Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

LETTER

User-Guided Global Explanations for Deep Image Recognition: A User Study

Mandana Hamidi-Haines | Zhongang Qi | Alan Fern | Fuxin Li | Prasad Tadepalli

School of Electrical Engineering and Computer Science, Oregon State University, Oregon, USA

Correspondence
Alan Fern, Oregon State University. Email: afern@oregonstate.edu

Summary

We study a user-guided approach for producing global explanations of deep networks for image recognition. The global explanations are produced with respect to a test data set and give the overall frequency of different “recognition reasons" across the data. Each reason corresponds to a small number of the most significant human-recognizable visual concepts used by the network. The key challenge is that the visual concepts cannot be predetermined and those concepts will often not correspond to existing vocabulary or have labelled data sets. We address this issue via an interactive-naming interface, which allows users to freely cluster significant image regions in the data into visually similar concepts. Our main contribution is a user study on two visual recognition tasks. The results show that the participants were able to produce a small number of visual concepts sufficient for explanation and that there was significant agreement among the concepts, and hence global explanations, produced by different participants.

KEYWORDS:
Explainable AI, Computer Vision, Human-Computer Interaction

1 | INTRODUCTION

Many explanation methods have been proposed for many types of deep networks. Among these, explanations can be divided into two broad types, local and global. Local explanations give insight into a network’s “reasons" for decisions on individual instances. In the simplest case, the reasons can simply be the parts of the input or image region responsible for the decision. More generally they can be abstracted into location and scale invariant image properties, e.g., presence of a hooked beak, which we call abstract local explanations. These explanations can be used by end-users or developers to adjust confidence in these decisions or to diagnose errors. In contrast, global explanations give insight into a network’s overall decision-making behavior. For example, a global explanation may provide summary statistics of abstract local explanations produced over a test data set. This can be useful for proactively assessing strengths and weaknesses of a network’s decision logic.

In this work, we consider the generation of global explanations for deep image recognition. Specifically, our global explanations are derived from local explanations that, for a given image, highlight the most significant image regions to a network’s decision. The global explanations are based on summarizing the local explanations over a test set by abstracting away the spatial information. In particular, the significant regions are abstracted by associating them with semantically-meaningful visual concepts, so that the abstracted local explanations are combinations of those concepts. The global explanation is then the frequency profile of the abstracted local explanations. This type of explanation can help to identify semantically-anomalous explanations or help assess the number of distinct decision types and their generality level.

There are two main challenges in generating the above explanations. First, the local explanations require computing the significant image regions for a decision. This problem has been addressed by computing activation maps (or heat maps) in an image for network nodes that are important to the decision. For example, activation maps have been derived via backward network analysis [1,2,3,4,5,6,4,7,8] or image perturbation techniques [9,10]. As described in Section 2, we use a novel combination of existing techniques to compute local explanations that ideally involve a small number of significant regions.

The second challenge, which is the main focus of this paper, is that global explanations require associating the significant image regions to human-meaningful concepts. One approach to this problem is to use labelled data to automatically map individual network nodes, and in turn their activation regions, to a set of predetermined concepts. While such approaches have shown some promise [11,12,13], current results suggest that often the activations of individual nodes do not consistently map to a single concept across different images. For example, 75% of a node's activations may correspond to a bird's beak, while the remaining 25% are distributed across other known and unknown concepts. In general, without significant research breakthroughs, node-level concept mappings should be expected to be noisy, which can lead to unsound and misleading global explanations.

Even if network nodes and/or their activations could be reliably mapped to predefined concepts, the explanations would be limited to just those concepts. Rather, we should expect that networks will uncover concepts outside of any predefined set, including visually-coherent concepts that don't correspond to existing vocabulary. For example, in our bird species recognition task, the significant image regions often correspond to concepts that combine parts, e.g. "beak and part of bird crown". We can also expect and hope that networks will uncover completely new concepts, which, once identified, may provide new domain insights to humans.

To address these challenges, we follow a human-guided approach for mapping significant image regions to meaningful visual concepts over test data. In particular, we provide an interface for a human to cluster the significant image regions in the data into meaningful groups called "visual concepts"—an activity we call "interactive naming." Drawing from the lessons of previous work [14], our interface provides maximum flexibility to human annotators by presenting them with significant image regions and allowing them to freely move the images around into clusters. Unlike previous work, however, which seeks to group images according to a predetermined label set, our approach allows annotators to create clusters that make the most sense to them and give them meaningful names. The flexibility of this approach is desirable since it avoids presupposing concepts. However, the flexibility also raises at least two practically important research questions.

RQ1 (Coverage of Interactive Naming): *What fraction of the significant image regions are covered by the human-defined visual concepts?* The annotators are not forced to cluster all significant image regions, since some regions may not be recognizable as coherent concepts. If coverage is low, then local explanations are not represented well by the visual concepts, resulting in lower-quality global explanations.

RQ2 (Inter-annotator Agreement): *How much do the visual-word clusters from different annotators overlap?* If different annotators produce semantically similar sets of visual concepts, then the resulting global explanations depend little on the annotator. Intuitively, this is highly desirable since the semantics of a global explanation should be primarily a function of the neural network and not the annotator.

Our main contribution is a user study that investigates these two research questions on two data sets: a bird species classification dataset [15], and a breast cancer classification dataset [16]. To the best of our knowledge this is the first time these practically important questions have been addressed in a user study. Our results reveal that for these datasets and trained networks, the annotators are able to cluster the vast majority of significant image regions into a small number of visual concepts. The results also show that there is significant agreement about the concepts between different annotators. Overall, the results suggest that interactive naming is a promising approach for generating global explanations and deserves further study in image recognition and other deep network applications.

2 | INTERACTIVE NAMING FOR GLOBAL EXPLANATIONS

Our overall motivation is to develop tools to help understand the decisions of deep neural networks (DNNs) that are trained for image recognition. In particular, we aim to generate meaningful global explanations of the decision-making behavior with respect to a representative set of test images. This can provide insight into the strengths and weaknesses of the learned DNN that may not be apparent by just observing test set accuracy. For example, one might hope to discover situations where the DNN is making the right decision, but for the wrong reason, which would identify potential future failure modes.

Figure 1 shows an overview of our *interactive naming* approach for producing test set explanations. At a high-level, each DNN decision for a test image is dominated by a set of the most significant activations of neurons in the penultimate layer. Thus,

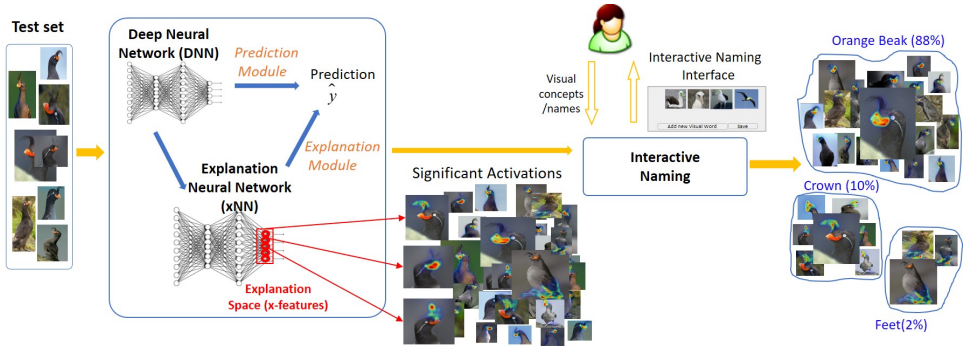

FIGURE 1 Given a test set the xNN is used to classify each image and identify the significant active x-features. The **local explanation** for the classification of an image is then a visualization of the salient image regions/activations of those features. Next interactive naming is used to abstract the local explanations by representing each significant activation by a meaningful visual concept. Specifically, the set of all significant activations are collected and a human uses our interface to cluster significant activations into semantically meaningful groups called visual concepts. These concepts are then used to produce **abstract local explanations**, each being the set of visual concepts associated with a local explanation. The **global explanation** (far right) is then the frequency profile of the usage of each abstract local explanation over the test data. For example, we see that on 88% of the test data, the visual concept “Orange Beak" is the sole contributing concept to the classification.

highlighting the image regions responsible for those activations is a useful type of local explanation. Further, as described in the introduction, abstracting the activations removing the spatial details and attaching meaningful concepts to them is useful for producing global explanations. However, typical DNNs use very large penultimate layers, which makes training easier, but can result in less compact explanations due to the large numbers of significant activations. For this reason, we attach an explanation neural network (xNN) to the penultimate layer of the DNN, which is trained to reproduce the decisions of the DNN, but dramatically reduces the number of activations. To further reduce the number of signficant activations we employ the notion of minimal sufficient explanation (MSX). MSX consists of a minimal number of units in the penultimate layer (here the xNN), whose sum of activations is sufficient to overcome any negative activation and classify the image.

In order to attach meaning to the significant xNN activations we developed an interactive-naming interface which displays visualizations of the significant activations (i.e. image regions) in a test set to a human annotator. The annotator is then able to cluster the activations into meaningful groups, called visual concepts, and attach linguistic labels to the groups if desired. Given one of the test images, the abstract local explanation for the decision is the group identities/names of the significant activations. A global explanation can then be formed by giving the statistics of how frequent different combinations of concepts appear as local explanations. This allows for a comprehensive understanding of the different qualitative decision types over the test set. The rest of this section explains the above steps in more detail.

2.1 | Explanation Neural Networks (xNNs)

An xNN[17] is an additional network module that can be attached to any intermediate layer of an original DNN, which typically has thousands of neurons. The xNN learns a lower dimensional embedding for the DNN layer, resulting in a vector of *X-features*, and then linearly maps the X-features to the output $\hat{y}$ in order to mimic the output y of the original DNN model. In our work, we apply xNNs to a convolutional DNN trained on the available multi-class data. The DNN outputs $p(c_i|I)$ for each given image I and category $c_i \in 1, \ldots, C$. The penultimate layer of the DNN can be considered as scoring functions for each category $s(c_i|I)$, where a softmax unit $p(c_i|I) = \frac{s(c_i|I)}{\sum_{i=1}^{C} s(c_i|I)}$ serves as the final layer of the DNN that computes the class-conditional probability from the scores. xNN is trained starting from the first fully-connected layer in the DNN for each class, aiming at being faithful to the scoring functions $s(c_i|I)$ for each category. The xNNs can then be used for multi-class prediction by computing the scores produced by each xNN and returning the highest scoring class.

It is desirable for X-features to have the following 3 properties: 1) *faithfulness*, the DNN predictions can be faithfully approximated from a simple linear transform of the X-features; 2) *sparsity*, a relatively small number of X-features are active per image, and 3) *orthogonality*, the X-features are as independent from each other as possible. Details of the optimization technique to achieve these properties are beyond the scope of this paper.

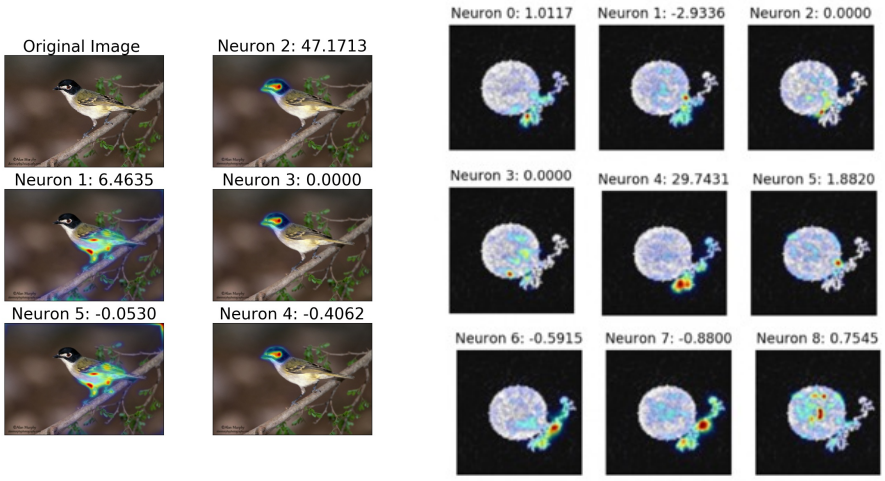

FIGURE 2 Examples of visualization of x-feature activations of bird species and breast cancer datasets.

2.2 | Explanations via Interactive Naming

Given a test image and a class c, we can use the xNN for c to produce a class score. This score is a linear combination $\sum_i w_i \cdot x_i$ of the X-features x_i and their associated weights. The positive terms (i.e. X-features with positive weights) in the linear combination sum to provide a positive score that can be viewed as providing positive evidence for c. Typically only a subset of the positive terms are significant. Thus, we define the *significant X-features* for the image to be *Minimal Sufficient Explanation (MSX)*. The MSX keeps the minimal subset of positive weights whose sum of activations exceeds the sum of all negative activation. The concept of minimal sufficient explanation was introduced in closely related work on explanations for optimal Markov Decision Process policies [18] and more recently in model-free reinforcement learning [19]. The significant X-features can be viewed as a type of local explanation of why the image might be assigned to class c. However, they do not have associated semantics, so the explanation is not very useful for human consumption.

To produce human-consumable local explanations, we produce an *activation map* for each significant X-feature in an image, which identifies the "salient" image region that is responsible for the X-feature activation. In this work, we applied the *(Integrated-Gradients Optimized Saliency (I-GOS)* algorithm [20] for computing activation maps. I-GOS is a new visualization method that optimizes an image mask, or heatmap, so that the classification scores on the masked image would maximally decrease. It computes descent directions based on the integrated gradients instead of the normal gradient, which avoids local optima and speeds up convergence. Compared with previous approaches, such as ExcitationBP [7], this method can flexibly compute heatmaps at any resolution. Figure 2 shows examples of 9 X-feature activations of breast cancer cell images, which are superimposed on the original image.

We call the maps of significant X-features the *significant activation maps* or simply the *significant activations*. We consider the set of significant activations for an image to be the local explanations, which can be easily viewed by a human. While local explanations give insight into a specific prediction, they do not provide a general understanding of the core semantic concepts and combinations of those concepts used for predictions across an entire test set, i.e. a global explanation, which is our goal.

The goal of interactive naming is to cluster the significant activations in the test set, where each group is intended to represent a semantically meaningful *visual concept* to the annotator. Activations that are assigned to a visual concept are considered to be *named*, while other activations are considered to be *unnamed*. The complete set of named activations resulting from interactive naming is called a *naming* of the test set. Given a test-set naming, we can now generate an *abstract local explanation* for each test image as the set of visual concepts for the significant activations. If a significant activation is unnamed, then the explanation includes "other" for the name of the activation. A global explanation can then be displayed, which depicts the frequency of the different abstract local explanations in the test data. Note that the global explanations ignore certain details of local explanations, such as the absolute and relative positions of activations. However, by ignoring this information, we find that the global explanations in our experiments are quite succinct and still yield significant insight into the DNN decision making.

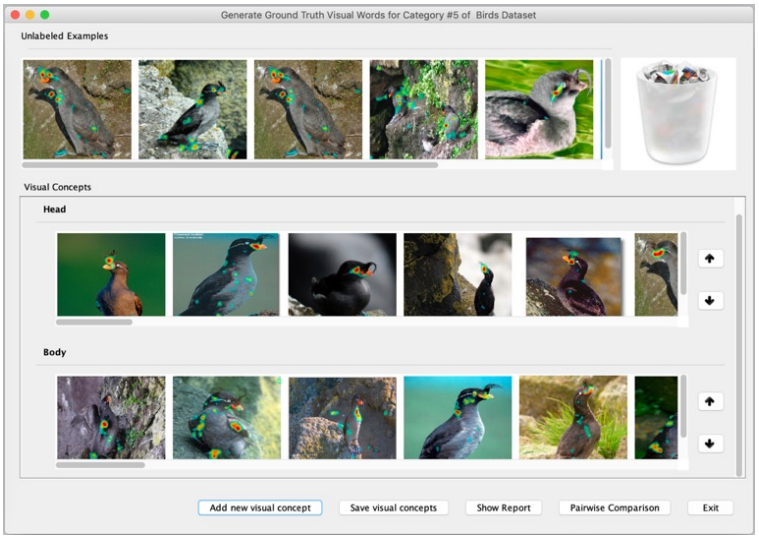

FIGURE 3 Annotation Interface: Our approach allows annotators to explore feature activations and group them into meaningful textual / visual concepts. The top row displays some number of currently unlabeled activation images. The lower rows correspond to user defined concepts for images they have assigned to the concept. The user can freely create new concepts (new rows) at any time or delete concepts. Unlabeled activation images can be moved to any of the concept rows, which indicates the assignment of that activation to the concept. Image that are determined to be noise can be put in the trashcan.

TABLE 1 The relevant X-feature activations of 12 bird categories: *(a) Laysan Albatross, (b)Crested Auklet, (c) Brewer Blackbird, (d) Red-winged Blackbird, (e) Northern Fulmar, (f) Green Jay, (g) Mallard, (h) Black Tern, (i) Common Tern, (j) Elegant Tern, (k) Green-tailed Towhee*, and *(l) Black-capped Vireo*. The table shows the number of images for which each feature makes a significant positive contribution.

Index of category	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
Number of images	60	44	59	60	60	57	60	60	60	60	60	51
True Positive images	52	41	40	57	53	54	53	45	39	45	55	49
False Positive images	10	3	23	5	15	9	0	10	18	11	9	7

2.3 | Interactive Naming Interface

One of the key aspects of interactive naming is that the set of visual concepts is not known beforehand and varies from person to person. Moreover, the visual concepts in an image are not immediately apparent until the annotator sees multiple images. In previous work it was shown that human labelers are more efficient and more consistent when they are presented with multiple instances at once and are allowed to choose the ones they want to label[14,21].

Following this previous work, we designed a flexible user interface (Figures 3) to group the significant activations into different visual concepts and give them textual labels/names. The set of X-feature activations is shown to the annotator in the "Unlabeled Examples" section of the interface. The annotator can cluster activations into visual concepts and give them names. The interface allows the annotators to compare all instances, create new visual concepts for which they are confident and leave the rest as unlabeled. The interface also allows for moving images across clusters, and merging clusters.

3 | INTERACTIVE-NAMING USER STUDY

Data Sets and Procedure. All our experiments were conducted on 12 bird categories of Caltech-UCSD Birds-200-2011 dataset[15], and 5 categories of microscopic images of breast cancer cells called CEL dataset[16]. Given a convolutional DNN trained on the available training data, we train a separate xNN for each category connected to the penultimate DNN layer. For the Birds and CEL data sets, we used 5 and 9 X-features respectively, which significantly reduces the dimensionality for the 4,096 feature penultimate layer.

TABLE 2 The relevant X-feature activations of 5 breast cancer cells categories. The table shows the number of images for which each feature makes a significant positive contribution.

Category	*Actinedge*	*Filopodia*	*Hemispherebleb*	*Lamellipodia*	*Smalbleb*
Number of images	275	268	231	324	250
True positive images	214	254	228	264	247
False positive images	4	4	62	34	37

Tables 1 and 2 show the number of images in each category of the Birds and CEL datasets along with the number of true and false positives of the DNN for each category. Somewhat surprisingly, for all images in both data sets, there was a single X-feature in the MSX and hence a single significant activation. This indicates that the xNN is highly effective at concisely capturing the "reason" for the original DNN decisions. Thus, each local explanation was a single activation map and each image produces a single significant activation for the interactive-naming process.

We conducted our user studies with 10 and 5 human subjects on the Birds and CEL data sets, respectively. We focus on producing global explanations for each category, which give the reasons that each category is predicted by the DNN. Thus, each user conducted the interactive-naming process for each category. In particular, for each category, the users were asked to use the interactive-naming interface to cluster and name the significant activations of images the DNN predicted as positive for that cateogory (the false positives and true positives). The annotators were instructed to only introduce visual concepts that contained at least three significant activations. Thus, if a signficant activation was visually coherent, but not like at least two other activations it received the default label "other". Otherwise, the participants were free to cluster and label as many activations as it made sense to them. However, not all subjects followed these instructions and included some clusters with less than 3 images. In the following analysis, we removed all such clusters.

RQ1: Coverage of Interactive Naming. Since the annotators are not forced to assign visual concepts to, or name all significant activations, some of the activations in the data are unnamed and treated as noise/outliers. Here we are interested in how well the annotations cover the activations and explanations and how this coverage varies across annotators.

Figure 4 (top) and Figure 4 (bottom) show the mean of the fraction of significant activations that are named by annotators for each category of Birds and CEL dataset, respectively. In addition, the bar labeled "Any Annotator", shows the fraction of significant activations that were assigned to a visual concept by at least one annotator. We see that within a particular class, there is relatively small variation among users and that the "Any Annotator" bar is not much higher than that of the typical individual annotator. This indicates that there is some consistency in the set of activations that users consider to be noise. For most categories there is a relatively small amount of activations not labeled by users. However, in *Hemispherebleb* category of CEL dataset in Figure 4 (bottom), in average 40% of the activations were not labeled by the users. The activations of this category are more difficult to annotate in comparison to the others.

We performed a qualitative analysis to understand some of the reasons that annotators were not able to assign names to activations. One of the major reasons was when activations were difficult to interpret and appeared to be noise. For example, this happens when activations highlight the edge of the image or fall on the background. Such activations are potential warning indicators about a classifier. Thus, uncovering these examples through interactive naming has value. In other cases, the activation map was interpretable to the annotator, but there were not enough similar activation maps to form a cluster. This case may be resolved by using a larger test set.

RQ2: Inter-annotator Agreement. We first characterize the fraction of images annotated by pairs of annotators. For this we use the Jaccard index, which is the ratio of the intersection to the union of the two sets of signification activations labeled by two annotators, to measure the fraction of the images both annotators annotated. This is shown in the last column of Table 3 averaged over different pairs of annotators. The Jaccard index is fairly high for all categories, indicating that there is a good overlap between the sets of activations chosen by different annotators to annotate.

Next we consider the extent to which the namings of different annotators can be translated to one another. Are there one-to-one correspondences, subsumption relationships, or cases of purely incompatible concepts? Understanding this issue is important for understanding the extent to which explanations are fundamentally annotator-specific.

Given two namings N_i and N_j of annotators i and j, we are interested in matching the clusters between the namings. For this purpose, we applied a cluster matching framework, called D-family matching[22]. It first defines the "intersection graph" G of N_i and N_j as a bipartite graph where the vertices in the two partite sets correspond to the clusters of N_i and N_j. Each pair

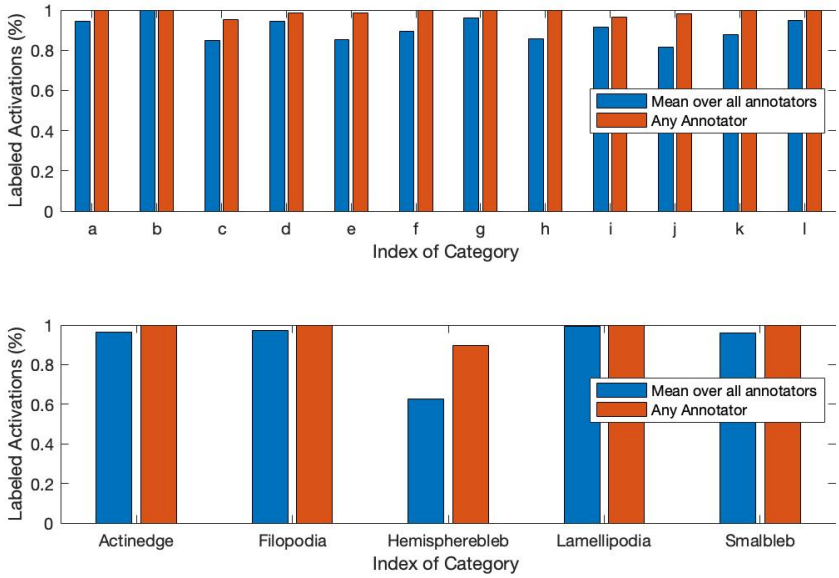

FIGURE 4 Fraction of labeled significant activations for each category of Birds dataset (top) and CEL dataset (bottom).

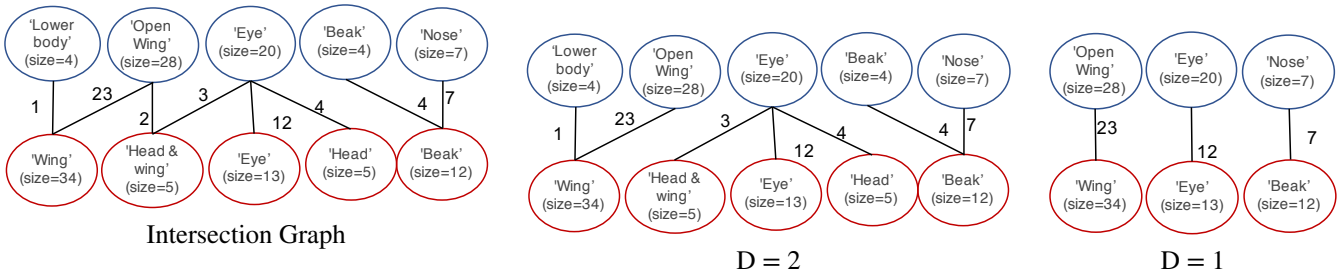

FIGURE 5 An example of pairwise similarity matching between two annotators. Blue circles represent the visual concepts created by Annotator i and Red circles are the visual concepts created by Annotator j

of clusters of N_i and N_j has an edge with the weight equal to the size of their intersection. D-family matching is a partition of all nodes of the bipartite graph into some number of disjoint sets $S_1, S_2, ...$ such that the diameter of all subgraphs of G over the nodes in S_i is $\leq D$. The best D-family matching maximizes the sum of the weights of all edges in all the subgraphs. Figure 5 shows an intersection graph for a pair of namings and its best D-family matchings for $D = 1$ and 2

We compute the agreement between the two annotators for $D = 1$ and 2 as the total weight of all edges in the D-family matching as a fraction of the number of activations labeled by both annotators. If we interpret the matchings as translations between namings, then the agreement is the fraction of activations that are translatable between namings. The columns labeled "Agreement" in Table 3 shows the statistics of 1-family and 2-family agreements for each category over the set of all annotator pairs for Birds and CEL datasets, respectively. The agreement numbers are fairly high across most categories except for a few categories in the birds dataset (b and k) for $D = 1$, where it ranges from 71-73 %. Since 2-family matching is more permissive than 1-family matching, the agreement numbers are higher for $D = 2$. Overall the high agreements show that there is reason to be optimistic about developing a common ontology for explanations.

Examples of Global Explanations. Recall that the final goal of the interactive naming task is to produce global explanations over a test set, which depend both on the category of interest and the annotator. For example, one global explanation of category g of the birds dataset consists of 29% of activations clustered as 'eye', 55% as 'green feathers' and 11% as 'yellow beak' while the remaining 5% unclustered. For category d a global explanation consists of 82% of cases clustered as 'red spots on wings', another 11% clustered as 'eye and red spots on wings' and the rest unclustered. Thus the global explanation of category g indicates that green feathers are an important indicator and for category b red spots on wings as important. Such global explanations might either confirm or contradict a practitioner's prior knowledge and help modulate their trust or provide new insight.

TABLE 3 Pairwise comparison between clusters generated by annotators over all categories.

Pairwise similarity scores for Bird Dataset			
Category	Agreement (D=1)	Agreement (D=2)	Jaccard index
a	0.83±0.12	0.95±0.03	0.93±0.04
b	0.73±0.17	0.95±0.1	0.96±0.04
c	0.95±0.04	0.98±0.02	0.89±0.05
d	0.96±0.02	0.98±0.02	0.95±0.03
e	0.88±0.07	0.94±0.04	0.86±0.06
f	0.9±0.05	0.97±0.03	0.86±0.09
g	0.8±0.16	0.94±0.03	0.92±0.05
h	0.9±0.09	0.97±0.03	0.77±0.16
i	0.79±0.13	0.91±0.05	0.92±0.05
j	0.8±0.13	0.99±0.02	0.86±0.07
k	0.71±0.24	0.92±0.06	0.87±0.07
l	0.99±0.01	1±0.01	0.93±0.04
Global Average	0.85	0.957	0.8924

Pairwise similarity scores for CEL Dataset			
Category	Agreement (D=1)	Agreement (D=2)	Jaccard index
Actinedge	0.91±0.03	0.96±0.02	0.94±0.02
Filopodia	0.63±0.1	0.75±0.07	0.94±0.03
Hemispherebleb	0.89±0.08	0.99±0.04	0.58±0.07
Lamellipodia	0.98±0.01	0.99±0.01	0.99±0.01
Smalbleb	0.96±0.02	0.98±0.01	0.91±0.08
Global Average	0.874	0.934	0.872

3.1 | Correspondence Between X-features Activations and Visual Concepts

Ideally, we would like X-features to correspond to distinct semantic concepts (i.e. visual words), since that would allow for fixed vocabulary terms to be associated with them. However, as discussed in Section 1, work toward such "disentanglement" has not achieved this ideal condition. While not the main objective of this paper, we now exploit the human-annotations of visual words to observe how close the X-features come to achieving the ideal.

We adopt the metric of *purity* to measure the correspondence between visual concepts and X-features[23]. The purity of a clustering is defined as the number of examples that belong to the plurality class of each cluster as a fraction of the total number of examples. Here, we employ purity to measure both the degree to which each visual concept maps to a single X-feature, i.e., visual concept → X-feature purity or *CX-purity*, and the degree to which each X-feature maps to a single visual concept, i.e., X-feature → concept or *XC-purity*.

We call the set of clusters of activations created by each annotator, a naming. To compute CX-purity of a naming, we first assign each visual concept to the X-feature to which a plurality of its significant activations belong. We call it *the majority X-feature* of that concept. CX-purity is the number of activations in the naming that belong to the majority X-feature of their concept as a fraction of all named activations. Similarly we define the majority concept of an X-feature to be the concept that is assigned to a plurality of activations of that X-feature. XC-purity is the number of activations in the naming that belong to the majority concept assigned to their X-feature as a fraction of all named activations.

As it may be clear, both CX-purity and XC-purity vary from category to category. Table 4 and Table **??** show the CX-purity and XC-purity values averaged over all annotators for each category of Birds and CEL dataset, respectively. The first observation is that the standard deviations of all these numbers are low, suggesting inter-annotator consistency. The CX-purity numbers in Table 4 are generally high ranging from 0.82 to 1.0 with an average of 0.92, with the exception of Category *b* where it is 0.61. The XC-purity numbers are generally worse than CX-purity, going down to 0.47 in one case (Category *k*) and a maximum of 0.96, with an average of 0.74. This suggests that the mapping from X-features to visual concepts is more one-to-many than the other way around in Birds dataset. There are two reasons for this. First the Birds dataset only has 5 X-features, which forces each X-feature to represent multiple concepts. Second, in most categories of Birds dataset, the significant activations are mostly covered by one or two X-features, which further reduces the number of available significant X-features to cover the visual concepts. Except for the difficult category of *Filopodia*, the CX-purity numbers vary from 0.82 to 0.98 and the XC-purity varies from 0.89 to 0.99 for the remaining categories of CEL dataset with averages of 0.82 and 0.86 respectively (Table **??**).

4 | SUMMARY AND FUTURE WORK

In this paper, we studied the use of human guidance for the purpose of grounding global explanations of DNNs in meaningful visual concepts. Our interactive-naming approach involves augmenting the original DNN with a sparser xNN, visualizing the significant activation maps for each decision of the xNN on a test set, and then allowing annotators to flexibly group the activations into recognizable visual concepts, while attaching names to the concepts if desired. The visual concepts can then be used as the basis for abstracting local explanations and generating corresponding global explanations relative to a test set. We reported on our experience of having annotators use our interface for DNNs trained to recognize different bird species and cell

TABLE 4 CX-purity and XC-purity results for both dataset.

Bird Dataset		
Category	CX-purity	XC-purity
a	0.84 ± 0.03	0.69 ± 0.08
b	0.61 ± 0.05	0.76 ± 0.17
c	0.89 ± 0.03	0.85 ± 0.05
d	1 ± 0	0.93 ± 0.03
e	0.95 ± 0.01	0.78 ± 0.08
f	0.82 ± 0.01	0.91 ± 0.05
g	1 ± 0	0.77 ± 0.13
h	0.93 ± 0.02	0.58 ± 0.07
i	1 ± 0	0.51 ± 0.1
j	1 ± 0	0.68 ± 0.16
k	1 ± 0	0.47 ± 0.22
l	0.99 ± 0.01	0.96 ± 0.04
Mean	0.92	0.74

CEL Dataset		
Category	CX-purity	XC-purity
Actinedge	0.98 ± 0.01	0.92 ± 0.04
Filopodia	0.56 ± 0.01	0.62 ± 0.08
Hemispherebleb	0.87 ± 0.08	0.89 ± 0.14
Lamellipodia	0.86 ± 0	0.99 ± 0.01
Smalbleb	0.83 ± 0.05	0.96 ± 0.02
Mean	0.82	0.86

types. Our results showed that 1) in all cases, the xNN is able to produce single significant activations that are sufficient for classification, 2) the annotators are able to assign names to a very high fraction of significant activations, and 3) there is very good agreement between the namings produced by different annotators.

In the current paper, we are not focused on optimizing the efficiency of labeling. There is significant room for future work on interfaces that partially automate the labeling, e.g. by inferring which unlabeled activations are likely to belong to the emerging concepts. Our work suggests that there is significant overlap between the namings of different annotators so that interfaces that can actively transfer namings across different annotators might prove useful.

ACKNOWLEDGMENTS

This work was supported under the DARPA XAI program under contract #N66001-17-2-4030. Any opinions, findings, conclusions, or recommendations expressed are the authors' and do not necessarily reflect the views of DARPA, the Army Research Office, or the US government.

References

1. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T., eds. *Computer Vision – ECCV 2014*Springer International Publishing; 2014; Cham: 818–833.

2. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ICLR Workshop* 2014.

3. Springenberg J, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity: The All Convolutional Net. In: ; 2015.

4. Shrikumar A, Greenside P, Shcherbina A, Kundaje A. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *CoRR* 2016; abs/1605.01713.

5. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Precup D, Teh YW., eds. *Proceedings of the 34th International Conference on Machine Learning*PMLR; 2017: 3319–3328.

6. Bach S, Binder A, Montavon G, Klauschen F, Müller K, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 2015; 10. doi: 10.1371/journal.pone.0130140

7. Zhang J, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. In: Springer. ; 2016: 543–559.

8. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: ; 2017: 618-626.

9. Dabkowski P, Gal Y. Real Time Image Saliency for Black Box Classifiers. In: ; 2017.

10. Fong RC, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In: ; 2017: 3449-3457.

11. Bau D, Zhou B, Khosla A, Oliva A, Torralba A. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In: ; 2017.

12. Kim B, Wattenberg M, Gilmer J, et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: ; 2018: 2673–2682.

13. Zhou B, Sun Y, Bau D, Torralba A. Interpretable Basis Decomposition for Visual Explanation. In: ; 2018.

14. Kulesza T, Amershi S, Caruana R, Fisher D, Charles D. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In: ACM; 2014.

15. Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology; 2011.

16. Eddy CZ, Wang X, Li F, Sun B. The morphodynamics of 3D migrating cancer cells. *arXiv:1807.10822* 2018.

17. Qi Z, Khorram S, Fuxin L. Embedding Deep Networks into Visual Explanations. *Artificial Intelligence* 2021; 292.

18. Khan OZ, Poupart P, Black JP. Minimal Sufficient Explanations for Factored Markov Decision Processes. In: ; 2009.

19. Juozapaitis Z, Koul A, Fern A, Erwig M, Doshi-Velez F. Explainable reinforcement learning via reward decomposition. In: ; 2019: 47–53.

20. Qi Z, Khorram S, Li F. Visualizing Deep Networks by Optimizing with Integrated Gradients. *CoRR* 2019; abs/1905.00954.

21. Sarkar A, Morrison C, Dorn JF, et al. Setwise Comparison: Consistent, Scalable, Continuum Labels for Computer Vision. In: CHI '16. ACM; 2016; New York, NY, USA: 261–271

22. Cazals F, Mazauric D, Tetley R, Watrigant R. Comparing two clusterings using matchings between clusters of clusters. Research Report RR-9063, INRIA Sophia Antipolis - Méditerranée ; Universite Cote d'Azur; 2017.

23. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press . 2008.

How to cite this article: Mandana Hamidi-Haines, Zhongang Qi, Alan Fern, Fuxin Li, and Prasad Tadepalli (2021), User-Guided Global Explanations for Deep Image Recognition:
A User Study, *AI Letters*, *2021;999:999–999*.