$\mathcal{M}^4$: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models

Xuhong Li
Baidu Inc.
lixuhong@baidu.com

Mengnan Du
New Jersey Institute of Technology
mengnan.du@njit.edu

Jiamin Chen
Baidu Inc.
chenjiamin01@baidu.com

Yekun Chai
Baidu Inc.
chaiyekun@baidu.com

Himabindu Lakkaraju
Harvard University
hlakkaraju@seas.harvard.edu

Haoyi Xiong[*]
Baidu Inc.
xionghaoyi@baidu.com

Abstract

While Explainable Artificial Intelligence (XAI) techniques have been widely studied to explain predictions made by deep neural networks, the way to evaluate the faithfulness of explanation results remains challenging, due to the heterogeneity of explanations for various models and the lack of ground-truth explanations. This paper introduces an XAI benchmark named $\mathcal{M}^4$, which allows evaluating various input feature attribution methods using the same set of faithfulness metrics across multiple data modalities (images and texts) and network structures (ResNets, MobileNets, Transformers). A taxonomy for the metrics has been proposed as well. We first categorize commonly used XAI evaluation metrics into three groups based on the ground truth they require. We then implement classic and state-of-the-art feature attribution methods using InterpretDL and conduct extensive experiments to compare methods and gain insights. Extensive experiments have been conducted to provide holistic evaluations as benchmark baselines. Several interesting observations are made for designing attribution algorithms. The implementation of state-of-the-art explanation methods and evaluation metrics of $\mathcal{M}^4$ is publicly available at `https://github.com/PaddlePaddle/InterpretDL`.

1 Introduction

Although deep neural networks (DNNs) have achieved state-of-the-art performance on numerous AI tasks, they are often regarded as black boxes due to their lack of transparency. This opacity hinders the adoption of deep models in high-stake applications that require explainability, such as healthcare, criminal justice, and law. Explainable AI (XAI) aims to address this limitation by developing techniques to provide explanations for predictions made by DNN models [15, 17]. In recent years, researchers have proposed various XAI algorithms to enable deeper understanding of DNNs [40, 46, 48, 6, 7, 9]. Among these techniques, post-hoc feature attribution is one of the most widely used paradigms, which could provide insight into the behaviors of trained DNNs.

Although these feature attribution methods can be helpful in understanding deep models, their **faithfulness** (i.e., how well explanations match model reasoning) is not always guaranteed. Unfaithful explanations fail to provide a complete or accurate description of the algorithm that the model implements, thus might yield futile or deceptive insights [41]. To address this, researchers have

[*] Corresponding author

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

https://doi.org/10.52202/075280-0081

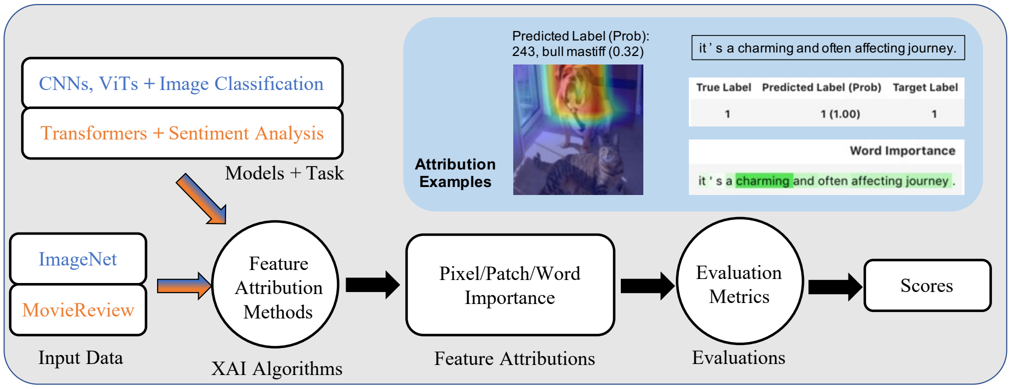


Figure 1: The benchmark pipeline $\mathcal{M}^4$, which supports evaluations on two modalities with more than ten deep models to holistically validate the faithfulness of existing feature attribution methods.

proposed methods to measure the faithfulness of explanations and filter out unfaithful algorithms. For example, Adebayo et al. [2] have proposed randomization tests on the model's parameters, on which explanation methods should depend. Such tests can easily filter out unfaithful explanation methods that rarely vary even when randomizing the model parameters. Furthermore, beyond simply binarizing the faithfulness of the explanation methods, recent work has built XAI benchmarks with evaluation metrics and datasets [14, 39, 3] to quantitatively compare the faithfulness among explanation methods. These evaluations are emerging as a guiding principle for selecting the most effective and appropriate explanation methods to elucidate models in specific tasks.

However, there are still several challenges towards this direction of research for faithfulness evaluations and benchmarks of XAI algorithms. First, assessing the faithfulness of feature attributions is difficult due to the lack of ground truths (of explanations). Although some attainable metrics have been proposed to evaluate feature attributions [42, 53], existing benchmarks rarely share or standardize these metrics. For instance, perturbation-based metrics are commonly used but differ in their perturbation methods, scales, and granularity across benchmarks. Alternative metrics also quantify faithfulness in different ways, but few studies have consolidated, classified, or analyzed the relationships between these metrics. Second, most benchmarks focus on a single data type and model, limiting their applicability and ability to validate new methods. Feature attribution techniques designed for specific models may not be applicable to others. For example, gradient-based explanations can be noisy for Vision Transformers [4]. To robustly evaluate explanation methods, XAI benchmarks should encompass assessments across multiple models and modalities.

To address these challenges, we propose developing benchmarks that cover various types of data and models, employ a taxonomy of standardized metrics, and analyze the relationships between different metrics. First, we categorize the commonly used faithfulness metrics into three groups, according to the status of the ground-truth explanations. Then, based on these metrics, we provide a unified XAI benchmark named $\mathcal{M}^4$, which supports the evaluations of two modalities (images and texts) with more than ten deep models to holistically validate the faithfulness of existing feature attribution methods (Figure 1). Through the proposed benchmark $\mathcal{M}^4$, we conducted comprehensive experiments and also obtained several interesting observations on the feature attribution methods that motivate future work in designing XAI methods and other applications related to XAI.

The contributions of this work are threefold:

- To the best of our knowledge, our work provides the first taxonomy of faithfulness evaluation metrics. We categorize the commonly used metrics into three sets that require no ground-truth of explanations, generate pseudo ground truth, and design synthetic ground truth, respectively.
- The proposed $\mathcal{M}^4$ benchmark allows the evaluation of diverse feature attribution methods with the same metrics across multiple data modalities (images and texts) and various architectures. We take advantage of the modular-designed implementations of XAI methods in InterpretDL [29] to build the $\mathcal{M}^4$ benchmark, making the evaluations easily extensible.
- We conduct extensive experiments to provide holistic comparisons with off-the-shelf baselines, yielding valuable observations that can inform future designs and applications of XAI methods.

2 Benchmark $\mathcal{M}^4$

In this section, we introduce $\mathcal{M}^4$, a unified benchmark for evaluating feature attribution methods.

2.1 Tasks, Datasets and Models

We consider two classic tasks: image classification from the computer vision domain and sentiment analysis from the NLP domain. These tasks have been used as testbeds for most explanation methods. To maximize the reuse of publicly available resources, we utilize two commonly used datasets for this benchmark, namely ImageNet [12] and MovieReview [55]. In fact, we do not need to train models on ImageNet because there are numerous pre-trained models publicly available. Furthermore, to enhance computational efficiency, we use a subset of 5,000 images from the ImageNet validation set, with 5 images per class for class balance. One training phase for images is required to quantify the Synthetic-based score (introduced in Section 2.3) because models need to be trained on a new synthetic dataset. For this training scenario, we take 10 random images per class from the ImageNet training set and randomly add synthetic patches to train the model. We then conduct the evaluations on the same 5,000 images (also with random synthetic patches) as the previous evaluations. For MovieReview, we fine-tune the pretrained language models on its training set and conduct the faithfulness evaluation on its validation set. Therefore, both the training and validation sets of MovieReview are required.

The reasons for choosing the ImageNet and MovieReview datasets, as well as the related tasks also include the availability of semantic segmentation labels [19, 31] and language reasoning labels [14] from public resources. Although these labels are not used in $\mathcal{M}^4$ as our purpose is for faithfulness evaluations, these labels can be used directly to measure human-labeled *interpretability* which is defined as the alignment between model explanations with human understanding.

Recent benchmarks rarely consider the choices of network structures when evaluating the faithfulness of explanation methods. In contrast, our proposed $\mathcal{M}^4$ considers a wide range of models to holistically evaluate faithfulness: VGG [45], three ResNets [21], Mobilenet-V3 [22], three ViT versions (small, base and large) [16] and MAE-ViT-base [20] for image classification, and two BERTs (base and large) [13], DistilBERT [43], ERNIE-2.0-base [47] and RoBERTa [34] for sentiment analysis.

2.2 Feature Attribution Methods

Our proposed benchmark $\mathcal{M}^4$ considers classic feature attribution methods that have been used in previous benchmarks, including model-agnostic explanations (LIME [40]), gradient-based (Integrated Gradient (IG) [48], SmoothGrad (SG) [46]), and model-specific (GradCAM [44]). In addition, our proposed benchmark $\mathcal{M}^4$ also evaluates the state-of-the-art ones especially for Transformer structures, e.g, Generic Attribution (GA) [6], Bidirectional Explanations (BT) [9], *etc.*, to comprehensively evaluate the feature attributions, as well as to revisit and confirm the progress of explanation methods.

Feature attribution methods are explainers that assign importance scores to input features of a machine learning model, elucidating the reasoning behind a model's specific prediction. Such explanations are readily comprehensible for human understanding. We note that there are more advanced forms of explanation, including prototype exemplars [8, 18, 37], concept vectors in the activation space [24, 54], and proxy models to simulate the rational process of deep models [27, 56], *etc*. Although these methods produce unique and probably more profound explanations, they may be either incomprehensible or laboriously evaluated. As an initial step toward establishing a benchmark for evaluating XAI methods, concentrating on feature attribution methods would be an achievable endeavor, albeit one that still presents its own set of challenges.

2.3 Metrics and Taxonomy

Due to the lack of explanation ground truths, there are no natural metrics to quantify the faithfulness of the explanations produced by feature attribution methods. Recent studies have proposed various metrics to evaluate explanation methods [42, 28, 53]. We review commonly used metrics and categorize them into three types based on whether they require explanation ground truths and how the ground truths are generated. Specifically, the three types of metrics are as follows.

No Ground Truth. Perturbation-based metrics offer a feasible approach that circumvents the need for ground-truth explanations. The underlying concept is based on the premise that important input

https://doi.org/10.52202/075280-0081

features will noticeably degrade the predicted probability of a model, while irrelevant features will have minimal impact on the probability mass. Several widely utilized metrics are built upon this fundamental idea. For example, Samek et al. [42] proposed *ordered perturbation*, which perturbs the input features gradually following the same order of values from the explanation results. They then calculate the area under the perturbation curve. The MoRF (most relevant first) metric represents the descending order, the LeRF (least relevant first) denotes the ascending order, and the ABPC (area between the perturbation curves) signifies their difference. Similarly deletion and insertion metrics [38] are an equivalent pair of MoRF and LeRF. Another example is *random perturbation*, such as Infidelity [53], which randomly perturbs the input, unlike ordered perturbation, and computes the empirical average of Eq.(4). Formally, their formulations are given as follows:

$$\text{MoRF}(\boldsymbol{x}) = \frac{1}{L+1}\sum_{k=0}^{L}(\boldsymbol{f}(\boldsymbol{x}_{\text{MoRF}}^{(0)}) - \boldsymbol{f}(\boldsymbol{x}_{\text{MoRF}}^{(k)}))\ , \tag{1}$$

$$\text{LeRF}(\boldsymbol{x}) = \frac{1}{L+1}\sum_{k=0}^{L}(\boldsymbol{f}(\boldsymbol{x}_{\text{LeRF}}^{(0)}) - \boldsymbol{f}(\boldsymbol{x}_{\text{LeRF}}^{(k)}))\ , \tag{2}$$

$$\text{ABPC}(\boldsymbol{x}) = \frac{1}{L+1}\sum_{k=0}^{L}(\boldsymbol{f}(\boldsymbol{x}_{\text{LeRF}}^{(k)}) - \boldsymbol{f}(\boldsymbol{x}_{\text{MoRF}}^{(k)}))\ , \tag{3}$$

$$\text{INFD}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{I}\sim\mu_{\boldsymbol{I}}}(\boldsymbol{I}^T\mathcal{A}(\boldsymbol{x},\boldsymbol{f}) - (\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}-\boldsymbol{I}))^2), \tag{4}$$

where $\boldsymbol{f}$ is the DNN model including the architecture and the trained parameters, $\boldsymbol{x}^{(0)}$ is the original input, $\boldsymbol{x}_{\text{MoRF}}^{(k)}$ is the perturbed input whose top-k features are masked, $\boldsymbol{x}_{\text{LeRF}}^{(k)}$ is the perturbed input whose bottom-k features are masked, and $\mathcal{A}$ is a feature attribution method taking a data sample $\boldsymbol{x}$ and a trained model $\boldsymbol{f}$ as input. Note that if the explanation is of better quality and of better loyalty to the model, the MoRF and ABPC scores are higher and the LeRF is lower. However, the metric of ABPC scores contains the information of both MoRF and LeRF. Without loss of completeness, we do not report the results of LeRF scores. INFD [53] follows the similar idea but its perturbation manner is quite different. Random perturbation on the input space is adopted (or an effective sampling strategy can be designed). This may lead to high computational complexity when the input space is large.

Pseudo Ground Truth. In certain cases, pseudo ground truths can serve as reasonable approximations of the actual ground truths for explanations. For example, pseudo ground truths of explanations can be generated through a consensus-based metric [28]. Here, the consensus refers to the aggregation of explanations from multiple deep models. We can consider this consensus as a pseudo ground truth for the explanation. To evaluate an explanation, we only need to measure its similarity score to this pseudo ground truth. We call this score the PScore. Formally, PScore can be formulated as follows:

$$\text{PScore}(\boldsymbol{x}) = \cos(\frac{1}{|\mathcal{M}|}\sum_{\boldsymbol{g}\in\mathcal{M}}\mathcal{A}(\boldsymbol{x},\boldsymbol{g}), \mathcal{A}(\boldsymbol{x},\boldsymbol{f}))\ , \tag{5}$$

where the cosine similarity is taken, $\mathcal{M}$ is a set of well-trained models and $\mathcal{A}$ is a feature attribution method that takes a data sample $\boldsymbol{x}$ and a trained model $\boldsymbol{f}$ as input.

Some clarifications may be required for elaborating the pseudo ground truth metric.

(1) Take an example of evaluating the faithfulness of a new attribution algorithm $\mathcal{A}$. On the image classification task, we use $\mathcal{A}$ to explain 15 models and get 15 attribution results respectively. Then we aggregate the 15 attribution results through normalization and average and obtain the pseudo ground truth. After that, we measure the similarity score between the pseudo ground truth and each of the 15 attributions, as the PScore for the evaluation result of $\mathcal{A}$ on each model. In this way, the faithfulness evaluation of $\mathcal{A}$ using the PScore metric is done.

(2) This metric requires some assumptions and preconditions, that

- The models in $\mathcal{M}$ should be well-trained, otherwise both the predictions and attributions can be random and bad. According to our experiments, ImageNet-pretrained models released in public are safe to use for image classification tasks.
- The number of models should be large. The original paper [28] suggests using 15 models, while we use 9 models for image tasks and 6 models for text tasks for reducing the computation complexity.

https://doi.org/10.52202/075280-0081

Synthetic Ground Truth. Synthetic datasets with sophisticated designs similar to "data poisoning" or "adversarial attacks" can provide synthetic ground truths of explanations [32]. Note that the "attacks" we intentionally apply here are noticeable and describable, simply consisting of painted patches. The idea is that the synthetic patches on the images serve as supervision signals for training models. Labels of the images with synthetic patches will be reversed. These patches constitute explanation ground truths because no other patterns can lead to correct predictions by design. Therefore, models trained on such datasets must attribute their predictions to the synthetic ground truths. To explain a well-trained model, effective feature attribution methods should produce explanations that closely match the synthetic ground truths. We propose using Average Precision (AP) and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to quantify the matching score, referred to as SynScore. Formally, SynScore can be defined as follows:

$$\text{SynScore}_{metric}(\boldsymbol{x}) = metric(Syn(\boldsymbol{x}), \mathcal{A}(\boldsymbol{x}, \boldsymbol{f}))\ , \tag{6}$$

where $metric$ can be AP or AUC-ROC and $Syn(\boldsymbol{x})$ is the synthetic ground truth of explanation for the sample $\boldsymbol{x}$.

These three categories of metrics are commonly used in recent work, but each has its own limitations. We do not expect that any single category accurately measures the faithfulness of feature attribution methods. Evaluating them together through cross-comparative analysis can help to understand their strengths and weaknesses. This could provide useful insights until more principled metrics emerge.

2.4 Benchmark Pipeline and Modular Implementations

We currently provide evaluations on two data modalities, i.e., images and texts. For images, we choose a small amount (5000) of images from ImageNet as the benchmark dataset, multiple pre-trained image classification networks as the benchmark models and multiple feature attribution methods as the benchmark explainers. Each explainer will produce an explanation given a model for each image in the dataset, and each explanation will be given a score by applying the evaluation metrics. Finally, the faithfulness of the explainer will be quantified by the average of scores across all images in the dataset. For texts it is very similar except that the base dataset is MovieReview and the base models are for the task of sentiment analysis. The pipeline is illustrated in Figure 1.

The benchmark pipeline is implemented following InterpretDL [29] in a modular-designed style and is publicly available [2]. This means that from deep models, feature attribution methods, to evaluation metrics, their implementations are independent modules. A code sample to obtain the explanation and the evaluation results is shown in Listing 1. Therefore, new methods and metrics can also be easily added and compatible across deep learning library frameworks [3]. For example, one can use Pytorch to obtain the explanations, e.g., Captum [26], and use our benchmark to do the evaluations. To show the benchmark utility, we have provided a user-case scenario of using a HuggingFace model [4] for feature attributions and faithfulness evaluations on the X-ray pneumonia classification task. See the source code in the supplementary material for details.

3 Experiments and Observations

Following the benchmark pipeline as described in the previous section, we have conducted evaluation experiments. In this section, we present the experimental results, where we address several interesting research questions and introduce observations that could guide future work.

To acquaint readers with our experimental results, we first present an illustrative example. We compare this example with constant and random baselines to confirm the effectiveness of all the feature attribution methods. The results show the scores obtained by applying all possible feature attributions and measuring them using all metrics on ResNet-50. Table 1 presents the results, where the constant and random baselines perform significantly worse than the other three attribution methods (i.e., GradCAM, IG and SG). Moreover, no single attribution method achieves the best performance across all faithfulness metrics, highlighting the need for our benchmark. For rigorous analysis, the statistics in the following subsections exclude the constant and random baselines.

[2] `https://github.com/PaddlePaddle/InterpretDL`

[3] See a simple demo using InterpretDL with Pytorch models: `https://colab.research.google.com/drive/1ZgI1ctCc2ryPk0bdPgkEwQCJ1tHZCq14`

[4] `https://huggingface.co/nickmuchi/vit-base-xray-pneumonia`

https://doi.org/10.52202/075280-0081

```python
import interpretdl as it

# Load a pretrained model from PaddlePaddle model zoo.
from paddle.vision.models import resnet50
model = resnet50(pretrained=True)

# Available feature attribution methods include but are not limited to
     SG, IG, LIME, BT, GA and etc. 'interpret' is the universal api.
algo = it.SmoothGradInterpreter(model, device="gpu:0")
expl_result = algo.interpret("test.jpg")

# Available faithfulness evaluation metrics include but are not
    limited to MoRF, ABPC, INFD and etc. 'evaluate' is the universal
    api. Note that some evaluators do not require the model.
evaluator = it.Infidelity(model)
eval_result = evaluator.evaluate("test.jpg", expl_result)
```

Listing 1: Codes for computing the SmoothGrad explanation and the INFD metric score, given the model of a pretrained ResNet-50 and a testing image.

Table 1: Evaluation results on the ImageNet-pretrained ResNet-50, with constant and random baselines, GradCAM, IG and SG. "Random-16" indicates that the random values are given at patch level (16×16 pixels) while "Random" is at pixel level.

Attribution Methods	**MoRF** ↑	**ABPC** ↑	**PScore** ↑	**INFD** ↓	**SynScore** ↑
Constant	0.000	0.000	N/A	3.015	0.072
Random-16	0.596	0.007	N/A	3.039	0.077
Random	0.599	0.008	N/A	3.015	0.078
GradCAM	0.628	**0.424**	**0.835**	2.496	**1.000**
IG	**0.709**	0.377	0.812	2.373	0.999
SG	0.701	0.369	0.820	**2.323**	0.998

Evaluating the faithfulness of attribution algorithms presents several challenges from three perspectives: metric-wise, attribution algorithm-wise, and classification model-wise. All of these perspectives can lead to variance in faithfulness benchmarking. We will analyze faithfulness performance from these three perspectives in the following subsections.

3.1 Whether There are Two Metrics that are Correlated?

Selecting appropriate metrics is critical but also challenging given the multitude of options. We introduced three families of metrics to evaluate the faithfulness of explainability methods. An intriguing research question is whether any two metrics are correlated. The short answer is yes. To investigate the inter-correlations between metrics, whether they are correlated or orthogonal, we compute the Pearson's correlation coefficient between metrics considering all possible pairs of models and explanation techniques. We can see from Figure 2 that first, the most positive correlation coefficient is located in the pair between ABPC and PScore (0.58, p-value $2.4e^{-4}$); Second, the most negative among all pairs is between INFD and PScore (-0.59, p-value $1.9e^{-4}$); Third, there is a near zero correlation for the MoRF-PScore, MoRF-INFD and PScore-SynScore pairs, indicating that the pairs are not correlated at all.

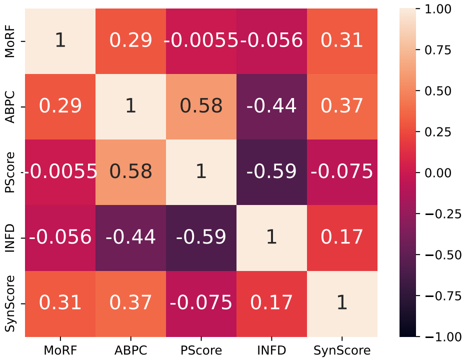
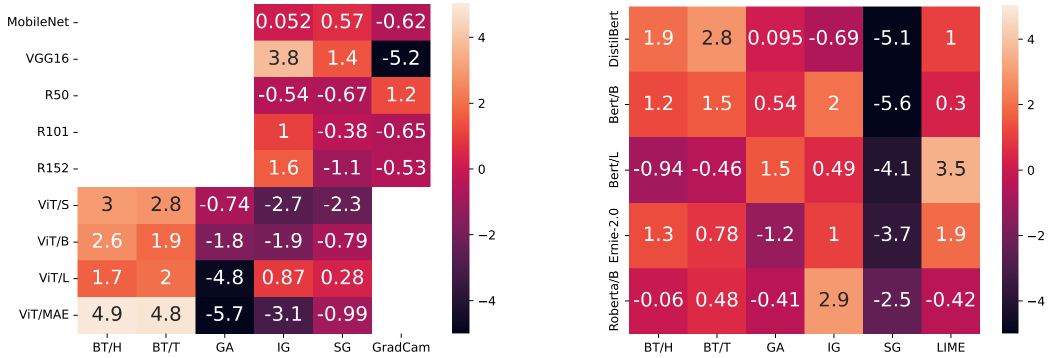


Figure 2: Correlation between metrics.

As noticed, ABPC, INFD, and PScore are potential alternatives to one another, each requiring a good amount of computation from different perspectives. Besides one trial of explanation algorithm, ABPC needs tens of forward passes of the original model. INFD requires generating random masks, which



(a) Modality of images.

(b) Modality of texts.

Figure 3: Averaged metric scores. A higher value indicates better faithfulness. The blanks indicate that the algorithm in a vanilla style is not suitable for the model.

is sampled from a very large space, especially for images. Meanwhile, PScore necessitates trained models' availability, with each model passing the explanation algorithm once. Practitioners should choose the most appropriate method based on the availability of models and computational resources.

There are two other key observations worth discussing. First, MoRF is weakly correlated with ABPC, with a correlation coefficient of 0.29, indicating that they are not measuring exactly the same characteristic. As discussed by Samek et al. [42], MoRF focuses only on the most important features, while ABPC also considers the ranking of the least important features. Therefore, if the goal is to filter out irrelevant features, ABPC scores should be more heavily weighted. Second, we found that ABPC, PScore, and INFD are strongly correlated with each other. In contrast, MoRF and SynScore evaluate faithfulness from different perspectives than the former three metrics.

3.2 Which Explanation Algorithm Demonstrates the Best Faithfulness?

Another interesting question is determining which explanation algorithm is the most faithful. Depending on the models and faithfulness metrics used, the optimal algorithm may differ. However, we still can draw several useful and instructive conclusions.

To simplify the comparison of faithfulness across explanation algorithms, we aggregate the multi-dimension metrics into a single one. First, we make the assumption that the metrics in our benchmark measure the faithfulness from different aspects since no pairs are perfectly correlated. Then we can propose a single score by averaging all of the metrics while negating the scores of INFD which is the only one ranking the faithfulness in a descent order. Moreover, we do the standardization for scores of all metrics within each model to balance the contributions of metrics. Specifically, the averaged metric is defined as:

$$\text{AvgScore} = z(\text{MoRF}) + z(\text{ABPC}) + z(\text{PScore}) - z(\text{INFD}) + z(\text{SynScore}) \ , \tag{7}$$

where $z(s) = (s - \bar{s})/\sigma(s)$ is the standardization within each model. Following the formulas, we compute the averaged faithfulness score for each model-algorithm pair and show the results in Figure 3. We provide the results of each metric in the supplementary materials.

We summarize the observations as follows:

- Overall, IG generally outperforms SG except for Vision Transformers. For Transformers in NLP, SG is not an optimal choice. One possible reason is that SG adds noise in the embedding layer, and the noise scale is difficult to tune. Though further investigation is needed, we believe that the theoretical guarantee (IG satisfies the Completeness axiom, i.e., the attributions add up to the difference between the output of the model at the input and a chosen baseline) may be one of the reasons to support wide applications.

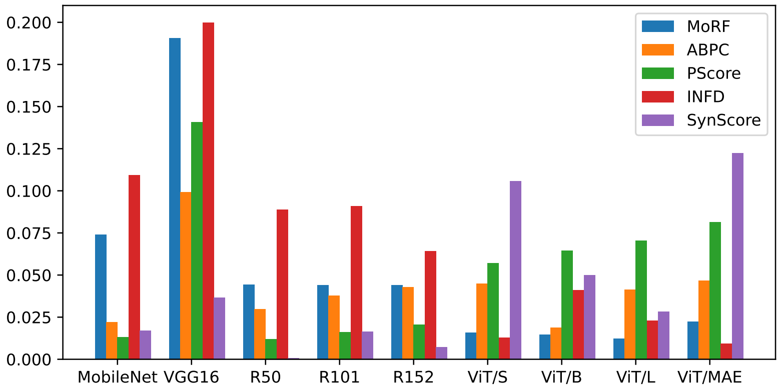


Figure 4: Sensibility measured by standard deviations across attribution methods. Note that the std for VGG16-INFD is very high (86.5) whereas we set the limit to 0.2 for better visualization.

- LIME steadily gets a high averaged score for NLP Transformers[5]. One reason is that the algorithmic computation of LIME overlaps *in some degree* with the evaluation of MoRF and ABPC. This is more obvious if we see the results of MoRF and ABPC, which are in the supplementary materials. Nevertheless, LIME is among the best algorithms measured by other metrics as well.
- For attention-based networks, including both ViTs and NLP Transformers, BT generally demonstrates higher faithfulness than others. This may stem from its accurate approximation of Transformer computations. This observation coincides with the implication of the first observation, motivating future work to design explanation algorithms through mathematical analysis of network structure, e.g., attention modules in Transformers.
- Even within the same network structure (here we have three sets of model families: ResNet-{50,101,152}, ViT-{Small,Base,Large} and Bert-{Distil,Base,Large}), no algorithm consistently achieved high faithfulness. For example, IG performed well on R101 and R152 but not R50. Although BT demonstrated the highest faithfulness in most cases, it did not do so for Bert/L. Developing a faithful algorithmic technique that works across different metrics, modalities and models remains an open challenge.

3.3 Which Model is the Most (In)sensitive to Explanation Algorithms?

The model's complexity and interpretability is also a key factor influencing the benchmarking performance. From the extensive evaluation results, we further investigate the sensitivity of the models to attribution methods. Given a model and a metric, we calculate the standard deviation of all metric scores across possible attribution methods. Take ResNet-50 as an example. For each evaluation metric, a standard deviation is computed among GradCAM, IG, and SG. The results are shown in Figure 4 for image classification models and Figure 5 for NLP models.

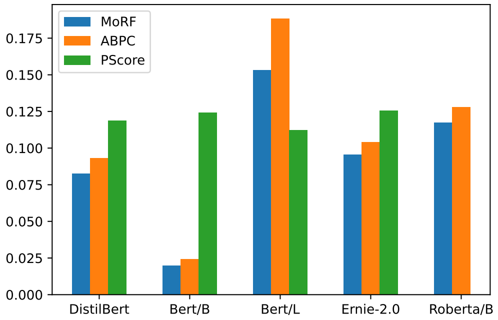


Figure 5: Sensibility measured by standard deviations across attribution methods. Since the vocabulary of Roberta/B is different from other models, thus PScore is excluded.

A model that is insensitive to attribution methods may indicate that the model can be easily explained, or all attribution methods fail to explain the model. Fortunately, the latter case does not occur in our experiments, because the attribution methods we selected for the models are relatively suitable. So, with the sensitivity, we can roughly estimate the difficulty of explaining a model. This may be helpful when designing novel network structures.

[5] For fair comparison, we did not involve the results of LIME in Figure 3 because the implementation of LIME involves a procedure of superpixel segmentation for reducing the computation complexity.

https://doi.org/10.52202/075280-0081

Although the most insensitive model is not easy to identify, the most sensitive one is VGG16, which gets the highest sensitivity in almost all metrics. For network families, we can find that ResNets get higher sensitivity than ViTs in MoRF and INFD but lower in PScore and SynScore. In either network family, the sensitivity does not vary much. As for NLP models, they have similar sensitivity as well, except Bert/B in the metrics of MoRF and ABPC. A similar observation is found in ViT/B by the ABPC metric. Existing attribution methods often use ViT/B or BERT as the primary model and achieve good evaluation results. However, some methods may work especially well for ViT/B or BERT/B but not as well for other models. Therefore, we encourage the research community to evaluate attribution methods on a variety of models with different network architectures.

4 Related Work

Existing work has developed benchmarks for evaluating and comparing explainability approaches. For example, Rathee et al. proposed BAGEL, a benchmark for evaluating explanation methods on graphical neural networks [39]. OpenXAI provides an open-source framework for evaluating post hoc explanation methods in tabular data [3]. Similarly, other benchmarks focus on NLP models, such as [14, 52]. The XAI-Bench library benchmarks feature attribution methods on synthetic datasets [33]. Chou et al. proposed a benchmark for counterfactual explanation methods on tabular data [11]. However, these benchmarks are limited to specific data modalities and explanation methods. A benchmark that considers multiple data modalities and explanation paradigms is still lacking. Our work addresses this gap by proposing a unified benchmark for explainable AI across different modalities, with the goal of facilitating holistic progress in the field of XAI.

5 Discussions, Limitations and Future Work

In this section, we provide discussions and limitations on the proposed benchmark $\mathcal{M}^4$. We also present our plans of future work for addressing the limitations.

In addition to including a wide range of feature attribution methods, faithfulness evaluation metrics, data modalities, and deep models, our benchmark $\mathcal{M}^4$ has two other properties. **The first one is efficiency and facility.** The benchmark $\mathcal{M}^4$ utilizes subsets of public datasets and evaluates the same datasets for each modality, i.e. ImageNet and MovieReview. Moreover, the models used during evaluation consist of publicly available pre-trained model weights, avoiding training new models from scratch, except for fine-tuning required in sentiment analysis. **The second one is objectiveness.** The $\mathcal{M}^4$ pipeline is objective and performed by evaluation algorithms because the faithfulness evaluation depends only on attribution methods and deep models.

We would also like to distinguish between faithfulness and interpretability. Interpretability refers to the alignment between the explanations of a model and human understanding [25]. Faithfulness is a prerequisite for interpretability and refers to how well an explanation reflects the model's functioning. This paper focuses on evaluating the faithfulness, specifically of feature attribution methods, across metrics, models, and modalities. We do not focus much on the interpretability of deep models in this work but our benchmark can be easily extended to its evaluations, e.g., with the help of ground-truth labels of image segmentation [19] and language reasoning [14].

We present the limitations and plans in the future work.

(1) The evaluation metrics in XAI contain several others beyond faithfulness, e.g., interpretability, sparsity, stability *etc*, while the current version of $\mathcal{M}^4$ only focuses on the faithfulness. For the comprehensive applicability of the XAI benchmark, we will progressively integrate other evaluation metrics for feature attributions. Some are easy to be plugged in. For example, the sparsity can be directly computed via the entropy of normalized attributions, but for the reason that the sparsity is not one of the faithfulness metrics, it is thus not reported in the current version of $\mathcal{M}^4$ benchmark. Another reason that we do not involve other aspects in the benchmark and focus on the faithfulness evaluations of feature attributions, is that we believe that based on faithful explanation results, we can more easily and accurately analyze other aspects of XAI. In the future, various metrics will be included in $\mathcal{M}^4$, contributing $\mathcal{M}^4$ for more comprehensive applicability in XAI.

(2) Our benchmark $\mathcal{M}^4$ did not include many good attribution methods, such as Shapley values based methods [35, 10], CAM variants [51, 36, 23], LIME variants [58, 30], LRPs [5, 49] attention-

https://doi.org/10.52202/075280-0081

based [1] and many others. However, explanations are of great variety. Although many advanced explanations beyond feature attributions have been proposed to facilitate deeper understanding of deep neural networks, their faithfulness is difficult to evaluate *a posteriori* and would be evaluated *ad hoc*. As an initial stride toward establishing a benchmark for evaluating XAI methods, concentrating on feature attribution methods would be an attainable endeavor, albeit one that still presents its own set of challenges.

(3) Language models and their explanations are only evaluated by the task of sentiment analysis. Though we are interested in the explanation faithfulness instead of language models' capacities and the sentiment analysis is one of the accessible tasks for faithfulness evaluations, it would be comprehensive to evaluate on other NLP tasks, e.g., those from the GLUE benchmark [50, 14, 57].

(4) Our benchmark $\mathcal{M}^4$ contains currently the image and text modalities. One of the future directions is to enhance the benchmark by integrating more data modalities, such as graphs, audio clips, tabular data, and multi-modality, whereas several pioneering studies [6, 9] have explored multi-modal explanations using some of the evaluation metrics in our benchmark pipeline.

(5) Social impacts and ethics. Our framework can assess bias and fairness issues of DNNs in high-stake applications involving sensitive attributes like gender, race, and age. This is a challenging topic and would be investigated in our future research.

6 Conclusions

Although existing benchmarks have advanced XAI in specific domains, a universal benchmark is lacking to compare explanation methods between models and modalities. Our work aims to address this gap by evaluating feature attribution methods on computer vision and NLP tasks using a variety of metrics. In our benchmark, we evaluated nine models using six of the most common explanation methods (LIME, SG, IG, GradCAM, GA and BT) on two modalities (image and texts) based on five evaluation metrics. We gain several observations that can inform the future design and applications of the XAI method. For future work, we plan to expand the benchmark to other modalities, including but not limited to graphs and audio. We also plan to incorporate additional evaluation perspectives, such as interpretability, stability, sparsity, *etc*.

Acknowledgments

Xuhong Li and Haoyi Xiong were supported in part by the National Key R&D Program of China under the gradnt No. 2021ZD0110303.

References

[1] Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Proceedings of the Advances in Neural Information Processing Systems*, 2018.

[3] Chirag Agarwal, Eshika Saxena, Satyapriya Krishna, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 2022.

[4] Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pages 435–451. PMLR, 2022.

[5] 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*. Springer, 2016.

https://doi.org/10.52202/075280-0081

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[8] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[9] Jiamin Chen, Xuhong Li, Lei Yu, Dejing Dou, and Haoyi Xiong. Beyond intuition: Rethinking token attributions inside transformers. *Transactions on Machine Learning Research*, 2022.

[10] 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.

[11] Yu-Liang Chou, Chihcheng Hsieh, Catarina Moreira, Chun Ouyang, Joaquim Jorge, and João Madeiras Pereira. Benchmark evaluation of counterfactual algorithms for xai: From a white box to a black box. *arXiv preprint arXiv:2203.02399*, 2022.

[12] 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.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

[15] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations, ICLR*, 2021.

[17] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

[18] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. Black box explanation by learning image exemplars in the latent feature space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019.

[19] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110:328–348, 2014.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

https://doi.org/10.52202/075280-0081

[23] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 2021.

[24] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the International Conference on Machine Learning*, 2018.

[25] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[26] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[27] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

[28] Xuhong Li, Haoyi Xiong, Siyu Huang, Shilei Ji, and Dejing Dou. Cross-model consensus of explanations and beyond for image classification models: An empirical study. *arXiv preprint arXiv:2109.00707*, 2021.

[29] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Zeyu Chen, and Dejing Dou. Interpretdl: Explaining deep models in paddlepaddle. *Journal of Machine Learning Research*, 23(197):1–6, 2022.

[30] Xuhong Li, Haoyi Xiong, Xingjian Li, Xiao Zhang, Ji Liu, Haiyan Jiang, Zeyu Chen, and Dejing Dou. G-lime: Statistical learning for local interpretations of deep neural networks using global priors. *Artificial Intelligence*, 314:103823, 2023.

[31] Xuhong Li, Haoyi Xiong, Yi Liu, Dingfu Zhou, Zeyu Chen, Yaqing Wang, and Dejing Dou. Distilling ensemble of explanations for weakly-supervised pre-training of image segmentation models. *Machine Learning*, pages 1–17, 2022.

[32] Yi-Shan Lin, Wen-Chuan Lee, and Z Berkay Celik. What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1027–1035, 2021.

[33] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. In *Advances in Neural Information Processing Systems Datasets Track*, 2021.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[35] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the Advances in Neural Information Processing Systems*, 2017.

[36] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[37] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021.

[38] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference*, 2018.

[39] Mandeep Rathee, Thorben Funke, Avishek Anand, and Megha Khosla. Bagel: A benchmark for assessing graph neural network explanations. *arXiv preprint arXiv:2206.13983*, 2022.

https://doi.org/10.52202/075280-0081

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[41] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[42] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[43] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[44] 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*, 2017.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[47] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975, 2020.

[48] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

[49] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5797–5808. Association for Computational Linguistics, 2019.

[50] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

[51] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[52] Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. A fine-grained interpretability evaluation benchmark for neural nlp. *arXiv preprint arXiv:2205.11097*, 2022.

[53] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I Inouye, and Pradeep Ravikumar. On the (in) fidelity and sensitivity for explanations. *arXiv preprint arXiv:1901.09392*, 2019.

[54] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020.

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

[56] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

https://doi.org/10.52202/075280-0081

[57] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*, 2023.

[58] Zhengze Zhou, Giles Hooker, and Fei Wang. S-lime: Stabilized-lime for model explanation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2429–2438, 2021.

https://doi.org/10.52202/075280-0081