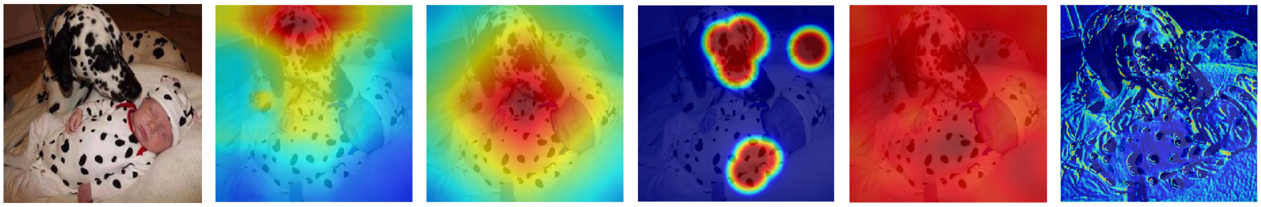
Compositional Explanations for Image Classifiers

Hana Chockler
King's College London
hana.chockler@kcl.ac.uk

Daniel Kroening*
Amazon.com, Inc.
daniel.kroening@gmail.com

Youcheng Sun
Queen's University Belfast
youcheng.sun@qub.ac.uk



(a) Input (b) CET (c) DeepCover (d) Extremal (e) RISE (f) Grad-CAM

Figure 1: Output from different tools for explaining 'dalmatian'. CET is the compositional explanation tool in this work.

Abstract

Existing algorithms for explaining the output of image classifiers perform poorly on inputs where the object of interest is partially occluded. We present a novel, black-box algorithm for computing explanations that uses a principled approach based on causal theory. We implement the method in the tool CET *(Compositional Explanation Tool). Owing to the compositionality in its algorithm,* CET *computes explanations that are much more accurate than those generated by the existing explanation tools on images with occlusions and delivers a level of performance comparable to the state of the art when explaining images without occlusions.*

1. Introduction

Deep neural networks (DNNs) are now a primary building block of many computer vision systems. DNNs are complex non-linear functions with algorithmically generated (and not engineered) coefficients. In contrast to traditionally engineered image processing pipelines it is difficult to retrace how the pixel data are interpreted by the layers of the DNN. This "black box" nature of DNNs creates demand for techniques that explain why a particular input yields the output that is observed.

Explanations for the results of image classifiers are typically given in the form of a *ranking* of the pixels, which is a numerical measure of importance: the higher the score, the more important the pixel is for the DNN classification outcome. Given an image that features some object, the good algorithms are able to generate rankings that identify that object with high accuracy. Leading tools in the area include (in order of publication) LIME [22], SHAP [5], Grad-CAM [24], RISE [21], Extremal [8] and DeepCover [26].

A typical proxy for the quality of a ranking is how many of the high-ranked pixels need to be masked before the classification generated by the DNN changes. Good explanations require very little masking.

The problem is computationally hard: the size of the space of rankings is exponential in the number of pixels, yet practical applications require performance that is roughly linear in the size of the image. It is therefore expected that algorithms in the domain approximate the solution in one way or another, and that they implement heuristics that are tuned for typical inputs. This assumption is entirely appropriate in many use cases, and in particular, works very well on the benchmark sets that are used in the area: the existing work has been evaluated using the ImageNet dataset, and ImageNet has been curated so are all objects clearly visible.

We argue that there is a use-case for explanations of the results of image classifiers for images where the trigger for the result is *not* contiguous. Obvious exemplars are cases with partial occlusion, say by a person walking into a scene or simply by dirt on your camera lens. To this end, we introduce a new image dataset we call *Photo Bombing*, in which we obscure ImageNet photos by masking parts of the object. The difference between the modified image and the original one is the ground truth for the "photobomber", and a good

[*] The work reported in this paper was done prior to joining Amazon.

explanation should not have any overlap with it. To avoid the overfitting done by prior work, we apply a principled approach grounded in causal theory. The algorithm implements an iterative, highly parallelizable approach that delivers significantly better accuracy on an existing dataset with partial occlusion and on our own photo boming data set. The tool, the new benchmark set, and the full set of results are available at https://sites.google.com/view/cet-tool/.

2. Related Work

There is a large body of work on explaining image classifiers. The existing approaches can be largely grouped into two categories: white-box and black-box.

White-box explanation methods are often regarded as more efficient. They back-propagate a model's decision to the input layer to determine the weight of each input feature for the decision. Grad-CAM [24] 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. In [25], the activation of each neuron is compared with some reference point, and its contribution score for the final output is assigned according to the difference. In [18], a unified framework is proposed for explanation methods, that is, an approximation of the model's local behavior using a simpler linear model and an application of the Shapley Value theory to solve this model.

By contrast, black-box explanation approaches explore the input space directly in search for an explanation. The exploration/search often requires a large number of inference passes, which incurs significant computational cost when compared to white-box methods. Many sampling methods have been proposed, but most of are based on random search or heuristics and lack rigor.

Given a particular input, LIME [22] samples the the neighborhood of this input and creates a linear model to approximate the model's local behavior; owing to the high computational cost of this approach, the ranking uses super-pixels instead of individual pixels. In [5], the natural distribution of the input is replaced by a user-defined distribution and the Shapley Value method is used to analyze combinations of input features and to rank their importance. In [3], the importance of input features is estimated by measuring the the flow of information between inputs and outputs. Both the Shapley Value and the information-theoretic approaches are computationally expensive. In RISE [21], the importance of a pixel is computed as the expectation over all local perturbations conditioned on the event that the pixel is observed. The concept of "extreme perturbations" has been introduced to improve the perturbation analysis by the Extremal algorithm [8]. More recently, spectrum-based fault localisation is applied to explaining image classifiers, which outperforms the other tools on explaining images without occlusions. While in RISE and DeepCover, input pixels are masked randomly, Extremal uses an area constraint to optimize the perturbation. The approach that is closest to the one we describe in this paper is implemented in the tool DeepCover and presented in [26]. Like CET, DeepCover constructs explanations greedily from a ranked list of pixels. The ranking, however, is calculated with statistical fault localisation, and is much less precise, as we demonstrate empirically.

The work presented in this paper is motivated by the fact that compositionality is a fundamental aspect of human cognition [1, 7] and by the compositional computer vision work in recent years [17, 27, 17, 30, 28, 29]. While it is well known that the performance of conventional convolutional neural networks degrades when given partially occluded objects, the impact of partial occlusion on algorithms for generating explanations has not been studied before.

The CET method presented in this paper is a back-box approach. The compositional explanation approach in CET addresses the limitations of existing black-box methods in two aspects. The feature masking in CET is based on causal reasoning that provides guarantee (subject to the assumption). Furthermore, the CET algorithm is highly parallel, which makes it ideal for large-scale computer vision problems. As we demonstrate in Section 6.1, CET constructs its explanations in a compositional manner and is a perfect fit for compositional computer vision pipelines.

3. Background

3.1. Deep neural networks (DNNs)

Let $f : \mathcal{I} \to \mathcal{O}$ be a deep neural network $\mathcal{N}$ with n layers. For a given input image $x \in \mathcal{I}$, $f(x) \in \mathcal{O}$ calculates the output label of the DNN.

$$f(x) = f_N(\dots f_2(f_1(x; W_1, b_1); W_2, b_2) \dots ; W_n, b_n) \tag{1}$$

where W_i and b_i for $i = 1, 2, \dots$ are learnable parameters, and $f_i(z_{i-1}; W_{i-1}, b_{i-1})$ is the layer function that maps the output of layer $(i-1)$, i.e., z_{i-1}, to the input of layer i. Our algorithm is independent of the specific internals of the DNN and treats a given DNN as a black box.

3.2. Actual causality

Our definition of cause is a simplified definition of *actual cause* introduced in [12]. For the lack of space, we do not present this definition here in full, but instead discuss the intuition informally. The definition of actual cause is based on the definition of *causal models*, which consists of the set of variables, the range of each variable, and the structural equations describing the dependencies between the variables. Actual causes are defined with respect to a given causal model, a given context (an assignment to the variables of the model), and a propositional logic formula that holds in the model in this context.

Actual causality extends the simple counterfactual reasoning [15] by considering *contingencies*, 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 logic formula φ being True if there exists a change in the current values of other values that creates a counterfactual dependency between the values of X and φ (that is, if we now change the values of variables in X, φ would be falsified). The formal definition is more complex and requires that the dependency is not affected by changing the values of variables not in the contingency, as well as requesting minimality.[1]

Responsibility, defined in [4], is a quantification of causality, attributing to each actual cause its *degree of responsibility*, which is based on the size of a smallest contingency required to create a counterfactual dependence. Essentially, 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 sets of variables that are not actual causes of 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 dependency.

4. Theoretical Foundations

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

4.1. Simplified causal theory

Our definition of causality for image classification follows the contingency-based approach and is based on the definition in [12] and its matching definition of responsibility [4].

We assume that the variables representing areas of the image are independent of each other, which significantly simplifies the definitions.

Definition 1 (Simplified cause)**.** *For an image x classified by the DNN 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* background *color does not change the classification;*

SC3. *changing the colour of P_j and the colour of p_i to the background 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)**.** *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 background color results in a change in the classification, we have $P_j = \emptyset$, and hence $r(p_i, x, o) = 1$.*

Lemma 1. *Definition 1 is equivalent to the definition of actual cause when all variables in the model are independent of each other.*

Proof sketch. The minimality requirement is satisfied immediately, given that our causes are singletons. The additional condition in [12] that requires subsets Z to be set to their original values is only relevant when there are dependencies between the variables. □

Observation 1. *Given an image x and its classification o, we can calculate the degree of responsibility of each pixel p_i of x by directly applying Def. 1, that is, by checking the conditions* **SC1**, **SC2**, *and* **SC3** *for all subsets P_j of pixels of x and then choosing a smallest witness subset. While there is an underlying Boolean formula that determines the classification o given the values of the pixels of x, we do not need to discover this formula in order to calculate the degree of responsibility of each pixel of x.*

4.2. Explanations

An explanation of an output of an automated procedure is essential in many areas, including verification, planning, diagnosis and the like. A good explanation can increase a user's confidence in the result. Explanations are also useful for determining whether there is a fault in the automated procedure: if the explanation does not make sense, it may indicate that the procedure is faulty. It is less clear how to define what a *good* explanation is. There have been a number of definitions of explanations over the years in various domains of computer science [2, 9, 20], philosophy [14] and statistics [23]. The recent increase in the number of machine learning applications and the advances in deep learning led to the need for *explainable AI*, which is advocated, among others, by DARPA [10] to promote understanding, trust, and adoption of future autonomous systems based on learning algorithms (and, in particular, image classification DNNs). DARPA provides a list of questions that a good explanation should answer and an epistemic state of the user after receiving a good explanation. The description of this epistemic state boils down to *adding useful information* about the output of the algorithm and *increasing trust* of the user in the algorithm.

In this paper, we are adapting the definition of explanations by Halpern and Pearl [13] to our setting. The definition in [12] is based on the definition of actual causality. Our definition is based on Def. 1.

[1] In [11], Halpern presents an updated definition of causality; the version in [12] is more suitable for our purposes, as we are interested in singleton causes.

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

We note that (1) the explanation cannot be too small (or empty), as a too small subset of pixels would violate the sufficiency requirement, and (2) there can be multiple explanations for a given input image.

The precise computation of an explanation in our setting is intractable, as it is equivalent to the earlier definition of explanations in binary causal models, which is DP-complete [6]. A brute-force approach of checking the effect of changing the color of each subset of pixels of the input image to the background color is exponential in the size of the image.

Instead, we introduce a *greedy compositional approach* to computing explanations. The approach is greedy because we rank the elements in the decreasing order of responsibility for the classification and greedily add them to the explanation one by one until the original classification is restored. Note, however, that computing the degree of responsibility of pixels is intractable, and is, in fact, not easier than computing an explanation.

To address the problem of intractability of computing the degree of responsibility, we introduce the notion of a *superpixel* P_i, which is a subset of pixels of a given image. Given an image x, we partition it to a small number of superpixels and compute their degree of responsibility in the output of the DNN. Then, we only refine the superpixels with a high responsibility (over a predefined threshold). The scalability of the approach depends on the following observation, which is heuristically true in our experiments.

Observation 2. *The pixels with the highest responsibility for the DNN's decision are located in super-pixels with the highest responsibility.*

Intuitively, the observation holds if pixels with high responsibility do not appear in the superpixels surrounded by other pixels with very low responsibility for the input image classification. While this can happen in theory, in practice due to the continuous nature of the image, we do not encounter this case (even when the explanation is non-contiguous). This property is a key to our compositional algorithm's success.

5. Compositional Explanations

In this section, we present our compositional explanation algorithm based on the causal and responsibility concepts in Section 4.1. The general idea is to calculate the responsibility of a super pixel and recursively distribute this responsibility to pixels within this super pixel. The compositional explanation (CE) approach in this work comprises of three steps.

1. Given a set of super pixels, compute the responsibility of each its super pixel (Section 5.1).

2. Following the responsibility result in Step 1, further refine the super pixel and calculate the responsibility for the refined super pixels (Section 5.2). This is where the compositionality of the explanation algorithm comes from.

3. As it is insufficient to explain an input by only using one particular set of super pixels, multiple sets will be selected and they will be analysed independently by Step 2. Finally, all their results will be merged and a ranking of pixels following their responsibility will be computed, from which an explanation will be constructed (Section 5.3).

5.1. Computing the responsibility of a super pixel

Given a set of pixels $\mathcal{P}$, we use P_i to denote a *partition* of $\mathcal{P}$, that is, a set $\{P_{i,j} : \bigcup P_{ij} = \mathcal{P}$ and $\forall j \neq q, P_{ij} \cap P_{jk} = \emptyset\}$. The number of elements in P_i is a parameter; in this work, we consider partitions of 4 elements. We note that P_{ij} can be further partitioned into smaller sets. We refer to P_{ij} as *superpixels*.

For a DNN $\mathcal{N}$, an input x, and a partition P_i, we can generalize Def. 1 to the set of *superpixels* defined by P_i. We denote by $r_i(P_{ij}, x, \mathcal{N}(x))$ the *degree of responsibility* of a super pixel P_{ij} for $\mathcal{N}$'s classification of x, given P_i.

For a partition P_i, we denote by X_i the set of *mutant images* obtained from x by masking subsets of P_i, and by $\tilde{X}_i$ the subset of X_i that is classified as the original image x. Formally,

$$\tilde{X}_i = \{x_m : \mathcal{N}(x_m) = \mathcal{N}(x)\}.$$

We compute $r_i(P_{ij}, x, \mathcal{N}(x))$, the responsibility of each superpixel P_{ij} in the classification of x, in Algorithm 1. For a superpixel P_{ij}, we define the set

$$\tilde{X}_i^j = \{x_m : P_{ij} \text{ is not masked in } x_m\} \cap \tilde{X}_i.$$

For a mutant image x_m, we define $\mathit{diff}_i(x_m, x)$ as the number of superpixels in the partition P_i that are masked in x_m (that is, the difference between x and x_m with respect to P_i). For an image y, we denote by $y(P_{i,j})$ an image that is obtained by masking the superpixel $P_{i,j}$ in y.

The degree of responsibility of a superpixel P_{ij} is calculated by Alg. 1 as a minimum difference between a mutant image and the original image over all mutant images x_m that do not mask P_{ij}, are classified the same as the original image x, and masking P_{ij} in x_m changes the classification.

Algorithm 1 *responsibility*(x, P_i)

```
INPUT: an image $x$, a partition $P_i$
OUTPUT: a responsibility map $P_i \rightarrow \mathbb{Q}$

for each $P_{ij} \in P_i$ do
    $k \leftarrow \min_{x_m}\{diff(x_m, x) \mid x_m \in \tilde{X}_i^j\}$
    $r_{ij} \leftarrow \frac{1}{k+1}$
end for
return $r_{i0}, \ldots, r_{i,|P_i|-1}$
```

5.2. Compositional refinement of the responsibility

Algorithm 1 calculates the responsibility of each super pixel, subject to a given partition. However, discovering a partition that increases the amount of information about the explanation is not a simple task. Consider, for example, a situation where the explanation is right in the middle of the image, and our partition divides the image into four quadrants. Each quadrant would be equally important for the classification, hence we would not gain any insight into why the image was classified in that particular way.

Our compositional algorithm (see Alg. 2) iteratively refines the high-responsibility superpixels until a precise explanation is constructed and recursively applies Algorithm 1 to each refinement.

Algorithm 2 *compositional_responsibility*(x, P_i)

```
INPUT: an image $x$ and a partition $P_i$
OUTPUT: a responsibility map $P_i \longrightarrow \mathbb{Q}$

1: $R \leftarrow responsibility(x, P_i)$
2: if $R$ meets termination condition then
3:     return $R$
4: end if
5: $R' \leftarrow \emptyset$
6: for each $P_{i,j} \in P_i$ s.t. $R(P_{i,j}) \neq 0$ do
7:     $R' \leftarrow R' \cup compositional\_resposibility(x, P_{i,j})$
8: end for
9: return $R'$
```

Given a partition, Algorithm 2 calculates the responsibility for each superpixel (Line 1). If the termination condition is met (Lines 2–3), the responsibility map Q is updated accordingly. Otherwise, for each superpixel in P_i with responsibility higher than 0, we refine it and call the algorithm recursively (0 is a parameter; we can replace it with a sufficiently low threshold without affecting the quality of the explanation). There algorithm terminates when: 1) the superpixels in P_i are sufficiently refined (containing only very few pixels), or 2) when all superpixels in P_i have the same responsibility (the later condition is for efficiency).

5.3. Compositional explanation algorithm

So far, we assume one particular partition P_i, which Algorithm 2 recursively refines and calculates the corresponding responsibility by calling Algorithm 1. In theory, the refinement in Algorithm 2 can continue until we reach the level of a single pixel. However, even in this case, different partitions may result in different values computed by Algorithm 2, as the partition determines the set of mutants in Algorithm 1. We ameliorate the influence of a particular partition by considering a set of partitions. In Algorithm 3, we consider N partitions and compute an average of the degrees of responsibility induced by each of these partitions. In the algorithm P^x stands for a partition chosen randomly from N possible partitions, and r_p denotes the degree of responsibility of a pixel p w.r.t. P^x.

Algorithm 3 *compositional_explanation*(x)

```
INPUT: an input image $x$, a parameter $N \in \mathbb{N}$
OUTPUT: an explanation $\mathcal{P}^{exp}$

 1: $r_p \leftarrow 0$ for all pixels $p$
 2: for $c$ in 1 to $N$ do
 3:     $P^x \leftarrow$ sample a partition
 4:     $R \leftarrow compositional\_responsibility(x, P^x)$
 5:     for each $P_{i,j} \in$ domain of $R$ do
 6:         $\forall p \in P_{i,j} : r_p \leftarrow r_p + \frac{R(P_{i,j})}{|P_{i,j}|}$
 7:     end for
 8: end for
 9: $pixel\_ranking \leftarrow$ pixels in $\mathcal{P}$ from high $r_p$ to low
10: $\mathcal{P}^{exp} \leftarrow \emptyset$
11: for each pixel $p_i \in pixel\_ranking$ do
12:     $\mathcal{P}^{exp} \leftarrow \mathcal{P}^{exp} \cup \{p_i\}$
13:     $x^{exp} \leftarrow$ mask pixels of $x$ that are not in $\mathcal{P}^{exp}$
14:     if $\mathcal{N}(x^{exp}) = \mathcal{N}(x)$ then
15:         return $\mathcal{P}^{exp}$
16:     end if
17: end for
```

Overall, the explanation algorithm in Algorithm 3 comprises of two parts: ranking all pixels (Lines 1–9) and constructing the explanation (Lines 10–17). The algorithm ranks the pixels of the image according to their responsibility for the model's output. Each time a partition is randomly selected (Line 3), the compositional refinement (Algorithm 2) is called to refine it into a set of fine-grained super pixels and calculate their responsibilities (Line 4). The superpixel's responsibility is evenly distributed to all its pixels, and the pixel-level responsibility is updated accordingly for each sampled partition (Lines 5–7). After N iterations, all pixels are ranked according to their responsibility r_p.

The remainder of Algorithm 3 for constructing an explanation follows the method for explaining the result of an image classifier in [26]. That is, we construct a subset of pixels $\mathcal{P}^{exp}$ to explain $\mathcal{N}$'s output on this particular input x *greedily*. We add pixels to $\mathcal{P}^{exp}$, while $\mathcal{N}$'s output on $\mathcal{P}^{exp}$ does not match $\mathcal{N}(x)$. This process terminates when $\mathcal{N}$'s output is the same as on the whole image x. The set $\mathcal{P}^{exp}$ is returned as the explanation.

Intuitively, our compositional explaining manner seems to be alike human for understanding an input image. In addition, there are several other advantages of the CE (compositional explanation) algorithm in this section.

- Given any partition, the responsibility result returned by Algorithm 1 has its guarantee following the theory in Section 4.1. This distinguish CE with other often random or heuristic based explanation approach.
- The CE approach simple and general. It is blackbox and it automatically refines and detects meaningful super pixels/features (Algorithm 2).
- The explanation algorithm in Algorithm 3 is highly parallel. The analysis of different partitions are independent from each other.

Comparison with existing work In this paragraph, we intend to justify the advantage of the compositional explanation (CE) approach proposed over the existing explanation work in theory. Experimental comparison between different approaches are available in the evaluation section.

The SHAP framework in [19] unifies the explanation theory behind the common and popular explanations of AI models. According to SHAP, to explain an input image with n pixels, in the worst case the number of combinations for masking some pixels while un-masking some other is exponential to the number of n. Though each particular explanation method (LIME [22], GradCAM [24]) will not adopt the wort-case scenario, by approximating the all possible combinations. Still, given the enormous problem space, there lacks a guarantee the important feature will be found. The problem becomes even more challenging when the important features are distributed in distant parts of an image.

Instead, if we think the same explanation problem of an n-pixel input and let us say each time a partition comprises of four super pixels (that is the setup in our evaluation). That is, given a particular partition, the responsibility method in Algorithm 1 at the worst case needs to consider 16 combinations of different kinds of masking. Subsequently, even if we consider the refinement in Algorithm 2 and the multiple partitioning in Algorithm 3, the overall complexity is till polynomial of 16. Thus, from the complexity perspective, CE approach is "easier" to find an explanation.

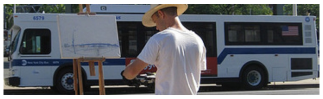


Figure 2: A partially occluded image (from the partial occlusion image data set [17, 27]) that is classified as 'bus' by the compositional net [16]

6. Evaluation

We have implemented the proposed compositional explanation approach in the publicly available tool CET[2]. In the evaluation, we compare CET with recent DNN explanation tools DEEPCOVER, Extremal and RISE. Both the compositional net for partially occluded images [16] (Section 6.2) and convolutional models for ImageNet (Sections 6.3, 6.4) are tested. Three data sets are used in the experiments: compositional images with partial occlusion [17, 27], a "Photobombing" data set with ground truth occlusions and the "Roaming Panda" data set that is a subset of ImageNet images with ground truth explanations [26].

There is no single best way to evaluate the quality of an explanation. Thereby, in this work, we evaluate the explanation quality from three complementary angles: 1) the explanation size, 2) the intersection with the planted occlusion part of an image the intersection of union (IoU) with the ground truth. Intuitively, a good explanation should be a part of the original input and it should not intersect much with the occlusion part of the input image, whereas it should overlap significantly with the ground truth explanation.

Setup When experimenting CET in this section, each partition has four super pixels (as in Algorithm 1). The termination conditions for the compositionality refinement in Algorithm 2 are: 1) the height/width of a super pixel is smaller than $\frac{1}{10}$ of the input image's or 2) the four super pixels share the same responsibility (this is more for efficiency). Inside the compositional explanation (Algorithm 3), N is set to 50.

6.1. Illustrative Example

We start from an example that illustrates how CET works. Figure 2 shows an image that is classified as 'bus' by the compositional net [16], even though there is an occlusion in the middle.

Initially, CET starts from an arbitrary partition of the image with four super pixels, as in Figure 3. This results in 15 combinations of super pixels masking for Algorithm 1 that calculates the responsibility of each super pixel. The refinement of a super pixel happens in Algorithm 2. As in Figure 4, an initial super pixel is further partitioned into four

[2] https://sites.google.com/view/cet-tool/

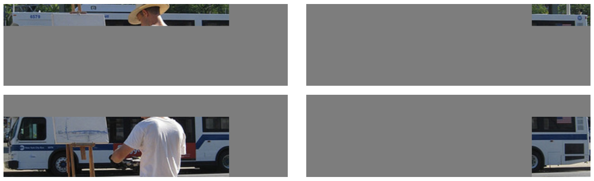


Figure 3: The initial four super pixels chosen by CET in Alg. 3

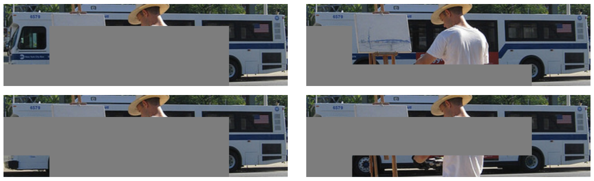
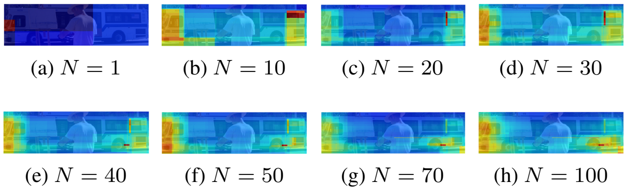


Figure 4: Further refinement of the (left bottom) super pixel in Figure 3 (c), by Alg. 2



(a) $N = 1$ (b) $N = 10$ (c) $N = 20$ (d) $N = 30$

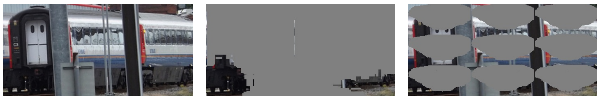
(e) $N = 40$ (f) $N = 50$ (g) $N = 70$ (h) $N = 100$

Figure 5: Improvement of the pixel ranking computed by CET as the number of initial partitions N increases

more fine-grained super pixels. Overall, the refinement done by Algorithm 2 for this particular example goes into two levels before the stopping condition is reached. By starting from this one particular partition, the importance of each pixel is depicted by the heat-map in Figure 5 (a). Though CET still highlights the important feature in the front of the bus, the result is overall coarse. However, when the number of iterations in the CET (Algorithm 3) increases, the result quickly converges as in Figure 5, and these important features identified successfully, avoiding the occlusion in the input image.

6.2. Images with partial occlusions

In this part of the evaluation, we consider explanations for classifications done by the compositional net [16] for partially occluded input images. Due to the lack of ground truth for this data set (Figure 6), we use the (normalised) size of the explanation as proxy for quality. As shown in Figure 7, CET's ranking yields 80% correct classification by when using the Compositional net on only 20% of the pixels of the input image. This is 20% better than what the latest explanation tool DEEPCOVER delivers.



(a) Input (b) CET (c) DEEPCOVER

Figure 6: Explaining the compositional net 'train' using CET and DEEPCOVER

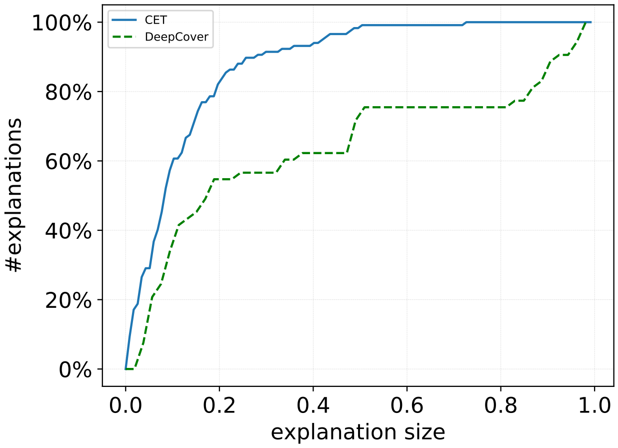
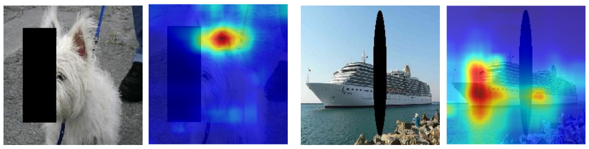


Figure 7: Sizes of explanations for the compositional net



(a) 'West Highland white terrier' (b) 'Ocean liner'

Figure 8: Photobombing images and output from CET

6.3. "Photo bombing" images

Similarly to the images with partial occlusion used in Section 6.2, we create an image data set named "Photo bombing", in which we plant occlusions (aka "photo bombers") into ImageNet images and we record these occlusion pixels so that we can compare the explanation with these pixels. Examples from the Photo bombing data set are given in Figure 8.

Figure 9 reports the explanation results from different tools on the photo bombing images. According to the results in Figure 9, more than 60% of the CET explanations do not overlap at all with the artificially planted occlusions, and this is almost 20% better than the second one, RISE. The explanation size from CET is also consistently smaller than other tools, as in Figure 10. Interestingly, even though DEEPCOVER identifies smaller explanations than RISE, its explanations overlap much more with the occlusions than

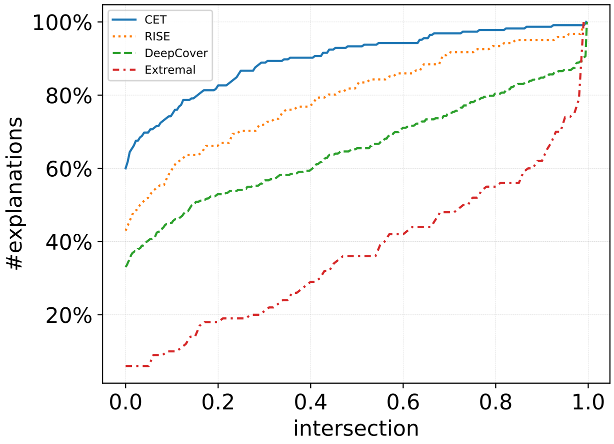


Figure 9: Intersection between the explanation and the occlusion on the Photo bombing dataset (smaller is better)

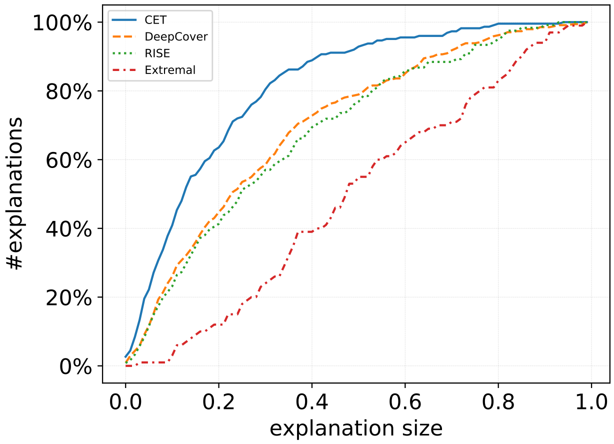


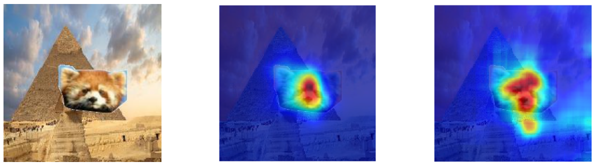
Figure 10: The size of the explanations for the Photobombing dataset (smaller is better)

those generated by RISE. This observation reconfirms that it is good to use multiple metrics to evaluate the quality of explanations.

6.4. ImageNet data with ground truth explanations

Experiments on CET so far shows leading performance on more effectively detecting the important features of input images, which have less intersection with the occlusions. On the other hand, we want to make sure that CET approach also works on "normal" images. In this experiment, we compare CET with other tools using the "Roaming Panda" data set [26], as in Figure 11.

We compare the number of cases that the ground truth explanation is successfully detected, with intersection of union (IoU) larger than 0.5, by each explanation tool. We confirm that CET has the best performance such that more than 70% of the cases the ground truth have been successfully detected (the number in the parenthesis is the percentage of successful detection by each tool): DEEPCOVER (76.7%) > CET (**72.3%**) > Extremal (70.7%) > RISE (55.8%). In contrast to the explanations for partial-occlusion images, DEEPCOVER delivers the best performance on the "roaming panda" data set. However, CET's results are nearly on par, and better than those by all other tools. We observe that with or without occlusion does impact the performance of explanation tools, and CET achieves an overall better performance than other tools that are validated using complementary metrics in the evaluation.



(a) Input (b) CET (VGG16) (c) CET (MobileNet)

Figure 11: The "roaming panda" serves as the ground truth for the label 'red panda'. CET explains it on VGG16 and MobileNet models.

6.5. Threats to validity

The following are the threats to the validity of our claims.

- There is a lack of benchmark images with ground truth explanations and/or occlusions. As a result, we measure the quality of explanations using several *proxy* metrics, including the explanation size and the intersection of the explanation with artificially added occlusion.
- We have set up CET with one particular heuristic configuration that may be overfitted.
- Even though we compare CET with the most recent explanation tools, there exist many other tools that might, in theory, deliver better performance.

7. Conclusion

Motivated by the inherent compositionality in computer vision, this paper proposes a compositional approach for explaining the result of image classifiers. Owing to its compositional approach, CET delivers the best explanations when explaining images with occlusions: its explanations feature the least amount of intersection with the occluded part of the image. CET delivers performance that is on par with the best existing tool (DEEPCOVER) on regular (un-occluded) ImageNet inputs. The algorithm is extremely suitable for parallelization, and it is straight forward to control the precision vs. compute cost trade-off.

References

[1] Elie Bienenstock, Stuart Geman, and Daniel Potter. Compositionality, mdl priors, and object recognition. In *Advances in neural information processing systems*, pages 838–844, 1997. 2

[2] Urszula Chajewska and Joseph Y. Halpern. Defining explanation in probabilistic systems. In *Uncertainty in Artificial Intelligence (UAI)*, pages 62–71. Morgan Kaufmann, 1997. 3

[3] Jianbo Chen, Le Song, Martin Wainwright, and Michael 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

[4] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *J. Artif. Intell. Res.*, 22:93–115, 2004. 3

[5] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (S&P)*, pages 598–617. IEEE, 2016. 1, 2

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

[7] Sanja Fidler, Marko Boben, and Ales Leonardis. Learning a hierarchical compositional shape vocabulary for multi-class object representation. *arXiv preprint arXiv:1408.5516*, 2014. 2

[8] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *International Conference on Computer Vision (ICCV)*, pages 2950–2958. IEEE, 2019. 1, 2

[9] Peter Gärdenfors. *Knowledge in Flux*. MIT Press, 1988. 3

[10] David Gunning. Explainable artificial intelligence (XAI) – program information. https://www.darpa.mil/program/explainable-artificial-intelligence, 2017. Defense Advanced Research Projects Agency. 3

[11] Joseph Y. Halpern. A modification of the Halpern–Pearl definition of causality. In *Proceedings of IJCAI*, pages 3022–3033. AAAI Press, 2015. 3

[12] Joseph Y. Halpern and Judea Pearl. Causes and explanations: a structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56(4), 2005. 2, 3

[13] Joseph Y. Halpern and Judea Pearl. Causes and explanations: a structural-model approach. Part II: Explanations. *British Journal for the Philosophy of Science*, 56(4), 2005. 3

[14] Carl Gustav Hempel. *Aspects of Scientific Explanation*. Free Press, 1965. 3

[15] D. Hume. *A Treatise of Human Nature*. John Noon, London, 1739. 3

[16] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8940–8949, 2020. 6, 7

[17] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1333–1341, 2020. 2, 6

[18] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017. 2

[19] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*, pages 4765–4774. Curran Associates, Inc., 2017. 6

[20] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988. 3

[21] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*. BMVA Press, 2018. 1, 2

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144. ACM, 2016. 1, 2, 6

[23] Wesley C. Salmon. *Four Decades of Scientific Explanation*. University of Minnesota Press, 1989. 3

[24] 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 *International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017. 1, 2, 6

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

[26] Youcheng Sun, Hana Chockler, Xiaowei Huang, and Daniel Kroening. Explaining image classifiers using statistical fault localization. In *ECCV, Part XXVIII*, volume 12373 of *LNCS*, pages 391–406. Springer, 2020. 1, 2, 6, 8

[27] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Vittal Premachandran, and Alan Yuille. Unsupervised learning of object semantic parts from internal states of CNNs by population encoding. *arXiv preprint arXiv:1511.06855*, 2015. 2, 6

[28] Mingqing Xiao, Adam Kortylewski, Ruihai Wu, Siyuan Qiao, Wei Shen, and Alan Yuille. Tdapnet: Prototype network with recurrent top-down attention for robust object classification under partial occlusion. *arXiv preprint arXiv:1909.03879*, 2019. 2

[29] Zhishuai Zhang, Cihang Xie, Jianyu Wang, Lingxi Xie, and Alan L Yuille. Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1372–1380, 2018. 2

[30] Hongru Zhu, Peng Tang, Jeongho Park, Soojin Park, and Alan Yuille. Robustness of object recognition under extreme occlusion in humans and computational models. *arXiv preprint arXiv:1905.04598*, 2019. 2