No] No potential negative social impacts.
   (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? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary materials.
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] See Section 5.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (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]
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] See Section 5.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
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? [No]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]