Multiple Different Black Box Explanations for Image Classifiers

Hana Chockler[a], David A. Kelly[a] and Daniel Kroening[b]

[a]King's College London, UK
[b]Amazon, USA

Abstract. Existing explanation tools for image classifiers usually give only a single explanation for an image's classification. For many images, however, image classifiers accept more than one explanation for the image label. These explanations are useful for analyzing the decision process of the classifier and for detecting errors. Thus, restricting the number of explanations to just one severely limits insight into the behavior of the classifier. In this paper, we describe an algorithm and a tool, MultiReX, for computing multiple explanations as the output of a black-box image classifier for a given image. Our algorithm uses a principled approach based on actual causality. We analyze its theoretical complexity and evaluate MultiReX against the state-of-the-art across three different models and three different datasets. We find that MultiReX finds more explanations and that these explanations are of higher quality.

1 Introduction

AI models are now a primary building block of most computer vision systems. The opacity of some of these models (e.g., neural networks) creates demand for explainability techniques, which attempt to provide insight into why a particular input yields a particular observed output. Beyond increasing a user's confidence in the output, and hence also their trust in the AI model, these insights help to uncover subtle classification errors that are not detectable from the output alone [10].

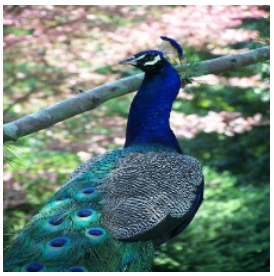
Existing explainability tools use a variety of definitions for explanations, often tied to a particular method of extracting them. The definition we use here is grounded in the theory of actual causality and, roughly speaking, defines an explanation as a smallest part of the input image that is sufficient for the classifier to yield the same top label as the original image (see Section 3). If the top classification for the model is, say, 'peacock' (Figure 1), then a causal explanation is a subset of image pixels also labeled 'peacock' as the top classification. Explanations need not be unique. A human could point to many parts of the image of a peacock and state that this part is sufficient to label the entire image 'peacock'. There is no reason, a priori, to assume that a model cannot do the same.

Why do we need to have multiple explanations? Previous work, using a user study, demonstrated that people value having multiple explanations of the model's decisions, as it increases their confidence in the classification and gives them more insight into the reasoning process of the model [35]. The increase in confidence and trust is reaffirmed in [21, 28]. Even, perhaps, more importantly, as we show in this paper, the prevalence of multiple explanations suggests that algorithms for computing more than one explanation are essential for understanding the reasoning process of image classifiers and uncovering subtle classification errors.

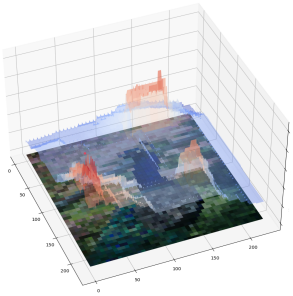
To illustrate the latter point, the image in Figure 2 is classified by a ResNet50 as 'tennis racket'. Figures 2b and 2c both show that sections of the image that overlap with the racket are sufficient for the overall classification. Figure 2d, however, shows a part of the player's shorts as being sufficient for the 'tennis racket' classification, with *higher* confidence (0.25) that for Figures 2b and 2c, which are both around 0.23. This is a concerning finding for the users of the model.

Most existing techniques provide only one explanation, potentially missing the error. The one notable exception is SAG (Structured Attention Graphs) [35], a pioneering work in the exploration of multiple explanations for image classifiers. However, SAG suffers from a number of shortcomings. First, it is not a true black-box tool, as it requires access to the gradient of the model. Such tools might more accurately be called *grey-box*, as a proprietary model may not expose the gradient during inference. Furthermore, SAG's concept of explanation is rather different from what a human might accept as an explanation, and is very different from a causal explanation. SAG is liberal in what it considers an explanation, resulting in potentially thousands of them found for a single image. [25] discovered this when investigating "compositionality", which they define as a conjunction of parts of an image (patches) that have high likelihood ratios for a particular classification. Intuitively though, it is unlikely that an image classified as, say 'dog', has thousands of explanations: it happens because SAG's explanation refers to *any* element in the output tensor above a predefined probability threshold. We argue that this renders SAG's results uninformative. Indeed, with a suitable threshold, *any* combination of patches is sufficient for the desired classification. Causality, on the other hand, says that a set of pixels explains label 'a' if that set is sufficient to be the actual (top) classification of the model. This is far stricter than SAG's explanations.

To overcome these problems, we present MultiReX, a black-box algorithm and a tool for computing multiple explanations for image classifiers. Using the theory of actual causality, MultiReX computes a causal responsibility ranking of the pixels of the image, from which it extracts multiple different explanations. Unlike with SAG, MultiReX is not fixed to a rigid grid, so its explanation discovery is much more flexible (Figure 1). MultiReX also allows an optional confidence threshold: unlike with SAG, where the threshold dictates what constitutes an ex-



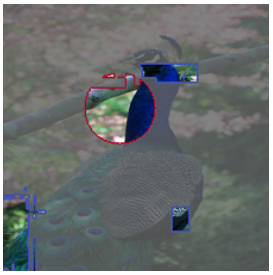
(a) Class 84: peacock



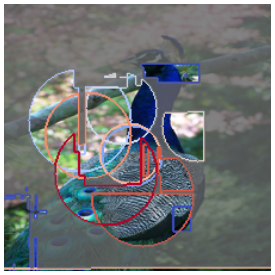
(b) Responsibility Map

(c) Explanations with no overlap

(d) Explanations with partial overlap



(e) Explanations with high confidence



(f) High confidence with overlap

Figure 1: MultiReX on a peacock. MultiReX computes a responsibility landscape (Figure 1b): this landscape encodes so much information about the image that, from it, we can compute non-overlapping explanations (Figure 1c, partially overlapping to any degree (Figure 1d), explanations which have higher confidence than the original image (Figure 1e and explanations with higher confidence than the original image, and total permissible overlap (Figure 1f).

planation, MultiReX's threshold affects the *quality* of an explanation (Figure 1e), picking out subsets of pixels which can have even higher confidence than the entire image. Furthermore, the explanations produced by MultiReX are *actual*, that is, they always explain the top (the most likely) classification of the input image. This is in contrast to SAG, which produces explanations that do not necessarily correspond to the top classification (see Section 6). In Section 4, we also present an exponential upper bound on the number of possible explanations and demonstrate that this bound is tight. In Section 6 we experimentally compare MultiReX with SAG on standard benchmarks. We show that MultiReX consistently finds more explanations than SAG. We also show that MultiReX performs well when we add in a confidence threshold to increase explanation quality.

We provide the details of the benchmark sets, the models, and the main results in the paper. The MultiReX tool is incorporated into ReX on GitHub at https://github.com/ReX-XAI/ and is on PyPI as https://pypi.org/project/rex-xai/. Please see the full version at [9] for the full set of reported as well as additional results.

2 Related Work

There is a large body of work on algorithms for computing one explanation for a given output of an image classifier. They can be largely grouped into white-box and black-box methods. White-box methods frequently use variations on propagation-based explanation methods to back-propagate a model's decision to the input layer to determine the weight of each input feature for the decision [38, 39, 2, 36, 29]. GradCam, a white-box technique which has spawned many variants, only needs one backward pass and propagates the class-specific gradient into the final convolutional layer of a DNN to coarsely highlight important regions of an input image [33].

Perturbation-based explanation approaches introduce perturbations to the input space directly in search for an explanation. These are typically found in black-box explanation methods. SHAP (SHapley Additive exPlanations) computes Shapley values of different parts of the input and uses them to rank the features of the input according to their importance [26]. LIME constructs a simple model to label the original input and its neighborhood of perturbed images and uses this model to estimate the importance of different parts of the input [32, 12, 5, 31, 16]. Finally, ReX [10] ranks elements of the image according to their responsibility for the classification and uses this ranking to greedily construct a small explanation. The ReX ranking procedure is based on an approximate computation of causal responsibility. None of these black-box tools provide multiple explanations of the same image.

Work on calculating more than one explanation for a given classification outcome is in its infancy. A recent paper on abductive explanations raises the problem of generating only one explanation and suggests *aggregating* multiple explanations to obtain the full information about the importance of different features of the input [4]. However, in the absence of a tool for reliably and systematically generating multiple explanations, they rely on re-executing LIME and SHAP multiple times, in the hope that their inherent non-determinism results in different explanations. This is clearly not a systematic approach.

To the best of our knowledge, there is only one algorithm and tool that specifically computes multiple explanations of image classifiers—SAG, described in [35]. We describe SAG's algorithm in more detail in Section 6 and argue that its definition of explanation is too permissive. Our experimental results show that, even with our stricter definition, we find more explanations than SAG.

Finally, we mention a growing body of work on logic-based explanations [24, 27, 11], where a symbolic encoding of the model is given. Their notion of abductive explanations is similar in spirit to the one used in this paper, except all possible values of pixels outside of the explanations are considered. The problem setting is very different from ours, considering the model as a logic formula. In contrast, our

approach is black box and is agnostic to the internal structure of the classifier, nor does it try to represent its decision process as a logic expression.

3 Background on Actual Causality

In this section we briefly review the definitions of causality and causal models introduced by Halpern and Pearl [20] and relevant definitions of causes and explanations in image classification [7]. The reader is referred to [19] for further reading.

We assume that the world is described in terms of variables and their values. Some variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. The structural equations describe how these values are determined.

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *(modifiable) structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y *ranges*). For simplicity, we assume here that $\mathcal{V}$ is finite, as is $\mathcal{R}(Y)$ for every endogenous variable $Y \in \mathcal{V}$. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X (i.e., $F_X = \mathcal{F}(X)$) such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$.

The structural equations define what happens in the presence of external interventions. Setting the value of some variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{X \leftarrow x}$, which is identical to M, except that the equation for X in $\mathcal{F}$ is replaced by $X = x$.

Probabilistic causal models are pairs $(M, \Pr)$, where M is a causal model and $\Pr$ is a probability on the contexts. A causal model M is *recursive* (or *acyclic*) if its causal graph is acyclic. If M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, the values of all the other variables are determined. In this paper we restrict to recursive models.

We call a pair $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$ a *(causal) setting*. A causal formula ψ is true or false in a setting. We write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \vec{u})$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique solution to the equations in M in context $\vec{u}$. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y}=\vec{y}}, \vec{u}) \models \varphi$, where $M_{\vec{Y} \leftarrow \vec{y}}$ is the causal model that is identical to M, except that the variables in $\vec{Y}$ are set to $Y = y$ for each $Y \in \vec{Y}$ and its corresponding value $y \in \vec{y}$.

A standard use of causal models is to define *actual causation*: that is, what it means for some particular event that occurred to cause another particular event. There have been a number of definitions of actual causation given for acyclic models (e.g., [3, 17, 18, 20, 19, 22, 23, 40, 41]). In this paper, we focus on what has become known as the *modified* Halpern-Pearl definition and some related definitions introduced by Halpern in 2019. We briefly review the relevant definitions below. The events that can be causes are arbitrary conjunctions of primitive events (formulas of the form $X = x$); the events that can be caused are primitive events, denoting the output of the model.

Definition 1 (Actual cause). $\vec{X} = \vec{x}$ *is an* actual cause *of* φ *in* $(M, \vec{u})$ *if the following three conditions hold:*

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ *and* $(M, \vec{u}) \models \varphi$.

AC2. *There is a a setting* $\vec{x}'$ *of the variables in* $\vec{X}$*, a (possibly empty) set* $\vec{W}$ *of variables in* $\mathcal{V} - \vec{X}'$*, and a setting* $\vec{w}$ *of the variables in* $\vec{W}$ *such that* $(M, \vec{u}) \models \vec{W} = \vec{w}$ *and* $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$*, and moreover*

AC3. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}''$ *can replace* $\vec{X} = \vec{x}'$ *in AC2, where* $\vec{x}''$ *is the restriction of* $\vec{x}'$ *to the variables in* $\vec{X}'$.

In the special case that $\vec{W} = \emptyset$, we get the but-for definition.

The notion of explanation, taken from [19], is relative to a set of contexts.

Definition 2 (Explanation). $\vec{X} = \vec{x}$ *is an* explanation *of* φ *relative to a set* $\mathcal{K}$ *of contexts in a causal model* M *if the following conditions hold:*

EX1a. *If* $\vec{u} \in \mathcal{K}$ *and* $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$*, then there exists a conjunct* $X = x$ *of* $\vec{X} = \vec{x}$ *and a (possibly empty) conjunction* $\vec{Y} = \vec{y}$ *such that* $X = x \wedge \vec{Y} = \vec{y}$ *is an actual cause of* φ *in* $(M, \vec{u})$.

EX1b. $(M, \vec{u}') \models [\vec{X} = \vec{x}]\varphi$ *for all contexts* $\vec{u}' \in \mathcal{K}$.

EX2. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}'$ *satisfies EX1, where* $\vec{x}'$ *is the restriction of* $\vec{x}$ *to the variables in* $\vec{X}'$*. (This is SC4).*

EX3. $(M, u) \models \vec{X} = \vec{x} \wedge \varphi$ *for some* $u \in \mathcal{K}$.

4 Theoretical Foundations of MultiRex

We view an image classifier (*e.g.* a neural network) as a probabilistic causal model. Specifically, the endogenous variables are taken to be the set $\vec{V}$ of pixels that the image classifier gets as input, together with an output variable that we call O. The variable $V_i \in \vec{V}$ describes the color and intensity of pixel i; its value is determined by the exogenous variables. The equation for O determines the output of the model as a function of the pixel values. Thus, the causal network has depth 2, with the exogenous variables determining the feature variables, and the feature variables determining the output variable. In this paper we also assume *causal independence* between the feature variables in $\vec{V}$, the set of pixels of the input image.

Pixel independence is a common assumption in explainability tools. This is a non-trivial assumption, and it might seem far-fetched, especially if we consider images capturing real objects: if a group of pixels captures, say, a cat's ear, then a group of pixels near it should capture a cat's eye. However, we argue that it is, in fact, accurate on images. Indeed, consider a partially obscured image, obtained by overlaying random color patches over an Imagenet image. Such an image is perfectly valid, as obscuring a part of the input image either by introducing an artificial object or by positioning a real object in front of the primary subject of the classification does not lead to any change in unobscured pixels. For example, obscuring a cat's ear does not lead to any change in an (unobscured) group of pixels, capturing a cat's eye. Thus, pixel independence holds on general images.

We refer the reader to [7] for a more in-depth discussion of causal independence between pixel values in image classification. We note that the causal independence assumption is not true in other types of inputs, such as tabular or spectral data. For those types of inputs, assuming independence is clearly an approximation and might lead to inaccurate results. In this paper, however, we focus on images, where causal independence between pixels holds.

(a) Class 752: racket (b) (c) (d)

Figure 2: Imagenet class 752: racket, according to ResNet50. Figures 2b to 2d show 3 minimal, sufficient explanations for class 752. Only Figures 2b and 2c contains part of the racket. The tennis players shorts are also classified as racket, with a higher confidence than either Figure 2b or Figure 2c.

Moreover, as the causal network is of depth 2, all parents of the output variable O are contained in $\vec{V}$. Given these assumptions, the probability on contexts directly corresponds to the probability on seeing various images (which the model presumably learns during training).

Given an input image I, the set of contexts $\mathcal{K}$ that we consider for an explanation is the set $\mathcal{K}_I$ obtained by *all partial occlusions of* I, where an occlusion sets a part $\vec{Y}$ of the image to a predefined masking color $\vec{y}$. The probability distribution over $\mathcal{K}_I$ is assumed to be uniform.

Under the assumptions above, the following definition is equivalent to Definition 2, as shown in [7].

Definition 3 (Explanation for image classification [8])**.** *An explanation is a minimal subset of pixels of a given input image that is sufficient for the model $\mathcal{N}$ to classify the image, where "sufficient" is defined as containing only this subset of pixels from the original image, with the other pixels set to the masking color.*

In the complexity discussion that follows, we discuss the complexity of decision problems matching the function problems of computing explanations. If the decision problem is A-complete, for some complexity class A, then the matching function problem is $\text{FP}^{A[\log n]}$-complete, assuming monotonicity of the function problem (see [6] for a more in-depth discussion of decision vs function problems in actual causality). Formally, the decision problem of explanation is, given a model, a context, an output φ, and a candidate explanation, to decide whether this candidate is indeed an explanation for φ in the given model and context.

[8] observe that the precise computation of an explanation in our setting is intractable, as the problem is equivalent to an earlier definition of explanations in binary causal models, which is DP-complete [14].[1]

The following lemma shows that computing a second (or any subsequent) explanation is not easier than computing the first one.

Lemma 1. *Given an input image and one explanation, the decision problem of finding a different explanation is DP-complete.*

Proof. Membership in DP is straightforward and follows from the membership in DP of the problem of deciding an explanation: adding a constraint that the output should be different from a given explanation does not increase the complexity class of the decision problem. For hardness in DP, we show a reduction from the decision problem of computing an explanation. Given an image I classified by a neural network (a black-box model) $\mathcal{N}$ as $\mathcal{N}(I)$, we define φ as "the output is an explanation of $\mathcal{N}(I)$ or it is exactly I". The reduction is a tuple $\langle I, \mathcal{N}(I), I\rangle$, viewed as an input to the different explanation problem, that is, $\langle$ input image, its label, a given explanation $\rangle$. The second I renders φ true. Then, an output that renders φ true and is different from I is an explanation of $\mathcal{N}(I)$ for I, completing the reduction. □

[1] DP is the class of languages that are an intersection of a language in NP and a language in co-NP and contains, in particular, the languages of unique solutions to NP-complete problems [30].

Chockler et al. [10] use a greedy approach to constructing approximate explanations, based on scanning the ranked list of pixels *pixel_ranking* (Figure 3). The pixels are ranked in the order of their approximate *degree of responsibility* for the classification, where responsibility is a quantitative measure of causality and, roughly speaking, measures the amount of causal influence on the classification. Formally, the *degree of responsibility* of a variable $X = x$ for the value of φ is $1/k$, where k is the size of a smallest set of variables $\vec{X}$ s.t. $X \in \vec{X}$ and has the value x in $\vec{x}$, and $\vec{X} = \vec{x}$ is an actual cause of φ according to Definition 1 [6]. The degree of responsibility is always between 0 and 1, with higher values indicating a stronger causal influence.

The precise computation of degrees of responsibility of pixels of I is intractable; its decision problem is NP-complete under our simplifying assumptions [6]. Hence, the ranking in [10] is based on the approximate degree of responsibility, which is computed by partitioning the set in iterations and computing the degrees of responsibility for each partition, while discarding low-responsibility elements (see Section 5 for details). The greedy explanation extraction adds pixels from the sorted ranked list until the original classification is obtained.

However, reducing the complexity of computing one explanation does not reduce the complexity of computing many explanations, as the number of explanations for a given image can be very high:

Lemma 2. *The number of explanations for an input image is bounded from above by $\binom{n}{\lfloor n/2 \rfloor}$, and this bound is tight.*

Proof. Since an explanation of the classification of x is a minimal subset of x that is sufficient to result in the same classification, the number of explanations is characterised by *Sperner's theorem*, which provides a bound for the number S of largest possible families of finite sets, none of which contain any other sets in the family [1]. By Sperner's theorem, $S \leq \binom{n}{\lfloor n/2 \rfloor}$, and the bound is reached when all subsets are of the size $\lfloor n/2 \rfloor$. The following example demonstrates an input on which this bound is reached. Consider a binary classifier $\mathcal{N}$ that determines whether an input image of size n has at least $\lfloor n/2 \rfloor$ green-colored pixels and an input image I that is completely green. Then, each explanation is of size $\lfloor n/2 \rfloor$, and there are $\binom{n}{\lfloor n/2 \rfloor}$ explanations. □

Finally, we note that given a set of explanations (sets of pixels) and an overlap bound, deciding a subset of a given number of explanations in which elements overlap for no more than the bound is NP-hard even assuming that constructing and training a binary classifier is $O(1)$,

by reduction from the independent set problem, which is known to be NP-complete. Indeed, let $(G = \langle V, E\rangle, n)$ be an input to the independent set problem, deciding whether G contains an independent set of nodes of size n. Then, G has an independent set of size n if and only if there exist n disjoint explanations of G having a connected component of size 1 (note that finding a connected component of size 1 is polynomial in the size of G).

5 The MultiReX Algorithm

In this section we present our algorithm for computing multiple explanations of an image. As shown in Section 4, the problem is intractable, motivating the need for efficient and accurate approximation algorithms. Due to the lack of space, some details and algorithms have been moved to the supplementary material.

The concept of a *superpixel* is used in a number of different explanation tools; SAG splits the image into a fixed grid of 7×7 superpixels. Dividing the image in this way greatly reduces the computational cost of searching for explanations. The rigidity of the grid, however, may lead to missing some explanations. MultiReX does not need a rigid grid: it starts with large, randomly selected superpixels which it iteratively refines. Furthermore, MultiReX repeats this procedure with different starting superpixels to reduce the influence of particular random choices. Responsibility landscapes from individual iterations are combined to produce a detailed responsibility landscape (see Figure 1b). As more iterations are added, the landscape becomes smoother. We separate explanations from this landscape using Algorithm 2. It is the existence of this landscape which gives MultiReX its edge: it provides a continuous search space over the entire image that is not confined to a discrete grid.

The high-level structure of the algorithm is presented in Figure 3, and the pseudo-code is in Algorithm 1. We discuss each component in more detail below.

Algorithm 1 MultiReX$(I, \mathcal{N}, r, n, \delta, s, p, q)$

INPUT: an image I, a model $\mathcal{N}$, a searchlight radius r, the maximum number of explanations n, maximum overlap between explanations δ, number of searchlights s, searchlight expansions p, expansion coefficient q
OUTPUT: a set $\mathcal{E}$ of up to n different explanations

1: $\mathcal{E} \leftarrow \emptyset$
2: $l \leftarrow \mathcal{N}(x)$
3: $\mathcal{S} \leftarrow CAUSAL\_RANK(I, \mathcal{N}, l)$
4: **for** i in $0 \ldots s-1$ **do**
5: $\quad E_i \leftarrow searchlight(I, \mathcal{N}, l, \mathcal{S}, r, n, p, q)$
6: $\quad E_i \leftarrow minimize(I, E_i, \mathcal{N}, \mathcal{S})$
7: $\quad \mathcal{E} \leftarrow \mathcal{E} \cup E_i$
8: **end for**
9: $\mathcal{E} \leftarrow separate(\mathcal{E}, \delta)$
10: **return** $\mathcal{E}$

The $CAUSAL\_RANK$ procedure in Line 3 of Algorithm 1 constructs a $pixel\_ranking$, which is a ranking of the pixels of the input image I by their causal responsibility. We use the algorithm in [10] as the basis for producing this landscape. The number of required explanations is given as an input parameter to the procedure, as the total number of explanations can be exponential (see Lemma 2).

The *Searchlight* procedure, called in Line 5, is described in Algorithm 2. It replaces the greedy explanation generation in ReX with a

Algorithm 2 $searchlight\ (I, \mathcal{N}, l, \mathcal{S}, r, n, p, q)$

INPUT: an image I, a model $\mathcal{N}$, a label l, a responsibility landscape $\mathcal{S}$, a searchlight radius r, number of steps n, number of expansions p, radius increase q
OUTPUT: an explanation E

1: $\mathcal{F} \leftarrow initialize(r)$
2: $\mathcal{E} \leftarrow \emptyset$
3: **for** i in $0 \ldots n-1$ **do**
4: $\quad$ **for** j in $0 \ldots p-1$ **do**
5: $\quad\quad l' \leftarrow \mathcal{N}(\mathcal{F}(I))$
6: $\quad\quad$ **if** $l = l'$ **then**
7: $\quad\quad\quad E \leftarrow \mathcal{F}(I)$
8: $\quad\quad\quad$ **return** E
9: $\quad\quad$ **else**
10: $\quad\quad\quad \mathcal{F} \leftarrow expand\_radius(r \times q)$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: $\quad \mathcal{F} \leftarrow neighbor$
14: **end for**
15: **return** $\mathcal{E}$

modified stochastic hill climb. In contrast to most hill-climb-based algorithms that look for the global maximum, we specifically search for *local maxima*, on the assumption that these are likely to correspond to explanations. The function $initialize$ in Line 2 creates a 'searchlight', $\mathcal{F}$, of radius r at a random position over the image I. The intuition is that only the pixels under $\mathcal{F}$ are exposed to the model (all other pixels being set to a baseline value).

If $\mathcal{F}$ contains an explanation (*i.e.* the exposed pixels already have the required classification), we invoke the $minimize$ procedure to remove redundant pixels (Line 6). The reason for potential redundancy is that $\mathcal{F}$ may have included too many pixels by virtue of its circular shape. The $minimize$ procedure consists of obtaining the responsibility ranking of all pixels inside $\mathcal{F}$, setting all those pixels outside $\mathcal{F}$ to responsibility 0, and using the greedy algorithm from [10] to add in pixels based on their responsibility. The stopping condition is the desired classification. This process is guaranteed to succeed because $\mathcal{F}$ already contains *at least* the sufficient pixels. Pixels with 0 responsibility are never added to an explanation.

A randomly placed searchlight might not contain an explanation, either due to its size or to its location. Rather than changing the location of $\mathcal{F}$ immediately, we first increase its size. If we start too small, we might miss all explanations (that is, $\mathcal{F}$ would never be large enough to capture all the necessary pixels). The number of expansions and size of expansion are controlled by hyperparameters.

If increasing the size of $\mathcal{F}$ still does not result in an explanation, the algorithm changes its location and resets to the original size. This step is guided by an objective function; by default, the objective function is the mean of the responsibility of the pixels contained in $\mathcal{F}$. Thus, $\mathcal{F}$ moves towards the areas with a higher average responsibility; these areas are more likely to contain an explanation.

Finally, the $separate$ procedure (Algorithm 3) separates a subset of at most n explanations that overlap on pixels up to the bound δ. δ is a value between 0 and 1, where 0 stands for "no permitted overlap", and 1 means "no overlap restrictions". As discussed in Section 4, the exact solution is NP-hard. The $separate$ procedure uses a greedy heuristic based on the Sørensen–Dice coefficient (SDC) [13, 37], typically used as a measure of similarity between samples. First, we create a list of all pairs in $\mathcal{E}$ which overlap by more that δ. We then iterate backwards

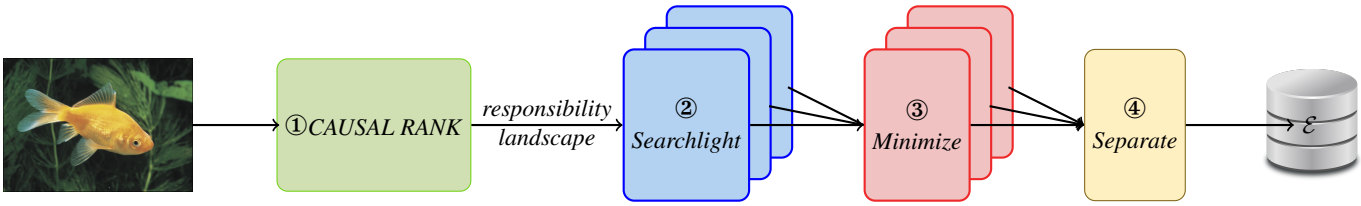


Figure 3: A schematic depiction of MultiReX, returning a set of explanations $\mathcal{E}$ for a given input image. Its components: ① *ranking* generates a responsibility landscape of pixels; ② *search* launches x *searchlight* searches over the landscape; ③ *minimize* minimizes the explanations founds in ②; ④ *separate* produces a maximal subset $\mathcal{E}$ from the output of ③, with the given overlap bound.

Algorithm 3 $separate(\mathcal{E}, \delta)$

INPUT: a set of explanations $\mathcal{E}$, a permitted degree of overlap δ
OUTPUT: a subset of explanations $\mathcal{E}' \subseteq \mathcal{E}$ with overlap at most δ

$all\_pairs \leftarrow \mathcal{E} \times \mathcal{E}$
$bad\_pairs \leftarrow \emptyset$
for (p_i, p_j) in all_pairs **do**
  $SDC \leftarrow dice\_coefficient(p_i, p_j)$
  if $SDC > \delta$ **then**
    $bad\_pairs \leftarrow bad\_pairs \cup (p_i, p_j)$
  end if
end for
for $e \in \mathcal{E}$ **do**
  if e does not contain any bad_pairs **then**
    return e
  end if
end for
return $\emptyset$

through the powerset of $\mathcal{E}$, $2^{\mathcal{E}}$ (*i.e.* starting from the complete $\mathcal{E}$ and not $\emptyset$) and stop at the first set $e \in 2^{\mathcal{E}}$ which does not contain one of the previously discovered 'bad' pairs. As an added optimization, when iterating through $2^{\mathcal{E}}$, we order all subsets e with the same cardinality by the total area of the contained explanations. Thus, the algorithm stops at the largest number of explanations with the smallest overall area and with overlap less that δ.

6 Experimental Results

Implementation We implemented Algorithm 1 in the tool MultiReX for generating multiple causal explanations. Given a responsibility landscape, by default, MultiReX attempts to find 10 explanations (a parameter). In practice, it is computationally inexpensive to find more explanations, but our experimental results demonstrate that images with more than 6 explanations are rare ($\approx 1\%$ of images).

Comparison to SAG As discussed earlier, SAG's definition of explanation is sufficiently different as to make exact comparison difficult. SAG uses a beam search over a 7×7 grid to discover multiple explanations. This strategy results in the search space of size 2^{49} (the number of subsets of regions of the SAG grid), and SAG attempts to solve the exponential explosion problem by reducing the search space *at random*.

Furthermore, SAG accepts a region as an explanation if the confidence of the original label, l, on this region is greater than a confidence bound, regardless of its position in the model's output tensor. That is, SAG might output an explanation to an entirely different label than the top classification of the input. The confidence bound of SAG is based on a hyperparameter 'probability threshold' and confidence on the original image. This is in contrast with MultiReX's much stricter definition of explanation, where a subset of pixels constitute an explanation is they alone are classified l by the model, where l is the top classification, regardless of the model's confidence. In order to perform a fair comparison, we evaluate both tools twice, each time with the default settings of one of them. Namely, we run SAG with default settings (specifically, with the probability threshold 0.9) and compare it against MultiReX where we set a minimum confidence threshold for an explanation at 0.9 as well. Note that is still not an entirely fair comparison, as MultiReX always explains the top classification, contrary to SAG. We also perform the comparative evaluation with setting SAG's probability threshold to 0[2] and leaving MultiReX's minimum confidence threshold at its default value of 0.

We set the maximum number of explanations in both tools to 10. SAG requires the user to set the maximal allowed overlap in grid squares, with suggested values of 0, 1, or 2. MultiReX does not measure overlap in blocks, as its responsibility landscape is continuous (or, rather, discrete at the level of a single pixel). For a fair comparison, we limit both tools to non-overlapping explanations. Apart from these changes, the experiments are performed on MultiReX and SAG with default settings.

Datasets and Models For our experiments, we use 3 different models from TORCHVISION with default weights. The models are 'resnet50', 'convnext_large' and 'vit_b_32'. We use 3 different publicly available datasets: ImageNet-mini validation[3], Pascal VOC2012 [15] and ECSSD [34].

Experimental Results A natural and quantifiable performance measure for multiple explanations is the number of significantly different explanations produced for each image. We also consider, for SAG, the position of the explanation in the model output (for MultiReX, this position is always 0, as MultiReX always explains the actual (top) classification). We also consider the size of explanations: in general, a good explanation should be close to minimal, *i.e.* having as few extraneous pixels as possible. This has an important bearing on multiple explanations if we do not have overlap: one cannot fit many large explanations in a standard-size image.

The experiment was performed on a 64-core machine running Ubuntu 20.04.6 with a 48GB RAM and several Nvidia A40 GPUs. For the sake of space, we present complete results for ResNet50 in the paper and other results in the supplementary material. The results for the other models follow the same pattern as for ResNet50. The

[2] In reality, we found setting it to 0.0 precisely resulted in no explanations, so for our experiment we used the value 0.01
[3] https://www.kaggle.com/datasets/ifigotin/imagenetmini-1000

No. Exp	Datasets					
	ImageNet1k		Voc		ECSSD	
	MULT	SAG	MULT	SAG	MULT	SAG
1	2052	2757	931	1138	548	732
2	1216	733	376	223	315	185
3	489	246	125	64	106	53
4	135	103	15	13	25	15
5	30	42	2	5	3	9
6	1	18	–	2	3	3
7	–	10	–	2	–	1
8+	–	14	–	2	–	2

(a) Default SAG and MULTIREX with at least 90% of original confidence and for the top classification. More than 1 explanation is good.

No. Exp	Datasets					
	ImageNet1k		Voc		ECSSD	
	MULT	SAG	MULT	SAG	MULT	SAG
1	333	517	291	263	158	112
2	688	602	423	257	248	150
3	1048	651	411	251	295	147
4	974	540	222	180	171	173
5	571	430	77	142	90	121
6	246	276	22	79	32	66
7	52	223	3	63	6	57
8+	11	684	–	214	–	174

(b) MULTIREX with default parameters; SAG with a probability threshold of 0.01. SAG produces more explanations, but they are not for the top classifications (see Table 2b), rendering this comparison non-informative.

Table 1: SAG and MULTIREX on ResNet50. Table 1a shows SAG with default probability threshold of 0.9. We force MULTIREX to use the same confidence threshold, while still producing the top classification. MULTIREX finds more explanations per image in general than SAG. Table 1b shows the effect of removing SAG's probability threshold, to bring it closer to MULTIREX's behavior. MULTIREX still performs well, but SAG's output is close to noise.

Position	Number of SAG Explanations		
	ImageNet	VOC	ECSSD
0	3698	1221	860
1	181	166	111
2	28	150	22
3	13	7	5
4	2	4	1
5+	1	1	1
Total in position 0 (%)	**94%**	**84%**	**86%**

(a) Position in the output tensor for ResNet50 explanations as produced by SAG at 0.9 probability threshold.

Position	Number of SAG Explanations		
	ImageNet	VOC	ECSSD
0	94	20	13
1	69	33	17
2	51	20	11
3	63	17	8
4	82	28	11
5+	3565	1331	940
Total in position 0 (%)	**2%**	**1%**	**1%**

(b) Position in the output tensor for ResNet50 explanations as produced by SAG at 0 probability threshold.

Table 2: Positions of SAG's explanations in the output of a ResNet50 model. The tables illustrate SAG's strong dependence on the probability threshold parameter: when it is 0.9, the majority of SAG's explanations (though not all) are for the top classification, as shown in Table 2a; when the probability threshold is close to 0, SAG's output is for classifications further down on the list, resulting in nonsensical explanations. For values between 0.9 and 0 SAG's output is between these two extremes.

'vit_b_32' model accepts a lower number of multiple explanations for both tools (see supplementary material).

Table 1a shows the comparison of SAG with default parameters against MULTIREX with a confidence threshold of 0.9. Table 2a shows the positions of SAG's explanations in the model output. It seems unreasonable to accept a set of pixels explaining a classification at the 6^{th} position as an explanation for the top classification.

At threshold 0.9, SAG's output mostly (though not always) refers to the top classification, and MULTIREX finds more multiple explanations in general. SAG's dependence on the probability threshold parameter is illustrated in Table 2. With the threshold close to 0, one of the SAG's explanations was in position 831 out of the possible thousand positions for an input in ImageNet.

The average size of an explanation for MULTIREX is $\approx 8\%$ of the input image across the 3 datasets. In contrast, the average size of SAG's explanations when run with 0.9 threshold (hence producing explanations mostly for the top classification) is $\approx 14\%$ of the input image, showing that MULTIREX produces much tighter explanations.

MULTIREX, unlike SAG, is consistent in its explanations: small and the top classification, regardless of probability threshold. If we increase the probability threshold for MULTIREX, the explanations become a little larger. We do not have the problem seen by [25]: a plethora of 'explanations' with little or no explanatory power. It becomes much clearer that the 'compositionality' of the different models is significantly different. In particular, the decision process of ResNet50 is similar to the one of ConvNext, and both are significantly different from that of the ViT model.

Timings MULTIREX takes an average of 15 seconds per image on our machine, for 20 iterations of the main algorithm. SAG, with the default probability threshold 0.9, takes on average 10 seconds; with a threshold of 0.01, SAG is faster, at 2 to 3 seconds, but produces uninformative results.

7 Conclusions

We have introduced MULTIREX, a novel explanation-discovery algorithm based on the responsibility landscape constructed in the ranking procedure and a "spotlight" search, ensuring different spatial locations for explanations.

Acknowledgments

Hana Chockler and David A. Kelly acknowledge support of the UKRI AI programme and the Engineering and Physical Sciences Research Council for CHAI – Causality in Healthcare AI Hub [grant number EP/Y028856/1].

References

[1] I. Anderson. *Combinatorics of Finite Sets*. Oxford University Press, 1987. 4

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS One*, 10(7), 2015. 2

[3] S. Beckers. Causal sufficiency and actual causation. *Journal of Philosophical Logic*, 50:1341–1374, 2021. 3

[4] G. Biradar, Y. Izza, E. A. Lobo, V. Viswanathan, and Y. Zick. Axiomatic aggregations of abductive explanations. In *Thirty-Eighth Conference on Artificial Intelligence, AAAI*, pages 11096–11104. AAAI Press, 2024. 2

[5] J. Chen, L. Song, M. Wainwright, and M. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning (ICML)*, volume 80, pages 882–891. PMLR, 2018. 2

[6] H. Chockler and J. Y. Halpern. Responsibility and blame: A structural-model approach. *J. Artif. Intell. Res.*, 22:93–115, 2004. 4

[7] H. Chockler and J. Y. Halpern. Explaining image classifiers. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR*, 2024. 3, 4

[8] H. Chockler, D. Kroening, and Y. Sun. Explanations for occluded images. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 1214–1223. IEEE, 2021. 4

[9] H. Chockler, D. A. Kelly, and D. Kroening. Multiple different black box explanations for image classifiers, 2024. URL https://arxiv.org/abs/2309.14309. 2

[10] H. Chockler, D. A. Kelly, D. Kroening, and Y. Sun. Causal explanations for image classifiers, 2024. URL https://arxiv.org/abs/2411.08875. 1, 2, 4, 5

[11] A. Darwiche and A. Hirth. On the (complete) reasons behind decisions. *J. Log. Lang. Inf.*, 32(1):63–88, 2023. 2

[12] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (S&P)*, pages 598–617. IEEE, 2016. 2

[13] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26:297—-302, 1945. 5

[14] T. Eiter and T. Lukasiewicz. Complexity results for explanations in the structural-model approach. *Artif. Intell.*, 154(1-2):145–198, 2004. 4

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012. 6

[16] R. Fong, M. Patrick, and A. Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *International Conference on Computer Vision (ICCV)*, pages 2950–2958. IEEE, 2019. 2

[17] C. Glymour and F. Wimberly. Actual causes and thought experiments. In J. Campbell, M. O'Rourke, and H. Silverstein, editors, *Causation and Explanation*, pages 43–67. MIT Press, Cambridge, MA, 2007. 3

[18] N. Hall. Structural equations and causation. *Philosophical Studies*, 132: 109–136, 2007. 3

[19] J. Y. Halpern. *Actual Causality*. The MIT Press, 2019. 3

[20] J. Y. Halpern and J. Pearl. Causes and explanations: a structural-model approach. Part I: causes. *British Journal for Philosophy of Science*, 56 (4):843–887, 2005. 3

[21] N. R. Hanson. *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, 2010. 1

[22] C. Hitchcock. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, XCVIII(6):273–299, 2001. 3

[23] C. Hitchcock. Prevention, preemption, and the principle of sufficient reason. *Philosophical Review*, 116:495–532, 2007. 3

[24] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*, pages 1511–1519. AAAI Press, 2019. 2

[25] M. Jiang, S. Khorram, and L. Fuxin. Comparing the decision-making mechanisms by transformers and cnns via explanation methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9546–9555, June 2024. 1, 7

[26] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017. 2

[27] J. Marques-Silva and A. Ignatiev. Delivering trustworthy AI through formal XAI. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pages 12342–12350. AAAI Press, 2022. 2

[28] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. 1

[29] W.-J. Nam, S. Gur, J. Choi, L. Wolf, and S.-W. Lee. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 2501–2508, 2020. 2

[30] C. Papadimitriou. The complexity of unique solutions. *Journal of ACM*, 31:492–500, 1984. 4

[31] V. Petsiuk, A. Das, and K. Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*. BMVA Press, 2018. 2

[32] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144. ACM, 2016. 2

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017. 2

[34] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2016. doi: 10.1109/TPAMI.2015.2465960. 6

[35] V. Shitole, F. Li, M. Kahng, P. Tadepalli, and A. Fern. One explanation is not enough: Structured attention graphs for image classification. In *Neural Information Processing Systems (NeurIPS)*, pages 11352–11363, 2021. 1, 2

[36] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*, volume 70, pages 3145–3153. PMLR, 2017. 2

[37] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab.*, 5:1—-34, 1948. 5

[38] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (Workshop Track)*, 2015. URL http://arxiv.org/abs/1412.6806. 2

[39] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 2

[40] B. Weslake. A partial theory of actual causation. *British Journal for the Philosophy of Science*, 2015. To appear. 3

[41] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, U.K., 2003. 3