Two-Stage Holistic and Contrastive Explanation of Image Classification

Weiyan Xie [*1] **Xiao-Hui Li** [2] **Zhi Lin** [1] **Leonard K. M. Poon** [3] **Caleb Chen Cao** [†1] **Nevin L. Zhang** [*1]

[1] The Hong Kong University of Science and Technology, Hong Kong, China
[2] Huawei Technologies Co., Ltd, Shenzhen, China
[3] The Education University of Hong Kong, Hong Kong, China[‡]

Abstract

The need to explain the output of a deep neural network classifier is now widely recognized. While previous methods typically explain a single class in the output, we advocate explaining the whole output, which is a probability distribution over multiple classes. A whole-output explanation can help a human user gain an overall understanding of model behaviour instead of only one aspect of it. It can also provide a natural framework where one can examine the evidence used to discriminate between competing classes, and thereby obtain contrastive explanations. In this paper, we propose a contrastive whole-output explanation (CWOX) method for image classification, and evaluate it using quantitative metrics and through human subject studies. The source code of CWOX is available at `https://github.com/vaynexie/CWOX`.

1 INTRODUCTION

The past few years have witnessed a surge of research activities on the explainability of deep neural networks, which is driven by the need for trust, fairness and accountability in high-stake applications [Samek et al., 2019, Li et al., 2020]. While there is some work on *ante hoc methods* that learn interpretable models to begin with [Zhang et al., 2018b], most efforts are spent on *post hoc methods* that explain complex models whose behaviours are not self-interpretable [Samek et al., 2019, Li et al., 2020]. A common way to explain image classification is to generate a saliency map that assigns a numerical value to each pixel to indicate its importance to an output class label. A variety of methods have been proposed [Simonyan et al., 2014, Springenberg et al., 2015, Zeiler and Fergus, 2014, Bach et al., 2015, Ribeiro et al., 2016, Shrikumar et al., 2017, Zhang et al., 2018a, Petsiuk et al., 2018]. Most methods are designed to explain one single output label, and hence we call them *individual output explanation (IOX)* methods.

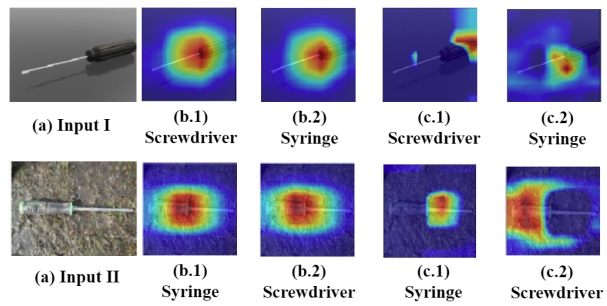


Figure 1: Explanations of the outputs of GoogleNet on two input images: (b.1-2) Grad-CAM is applied to each top output class separately (SWOX); (c.1-2) The top output classes are contrasted against each other (CWOX).

IOX methods are unable to provide users with an overall understanding of model behavior [Kim and Doshi-Velez, 2021], and might mislead users to unjustified confidence in the explanation and the model [Rudin, 2019, Adebayo et al., 2018, 2022]. Consider the two input images in Fig. 1, both with ground-truth label `screwdriver`. The outputs of GoogleNet [Szegedy et al., 2015] are {`screwdriver` (0.49), `syringe` (0.38)} (*input I*), and {`syringe` (0.50), `screwdriver` (0.38)} (*input II*) respectively. The saliency maps created using Grad-CAM [Shrikumar et al., 2017] for the output classes are shown in (b.1-2).

Consider two scenarios for the first input in Fig. 1 : (1) Present a user only with the heatmap for the top class (b.1), or (2) present a user with the heatmaps for both top classes (b.1-2). Clearly, the user would gain a better understanding of the model in the second scenario and realize that the model has difficulty in discriminating `screwdriver` and `syringe`. In addition, the user would realize that the

[*]Equal Contribution.
[†]This work is done when Caleb was in Huawei Research HK.
[‡]Previous affiliation where the work was carried out.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

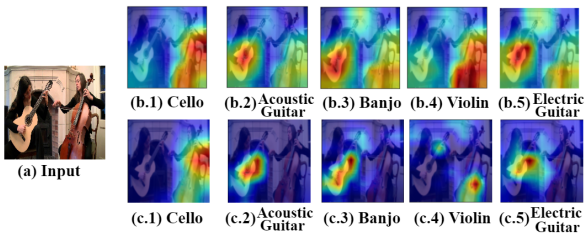


Figure 2: The output of ResNet50 on the input image includes 5 top classes. The top row (b.1-5) shows their SWOX saliency maps while the bottom row (c.1-5) show their CWOX-1s saliency maps.

two heatmaps, being almost identical, do not help understand what evidence the model uses to discriminate the two classes. To appreciate the point better, imagine a scenario where a user is presented with the two heatmaps and the two labels *separately*, and is asked to match them. This would be virtually an impossible task. The same is true for the second input, where the order of the top 2 labels is reversed.

It is clear that we need *whole-output explanation (WOX)* methods that explain all top output classes. It is also evident that a *simple WOX (SWOX)* method, which explains the top classes one by one independently, is not sufficient. It is necessary to reveal the evidence that supports each top class against other top classes [Wang and Vasconcelos, 2020]. This leads to what we call *Contrastive Whole-Output Explanation (CWOX)*. For the first example in Fig. 1, the CWOX explanations are shown in top row (c.1-2). We see that the handle is highlighted for `screwdriver` and the shaft is highlighted for `syringe`. Those can evidently help a user understand why there are two possible output classes instead of one, and correctly match the heatmaps with the label in the case where there are presented separately. The same is true for the second example.

Images often contain multiple objects of interest. Compared with those with a single object, such images usually lead to more classes with significant probabilities in model output. For example, the output of ResNet50 [He et al., 2016] on the input image shown in Fig. 2 consists of 5 top classes: `cello` (0.839), `acoustic-guitar` (0.081), `banjo` (0.036), `violin` (0.021), `electric-guitar` (0.008). From the SWOX saliency maps (top row in Fig. 2), we see that different top classes (e.g., `cello` and `violin`) might refer to the same object in the input image and are competing labels for that object. Such classes are *confusing* to the classifier in the sense that the classifier is uncertain as to which of the classes to use when labeling the object.

Our main contribution in this paper is to show that the quality of explanations can be substantially improved by utilizing this observation. Specifically, we propose to divide the top class labels into confusion clusters based on the object they refer to, and perform the explanation in two

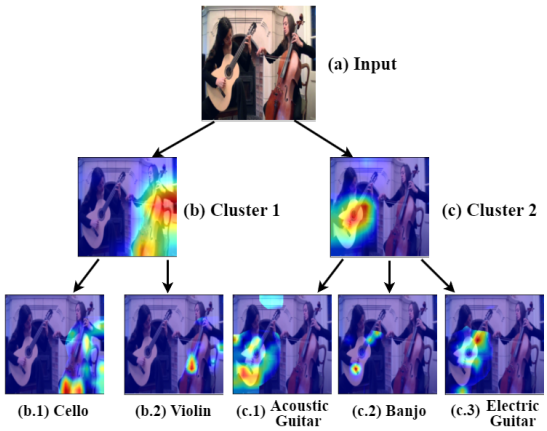


Figure 3: CWOX-2s explanation of the output of ResNet on the input image (with Grad-CAM as the base explainer).

steps: (1) Generate heatmaps to contrast different confusion clusters, and (2) generate heatmaps to contrast classes within each cluster. We call this method *two-stage contrastive whole-output explanation (CWOX-2s)*. On the other hand, the method alluded to in previous paragraphs contrasts each class directly against all other classes. We call it *one-stage contrastive whole-output explanation (CWOX-1s)*. Note that CWOX-2s reduces to CWOX-1s when there is only one confusion cluster, as in the case of Fig. 1. However, CWOX-2s makes significantly different explanations when there are more than one confusion clusters.

For instance, in the example shown in Fig. 2, CWOX-2s divides the top five classes into two clusters: {`cello`, `violin`} and {`acoustic guitar`, `banjo`, `electric guitar`}. It first contrasts the two clusters, and then contrasts classes within each cluster against the other classes in the same cluster. This approach is more reasonable than CWOX-1s. In Fig. 2, it is clear that `violin` should have more contrastive value to `cello` than other classes. This observation is ignored by CWOX-1s.

The explanation given by CWOX-2s is as shown in Fig. 3. It first shows that evidence for the two clusters comes from the left and right part of the input image, respectively. `Cello` and `violin` are competing labels for the right part of the image. The evidence that supports `cello` relative to `violin` is the body bottom of the instrument (b.1), and the evidence that supports `violin` relative to `cello` is the middle section of the strings (b.2). Those make sense intuitively because cellos have large bottoms and the middle section of the strings on a cello is visually similar to that on a violin. The supportive evidence for the three labels in the other cluster relative to each other are displayed in (c.1) (lower body), (c.2) (bridge), and (c.3) (strings), respectively. Those are intuitively more informative than the heatmaps by CWOX-1s shown in the second row of Fig. 2. Later we will show that CWOX-2s is superior to SWOX and CWOX-1s in both quantitative evaluations and human subject studies.

2 RELATED WORK

CWOX-2s aims to provide contrastive explanations for the top predicted classes. There are previous works on contrastive explanations. Miller [2019] surveyed over 250 papers in philosophy, psychology, and cognitive science and found that humans prefer contrastive explanations that explain *why class A but not class B* to non-contrastive ones that only explain *why A*. In XAI, this is achieved through *counterfactual explanation* or *discriminative explanation*. Counterfactual explanations identify necessary modifications to change the prediction from A to B [Wachter et al., 2017], while discriminative explanations provide the evidence in the input that supports A over B [Wang and Vasconcelos, 2020, Prabhushankar et al., 2020, Jacovi et al., 2021]. As illustrated in Fig. 3, CWOX-2s is a systematic and organized way to apply discriminative explanation to the top classes in the classification output.

In both types of contrastive explanations, there is a need to identify a *contrast class (foil) B* for the *target class (fact) A*. Previous works let the foil be: (1) all other classes (i.e., non-A) [Zhang et al., 2018a, Jacovi et al., 2021]; (2) any other class [Dhurandhar et al., 2018, Goyal et al., 2019, Wang and Vasconcelos, 2020]; (3) the class with second highest probability [Wang and Wang, 2022]; (4) another class picked by users [Liu et al., 2019, Akula et al., 2020]; or (5) the prediction of another smaller model [Wang and Vasconcelos, 2020]. In CWOX-2s, we propose a principled method for determining how to contrast the top classes against each other. Specifically, we divide the top classes into confusion clusters. We first contrast different confusion clusters against each other, and then contrast different classes within the same confusion cluster.

In XAI literature, methods explaining "*why class A*" and methods explaining "*why class A but not class B*" are regarded as two separate lines of work. The first line of work is essentially about localizing the object (or region) that class A refers to [Selvaraju et al., 2017, Shrikumar et al., 2017, Zhang et al., 2018a]. The second line of work is about deciding whether the object in focus belongs to class A or B [Dhurandhar et al., 2018, Prabhushankar et al., 2020, Goyal et al., 2019, Wang and Vasconcelos, 2020]. The latter is carried out in the context of fine-grained image classification. CWOX-2s can be viewed as a junction where the two lines of work meet. The first step of CWOX-2s is about object localization, and the second step targets class discrimination.

3 GROUPING CLASS LABELS INTO CONFUSION CLUSTERS

As mentioned in Section 1, the notion of confusion clusters is of pivotal importance to CWOX-2s. The question arises: how do we divide the top output labels for a given input into confusion clusters? A straightforward approach is to examine the IOX heatmaps for all the top labels and group two labels together into the same cluster if their IOX heatmaps overlap substantially. A threshold is required for this approach. We have found it difficult to determine a threshold that suits all cases. Consequently, we step back and ask how to tell if two classes are confusing to a classifier without using XAI? One answer is described below. It is one of the main innovative aspects of this paper.

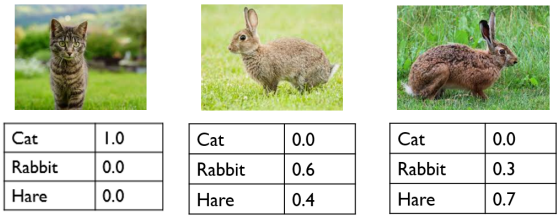


Figure 4: Rabbit and hare are confusing to humans. A person would find it difficult to decide whether to label the second and third images as rabbit or hare. Similarly, a neural network classifier would give the two class labels high probabilities in either case.

Rabbit and hare are confusing to humans. When presented with an image of either class, a person would find it difficult to decide whether to label it as a rabbit or a hare (Fig. 4). Similarly, two classes are confusing to a classifier if, when processing images containing objects of either class, it has trouble determining which of them to use as the output label. Consequently, it gives high probabilities to both classes. It is therefore possible to determine if two classes are confusing to a classifier by checking if they *often* co-occur as top classes in classification outputs.

To partition classes into confusion clusters, we first run the target classifier on a set of examples, typically the training examples. For each example, we get a list of top class labels, which we regard as a short document. For the example in Fig. 3, the document consists of five words: {`cello`, `acoustic-guitar`, `banjo`, `violin`, `electric-guitar`}. Suppose there are N training examples. Then we have N short documents. The task now becomes a word clustering problem. We want to partition the words (class labels) into clusters such that words from the same cluster co-occur more often in the N documents than words from different clusters.

There are many methods that can be used for word clustering. We choose to use hierarchical latent tree analysis (HLTA) [Chen et al., 2017, Zhang and Poon, 2017] because it is developed specifically to model word co-occurrence in documents. We leave it to future work to evaluate other methods for this step.

HLTA is based on hierarchical latent tree models (HLTM), which are Bayesian networks with multiple levels of latent variables. An example is shown in Fig. 5. The idea is

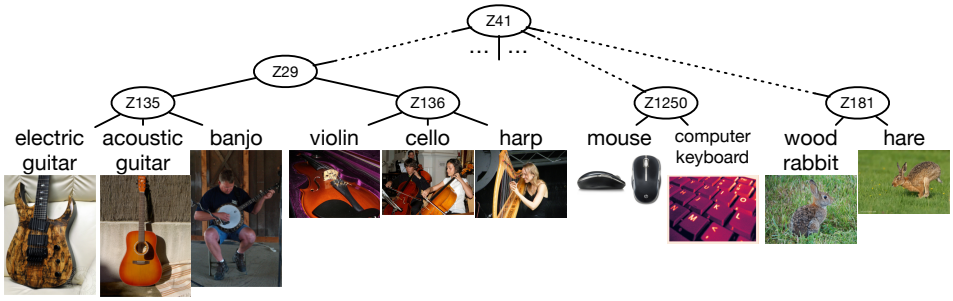


Figure 5: A part of the latent tree model built from the outputs of ResNet50 on ImageNet training examples. Solid lines are direct connections, and dashed lines are indirect connections with intermediate nodes removed. Images of the classes are displayed for visual reference. They are not part of the model. The tree reveals co-occurrence patterns of class labels in classification outputs.

to model correlations among observed variables (the leaf nodes) using a tree of latent variables. Given a dataset on the observed variables, HLTA aims to find the model that maximizes the Bayes Information Criterion [Schwarz, 1978].

We performed HLTA on a collection of short documents obtained using ResNet50 on the training examples of ImageNet. Fig. 5 shows a part of the structure of the resulting model. [1] The variables at the bottom level, level 0, are binary variables that represent the presence/absence of words in a document. The latent variables at level 1 are introduced during the analysis to model word co-occurrence patterns, e.g., $Z181$ for the co-occurrence of `hare` and `wood-rabbit`, and $Z1250$ for the co-occurrence of `mouse` and `computer-keyboard`. Latent variables at level 2 are introduced during the analysis to model the co-occurrence of the patterns at level 1, e.g., $Z29$ for the co-occurrence of the patterns $Z135$ and $Z136$.

Each node in the tree defines a cluster of class labels, which consists of the labels in the subtree rooted at the node. Some of the clusters given by level-1 nodes, for instance {`hare`, `wood-rabbit`} and {`electric-guitar`, `accoustic-guitar`, `banjo`}, consist of visually similar classes that are difficult for the classifier to discriminate. They are often competing labels for the same object/region in the input image, and hence all appear as top classes in classification output.

Class labels in some other clusters are not visually similar. One example is {`mouse`, `computer-keyboard`}. The two classes are grouped together nonetheless because mouses and keyboards tend to co-occur in images, and hence co-occur as top classes in classification output. Due to the co-occurrence, classifiers often have trouble in deciding which of them to use to label an image. If we think of a *composite object* `mouse+computer-keyboard`, then this co-occurrence cluster is no different from the visually similar clusters above: Different labels in the cluster are competing labels for the same (composite) object. See Fig. 8 for example of explanations on the composite object.

There is a hierarchy behind the ImageNet classes that was derived from WordNet [Miller, 1995]. While there are some similarities, our latent tree differs from the WordNet hierarchy significantly. For example, `screwdriver` and `syringe` are far apart in WordNet, but close to each other in our latent tree due to their visual similarity. See Fig. 1.

4 CREATING CONTRASTIVE EXPLANATIONS

Suppose we want to explain the behaviors of a classification model m. In our approach, the first step is to build a latent tree T for all the class labels as described in the previous section. This is done in an *offline phase*.

During the *online phase*, we create explanations for the outputs of m on individual inputs. For each input image $\mathbf{x}$, we feed it to m to get the K top classes in the output. The value of K can either be a predetermined number (e.g.,5), or the number of top classes whose total probability exceeds a threshold (e.g., 0.95).

To divide the top classes into confusion clusters, we first restrict the latent tree T from the offline phase onto those classes to obtain a subtree, and then cut the subtree at level 1 to get clusters of labels. In our running example in Fig. 3, there are 5 top classes: `cello`, `violin`, `acoustic-guitar`, `banjo` and `electric-guitar`. By restricting the latent tree in Fig. 5 onto those classes and cutting the resulting subtree at level 1, we get the following two clusters: {`cello, violin`} and {`acoustic-guitar, banjo, electric-guitar`}.

In general, suppose the K top classes are divided into I confusion clusters $\mathbf{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_I\}$, and each cluster $\mathbf{C}_i$ consists of J_i class labels $\mathbf{C}_i = \{c_{i1}, \ldots, c_{iJ_i}\}$. To explain the top classes, CWOX-2s generates a collection of *contrastive heatmaps* in two stages:

[1] The entire model structure will be posted online later.

1. For each confusion cluster $\mathbf{C}_i$, create a heatmap to highlight the pixels that support $\mathbf{C}_i$ over other clusters;
2. In each $\mathbf{C}_i$, create a heatmap for each class c_{ij} to highlight the pixels that support c_{ij} over other classes.

4.1 BASE EXPLAINERS

In CWOX-2s, contrastive heatmaps are created from saliency maps for individual classes. A *saliency map* for a classes c aims to highlight the pixels that are, according to the model m, important for the class. The more important a pixel is to the class, the higher its saliency value. It is usually computed from either the probability $P_m(c|\mathbf{x})$ or the logit $z_c(\mathbf{x})$ of the class. Saliency maps can be generated by a variety of IOX methods, including backpropagation-based techniques such as Guided Backpropagation [Springenberg et al., 2015], DeepLIFT [Shrikumar et al., 2017], Grad-CAM [Selvaraju et al., 2017]; forward propagation-based techniques like RISE [Petsiuk et al., 2018], and local approximation methods like LIME [Ribeiro et al., 2016]. They will be referred to as *base explainers* in the context of CWOX-2s.

The concept of saliency map can easily be generalized to clusters of classes. A cluster $\mathbf{C}$ of classes can be viewed as a *compound class* with probability and logit given as follows:

$$P_m(\mathbf{C}|\mathbf{x}) = \sum_{c \in \mathbf{C}} P_m(c|\mathbf{x}), \quad z_{\mathbf{C}}(\mathbf{x}) = \log \sum_{c \in \mathbf{C}} e^{z_c(\mathbf{x})}.$$

Saliency maps can be generated for the cluster $\mathbf{C}$ in the same way as for individual classes.

4.2 CONTRASTIVE HEATMAPS

Let $H_{\mathbf{C}_i}$ and $H_{\mathbf{C}\backslash\mathbf{C}_i}$ be saliency maps for a confusion cluster $\mathbf{C}_i$ and the union of all other confusion clusters respectively. $H_{\mathbf{C}_i}$ and $H_{\mathbf{C}\backslash\mathbf{C}_i}$ presumably highlight the pixels that are, according to the model m, important for $\mathbf{C}_i$ and $\mathbf{C}\backslash\mathbf{C}_i$ respectively. For a given pixel x, the difference $H_{\mathbf{C}_i}(x) - H_{\mathbf{C}\backslash\mathbf{C}_i}(x)$ measures the importance of x to $\mathbf{C}_i$ relative to $\mathbf{C}\backslash\mathbf{C}_i$. Consequently, we use the following heatmap to contrast $\mathbf{C}_i$ against other confusion clusters:

$$\hat{H}_{\mathbf{C}_i} = \begin{cases} ReLU[H_{\mathbf{C}_i} - H_{\mathbf{C}\backslash\mathbf{C}_i}] & \text{if } I > 1; \\ H_{\mathbf{C}_i} & \text{if } I = 1, \end{cases} \quad (1)$$

Note that ReLU is used so as to focus on the evidence for cluster $\mathbf{C}_i$ rather than that against it.

Next, consider all the classes in a confusion cluster $\mathbf{C}_i$. Let $H_{c_{ij}}$ and $H_{\mathbf{C}_i\backslash c_{ij}}$ be saliency maps for a class $c_{ij} \in \mathbf{C}_i$ and all other classes in the same cluster respectively. We use the following heatmap to contrast c_{ij} against the other classes:

$$\hat{H}_{c_{ij}} = \begin{cases} \text{supp}(\hat{H}_{\mathbf{C}_i}) \times ReLU[H_{c_{ij}} - H_{\mathbf{C}_i\backslash c_{ij}}] \\ \text{if } J_i > 1; \\ \text{supp}(\hat{H}_{\mathbf{C}_i}, \epsilon) \times H_{c_{ij}} \\ \text{if } J_i = 1. \end{cases} \quad (2)$$

Algorithm 1 CWOX-2s

I. OFFLINE PHASE

Input: A classification model m; a dataset S.

Do:

1: Feed each example $\mathbf{x}$ in S to m to get a list (document) of top class labels.
2: Run HLTA on the documents to get a latent tree T.

II. ONLINE PHASE

Input: A test example $\mathbf{x}$; a base explainer.

Do:

1: Feed $\mathbf{x}$ to m to get a list of top class labels.
2: Restrict T to those labels to get a subtree
3: Partition the labels into confusion clusters by cutting the subtree at level 1.
4: Create a heatmap to contrast each confusion cluster against other clusters using Equation (1).
5: In each cluster, create a heatmap to contrast each class in the cluster against other classes using Equation (2).

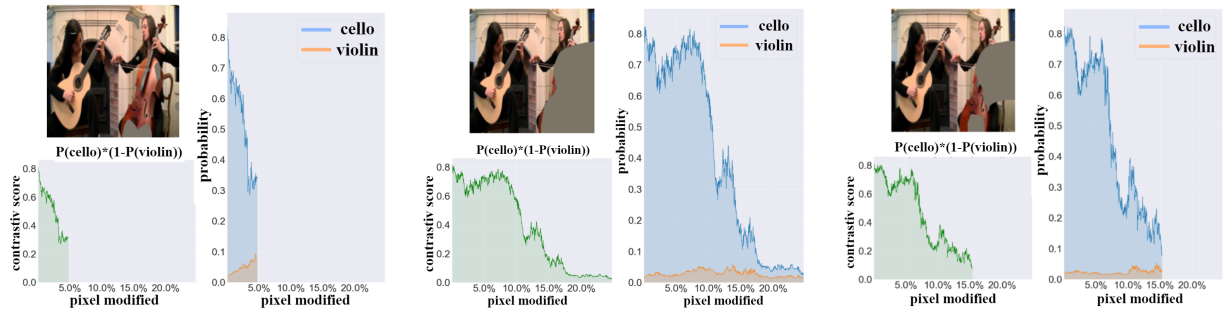
Note that the contrastive heatmap for c_{ij} is restricted to $\text{supp}(\hat{H}_{\mathbf{C}_i})$. This means that, when identifying contrastive evidence for classes in the $\mathbf{C}_i$, we focus only on the evidence supportive of the cluster.

An overall description of CWOX-2s is given in Algorithm 1. As alluded to earlier, a confusion cluster $\mathbf{C}_i$ consists of classes that are competing labels for the same region in the input image. The first step of CWOX-2s aims to highlight that region, and hence is about object localization. The second step of CWOX-2s aims to pinpoint at the evidence for each of the competing classes. It is about class discrimination. Class discrimination requires more fine-grained information. Some base explainers can facilitate this desiderata. For instance in Grad-CAM, one needs to specify a pivot layer where multiple feature maps are aggregated into one heatmap using gradients from the output layer. The further away the pivot layer is from the output layer, the more fine-grained is the heatmap. In RISE, one needs to specify a mask size and a pixel mask probability. The smaller the mask size and the pixel mask probability, the more fine-grained the resulting heatmap.

The idea of subtracting two saliency maps to create a contrastive heatmap was first proposed by Shrikumar et al. [2017], Zhang et al. [2018a]. Alternatively, one can multiply one saliency map with the “inverse" of the other [Wang and Vasconcelos, 2020].

5 EMPIRICAL EVALUATIONS

In this section, we evaluate CWOX-2s against several other WOX methods to explain all top classes. The evaluations are in terms of the faithfulness and interpretability of the explanations. Here, *faithfulness* refers to an explanation’s ability



(i) CWOX-2s: $n_\delta = 2,377$, CAUC = 0.023, CDROP = 0.505; (ii) SWOX: $n_\delta = 12,402$, CAUC = 0.033, CDROP = 0.309; (iii) CWOX-1sA: $n_\delta = 7,686$, CAUC = 0.033, CDROP = 0.364.

Figure 6: Changes in the probabilities $P(\texttt{cello})$ and $P(\texttt{violin})$ and the contrastive score $P(\texttt{cello}) \times (1 - P(\texttt{violin}))$ as δ-salient pixels are deleted according to the order induced by: (i) the CWOX-2s heatmap in Fig. 3 (b.1); (ii) the SWOX heatmap in Fig. 2 (b.1); and (iii) the CWOX-1sA heatmap in Fig. 2 (c.1).

to accurately reflect the function learned by the model [Selvaraju et al., 2017, Petsiuk et al., 2018], while *interpretability* refers to its ability to provide a clear understanding of the relationship between input and output for human users [Ribeiro et al., 2016, Doshi-Velez and Kim, 2017].

We have presented three methods, SWOX, CWOX-1s and CWOX-2s, for explaining all top classes. CWOX-1s has two possible variants. CWOX-1sA obtains a heatmap for each class by subtracting saliency maps, i.e., $ReLU[H_c - H_{\mathbf{C}\setminus c}]$, similar to the contrastive heatmaps created in CWOX-2s. On the other hand, CWOX-1sB multiplies H_c with the "inverse" of $H_{\mathbf{C}\setminus c}$. The second variant was proposed earlier in [Wang and Vasconcelos, 2020], where it is called *SCOUT*. As will be seen, CWOX-1sA significantly outperforms the CWOX-1sB. Hence, we do not consider the B-variant of CWOX-2s.

To evaluate CWOX-2s and the three baselines, we use them to explain the outputs of GoogleNet [Szegedy et al., 2015] and ResNet50 [He et al., 2016] on a subset of randomly selected 10,000 images from the ImageNet validation set [Deng et al., 2009]. For each image, we apply the WOX methods to explain its top K predicted classes with $K = \min\{5, Cum(0.95)\}$, where $Cum(0.95)$ is a function to return the smallest number of top classes with a cumulative probability greater than 0.95. Two base explainers, namely Grad-CAM [Selvaraju et al., 2017] and RISE [Petsiuk et al., 2018], are used in the experiments.

5.1 FAITHFULNESS TO MODEL

Rationale for Evaluation Metrics: An IOX method aims to reveal the evidence a model relies on to predict a particular class. IOX explanations (saliency maps) are often evaluated in terms of their *faithfulness* to a model. Ideally, a faithful saliency map should highlight important pixels for the class, and removing pixels with high saliency values should decrease the class probability. This concept gives rise to a widely-used metric, the *deletion AUC* metric [Samek et al., 2016, Petsiuk et al., 2018].

Different from IOX, CWOX-2s aims to reveal the evidence that a model uses to discriminate between classes. Consequently, CWOX-2s explanations should be evaluated in terms of their *contrastive faithfulness* to a model m, i.e., how effective they are at revealing the evidence that the model relies on to discriminate between different classes.

Suppose that a model m has reasons to believe that an input $\mathbf{x}$ belongs to a class c, but cannot rule out the possibility of it belonging to some other classes $\mathbf{C}'$. If a heatmap H is contrastively faithful to m, then it should give high values to the pixels that m considers strongly supportive of c relative to $\mathbf{C}'$. The deletion of such high-value pixels should lead to fast decrease in the probability of c and an increase in that of $\mathbf{C}'$. To be more specific, let there be totally n pixels, and $x_1, \ldots, x_n$ be an enumeration of pixels in descending order of $H(x)$. Let $\mathbf{x}_{[r,n]}$ be the resulting image of deleting the first $r-1$ pixels from the input image $\mathbf{x}$. If H is contrastively faithful to m, then the probability $P_m(c|\mathbf{x}_{[r,n]})$ would decrease quickly with r and $P_m(\mathbf{C}'|\mathbf{x}_{[r,n]})$ would increase with it. Thus, the *contrastive score* defined below would decrease quickly:

$$s(r) = P_m(c|\mathbf{x}_{[r,n]})(1 - P_m(\mathbf{C}'|\mathbf{x}_{[r,n]})). \tag{3}$$

As an example, consider the CWOX-2s heatmap (b.1) in Fig. 3. It presumably reveals the evidence that ResNet50 considers supportive of `cello` relative to `violin`. Fig. 6 (i) shows what happens when pixels are deleted from the input image according to order induced by the CWOX-2s heatmap. We see that $P(\texttt{cello})$ decreases and $P(\texttt{violin})$ increases. Consequently, the contrastive score $P(\texttt{cello}) \times (1 - P(\texttt{violin}))$ decreases. Fig. 6 (ii) shows what happens when pixels are deleted according to the order induced by SWOX heatmp shown in Fig. 2 (b.1). We see that, compared to the CWOX-2s heatmap, the contrastive score of the SWOX heatmap drops more slowly at the beginning. Although a bigger drop is achieved later, it is at the expense of

Table 1: Average CAUC scores on the ImageNet examples (**smaller** ↓ CAUC indicates better contrastive faithfulness).

	ResNet50		GoogleNet	
	Grad-CAM	RISE	Grad-CAM	RISE
SWOX	7.54×10^{-3}	5.18×10^{-3}	5.93×10^{-3}	3.36×10^{-3}
CWOX-1sA	7.19×10^{-3}	4.65×10^{-3}	5.37×10^{-3}	3.12×10^{-3}
CWOX-1sB	7.68×10^{-3}	4.96×10^{-3}	6.12×10^{-3}	3.24×10^{-3}
CWOX-2s	$\mathbf{5.78 \times 10^{-3}}$	$\mathbf{4.08 \times 10^{-3}}$	$\mathbf{4.47 \times 10^{-3}}$	$\mathbf{2.78 \times 10^{-3}}$

Table 2: Average CDROP scores on the ImageNet examples (**larger** ↑ CADROP indicates better contrastive faithfulness).

	ResNet50		GoogleNet	
	Grad-CAM	RISE	Grad-CAM	RISE
SWOX	6.84×10^{-2}	8.19×10^{-2}	6.56×10^{-2}	7.68×10^{-2}
CWOX-1sA	7.01×10^{-2}	8.35×10^{-2}	6.59×10^{-2}	7.73×10^{-2}
CWOX-1sB	5.22×10^{-2}	7.01×10^{-2}	5.05×10^{-2}	6.32×10^{-2}
CWOX-2s	$\mathbf{8.21 \times 10^{-2}}$	$\mathbf{8.97 \times 10^{-2}}$	$\mathbf{7.64 \times 10^{-2}}$	$\mathbf{8.32 \times 10^{-2}}$

Table 3: Performances of CWOX-2s with four base explainer. Here, $\bar{n}_\delta$ stands for average number of δ-salient pixels.

	ResNet50			GoogleNet		
Base Explainer	$\bar{n}_\delta$	CAUC ↓	CDROP ↑	$\bar{n}_\delta$	CAUC ↓	CDROP ↑
Grad-CAM	2,029	3.11×10^{-3}	8.01×10^{-2}	2,181	1.80×10^{-3}	7.46×10^{-2}
MWP	4,194	3.12×10^{-3}	7.37×10^{-2}	3,026	1.77×10^{-3}	5.85×10^{-2}
LIME	2,464	3.40×10^{-3}	4.89×10^{-2}	2,351	1.92×10^{-3}	3.64×10^{-2}
RISE	1,282	$\mathbf{3.07 \times 10^{-3}}$	$\mathbf{8.97 \times 10^{-2}}$	1,105	$\mathbf{1.72 \times 10^{-3}}$	$\mathbf{8.32 \times 10^{-2}}$

deleting many more pixels. Those indicate that the SWOX heatmap is less effective than the CWOX-2s at pinpointing at the evidence that ResNet50 relies on to discriminate `cello` from `violin`. Additionally, Fig.6 (iii) shows the results when pixels are removed based on the order suggested by the CWOX-1sA heatmap shown in Fig.2 (c.1). Although it has a smaller n_δ and a larger CDROP compared to SWOX, its overall performance is still notably inferior to the CWOX-2s results that are shown in (i).

Evaluation Metrics: We propose two quantitative metrics to evaluate the contrastive faithfulness of a heatmap to the target model. The first one is the area under the contrastive score curve, or *contrastive AUC (CAUC)* for short:

$$CAUC(H, m|\mathbf{x}, c, \mathbf{C}') = \frac{1}{n}\sum_{r=1}^{n_\delta} s(r), \qquad (4)$$

where $s(r)$ is the contrastive score defined in Equation (3) and n_δ is the number of *δ-salient pixels*. A pixel x is considered δ-salient if its saliency value, denoted by $H(x)$, is greater than or equal to δ times the maximum saliency value across all pixels in the heatmap H. In other words, such pixels have a saliency value of $H(x) \geq \delta \max_{x' \in \mathbf{x}} H(x')$. The rationale behind the δ-salient pixels is twofold: First, it enables the evaluation to concentrate on the most salient pixels, and second, it helps to exclude the numerous zero-valued pixels that are frequently presented in CWOX-2s heatmaps and lack any meaningful ordering. In our experiments, δ is set to 0.5.

Similar to the deletion AUC, smaller CAUC scores indicate better heatmaps in terms of contrastive faithfulness. In Fig. 6, the curves are shown only for the first n_δ pixels. Different heatmaps may have different n_δ. CAUC scores computed over different numbers of pixels are not comparable. Consequently, when comparing two or more heatmaps, we use the minimum of numbers of salient pixels to calculate the CAUC scores for all of the heatmaps.

One drawback of the CAUC metric is that, when comparing two heatmaps, it does not consider all the δ-salient pixels in some of the heatmaps. To take all the δ-salient pixels into consideration, we propose another metric called the *weighted drop in contrastive score (CDROP)*:

$$CDROP(H, m|\mathbf{x}, c, \mathbf{C}') = \frac{s(1) - s(n_\delta + 1)}{\log_2(1 + \frac{\max\{n_\delta, \tau\}}{\tau})}, \qquad (5)$$

where τ is a hyperparameter. The score is a combination of two factors. The first factor $s(1) - s(n_\delta + 1)$ is the drop in the contrastive score due to the deletion of all the salient pixels. The second factor $\log_2(1 + \frac{\max\{n_\delta, \tau\}}{\tau})$ is a logarithmic penalty factor for n_δ when it exceeds τ, which is set at $0.05n$ in our experiments (i.e., 5% of the total number of pixels). It captures the intuition that too many salient pixels can be distracting to a human user, and is motivated by the Weber-Fechner law. This Weber-Fechner law posits that the intensity of human sensation grows in proportion to the *logarithm* of an increase in energy, rather than increasing at the same rate as the energy. Larger CDROP scores indicate

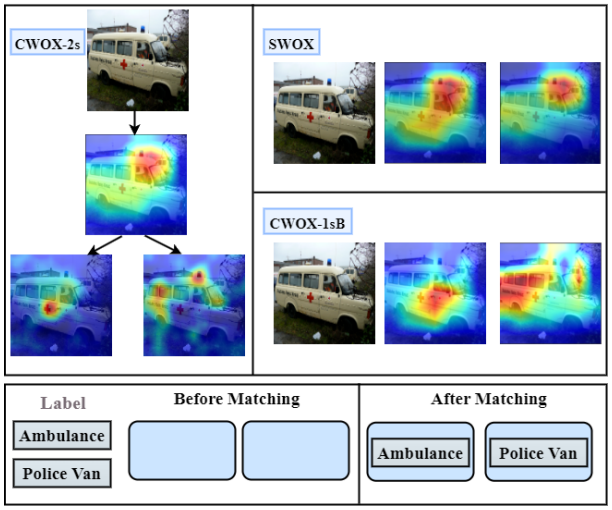


Figure 7: In the user study, heatmaps for pairs of confusing labels are displayed. A user is asked to match the labels with the heatmaps.

better heatmaps in terms of contrastive faithfulness.

Results: Tab. 1 presents the CAUC scores of the four methods. These scores are averaged across all 10,000 test images, and for each test image, all pairs $(c_{ij}, \mathbf{C}_i \backslash c_{ij})$ are considered. These pairs are composed of a top class c_{ij} and all other classes $\mathbf{C}_i \backslash c_{ij}$ from the same confusion cluster $\mathbf{C}_i$. Tab. 2 show the corresponding CDROP scores. We see that the CAUC scores of CWOX-2s are significantly lower than those of the baselines, and its CRDOP scores are significantly higher. Those indicate that the explanations produced by CWOX-2s are more contrastively faithful to the models than the baselines. It is also interesting to note that CWOX-1sA is inferior to CWOX-1sB in all cases, and sometimes it is even inferior to SWOX.

Apart from comparing various WOX methods, Tab. 3 presents a performance comparison of CWOX-2s using four different IOX methods as base explainers, including two backpropagation-based methods, Grad-CAM [Selvaraju et al., 2017] and MWP [Zhang et al., 2018a], as well as two forward propagation-based methods, LIME [Ribeiro et al., 2016] and RISE [Petsiuk et al., 2018]. We can see that RISE outperforms others in contrastive faithfulness metrics, with the lowest CAUC score and the highest CDROP score. It also has the fewest salient pixels, which indicates the precision in identifying crucial evidence. Grad-CAM follows, then MWP, which produces numerous salient pixels and is less precise. LIME has the weakest performance among the four. Appendix A details the setup for each base explainer in CWOX-2s and provides examples for comparing the contrastive faithfulness across them.

Table 4: Results of the user study in the expert group ($\pm$ 95% confidence interval).

	SWOX	CWOX-1sB	CWOX-2s
Accuracy	0.45±0.048	0.57±0.088	0.83±0.092
Confidence	1.60±0.241	2.60±0.241	3.60±0.237

Table 5: Results of the user study in the non-expert group ($\pm$ 95% confidence interval).

	SWOX	CWOX-1sB	CWOX-2s
Accuracy	0.40±0.075	0.51±0.102	0.75±0.119
Confidence	1.40±0.108	2.80±0.172	3.40±0.163

5.2 INTERPRETABILITY TO USERS

How well does a WOX method help human users understand the evidence that a model relies on to discriminate between classes? To answer this question, we have conducted an user study following the *forward simulation* protocol [Ribeiro et al., 2016, Doshi-Velez and Kim, 2017, Nunes and Jannach, 2017, Lage et al., 2019, Tjoa and Guan, 2020]. As shown in Fig. 7, we display the heatmaps for pairs of confusing labels alongside the input image, and ask users to match the heatmaps with the labels. A correct matching would indicate that a user understands what pixels the model considers important for each of the two labels.

The study was conducted on the predictions by ResNet50 on a collection of images from the ImageNet validation set. For each image, a pair of confusing top classes was selected based on the latent tree T from the offline phase. CWOX-2s, CWOX-1sB and SWOX were included in the study. To make the study manageable, we did not consider all possible combinations of image classification models, WOX methods, and base explainers. We also limited the choices of input images and confusing class labels. See Appendix B for the details. CWOX-1sB was chosen over CWOX-1sA because it is based on previous work [Wang and Vasconcelos, 2020], and also because CWOX-1sA would simplify to CWOX-2s when there is only one confusion cluster for the inputs. RISE was used as the base explainer due to its proven superior contrastive faithfulness compared to other base explainers (as shown in Tab. 3).

The user study consisted of two groups of participants. The first group, referred to as the expert group, included postgraduate students enrolled in a machine learning course. These students had hands-on experience with deep computer vision models, such as training a CNN model. In contrast, the second group, known as the non-expert group, consisted of first-year undergraduate students who had no prior experience or knowledge in training deep learning models. Both groups had an equal number of participants, with 60 individuals in each group.

Each group was randomly divided into three subgroups, and

each subgroup was responsible for only *one* of the three WOX methods. As in previous XAI user studies [Doshi-Velez and Kim, 2017, Nguyen, 2018, Hase and Bansal, 2020, Fel et al., 2021], the participants went through a training phase before handling explanations for *new unseen examples*. Besides the matching task, they were also asked to rate their confidence in their answers on a scale of 1 to 5, with 1 meaning "not sure" and 5 meaning "completely sure".

The results are shown in Tab. 4 and 5. In both groups, the accuracy and confidence of CWOX-2s were significantly better than CWOX-1sB and SWOX. These results indicate that CWOX-2s is more effective in helping both novice and expert users understand the evidence used by the model to discriminate between classes. Interestingly, compared to the non-expert group, the expert users showed higher accuracy with narrower confidence intervals, especially among those responsible for CWOX-2s. This suggests that expert users can acquire more information about model behaviors from CWOX-2s explanations than non-expert users.

To get a concrete feeling about the superiority of CWOX-2s, imagine completing the matching tasks shown in Fig. 7. Among the two second-stage CWOX-2s heatmaps (those at the bottom), the one on the left highlights the red cross, while the one on the right does not. Hence, the former should be obviously matched with `ambulance` and the latter with `police van`. The matching task is relatively more challenging with heatmaps by SWOX and CWOX-1sB.

5.3 VISUAL EXAMPLES

In this section, we provide two more visual examples with Grad-CAM as the base explainer. More examples including examples with different base explainers are in Appendix C.

The first example (Fig. 8) illustrates the differences between SWOX and CWOX-2s when used to explain the output of an an image that contains both keyboard and mouse. As discussed in Section 3, `mouse` and `computer keyboard` are grouped together in the latent tree as they often co-occur in images. Consequently, CWOX-2s first identifies evidence for the composite object `mouse+computer-keyboard` (c), and then the evidence for `keyboard` against `mouse` (c.1) and the evidence for `mouse` against `keyboard` (c.2). The CWOX-2s explanations are clear and discriminative. However, the SWOX explanations for `mouse` and `computer keyboard` are very similar to each other, and hence are not discriminative.

The second example (Fig. 9) shows the difference between CWOX-1sA and CWOX-2s when explaining the output of an image that include the `electric guitar` and `acoustic guitar` among the top predicted classes. While the CWOX-2s explanations (b.1-2) provide discriminative information for the two visually similar class, CWOX-1sA explanations (d.1-2) exhibit noticeable overlap in highlighted regions, making it difficult to understand what features the model relies on to distinguish the two classes.

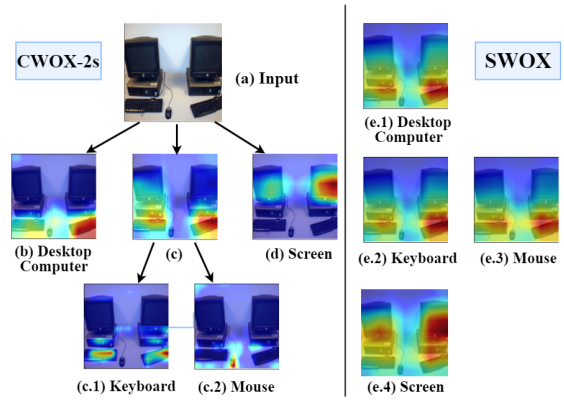


Figure 8: SWOX and CWOX-2s explanations of the output of ResNet50 on an image that contains keyboard and mouse.

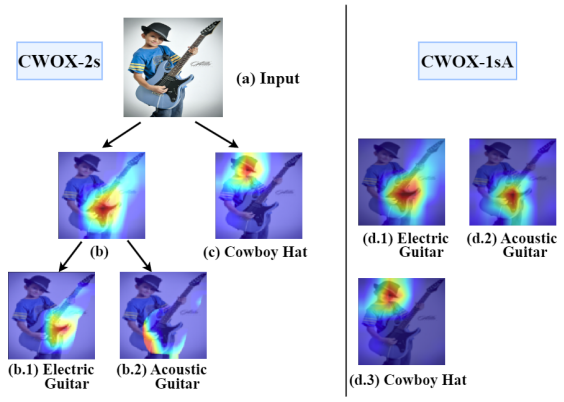


Figure 9: CWOX-2s and CWOX-1sA explanations of the output of ResNet50 on an image with `electric guitar` and `acoustic guitar` among the top classes.

6 CONCLUSION

We propose a novel post-hoc local explanation method called CWOX-2s for image classification. Unlike most previous methods, CWOX-2s explains all top classes in the output rather than one individual class. The key technical contribution is a principled method for determining how to contrast the top classes against each other. Recently, a new conceptual framework for XAI termed *evaluative AI* is proposed [Miller, 2023], which stresses the use of XAI to "provide evidence for and against decisions made by people, rather than provide recommendations to accept or reject". CWOX-2s aligns with this framework nicely. Empirical results show that, in comparison with alternative methods that explain all top classes, CWOX-2s produces explanations that are more faithful to the model and more interpretable to human users. Furthermore, we propose two metrics for evaluating contrastive explanations, namely Contrastive AUC (CAUC) and Weighted Drop in Contrastive Score (CDROP).

Acknowledgements

We thank the deep learning computing framework MindSpore (https://www.mindspore.cn) and its team for the support on this work. Research on this paper was supported in part by Hong Kong Research Grants Council under grant 16204920. Weiyan Xie was supported in part by the Huawei PhD Fellowship Scheme. We thank Prof. Janet Hsiao, Yueyuan Zheng, Luyu Qiu and Yunpeng Wang for valuable suggestions and discussions. We thank April Hua Liu for organizing the user study with non-experts.

References

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

Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International Conference on Learning Representations*, 2022.

Arjun Akula, Shuai Wang, and Song-Chun Zhu. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2594–2601, 2020.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Peixian Chen, Nevin L Zhang, Tengfei Liu, Leonard KM Poon, Zhourong Chen, and Farhan Khawar. Latent tree models for hierarchical topic detection. *Artificial Intelligence*, 250:105–124, 2017.

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

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pages 592–603, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

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

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*, 2021.

Been Kim and Finale Doshi-Velez. Machine learning techniques for accountability. *AI Magazine*, 42(1):47–52, 2021.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

Xiao-Hui Li, Caleb Chen Cao, Yuhan Shi, Wei Bai, Han Gao, Luyu Qiu, Cong Wang, Yuanyuan Gao, Shenjia Zhang, Xun Xue, et al. A survey of data-driven and knowledge-aware eXplainable AI. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. *GlobalSIP*, pages 1–5, 2019.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

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

Tim Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support. *arXiv preprint arXiv:2302.12389*, 2023.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, 2018.

Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3):393–444, 2017.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

Mohit Prabhushankar, Gukyeong Kwon, Dogancan Temel, and Ghassan AlRegib. Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3289–3293. IEEE, 2020.

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

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

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

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3145–3153, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *ICLR*, 2014.

J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32 (11):4793–4813, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990, 2020.

Yipei Wang and Xiaoqian Wang. "why not other classes?": Towards class-contrastive back-propagation explanations. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=X5eFS09r9hm`.

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

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018a.

Nevin L Zhang and Leonard KM Poon. Latent tree analysis. *AAAI*, pages 4891–4898, 2017.

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2018b.