CRAFT: Concept Recursive Activation FacTorization for Explainability

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite their considerable potential, concept-based explainability methods have received relatively little attention, and explaining *what*'s driving models' decisions and *where* it's located in the input is still an open problem. To tackle this, we revisit unsupervised concept extraction techniques for explaining the decisions of deep neural networks and present CRAFT – a framework to generate concept-based explanations for understanding individual predictions and the model's high-level logic for whole classes. CRAFT takes advantage of a novel method for recursively decomposing higher-level concepts into more elementary ones, combined with a novel approach for better estimating the importance of identified concepts with Sobol indices. Furthermore, we show how implicit differentiation can be used to generate concept-wise attribution explanations for individual images. We further demonstrate through fidelity metrics that our proposed concept importance estimation technique is more faithful to the model than previous methods, and, through human psychophysic experiments, we confirm that our recursive decomposition can generate meaningful and accurate concepts. Finally, we illustrate CRAFT's potential to enable the understanding of predictions of trained models on multiple use-cases by producing meaningful concept-based explanations.*

1 Introduction

Interpreting the decisions of modern machine learning models such as neural networks remains a major challenge. The need for robust and reliable explainability methods has never been more urgent as machine learning is being applied to an ever increasing range of domains, including safety critical ones. The application of the General Data Protection Regulation law (GDPR) [1] in the European Union has drawn the attention of the general public to the rights they should have on their data. This kickstarted a race for other needs, with more and more regulation agencies asking for the right for AI decisions to be explainable to users – e.g. European AI act [2], EASA concepts for design assurance [3].

In order to try to meet this need, an array of explainability methods have already been proposed. Most of these methods aim at explaining what inputs (or pixels in an image) are driving the model's decision. These so-called attribution methods yield heatmaps that indicate the importance of individual pixels. Among the most notable ones is LIME [4], which was initially developed to try to locally – that is, at an instance level – understand models' predictions to identify possible biases in vision models. Multiple improvements have since been introduced – either by better harnessing the information provided by gradients to estimate the importance of individual pixels [5, 6, 7, 8, 9, 10, 11, 12], leveraging image perturbations to evaluate the sensitivity of a model's output [13, 14] or, more recently, via the use of formal methods to generate explanations [15].

*Our code is available at `anonymous.4open.science/r/craft-concept-explanation-4351`.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

However, all the aforementioned methods focus on one side of explainability – answering the question of *where* – i.e., where in an image are the pixels that are critical to the decision located. They leave the question of *what* – i.e., what visual features are actually driving decisions – entirely open. We argue that this limitation is one of the main reasons why these methods fail in some cases to help users, for instance, identify the source of a system's bias or its failure cases as shown in [16]. Feature visualization methods [17, 18] characterize the selectivity of individual neurons (or neural channels or arbitrary directions in the neural activation space) via the synthesis of input stimuli which maximize their responses and can partially answer this question. Still in this vein, [19, 20, 21] proposed to use the training dataset to identify the samples that contribute the most to the model's decision. Finally, closer to our work, a new line of research has recently been initiated [22] based on high-level concepts. The goal of this branch is to find humanly interpretable concepts in the activation space of a layer in a neural network. This approach can give positive results, but in its original formulation, it requires prior knowledge on the relevant concepts, and more importantly, the labeling of a dataset for each of the concepts we want to extract. Hence, several works have proposed to automate the concept search based only on the training dataset and without explicit human supervision. The most prominent of these techniques, ACE [23], uses a combination of segmentation and clustering techniques, but requires heuristics to remove outliers. This method unlocks the possibility of large scale concept extraction without additional labeling or human supervision. Nevertheless, it suffers from several problems: each segment can only belong to one cluster, the choice of the layer from which to retrieve the concepts is not clear, and the amount of information lost during the outlier rejection phase can be a cause of concern. More recently, [24] proposes to leverage matrix decompositions on internal feature maps to discover concepts.

It is important to note that current work does not offer a link between their global and local explanations, nor do they offer an answer to the question of which layer to choose to perform the decomposition. Building up on these conclusions, we revisit these concept extraction techniques by using Non-Negative Matrix Factorization (NMF) and propose 3 different ingredients to answer these questions simultaneously, thereby introducing CRAFT, a new automatic concept extraction method.

We can summarize our main contributions as follows:

- A novel approach for the automated extraction of high-level concepts learned by deep neural networks.
- A recursive procedure to automatically decompose concepts into sub-concepts, starting with the last layer of the model and working our way inwards. We validate the benefit of this recursivity – i.e. decomposing concepts into sub-concepts – with human psychophysic experiments which show that (i) that the decomposition of a concept yields more coherent sub-concepts (ii) the groups of points formed by these sub-concepts are more refined and appear meaningful to humans (expert or non-expert).
- A novel technique to quantify the importance of individual concepts on a model's predictions using Sobol indices coming from the field of Sensitivity Analysis.
- A novel *Concept Attribution Map* (CAM) method to backpropagate each of the concept values independently into the pixel space by leveraging the implicit function theorem, allowing us to locate the concept in a given input image. This effectively unlocks the ability to apply all the white-box [5, 6, 7, 8, 9, 12, 25] and black-box [4, 26, 13, 14] explainability techniques in the literature to obtain concept-wise attribution maps.
- A demonstration of the approach combining local and global explanations to accurately explain predictions and understand complex failure cases.

2 Related Work

Explaining *where* The widespread use of black-box machine learning methods including deep convolutional neural networks in myriads of computer vision tasks prompted a need to understand where in the input image the model looked to make predictions. These explanatory heatmaps can be generated through completely different approaches depending on whether access to gradients is provided. If it is indeed the case, there's a plethora of different methods that harnesses intermediary information inside the neural network to create these explanations [5, 8, 7, 28, 6, 29, 9, 12]. However, they have been found to induce confirmation bias [30] and to be vulnerable to adversarial attacks [31]. Somewhat differently, there are other methods [10, 11] that harness gradients to optimize masks to maximize the impact on the predictions, and thus determine the most important parts of the input for the model. However, if only the input and its corresponding output are available, other

Figure 1: **CRAFT Results for the prediction 'Chain Saw'.** First, our method uses NMF to extract from the train set (ILSVRC2012 [27]) the most relevant concepts used by the network (ResNet50V2). Then, the global influence of these concepts on predictions is measured using Sobol indices (right panel). Finally, the method provides local explanations through *Concept Attribution Maps* (heatmap associated to 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: $\mathcal{C}_0$ for the chainsaw engine, $\mathcal{C}_2$ for the saw blade, $\mathcal{C}_4$ for the human head, $\mathcal{C}_{18}$ for the vegetation, $\mathcal{C}_{21}$ for the jeans and $\mathcal{C}_{22}$ for the tree trunk.

techniques exist that enable the generation of attribution maps by locally estimating the importance of each input pixel: LIME [4], RISE [13], and more recently, an attribution method based on Sobol indices [14, 32]. Crucially, they propose to input perturbed versions of the example one wishes to explain and either construct a linear model to determine the importance of each region of the input, leverage Monte-Carlo methods to this end, or compute the Sobol indices [32] associated to them as a measure of their influence on the model. Concretely, we will be exploiting all this literature to locate the important parts of images with respect to what we will call "high-level" concepts by generating concept-wise attribution maps.

Explaining «what» There have been studies [33, 34] that indicate that CNNs trained on the ImageNet dataset [27] rely heavily on textures to classify, and largely disregard the shapes. For this reason, some researchers suggest that attribution maps might not be enough to explain models' predictions [17], and that explainability methods revealing the role of the textures are a must. Namely, in [18] and [17], explanations are generated as the inputs that would maximize the neural activation of a given layer with respect to a given class. However, these explanations may not be easily interpretable by humans. Finally, other approaches suggest to modify the structure of the neural network, either by constraining the convolutional layers to naturally provide visual explanations [35], or by forcing it to generate prototypes for the classes [36], but our main focus are post-hoc methods that can be applied to pre-trained neural networks and don't need further training.

Concept discovery. In [22], Kim et al. proposed an alternative to explaining the *what*: they built a database with different concepts (such as "stripes") to extract a concept vector in the latent space of a given layer. Then they proposed to estimate the importance of this concept vector using the directional derivative of the model's predictions with respect to this concept vector. However, it is a supervised approach, and thus, only applicable when we have prior knowledge of the concepts in play. The natural extension of this idea is automatic discovery of concepts in an unsupervised fashion, without the need for prior knowledge or labelled concept datasets. As such, in [23], a technique is proposed to discover these "high-level" concepts: they perform segmentation at different resolutions on patches of images, cluster them and select the most significant based on perception and Testing with Concept Activation Vectors (TCAV) [22] scores. However, the quality of the result is highly dependent on the segmentation scheme and on the layer used for perception scores. Building up on this technique, [24] propose to generate a bank of concepts for each class by performing dimensionality reductions on the activation maps flattened over the channel dimension. Once the factorization done, the reconstruction of the activation of the image can then be interpreted as a combination of a set of concepts and a coefficient associated to these concepts. Not all factorization-based methods are equal though. Their large-scale human experiments show an interesting trend: Non-negative Matrix Factorization (NMF) is widely preferred over Principal Component Analysis (PCA) or ACE for generating meaningful

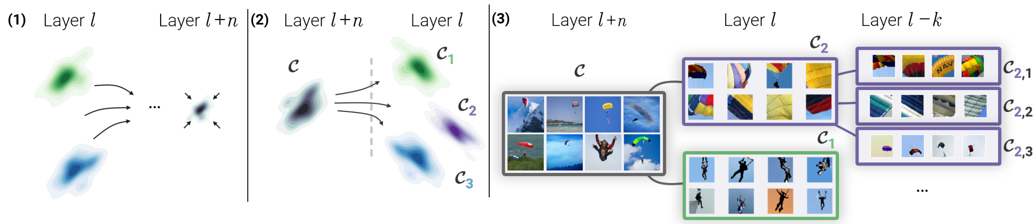

Figure 2: **(1) Neural Collapse (Amalgamation)** classifiers need to be able to linearly separate each class at the last layer, and to do this, the activations of the same class must merge during the forward pass until they all converge to the one-hot vector of the class in the logits layer. This may result in activations that are too concentrated to be broken down into meaningful concepts. **(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 concept recursive decomposition** using CRAFT on the class 'Parachute' of ILSVRC2012 [27].

concepts for humans. Finally, [37] defines the notion of *completeness* of a concept bank and proposes a method to learn a complete set of concepts using Shapley values [26].

3 Overview of the method

In this section, we first describe our Concept Activations Factorization method by pointing out the differences that set our technique apart from previous work. We then proceed to introduce the three new ingredients that make up CRAFT: (1) a method to recursively decompose concepts into sub-concepts, (2) a new approach to better estimate the importance of extracted concepts and (3) how we unlock any attribution method to create *Concept Attribution Maps*, using implicit differentiation [38, 39, 40].

Notation In this work, we consider a general supervised learning setting, where $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \in \mathcal{X}^n$ are n points and $(y_1, ..., y_n) \in \mathcal{Y}^n$ their associated labels. Unless specified, all points are assumed to have the same 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 assume that $\boldsymbol{f}$ is a neural network composed of k layers, and we denote $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{h}_k \circ \boldsymbol{h}_{k-1} \circ ... \circ \boldsymbol{h}_1(\boldsymbol{x})$ with $\boldsymbol{h}_l(\boldsymbol{x}) \subseteq \mathbb{R}^p$ being the intermediate activations for the layer l and $\boldsymbol{h}_l(\boldsymbol{x})_i$ an activation for the same layer. Further, we require non-negative activations: $\boldsymbol{h}_l(\boldsymbol{x})_i \geq 0 : \forall i \in \{1, ..., p\}$, which amounts to choosing a layer whose activation function $\boldsymbol{\sigma}_l(\boldsymbol{x}) \geq 0$. In particular, this assumption is verified by any architecture that utilizes *ReLU*, but any non-negative activation function works. Finally, we denote $\boldsymbol{h}_{l,k}$ the function going from the layer l to the output of the model $\boldsymbol{f}$.

3.1 Concept Activations Factorization

As illustrated in Fig.3, we propose to use Non-negative matrix factorization activations to find a basis of concepts. Inspired by ACE [23], we will use sub-regions of images to attempt to identify coherent concepts. Instead of using segmentation – which naturally introduces artifacts due to the inpainting required by a baseline value –, we start by taking random crops of each image in our dataset (e.g, a set of points that the model predicts as belonging to the same class) to form an auxilary dataset $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ such that $\boldsymbol{X}_i = \boldsymbol{\tau}(\boldsymbol{x}_i)$ with $\boldsymbol{\tau}$ a crop function. Given a layer l, we obtain the activations for the random crops $\boldsymbol{A} = \boldsymbol{h}_l(\boldsymbol{X}) \in \mathbb{R}^{n \times p}$. In the case where $\boldsymbol{f}$ is a convolutional neural network, a global average pooling is applied on the activations. We recall that all the elements of $\boldsymbol{A}$ are non-negative real numbers.

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

$$\min_{\boldsymbol{U} \geq 0, \boldsymbol{W} \geq 0} \frac{1}{2} \|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T\|_F^2 \quad (1)$$

Where $|| \cdot ||_F$ denotes the Frobenius norm. One of the appealing properties of NMF is the low rank constraint $r \ll \mathbf{min}(n, p)$. Simply put, NMF can be understood as the joint learning of $\boldsymbol{W}$, a dictionary of CAVs – "concept bank" in Figure 3 – that maps a $\mathbb{R}^p$ basis onto $\mathbb{R}^r$, and $\boldsymbol{U}$ the coefficients of vectors $\boldsymbol{A}$ expressed in this new basis. The minimization of the reconstruction error

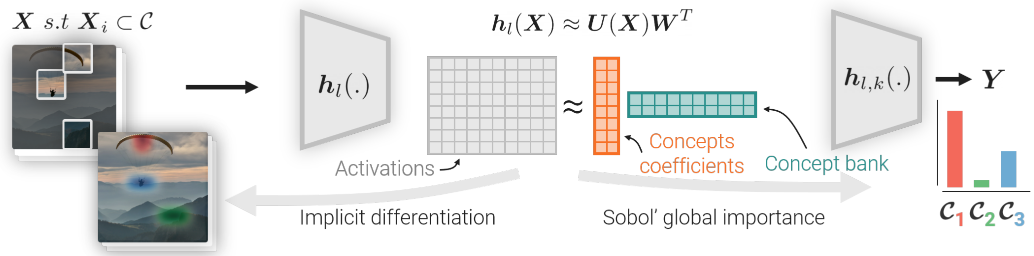

Figure 3: **Overview of CRAFT.** Starting from a set of crops $\boldsymbol{X}$ containing a concept $\mathcal{C}$ (e.g., crops images of the class Parachute), we send random crops to a layer l to get activations it $\boldsymbol{h}_l(\boldsymbol{X})$. We then factorize the activation into two lower rank matrices, $\boldsymbol{U}$ and $\boldsymbol{W}$. $\boldsymbol{W}$ is what we call a concept bank (a base of concepts), while $\boldsymbol{U}$ corresponds to the coefficients in this new basis. We then extend the method with 3 new ingredients: (1) the recursivity by proposing to re-decompose a concept (e.g., take a new set of point containing $\mathcal{C}_1$) at an earlier layer $l' < l$, (2) a better importance estimation using Sobol indices and (3) leveraging implicit differentiation to generate *Concept Attribution Maps* allowing to localize concepts in an image.

$\frac{1}{2}\|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}\|_F^2$ ensures that the new basis contains (mostly) relevant concepts. Intuitively, the non-negativity constraints $\boldsymbol{U} \geq 0, \boldsymbol{W} \geq 0$ encourage (i) the sparsity of $\boldsymbol{W}$ (useful for creating disentangled concepts), (ii) the sparsity of $\boldsymbol{U}$ (convenient for selecting a minimal set of useful concepts) and (iii) the imputation of missing data [41], which corresponds to the sparsity pattern of *post-ReLU* activations $\boldsymbol{A}$. We shall also note that each original activation $\boldsymbol{A}_i$ coming from the input $\boldsymbol{x}_i$ can be approximated by its reconstruction $\boldsymbol{h}_l(\boldsymbol{\tau}(\boldsymbol{x}_i)) = \boldsymbol{U}_i \boldsymbol{W}^T = \sum_{j=1}^{r} \boldsymbol{U}_{i,j} \boldsymbol{W}_j^T$. This approach is attractive as each activation can be understood as a composition of concepts.

While other methods in the literature solve a similar problem (such as low rank factorization using SVD or ICA), the NMF has stepped up as both fast, effective and is known to yield meaningful concepts to humans [42, 43, 24]. Finally, once the concept bank $\boldsymbol{W}$ is precomputed, we can associate the concept coefficients $\boldsymbol{U}(\boldsymbol{x})$ to any new input $\boldsymbol{x}$ (e.g a full image) by solving the underlying Non-Negative Least Squares (NNLS) problem $\min_{\boldsymbol{U} \geq 0} \frac{1}{2}\|\boldsymbol{h}_l(\boldsymbol{x}) - \boldsymbol{U}(\boldsymbol{x})\boldsymbol{W}^T\|_F^2$, and therefore have its decomposition in the concept base.

In essence, the core of our method can be summarized as follows: using a set of images, we re-interpret their embedding at a given layer l as a composition of concepts that can be easily understood by humans. In the next section we show how we can recursively apply concept activation factorizations on a layer $l' < l$ for an image containing a previously computed concept.

3.2 Ingredient 1: A Recursive Flavor

One of the most apparent issues in previous works [23, 24] is the choice of the layer at which the activation maps are extracted. Depending on this, certain concepts start getting amalgamated [44] into one, resulting in incoherent and indecipherable clusters, as illustrated in Fig 2. 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, look for their sub-concepts at 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 points 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 see Fig.2 for an illustrative example. Note that we generalize current methods in the sense that taking points $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ that are clustered in the logits layer (belonging to the same class) and decomposing them in a previous layer – as done in [23, 24] – is a valid recursive step. For a more general case, let us assume that a set of points that contain a common concept is obtained using a first step of Concept Activation Factorization. We then look for a set of points with a high coefficient for the concept of our choice to perform the next factorization. Formally, with a factorization for a layer l $\boldsymbol{U}\boldsymbol{W}^T$ and a concept index i, this set of points is defined as $\mathcal{C} = \{\boldsymbol{\tau}(\boldsymbol{x}_j) \ : \ \boldsymbol{U}(\boldsymbol{A}_j)_i \geq \lambda\}$ In practice, we assume λ to be equal to the 90th percentile of the values of $\boldsymbol{U}_i$. Given this new set of points, we can then re-apply the Concept Matrix Factorization method to a earlier layer l' – with $l' < l$ – to obtain the sub-concept's decomposition from the initial concept.

3.3 Ingredient 2: Sobol indices for enhanced concept importance estimation

A common concern with concept extraction methods is that what makes sense to humans is not necessarily what is being used by the model to predict. To avoid this kind of confirmation bias during our concept analysis phase, we can estimate the global importance of the extracted concepts. To do so, [22] proposed an estimator based on directional derivatives: the partial derivative of the model output with respect to the concept vector. While this measure is theoretically well founded, it relies on the same principle as gradient-based methods, and thus, suffers from the same pitfalls: neural network models have noisy gradients [5, 7]. 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 [45, 46, 32, 47]. In this section, we briefly recall the classical total Sobol indices and how to apply it to our problem. The complete derivation of the Sobol-Hoeffding decomposition is presented in the appendix D.

Formally, we place ourselves at layer l and perform our Concept Activation Factorization, providing us with $\boldsymbol{U}, \boldsymbol{W}$. A natural way to estimate the importance of a concept $\boldsymbol{U}_i$ is to measure the fluctuation of the model's output $\boldsymbol{h}_{l,k}(\boldsymbol{U}\boldsymbol{W}^T)$ in response to meaningful perturbations of the concept coefficient $\boldsymbol{U}_i$. Concretely, with $\boldsymbol{M} = (M_1, ..., M_r) \in [0,1]^r$, here an i.i.d sequence of real-valued random variables, we introduce a concept fluctuation to reconstruct a perturbated activation $\tilde{\boldsymbol{A}} = (\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T$ (e.g., the masks can be used to put a concept value to zero). We can then propagate this perturbated activation to the model output $\boldsymbol{Y} = \boldsymbol{h}_{l,k}(\tilde{\boldsymbol{A}})$. Thus, an important concept will have a large variance on the model output while an unused concept will barely change it.

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 indices of the masks except the $\boldsymbol{M}_i$ were to be fixed. This yields the general definition of the Total Sobol indices.

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

$$\mathcal{S}_{T_i} = \frac{\mathbb{E}_{\boldsymbol{M}_{\sim i}}(\mathbb{V}_{M_i}(\boldsymbol{Y}|\boldsymbol{M}_{\sim i}))}{\mathbb{V}(\boldsymbol{Y})} = \frac{\mathbb{E}_{\boldsymbol{M}_{\sim i}}(\mathbb{V}_{M_i}(\boldsymbol{h}_{l,k}((\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T)|\boldsymbol{M}_{\sim i}))}{\mathbb{V}((\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T)} \quad (2)$$

In a practical way, this index can be calculated efficiently [48, 49, 50, 51, 52], more details on the sampling (Quasi-Monte Carlo) and the estimator used are left in appendix D.

3.4 Ingredient 3: Unlocking *Concept Attribution Map*

Attribution methods are useful for determining the regions deemed important by the model for the decision, but they lack the information about what exactly triggered it. We have seen that we can already extract this information from the matrices $\boldsymbol{U}$ and $\boldsymbol{W}$, but as it is, we cannot know to which part of the image the model associates each concept, and thus, better comprehend the model's decisions. In this section, we will show how we can unlock the set of attribution methods (forward and backward mode) to find where a concept is located in the input image (see Fig.1). Forward attribution methods don't rely on gradients and only use inference information, whereas backward methods require to back-propagate through the network's layers. By application of the chain rule, computing $\frac{\partial \boldsymbol{U}}{\partial x}$ requires access to $\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{A}}$.

To do so, it could be tempting to solve the linear system $\boldsymbol{U}\boldsymbol{W}^T = \boldsymbol{A}$. However, this problem is ill-posed since $\boldsymbol{W}^T$ is low rank. A standard approach is to calculate the Moore-Penrose pseudo-inverse $(\boldsymbol{W}^T)^+$, which solves rank deficient systems by looking at the minimum norm solution [53]. In practice $(\boldsymbol{W}^T)^+$ is computed with the Singular Value Decomposition (SVD) of $\boldsymbol{W}^T$. Unfortunately, SVD is also the solution to the *unstructured minimization* of $\frac{1}{2}\|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T\|_F^2$ by the Eckart–Young–Mirsky theorem [54]. Hence, the non negativity constraints – i.e $\boldsymbol{U} \geq 0, \boldsymbol{W} \geq 0$ – of the NMF are ignored, which prevents approaches based on solving $\boldsymbol{U}^T\boldsymbol{W} = \boldsymbol{A}^T$ from succeeding. Other issues stem from the fact that the $\boldsymbol{U}, \boldsymbol{W}$ decomposition is generally not unique.

Our third contribution consists on tackling this problem to allow the use of attribution methods – i.e. *Concept Attribution Maps* – by proposing a strategy to differentiate through the NMF layer.

Implicit differentiation of NMF layers The NMF problem 1 is NP-hard [55], and it is not convex with respect to the input pair $(\boldsymbol{U}, \boldsymbol{W})$. However, fixing the value of one of the two factors and

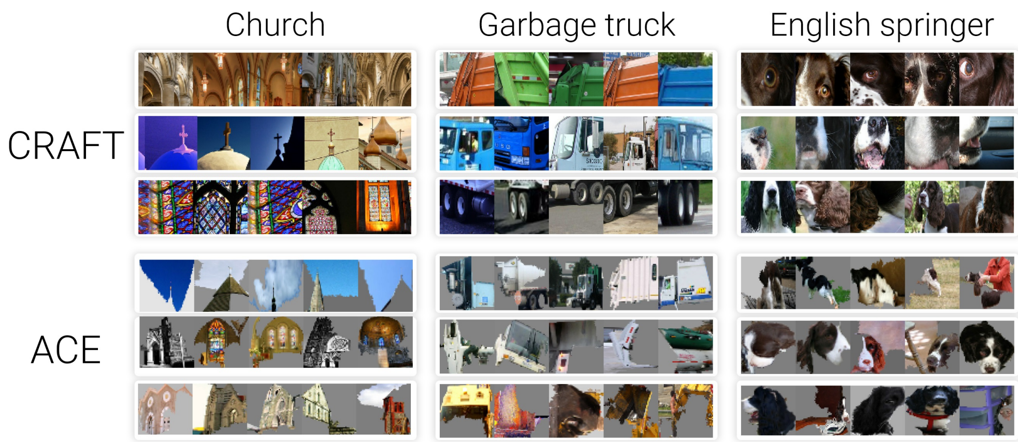

Figure 4: **Qualitative comparison.** We compare concepts found by our method (top) to those extracted with ACE [23] (bottom) for the classes *Church*, *Garbage truck* and *English springer* from ILSVRC2012 [27].

optimizing the other turns the NMF formulation into a pair of Non Negative Least Squares (NNLS) problems (see Equation 3), which are convex. This ensures that alternating minimization (a standard approach for NMF) of $(\boldsymbol{U}, \boldsymbol{W})$ factors will eventually reach a local (and global) minimum:

$$\boldsymbol{U}_{t+1} = \underset{\boldsymbol{U} \geq 0}{\arg\min} \, \frac{1}{2} \|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}_t^T\|_F^2 \qquad \boldsymbol{W}_{t+1} = \underset{\boldsymbol{W} \geq 0}{\arg\min} \, \frac{1}{2} \|\boldsymbol{A} - \boldsymbol{U}_t\boldsymbol{W}^T\|_F^2 \tag{3}$$

Each of the NNLS problem fulfills the KKT conditions[56, 57], which can be encoded in the so-called *optimality function* $\boldsymbol{F}$, see Equation 10 Appendix C.2. The implicit function theorem [39] allows us to use implicit differentiation [38, 39, 58] to efficiently compute the Jacobians $\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{A}}$ and $\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{A}}$ without requiring to back-propagate through each of the iterations of the NMF solver:

$$\frac{\partial(\boldsymbol{U}, \boldsymbol{W}, \bar{\boldsymbol{U}}, \bar{\boldsymbol{W}})}{\partial \boldsymbol{A}} = -(\partial_1 \boldsymbol{F})^{-1} \partial_2 \boldsymbol{F} \tag{4}$$

However, this requires the dual variables $\bar{\boldsymbol{U}}$ and $\bar{\boldsymbol{W}}$, which are not computed by Scikit-learn's [59] popular implementation[†]. Consequently, we leverage the work of [62] and we re-implement our own solver with Jaxopt [40] based on ADMM [63], a GPU friendly algorithm (see Appendix C.2).

We start by performing Concept Activations Factorization – i.e we precompute the concept bank $\boldsymbol{W}$ by solving the NMF. Concept Attribution Maps of a new input $\boldsymbol{x}$ are calculated by solving the NNLS problem $\min_{\boldsymbol{U} \geq 0} \frac{1}{2} \|\boldsymbol{h}_l(\boldsymbol{x}) - \boldsymbol{U}\boldsymbol{W}^T\|_F^2$. The implicit differentiation of NMF layer $\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{A}}$ is integrated into classical back-propagation to obtain $\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{x}}$. Most interestingly, this technical advance unlocks all white-box explainability methods [5, 6, 7, 8, 9, 12] to generate concept-wise attribution maps and trace the part of the image that triggered the detection of the concept. Additionally, it is even possible to employ black-box methods [4, 13, 26, 14] since it only amounts to solving an NNLS problem.

4 Experimental evaluation

We used CRAFT to explain a ResNet50V2 trained on the ILSRVC2012 [27] data set (ImageNet). We selected a subset of 10 classes, each containing 1000 images (those recommended by ImageNette[‡]). In all of our experiments, $r = 25$, like in [23] and the cropping function $\boldsymbol{\tau}$ consists on randomly choosing 10 square 64×64 patches for each image.We start by qualitatively validating CRAFT by showing that: (1) the method yields concepts that are easy to interpret (see Fig. 4), (2) the combination of local and global explanations allows to explain complex failure cases otherwise unexplainable with only the attribution methods (see Fig. 5). Then, we validate independently the new ingredients brought by the method by showing quantitatively that (3) recursivity allows us to refine concepts, making them more meaningful to humans with the help of two psychophysics experiments, and (4) Sobol indices allow for a better estimation of concept importance. Additional experiments, including

[†]Scikit-learn uses a Block coordinate descent algorithm [60, 61], with a randomized SVD initialization.
[‡]https://github.com/fastai/imagenette

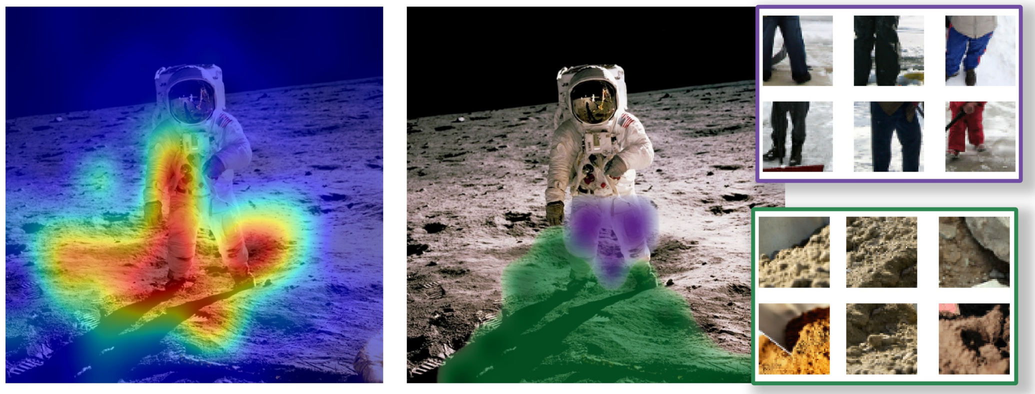

Figure 5: **This is a Shovel.** We compare a heatmap generated by RISE [13] (left) with the *Concept Attribution Maps* generated with our implicit differentiation pipeline and Grad-CAM (right) on the explanations of the two most influential concepts that drove the ResNet50's decision. We found a first concept that seems to be associated with textures of dirt commonly found in the images of the class *Shovel*. The second concept elucidated by CRAFT is located on the astronaut's pants, which he confuses with the ski suits of people clearing snow from their driveway with a shovel.

a sanity check and an example of activation maximization (Deep dream) on the concept bank, as well as many other examples of local explanations for randomly picked images from ILSVRC2012, are included in appendix B.

We leave a discussion on the limitations of this method and on the broader impact in appendix A.

4.1 Example of CRAFT concepts

Figure 4 compares the examples of concepts found by CRAFT against those found by ACE [23] for 3 classes of Imagenet. For each class the concepts are ordered by importance (the highest being the most important). ACE uses a clustering technique and TCAV to estimate importance, while CRAFT uses the method introduced in 3 and Sobol to estimate importance. These examples illustrate one of the weaknesses of ACE: the segmentation used can introduce biases through the baseline value used [64, 10]. The concepts found by CRAFT seem distinct: (vault, cross, stained glass) for the Church class, (dumpster, truck door, two-wheeler) for the garbage truck, and (eyes, nose, fluffy ears) for the English Springer. More examples can be found in the appendix.

4.2 Explaining complex failure cases

One of the goals of explainability is to explain the failure cases of the models studied. Figure 5 shows an example of an incorrect prediction: the model in question – here still a ResNet50 – predicts 'shovel'. Moreover, the attribution method on the left – here RISE [13] – does not tell us much except that the evidence for shovel seems to be located at the level of the ground and the lower torso and legs of the astronaut. With CRAFT, we can however study the concepts found by the model at these locations. There are two of them: the first concept in green, aims at the lunar ground and refers to the rocks often seen next to shovels in the dataset. The second concept in purple is aimed at the legs of the astronaut and refers to the legs of a person, often in a ski suit, which he takes for the astronaut's.

	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 1: **Results from the psychophysics experiments.**

4.3 Validation of Recursivity

To evaluate the meaningfulness of the extracted high-level concepts, we performed psychophysic experiments with human subjects, to whom we requested 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 in computer vision. Both groups of participants were volunteers and didn't receive any monetary compensation. Some examples of the developed interface are available the appendix E.

Intruder detection experiment, we make users identify the intruder out of a series of five segments belonging to a certain class, with the odd one being taken from a different concept but from the same class. Now, we compare the results of this intruder detection with 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.2) extracted using our recursive method. If the concept (or sub-concept) is meaningful, then it should be easy for the users to find the intruder. Table 1 summarizes our results, showing that indeed both concepts and sub-concepts are meaningful, and that recursivity can lead to a slightly higher understanding of the generated concepts (significant for non-experts, not significant for experts) and might suggest a way to make concepts more interpretable.

Binary choice experiment, In order to test the improvement of the meaningfulness of the sub-concept generated with recursivity with respect to the larger parent concept, we showed participants a segment belonging to a subcluster and to the parent cluster (e.g., $\boldsymbol{\tau}(\boldsymbol{x}) \subset \mathcal{C}_{11} \subset \mathcal{C}_1$) without specifying why those images are grouped together. We then we asked 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 1 are satisfying, since in both the expert and non-expert groups, the participants chose the sub-cluster by more than 74% of the times. 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 $p < 0.001$).

4.4 Fidelity analysis

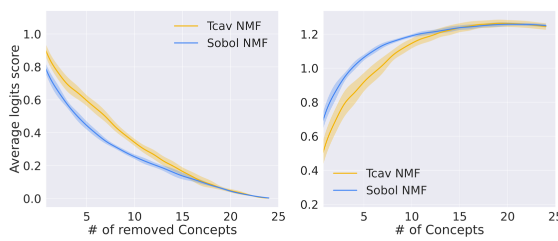

Figure 6: **(Left)** Deletion curve (lower is better). **(Right)** Insertion curves (higher is better). Whether in deletion or insertion, the score – calculated on more than 100,000 images – shows that using Sobol indices yield to better estimates of important concepts.

We propose to simultaneously verify that the concepts are faithful to the model and that the concept importance estimator performs better than TCAV [22] by using the fidelity metrics introduced in [23, 24]. These metrics are similar to the one used for attribution methods, which consist on studying the change of the logit score when removing/adding pixels considered important. Nevertheless, we do not make these modifications in the pixel space but in the concept space: once $\boldsymbol{U}, \boldsymbol{W}$ are computed, we reconstruct the matrix $\boldsymbol{A} \approx \boldsymbol{U}\boldsymbol{W}^T$ using only the most important concept (or removing the most important concept for deletion), and study the score in output of the model. As can be seen from Fig. 6, ranking the extracted concepts using Sobol's importance score results in much steeper curves than when they are sorted by their TCAV scores. We confirm these results with other matrix factorization techniques (PCA, ICA, RCA) in the Appendix F.

5 Conclusion

In this paper, we introduced a method for automatically extracting human-scrutable concepts from Deep Neural Network: CRAFT. Our method allows to explain a pre-trained model both in a per-class and per-instance basis by highlighting both *what* the model saw when predicting the class label and *where* it is located, which, as we have shown, exhibits complementary benefits. The approach relies on three novel ingredients: 1) exploiting the recursive nature of the feature extraction chains in CNNs to find decompositions where each concept is clearly understandable; 2) measuring concept importance through Sobol indices to more accurately identify which concepts influence a model's decision for a given class; and 3) harnessing implicit differentiation to backpropagate through NMF blocks, thus enabling the use of any attribution method to create concept-wise local explanations that we call *Concept Attribution Maps*. Human experiments confirmed the validity of the approach and that concepts identified by CRAFT are meaningful. We hope that this work will guide further efforts in the search for concept-based explainability methods and that further connections between local and global explanations will be made.

References

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

[2] Mauritz Kop. Eu artificial intelligence act: The european approach to ai. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust . . . , 2021. 1

[3] Christoph Torens, Umut Durak, and Johann C Dauer. Guidelines and regulatory framework for machine learning in aviation. In *AIAA Scitech 2022 Forum*, page 1132, 2022. 1

[4] 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. 1, 2, 3, 7, 24

[5] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1, 2, 6, 7, 28

[6] 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, 2, 7

[7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 2, 6, 7

[8] 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, 2, 7

[9] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 1, 2, 7

[10] 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. 1, 2, 8

[11] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019. 1, 2

[12] 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. 1, 2, 7

[13] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 1, 2, 3, 7, 8, 24, 26

[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, 2, 3, 7, 24

[15] Thomas Fel, Melanie Ducoffe, David Vigouroux, Remi Cadene, Mikael Capelle, Claire Nicodeme, and Thomas Serre. Don't lie to me! robust and efficient explainability with verified perturbation analysis. *arXiv preprint arXiv:2202.07728*, 2022. 1

[16] Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *arXiv preprint arXiv:2112.04417*, 2021. 2

[17] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. https://distill.pub/2018/building-blocks. 2, 3

[18] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization. 2, 3, 16, 21

[19] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017. 2

[20] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020. 2

[21] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31, 2018. 2

[22] 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, 6, 9, 21, 26

[23] 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, 9, 26

[24] 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, 9, 26

[25] 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. 2

[26] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2, 4, 7

[27] 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. 3, 4, 7, 26, 27

[28] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015. 2

[29] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2

[30] 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, 28

[31] 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

[32] 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. 3, 6

[33] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020. 3

[34] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 3

[35] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836, 2018. 3

[36] 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. 3, 16

[37] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020. 4

[38] Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002. 4, 7, 24

[39] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008. 4, 7, 24

[40] 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. 4, 7, 24

[41] 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. 5

[42] 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

[43] 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

[44] 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

[45] 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. 6

[46] Ilya M Sobol. Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment*, 1:407–414, 1993. 6, 25

[47] 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. 6

[48] 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. 6

[49] 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. 6

[50] 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. 6, 25

[51] Art B Owen. Better estimation of small sobol'sensitivity indices. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(2):1–17, 2013. 6

[52] 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. 6

[53] 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

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

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

[56] William Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939. 7, 23

[57] 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. 7, 23

[58] 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. 7, 24

[59] 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. 7

[60] 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. 7, 24

[61] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011. 7, 24

[62] 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. 7, 23

[63] 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. 7, 16, 23, 24

[64] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020. 8

[65] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. *arXiv preprint arXiv:2103.06191*, 2021. 14

[66] 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. 16

[67] 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. 21

[68] 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. 23, 24

[69] 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. 26

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] As a URL
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For the experiments comparing TCAV scores to our concept importance score based on Sobol indices
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The ILSVRC2012 dataset contain personally identifiable information [65]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Screenshot of the experiments are in the appendix
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]