https://doi.org/10.2352/EI.2026.38.1.VDA-349

Easy Interpretation of Image Classification Results with Feature-Level Visualization

Deepshikha Bhati; Department of Computer Science, Kent State University, Kent, OH, USA
Ye Zhao; Department of Computer Science, Kent State University, Kent, OH, USA
Tsung-Heng Wu; Department of Computer Science, Kent State University, Kent, OH, USA
Md Amiruzzaman; Department of Computer Science, West Chester University, West Chester, PA, USA
Jing Yang; Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

Abstract

Given a picture classified as a Persian cat by an AI model, users may ask questions such as, "What are the contributions of the eyes and ears to the classification result?" or "Which features contribute the most?" While existing post-hoc XAI methods effectively explain model predictions at the pixel or patch level, they are limited in directly quantifying the contributions of human-interpretable semantic features. In this paper, we propose a visual analytics approach for feature-level interpretation of image classification results. Our contributions are twofold. First, we introduce a semantic contribution quantification method that builds upon existing pixel-level attribution techniques (e.g., Layer-wise Relevance Propagation, Grad-CAM). Specifically, we aggregate and normalize pixel-level relevance scores over predefined semantic regions (such as eyes, ears, and body) to compute comparable contribution scores for each semantic feature within an image. Second, we present an interactive visual interface that leverages these quantified semantic feature contributions to support exploration, comparison, and analysis of AI outputs across image collections. Through illustrative scenarios and expert feedback, we demonstrate that our approach provides an intuitive, scalable, and semantically meaningful means to interpret image classification explanations.

Introduction

AI-based image classification has achieved considerable success in various applications. Meanwhile, numerous post-hoc eXplainable AI (XAI) methods have been developed to provide insights into the decision-making process of AI models after they have been trained, offering a means to understand how specific input image components contribute to a model's predictions. These methods can be broadly classified into two categories:

- *Pixel-level Methods:* These methods compute a *score of contribution* for each pixel in the input image (often referred to as importance, attribution, saliency, relevance, etc.) with respect to a specific classification [27]. Then, the contribution scores are visualized as relevance maps (also known as attribution maps, saliency maps, etc.), where a heatmap colorization (e.g., blue to red) is applied to indicate varying levels of contribution (from low to high). The contribution scores are typically visualized as relevance maps (also known as attribution maps, saliency maps, etc.), where a heatmap colorization (e.g., blue to red) is applied to indicate varying levels of contribution (from low to high). Fig. 1 shows an example of a cat image. These heatmaps, when paired with the original input image, allow users to discern the features that contribute to the AI model's classification.
- *Patch-level methods:* These methods automatically cluster image patches, manually associate these clusters with abstract "concepts", and then automatically evaluate the influence of the concepts on the prediction of a class of input images (e.g., [1, 3, 4, 10, 21]). For instance, the concept of "stripes" might be discovered for its critical impact on classifying images as "zebras".

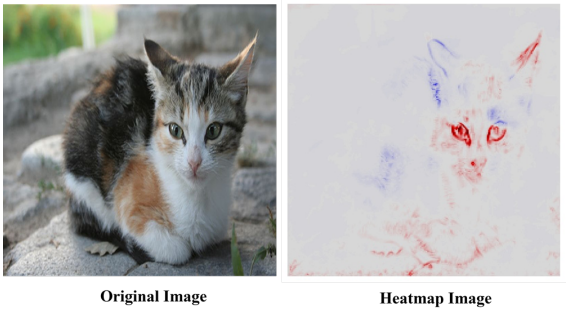


Figure 1: *Original image of a cat (left) and its corresponding Layer-wise Relevance Propagation (LRP) heatmap (right), where red pixels refer to positive relevance and blue pixels refer to negative relevance.*

Both of these post-hoc XAI approaches have demonstrated success in explaining image classification results. However, the pixel-level methods usually convey computational results in the form of heatmaps. It is the observer's responsibility to visually interpret this information, identifying high-level features such as the ears or eyes based on pixel color intensity. This visual perception process can be error-prone and time-consuming, especially when analyzing a large number of images. Furthermore, the choice of color mapping schemes of the heatmap (e.g., different color palettes [23]) can introduce biases and undermine the effective perception of the pixel values and their differences.

On the other hand, the image patch clusters generated by patch-level methods sometimes do not lead to meaningful concepts. Moreover, the discovered concepts often do not align with human-recognizable, intuitive features such as eyes or ears, so that users cannot specify features of interest to evaluate their contributions.

Since interpreting XAI results at the pixel or image patch

levels may hinder effective information conveyance to AI users, we propose a new method to interpret XAI results in terms of predefined, intuitive semantic features, such as ears, eyes, and body. This process transforms low-level explanations into human-interpretable summaries aligned with familiar semantic features. AI users, including the general public and practitioners, can quickly understand and examine the AI output with the semantic features they are familiar with and interested in. We calculate the contribution scores of these semantic features to the image classification results. These contribution scores can be used to enable intuitive and efficient comprehension, query, and exploration of XAI results in the semantic feature space. Our method aggregates pixel-level relevance scores over predefined semantic regions (e.g., ears, eyes, body), which can be built up on established computational methods such as *Grad-CAM* [28], *SHAP* [20], *Integrated Gradients (IG)* [32], and *Layer-wise Relevance Propagation (LRP)* [25]. We also built a prototype named *CatViz* to demonstrate how to employ these contribution scores to visually explore the XAI results of a collection of images.

In summary, the main contributions of this paper include:

- We propose a new way to interpret AI image classification results at the semantic feature level, to address the limitations of existing pixel-level and patch-level methods, such as the lack of direct connections between the interpretation and familiar semantic features.
- We present a computational approach to calculating the contribution scores of predefined semantic features to the classification results of individual images, based on the XAI results of established XAI methods.
- We have developed a working visualization prototype to demonstrate how to employ semantic contributors and their contribution scores to explore the XAI results of an image collection with visual analytics tasks, such as clustering, filtering, and comparison of groups of images.

Related Work

XAI is drawing more attention from research communities because of its ability to enhance data and model explainability [19] in many applications. In explaining image classification results, two major categories of approaches are pixel-level and patch-level methods (see a recent survey [5]).

Pixel-level XAI Methods: Four major categories are discussed in the survey [27], including local surrogates, occlusion analysis, integrated gradient, and Layerwise Relevance Propagation (LRP) techniques. Most of these methods are considered local methods [1] since they compute relevance scores (importance, saliency, etc.) for image pixels on individual images given a classification class. Examples include Deconvolutional Networks (DeconvNets) [29], Guided Back Propagation [31], Class Activation Mapping (CAM) [18], LRP [2], and Class-discriminative LRPs [13].

Pixel-level heatmaps are often used to visualize the computed pixel-level relevance scores [27]. For example, Polarized-LRP [17] uses both positive and negative heatmaps for a specific application of interpreting galaxy deblender GAN. VisLRP [12] presents a visual model designer that helps LRP designers and students efficiently design, explore, and find suitable LRP models with preferred parameters.

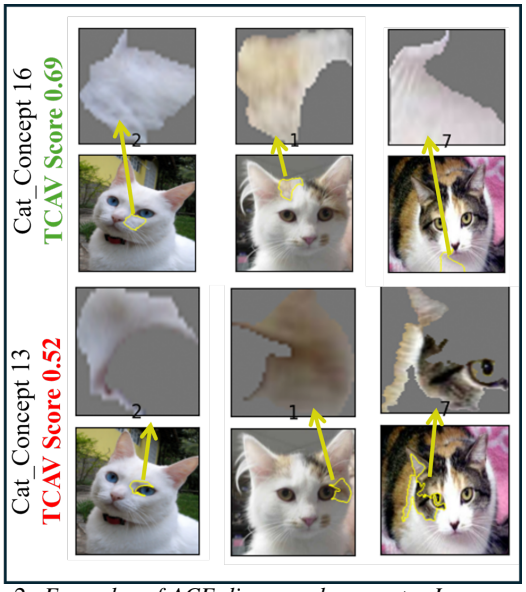


Figure 2: *Examples of ACE discovered concepts. In our experiments with a cat image dataset, the ACE algorithm [10] generated concepts such as "fur texture" and "eye region texture".*

Our method leverages the pixel-level relevance scores to compute the contribution scores at the semantic feature level for individual images. Thus, it explains the images at a higher level of abstraction and enables more effective, efficient, and scalable visual explorations.

Patch-Level XAI Methods: Patch-level methods detect higher-level features or concepts learned by a model on entire classes of images [1, 3, 15, 21, 33]. Given a set of images from the same class, each image is segmented into a pool of patches. These patches are clustered and the clusters of patches form a "concept" and its importance score is computed over the images in the given class. Concept-based interpretability methods such as TCAV [14] and prototype-based approaches [6, 7, 26] also aim to connect human-defined concepts or prototypical parts with model predictions. These methods provide complementary perspectives on concept attribution, but they rely on either learned or post-hoc concept definitions.

The image patches are visually presented to users for their understanding of the concepts with respect to the given image classes. Visual analytics tools can help users perform interactive editing and exploration of these concepts and patches. ConceptExtract [36] develops a visual system that helps users analyze model behavior, extract human-friendly concepts, and use these concepts to understand the predictions. It aims to add human feedback to a concept extractor from the clusters of image patches, so as to reduce human labeling effort. In addition, ConceptExplainer [11], a visual analytics system, allows users to explore the clustered image patches and classes for the explanation of the "concepts". The usage scenarios include reviewing related image clusters, identifying unreasonable concepts, and studying multiple concepts from image classes.

Our approach focuses on the semantic feature-level contribution, which is different from the above patch-level methods.

Rationale and Tasks of Our Approach

We have implemented a widely-used patch-level method - ACE [11] to an image dataset including 50 cat images from ImageNet [9]. ACE segments the input images into patches and the latent representations of these patches are clustered. The top 25 clusters are considered as the concepts and each concept is measured by a TCAV (Testing with Concept Activation Vectors) score for its influence on the classification. These top concepts include both meaningful and irrelevant concepts. The irrelevant concepts are difficult for users to explain. For meaningful concepts, two important ones are shown in Fig. 2, which can be named "fur texture" or "eye region texture". However, these concepts do not align with commonly recognized semantic features of cats (e.g. eyes or ears). This experiment reveals several limitations of the patch-level methods: 1) The unsupervised patch-level methods often capture features that are statistically prominent but lack direct interpretability; 2) The identified concepts require human interaction to extract meaningful results; 3) Users need to explain the image patches and provide semantic names or labels; 4) Users cannot specify features of interest to examine their contributions.

To address these limitations, we present a new method which discovers the contribution scores of **semantic features**, such as distinct parts of animals (e.g., eyes, ears, body), to AI classification results. We refer to these features **semantic contributors**, as they enable explanations that align with human expectations and offer insights directly relevant to the features of interest. The following analysis tasks, which are important but require tremendous human efforts in current XAI practice, can benefit from utilizing the contribution scores of the semantic contributors:

T1 - Contributor Ranking and Comparison: Contribution scores allow users to compare and rank the semantic contributors within an image based on their significance in the model's decision-making process. This quantitative analysis is crucial for understanding feature relevance. For instance, a feature such as the "Left Eye" may be more relevant than the "Left Ear" in predicting a cat image, which might not be immediately apparent from a traditional relevance heatmap.

T2 - Group Analysis Using Semantic Contributors: Contribution scores enable users to identify groups of images with similar semantic features (e.g., cats characterized by strong Left Eye contributions or birds influenced by prominent Beak features) and analyze the reasons behind correct or incorrect model classifications. By leveraging contribution scores, users can perform various data exploration operations, such as filtering, ranking, and outlier detection, in the semantic feature space for an image collection, to reveal patterns and insights that would be difficult to discern using other methods.

Algorithms for Computing Semantic Contribution Scores

Our method computes the contribution scores of semantic contributors in an input image, given a classification decision, based on two sources of prior information: (1) a relevance map (also referred to as a saliency map or attribution map) with pixel-level relevance values generated by any XAI algorithm (e.g., Grad-CAM, SHAP, IG, LRP); and (2) semantic features of the input image that are obtained through semantic segmentation algorithms. Numerous methods exist for partitioning an image into semantically meaningful regions or objects (see the survey in [22]). Next, we introduce the algorithms for computing semantic contributor scores.

Computing Pixel-Level Relevance Values

In our implementation, we utilize the popular LRP method to generate a relevance map. LRP computes relevance values that quantify the contribution of network components and input pixels to produce a classification decision of an input image [24].

The algorithm initiates the relevance values based on a selected output class. In particular, a backward propagation from the output layer to the lower layers is employed to compute relevance values at each layer and towards the input pixels. Each neuron receives a share of the relevance values from its successor nodes and redistributes its relevance to predecessor neurons, while the conservation of relevance is ensured. During this backpropagation process, the distribution of relevance can be computed by using different relevance propagation rules (i.e., functions) that utilize the forward neuron activations and a set of artificial parameters.

Specifically, a relevance vector $\vec{r}$ is defined by the values corresponding to a given classification category C such as:

$$\vec{r} = \begin{cases} s_i, & i = C \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Here, N is the number of classification categories and $i \in [1 \cdots N]$. s_c is the pre-softmax value of category C.

Then, a backward propagation from the output layer to the lower layers is employed to compute relevance values at each layer and towards the input pixels. The computation from layer $l+1$ to layer l follows an LRP function F as:

$$\vec{r}_i^{\,l} = \sum_j F(a_i, w_{ij}) \vec{r}_j^{\,l+1}, \tag{2}$$

while $\sum_i \vec{r}_i^{\,l} = \sum_j \vec{r}_j^{\,l+1}$ for the preservation of relevance values. Here w_{ij} is the network weight from neuron x_i^l to neuron x_j^{l+1}, and a_i is the activation of each neuron x_i^l. The LRP function can be defined by different rules. One widely used LRP rule, $LRP-\varepsilon$, implements F as:

$$\textit{LRP-}\varepsilon: \quad \vec{r}_i^{\,l} = \sum_j \frac{a_i w_{ij}}{\varepsilon + \sum_i a_i w_{ij}} \vec{r}_j^{\,l+1}, \tag{3}$$

It uses a constant ε to keep numerical instability, a condition where division by near-zero values may cause extreme fluctuations or undefined outputs. This ensures stable and consistent propagation of relevance values during backpropagation, particularly when input activations or weights are very small.

This iterative relevance propagation process reaches the input layer (i.e., input image) and it generates a relevance map of the input image for the category c.

In our experiments, we apply $LRP-\varepsilon$ on a CNN model of VGG16 for a group of cat images from the ImageNet database. The original image $I_{i,j}, i \in [0 \cdots 223], j \in [0 \cdots 223]$ with 224×224 pixels. Then the resulting relevance map is defined as $IR_{i,j} \in (\alpha, \beta)$ with $i \in [0 \cdots 223], j \in [0 \cdots 223]$. Here, β is the maximum relevance value, and α is the minimum. Note that these values include positive and/or negative scores showing the pixels may have either positive or negative (i.e., adverse) contributions to the classification.

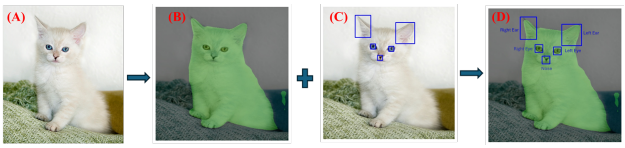


Figure 3: *Illustration of semantic feature extraction for a cat image.*

Semantic Feature Extraction

Extraction of meaningful features or objects from an input image is a key topic in semantic image segmentation [22]. Deep Learning (DL) models are trained with vast datasets for accurate and multiscale object extraction. However, there is no universal model that can accurately extract the preferred semantic features from any image.

We conducted experiments using raw images of cats from the ImageNet database and bird images from the CUB-200 dataset. The target features for cats included body parts such as the eyes, ears, and nose, while for birds they included the eyes, wings, and beak. However, we did not identify an optimal deep learning (DL) model capable of reliably extracting these features. Since the design of a specialized automatic DL model is beyond the scope of this paper, we instead adopt a human-in-the-loop strategy, commonly applied in practical settings, that combines DL tools with user intervention.

While our focus is on quantifying contributions of fine-grained semantic features (e.g., left ear, right ear, eyes), current semi-automatic segmentation tools, such as the Segment Anything Model (SAM) [16], are not designed to reliably extract such detailed features, since they typically operate at the object level. Therefore, manual annotation remains necessary in this study to ensure semantic fidelity. Future work may explore domain-specific annotation strategies to improve scalability. At this stage, we did not attempt to define thresholds for manual intervention; instead, we relied on manual annotation to ensure semantic fidelity. Future work may investigate human–computer collaboration strategies to balance automation efficiency with semantic accuracy, including threshold design for semi-automatic correction.

First, we apply a popular DL model named DeepLabV3 [8] to extract the whole shape of the cat/bird from its background. As illustrated in Fig. 3, the input image Fig. 3(A) is segmented to extract the body of a cat shown in green in Fig. 3(B). Second, we adopt a human annotation-based approach to further identify the body parts. In particular, human annotators draw boxes over the image to annotate regions such as the right ear and left ear, as shown in Fig. 3(C). These bounding boxes are joined with the cat shape to finally extract pixels belonging to these features (Fig. 3(D)). Please note that other AI-based automatic and semi-automatic tools of image object segmentation [35] may also be applied in this preprocessing stage. We selected DeepLabV3 as it provides robust semantic segmentation with manageable complexity, suitable as a baseline for our manual refinement. While models such as Mask R-CNN excel at object- and instance-level segmentation, they are not optimized for fine-grained, part-based semantic features required in this study.

After applying these operations to an input image, we achieve a set of K semantic features $O_k, k \in [1 \cdots K]$. Each O_k has its semantic label and refers to a specific region on the original image I and on the relevance map IR. This set covers the whole image region. For instance, the cat image in Fig. 3 has $K = 7$ semantic features {Right Ear, Left Ear, Right Eye, Left Eye, Nose, Body Part, Background}. Here, the background feature represents the remainder after Fig. 3(B) extracts the shape of a cat. For bird images, we similarly set $K = 7$ semantic features {Left Eye, Right Eye, Left Wing, Right Wing, Beak, Tail, Body} as shown in Fig. 7

Semantic Contributors and Contribution Scores

For each feature O_k, we identify its corresponding region on IR and define a feature-level contribution score H_k. Defining the computing algorithm of H_k is not an easy task. We tested several approaches and found that the following definition meets our requirement, allowing quantitative comparison among features in different images.

First, we compute the average relevance value of a feature as:

$$A_k = \frac{\sum_{i,j}^{N} IR_{i,j}}{N}, i, j \in O_k \tag{4}$$

Here $IR_{i,j}$ is the relevance value for a pixel at position (i, j) and N is the total number of pixels within O_k.

However, the average relevance value A_k cannot be directly used as the contribution score of O_k. Although it facilitates a quantitative comparison of two features on the same image, it does not support a direct comparison of the same semantic features on different images. The reason comes from the fact that the LRP computation on different images IR generates relevance values with different value ranges α, β and which are not quantitatively comparable. Therefore, we define the **contribution score** of a semantic contributor as the ratio of a feature's A_k value to the sum of all $A_k$on an image:

$$H_k = \frac{A_k}{\sum_{k=1}^{K} A_k}. \tag{5}$$

In implementation, this contribution score should be computed independently for negative and positive contributions. In the following, we represent contribution scores as percentages. They represent the relative contribution percentage of each feature O_k with respect to all features. It enables ranking and comparison of semantic feature contributions among different images. For instance, if a feature O_1 on a cat image contributes $H_k = 30\%$ to its classification and a feature O_2 in another cat image contributes $H_k = 40\%$, then we can consider O_2 to have a larger contribution than O_1.

We intentionally compute the average attribution per semantic region so that each predefined feature (e.g., left ear, right ear) is treated as a meaningful unit, preventing large but uninformative areas (such as background) from dominating the score.

Note that negative scores are included in the denominator to capture inhibitory contributions, ensuring that both positive and negative influences are proportionally reflected.

Contribution Vectors and Their Similarities

The contribution scores H_k for $k \in [1 \cdots K]$ form a vector $\vec{H}$ representing the contributions of semantic features. Consequently, this vector provides a quantitative representation of IR.

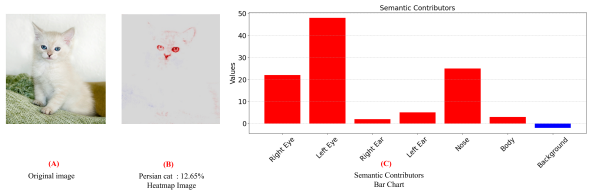


Figure 4: *Visualizing the contribution scores of seven semantic features with a bar chart, together with the original image and the heatmap image. They show the quantitative contribution information to the classification category of Persian cat.*

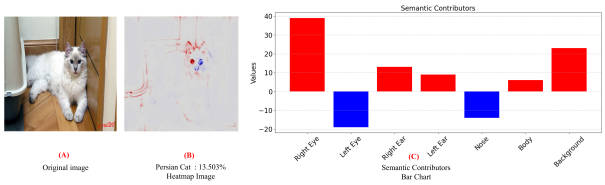


Figure 5: *Visualizing positive and negative contribution scores of semantic features towards a Persian cat classification.*

We can compute similarities among images based on their contribution vectors, allowing clusters and outliers to be visually explored. This facilitates a new way of analyzing XAI results, which has not been examined in previous work. In our experiments on image groups (e.g., multiple cat or bird images), we use cosine similarity within an embedded visualization interface.

Visual Examples

Using the semantic contributors makes it easy to visualize their scores with bar charts leading to benefits of simplicity, user familiarity, and ease of quantitative comparison. Next we show examples with several images from ImageNet.

Example 1: Fig. 4 shows an original image, its heatmap image of relevance map from LRP, together with a bar chart visualizing the contribution scores of seven semantic contributors. These are the computing results based on the relevance map of the top classification Persian cat of Fig. 3. It can be observed from the bar chart that Left Eye has the largest score, Nose is the second, and Right Eye is the third contributor. This provides an easy insight from the bar chart. Without the contribution scores, users can only examine the heatmap image, and it is not easy to discover whether Nose has a stronger or weaker relevance than Left Eye and Right Eye.

Example 2: In Fig. 5, the red and blue bars clearly indicate that this image has both positive and negative contributors to a Persian cat classification. Positive contributors, such as Right Eye (39%), Right Ear (13%), Left Ear (9%), Body (6%), and Background (23%), can be easily identified, along with negative contributors like Left Eye (-19%) and Nose (-14%). Despite Left Eye having a large negative effect, Right Eye has a stronger positive effect, resulting in the correct classification of this image as a Persian cat. Human observers may not see a big difference in the original image, but the AI model may interpret Left Eye and Nose in this image as atypical styles for a Persian cat. Another interesting finding is that the Background region has a relatively large relevance to the classification.

Example 3: Fig. 6 shows an image with a top classification of Pumpkin, although a cat sits close to the pumpkin. The bar chart shows the contribution scores to this top decision. Here, the scores of the cat features are Right Eye (-14%), Left Eye (-36%), Right Ear (-7%), Left Ear (-8%), Nose (-23%), and Body (-8%), juxtaposed with a positive contribution from the background (46%). The features on the cat all have negative contributions to the class pumpkin. However, the background (including the pumpkin) has a larger quantity of contribution score (46%), which prevents the image from being identified as a cat. With the semantic contributors, users can quantify this explanation through the visualization.

Summary of Benefits: In the three examples, the visualization of semantic contributors gives users an immediate gauge to explain the behaviors of the AI model. It overcomes potential problems that may be introduced when creating and interpreting a heatmap. It also allows users to perform rapid comparisons based on quantitative values and, therefore, enhance the understanding of AI classification output.

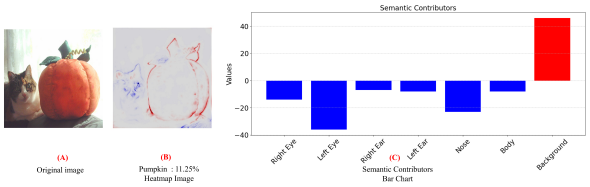


Figure 6: *Visualization of contribution scores for semantic features in an image of a cat sitting next to a pumpkin, illustrating their relative contributions to the classification category Pumpkin.*

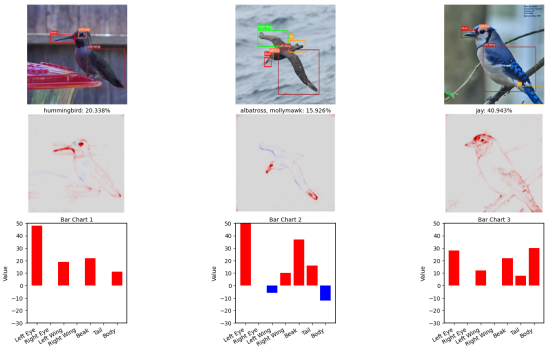


Figure 7: *Three randomly selected bird images (top), their LRP heatmaps (middle; red indicates positive relevance to the predicted class and blue indicates negative relevance), and the corresponding semantic contribution summaries (bottom). The bar charts report signed contribution scores for $K = 7$ features:{Left Eye,Right Eye,Left Wing, Right Wing, Beak, Tail, Body}*

CatViz for Exploring Semantic Contributions of an Image Collection

We study feature-level semantic contributors to AI predictions across image collections through *CatViz*, an experimental visual analytics prototype. In its current form, CatViz is intentionally scoped to a single curated dataset of cat images, and all examples, figures, and analyses in this paper are drawn exclusively from this cat collection. Restricting the interface to one class keeps the semantics consistent and enables controlled, within-class comparisons of contribution patterns across images; extending CatViz to additional classes is left for future work.

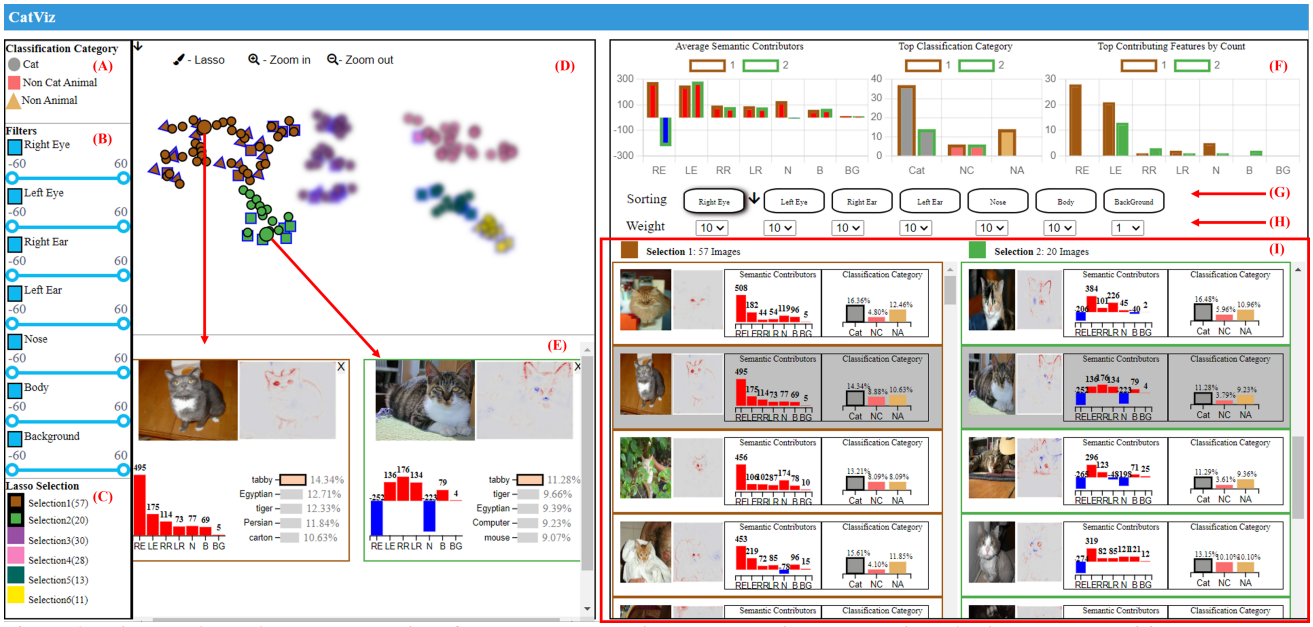


Figure 8: *The interface of CatViz. (A) Classification category selectors; (B) Filters; (C) A list of selections created by a user; (D) A scatterplot where each circle/square/triangle represents an image. Images in the brown cluster and the yellow clusters are highlighted; (E) Details of two images that a user clicked on the scatterplot; (F) Statistics bar charts of the brown cluster and the yellow cluster. Their bars are highlighted by brown and yellow boundaries, respectively. From left to right: the average scores (multiplied by weights) on the semantic contributors, the top classification categories, and the top contributing features by count; (G) Sorting buttons; (H) Weight controllers; (I) Left/right: Details of images in the brown/yellow clusters. Both clusters are sorted by Right Eye score from high to low.*

CatViz Tasks

CatViz allows users to interactively explore a group of images and their semantic contributors. They can easily conduct an in-depth investigation of an AI model's prediction output in the space defined by contribution scores for multiple semantic contributors. The system aims to address the following tasks that are not well supported in existing XAI approaches:

- At the group level, to learn how each semantic feature contributes to the prediction results, identify the major contributors, and discover contributors negatively contributing to the results. Such information about an image group will help users get a comprehensive understanding of the output of the AI model.
- To discover clusters and outliers of images identified using their similarities in the contribution scores, and learn their contribution score patterns. Such insights can help users in tasks such as selecting training images.
- To examine details of a cluster, and compare multiple clusters for their shared and distinct contribution score patterns. This will help users discover and understand different paths to the prediction results.
- To conduct detailed analyses of the relationships between a prediction result and the contribution scores at the individual image level. Such a drill-down study can help users get insights, such as why specific images are misclassified.
- To retrieve images and heatmaps based on their contribution scores, and rank them by scores on one or more semantic contributors of interest. This allows users to delve into a smaller subset of images and heatmaps with interesting semantic features for further analysis.

As an experimental system for exemplifying the usages of the contribution scores of semantic contributors, CatViz provides a set of coordinated visualizations and interactions based on the contribution scores of semantic features. They are introduced in the following sections.

CatViz Visualization Interface

Scatterplot: In the scatterplot (see Fig. 8D), the images are mapped from the high dimensional space defined by the contribution score vectors to a 2D embedded space using t-SNE [34], a dimensionality reduction technique. Each image is shown as a symbol representing their classification categories, and the adjacency of their positions reflects their similarities in contribution scores. From this view, users can visually discover clusters of images with similar contributions on semantic features and identify outliers. Users can use a lasso tool to define multiple image selections, each having a set of interesting images for further analysis. Users can click on any image in this view, and its raw image, heatmap, contribution score bar chart, and classification category will be displayed for detailed analyses.

Statistics bar charts: A set of bar charts showing aggregated information of one or two selections of multiple images (from the lasso tool), including their average contribution scores on semantic contributors, top classification categories, and top contributors (see Fig. 9A, Fig. 8F). These bar charts allow users to learn and compare the contribution scores of the image selections side-by-side.

Detail panel: This panel presents detailed information about an image selection, including their raw images, heatmaps, contribution score bar charts, and classification categories. For two selections, the information will be displayed side-by-side in the detail

panel for comparison (see Fig. 8I). This panel enables users to closely inspect individual images to gain deeper insights into their characteristics and the underlying reasons for their classification.
Control panel: This panel allows users to manage the classification categories (Fig. 8A) highlighting on the scatterplot. It also includes a set of filters (Fig. 8B) for interactive control of images displayed. Moreover, the existing selections of images are listed and users can manage them in Fig. 8C.

CatViz Visual Interactions

CatViz provides the following interactives based on the contribution scores of features:
Filtering: Users can filter images using a set of score range filters, one for each semantic contributor as shown in (Fig. 8B). This allows users to define thresholds of contribution scores (H_k in Eqn. 5) of all semantic contributors, so as to narrow down to subsets of images with interesting contribution score patterns. Here, the contribution scores can be changed in the range of [-60% $\cdots$ 60%]. The threshold limits are defined since for this dataset the largest possible contribution score for all features is lower than 60%.

CatViz provides real-time updates on the scatterplot and the detailed image display when users move slides on the filters for instant visual feedback. The positions of the images in the scatterplot remain the same for visual consistency.
Ranking: Users can rank images in the detail panel in ascending or descending order based on their contribution scores of any semantic contributors as shown in (Fig. 8G, Fig. 10A, B). This interaction allows users to identify score trends and patterns on the sorting contributor, as well as explore the relationship between the sorting contributor and the selection criteria (e.g. lasso selected clusters, selection by classification outcomes, and more).
Dynamic clustering with user-adjusted semantic contributor weights: CatViz allows users to interactively assign external weights to the semantic contributors (Fig. 8H). Semantic contributors with higher weights will have bigger influences on the t-SNE projection. In similarity computation, the classic cosine similarity of two contribution vectors $\vec{H}_l, \vec{H}_m$ of two images l and m is improved as:

$$\delta(\vec{H}_l, \vec{H}_m) = \frac{\sum_{k=1}^{K}(H_{lk}W_k)(H_{mk}W_k)}{\sqrt{\sum_{k=1}^{K}(H_{lk}W_k)^2 . \sum_{k=1}^{K}(H_{mk}W_k)^2}} \tag{6}$$

where W_k is the weight of feature k, K is the number of features. Here, δ denotes the similarity measure between the semantic contribution vectors of two images, computed using cosine similarity.

Consequently, higher-weighted semantic contributors shape the layout of the t-SNE projection, and the scatterplot is updated accordingly. This interaction allows users to adapt the cluster structure in the scatterplot based on their analysis interests for more customized cluster and outlier identification.

Cosine similarity is employed because it captures the relative distribution of contributions across semantic features, consistent with our normalized score design, whereas Euclidean distance is sensitive to absolute scale.

Use Scenarios of CatViz

This section demonstrates the functional validation of CatViz through usage scenarios, highlighting how the system supports exploration and interpretation of semantic feature contributions.

In this section, we present a few example scenarios to demonstrate the usefulness of CatViz. In these scenarios, a user (named Alice for easy description) explores the cat dataset. Among them, 160 images are classified as cats, 25 images are classified as non-cat animals, and 25 images are classified as non-animal categories by the VGG16 model [30].

In our examples, seven semantic contributors of these cat images are processed by the proposed algorithms including Left Eye (LE), Right Eye (RE), Left Ear (LE), Right Ear (RE), Nose (N), Body (B), and Background (BG). Their contribution scores are all computed to a top category of cats, which are used for visual analysis. Please note that: (1) For images with a top classification as non-animal objects or non-cat animals, we use its first cat category in classification instead of using their top non-cat category. Because it is not meaningful to compute the similarity between two contribution vectors towards different categories such as a cat and a non-cat object (e.g., pumpkin). More importantly, our goal is to study these images labeled as cats in the database, based on the contributions of semantic features to the classification of cats. (2) For the images predicted as cats, we do not distinguish different types of cats (e.g., tiger cat, Persian cat). The top cat category is used to generate the heatmap and bar chart. The reason is that there exists no fact label for each image for its cat type. Then it is difficult to study the differences among different cat types in CatViz. We consider them as a category cat in the following study.

Scenario 1: Browsing and Cluster Analysis

Alice loads the cat dataset into CatViz and starts her exploration by browsing the scatterplot (Fig. 9A). All 160 images are projected to the embedded space using t-SNE. The circles, squares, and triangles represent the images classified as cats (Cat), non-cat animals (NC), and non-animals (NA), respectively. She observes that the images are separated into several clusters, but their boundaries are vague. Images classified as non-cat animals and non-animals are scattered in most clusters. Looking at the statistics bar charts (Fig. 9A), she notices that Background has the highest average scores in semantic contributors, and it is the top contribution feature for a large number of images. She wants to focus on the anatomical features of cats in the scatterplot. Thus she increases the weights (W_k in Eqn. 6) of the Left Eye, Right Eye, Left Ear, Right Ear, Nose, and Body from the default value of 1 to 10 while keeping the weight of Background 1. As shown in Fig. 8D, after the weight adjustment, there are coherent clusters in the scatterplot with clear boundaries between most of them.

Using the lasso tool, Alice creates six selections (Fig. 8C), each for one cluster, and examines them one by one. From the statistics bar charts of these clusters (Fig 9B-G), she finds the images in the brown cluster generally indicate strong positive contributors on all the anatomical features (Fig. 9B). On the other hand, the images in the green cluster show negative contributions from the Right Eye (Fig. 9C), while other features give positive contributions. Alice is keen on comparing the two clusters, so she sends both to the detail panel for a comparative study (Fig. 8). The statistical bar charts in the comparison mode (Fig. 8F) reveal that the main difference between these two clusters is their scores on the Right Eye. Alice thus sorts the images by Right Eye scores from high to low. From the detail panel (Fig. 8I), she confirms that the brown cluster only contains images with positive scores on Right Eye while images in the green clusters all have high neg-

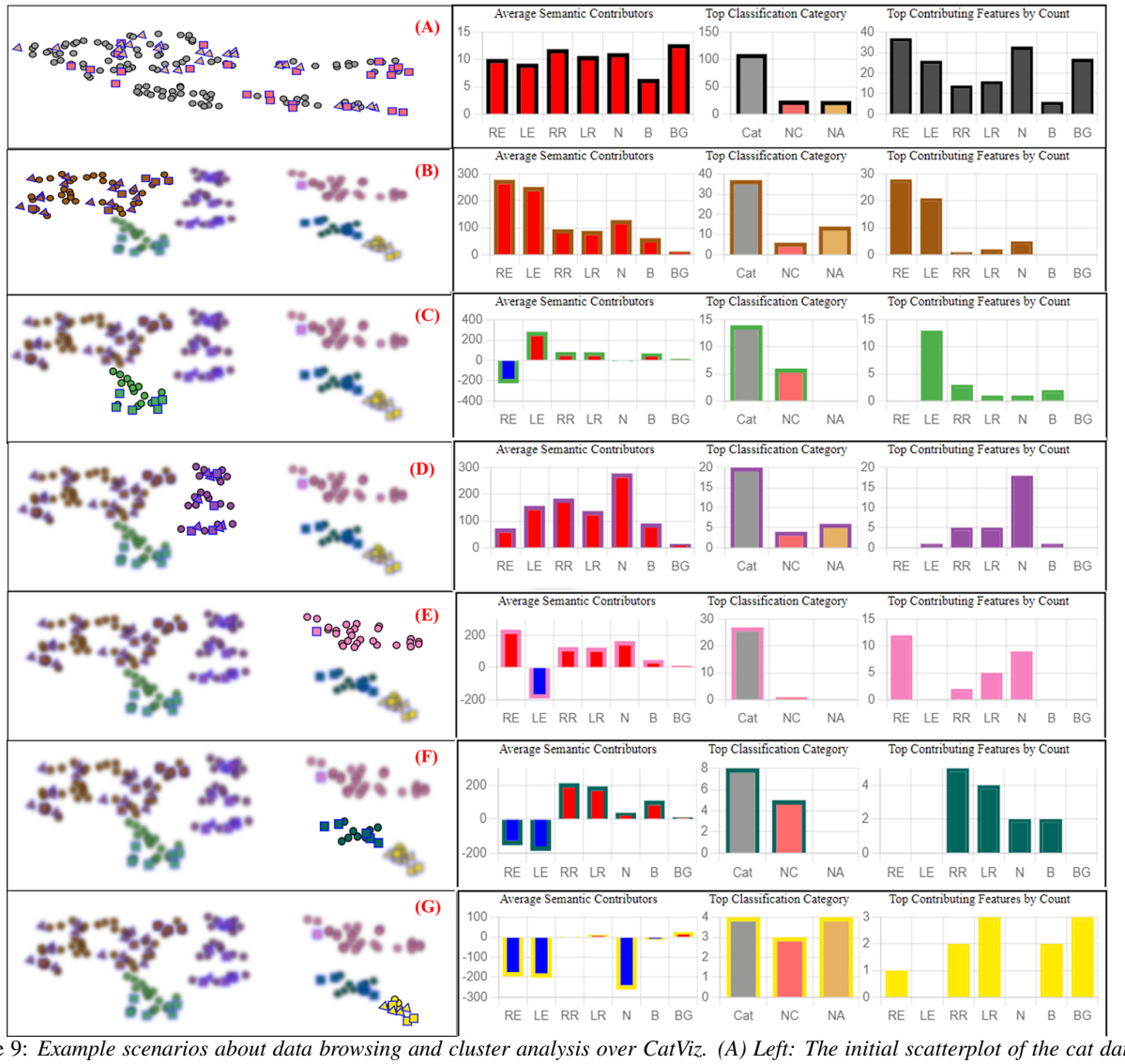


Figure 9: *Example scenarios about data browsing and cluster analysis over CatViz. (A) Left: The initial scatterplot of the cat dataset, where all features have the same weight. Right: The statistics bar charts for the whole dataset. (B)-(G) Left: The scatterplots with heavier weights on anatomical features. One cluster is highlighted in each figure. Right: The statistics bar charts for the cluster selected.*

ative scores on Right Eye.

Moving on, she proceeds to examine the rest of the clusters. The yellow cluster (Fig. 9G) looks interesting to her. Notably, she discovers that images within this cluster exhibit negative scores across all anatomical features but positive values in the Background. Furthermore, she observes that many images in this cluster are categorized as non-animal. Intrigued, she decides to delve deeper into the study of the Background feature at a later time. She also finds in the purple cluster (Fig. 9D), all features contribute positively, with Nose being the most significant contributor.

Next, Alice conducts further investigation by directly clicking images on the scatterplot at interesting positions to examine their details (Fig. 8E). This interactive approach aids in pinpointing outliers that lay far away from the cluster centers.

Through the above explorations, Alice gained knowledge that the classification output can be attributed to a variety of compositions of features. In other words, there exist latent subgroups of these cat images as inferred by the AI model, which may be employed in model refinement. This finding cannot be easily achieved without the proposed method and interface.

Scenario 2: Analyzing the Background Feature

To conduct a focused analysis of the Background feature, Alice resets the weights, sets all anatomical filters to the range of [0, 60] to exclude the influence of misclassified features, sends the resulting images to the detail panel, and sorts them by Background score from high to low. From the top of the detail panel (they are the same images as shown in Fig. 10A), where images exhibit high Background scores, she discovers intriguing images where the background prominently stands out, such as an image with a cat sitting on an office table. In these images, the cat's anatomical features are positive contributors, while the Background is the top contributor. Alice is particularly intrigued by such images, prompting her to adjust the Background filter by dragging the left slider towards the right. Interestingly, she observes that as

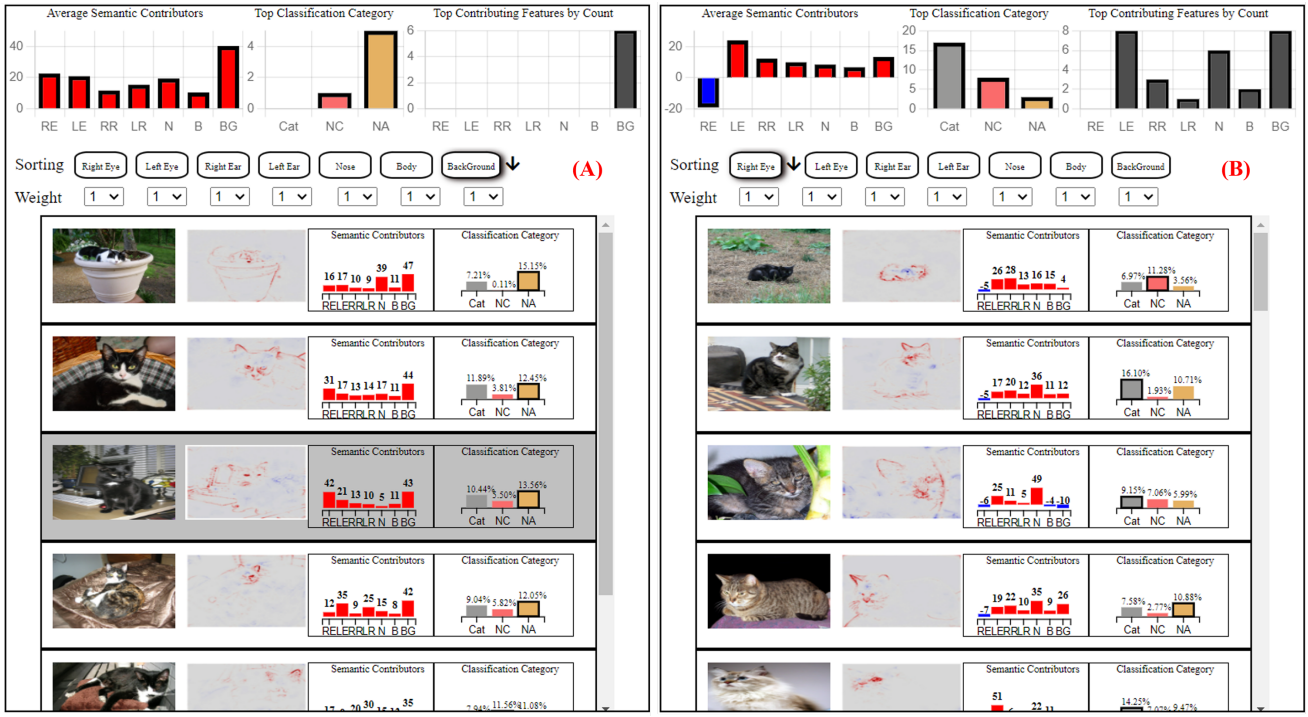


Figure 10: *Example scenarios about background feature analysis and image retrieval of CatViz. (A) Analyzing the Background feature. The images are retrieved using the Background filter (within the range [30, 60]) and the anatomical filters (within the range [0,60]). The statistics bar charts show that most images are classified as non-animal, and Background is the top contributing feature for all images. The images are sorted by Background scores from high to low, and those with dominant backgrounds stand out. (B) Retrieving images of cats with turned heads using the Left Eye filter (positive contribution within the range [0, 60]) and the Right Eye filter (negative contributions within the range [-60, 0]).*

the lower limit threshold of the filter approaches 30, the filtered images are consistently categorized as non-animal with a few exceptions of non-cat animals, as shown in the statistics bar charts in Fig. 10A. These images indicate an interesting finding: the AI model considers that the background objects also contribute a positive value to the classification of a cat. This is not an error since the heatmaps directly from LRP also show positive pixels in red on the backgrounds. However, it can be observed from the classification category bar chart that these images' top classification category is non-animal (NA) because these background objects have higher contribution scores to that category. Model designers may need to examine the corresponding neural network to further study the roles of background objects. CatViz allows them to easily find these interesting images for further investigation.

Scenario 3: Image Retrieval

Alice wonders if she can retrieve images where a cat's head is turned away from the camera. This task can be conducted by making use of the Left Eye and Right Eye contributors. Toward this goal, she sets the range of the Left Eye filter to [0, 60] and the range of the Right Eye filter to [-60, 0]. A positive contribution from the Left Eye and a negative contribution from Right Eye may indicate that the Right Eye is hidden or not clear while a cat turning its head. When Alice examines the results (shown in Fig. 10B), she finds that most of the images depict cats whose heads are turning away from the camera, which is what she desires. However, there are also images where the cats are facing the camera selected by the filters. She examines their heatmap images and finds their Right Eye regions are colored blue, indicating negative pixels. In addition, many of them are classified with a top category as non-cat animals, such as "Hamster". She realizes that the right eyes of the cats in these images are not detected as cat eyes by the AI model. They may be recognized as other animals' eyes. These unexpected outcomes are worth studying by the designers of the AI model. This example shows that the contribution scores of semantic contributors can help users quickly detect specific input images related to AI behaviors.

Benefits

This section complements the functional validation by presenting expert feedback on CatViz, focusing on its usability, interpretability, and practical value.

The example use scenarios demonstrate a set of benefits from its new visual analytical capabilities enabled by semantic contributors. First, users can perform analysis of the XAI output over many images intuitively and efficiently, which was not feasible with existing tools. Second, users can gain knowledge about semantic feature composition (or combination) leading to AI classification results, either correct or incorrect. This can be used to improve the AI model. Third, users can separate the images into subgroups with different classification behaviors, and also detect outlier images. This can help them further fine-tune a training process. Fourth, the visual exploration results can be quantitatively summarized to support decision-making and precisely conveyed to collaborators and stakeholders.

Expert Feedback and Discussion

We interviewed an AI domain scientist, who is a university professor teaching AI, NLP, and ML-related classes and has abundant knowledge and skills in image classification. At the beginning of our interview, we provided a thorough introduction to our approach, CatViz, and the example scenarios. During the interview, we discussed several topics regarding usability, impact, and potential improvements of CatViz. The expert provided insights based on his experience in the domain of image classification and shared valuable feedback on CatViz's features and functionalities.

Overall, the expert emphasized that CatViz offers a useful bridge between pixel-level explanations and higher-level semantic features, making it easier for both students and researchers to interpret model reasoning. He highlighted the strength of visualizing contributions at the semantic feature level, which provides more intuitive insights than traditional heatmaps. At the same time, he suggested that future work could broaden the evaluation to include diverse datasets and user groups to further strengthen its generalizability. This feedback complements the student user study by offering an expert perspective, thereby supporting the relevance and applicability of CatViz in both research and teaching contexts.

Feedback on Usability: The expert highlighted several positive aspects of CatViz, including its embedded view, ranking, filtering, and the study of clusters. He appreciated the system's ability to visualize image groups in an embedded space, enabling users to explore similarities and patterns effectively. He said, "The semantic features for the cat dataset seem spot on." In addition, he mentioned that the ranking and filtering functionalities provide flexibility in analyzing semantic features and refining selections. Furthermore, the expert found the study of clusters valuable for understanding feature contributions and model predictions. As he noted, "I particularly found valuable the ability to adjust thresholds of each of the semantic features to further analyze the data and understand how those features were contributing to the model's prediction." Here is the breakdown of the expert's opinion on individual functionalities of our system:

- *Embedded view*: Within the embedded space, the expert easily explored images for visual examination, enhancing his understanding of the structures and relationships within the datasets.
- *Ranking*: The expert appreciated that the ranking provided an opportunity to see patterns within the dataset and to understand how these relate to the semantic features affecting image classification outcomes.
- *Filtering*: The expert found the filtering feature especially beneficial for isolating images based on specific semantic features. This allowed him to focus on images with strong background features, which may predominantly classify as non-animal or non-cat.
- *Study of clusters*: The expert deemed the study of clusters feature valuable for identifying similarities among images and understanding feature contributions and model predictions.
- *Weight assignment*: This functionality allows weight assignment on semantic features. The expert commented that this feature could help users customize the analysis and find outliers.

Feedback on System Impact: The expert emphasized the importance of XAI techniques in improving intelligibility and interpretation within image categorization. He was optimistic about the CatViz's ability to reveal semantic contributors and their significance in image classification. He commented that through innovative visualization methods, CatViz offered a promising route toward analyzing image datasets and understanding the main factors affecting classification results. He commented that CatViz is a powerful tool that users can leverage to explore hidden patterns behind machine learning predictions.

Limitations and Suggestions: As for the limitations, the expert raised concerns about the adaptability of CatViz to datasets with less intuitive semantic features, such as medical data. He suggested considering how the system could discover semantic features for such datasets or datasets understood only by domain experts. Additionally, the expert questioned whether the visualization could accommodate more than 6-7 semantic features, highlighting the need for improvement in handling increasing dimensions and scales. Moreover, the expert emphasized the importance of a robust assessment process. He highlighted how crucial it is to put the system through extensive testing, particularly with the involvement of domain experts.

Conclusion and Future Work

This paper presented a novel approach to enhancing explainable AI (XAI) for image classification by quantifying the contributions of semantic features within input images. Unlike pixel- or patch-level methods, our feature-level approach computes contribution scores for predefined semantic features, enabling intuitive interpretation and visual exploration, as demonstrated in the CatViz prototype. Applied to the ImageNet cat dataset and the CUB-200 bird dataset, the method produced interpretable feature-level contribution patterns, suggesting potential portability across classes.

Future work will focus on several directions. First, integrating advanced semantic segmentation and user refinement tools will help define specialized semantic features. Second, we plan to evaluate the approach on more diverse datasets to assess generalizability and scalability. Third, improvements in computation and visualization will support higher-dimensional features and larger datasets. In domains such as medical imaging and industrial inspection, we will collaborate with experts to define and validate domain-specific semantic contributors. To address current limitations, we will explore semi-automated annotation protocols to reduce manual effort and investigate alternatives like UMAP for more stable embeddings. Finally, systematic quantitative validation, such as occlusion-based tests and ground-truth comparisons, will be performed to assess the alignment between feature contributions and model reasoning.

References

[1] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, sep 2023. doi: 10.1038/s42256-023-00711-8

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek. On pixel-wise explanations for non-

linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), 2015. doi: 10.1371/journal.pone.0130140

[3] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, pp. 3319–3327, 2017.

[4] D. Bau, J. Zhu, H. Strobelt, À. Lapedriza, B. Zhou, and A. Torralba. Understanding the role of individual units in a deep neural network. *PNAS*, 117:30071–30078, 2020.

[5] D. Bhati, M. Amiruzzaman, Y. Zhao, A. Guercio, and T. Le. A survey of post-hoc XAI methods from a visualization perspective: Challenges and opportunities. *IEEE Access*, 2025.

[6] W. Brendel and M. Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[7] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[8] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2018.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[10] A. Ghorbani, J. Wexler, J. Zou, and B. Kim. Towards automatic concept-based explanations. *33rd Conference on Neural Information Processing Systems*, 2019.

[11] J. Huang, A. Mishra, B. C. Kwon, and C. Bryan. Conceptexplainer: Interactive explanation for deep neural networks from a concept perspective. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):831–841, 2023. doi: 10.1109/TVCG.2022.3209384

[12] X. Huang, W. Xu, S. Jamonnak, Y. Zhao, and T. H. Wu. VisLRP: A visual designer of layer-wise relevance propagation models. *Computer Graphics Forum*, 3(40):227–238, 6 2021. doi: doi.org/10.1111/cgf.14302

[13] B. K. Iwana, R. Kuroki, and S. Uchida. Explaining Convolutional Neural Networks using Softmax Gradient Layer-wise Relevance Propagation. *ICCV 2019 XAIC Workshop arXiv:1908:04351*, 2019.

[14] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

[15] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. TCAV: Relative concept importance testing with Linear Concept Activation Vectors. *ICLR*, 2018.

[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W. Lo, et al. Segment anything (arxiv: 2304.02643). arxiv, 2023.

[17] H. Li, Y. Lin, K. Mueller, and W. Xu. Interpreting Galaxy Deblender GAN from the Discriminator's Perspective. *arXiv:2001.06151*, 2020.

[18] M. Lin, Q. Chen, and S. Yan. Network In Network. *International Conference on Learning Representations*, 2014.

[19] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. Stumpf. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024. doi: 10.1016/j.inffus.2024.102301

[20] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 30, pp. 4768–4777, 2017.

[21] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, pp. 5188–5196, 2015.

[22] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[23] E. Molina, C. Middel, and P. Vazquez. Effect of color palettes in heatmaps perception: a study. *25th EuroGraphics Conference on Visualization*, 2023.

[24] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. R. Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. doi: 10.1016/j.patcog.2016.11.008

[25] G. Montavon, W. Samek, and K. R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73:1–15, 2018. doi: 10.1016/j.dsp.2017.10.011

[26] M. Nauta, R. Van Bree, and C. Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14933–14943, 2021.

[27] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021. doi: 10.1109/JPROC.2021.3060483

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

[29] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, 2014.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. *International Conference on Learning Representations*, 2015.

[32] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328, 2017.

[33] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan,

I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *The 2nd International Conference on Learning Representations*, 2014.

[34] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

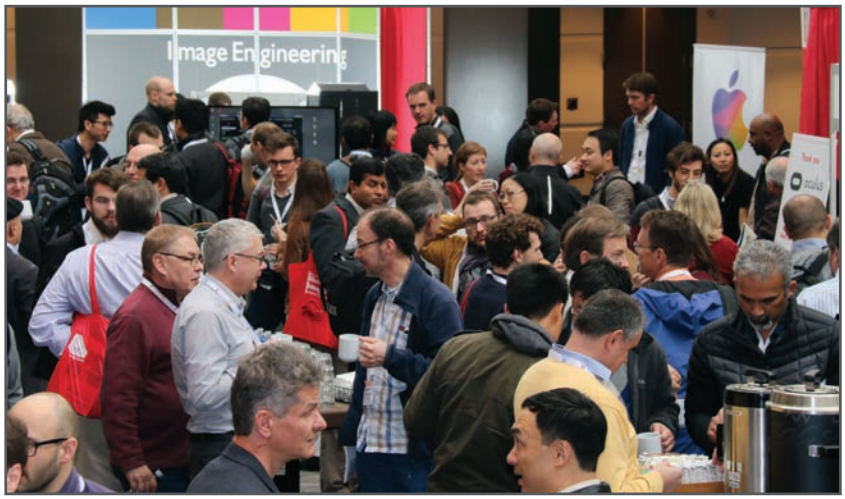
[35] Y. Wang, U. Ahsan, H. Li, and M. Hagen. A comprehensive review of modern object segmentation approaches. *Foundations and Trends in Computer Graphics and Vision*, 13(2–3):111–283, 2023. doi: 10.1561/0600000097

[36] Z. Zhao, P. Xu, C. Scheidegger, and L. Ren. Human-in-the-loop extraction of interpretable concepts in deep learning models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):780–790, 2021.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org