CRAFT: Concept Recursive Activation FacTorization for Explainability

Thomas Fel[1,3,5*] **Agustin Picard**[3,6*] **Louis Bethune**[3*] **Thibaut Boissin**[3,4*]
David Vigouroux[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] Institut de Recherche Technologique Saint-Exupery, France
[5] Innovation & Research Division, SNCF , [6] Scalian

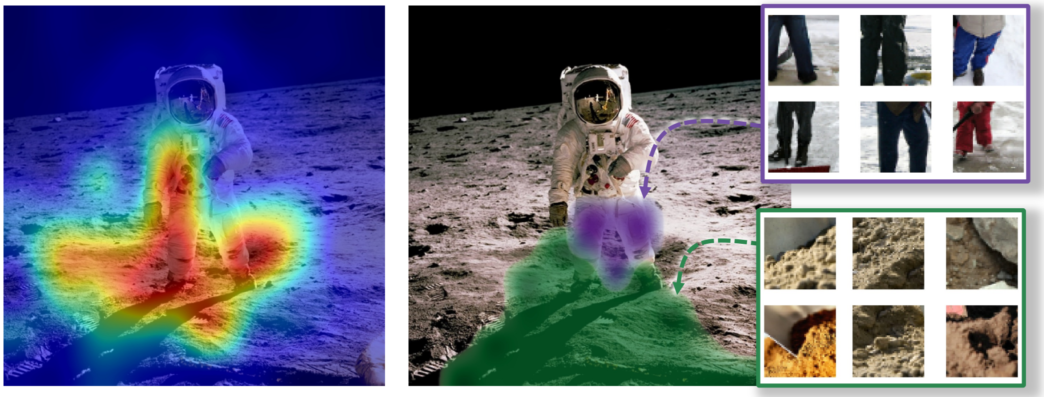

Figure 1. **The "Man on the Moon" incorrectly classified as a "shovel" by an ImageNet-trained ResNet50.** Heatmap generated by a classic attribution method [55] (left) vs. *concept attribution maps* generated with the proposed CRAFT approach (right) which highlights the two most influential concepts that drove the ResNet50's decision along with their corresponding locations. CRAFT suggests that the neural net arrived at its decision because it identified the concept of "dirt" ● commonly found in members of the image class "shovel" and the concept of "ski pants" ● typically worn by people clearing snow from their driveway with a shovel instead the correct concept of astronaut's pants (which was probably never seen during training).

Abstract

*Attribution methods, which employ heatmaps to identify the most influential regions of an image that impact model decisions, have gained widespread popularity as a type of explainability method. However, recent research has exposed the limited practical value of these methods, attributed in part to their narrow focus on the most prominent regions of an image – revealing "where" the model looks, but failing to elucidate "what" the model sees in those areas. In this work, we try to fill in this gap with CRAFT – a novel approach to identify both "what" and "where" by generating concept-based explanations. We introduce 3 new ingredients to the automatic concept extraction literature: (**i**) a recursive strategy to detect and decompose concepts across layers, (**ii**) a novel method for a more faithful estimation of concept importance using Sobol indices, and (**iii**) the use of implicit differentiation to unlock Concept Attribution Maps.*

We conduct both human and computer vision experiments to demonstrate the benefits of the proposed approach. We show that the proposed concept importance estimation technique is more faithful to the model than previous methods. When evaluating the usefulness of the method for human experimenters on a human-centered utility benchmark, we find that our approach significantly improves on two of the three test scenarios.

1. Introduction

Interpreting the decisions of modern machine learning models such as neural networks remains a major challenge. Given the ever-increasing range of machine learning applications, the need for robust and reliable explainability methods continues to grow [11, 35]. Recently enacted European laws (including the General Data Protection Regulation (GDPR) [37] and the European AI act [43]) require the assessment of explainable decisions, especially those made by algorithms.

In order to try to meet this growing need, an array of explainability methods have already been proposed [14, 50,

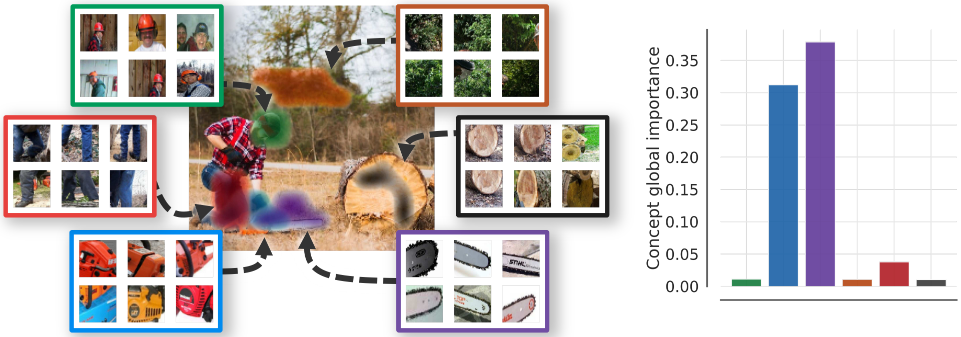

Figure 2. **CRAFT results for the prediction "chain saw".** First, our method uses Non-Negative Matrix Factorization (NMF) to extract the most relevant concepts used by the network (ResNet50V2) from the train set (ILSVRC2012 [10]). The global influence of these concepts on the predictions is then measured using Sobol indices (right panel). Finally, the method provides local explanations through *concept attribution maps* (heatmaps associated with a concept, and computed using grad-CAM by backpropagating through the NMF concept values with implicit differentiation). Besides, concepts can be interpreted by looking at crops that maximize the NMF coefficients. For the class "chain saw", the detected concepts seem to be: ● the chainsaw engine, ● the saw blade, ● the human head, ● the vegetation, ● the jeans and ● the tree trunk.

59, 61, 62, 66, 69, 71, 76]. One of the main classes of methods called attribution methods yields heatmaps that indicate the importance of individual pixels for driving a model's decision. However, these methods exhibit critical limitations [1, 27, 63, 65], as they have been shown to fail – or only marginally help – in recent human-centered benchmarks [8, 26, 41, 49, 60, 64]. It has been suggested that their limitations stem from the fact that they are only capable of explaining *where* in an image are the pixels that are critical to the decision but they cannot tell *what* visual features are actually driving decisions at these locations. In other words, they show where the model looks but not what it sees. For example, in the scenario depicted in Fig. 1, where an ImageNet-trained ResNet mistakenly identifies an image as containing a shovel, the attribution map displayed on the left fails to explain the reasoning behind this misclassification.

A recent approach has sought to move past attribution methods [40] by using so-called "concepts" to communicate information to users on how a model works. The goal is to find human-interpretable concepts in the activation space of a neural network. Although the approach exhibited potential, its practicality is significantly restricted due to the need for prior knowledge of pertinent concepts in its original formulation and, more critically, the requirement for a labeled dataset of such concepts. Several lines of work have focused on trying to automate the concept discovery process based only on the training dataset and without explicit human supervision. The most prominent of these techniques, ACE [24], uses a combination of segmentation and clustering techniques but requires heuristics to remove outliers. However, ACE provides a proof of concept that it might be possible to discover concepts automatically and at scale – without additional labeling or human supervision. Nevertheless, the approach suffers several limitations: by construction, each image segment can only belong to a single cluster, a layer has to be selected by the user to be used to retrieve the relevant concepts, and the amount of information lost during the outlier rejection phase can be a cause of concern. More recently, Zhang et al. [77] proposes to leverage matrix decompositions on internal feature maps to discover concepts.

Here, we try to fill these gaps with a novel method called CRAFT which uses Non-Negative Matrix Factorization (NMF) [46] for concept discovery. In contrast to other concept-based explanation methods, our approach provides an explicit link between their global and local explanations (Fig. 2) and identifies the relevant layer(s) to use to represent individual concepts (Fig. 3). Our main contributions can be described as follows:

(i) A novel approach for the automated extraction of high-level concepts learned by deep neural networks. We validate its practical utility to users with human psychophysics experiments.

(ii) A recursive procedure to automatically identify concepts and sub-concepts at the right level of granularity – starting with our decomposition at the top of the model and working our way upstream. We validate the benefit of this approach with human psychophysics experiments showing that (i) the decomposition of a concept yields more coherent sub-concepts and (ii) that the groups of points formed by these sub-concepts are more refined and appear meaningful to humans.

(iii) A novel technique to quantify the importance of individual concepts for a model's prediction using Sobol indices [34, 67, 68] – a technique borrowed from Sensitivity Analysis.

(iv) The first concept-based explainability method which produces *concept attribution maps* by backpropagating concept scores into the pixel space by leveraging the implicit function theorem in order to localize the pixels associated with the concept of a given input image. This effectively opens up the toolbox of both white-box [15, 59, 62, 66, 69, 71,78] and black-box [14,47,55,57] explainability methods to derive concept-wise attribution maps.

2. Related Work

Attribution methods Attribution methods are widely used as post-hoc explainability techniques to determine the input variables that contribute to a model's prediction by generating importance maps, such as the ones shown in Fig.1. The first attribution method, Saliency, introduced in [62], generates a heatmap by utilizing the gradient of a given classification score with respect to the pixels. This method was later improved upon in the context of deep convolutional networks for classification in subsequent studies, such as [66, 69, 71, 76]. However, the image gradient only reflects the model's operation within an infinitesimal neighborhood around an input, which can yield misleading importance estimates [22] since gradients of large vision models are notoriously noisy [66]. That is why several methods leverage perturbations on the input image to probe the model and create importance maps that indicate the most crucial areas for the decision, such as Rise [55], Sobol [14], or more recently HSIC [50].

Unfortunately, a severe limitation of these approaches – apart from the fact that they only show the "*where*" – is that they are subject to confirmation bias: while they may appear to offer useful explanations to a user, sometimes these explanations are actually incorrect [1, 23, 65]. These limitations raise questions about their usefulness, as recent research has shown by using human-centered experiments to evaluate the utility of attribution [8, 26, 41, 49, 60].

In particular, in [8], a protocol is proposed to measure the usefulness of explanations, corresponding to how much they help users identify rules driving a model's predictions (correct or incorrect) that transfer to unseen data – using the concept of meta-predictor (also called simulatability) [11, 18, 39]. The main idea is to train users to predict the output of the system using a small set of images along with associated model predictions and corresponding explanations. A method that performs well on this this benchmark is said useful, as it help users better predict the output of the model by providing meaningful information about the internal functioning of the model. This framework being agnostic to the type of explainability method, we have chosen to use it in Section 4 in order to compare CRAFT with attribution methods.

Concepts-based methods Kim et al. [40] introduced a method aimed at providing explanations that go beyond attribution-based approaches by measuring the impact of pre-selected concepts on a model's outputs. Although this method appears more interpretable to human users than standard attribution techniques, it requires a database of images describing the relevant concepts to be manually curated. Ghorbani et al. [24] further extended the approach to extract concepts without the need for human supervision. The approach, called ACE [24], uses a segmentation scheme on images, that belong to an image class of interest. The authors leveraged the intermediate activations of a neural network for specific image segments. These segments were resized to the appropriate input size and filled with a baseline value. The resulting activations were clustered to produce prototypes, which they referred to as "concepts". However, some concepts contained background segments, leading to the inclusion of uninteresting and outlier concepts. To address this, the authors implemented a post-processing cleanup step to remove these concepts, including those that were present in only one image of the class and were not representative. While this improved the interpretability of their explanations to human subjects, the use of a baseline value filled around the segments could introduce biases in the explanations [28, 31, 42, 70].

Zhang et al. [77] developed a solution to the unsupervised concept discovery problem by using matrix factorizations in the latent spaces of neural networks. However, one major drawback of this method is that it operates at the level of convolutional kernels, leading to the discovery of localized concepts. For example, the concept of "grass" at the bottom of the image is considered to be distinct from the concept of "grass" at the top of the image.

3. Overview of the method

In this section, we first describe our concept activations factorization method. Below we highlight the main differences with related work. We then proceed to introduce the three novel ingredients that make up CRAFT: (1) a method to recursively decompose concepts into sub-concepts, (2) a method to better estimate the importance of extracted concepts, and (3) a method to use any attribution method to create *concept attribution maps*, using implicit differentiation [4, 25, 44].

Notations In this work, we consider a general supervised learning setting, where $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \in \mathcal{X}^n \subseteq \mathbb{R}^{n \times d}$ are n inputs images and $(y_1, ..., y_n) \in \mathcal{Y}^n$ their associated labels. We are given a (machine-learnt) black-box predictor $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$, which at some test input $\boldsymbol{x}$ predicts the output $\boldsymbol{f}(\boldsymbol{x})$. Without loss of generality, we establish that $\boldsymbol{f}$ is a neural network that can be decomposed into two distinct components. The first component is a function $\boldsymbol{g}$ that maps the input to an intermediate state, and the second component is $\boldsymbol{h}$, which takes this intermediate state to the output, such that $\boldsymbol{f}(\boldsymbol{x}) = (\boldsymbol{h} \circ \boldsymbol{g})(\boldsymbol{x})$. In this context, $\boldsymbol{g}(\boldsymbol{x}) \subseteq \mathbb{R}^p$ represents the intermediate activations of $\boldsymbol{x}$ within the network.

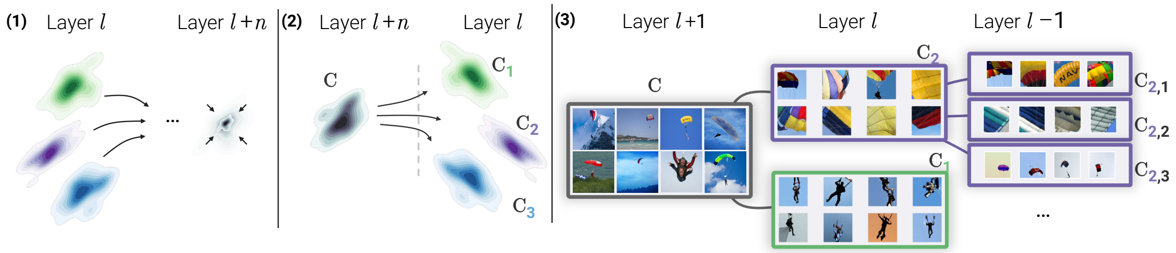

Figure 3. **(1) Neural collapse (amalgamation).** A classifier needs to be able to linearly separate classes by the final layer. It is commonly assumed that in order to achieve this, image activations from the same class get progressively "merged" such that these image activations converge to a one-hot vector associated with the class at the level of the logits layer. In practice, this means that different concepts get ultimately blended together along the way. **(2) Recursive process.** When a concept is not understood (e.g., $\mathcal{C}$), we propose to decompose it into multiple sub-concepts (e.g., $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$) using the activations from an earlier layer to overcome the aforementioned neural collapse issue. **(3) Example of recursive concept decomposition** using CRAFT on the ImageNet class "parachute".

Further, we will assume non-negative activations: $\boldsymbol{g}(\boldsymbol{x}) \geq 0$. In particular, this assumption is verified by any architecture that utilizes *ReLU*, but any non-negative activation function works.

3.1. Concept activation factorization.

We use Non-negative matrix factorization to identify a basis for concepts based on a network's activations (Fig.4). Inspired by the approach taken in ACE [24], we will use image sub-regions to try to identify coherent concepts.

The first step involves gathering a set of images that one wishes to explain, such as the dataset, in order to generate associated concepts. In our examples, to explain a specific class $y \in \mathcal{Y}$, we selected the set of points $\mathcal{C}$ from the dataset for which the model's predictions matched a specific class $\mathcal{C} = \{\boldsymbol{x}_i : \boldsymbol{f}(\boldsymbol{x}_i) = y, 1 \leq i \leq n\}$. It is important to emphasize that this choice is significant. The goal is not to understand how humans labeled the data, but rather to comprehend the model itself. By only selecting correctly classified images, important biases and failure cases may be missed, preventing a complete understanding of our model.

Now that we have defined our set of images, we will proceed with selecting sub-regions of those images to identify specific concepts within a localized context. It has been observed that the implementation of segmentation masks suggested in ACE can lead to the introduction of artifacts due to the associated inpainting with a baseline value. In contrast, our proposed method takes advantage of the prevalent use of modern data augmentation techniques such as randaugment, mixup, and cutmix during the training of current models. These techniques involve the current practice of models being trained on image crops, which enables us to leverage a straightforward crop and resize function denoted by $\boldsymbol{\pi}(\cdot)$ to create sub-regions (illustrated in Fig.4). By applying $\boldsymbol{\pi}$ function to each image in the set $\mathcal{C}$, we obtain an auxiliary dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ such that each entries $\mathbf{X}_i = \boldsymbol{\pi}(\boldsymbol{x}_i)$ is an image crop.

To discover the concept basis, we start by obtaining the activations for the random crops $\mathbf{A} = \boldsymbol{g}(\mathbf{X}) \in \mathbb{R}^{n \times p}$. In the case where $\boldsymbol{f}$ is a convolutional neural network, a global average pooling is applied to the activations.

We are now ready to apply Non-negative Matrix Factorization (NMF) to decompose positive activations $\mathbf{A}$ into a product of non-negative, low-rank matrices $\mathbf{U} \in \mathbb{R}^{n \times r}$ and $\mathbf{W} \in \mathbb{R}^{p \times r}$ by solving:

$$(\mathbf{U}, \mathbf{W}) = \underset{\mathbf{U} \geq 0, \mathbf{W} \geq 0}{\arg\min} \ \frac{1}{2} \|\mathbf{A} - \mathbf{U}\mathbf{W}^{\mathsf{T}}\|^2_F, \qquad (1)$$

where $|| \cdot ||_F$ denotes the Frobenius norm.

This decomposition of our activations $\mathbf{A}$ yields two matrices: $\mathbf{W}$ containing our Concept Activation Vectors (CAVs) and $\mathbf{U}$ that redefines the data points in our dataset according to this new basis. Moreover, this decomposition in this new basis has some interesting properties that go beyond the simple low-rank factorization – since $r \ll \min(n, p)$. First, NMF can be understood as the joint learning of a dictionary of Concept Activation Vectors – called a "concept bank" in Fig. 4 – that maps a $\mathbb{R}^p$ basis onto $\mathbb{R}^r$, and $\mathbf{U}$ the coefficients of the vectors $\mathbf{A}$ expressed in this new basis. The minimization of the reconstruction error $\frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}\|^2_F$ ensures that the new basis contains (mostly) relevant concepts. Intuitively, the non-negativity constraints $\mathbf{U} \geq 0, \mathbf{W} \geq 0$ encourage ***(i)*** $\mathbf{W}$ to be sparse (useful for creating disentangled concepts), ***(ii)*** $\mathbf{U}$ to be sparse (convenient for selecting a minimal set of useful concepts) and ***(iii)*** missing data to be imputed [56], which corresponds to the sparsity pattern of *post-ReLU* activations $\mathbf{A}$.

It is worth noting that each input $\boldsymbol{x}_i$ can be expressed as a linear combination of concepts denoted as $\mathbf{A}_i = \sum_{j=1}^{r} \mathbf{U}_{(i,j)} \mathbf{W}_j^{\mathsf{T}}$. This approach is advantageous because it allows us to interpret each input as a composition of the underlying concepts. Furthermore, the strict positivity of each term – NMF is working over the anti-negative semiring, – enhances the interpretability of the decomposition. Another interesting interpretation could be that each input is represented as a superposition of concepts [13].

While other methods in the literature solve a similar problem (such as low-rank factorization using SVD or ICA), the NMF is both fast and effective and is known to yield concepts that are meaningful to humans [20,74,77]. Finally, once the concept bank $\mathbf{W}$ has been precomputed, we can associate the concept coefficients $\boldsymbol{u}$ to any new input $\boldsymbol{x}$ (e.g., a full image) by solving the underlying Non-Negative Least Squares (NNLS) problem $\min_{\boldsymbol{u} \geq 0} \frac{1}{2} \|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{u}\mathbf{W}^\intercal\|_F^2$, and therefore recover its decomposition in the concept basis.

In essence, the core of our method can be summarized as follows: using a set of images, the idea is to re-interpret their embedding at a given layer as a composition of concepts that humans can easily understand. In the next section, we show how one can recursively apply concept activation factorizations to preceding layer for an image containing a previously computed concept.

3.2. Ingredient 1: A pinch of recursivity

One of the most apparent issues in previous work [24,77] is the need for choosing a priori a layer at which the activation maps are computed. This choice will critically affect the concepts that are identified because certain concepts get amalgamated [53] into one at different layers of the neural network, resulting in incoherent and indecipherable clusters, as illustrated in Fig. 3. We posit that this can be solved by iteratively applying our decomposition at different layer depths, and for the concepts that remain difficult to understand, by looking for their sub-concepts in earlier layers by isolating the images that contain them. This allows us to build hierarchies of concepts for each class.

We offer a simple solution consisting of reapplying our method to a concept by performing a second step of concept activation factorization on a set of images that contain the concept $\mathcal{C}$ in order to refine it and create sub-concepts (e.g., decompose $\mathcal{C}$ into $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$) see Fig. 3 for an illustrative example. Note that we generalize current methods in the sense that taking images $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ that are clustered in the logits layer (belonging to the same class) and decomposing them in a previous layer – as done in [24,77] – is a valid recursive step. For a more general case, let us assume that a set of images that contain a common concept is obtained using the first step of concept activation factorization.

We will then take a subset of the auxiliary dataset points to refine any concept j. To do this, we select the subset of points that contain the concept $\mathcal{C}_j = \{\mathbf{X}_i : \mathbf{U}_{(i,j)} > \lambda_j, 1 \leq i \leq n\}$, where λ_j is the 90th percentile of the values of the concept $\mathbf{U}_{(1,j)}, \ldots, \mathbf{U}_{(n,j)}$. In other words, the 10% of images that activate the concept j the most are selected for further refinement into sub-concepts. Given this new set of points, we can then re-apply the Concept Matrix Factorization method to an earlier layer $\boldsymbol{g}'(\cdot)$ to obtain the sub-concepts decomposition from the initial concept – as illustrated in Fig.3.

3.3. Ingredient 2: A dash of sensitivity analysis

A major concern with concept extraction methods is that concepts that makes sense to humans are not necessarily the same as those being used by a model to classify images. In order to prevent such confirmation bias during our concept analysis phase, a faithful estimate the overall importance of the extracted concepts is crucial. Kim et al. [40] proposed an importance estimator based on directional derivatives: the partial derivative of the model output with respect to the vector of concepts. While this measure is theoretically grounded, it relies on the same principle as gradient-based methods, and thus, suffers from the same pitfalls: neural network models have noisy gradients [66, 71]. Hence, the farther the chosen layer is from the output, the noisier the directional derivative score will be.

Since we essentially want to know which concept has the greatest effect on the output of the model, it is natural to consider the field of sensitivity analysis [9,33,34,67,68]. In this section, we briefly recall the classic "total Sobol indices" and how to apply them to our problem. The complete derivation of the Sobol-Hoeffding decomposition is presented in Section D of the supplementary materials. Formally, a natural way to estimate the importance of a concept i is to measure the fluctuations of the model's output $\boldsymbol{h}(\mathbf{U}\mathbf{W}^\intercal)$ in response to meaningful perturbations of the concept coefficient $\mathbf{U}_{(1,i)}, \ldots, \mathbf{U}_{(n,i)}$. Concretely, we will use perturbation masks $\mathbf{M} = (M_1, ..., M_r) \sim \mathcal{U}([0,1]^r)$, here an i.i.d sequence of real-valued random variables, we introduce a concept fluctuation to reconstruct a perturbed activation $\tilde{\mathbf{A}} = (\mathbf{U} \odot \mathbf{M})\mathbf{W}^\intercal$ where $\odot$ denote the Hadamard product (e.g., the masks can be used to remove a concept by setting its value to zero). We can then propagate this perturbed activation to the model output $\mathbf{Y} = \boldsymbol{h}(\tilde{\mathbf{A}})$. Simply put, removing or applying perturbation of an important concept will result in a substantial variation in the output, whereas an unused concept will have minimal effect on the output.

Finally, we can capture the importance that a concept might have as a main effect – along with its interactions with other concepts – on the model's output by calculating the expected variance that would remain if all the concepts except the i were to be fixed. This yields the general definition of the total Sobol indices.

Definition 3.1 (Total Sobol indices). *The total Sobol index $\mathcal{S}_i^T$, which measures the contribution of a concept i as well as its interactions of any order with any other concepts to the model output variance, is given by:*

$$\mathcal{S}_i^T = \frac{\mathbb{E}_{\boldsymbol{M}_{\sim i}}(\mathbb{V}_{M_i}(\boldsymbol{h}((\mathbf{U} \odot \mathbf{M})\mathbf{W}^\intercal)|\mathbf{M}_{\sim i}))}{\mathbb{V}(\boldsymbol{h}((\mathbf{U} \odot \mathbf{M})\mathbf{W}^\intercal))}. \quad (2)$$

In practice, this index can be calculated very efficiently [36, 48, 52, 58, 72], more details on the Quasi-

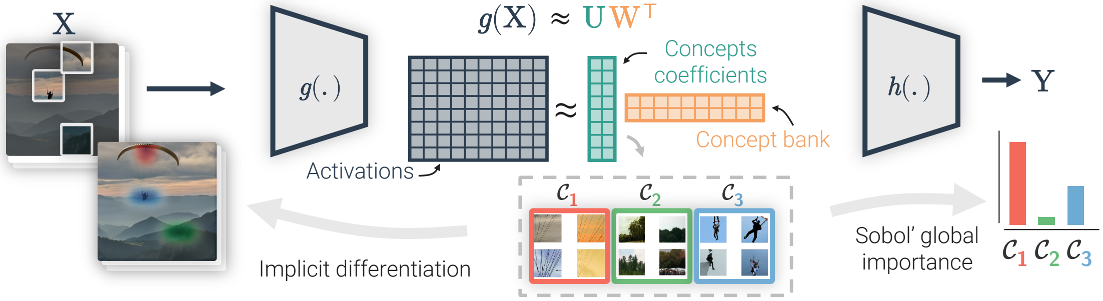

Figure 4. **Overview of CRAFT.** Starting from a set of crops $\mathbf{X}$ containing a concept $\mathcal{C}$ (e.g., crops images of the class "parachute"), we compute activations $\boldsymbol{g}(\mathbf{X})$ corresponding to an intermediate layer from a neural network for random image crops. We then factorize these activations into two lower-rank matrices, $(\mathbf{U}, \mathbf{W})$. $\mathbf{W}$ is what we call a "concept bank" and is a new basis used to express the activations, while $\mathbf{U}$ corresponds to the corresponding coefficients in this new basis. We then extend the method with 3 new ingredients: (1) recursivity – by proposing to re-decompose a concept (e.g., take a new set of images containing $\mathcal{C}_1$) at an earlier layer, (2) a better importance estimation using Sobol indices and (3) an approach to leverage implicit differentiation to generate *concept attribution maps* to localize concepts in an image.

Monte Carlo sampling and the estimator used are left in appendix D.

3.4. Ingredient 3: A smidgen of implicit differentiation

Attribution methods are useful for determining the regions deemed important by a model for its decision, but they lack information about what exactly triggered it. We have seen that we can already extract this information from the matrices $\mathbf{U}$ and $\mathbf{W}$, but as it is, we do not know in what part of an image a given concept is represented. In this section, we will show how we can leverage attribution methods (forward and backward modes) to find where a concept is located in the input image (see Fig. 2). Forward attribution methods do not rely on any gradient computation as they only use inference processes, whereas backward methods require back-propagating through a network's layers. By application of the chain rule, computing $\partial \mathbf{U}/\partial \mathbf{X}$ requires access to $\partial \mathbf{U}/\partial \mathbf{A}$.

To do so, one could be tempted to solve the linear system $\mathbf{U}\mathbf{W}^\intercal = \mathbf{A}$. However, this problem is ill-posed since $\mathbf{W}^\intercal$ is low rank. A standard approach is to calculate the Moore-Penrose pseudo-inverse $(\mathbf{W}^\intercal)^\dagger$, which solves rank deficient systems by looking at the minimum norm solution [2]. In practice, $(\mathbf{W}^\intercal)^\dagger$ is computed with the Singular Value Decomposition (SVD) of $\mathbf{W}^\intercal$. Unfortunately, SVD is also the solution to the *unstructured minimization* of $\frac{1}{2}\|\mathbf{A}-\mathbf{U}\mathbf{W}^\intercal\|_F^2$ by the Eckart-Young-Mirsky theorem [12]. Hence, the non-negativity constraints of the NMF are ignored, which prevents such approaches from succeeding. Other issues stem from the fact that the $\mathbf{U}, \mathbf{W}$ decomposition is generally not unique.

Our third contribution consists of tackling this problem to allow the use of attribution methods, i.e., *concept attribution maps*, by proposing a strategy to differentiate through the NMF block.

Implicit differentiation of NMF block The NMF problem 1 is NP-hard [73], and it is not convex with respect to the input pair $(\mathbf{U}, \mathbf{W})$. However, fixing the value of one of the two factors and optimizing the other turns the NMF formulation into a pair of Non-Negative Least Squares (NNLS) problems, which are convex. This ensures that alternating minimization (a standard approach for NMF) of $(\mathbf{U}, \mathbf{W})$ factors will eventually reach a local minimum. Each of this alternating NNLS problems fulfills the Karush-Kuhn-Tucker (KKT) conditions [38, 45], which can be encoded in the so-called *optimality function* $\mathbf{F}$ from [4], see Eq. 9 Appendix C.2. The implicit function theorem [25] allows us to use implicit differentiation [3, 25, 44] to efficiently compute the Jacobians $\partial \mathbf{U}/\partial \mathbf{A}$ and $\partial \mathbf{W}/\partial \mathbf{A}$ without requiring to back-propagate through each of the iterations of the NMF solver:

$$\frac{\partial(\mathbf{U}, \mathbf{W}, \bar{\mathbf{U}}, \bar{\mathbf{W}})}{\partial \mathbf{A}} = -(\partial_1 \mathbf{F})^{-1} \partial_2 \mathbf{F}. \tag{3}$$

However, this requires the dual variables $\bar{\mathbf{U}}$ and $\bar{\mathbf{W}}$, which are not computed in scikit-learn's [54] popular implementation[*]. Consequently, we leverage the work of [32] and we re-implement our own solver with Jaxopt [4] based on ADMM [5], a GPU friendly algorithm (see Appendix C.2).

Concretely, given our concepts bank $\mathbf{W}$, the concept attribution maps of a new input $\boldsymbol{x}$ are calculated by solving the NNLS problem $\min_{\mathbf{U}\geq 0} \frac{1}{2}\|\boldsymbol{g}(\boldsymbol{x}) - \mathbf{U}\mathbf{W}^\intercal\|_F^2$. The implicit differentiation of the NMF block $\partial \mathbf{U}/\partial \mathbf{A}$ is integrated into the classic back-propagation to obtain $\partial \mathbf{U}/\partial \boldsymbol{x}$. Most interestingly, this technical advance enables the use of all white-box explainability methods [59, 62, 66, 69, 71, 78] to generate concept-wise attribution maps and trace the part of an image that triggered the detection of the concept by the network. Additionally, it is even possible to employ black-box

[*] Scikit-learn uses a block coordinate descent algorithm [7, 17], with a randomized SVD initialization.

		Husky vs. Wolf				*Leaves*				*“Kit Fox” vs “Red Fox”*			
	Session n°	1	2	3	*Utility*	1	2	3	*Utility*	1	2	3	*Utility*
	Baseline	55.7	66.2	62.9		70.1	76.8	78.6		58.8	62.2	58.8	
	Control	53.3	61.0	61.4	0.95	72.0	78.0	80.2	1.02	60.7	59.2	48.5	0.94
Attributions	Saliency [62]	53.9	69.6	73.3	1.06	83.2	88.7	82.4	1.13	61.7	60.2	58.2	1.00
	Integ.-Grad. [71]	67.4	72.8	73.2	1.15	82.5	82.5	85.3	1.11	59.4	58.3	58.3	0.98
	SmoothGrad [66]	68.7	75.3	78.0	1.20	83.0	85.7	86.3	1.13	50.3	55.0	61.4	0.93
	GradCAM [59]	77.6	85.7	84.1	1.34	81.9	83.5	82.4	1.10	54.4	52.5	54.1	0.90
	Occlusion [76]	71.0	75.7	78.1	1.22	78.8	86.1	82.9	1.10	51.0	60.2	55.1	0.92
	Grad.-Input [61]	65.8	63.3	67.9	1.06	76.5	82.9	79.5	1.05	50.0	57.6	62.6	0.95
Concepts	ACE [24]	68.8	71.4	72.7	1.15	79.8	73.8	82.1	1.05	48.4	46.5	46.1	0.78
	CRAFTCO (ours)	82.4	87.0	85.1	1.38	78.8	85.5	89.4	1.12	55.5	49.5	53.3	0.88
	CRAFT (ours)	90.6	97.3	95.5	**1.53**	86.2	86.6	85.5	**1.15**	56.5	50.6	49.4	0.87

Table 1. ***Utility* scores on 3 datasets from [8]**. Their *Utility* benchmark evaluates how well explanations help users identify general rules driving classifications that readily transfer to unseen instances. At training time, users are asked to infer rules driving the decisions of the model given a set of images, and their associated predictions and explanations. At test time, the *Utility* metric measures the accuracy of users at predicting the model decision on novel images averaged over 3 sessions, and normalized by the baseline accuracy of users trained without explanations. The higher the *Utility* score, the more useful the explanation, and the more crucial the information provided is for understanding –and thus predicting the model's output– on novel samples. CRAFTCO stands for “CRAFT Concept Only” and designates an experimental condition where only global concepts are given to users, without local explanations (i.e., the concept attribution maps). The first and second best results above the baseline are in **bold** and underlined, respectively.

methods [14, 47, 55, 57] since it only amounts to solving an NNLS problem.

4. Experimental evaluation

In order to evaluate the interest and the benefits brought by CRAFT, we start in Section 4.1 by assessing the practical utility of the method on a human-centered benchmark composed of 3 XAI scenarios.

After demonstrating the usefulness of the method using these human experiments, we independently validate the 3 proposed ingredients. First, we provide evidence that recursivity allows refining concepts, making them more meaningful to humans using two additional human experiments in Section 4.2. Next, we evaluate our new Sobol estimator and show quantitatively that it provides a more faithful assessment of concept importance in Section 4.3. Finally, we run an ablation experiment that measures the interest of local explanations based on concept attribution maps coupled with global explanations. Additional experiments, including a sanity check and an example of deep dreams applied on the concept bank, as well as many other examples of local explanations for randomly picked images from ILSVRC2012, are included in Section B of the supplementary materials. We leave the discussion on the limitations of this method and on the broader impact in appendix A.

4.1. Utility Evaluation

As emphasized by Doshi-Velez et al. [11], the goal of XAI should be to develop methods that help a user better understand the behavior of deep neural network models. An instantiation of this idea was recently introduced by Colin & Fel et al. [8] who described an experimental framework to quantitatively measure the practical usefulness of explainability methods in real-world scenarios. For their initial setup, these authors recruited $n = 1,150$ online participants (evaluated over 8 unique conditions and 3 AI scenarios) – making it the largest benchmark to date in XAI. Here, we follow their framework rigorously to allow for the robust evaluation of the utility of our proposed CRAFT method and the related ACE. The 3 representative real-world scenarios are: (1) identifying bias in an AI system (using Husky vs Wolf dataset from [57]), (2) characterizing the visual strategy that are too difficult for an untrained non-expert human observer (using the Paleobotanical dataset from [75]), (3) understanding complex failure cases (using ImageNet “Red fox” vs “Kit fox” binary classification). Using this benchmark, we evaluate CRAFT, ACE, as well as CRAFT with only the global concepts (CRAFTCO) to allow for a fair comparison with ACE. To the best of our knowledge, we are the first to systematically evaluate concept-based methods against attribution methods.

Results are shown in Table 1 and demonstrate the benefit of CRAFT, which achieves higher scores than all of the attribution methods tested as well as ACE in the first two scenarios. To date, no method appears to exceed the baseline on the third scenario suggesting that additional work is required. We also note that, in the first two scenarios, CRAFTCO is one of the best-performing methods and it always outperforms ACE – meaning that even without the local explanation of the concept attribution maps, CRAFT

	Experts ($n = 36$)	Laymen ($n = 37$)
Intruder		
Acc. Concept	70.19%	61.08%
Acc. Sub-Concept	74.81% ($p = 0.18$)	**67.03%** ($p = 0.043$)
Binary choice		
Sub-Concept	**76.1%** ($p < 0.001$)	**74.95%** ($p < 0.001$)
Odds Ratios	3.53	2.99

Table 2. **Results from the psychophysics experiments to validate the recursivity ingredient.**

largely outperforms ACE. Examples of concepts produced by CRAFT are shown in the Appendix E.1.

4.2. Validation of Recursivity

To evaluate the meaningfulness of the extracted high-level concepts, we performed psychophysics experiments with human subjects, whom we asked to answer a survey in two phases. Furthermore, we distinguished two different audiences: on the one hand, experts in machine learning, and on the other hand, people with no particular knowledge of computer vision. Both groups of participants were volunteers and did not receive any monetary compensation. Some examples of the developed interface are available the appendix E. It is important to note that this experiment was carried out independently from the utility evaluation and thus it was setup differently.

Intruder detection experiment First, we ask users to identify the intruder out of a series of five image crops belonging to a certain class, with the odd one being taken from a different concept but still from the same class. Then, we compare the results of this intruder detection with another intruder detection, this time, using a concept (e.g., $\mathcal{C}_1$) coming from a layer l and one of its sub-concepts (e.g., $\mathcal{C}_{12}$ in Fig.3) extracted using our recursive method. If the concept (or sub-concept) is coherent, then it should be easy for the users to find the intruder. Table 2 summarizes our results, showing that indeed both concepts and sub-concepts are coherent, and that recursivity can lead to a slightly higher understanding of the generated concepts (significant for non-experts, but not for experts) and might suggest a way to make concepts more interpretable.

Binary choice experiment In order to test the improvement of coherence of the sub-concept generated by recursivity with respect to the larger parent concept, we showed participants an image crop belonging to both a subcluster and a parent cluster (e.g., $\boldsymbol{\pi}(\boldsymbol{x}) \in \mathcal{C}_{11} \subset \mathcal{C}_1$) and asked them which of the two clusters (i.e., $\mathcal{C}_{11}$ or $\mathcal{C}_1$) seemed to accommodate the image the best. If our hypothesis is correct, then the concept refinement brought by recursivity should help form more coherent clusters. The results in Table 2 are satisfying since in both the expert and non-expert groups, the participants chose the sub-cluster more than 74% of the time. We measure the significance of our results by fitting a binomial logistic regression to our data, and we find that both groups are more likely to choose the sub-concept cluster (at a $p < 0.001$).

4.3. Fidelity analysis

We propose to simultaneously verify that identified concepts are faithful to the model and that the concept importance estimator performs better than that used in TCAV [40] by using the fidelity metrics introduced in [24, 77]. These metrics are similar to the ones used for attribution methods, which consist of studying the change of the logit score when removing/adding pixels considered important. Here, we do not introduce these perturbations in the pixel space but in the concept space: once $\mathbf{U}$ and $\mathbf{W}$ are computed, we reconstruct the matrix $\mathbf{A} \approx \mathbf{U}\mathbf{W}^\mathsf{T}$ using only the most important concept (or removing the most important concept for deletion) and compute the resulting change in the output of the model. As can be seen from Fig. 5, ranking the extracted concepts using Sobol's importance score results in steeper curves than when they are sorted by their TCAV scores. We confirm that these results generalize with other matrix factorization techniques (PCA, ICA, RCA) in Section F of the Appendix.

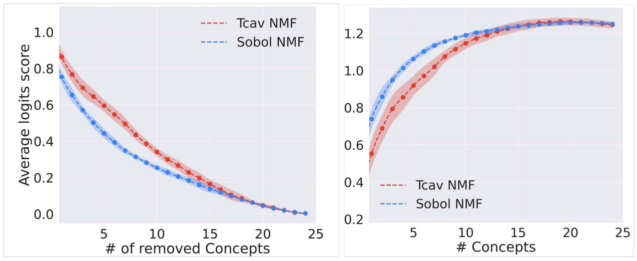

Figure 5. **(Left)** Deletion curves (lower is better). **(Right)** Insertion curves (higher is better). For both the deletion or insertion metrics, Sobol indices lead to better estimates (calculated on >100K images) of important concepts.

5. Conclusion

We presented CRAFT, a method to automatically extract understandable concepts from deep networks, which provides insights into the model's decision-making process. Our approach uses a recursive formulation to identify concepts at the correct level of granularity, a novel method for measuring concept importance, and implicit differentiation to generate concept attribution maps. We conducted psychophysics experiments to validate the approach's usefulness and meaningfulness to human experimenters. We anticipate that our work will inspire future research in concept-based explainability methods.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 2, 3, 24

[2] João Carlos Alves Barata and Mahir Saleh Hussein. The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1):146–165, 2012. 6

[3] Bradley M Bell and James V Burke. Algorithmic differentiation of implicit functions and optimal values. In *Advances in Automatic Differentiation*, pages 67–77. Springer, 2008. 6, 19

[4] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *arXiv preprint arXiv:2105.15183*, 2021. 3, 6, 19

[5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. 6, 13, 18, 19

[6] 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. 13

[7] Andrzej Cichocki and Anh-Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009. 6, 19

[8] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3, 7, 21

[9] RI Cukier, CM Fortuin, Kurt E Shuler, AG Petschek, and J Ho Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. *The Journal of chemical physics*, 59(8):3873–3878, 1973. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 14, 23

[11] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *ArXiv e-print*, 2017. 1, 3, 7

[12] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. 6

[13] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. 4

[14] Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3, 7, 19

[15] Thomas Fel, Mélanie Ducoffe, David Vigouroux, Rémi Cadène, Mikael Capelle, Claire Nicodème, and Thomas Serre. Don't lie to me! robust and efficient explainability with verified perturbation analysis. *arXiv preprint arXiv:2202.07728*, 2022. 3

[16] Thomas Fel, Lucas Hervier, David Vigouroux, Antonin Poche, Justin Plakoo, Remi Cadene, Mathieu Chalvidal, Julien Colin, Thibaut Boissin, Louis Béthune, Agustin Picard, Claire Nicodeme, Laurent Gardes, Gregory Flandin, and Thomas Serre. Xplique: A deep learning explainability toolbox. *Workshop, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 16

[17] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011. 6, 19

[18] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 3

[19] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 13

[20] Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, and Wing-Kin Ma. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Process. Mag.*, 36(2):59–80, 2019. 5

[21] Yunhao Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyan Wu. A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2195–2204, 2021. 13

[22] Sahra Ghalebikesabi, Lucile Ter-Minassian, Karla DiazOrdaz, and Chris C Holmes. On locality of local explanation models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[23] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 3

[24] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 4, 5, 7, 8, 13, 14, 23

[25] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008. 3, 6, 19

[26] 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

[27] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[28] Johannes Haug, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. On baselines for local feature attributions. *arXiv preprint arXiv:2101.00905*, 2021. 3

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 23

[30] Magnus R Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving. *Journal of research of the National Bureau of Standards*, 49(6):409, 1952. 18, 19

[31] 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3

[32] Kejun Huang, Nicholas D Sidiropoulos, and Athanasios P Liavas. A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. *IEEE Transactions on Signal Processing*, 64(19):5052–5065, 2016. 6, 18

[33] Marouane Il Idrissi, Vincent Chabridon, and Bertrand Iooss. Developments and applications of shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling & Software*, 2021. 5

[34] Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer, 2015. 2, 5

[35] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021. 1

[36] Alexandre Janon, Thierry Klein, Agnes Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364, 2014. 5, 20

[37] Margot E Kaminski and Jennifer M Urban. The right to contest ai. *Columbia Law Review*, 121(7):1957–2048, 2021. 1

[38] William Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939. 6, 18

[39] 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. 3

[40] 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 *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 2, 3, 5, 8, 16, 23

[41] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[42] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham, 2019. 3

[43] Mauritz Kop. Eu artificial intelligence act: The european approach to ai. In *Stanford - Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments, Stanford University, Issue No. 2/2021. https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/.* Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust , 2021. 1

[44] Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002. 3, 6, 19

[45] Harold W Kuhn and Albert W Tucker. Nonlinear programming proceedings of the second berkeley symposium on mathematical statistics and probability. *Neyman*, pages 481–492, 1951. 6, 18

[46] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 2

[47] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 3, 7

[48] Amandine Marrel, Bertrand Iooss, Beatrice Laurent, and Olivier Roustant. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3):742–751, 2009. 5

[49] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[50] Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3

[51] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization. 13, 16

[52] Art B Owen. Better estimation of small sobol'sensitivity indices. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(2):1–17, 2013. 5

[53] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 5

[54] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011. 6

[55] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 1, 3, 7, 19, 23

[56] Bin Ren, Laurent Pueyo, Christine Chen, Élodie Choquet, John H Debes, Gaspard Duchêne, François Ménard, and Marshall D Perrin. Using data imputation for signal separation in high-contrast imaging. *The Astrophysical Journal*, 892(2):74, 2020. 4

[57] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 3, 7, 19

[58] Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer physics communications*, 181(2):259–270, 2010. 5

[59] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 3, 6, 7

[60] 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

[61] 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. 1, 7

[62] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014. 1, 3, 6, 7

[63] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 2

[64] Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. Do users benefit from interpretable vision? a user study, baseline, and dataset. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2

[65] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. 2, 3

[66] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1, 3, 5, 6, 7

[67] Ilya M Sobol. Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment*, 1:407–414, 1993. 2, 5, 20

[68] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001. 2, 5

[69] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 1, 3, 6

[70] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. 3, 13

[71] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 3, 5, 6, 7

[72] Stefano Tarantola, Debora Gatelli, and Thierry Alex Mara. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, 91(6):717–727, 2006. 5

[73] Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010. 6

[74] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013. 5

[75] 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. 7

[76] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1, 3, 7

[77] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. *arXiv preprint arXiv:2006.15417*, 2020. 2, 3, 5, 8, 23

[78] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3, 6