Interactive Naming for Explaining Deep Neural Networks: A Formative Study

Mandana Hamidi-Haines, Zhongang Qi, Alan Fern, Fuxin Li, Prasad Tadepalli
School of Electrical Engineering and Computer Science, Oregon State University
Corvallis, Oregon, USA
{hamidim,qiz,Alan.Fern,Fuxin.Li,Prasad.Tadepalli}@oregonstate.edu

ABSTRACT

We consider the problem of explaining the decisions of deep neural networks for image recognition in terms of human-recognizable visual concepts. In particular, given a test set of images, we aim to explain each classification in terms of a small number of image regions, or activation maps, which have been associated with semantic concepts by a human annotator. The main contribution of this paper is a systematic study of the visual concepts produced by five human annotators using an interactive naming interface in terms of the adequacy of the concepts for explaining the test images and the inter-annotator agreement of visual concepts. Our work is an exploratory study of the interplay between machine learning and human recognition mediated by visualizations of the results of learning.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Neural networks.*

KEYWORDS

Explainable AI; Explanation Networks; Deep Neural Network; Interactive Machine Learning.

ACM Reference Format:
Mandana Hamidi-Haines, Zhongang Qi, Alan Fern, Fuxin Li, Prasad Tadepalli. 2019. Interactive Naming for Explaining Deep Neural Networks: A Formative Study. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019* , 7 pages.

1 INTRODUCTION

Deep neural networks (DNNs) are powerful learning models that achieve excellent performance on many problems ranging from object recognition to machine translation. However, the potential utility of DNNs is limited by the lack of human interpretability of their decisions, which can lead to a lack of trust. The goal of this paper is to study an approach, called *interactive naming*, for improving our understanding of the decision-making process of DNNs. In particular, this approach allows a human annotator to visualize and organize activation maps of critical neurons into meaningful visual concepts, which can then be used to explain decisions made over the test data.

Interpreting the role of neurons in the decisions of DNNs has been a long-standing problem in artificial intelligence[6]. Much recent work on interpretibility are based on the following methods: *1) heatmap-based methods*, which focus on visualizing activation maps that highlight parts of the input that are most important to the final decision of the DNN or the output of an individual neuron [1, 11, 12, 12–15, 17, 18]. *2) perturbation-based methods*, which perturb parts of the input to see which ones are most important to preserve the final decision [4, 5]. *3) concept-based methods*, which analyzes the alignment between individual hidden neurons and a set of semantic concepts [2, 7, 19]. While this provides additional insight into the semantics of neurons, they requires large sets of data labeled by the semantic concepts and is limited to the semantic concepts in that data. Importantly, none of the current approaches support human interaction in recognizing, clustering, and naming the concepts implicitly employed by the neural network in making its decisions. While some methods do employ human recognizable concepts, they are learned by the system offline from a large amount of labeled data that may or may not be relevant to the task at hand.

In this work, we make progress toward this goal by building an interface for interactive naming, and conducting a formative study on a set of non-trivial image classification tasks. In particular, our approach is based on the idea that the final decision of a DNN is dominated by the most highly-weighted neuron activations (the *significant activations*) in the penultimate network layer. Explanations of the decisions can thus be formed by 1) identifying the significant activations for each decision, and 2) attaching meaningful concepts to the significant activations. Since DNNs typically have thousands of units in the penultimate layer, (1) can result in an overwhelming number of activations. To address this issue we draw on recent work that augments the original DNN with a learned *explanation Neural Network* (*xNN*), which mimics the predictions of the DNN using a much smaller penultimate layer of *X-features*. Since the xNN is effectively equivalent to the original DNN, we can use it to make predictions on test instances with no loss in accuracy, but with a dramatic reduction in the number of significant activations to be considered for explanations.

To deal with (2), our interface displays the (significant) activation maps of X-features for decisions made on a test set and allows an annotator to cluster the activations into meaningful groups called "visual concepts." Even though there are a small number of significant activations that sufficiently explain the final decisions, there may not be a one-to-one correspondence between them and human-recognizable visual concepts. Indeed, unlike in the standard supervised learning setting, where the number of classes/concepts is typically fixed beforehand, the number of visual concepts covered by the set of all significant activations is unknown. To make matters more interesting, the set of visual concepts might be different for different annotators. Finally, the annotators may not be able to label a map in isolation, and might need to see multiple images and find

IUI Workshops'19, March 20, 2019, Los Angeles, USA

similarities and differences before labeling them. Indeed, this last problem has been studied under the name of "structured labeling," in the context of active learning and provides an inspiration for our work [8]. Drawing from the lessons of previous work, our interface provides maximum flexibility to the human annotators by presenting them with the activation maps of all X-features of all test images that belong to each category. Unlike the previous work on supervised and active learning which seek labels from a fixed label set, the annotators are asked to cluster the maps in a way that makes most sense to them and give them meaningful names.

The result of interactive naming is a set of explanations of test set predictions in terms of visual concepts. This enables summarizing the types of predictions that are made to gain confidence in the predictor and/or identify potential flaws in the predictor. Importantly, this type of summary is dependent on the human annotator, which raises interesting questions about differences in explanations that might result from different annotators. Specifically, we seek answers to the following research questions (RQs): through our study:

RQ1 (Coverage of Interactive Naming): What fraction of the examples are explainable using human recognizable visual concepts? If a significant fraction of the examples are not explainable via visual concepts, it might mean that the X-features are not properly aligned with human concepts and will have to be retrained from human data.

RQ2 (Inter-annotator Agreement): How much overlap exists between the annotated sets of activations between different subjects? How much do the clusters of different subjects overlap? Existence of significant overlaps might suggest that we can move toward building a standardized ontology of visual concepts for explanations. Lack of significant agreement might mean that we will have to personalize explanations to different annotators.

We explore the above questions through empirical experiments and annotator studies based on data from 5 annotators on a bird species classification dataset [16]. The studies reveal that a significant fraction of the images are human recognizable with some individual differences among different annotators.

2 INTERACTIVE NAMING FOR TEST SET EXPLANATIONS

We first give an overview of the overall approach and then describe each component of the system.

2.1 Overview

Our overall goal is to develop tools to help understand the decisions of DNNs that are trained for image recognition via supervised learning. In particular, we aim to generate meaningful explanations for decisions made over 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, attaching meaningful concepts to those activations is one way to explain decisions. 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 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. Thus, explanations can be formed in terms of much smaller number of activations.

In order to attach meaning to the significant xNN activations we developed an interactive naming interface which displays visualizations of the significant activations 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 a test instance, we can then form an explanation by producing the significant xNN activations and displaying the group identities/names of those activations. Qualitatively different decisions will tend to have different explanations. A key functionality of the system is to allow for the investigation into the different qualitative decision types over the test set. The rest of this section explains the above steps in more detail.

2.2 Explanation Neural Networks (xNNs)

An xNN [9] is an additional network module that can be attached to any intermediate layer of an original DNN, which typically have 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.

2.3 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 minimum subset of X-features that account for at least 90%

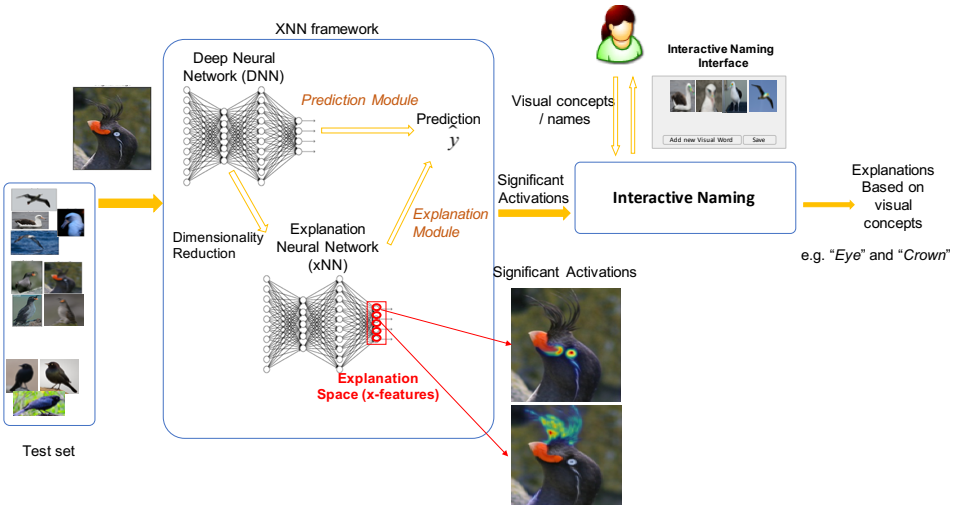

Figure 1: Interactive Naming Framework.

of the positive score. The significant X-features can be viewed as a type of 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 assign semantics to explanations, we can first produce an *activation map* for each significant X-feature in an image for the class under consideration, which identifies the "salient" image region that is responsible for the X-feature activation. In this work, we use the ExcitationBP algorithm for computing activation maps [18]. We call these maps the *significant activation maps* or simply the *significant activations*. While one can gain insight into a prediction by simply viewing the significant activations, it is difficult to obtain a general understanding of the core semantic concepts and combinations of those concepts used for predictions across an entire test set, which is our goal. Figure 2 (left) shows an example of a bird's original image followed by its 5 X-feature activations, which are superimposed on the original image.

Our interface is designed to attach semantics to all the significant activations across a test set. In particular, the interface allows an annotator to cluster the significant activations, 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 the *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 naming of a test set, we can now generate an *explanation* for each test image by generating the significant activations of the image and outputting the visual concept names for those activations. Thus, an explanation is just a set of names.

2.4 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 [8], it was shown that human labelers are more efficient when they are presented with multiple instances at once and are allowed to choose the ones they want to label. In another study [10], it was demonstrated that not only labeling multiple images is more efficient, but also elicits more consistent labels.

Following the previous work, we designed a flexible user interface(Figure 2 (Right)) 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 freely cluster activations into visual concepts and give them names. The interface allows the annotators to compare all instances, and create new visual concepts when they are confident. If the annotator is not comfortable with grouping or labeling some activations, they can leave them in the unlabeled section. The subjects can move images across clusters, and merge clusters. They are also allowed to discard the activations that they consider noisy.

2.5 Data Preparation

All our experiments were conducted on 12 categories of Caltech-UCSD Birds-200-2011 dataset [16]. The first row of Table 1 shows the number of images in each category. Given a convolutional DNN trained on the available multi-class data, we train xNN starting from the first fully-connected layer in the DNN for each category. This approach reduces the dimensionality from 4,096 features in the DNN to 5 X-features in the xNN without significant loss of accuracy. The fifth and sixth rows of Table 1 shows the multi-class classification accuracy of the xNN on the original 200 categories after replacing the DNN score with the one generated by xNN for each respective category, as well as the original DNN accuracy on those categories. It can be seen that xNN has almost identical accuracies as the DNN, even performing better than DNN in one case. Also, the last row shows the Root Mean Squared Error (RMSE) of xNN, which approximates exact scores of DNN well. xNN has an RMSE between $0.2 - 0.5$ while the range of the scoring function is usually $0 - 50$.

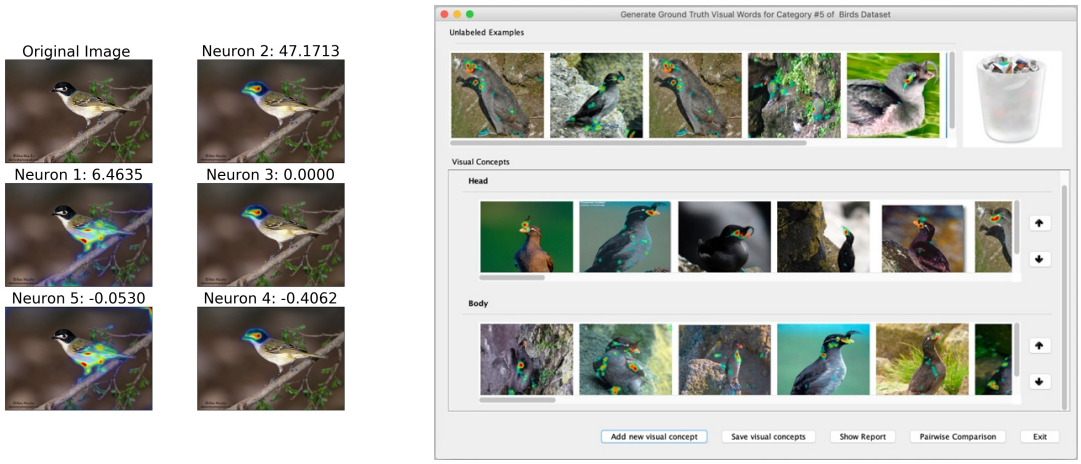

Figure 2: (Left) Examples of visualization of x-feature activations. (Right) Annotation Interface: Our approach allows annotators to explore feature activations and group them into different meaningful textual / visual concepts.

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
Total significant activations	118	108	158	120	73	167	138	125	170	124	120	81
Average of significant activations	1.97	2.45	2.68	2	1.22	2.93	2.3	2.08	2.83	2.07	2	1.49
DNN accuracy (%)	86.7	93.18	67.8	95.0	88.33	94.74	88.33	75.0	65.0	75.0	91.67	96.08
xNN accuracy (%)	86.7	93.18	67.8	95.0	88.33	96.49	88.33	75.0	65.0	75.0	91.67	96.08
xNN RMSE	0.23	0.23	0.51	0.41	0.35	0.30	0.50	0.20	0.21	0.34	0.45	0.33

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.*

Since the X-features with negative weights do not provide positive evidence to the class at hand, their activation maps are not used for annotation. We further filter the activation maps to only those maps that contribute to 90% of the total positive weight for the final decision. We call these *significant activations*. The second row of Table 1 shows the total number of significant activations in each category. The third shows the average number of significant activations per image.

3 HUMAN SUBJECT STUDY

We had the activation maps of the different images annotated by 5 different subjects using the annotation interface. The activation maps were separated by the class, but not by the X-feature. The annotators were instructed to not introduce visual concepts that only applied to one or two images, but were otherwise free to cluster and label as many images as it made sense to them. However, not all subjects followed instructions and left some clusters with less than 3 images. In the following analysis, we first cleaned the data by removing a small number of clusters with less than 3 images.

3.1 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 3 shows the fraction of significant activations that are named by each annotator for each bird category. In addition, the last bar for each category, 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. Also for most categories there is a relatively significant amount of activations not labeled by users, approximately ranging from 20% to 40%.

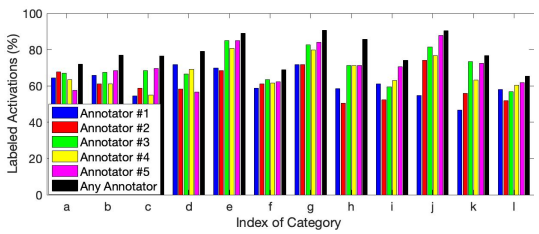

Figure 3: Fraction of labeled significant activations for each category across annotators.

We now consider how well the annotations cover explanations, which gives a better sense of how useful they will be for analyzing explanations. In particular, we consider an explanation for an image to be *completely (partially) covered* by an annotation if all (at least one) of the significant activations for that image are named. Figures 4 and 5 show the partial and complete coverage for each annotator and the "Any Annotator". We see that for most annotators the fraction of explanations that are at least partially covered is quite high. This means that at least partial semantics will be available for explanations on the vast majority of cases. We also see that the "Any Annotator" bar is similar to the individual annotators, which indicates that the sets of partially covered explanations across annotators is similar. The complete coverage percentages drop substantially, which is not surprising given the results for activation coverage from Figure 3. Once again the "Any Annotator" bar is not significantly different from the rest.

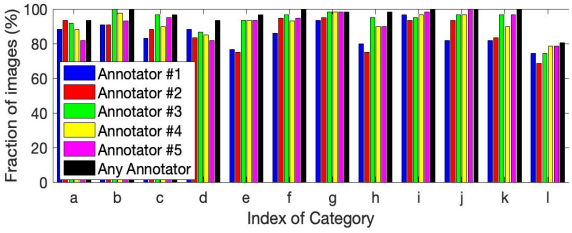

Figure 4: Partial explanation coverage for each category across annotators.

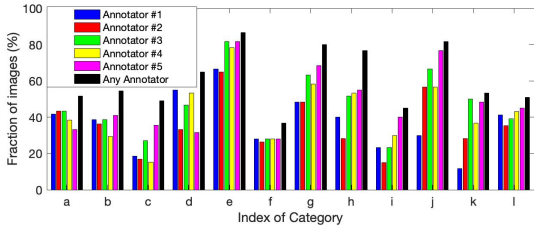

Figure 5: Complete explanation coverage for each category across annotators.

Overall, we see that a non-trivial fraction of significant annotations are not named by users. This necessarily resulted in less than 50% complete coverage rates for most users. From further analysis, the low rates of complete coverage seems to be due to the fact that in most cases the final decision is dominated by a single significant activation. The other significant activations appear to be not strong enough to lead to easily recognizable concepts and hence are not named by the annotators. Still there are a non-trivial number of examples that are completely covered, which allows for analysis of full explanations for a fraction of the test data. Even this can help build trust and identify potential flaws. On the other hand, most explanations are at least partially covered, which allows for gaining some semantic insight into most decisions. Even partial explanations can be useful in identifying flaws and building trust. These results suggest future work on increasing the coverage by bootstrapping from partial coverage, e.g. by incorporating the annotations into retrained DNNs.

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, when activations highlight the edge of the image or fall on background with unclear semantics. 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.

3.2 RQ2: Inter-annotator Agreement

In general we can expect different annotators to produce different namings for a test set, where at least some of the visual concepts differ. Here we consider the extent that these different namings agree and in turn whether explanations produced by different namings are semantically similar. Understanding this issue is important for understanding the extent to which explanations are fundamentally annotator specific.

First, we consider annotator agreement about which significant activations should be named. Figure 6 shows, for each bird category, the fraction of significant activations that were named by different numbers of annotators - 0 thru 5. Interestingly, the largest fraction of activations are annotated by all 5 annotators and the second largest are annotated by 0 annotators. This confirms, once again, that for most significant activations, either all annotators choose to assign a name or none of them do. There is strong agreement about the set of activations that should be named.

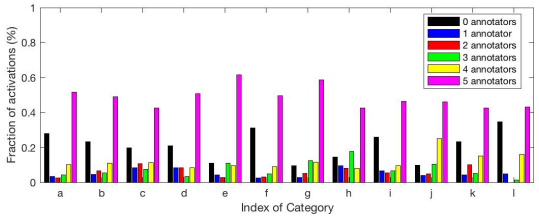

Figure 6: Fraction of significant activations that are named by exactly n of the annotators, where $n \in \{0, 1, 2, 3, 4, 5\}$

Next, we want to know how similar are the concepts and find potential translations between the concepts of different annotators. Are there one-to-one correspondences, subsumption relationships, or cases of purely incompatible concepts?

Given two namings N_i and N_j produced by two 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 [3]. The framework first defines the "intersection graph" G of N_i and N_j. G is a bipartite graph where the vertices in each partite set correspond to the clusters of N_i and N_j respectively. Thus, a large weight between two visual concepts indicates that they represent many of the same activations in the test set. D-family matching is a partition of all nodes (that belong to either naming) of the bipartite graph into some number of disjoint sets $S_1, S_2, \ldots$ 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 7).

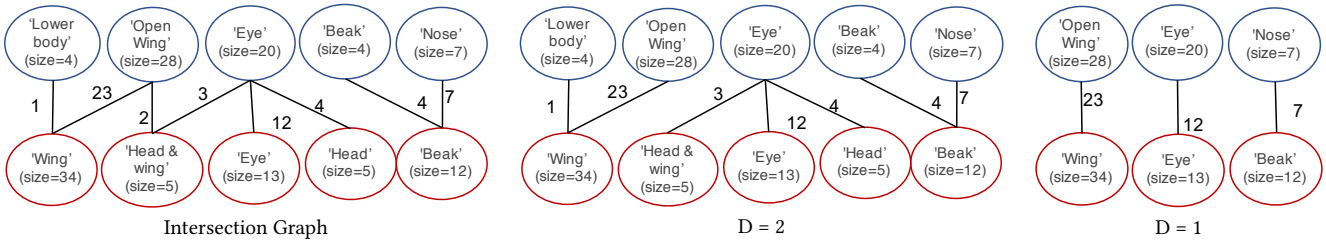

Figure 7: An example of pairwise similarity matching between two annotators. Blue and red circles represent the namings produced by two different Annotators.

Since not all significant activations are labeled by all annotators, we first try to characterize the fraction of common annotations between pairs of annotators. Thus, 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 2 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.

	Pairwise similarity scores		
Category	Agreement (D=1)	Agreement (D=2)	Jaccard index
a	0.74±0.09	0.96±0.04	0.82±0.04
b	0.82±0.05	0.91±0.04	0.81±0.04
c	0.79±0.07	0.92±0.05	0.75±0.05
d	0.96±0.02	0.98±0.01	0.8±0.04
e	0.88±0.06	0.97±0.04	0.83±0.05
f	0.77±0.13	0.94±0.04	0.86±0.04
g	0.69±0.07	0.91±0.07	0.8±0.05
h	0.73±0.06	0.86±0.07	0.7±0.08
i	0.69±0.07	0.85±0.07	0.8±0.04
j	0.74±0.1	0.85±0.05	0.75±0.13
k	0.71±0.1	0.92±0.05	0.77±0.08
l	0.86±0.09	0.95±0.05	0.85±0.06
Average	0.78	0.91	0.79

Table 2: Pairwise comparison between clusters generated by annotators over all categories

We compute the agreement between the two annotators 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 matchines as translations between namings, then the agreement is the fraction of activations that are translatable between namings. The columns labeled "Agreement" in Table 2 shows the statistics of 1-family and 2-family agreements for each category over the set of all annotator pairs. The agreement numbers are fairly high across most categories, although the minimum values for some categories for $D = 1$ are low. Since 2-family matching is more permissive than 1-family matching, the agreement numbers are higher for $D = 2$ as we expect. Even for $D = 1$ the agreement in most categories is reasonably high, which shows that there is reason to be optimistic about developing a common ontology for explanations.

Test Set Explanation Summaries. One of the motivations for naming a test set is to produce summaries of the explanation types used to predict test images. For example, for category'a' over 56% of the test set predictions had the explanation ('eye', 'close wing'), which indicates that the network was focusing on the bird eye and closed wing area for those examples. As another example, for category 'l' over 88% of the predictions had the explanation ('eye'), which means the network only looked at the eye area to make the prediction. This type of insight may cause a practitioner to either question the robustness of the classifier if they have reason to believe the eye alone is not discriminative enough. Alternatively, an expert may gain insight from this explanation and realize that the eye is discriminative enough for the task.

4 DISCUSSION AND CONCLUSIONS

In this paper we studied the problem of understanding the decisions of DNNs in terms of human-recognizable visual concepts. Our interactive-naming approach involved 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 producing concise meaningful explanations for test set images. We reported on our experience of having 5 annotators use our interface for DNNs trained to recognize different bird species. Our results showed that: 1) annotators were able to assign names to a non-trivial fraction of activations, which allows for at least partial semantic explanations for most test images; 2) the annotators had strong agreement about which activations should and should not be named; 3) there was a non-trivial amount of agreement between the namings produced by different annotators,

This formative study has set the stage for a variety of future work. Our current interactive naming interface is flexible, but does not attempt to actively reduce the annotator effort. Thus, there is potential to improve the speedup of naming a test set via active learning techniques. We are also interested in interactively training the system based on named concepts, which might reduce the number of activations that cannot be named. In addition, investigations on other datasets with even larger varieties of visual concepts is important for understanding the general characteristics of annotator produced namings.

ACKNOWLEDGMENTS

The authors acknowledge the support of grants from NSF (grant no. IIS-1619433), ONR (grant no. N00014-11-1-0106), and DARPA (grant no. DARPA N66001-17-2-4030).

REFERENCES

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10 (10 7 2015). https://doi.org/10.1371/journal.pone.0130140

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Computer Vision and Pattern Recognition*.

[3] Frédéric Cazals, Dorian Mazauric, Romain Tetley, and Rémi Watrigant. 2017. *Comparing two clusterings using matchings between clusters of clusters.* Research Report RR-9063. INRIA Sophia Antipolis - Méditerranée ; Universite Cote d'Azur. https://hal.inria.fr/hal-01514872

[4] Piotr Dabkowski and Yarin Gal. 2017. Real Time Image Saliency for Black Box Classifiers. In *NIPS*.

[5] R. C. Fong and A. Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 3449–3457.

[6] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. 1986. Distributed representations. In *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*, D. E. Rumeihart, J. L. McClelland, and the PDP Research Group (Eds.). MIT Press, Cambridge, MA, 77âĂŞ–109.

[7] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. 2673–2682. http://proceedings.mlr.press/v80/kim18d.html

[8] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. ACM.

[9] Zhongang Qi, Saeed Khorram, and Fuxin Li. 2017. Embedding Deep Networks into Visual Explanations. *CoRR* abs/1709.05360 (2017). arXiv:1709.05360

[10] Advait Sarkar, Cecily Morrison, Jonas F. Dorn, Rishi Bedi, Saskia Steinheimer, Jacques Boisvert, Jessica Burggraaff, Marcus D'Souza, Peter Kontschieder, Samuel Rota Bulò, Lorcan Walsh, Christian P. Kamm, Yordan Zaykov, Abigail Sellen, and Siân Lindley. 2016. Setwise Comparison: Consistent, Scalable, Continuum Labels for Computer Vision. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 261–271. https://doi.org/10.1145/2858036.2858199

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

[12] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *CoRR* abs/1605.01713 (2016). http://arxiv.org/abs/1605.01713

[13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ICLR Workshop* (2014).

[14] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR Workshop*. http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a

[15] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3319–3328.

[16] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset.* Technical Report CNS-TR-2011-001. California Institute of Technology.

[17] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 818–833.

[18] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2016. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*. Springer, 543–559.

[19] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable Basis Decomposition for Visual Explanation. In *Computer Vision – ECCV 2018*.