Multiple Different Explanations for Image Classifiers

Hana Chockler[1], David A. Kelly[2], Daniel Kroening[3*]

[1] King's College London, UK
hana.chockler,david.a.kelly@kcl.ac.uk
[2]Amazon, UK
daniel.kroening@magd.ox.ac.uk

(a) Input (b) Heatmap (c) Multiple disjoint explanations

Figure 1: An example of an image ('peacock') and its maximally disjoint explanations.

"If one is investigating things that are not directly perceptible, and if one sees that several explanations are possible, it is reckless to make a dogmatic pronouncement concerning any single one; such a procedure is characteristic of a seer rather than a wise man."

Diogenes

Abstract

Existing explanation tools for image classifiers usually give only one single explanation for an image. For many images, however, both humans and image classifiers accept more than one explanation for the image label. Thus, restricting the number of explanations to just one severely limits the insight into the behavior of the classifier. In this paper, we describe an algorithm and a tool, REX, for computing multiple explanations of the output of a black-box image classifier for a given image. Our algorithm uses a principled approach based on causal theory. We analyse its theoretical complexity and provide experimental results showing that REX finds multiple explanations on 7 times more images than the previous work on the ImageNet-mini benchmark.

1 Introduction

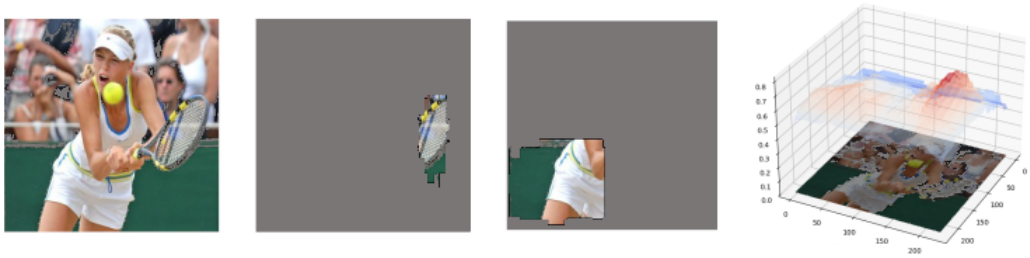
Neural networks (NN) are now a primary building block of most computer vision systems. The opacity of NNs creates demand for explainability techniques, which attempt to provide insight into why a particular input yields a particular observed output. Beyond increasing a user's confidence in the output, and hence also their trust in the AI system, these insights help to uncover subtle classification errors that are not detectable from the output alone (Chockler, Kroening, and Sun 2021).

The most commonly used definition of an explanation is a part of the input image that is sufficient for the classifier to yield the same label as the original image. Explanations, according to this definition, are obviously not unique. Images often have several explanations for their classification as illustrated in Figure 1. Initial results in the study of multiple explanations by Shitole et al. (2021) demonstrate that images have multiple explanations in all but degenerate cases.

There is a growing body of work analysing the human perception of images (Fan et al. 2020; van Dyck et al. 2021) and how this differs from what NNs do. Roughly speaking, humans do not detect small differences. In particular, there is little sense in checking the effect of changing one pixel or any small number of pixels, as a new explanation would be indistinguishable to the human eye from the previous one. Therefore, we focus our effort on the search for *sufficiently different* explanations (see Section 4.1).

The prevalence of multiple explanations suggests that algorithms for computing more than one explanation are essential for understanding image classifiers and uncovering subtle classification errors. The image in Figure 2 is classified by VGG19 as a tennis racket, with the first explanation being indeed a part of the racket. However, the second explanation is the player's shorts, uncovering a misclassification. Yet, existing techniques provide only one explanation of an output of the classifier. The one notable exception is the tool SAG (Shitole et al. 2021), outlined in Section 2, which constructs multiple explanations by using a beam search over a fixed grid. However, as SAG searches an exponential space (the number of combinations of cells of the grid is exponential), it either runs into the exponential explosion problem or drops a part of the state space. This is hardly surprising; as

[*]This work was done prior to joining Amazon.



(a) (b) (c) (d)

Figure 2: The image (a) is classified as 'tennis racket'. Its disjoint explanations found by REX are in (b) and (c). The first explanation is a part of a racket, and the second explanation uncovers a misclassification, as it is the player's shorts. (d) is the saliency landscape.

we prove in Section 3.3, the problem of computing multiple explanations is intractable. Specifically, we present an exponential upper bound on the number of possible explanations and demonstrate that this bound is tight.

In view of these theoretical results, we present REX, an approximation algorithm and a tool for computing multiple explanations for black-box image classifiers. Using the formal mathematical theory of actual causality, REX computes a ranking of the pixels of the image. This ranking is used to construct a refined search landscape (Figure 5), which REX explores in order to generate multiple, different, explanations. Unlike SAG, our search is not limited to exploration from highly ranking parts of the image and allows even unlikely (low ranking regions) to be fruitfully exploited for explanations. Whereas SAG uses a fixed square and beam width for its search, REX expands and contracts its search width to minimise explanation size. For instance, for the image in Figure 1, REX produces 4 explanations whereas SAG produces only 2.

In Section 5 we experimentally compare REX with SAG and with DEEPCOVER on standard benchmarks. The results demonstrate that REX produces finer-grained explanations and is superior to SAG wrt the number of sufficiently different explanations it produces. We provide the details of the benchmark sets, the models, and the main results in the paper. The tool, all datasets, and the full set of results are submitted as the supplementary material.

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 two categories: propagation and perturbation. Propagation-based explanation methods back-propagate a model's decision to the input layer to determine the weight of each input feature for the decision (Springenberg et al. 2015; Sundararajan, Taly, and Yan 2017; Bach et al. 2015; Shrikumar, Greenside, and Kundaje 2017; Nam et al. 2020). GRAD-CAM 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 (Selvaraju et al. 2017).

Perturbation-based explanation approaches introduce perturbations to the input space directly in search for an explanation. 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 (Lundberg and Lee 2017). LIME constructs a small neural network to label the original input and its neighborhood of perturbed images and uses this network to estimate the importance of different parts of the input (Ribeiro, Singh, and Guestrin 2016; Datta, Sen, and Zick 2016; Chen et al. 2018; Petsiuk, Das, and Saenko 2018; Fong, Patrick, and Vedaldi 2019). ANCHORS uses a similar approach to find parts of the inputs sufficient for the classification, regardless of the values of other parts (Ribeiro, Singh, and Guestrin 2018). Finally, DEEPCOVER ranks elements of the image according to their importance for the classification and uses this ranking to greedily construct a small explanation. The DEEPCOVER ranking procedure in (Sun et al. 2020) uses SFL, and is replaced in (Chockler, Kroening, and Sun 2021) by the approximate computation of causal responsibility.

Work on calculating more than one explanation for a given classification outcome is in its infancy. To the best of our knowledge, there is only one algorithm and tool that computes multiple explanations of image classifiers – SAG, described in (Fuxin et al. 2021; Shitole et al. 2021). The motivation for SAG is the same as ours: increasing human confidence and trust as well as our understanding of image classification algorithms.

SAG partitions the input image into a fixed grid (by default 7×7). A beam search algorithm is used to search for the initial w (i.e., the beam width) root nodes in the graph. The search starts with w distinct highest weighted image regions. Their children nodes are perturbed, until the resulting mask causes a unacceptable drop in the label's probability. Explanations are identified from the SAG as multiple minimal regions of the input image sufficient for the correct classification with a high confidence. Explanations are presented in the form of a directed acyclic graph, or Structured Attention Graph (SAG). Multiple explanation diversity is enforced by bounding the maximal overlap in terms of a number of regions shared between explanations.

3 Theoretical Results

In this section we describe the theoretical foundations of our approach.

3.1 Background on Actual Causality

Our definitions are based on the framework of *actual causality* introduced by Halpern and Pearl (2005). The reader is referred to that paper and to Halpern (2019) for an updated overview and more information on actual causality. Due to the lack of space, we omit formal definitions and instead discuss the intuition informally. This is sufficient for our purposes, as we explain below.

The definition of an *actual cause* is based on the concept of *causal models*, which consist of a set of variables, a range of each variable, and structural equations describing the dependencies between the variables. Actual causes are defined with respect to a given causal model, a given assignment to the variables of the model (a context), and a propositional logic formula that holds in the model in this context.

Actual causality extends the simple counterfactual reasoning (Hume 1739) by considering the effect of *interventions*, which are changes of the current setting. Roughly speaking, a subset of variables X and their values in a given context is an actual cause of a Boolean formula φ being True if there exists a change in the values of other values that creates a counterfactual dependency between the values of X and φ (that is, if we change the values of variables in X, φ would be falsified). The formal definition by Halpern and Pearl (2005) and in its modifications, the latest of which is by Halpern (2015), are far more complex due to the potential dependencies between the variables and considering causes of more than one element. In our setup, where we are only interested in singleton causes and in interventions only on the input variables, all versions of the definition of (a part of) an actual cause are equivalent to our definition under the assumption of independence between the input variables.

Responsibility, as defined by Chockler and Halpern (2004) and adapted to the modified definition of causality by Halpern (2015), is a quantification of causality, attributing to each actual cause its *degree of responsibility*, which is derived from the size of a smallest contingency required to create a counterfactual dependence. The degree of responsibility is defined as $1/(k+1)$, where k is the size of a smallest contingency. The degree of responsibility of counterfactual causes is therefore 1 (as $k = 0$), and the degree of responsibility of variables that have no causal influence on φ is 0, as k is taken to be ∞. In general, the degree of responsibility is always between 0 and 1, with higher values indicating a stronger causal dependence.

3.2 Causes and explanations in image classification

We follow the approach by Chockler, Kroening, and Sun (2021) (CKS from now on) to defining causes and constructing explanations in image classification. We view the NN as a black-box causal model in the Halpern and Pearl (2005)-sense of the word, with its inputs being the individual pixels of an input image. The variables are defined as Boolean, with the values being the original color and the masking color (as shown by CKS, a specific masking color does not have almost any effect on the results). Following Halpern (2019), we further augment the model by limiting the allowed interventions to masking the colors of input's pixels. Moreover, we are only interested in singleton causes (recall that we assume independence between the input variables).

Definition 1 (Singleton cause for image classification, CKS)**.** *For an image x classified by the NN as $f(x) = o$, a pixel p_i of x is a* cause *of o iff there exists a subset P_j of pixels of x such that the following conditions hold:*

SC1. $p_i \notin P_j$*;*
SC2. *changing the color of any subset $P'_j \subseteq P_j$ to the masking color does not change the classification;*
SC3. *changing the color of P_j and the color of p_i to the masking color changes the classification.*

We call such P_j a witness *to the fact that p_i is a cause of x being classified as o.*

Definition 2 (Simplified responsibility, CKS)**.** *The* degree of responsibility *$r(p_i, x, o)$ of p_i for x being classified as o is defined as $1/(k+1)$, where k is the size of the smallest witness set P_j for p_i. If p_i is not a cause, k is defined as ∞, and hence $r(p_i, x, o) = 0$. If changing the color of p_i alone to the masking color results in a change in the classification, we have $P_j = \emptyset$, and hence $r(p_i, x, o) = 1$.*

Lemma 1 (CKS)**.** *Definition 1 is equivalent to the definition of an actual cause when input variables in the model are independent of each other, and we do not consider interventions on internal variables.*

Corollary 1 (CKS)**.** *The problem of detecting causes in image classification is NP-complete.*

Definition 3 (Explanation for image classification, CKS)**.** *An explanation in image classification is a minimal subset of pixels of a given input image that is sufficient for the NN 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.*

3.3 New Theoretical Results

In this section we prove new complexity bounds for computing multiple explanations.

CKS 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 (Eiter and Lukasiewicz 2004). 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 (Papadimitriou 1984). The following lemma shows that computing a second (or any subsequent) explanation is not easier than computing the first one. For the purposes of proving theoretical results, a subsequent explanation is one that differs from the previous ones in at least one pixel; the algorithm in Section 4 constructs spatially different explanations, more suitable to the human perception.

Lemma 2. *Given an explanation, constructing a different one is DP-complete.*

Proof Sketch. Membership in DP is straightforward. For the hardness part we show a reduction from the problem of computing an explanation. Given an image x classified as $\mathcal{N}(x)$, we construct a *chimera* image from x and an existing explanation of $\mathcal{N}(x)$ (taken from another image) attached to it without obscuring it. Then, our existing explanation is the first explanation to the image being classified as $\mathcal{N}(x)$, and a second one is an explanation of the classification of the original image. □

We note that the chimera image constructed in the reduction does not have a rectangular shape; however the complexity of the explanation problem does not depend on the shape of the input image.

CKS use a greedy approach to constructing approximate explanations, based on scanning the ranked list of pixels $pixel\_ranking$. We note that the construction of the ranked list is intractable as well (NP-complete), even when the ranking is based on Definition 2, rather than the general definition of responsibility by Chockler and Halpern (2004). Hence, CKS construct an approximate ranked list by partitioning the set in iterations and computing approximate degrees of responsibility for each partition while discarding low-responsibility elements.

However, this approach does not help in reducing the complexity of computing many explanations, as the number of explanations for a given image can be very large, as proven in the following lemma.

Lemma 3. *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 (Anderson 1987). 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 that determines whether an input image of size n has at least $\lfloor n/2 \rfloor$ green-coloured pixels and an input image 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, finding 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 $O(1)$. Indeed, let $\mathcal{N}$ be a binary classifier that determines whether an input graph $G = \langle V, E \rangle$ contains any connected components of size more than 1. An explanation would be a node $v \in V$ with its adjacent edges. Now, G contains an independent set of size n iff there exist n disjoint explanations of the non-empty label of G given by the classifier, thus proving NP-hardness of the problem.

4 Multiple Explanations

In this section we present our algorithm for computing multiple, different explanations. As shown in Section 3, 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 appendix (see the supplementary material).

4.1 What is a "Different Explanation"?

As the goal of constructing different explanations is presenting them to humans for analysis, we need to ensure that the explanations are indeed perceived as different by humans. Consider the explanations in Figure 4. As discussed by Zhang et al. (2015), a human eye fills in the gaps of hazy and low-resolution images. Hence, if we remove a small subset of pixels from a given explanation, it would be sufficiently different from the original one according to many distance measures, yet would likely not be different at all to the human eye (in Figure 4 the gaps are increased for illustrative purposes).

To avoid this problem, we define an *atomic superpixel*, the smallest set of contiguous pixels (a square) that is distinguishable to a human, as a parameter of the algorithm. The concept of a superpixel is used in a number of different explanation tools. Both SAG and GRAD-CAM split the image into a $7x7$ grid of squares. Dividing the image in this way greatly reduces the computational cost of searching for explanations. The rigidity of the grid, however, leads to a strict bound on the minimum explanation size: an explanation cannot be less than the size of a square, and a square may be significantly bigger than the smallest superpixel responsible for the classification. REX overcomes this problem by generating a random grid. We also allow the minimum size of the superpixel as a parameter. We discuss this further in Section 5. To overcome the limitations of just one grid, we allow for multiple iterations of the algorithm, each with a different grouping of superpixels. The results of the different grids are automatically combined to produce a detailed saliency landscape, as in Figure 5. As one can see, the more iterations are added, the smoother the saliency landscape becomes.

4.2 The REX Algorithm

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.

The $RANK$ procedure in Line 3 of Algorithm 1 constructs a $pixel\_ranking$, which is a ranking of the pixels of the input image x. While any pixel ranking mechanism can be used (e.g., an SFL-based ranking in Sun et al. (2020) or LIME or SHAP heatmaps), the quality and the granularity of the final results depend on the quality and the granularity of the ranking. We implemented and tested REX with causal responsibility-based ranking described in CKS. 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 3).

The $Floodlight$ procedure called in Line 5 is described in Algorithm 2. It replaces the greedy explanation generation in DEEPCOVER with a spatially delimited stochastic

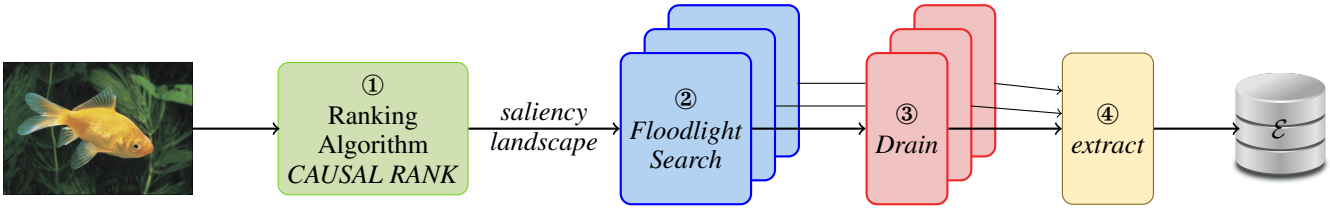


Figure 3: A schematic depiction of our algorithm, returning a set of explanations $\mathcal{E}$ for a given input image. Its components: ① *ranking* generates a saliency landscape of pixels; ② *search* launches x *floodlight* searches over the landscape; ③ *drain* minimizes the explanations founds in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

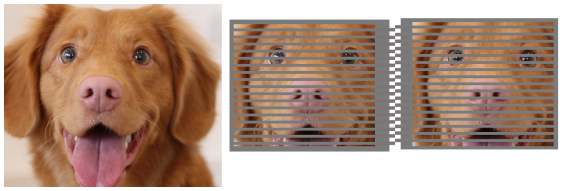


Figure 4: To the human eye, these two explanations for a dog are equivalent, but they do not have any non-background pixels in common. Naively calculating the pixel overlap is insufficient in this case; we must take into account spatial location.

Algorithm 1: REX$(x, \mathcal{N}, r, n, \delta, p, q)$

INPUT: an image x, a network $\mathcal{N}$, a floodlight radius r. the maximal number of explanations n, δ maximum overlap between explanations, p number of floodlight expansions, q expansion coefficient
OUTPUT: a set of up to n different explanations $\mathcal{E}$

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

hill climb. In contrast to most hill-climb-based algorithms that look for the global maximum, we search for *local maxima*, as these are likely to correspond to explanations. The global maximum usually matches the explanation computed by DEEPCOVER, though it is not guaranteed. The function *initialize* in Line 1 creates a floodlight of radius r at a random position over the image x. We call the model on this masked image. If the initial size of the floodlight, $\mathcal{F}$, is too small to encompass an explanation, before taking a random step, $\mathcal{F}$ expands in position a fixed number of times. If this increased flooding still does not result in an explanation, the floodlight takes a random step, returning to its original radius. This random step is mediated by an objective function. By default, REX uses the mean of the responsibility of the pixels under the floodlight.

Algorithm 2: *Floodlight* $(x, \mathcal{N}, l, \mathcal{S}, r, n, p, q)$

INPUT: an image x, a network $\mathcal{N}$, a label l, a saliency landscape $\mathcal{S}$, a floodlight 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: $\quad\quad l' \leftarrow \mathcal{N}(\mathcal{F}(x))$
6: $\quad\quad$ **if** $l = l'$ **then**
7: $\quad\quad\quad E \leftarrow \mathcal{F}(x)$
8: $\quad\quad\quad$ **return** E
9: $\quad\quad$ **else**
10: $\quad\quad\quad \mathcal{F} \leftarrow expand\_radius(r * q)$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: $\quad \mathcal{F} \leftarrow neighbor$
14: **end for**
15: **return** E

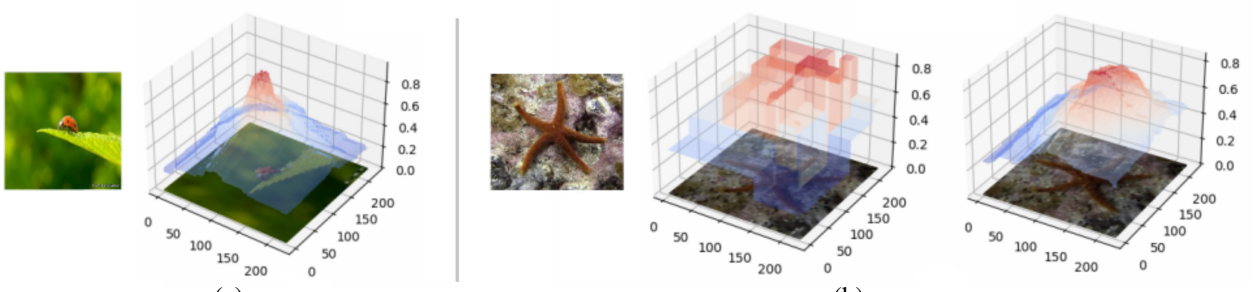
Once an explanation is found, REX performs a local ablation *drain* (see the appendix for details).

Finally, the *extract* procedure (Algorithm 3), extracts a subset of at most n explanations that pairwise overlap up to the input bound δ. As discussed in Section 3.3, the exact solution is NP-hard. The procedure uses a greedy heuristic based on the Sørensen–Dice coefficient (SDC) (Dice 1945; Sørensen 1948), typically used as a measure of similarity between samples. First, we calculate the matrix SDC for all pairs of explanations: $SDC(i,j) = 0$ iff $SDC(E_i, E_j) \leq \delta$, and is $SDC(E_i, E_j)$ otherwise. [1] Columns that sum to 0 correspond to explanations that do not overlap others in more than δ, and are hence added to $\mathcal{E}$. We then greedily remove the most overlapping explanations and recalculate the overlap matrix, adding the columns summing to 0 to $\mathcal{E}$. The procedure iterates until SDC is empty.

5 Experimental Results

Implementation We implemented Algorithm 1 in the tool REX for generating multiple explanations. Given a saliency landscape, by default, REX attempts to find 10 explanations.

[1] For disjoint explanations, i.e., $\delta = 0$, we can simply take $E_i \cap E_j$ instead of $SDC(i,j)$.



(a) (b)

Figure 5: Different saliency landscapes. The image on the left shows shows an image with a single explanation, clearly indicated by the single central peak. The other images show the saliency landscapes of a starfish after 1 and 20 iterations. The landscape here is flatter, with multiple separate peaks. These peaks are likely to correspond with different explanations.

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

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

$all\_pairs \leftarrow \mathcal{E} \times \mathcal{E}$
$m \leftarrow 0_{|\mathcal{E}|,|\mathcal{E}|}$
for (p_i, p_j) in all_pairs **do**
  $SDC \leftarrow dice\_coefficient(p_i, p_j)$
  $m_{p_i,p_j} = SDC > \delta \; ? \; SDC : 0$
end for
$\mathcal{E}' \leftarrow \emptyset$
for i in $0 \ldots |\mathcal{E}| - 1$ **do**
  for j in $0 \ldots cols(m)$ **do**
    if $sum(col_j) = 0$ **then**
      $\mathcal{E}' \leftarrow \mathcal{E}' \cup \mathcal{E}_j$
    end if
  end for
  $m \leftarrow remove\_most\_overlapping\_explanation(m)$
end for
return $\mathcal{E}'$

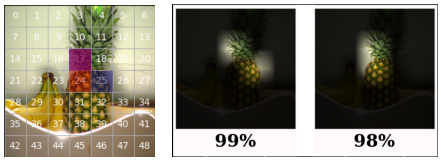
While it is computationally relatively inexpensive to search for more explanations than this, on our dataset we observe that images with more than 6 sufficiently different explanations are extremely rare ($\approx 1\%$ of images). The algorithm computes multiple maximally different approximations of causal explanations according to Definition 3.

Datasets and Models For our experiments, we used two standard image datasets. The first dataset is ImageNet-mini[2], consisting of 3923 images representing 1000 different labels. The second dataset consists of all images, 81 in total, labeled *starfish* from the Caltech256 dataset (Griffin, Holub, and Perona 2007). The second dataset is used as an additional independent source of images to mitigate the risk of overfitting to the ImageNet-mini dataset. While REX is agnostic to the model, SAG uses VGG19 by default. To enable comparison, we tested REX with the same model.

[2]https://www.kaggle.com/datasets/ifigotin/imagenetmini-1000

5.1 Tool setup and parameters

We use both REX and SAG with default settings. In particular, REX offers a large number of tunable parameters. We set the total number of iterations at 20, and the total number of floodlights at 10 for all images. SAG requires the user to set the maximal allowed overlap in squares, with suggested values of 0, 1, or 2. REX has no required bound parameter, but has the option of returning all explanations. We present the results for disjoint explanations (that is, SAG with 0 overlaps). See the supplementary material for the results for explanations with a small overlap.



(a) SAG's grid (b) Explanations

Figure 6: Two explanations from SAG overlapping on the square no.17. SAG has a rigid overlap size (one square on the grid). REX's overlap is a parameter and depends on the size of both explanations.

SAG divides the image into a rigid grid, by default 7×7, whereas REX iteratively refines the random image partitioning. We used a betabinomal distribution, with both α and β set to 1.1 for the random partitioning, to reduce the probability of extremely unbalanced partitions. SAG takes a probability threshold when considering whether a combination of squares is an explanation. REX, by default, takes the top prediction from the model, without reference to probability. Setting a probability threshold is arbitrary, whereas taking the top prediction is more consistent with the model's "best guess". More importantly, by setting a probability threshold, we deliberately ignore inconsistent classification or misclassification (Figure 2).

5.2 Experimental Results and Comparison with SAG

A natural performance measure for multiple explanations is the number of multiple significantly different explanations

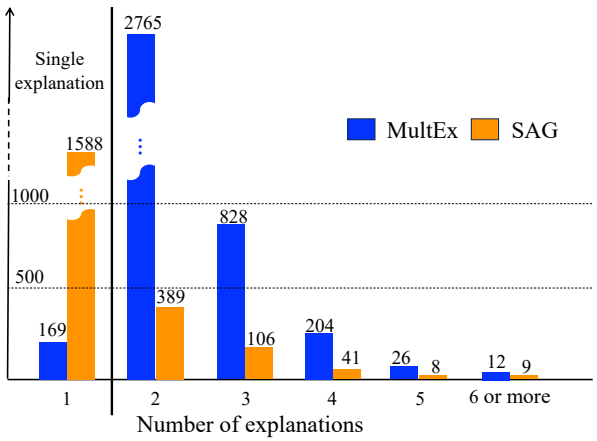


Figure 7: Number of disjoint explanations found by SAG and REX on Imagenet-mini and Caltech256 starfish datasets (The fewer images with 1 explanation, the better).

produced for each image. We tested SAG and REX with the option of producing completely disjoint explanations and with the option of having a small overlap. Note that as a square in SAG has a fixed size, $1/49$-th of the input image, it can result in very similar explanations if the explanations are small (see an illustration in Figure 6).

The main experiment was run on AWS, using a cluster of Inter Xeon Platinum 8375C CPU @ 2.90 GHz, without GPU support (which equally disadvantages both tools). The timeout (TO) is set to 10 minutes for each tool on each image. Within the given TO, SAG did not terminate on just under half of the images in Imagenet-mini, and on one image in the Caltech256 starfish dataset. REX terminated on all images in both datasets.

The results show that REX computes multiple explanations on **7X** more images in the benchmark set than SAG (**3835** for REX vs **553** for SAG). Figure 7 shows the breakdown of the results by to the number of images having a particular number of disjoint explanations found by REX and by SAG, respectively. The results are also presented in the tabular form in Table 2. Table 1 presents the analysis of percentage of termination and the average number of explanations for each tool. The results on the starfish dataset are similar to those on Imagenet-mini, demonstrating robustness of our approach on unseen images. 1-square overlap in explanations produces similar results (see the supplementary material).

6 Conclusions and Future Work

Motivated by studies in human cognition and the need for thorough debugging of image classifiers, this paper proposes an algorithm and a tool REX for constructing multiple explanations for the outputs of image classifiers. The algorithm is based on a solid mathematical theory of causal reasoning and is agnostic to the classifier, viewing it as a black box. The tool REX is modular and borrows its ranking procedure from the existing tool DEEPCOVER. We introduce a a novel

Tools	Datasets					
	Imagenet-mini			Starfish		
	% term	avg term	avg	% term	avg term	avg
REX	100	2.34	2.3	100	2.12	2.12
SAG	53	1.4	0.7	98	1.07	1.08

Table 1: Results for REX and SAG on Imagenet-mini and Caltech256 starfish. '% term' is % of images on which each tool terminated within the TO; 'avg term' is the mean number of explanations for images, where the tool terminated; 'avg' is the mean taken over all images.

Tools	No. Exp						Total
	1	2	3	4	5	6+	
REX	169	2765	828	204	26	12	**4004**
SAG	1588	389	106	41	8	9	**2141**

Table 2: Total number of images with corresponding number of explanations for both datasets.

explanation-discovery algorithm based on the saliency landscape and a "floodlight" search, ensuring different spatial locations for explanations.

REX is built as a command-line tool and a Python library with pluggable components. Owing to its systematic and compositional approach, REX finds multiple different explanations of image labels on standard benchmark sets and is fully configurable. Moreover, by default REX does not depend on the probabilities assigned to the labels by the classifier. We compare our results with SAG, the only other tool for multiple explanations, and demonstrate that REX finds significantly more explanations than SAG. Moreover, REX terminates on the whole benchmark set, in contrast to SAG that timed out on 50% of it. The algorithm is completely parallelized, which, together with the efficient floodlight search, leads to 17X speedup compared to DEEPCOVER (see the appendix).

There is a number of promising directions for future work. Due to the modularity of REX, it is possible to plug in any other ranking procedure and to experiment with different algorithms for explanation discovery based on the saliency landscape. Furthermore, we hypothesise that the precision of ranking drops for the lower ranked elements, affecting the quality of explanations that are lower on the saliency landscape. While the intractability of computing an explanation implies a tradeoff between the quality of the approximation and the precision of the result, we will search for new heuristics to improve the saliency landscape at low levels.

References

Anderson, I. 1987. *Combinatorics of Finite Sets*. Oxford University Press.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller,

K.-R.; and Samek, W. 2015. On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation. *PLOS One*, 10(7).

Chen, J.; Song, L.; Wainwright, M.; and Jordan, M. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *International Conference on Machine Learning (ICML)*, volume 80, 882–891. PMLR.

Chockler, H.; and Halpern, J. Y. 2004. Responsibility and Blame: A Structural-Model Approach. *J. Artif. Intell. Res.*, 22: 93–115.

Chockler, H.; Kroening, D.; and Sun, Y. 2021. Explanations for Occluded Images. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 1214–1223. IEEE.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *Security and Privacy (S&P)*, 598–617. IEEE.

Dice, L. R. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26: 297—-302.

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

Fan, S.; Koenig, B. L.; Zhao, Q.; and Kankanhalli, M. S. 2020. A Deeper Look at Human Visual Perception of Images. *SN Computer Science*, 1(58).

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

Fuxin, L.; Qi, Z.; Khorram, S.; Shitole, V.; Tadepalli, P.; Kahng, M.; and Fern, A. 2021. From Heatmaps to Structured Explanations of Image Classifiers. *Applied AI Letters*, 2(4): e46.

Griffin, G.; Holub, A.; and Perona, P. 2007. Caltech-256 object category dataset.

Halpern, J. Y. 2015. A Modification of the Halpern–Pearl Definition of Causality. In *Proceedings of IJCAI*, 3022–3033. AAAI Press.

Halpern, J. Y. 2019. *Actual Causality*. The MIT Press.

Halpern, J. Y.; and Pearl, J. 2005. Causes and Explanations: A Structural-model Approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56(4).

Hume, D. 1739. *A Treatise of Human Nature*. John Noon.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 4765–4774.

Nam, W.-J.; Gur, S.; Choi, J.; Wolf, L.; and Lee, S.-W. 2020. Relative Attributing Propagation: Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks. In *AAAI Conference on Artificial Intelligence*, volume 34, 2501–2508.

Papadimitriou, C. 1984. The Complexity of Unique Solutions. *Journal of ACM*, 31: 492–500.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*. BMVA Press.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Knowledge Discovery and Data Mining (KDD)*, 1135–1144. ACM.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of AAAI*, 1527–1535. AAAI Press.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *International Conference on Computer Vision (ICCV)*, 618–626. IEEE.

Shitole, V.; Li, F.; Kahng, M.; Tadepalli, P.; and Fern, A. 2021. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. In *Neural Information Processing Systems (NeurIPS)*, 11352–11363.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. In *International Conference on Machine Learning (ICML)*, volume 70, 3145–3153. PMLR.

Sørensen, T. 1948. 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.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. A. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR (Workshop Track)*.

Sun, Y.; Chockler, H.; Huang, X.; and Kroening, D. 2020. Explaining Image Classifiers using Statistical Fault Localization. In *ECCV, Part XXVIII*, volume 12373 of *LNCS*, 391–406. Springer.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*, 3319–3328. PMLR.

van Dyck, L. E.; Kwitt, R.; Denzler, S. J.; and Gruber, W. R. 2021. A Deeper Look at Human Visual Perception of Images. *Front. Neurosci.*, 15.

Zhang, X.-S.; Gao, S.-B.; Li, C.-Y.; and Li, Y.-J. 2015. A Retina Inspired Model for Enhancing Visibility of Hazy Images. *Front. Comput. Neurosci.*, 9.