What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods

Thomas Fel, Julien Colin, Rémi Cadène, Thomas Serre

What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods

Thomas Fel[1,3,4 *] **Julien Colin**[1,3 *] **Rémi Cadène**[1,2 †] **Thomas Serre**[1,3]

[1]Carney Institute for Brain Science, Brown University, USA [2]Sorbonne Université, CNRS, France
[3]Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
[4] Innovation & Research Division, SNCF

{thomas_fel, julien_colin, remi_cadene}@brown.edu

Abstract

A multitude of explainability methods and theoretical evaluation scores have been proposed. However, it is not yet known: (1) how useful these methods are in real-world scenarios and (2) how well theoretical measures predict the usefulness of these methods for practical use by a human. To fill this gap, we conducted human psychophysics experiments at scale to evaluate the ability of human participants $(n = 1,150)$ *to leverage representative attribution methods to learn to predict the decision of different image classifiers. Our results demonstrate that theoretical measures used to score explainability methods poorly reflect the practical usefulness of individual attribution methods in real-world scenarios. Furthermore, the degree to which individual attribution methods helped human participants predict classifiers' decisions varied widely across categorization tasks and datasets.*

Overall, our results highlight fundamental challenges for the field – suggesting a critical need to develop better explainability methods and to deploy human-centered evaluation approaches. We will make the code of our framework available to ease the systematic evaluation of novel explainability methods.

1. Introduction

There is now broad consensus that simply evaluating the test accuracy of computer vision systems does not provide a sufficient guarantee that these systems are safe to be deployed in the real-world [34,42] as those systems are capable of exploiting dataset biases and other statistical shortcuts to achieve unprecedented levels of accuracy [13, 21]. A growing body of research thus focuses on making modern computer vision systems more trustworthy – in part via the development of explainability methods to help interpret their predictions [31, 37, 39, 41, 46, 49, 51]. Such explainability methods will find broad societal uses – like easing the debugging of self-driving vehicles [54] and helping to fulfill the "right to explanation" that European laws guarantee to its citizens [22].

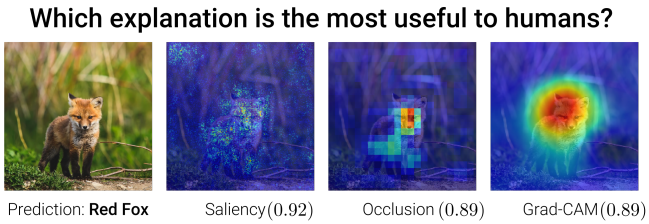

Figure 1. We study the relationship between the theoretical scores used to evaluate explainability methods versus their usefulness in practical cases. A motivating example is proposed here showing three commonly used attribution methods ranked from most faithful (Saliency) to least faithful (Grad-CAM) according to the commonly used Deletion metric [37, 51]. The image on the left is correctly classified by the classifier as a "Red Fox". Surprisingly, the most faithful explanation (Saliency) seems more complex to grasp for a human observer and is likely to be of little use in real-life situations. Here we conducted large-scale psychophysical experiments to evaluate the practical utility of these explanations and contrast the results with those of commonly used metrics.

Here, we consider the problem of evaluating state-of-the-art explainability methods. Arguably, these methods should be evaluated directly according to their *usefulness* to a human experimenter, i.e., how much they help understand how a system classifies arbitrary inputs. However, in practice, these methods are instead evaluated using surrogate measures defined axiomatically [6, 9, 32, 54]. Building on previous work [20,30], we propose a practical and actionable definition of an explainable model which we use to introduce a human-centered evaluation framework.

[*] Equal contribution [†] Work done before April 2021 and joining Tesla

A Meta-predictor perspective. Before providing a rigorous definition of interpretability, let us motivate our approach with an example: a linear classifier is often considered to be readily interpretable by a trained user because its inner working is sufficiently intuitive enough that it can be comprehended by a human user. A user can in turn build a mental model of the classifier such that it can potentially successfully predict the classifier's output for arbitrary inputs (i.e., by visually matching the degree to which "important" features associated to high classification weights are present in the input). In essence, we suggest that the model is interpretable because the output can be predicted. This concept of predicting the classifier's output is central to our approach and we conceptualize the human user as a *Meta-predictor* of the machine learning model. This notion of Meta-predictor is also closely related to the notion of *simulatability* [15, 20, 23, 30, 38]. We will define the term more formally in Sec. 3.1.

In contrast to linear classifiers, deep neural network models are highly non-linear and possess numerous parameters which render them hardly interpretable to human users. To make these models more interpretable, many different explainability methods have been introduced [11, 20, 37, 39, 41, 45, 48, 49, 51, 55]. In vision, these methods typically render an explanation by calculating an importance score for each input pixel which are then displayed as a heatmap – also called an attribution map – revealing regions of the image that support the model's decision. Despite the existence of a wide range of explainability methods, assessing the quality and reliability of these methods remains an open problem. The community has taken up the issue and proposed two commonly used approaches to evaluate methods: one based on ground truth annotations [17, 19, 20, 37, 41, 51, 56] and a second more commonly used on *faithfulness* metrics [19, 20, 29, 37, 40, 51].

In line with previous approaches [23, 35, 39, 43], we put humans at the center of the evaluation procedure. More specifically, we focus on measuring how useful explanations from common methods are to human users in the sense that they can be leveraged to build a more accurate meta-predictor of the model to be explained. We ran a series of large-scale online psychophysics experiments to evaluate the usefulness of 6 representative attribution methods for several binary categorization tasks derived from 3 different datasets (a standard biased dataset, a fine-grained dataset of leaf images, and ImageNet). We evaluated the ability of human participants ($n = 1,150$) to use attribution maps to learn to predict the decision of different image classifiers. Our results demonstrate that (1) the degree to which attribution methods were helpful to human observers varied widely across categorization tasks; (2) the *faithfulness* [37, 40] and our usefulness measures appeared to exhibit only a very weak correlation. These results suggest that faithfulness measures are surprisingly poor predictors of the practical usefulness of current attribution methods. In the next section, we highlight the originality of our proposed human-centered evaluation framework compared to related approaches. Then, we formalize our usefulness measure and describe our experimental design with its control experiments. Finally, we report our main experiments on 3 datasets including ImageNet and discuss how our results suggest a critical need for better explainability methods and better quality measures for these methods.

2. Related work

Evaluations based on ground-truth annotations A first class of evaluation approaches scores explainability methods according to their ability to identify image locations that overlap with the target object defined either by a human-derived bounding box or a segmentation mask [17, 19, 20, 37, 41, 51]. More recently, the evaluation method called Pointing Game [56] counts the number of times the most important region according to the explanation intersects with the location of the object to classify.

All the aforementioned evaluation methods suffer from a critical flaw. They assume that image classifiers rely on visual features that belong to the target object. This is obviously not the case as they often rely on contextual features and all kinds of other visual shortcuts [21]. Indeed, for a model not using the target object but shortcuts for its decision, a proper explainability method which points to this shortcut will be unfairly penalized (for not having pointed to the mask of the object). Conversely, our approach does not make any such assumptions. Irrespective of whether image classifiers provide correct or erroneous predictions, our approach assesses whether explanations ultimately help (or not) a user better understand the basis for classifying images.

Evaluations based on faithfulness measures A second class of approaches has recently started to emerge and do not make use of any ground-truth annotations – specifically to avoid some of the aforementioned limitations. They introduce measures of faithfulness to the model. Common approaches [37, 40] measure the change in the classification score when the most important pixels are progressively removed. The bigger the drop, the more faithful the explanation method is. To ensure that the drop in score does not come from a change in distribution of the perturbated images, the ROAR [25] methods include an additional step where the image classifier is re-trained in between each removal step. These methods do not require ground-truth annotations and hence can be used even for datasets that do not include object masks or bounding boxes. As a consequence, they are quite popular in computer vision [19, 20, 29, 37, 40, 51] and natural language processing [4, 5, 51].

Nevertheless, *faithfulness* measures have recently been criticized as they all rely on a baseline to remove important areas, a baseline that will obviously give better scores to methods relying internally on the same baseline. [26]. More importantly, they do not consider humans at any time in the evaluation. As a result, it is not clear if the most faithful explanation methods can be practically useful to humans.

Evaluations based on human data A third class of approaches consists in evaluating the ability of humans to leverage explanations for different purposes [2, 8, 10, 23, 33, 35, 36, 39, 41, 43]. For instance, in [41] and [39], explanations are evaluated according to how much they help human participants' to identify biases in models. In [39], a classifier was trained on a biased dataset of wolves and huskies: the model consistently uses the background to classify. They asked participants if they would trust the model a first time after showing images and predictions without explanations, and a second time after showing the explanations on the same images. Showing explanations had an effect on the reported trust. We also report results on this dataset. However, our psychophysical experiments differ greatly from those presented in [39] in that we do not ask users if they feel a sentiment of trust, but directly measure their ability to predict the output of the model in the absence of explanations on unseen test images. We also introduce several controls with noisy explanations.

Another line of closely related work [23, 35, 43] centers around the notion of *simulatability* which was first introduced in [30] and later refined in [15]. They introduce different experimental procedures to measure if humans can learn from the explanations how to copy the model prediction on unseen images. Some provide the explanations at test time [35, 43]. Similar to ours but for tabular data, [23] proposes a more difficult procedure where the explanations are hidden at test time. This forces the participants to learn the mechanisms to simulate the model at training time where the explanations are shown. Differently to ours, they also provide the groundtruth class associated with the input image during training. We argue that it should be removed since the end goal is to predict the model outputs and not to identify when the model gets an accurate prediction. A second key difference is that they have two phases with each time a training and testing time. They measure the difference between the first phase where the participants do not see any explanations and the second phase with the same examples where explanations are shown at training time. We argue that any improvement observed during the experimental condition could be the result of a learning effect because participants saw the same samples and the associated class label twice. Instead, we only have a single phase and we measure the impact of explanations by comparing results of participants that had explanations at training time with others that did not. We made sure that our results are statistically significant by having enough participants in our study.

3. A Meta-predictor evaluation framework

3.1. Formalism

We consider a standard supervised learning setting where $\boldsymbol{f}$ is a black-box predictor that maps an input $\boldsymbol{x} \in \mathcal{X}$ (e.g., an image) to an output $\boldsymbol{f}(\boldsymbol{x}) \in \mathcal{Y}$ (e.g., a class label). One of the main goals of explainable AI is to yield useful rules to understand the inner-working of a model $\boldsymbol{f}$ such that it is possible to infer its behavior on unseen data points. To this end, a current approach consists in studying explanations (attribution map, concept vectors, feature visualization) for several predictions. Formally, $\boldsymbol{\Phi}$ is any explanation functional which, given a predictor $\boldsymbol{f}$ and a point $\boldsymbol{x}$, provides an information $\boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x})$ about the prediction of the predictor. In our experiments, $\boldsymbol{\Phi}$ is an attribution method.

Understandability-Completeness trade-off. Different attribution methods will typically produce different heatmaps – potentially highlighting different image regions and/or presenting the same information in a different format. The quality of an explanation can thus be affected by two factors: *faithfulness* of the explanation (i.e., how much pixels or input dimensions deemed important effectively drive the classifier's prediction) and the understandability of the explanation for a user.

At one extreme, an explanation can be entirely *faithful* and provide all the information necessary to predict how a classifier will assign a class label to an arbitrary image (by rendering all the parameters of the classifiers). However, such information will obviously be overly complex to be understood by a user and hence it is not *understandable*. It should be noted that when this is understandable, for a linear model, for example, it is generally the most suitable explanation. Conversely, a simple explanation will be *understandable* but it may ultimately mislead a user if it is not *faithful*. That is to say, just because a human agrees with an explanation does not necessarily mean that it reflects how the model is working.

In the end, this means that there is a trade-off between the amount of information provided by an explanation and its comprehensibility to humans. In the middle of this trade-off lies the most useful explanation.

Focus on *usefulness*. We describe a new human-centered measure that incorporates this trade-off into a single *usefulness* measure by empirically evaluating the ability of human participants to learn to "predict the predictor", i.e., to be an accurate Meta-predictor. Indeed, if an explanation allows users to infer precise rules for the functioning of the predictor on past data, the correct application of these same rules should allow the user to correctly anticipate the model's

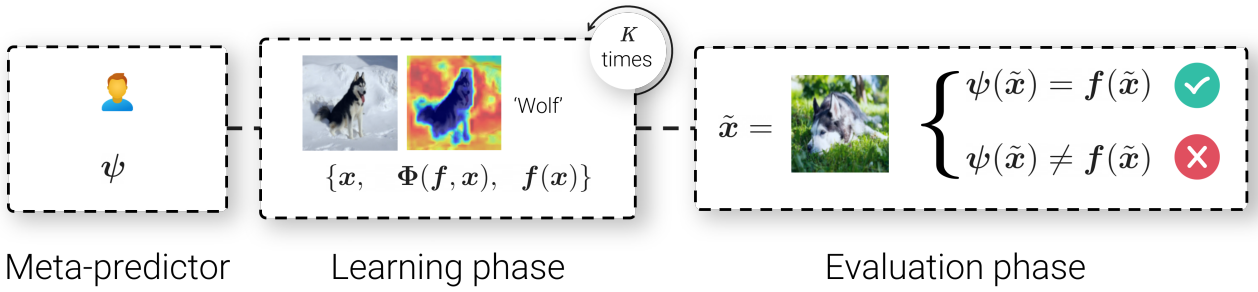

Figure 2. We describe a human-centered framework to evaluate explainability methods borrowing from the concept of Meta-predictor. The framework requires a black box model $\boldsymbol{f}$, an explanation method $\boldsymbol{\Phi}$ and a human subject $\boldsymbol{\psi}$ which will try to predict the predictor, hence, the name Meta-predictor. The first step is the learning phase where the Meta-predictor is training using K samples $\boldsymbol{x}$, together with the associated model predictions $\boldsymbol{f}(\boldsymbol{x})$ and explanations $\boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x})$. The goal of this learning phase is for the Meta-predictor to try to uncover a rule describing the functioning of the model from the triplets $(\boldsymbol{x}, \boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}))$. Then, the second step is the evaluation phase where we test the Meta-predictor accuracy on new samples $\tilde{\boldsymbol{x}}$ by comparing its predictions $\boldsymbol{\psi}(\tilde{\boldsymbol{x}})$ to those of $\boldsymbol{f}(\tilde{\boldsymbol{x}})$. The better his predictions are, the more useful the rules learned during his training are and therefore the method is relevant.

decisions on future data. Scrutable but inaccurate explanations will result in an inaccurate Meta-predictor – just like accurate inscrutable ones. This Meta-predictor framework avoids current pitfalls such as confirmation bias - just because a user likes the explanation does not mean they will be a better Meta-predictor - or prediction leakage on the explanation - in simulatability experiments, as the explanation is available during the test phase, any explanation that leak the prediction would have a perfect score, without giving us any additional information about the model-. We will now formally describe the metric build using this framework.

We assume a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x}_i)\}_{i=1}^{K}$ used to train human participants to learn to predict a classifier's output $\boldsymbol{f}$ from K samples made of an input image $\boldsymbol{x}_i$, the associated predictions $\boldsymbol{f}(\boldsymbol{x}_i)$ and explanations $\boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x}_i)$. We denote $\boldsymbol{\psi}^{(K)}$ a human Meta-predictor after being trained on the dataset $\mathcal{D}$ (see Fig. 2) using explanations. We denote $\boldsymbol{\psi}^{(0)}$ human meta-predictions after participants were trained on the same dataset but without explanations to offer baseline accuracy scores.

We can now define the usefulness of an explainability method $\boldsymbol{\Phi}$ after training participants on K samples through the accuracy score of the Meta-predictor normalized by the baseline Meta-predictor accuracy:

$$\textit{Utility-}K = \frac{\mathbb{P}(\boldsymbol{\psi}^{(K)}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}))}{\mathbb{P}(\boldsymbol{\psi}^{(0)}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}))} \tag{1}$$

Utility-K score thus measures the improvement in accuracy that the explanation has brought. It is important to emphasize that this *Utility* measure only depends on the classifier prediction and not on the ground-truth label as recommended by [28]. After fixing the number of training samples K, we compare the normalized accuracy of different Meta-predictor. The Meta-predictor with the highest score is then the one whose explanations were the most useful as measures compared to a no-explanation baseline.

***Utility* metric.** In practice, we propose to vary the number of observations $K \in \{K_0, ..., K_n\}$ and to consider an aggregated Utility score by computing the area under the curve (AUC) of the *Utility-K*. The higher the AUC the better the corresponding explanation method is. Formally, given a curve represented by a set of n points $\mathcal{C} = \{(K_0, \textit{Utility-}K_0), ..., (K_n, \textit{Utility-}K_n)\}$ where $K_{i-1} < K_i$ we define our main metric $\textit{Utility} = AUC(\mathcal{C})$.

3.2. Experimental design

We first describe how participants were enrolled in the study, the exclusion criteria used to filter out uncooperative online participants, and our general experimental design (including both experimental and control conditions).

Participants Behavioral accuracy data were gathered from $n = 1,150$ participants using the Amazon Mechanical Turk (AMT) platform (www.mturk.com). AMT is a powerful tool that allows the recruitment of massive trials of anonymous workers screened with a variety of criteria [12]. All participants provided informed consent electronically and were compensated \$1.4 for their time ($\sim$ 5–8 min). The protocol was approved by the University IRB and was carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

For each of the three tested datasets, we make sure -after filtering- to have the results from at least $n = 240$ participants (30 per condition, 8 conditions) in order to obtain a statistical power* of 80% to detect a medium effect size with our experimental design.

*The statistical power is the probability that we correctly reject the null hypothesis (H_0) if a specific alternative hypothesis (H_1) is true

Pruning out uncooperative participants To prune out uncooperative participants, we subjected them to a 3-stage screening process. First, participants completed a short practice session to make sure they understood the task and how to use the attribution methods to infer the rules used by the model. Second, as done in [16], we asked participants to answer a few questions regarding the instructions provided to make sure they actually read and understood them. Third, during the main experiment, we took advantage of the reservoir to introduce a catch trial. The reservoir is the place where we store the training example of the current session, which can be accessed during the testing phase. We added a trial in the testing phase of each session where the input image corresponded to one of the training samples used in the current session: since the answer is still on the screen (or a scroll away) we expect participants to be correct on these catch trials. If they answered incorrectly, the participants were excluded from further analysis. Additional details regarding the selection criteria are given in the Appendix.

General study design It included 3 conditions: an experimental condition where an explanation is provided to human participants during their training phase (see Fig. 2), a baseline condition where no explanation was provided to the human participants, and a control condition where a bottom-up saliency map [27] was provided as a non-informative explanation. This last condition provides a control for the possibility that providing explanations along with training images simply increases participants' engagement in the task. As we will show in Sec. 4, such non-informative explanation actually led to a decrease in participants' ability to predict the classifier's decisions.

Each participant was only tested on a single condition to avoid possible experimental confounds (including learning confounds introduced in [23] by testing the same participants on the same images with and without explanations). The main experiment was divided into 3 training sessions (with 5 training samples in each) each followed by a brief test. In each individual trial, an image was presented with the associated prediction of the model, either alone for the baseline condition or together with an explanation for the experimental and control condition. After a brief training phase (5 samples), participants' ability to predict the classifier's output was evaluated on 7 new samples during a test phase. During the test phase, no explanation was provided to participants to assess their understanding of the classifiers while limiting confounds. One particularly problematic confound could arise from the presentation of the explanation in case the explanation leaks information about the class label.[†] We also propose to use a reservoir that subjects can refer to during the testing phase to minimize memory load as a confounding factor which was reported in [23](see Appendix for an illustration).

Controlling for prior class knowledge To control for users' own semantic knowledge, we balanced the samples shown to participants so that the classifiers were correct 50% of the time and incorrect 50% of the time. In this way, a participant who tries to simply predict the true class label of an image as opposed to learning to predict the model's outputs would only be correct 50% of the time. This allows us to control for a semantic confound in our experiments and avoid having participants receive a high score because they simply guess the real label.

4. Experiments

We performed three distinct experiments in total – using a variety of neural network architectures and 6 representative attributions methods. Each of these experiments aimed at testing the usefulness of the explanation in a different context or for a different purpose.

Setup. One of the first uses of attribution methods was for bias detection and [39] were the first to build an evaluation around the usefulness of explanation for humans to detect biases. The already existing positive results make it a good control experiment to measure the effectiveness of the framework proposed in Sec. 3.1. For this experiment, we used the same model as in the original paper: InceptionV1 [50], and a similar dataset of Husky and Wolf to bias the model. In this situation where prior knowledge of subjects can affect their score, we balance data by showing 50% of correct prediction and 50% of incorrect prediction. A subject relying only on their prior knowledge will therefore end up as a bad Meta-predictorof the model. For this experiment, the results come from 242 subjects who all passed our screening process.

As the first dataset was artificially designed, our second experiment is a real dataset with practical application proposed by [53]. This botany dataset contains over $5,000$ images divided into 19 classes. The predictor used is a VGG-16 [45]. We then selected 2 classes that could not be classified using shape in order to avoid that subjects solve the task too easily and force them to rely on non-trivial features provided by the explanations (veins, leaf margin...). Let us note that this scenario is far from being artificial and is a genuine problem for the paleobotanist [53]. The classes Betulaceae and Celastraceae fulfill this condition and have been chosen. As subjects are lay participants from Amazon Mechanical Turk we do not expect them to be experts in botany, therefore we do not control for prior knowledge. In this experiment, 240 subjects have passed all our screening processes.

Finally, for our third experiment we used ImageNet [14], a very large image dataset widely used in computer vision

[†]Imagine an attribution method that would solely encode the classifiers' prediction. Participants would be able to guess the classifier's prediction perfectly from the explanation but the explanation per se would not help participants understand how the classifiers work.

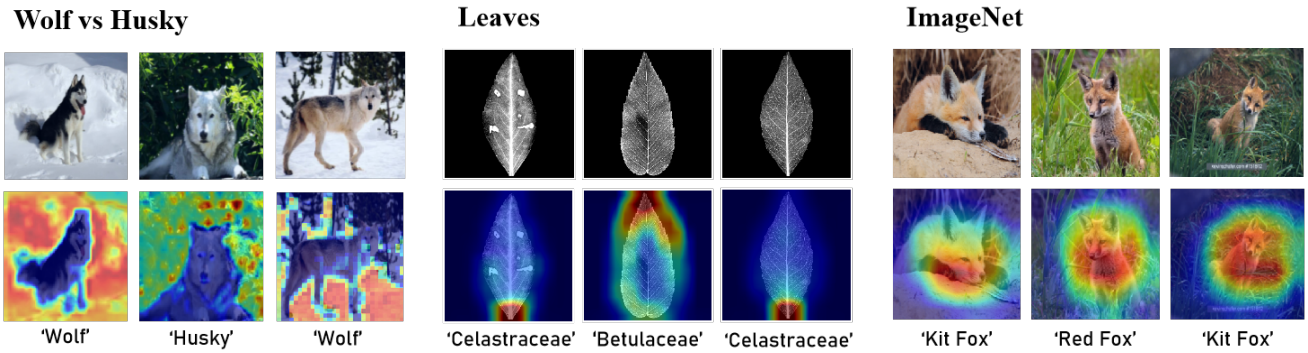

Figure 3. **Example of experiments data**. Triplets: image, explanation (Grad-CAM here), prediction shown to the participants at training time. In our Meta-predictor framework, we suggest not to show the actual label during training since the goal is to predict the model. Moreover, during the test phase (on images not seen at the training time) the explanation is not provided since we aim to measure how well a user is able to predict the model and not the ability of the explanation to leak the label.

and the field of Explainable AI is no exception [18, 20, 25, 36, 43, 51]. We use this dataset because we expect it to be representative of real-world scenarios where it is difficult to understand what the model is relying on for its decisions. Moreover, previous works have pointed out that attribution methods were not useful on this dataset [43], we have thus chosen to extend our analysis to this particular case. We use a ResNet50 [24] pretrained on this dataset as predictor. Because prior knowledge is a major confounding factor on ImageNet, we select a pair of classes that was heavily misclassified by the model, to be able to show subjects 50% of correct prediction and 50% of incorrect prediction to control for prior knowledge: the pair Kit Fox and Red Fox fits this requirement. In this experiment, the results come from 241 subjects who all passed our screening process.

For all experiments, we compared 6 representative attribution methods: Saliency (SA) [45], Gradient $\odot$ Input (GI) [3], Integrated Gradients (IG) [49], Occlusion (OC) [55], SmoothGrad (SG) [46] and Grad-CAM (GC) [41]. Further information on these methods can be found in the Appendix.

As a *faithfulness* evaluation, we used the Deletion [37] metric which is commonly used to compare attribution methods [19, 20, 29, 37, 40, 51]. A low Deletion score indicates a good *faithfulness*, and we report the *faithfulness* score as $1-$ Deletion such that a higher *faithfulness* score is better.

4.1. Sanity check for Usefulness

The *Utility* score encodes the quality of the explanations provided by a method, the higher the score, the better the method. Fig. 4 shows the *Utility*-K scores for each method after different number of training samples were used to train participants for the biased dataset of Husky vs Wolf.

A first observation is that the explanations have a positive effect on the *Utility*-K score: the explanation allows participants to better predict the model's output (as the *Utility* scores are above 1). This is confirmed with an Analysis of Variance (ANOVA) for which we found a significant main effect, with a medium effect size for our conditions ($F(7, 234) = 4.58,\ p < .001,\ \eta^2 = 0.121$). Moreover, the only score below the baseline is that of the control explanation (based on Gabor filters), which do not make use of the model. We further explore the effect of our conditions by performing pairwise comparisons using Tukey's Honestly Significant Difference [52] to compare the different explanations between themselves. The users in the control condition did not perform better than the condition without

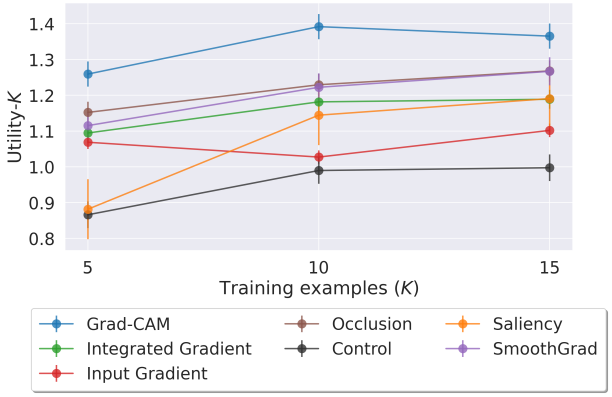

Figure 4. ***Utility*-K for Husky vs Wolf dataset.** The *Utility*-K of the explanation, or the accuracy of the human Meta-predictor after training, is measured after each training session (3 in total). All methods have a positive effect on the score obtained - they improve the subjects' ability to predict the model - and are thus useful. Moreover, the difference between the scores of the methods and the baseline implies that the humans used the explanations well. This is confirmed by the fact that the explanation based on Gabors filters (Control) misled the subjects, causing them to do worse than the baseline. Grad-CAM is the method with the most positive effect and is statistically significantly better than the other methods.

explanation.

The control condition performing worse than the baseline is an important result, as it suggests that attributions methods suspected to be more image-based than model-based [1] can be misleading. This finding also reinforces the idea that participants did make use of the explanations to try to understand the model, which led them to either get better at predicting when the explanations highlighted a behavior of the model, or misleading them when the explanations were not unrelated to the model's decision. Here Grad-CAM seems to be the best method as it is the only method significantly better than our baseline condition ($p = .002$) and the control condition ($p < .001$). Thus, participants who received the Grad-CAM explanations performed much better than those who did not receive them. The attributions maps, therefore, provided participants with essential information about the model that allowed them to train a more accurate Meta-predictor.

Finally, some attribution methods seem more useful than others. We found that Integrated Gradients and Saliency are the worst explanations as both of them are significantly less useful than Grad-CAM ($p = .01$, and $p = .02$ resp.). In between those 2 groups lies Integrated Gradients, SmoothGrad and Occlusion which do not seem to have any significant difference in their usefulness to subjects.

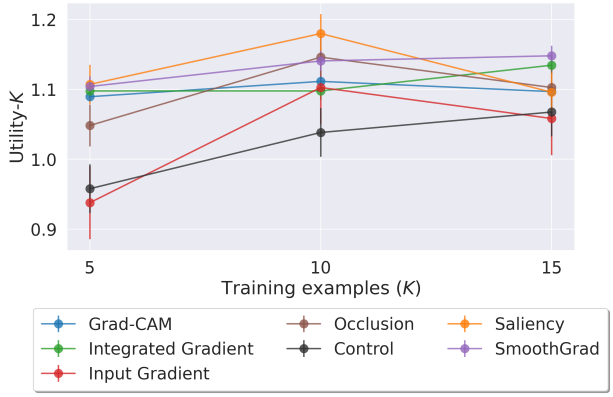

Figure 5. ***Utility-K* for Leaves dataset.** The *Utility-K* score is measured on [53] botany dataset. As in the previous series of experiments, the participants were able to leverage the explanations to learn to predict the model's predictions. All the conditions tested here were better than the baseline.

In Fig. 5, we show results on the Leaves dataset obtained with 240 subjects. The ANOVA across all conditions revealed a significant main effect, with a medium effect for our conditions ($F(7, 232) = 3.12$, $p = .004$, $\eta^2 = 0.086$). This implies that explanation also had an effect on the construction of a better Meta-predictor in this case. More precisely, according to Tukey's Honestly Significant Difference test we found that the best explanations are Saliency and SmoothGrad as they are the only ones to be significantly better than our baseline (WE) ($p = .014$ and $p = .48$ respectively). Concerning the ranking of the methods, none is significantly better than the others. However, the *Utility* score of Gradient $\odot$ Input seems to be worse than the other method and SmoothGrad slightly leads the ranking. A surprising result is that Saliency which was one of the worst explanation on the first use case, is now the best explanation on this use case.

These two experiments allow us to verify that attributions maps could allow humans to better predict and thus to better understand their model. We now wish to analyze the relationship between the methods being the most useful and the current regularly used *faithfulness* metrics.

4.2. Faithfulness metric as a proxy for Usefulness?

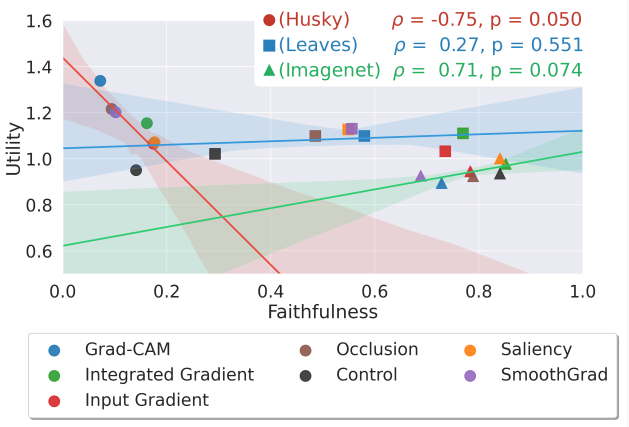

Figure 6. ***Utility* vs Faithfulness correlation.** The utility scores on the two datasets Husky vs. Wolf (point marker) and Leaves (square marker) are plotted showing a poor or anti-correlation between the two measures. Concerning the ImageNet dataset (triangle marker), the *Utility* scores are insignificant since none of the methods improves the baseline.

The results of our utility-based human evaluation show that not only are attribution methods useful in helping users understand their model, but that there are substantial differences among these methods. It would then be useful to be able to predict this ranking without resorting to human experiments which can be costly. The literature currently uses automatic metrics (which do not require humans) of *faithfulness* to compare methods between them, it seems legitimate to ask if the best method on these benchmarks is the one that helps humans the most.

On the first dataset, where Grad-CAM is the most useful method and Integrated Gradients, Saliency the worst ones, the faithfulness ranking in Tab. 1 designates Grad-CAM as the least faithful method behind Saliency which gets the best score. Surprisingly, we can observe that these results do not seem to be only independent of the utility score but negatively correlated. Thus, the *faithfulness* score does not

Method	*Husky vs Wolf*	*Leaves*	*ImageNet*
Control (Random)	0.14	0.29	0.83
Saliency [45]	**0.18**	0.55	<u>0.84</u>
Grad.-Input [7]	<u>0.17</u>	<u>0.74</u>	0.78
Integ.-Grad. [49]	0.16	**0.77**	**0.85**
SmoothGrad [46]	0.10	0.56	0.69
GradCAM [41]	0.07	0.58	0.73
Occlusion [55]	0.09	0.49	0.79

Table 1. ***Faithfulness* scores.** Average faithfullness score for each of the 3 datasets considered. Higher is better. The first and second best results are respectively **bolded** and <u>underlined</u>.

indicate how useful a method will be on the bias detection task.

On the second dataset, the best explanations are SmoothGrad and Saliency , with Gradient $\odot$ Input being the worst. From a *faithfulness* point of view, it is the opposite (see Tab. 1) since Gradient $\odot$ Input appears as one of the most faithful, behind Integrated Gradients. Moreover the most useful method: Saliency is designated as the less faithfull, just before Occlusion. Overall, all the methods tested here are more faithful than the baseline.

In Fig. 6, we summarize our results on all 3 datasets by showing that explanations are useful to the subjects on the Husky versus Wolf and Leaves datasets as the *Utility* scores are above 1. However, our scores are not correlated with the faithfulness score. Indeed, we observe a weak correlation ($\rho = 0.27$) for the Leaves experiment and a strong anti-correlation for the Husky experiment ($\rho = -0.75$) indicating that the actual *faithfulness* metrics are a poor proxy of usefulness.

4.3. When explanations are Useless

Fig. 6 shows that, on the ImageNet dataset, the set of methods tested (triangle marker) does not exceed 1, the baseline accuracy. Indeed, the experiment carried out, even with an improved experimental design, led us to the same conclusion as previous works [43]: none of the tested attribution methods are useful. The ANOVA done on the $N = 241$ subjects tells us that there is no significant main effects of our conditions ($F(7, 233) = 1.2,\ p > .05$). And as expected, the follow-up Tukey's Honestly Significant Difference test showed that none of the methods tested are useful against the baselines. Therefore no ranking based on usefulness could be established. Our previous 2 experiments indicate that, when explanations are useful to humans, their faithfullness seems to be either weakly or inversely correlated. In contrast to the *Utility* results, the *faithfulness* metric gives us a clear ranking (see Tab. 1) of the methods, indicating that once again it fails to predict the cases where none of the methods is useful.

We note that the control condition -that only needs the image to be generated- is the 3rd most faithfull explanation and is very close to the leading method. This may indicate the necessity for better baselines in the *faithfullness* evaluation to put into perspective the results.

4.4. Discussion

In summary, the proposed experimental protocol allowed us to verify two things: (1) that attributions methods can be useful -some more than others-, and (2) that the current faithfulness score of a method poorly reflects its practical usefulness to understand a model.

Regarding (1), the experiments conducted on the Husky vs Wolf dataset and on the Leaves dataset allowed us to see that our experimental design effectively captures the usefulness of explanations: humans perform better with explanation than without it. Moreover, with our control condition, we have results that highlight the practical risk of attributions methods that rely too heavily on image and not enough on model [1]. This reinforces the idea that it is critically important to evaluate the usefulness of an explanation.

The second point concerns the evaluation currently used to evaluate attribution methods. These faithfulness evaluations are poor substitutes for utility. They can therefore only be considered as sanity checks, in the strict and limited sense of the term. Since these evaluations do not give an equivalence of utility, to properly evaluate attribution methods, it is necessary to correctly put the human in the loop to avoid the previous pitfalls.

Finally, as pointed out in previous works [43], in some cases the explanation is not useful to humans. More importantly, this weakness cannot be predicted from current metrics. One of the tracks considered to explain this dysfunction is linked to the attribution methods which are limited to describe a spatial area, thus leaving to the human the care to bring their semantics, leaving a possible ambiguity. These inherent weaknesses of attribution methods suggest that other (potentially complementary) methods of explanation should be considered.

5. Conclusion

In this work, we propose a new experimental design to evaluate whether explanations are useful to humans using a Meta-predictor perspective. We also perform an online experiment on 3 datasets to obtain a ranking of the most useful explanations for humans. In parallel, we evaluate these same attribution methods according to Deletion, a widely used *faithfullness* metric. We find that the ranking of explainability methods based on their faithfullness does not correlate with their practical usefulness, which may indicate

that we are straying too far away from the original goal of explainable AI. Our results suggest a need to develop better metrics and human-centered methods.

Acknowledgments

This work was conducted as part the DEEL[‡] project. It was funded by the Artificial and Natural Intelligence Toulouse Institute (ANITI) grant #ANR19-PI3A-0004. MC was funded by ANR grant VISADEEP (ANR-20-CHIA-0022). TS was funded by ONR (N00014-19-1-2029) and NSF (IIS-1912280). The computing hardware was supported in part by NIH Office of the Director grant #S10OD025181 via the Center for Computation and Visualization.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 7, 8

[2] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. Does explainable artificial intelligence improve human decision-making? In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 3

[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 6, 18

[4] 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(8):e0181142, 2017. 2

[5] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) in ENMLP*, 2017. 2

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. 1

[7] Peyton Greenside Avanti Shrikumar and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 8, 18

[8] Felix Biessmann and Dionysius Refiano. Quality metrics for transparent machine learning with and without humans in the loop are not correlated. In *Proceedings of the ICML Workshop on Theoretical Foundations, Criticism, and Application Trends of Explainable AI held in conjunction with the 38th International Conference on Machine Learning (ICML),*, 2021. 3

[9] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019. 1

[10] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? *arXiv preprint arXiv:1810.12366*, 2018. 3

[11] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2

[12] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, 2013. 4

[13] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020. 1

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5

[15] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *ArXiv e-print*, 2017. 2, 3

[16] Julie S Downs, Mandy B Holbrook, Steve Sheng, and Lorrie Faith Cranor. Are your participants gaming the system? screening mechanical turk workers. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2399–2402, 2010. 5

[17] Andrew Elliott, Stephen Law, and Chris Russell. Explaining classifiers using adversarial perturbations on the perceptual ball. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[18] Andrew Elliott, Stephen Law, and Chris Russell. Explaining classifiers using adversarial perturbations on the perceptual ball. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10693–10702, 2021. 6

[19] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 6

[20] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 6

[21] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020. 1, 2

[‡]https://www.deel.ai/

[22] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017. 1

[23] Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)*, 2020. 2, 3, 5

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[25] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 2, 6

[26] Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. Evaluations and methods for explanation through robustness analysis. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 3

[27] L. Itti. Models of bottom-up attention and saliency. *Neurobiology of attention*, 2005. 5

[28] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)*, 2020. 4

[29] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 6

[30] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 1, 2, 3

[31] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1

[32] Zachary C. Lipton. The mythos of model interpretability. In *Workshop on Human Interpretability in Machine Learning, ICML*, 2016. 1

[33] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[34] Franck Mamalet, Eric Jenn, Gregory FLANDIN, Hervé Delseny, Christophe Gabreau, Adrien Gauffriau, Bernard Beaudouin, Ludovic Ponsolle, Lucian Alecu, Hugues Bonnin, Brice Beltran, Didier Duchel, Jean-Brice Ginestet, Alexandre Hervieu, Sylvain Pasquet, Kevin Delmas, Claire Pagetti, Jean-Marc Gabriel, Camille Chapdelaine, Sylvaine Picard, Mathieu Damour, Cyril Cappi, Laurent Gardès, Florence De Grancey, Baptiste Lefevre, Sébastien Gerchinovitz, and Alexandre Albore. White Paper Machine Learning in Certified Systems, 2021. 1

[35] 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. 2, 3

[36] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *arXiv preprint arXiv:2105.14944*, 2021. 3, 6

[37] 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 (BMVC)*, 2018. 1, 2, 6

[38] Danish Pruthi, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. Evaluating explanations: How much do explanations from the teacher aid students? In *ArXiv e-print*, 2020. 2

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016. 1, 2, 3, 5

[40] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. In *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2015. 2, 6

[41] 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 (ICCV)*, 2017. 1, 2, 3, 6, 8, 18

[42] Thomas Serre. Deep learning: The good, the bad, and the ugly. *Annual review of vision science*, 2019. 1

[43] Hua Shen and Ting-Hao Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172, 2020. 2, 3, 6, 8

[44] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 18

[45] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 2, 5, 6, 8, 18

[46] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1, 6, 8, 18

[47] Matthew Sotoudeh and Aditya V. Thakur. Computing linear restrictions of neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 18

[48] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 2

[49] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1, 2, 6, 8, 18

[50] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[51] Fel Thomas, Cadene Remi, Chalvidal Mathieu, Cord Matthieu, Vigouroux David, and Serre Thomas. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 6

[52] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 1949. 6

[53] Peter Wilf, Shengping Zhang, Sharat Chikkerur, Stefan A Little, Scott L Wing, and Thomas Serre. Computer vision cracks the leaf code. *Proceedings of the National Academy of Sciences*, 113(12):3305–3310, 2016. 5, 7

[54] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Explainability of vision-based autonomous driving systems: Review and challenges. *arXiv preprint arXiv:2101.05307*, 2021. 1

[55] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014. 2, 6, 8, 18

[56] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 2

A. Human experiments

A.1. Experimental design

Figure 7 summarizes the experimental design used for our experiments. The participants that went through our experiments are users from the online platform Amazon Mechanical Turk (AMT). We prioritized users with a Master qualification (which is a qualification attributed by AMT to users who have proven to be of excellent quality) or normal users with high qualifications (number of HIT completed $= 10000$ and HIT accepted $> 98\%$) .

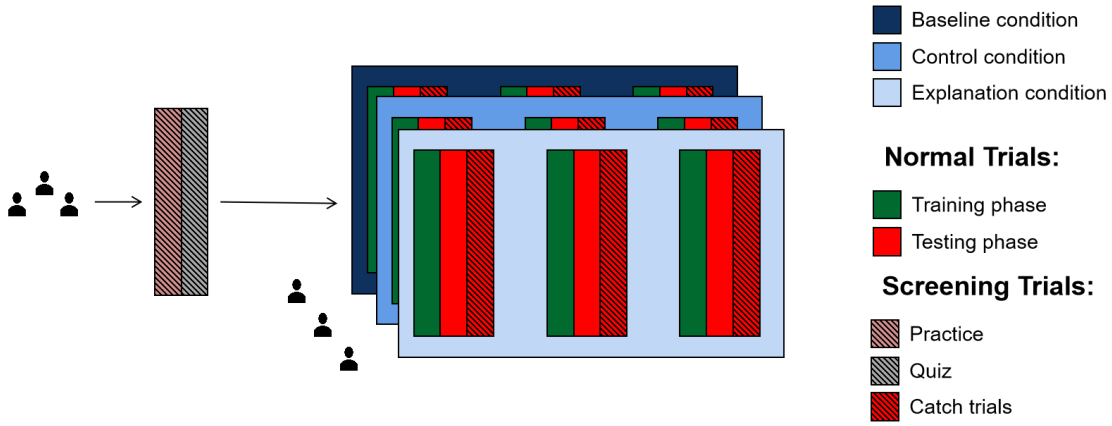

Figure 7. **Experimental design.** First, every participant goes through a practice session (fig 8) to make sure they understand how to use attribution methods to infer the rules used by a model, and a quiz (fig 9) to make sure they actually read and understand the instructions. Then, participants are split into the different conditions – every participant will only go through one condition. The 3 possible conditions are: an Explanation condition where an explanation is provided to human participants during their training phase, a Baseline condition where no explanation was provided to the human participants, and a Control condition where a non-informative explanation was provided. The main experiment was divided into 3 training sessions each followed by a brief test. In each individual training trial, an image was presented with the associated prediction of the model, either alone for the baseline condition or together with an explanation for the experimental and control condition. After a brief training phase (5 samples), participants' ability to predict the classifier's output was evaluated on 7 new samples (only the image, no explanation) during a test phase. To filter out uncooperative participants we also add a catch trial (fig 10) in each test session.

A.2. 3-stage screening proccess

In this section, we show screens of our experiments, for each screening process: the practice session (fig 8), the quiz (fig 9), and the catch trials (fig 10). Participants that failed any of the 3 screening process were excluded from further analysis.

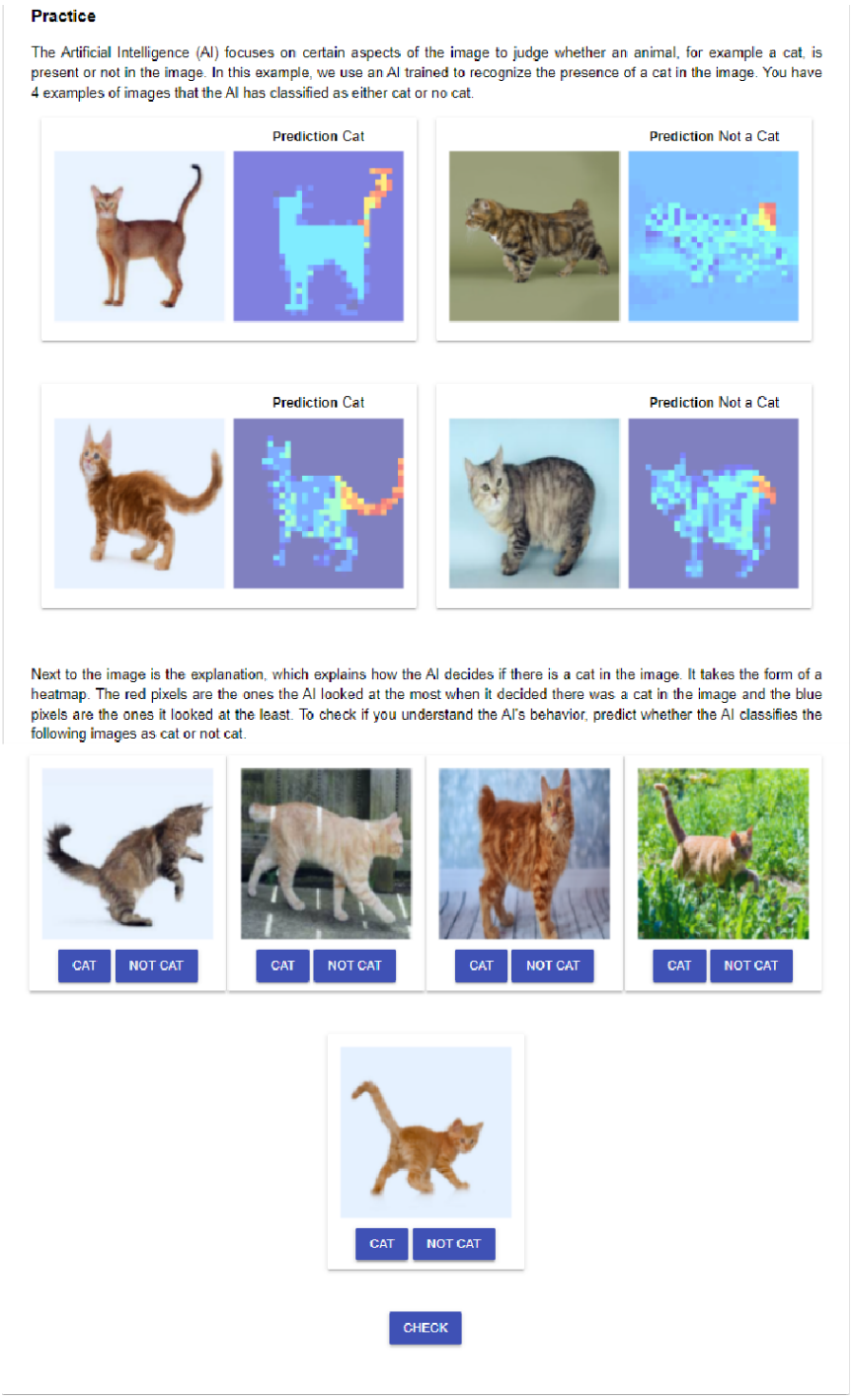

Figure 8. **Practice session.** Through a practice session, which is a simplified version of the main experiment, we evaluate if users understand how to read and use explanations. Participants that failed to predict correctly any of the 5 cat test images on the first try were excluded from further analysis.

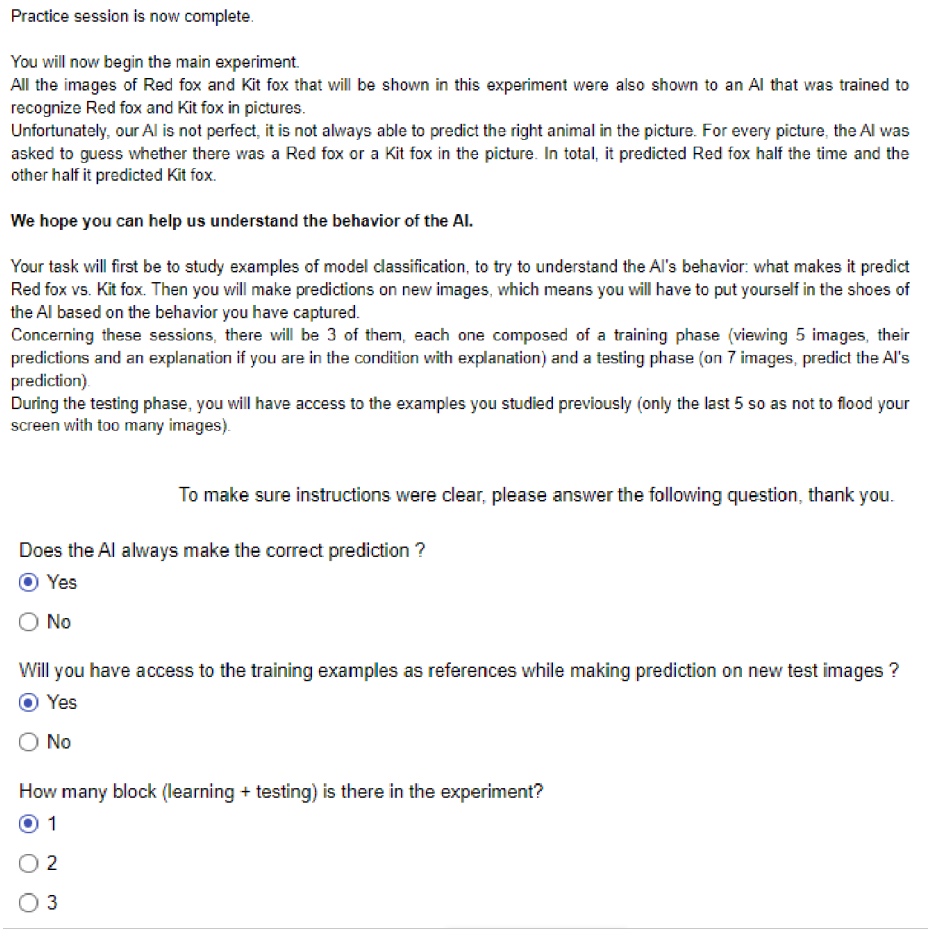

Figure 9. **Quiz.** Through a quiz, we make sure that users read and understood the instructions. Participants that did not answer correctly every question on the first try were excluded from further analysis.

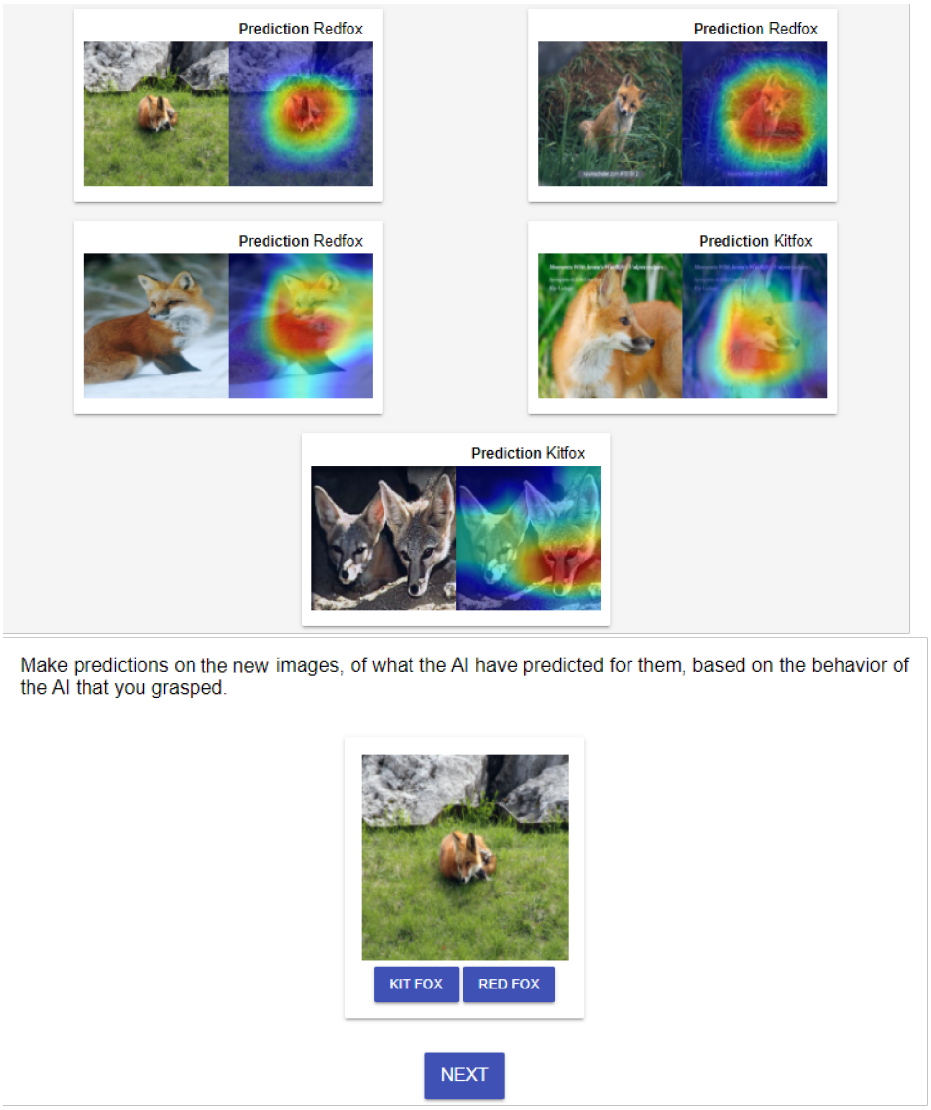

Figure 10. **Catch trial.** We use a reservoir (to store all the examples of the current training session) that participants can refer to during the testing phase to minimize memory load. At the top of the screen is the reservoir, at the bottom of the screen is a trial from the testing phase. We take advantage of the reservoir to introduce a catch trial. We added a trial in the testing phase of each session where the input image corresponded to one of the training samples used in the current session: since the answer is still on the screen (or a scroll away) we expect participants to be correct on these catch trials. Participants that failed any of the 3 catch trials (one per session) were excluded from further analysis.

B. Attribution methods

B.1. Methods

In the following section, the formulation of the different methods used in the experiment is given. We define $\boldsymbol{f}(\boldsymbol{x})$ the logit score (before softmax) for the class of interest. An explanation method provides an attribution score for each input variables. Each value then corresponds to the importance of this feature for the model results.

Saliency [45] is a visualization technique based on the gradient of a class score relative to the input, indicating in an infinitesimal neighborhood, which pixels must be modified to most affect the score of the class of interest.

$$\boldsymbol{\Phi}^{SA}(\boldsymbol{x}) = ||\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})||$$

Gradient $\odot$ Input [7] is based on the gradient of a class score relative to the input, element-wise with the input, it was introduced to improve the sharpness of the attribution maps. A theoretical analysis conducted by [3] showed that Gradient $\odot$ Input is equivalent to ϵ-LRP and DeepLIFT [44] methods under certain conditions: using a baseline of zero, and with all biases to zero.

$$\boldsymbol{\Phi}^{GI}(\boldsymbol{x}) = \boldsymbol{x} \odot ||\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})||$$

Integrated Gradients [49] consists of summing the gradient values along the path from a baseline state to the current value. The baseline is defined by the user and often chosen to be zero. This integral can be approximated with a set of m points at regular intervals between the baseline and the point of interest. In order to approximate from a finite number of steps, we use a Trapezoidal rule and not a left-Riemann summation, which allows for more accurate results and improved performance (see [47] for a comparison). The final result depends on both the choice of the baseline $\mathbf{x}_0$ and the number of points to estimate the integral. In the context of these experiments, we use zero as the baseline and $m = 80$.

$$\boldsymbol{\Phi}^{IG}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}_0) \int_0^1 \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}_0 + \alpha(\boldsymbol{x} - \boldsymbol{x}_0)))d\alpha$$

SmoothGrad [46] is also a gradient-based explanation method, which, as the name suggests, averages the gradient at several points corresponding to small perturbations (drawn i.i.d from a normal distribution of standard deviation σ) around the point of interest. The smoothing effect induced by the average helps reduce the visual noise and hence improve the explanations. In practice, Smoothgrad is obtained by averaging after sampling m points. In the context of these experiments, we took $m = 80$ and $\sigma = 0.2$ as suggested in the original paper.

$$\boldsymbol{\Phi}^{SG}(\boldsymbol{x}) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \boldsymbol{I}\sigma)}(\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x} + \varepsilon))$$

Grad-CAM [41] can be used on Convolutional Neural Network (CNN), it uses the gradient and the feature maps $\mathbf{A}^k$ of the last convolution layer. More precisely, to obtain the localization map for a class, we need to compute the weights α_c^k associated to each of the feature map activation $\mathbf{A}^k$, with k the number of filters and Z the number of features in each feature map, with $\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \mathbf{A}_{ij}^k}$ and

$$\boldsymbol{\Phi}^{GC} = max(0, \sum_k \alpha_k^c \mathbf{A}^k)$$

Notice that the size of the explanation depends on the size (width, height) of the last feature map, a bilinear interpolation is performed in order to find the same dimensions as the input.

Occlusion [55] is a sensitivity method that sweeps a patch that occludes pixels over the images, and uses the variations of the model prediction to deduce critical areas. In the context of these experiments, we took a patch size and a patch stride of of 1 tenth of the image size.

$$\boldsymbol{\Phi}_i^{OC} = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{[\boldsymbol{x}_i = 0]})$$

B.2. Examples of explanations

Examples of explanations from the different attributions methods evaluated through our experiments on the Husky vs Wolf dataset (fig 11), the Leaves dataset (fig 12) and the ImageNet dataset (fig 13).

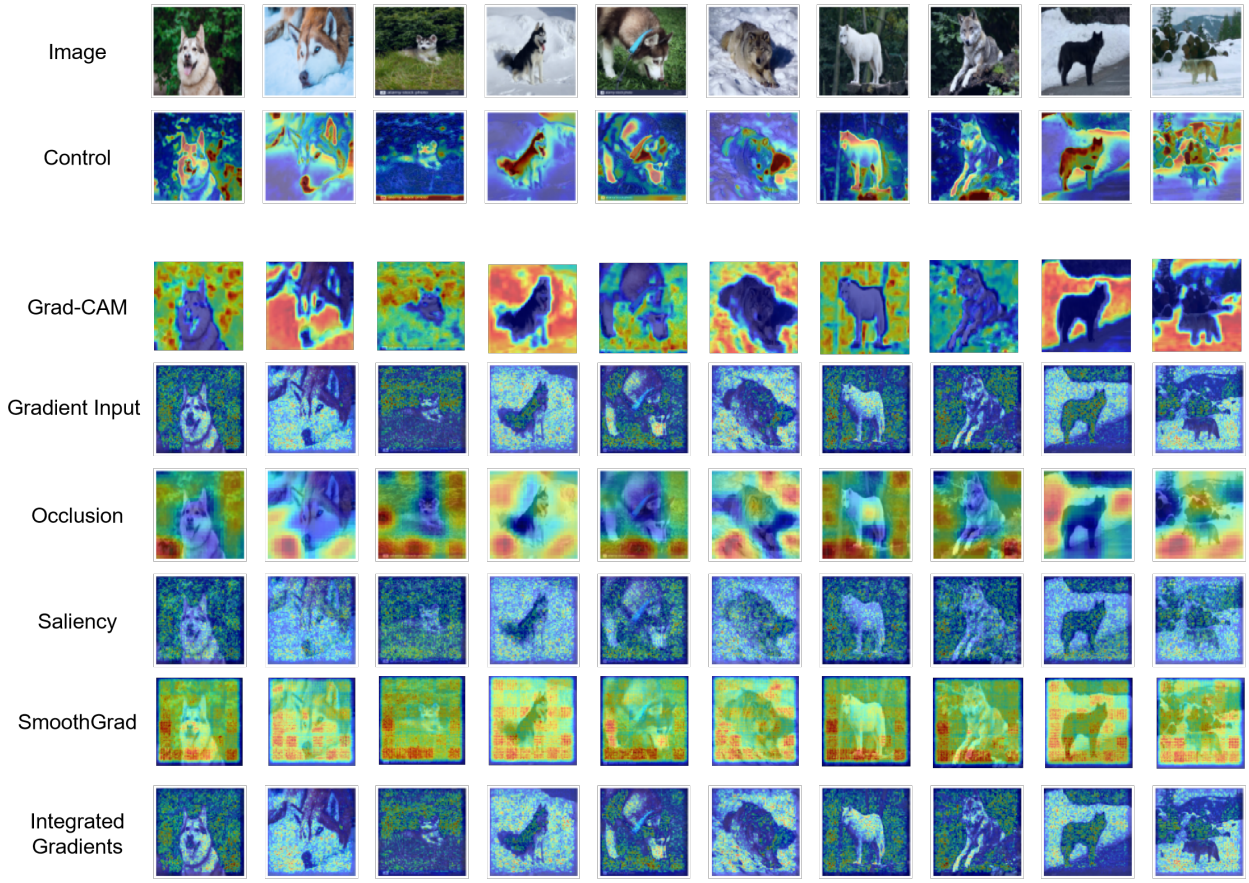

Figure 11. Examples of images from the Wolf vs Husky experiment, alongside their respective: Control explanation (which is a non-informative explanation) as well as the different Attribution methods evaluated in our experiment.

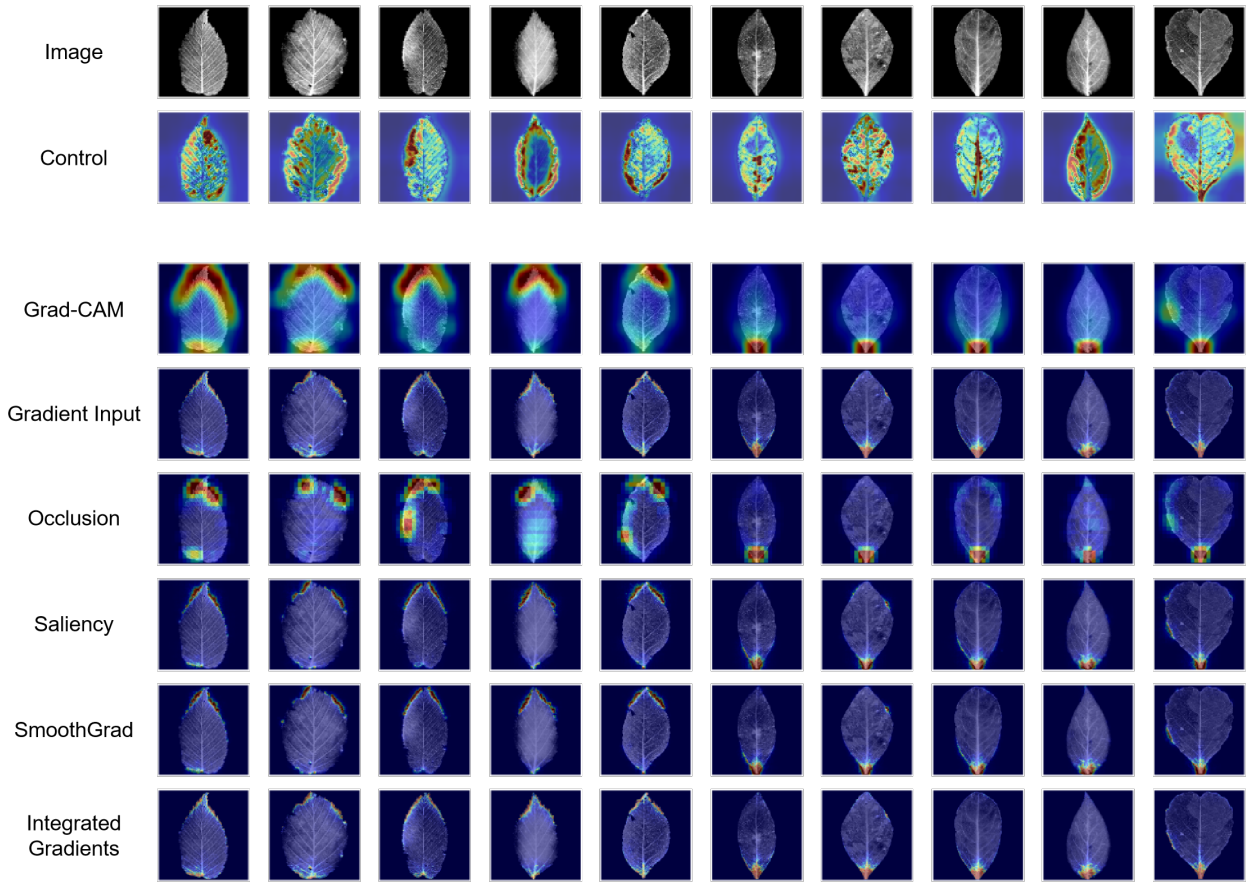

Figure 12. Examples of images from the Leaves experiment, alongside their respective: Control explanation (which is a non-informative explanation) as well as the different Attribution methods evaluated in our experiment.

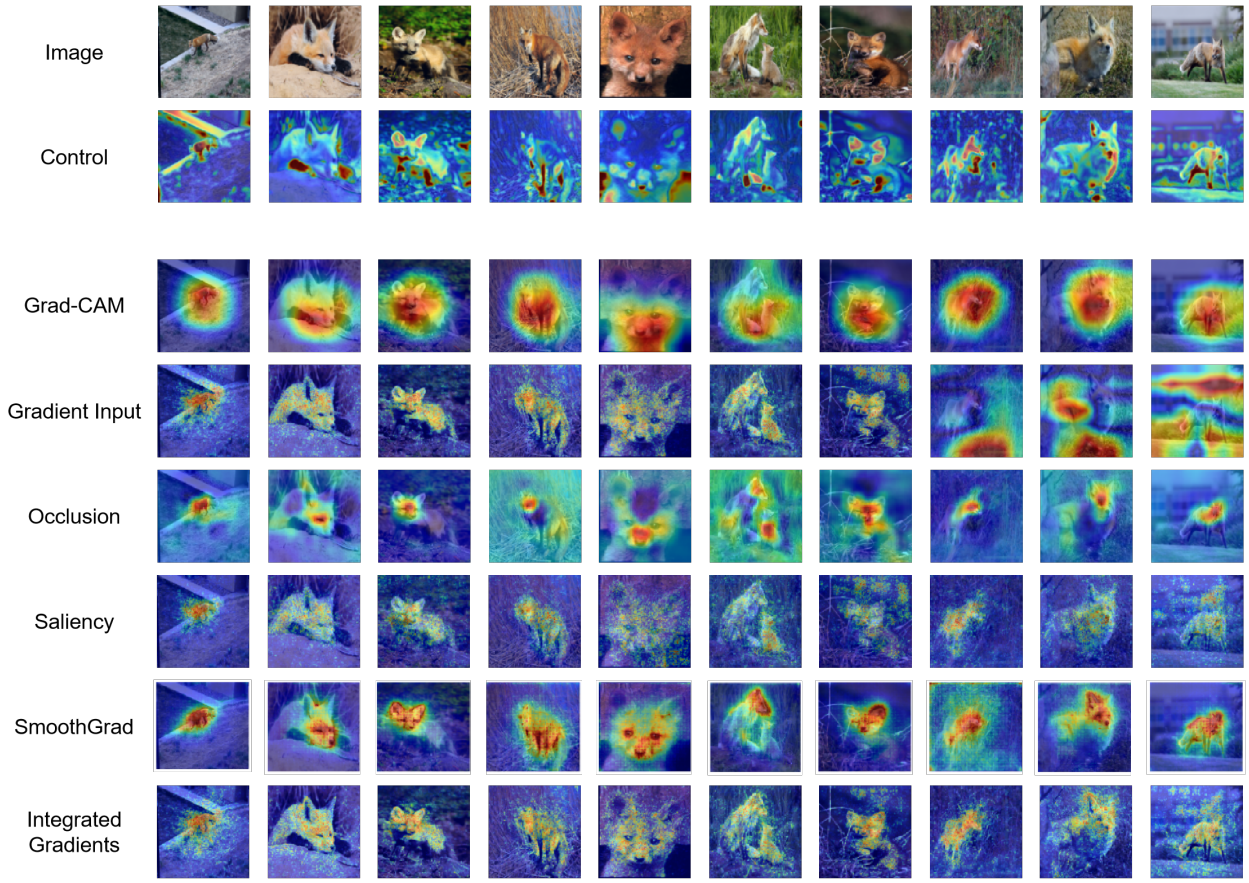

Figure 13. Examples of images from the ImageNet experiment, alongside their respective: Control explanation (which is a non-informative explanation) as well as the different Attribution methods evaluated in our experiment.