

This Looks Like That: Deep Learning for Interpretable Image Recognition

Chaofan Chen^{1*}

cfchen@cs.duke.edu

Oscar Li^{1*}

runliang.li@duke.edu

Chaofan Tao¹

chaofan.tao@duke.edu

Alina Jade Barnett¹

abarnett@cs.duke.edu

Jonathan Su³

su@ll.mit.edu

Cynthia Rudin^{1,2}

cynthia@cs.duke.edu

¹Department of Computer Science, Duke University, Durham, NC, USA 27708

²Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA 27708

³MIT Lincoln Laboratory, Lexington, MA 02421-6426[†]

Abstract

When we are faced with challenging image classification tasks, we often explain our reasoning by dissecting the image, and pointing out prototypical aspects of one class or another. The mounting evidence for each of the classes helps us make our final decision. In this work, we introduce a deep network architecture that reasons in a similar way: the network dissects the image by finding prototypical parts, and combines evidence from the prototypes to make a final classification. The model thus reasons in a way that is qualitatively similar to the way ornithologists, physicians, geologists, architects, and others would explain to people on how to solve challenging image classification tasks. The network uses only image-level labels for training, meaning that there are no labels for parts of images. We demonstrate our method on the CUB-200-2011 dataset and the CBIS-DDSM dataset. Our experiments show that our interpretable network can achieve comparable accuracy with its analogous standard non-interpretable counterpart as well as other interpretable deep models.

1. Introduction

How would you describe why the image in Figure 1 looks like a clay colored sparrow? Perhaps the bird’s head looks like that of a prototypical clay colored sparrow, even though its belly might look like that of either a grasshopper

*Contributed equally

[†]DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

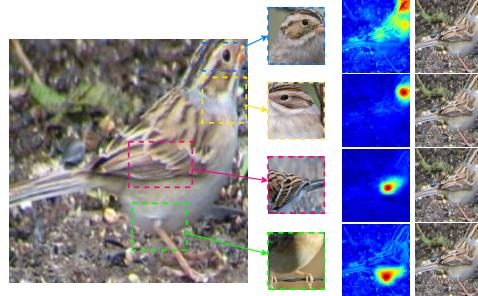


Figure 1. Image of a clay colored sparrow and the learned prototypical parts of a clay colored sparrow used to classify the bird’s species. The smaller images in the second column are the prototypical parts of a clay colored sparrow learned by our algorithm. Our model compares these prototypical parts with a test image (leftmost), and generates heat maps (in the third column) that indicate how similar each prototypical part resembles part of the test bird. The original test image is replicated (in the fourth column) beside each heat map for easy reference of which part of the original image is activated by each prototype.

sparrow or a clay colored sparrow. When we describe how we classify images, we might focus on parts of the image and compare them with prototypical parts of images from a given class. This method of reasoning is commonly used in difficult identification tasks: for example, radiologists compare suspected tumors in X-ray scans with prototypical tumor images for diagnosis of cancer. The question is whether we can ask a machine learning algorithm to imitate this way of thinking, in order to explain its reasoning process to humans in a way that they can understand.

The goal of this work is to define a form of interpretability in image processing (*this looks like that*) that agrees with the way humans describe their own thinking in classification tasks. In this work, we introduce a network architecture that accommodates this definition of interpretability, where the

comparison of image parts to learned prototypes is integral to the way our network reasons about new examples – our learning algorithm selects, from the training set, a limited number of prototypical parts that are useful in classifying a new image, and learns an internal notion of distance for comparing parts of the new image to those learned prototypes. Given a new bird image as in Figure 1, our model is able to identify several parts of the image where it thinks that *this* identified part of the image looks like *that* prototypical part of some training image, and makes its prediction based on a weighted combination of the similarity scores between parts of the image and the learned prototypes. In this way, our model is *interpretable*, in the sense that it has a transparent reasoning process that is actually used to make predictions. Our experiments show that interpretability can be gained without losing much accuracy: our interpretable network can achieve comparable accuracy with its analogous non-interpretable counterpart and other interpretable deep models, on datasets such as CUB-200-2011 [36].

1.1. Related Work

Our work relates to (but contrasts with) those that perform *posthoc* interpretability analysis for trained networks. In posthoc analysis, one interprets a trained network by fitting explanations to how the network performs classification. There are two general approaches to understanding networks posthoc: one is class-specific activation maximization [8, 12, 15, 35, 23, 30, 39], and the other is input-specific posthoc visualization such as deconvolution [40] and gradient-based saliency visualization [30, 33, 32, 27]. All of these posthoc visualization methods do not explain the reasoning process of how a network *actually* makes its decisions. In contrast, our network has a built-in case-based reasoning process, and the explanations generated by our network are actually used during classification and are not created posthoc.

Our work relates closely to works that build interpretability into deep neural networks. Attention mechanisms that identify the most relevant parts of an input for various tasks have been integrated into neural networks: various methods have been proposed to jointly train networks with integrated class-specific attention maps [24, 43]. There are also works that not only identify the important parts but also make use of them directly for classification: these works usually “single” out the important parts and use only these parts in the downstream reasoning process. They either use heavy supervision to locate the most relevant parts for classification (e.g. [41, 13, 44]), or rely on an auxiliary (pre-trained) network to extract image patches for unsupervised identification of important parts (e.g. [29, 38]), or propose a number of candidate parts using selective search-based region proposal network [34, 11, 10, 26] or Monte Carlo sampling [18]. However, none of these works

learn prototypical cases for comparison and prediction as we do in our work.

Recently there have also been attempts to quantify the interpretability of visual representations learned by a convolutional neural network (CNN). Bau et al. proposed the network dissection framework that uses the overlap between the receptive field of top activations and regions corresponding to labeled visual concepts as a measure of the interpretability of the convolutional unit [2]. Zhang et al. used this measure of interpretability and proposed architectural modifications to traditional CNNs [42]. These are useful, but to quantitatively measure the interpretability of a convolutional unit in a network requires fine-grained labeling for a significantly large dataset specific to the purpose of the network. While Bau et al. have built the Broden dataset for scene/object classification networks [2], this dataset is not well-suited to measure the unit interpretability of a network trained for fine-grained classification or medical applications, because the concepts detected by that network may not be present in the Broden dataset. Hence, in our work, we do not focus on quantifying unit interpretability of our network, but instead look at the *reasoning process* of our network which is qualitatively similar to that of humans. We do not aim to compare everything identified in the image to a known, labeled, visual concept; instead, we aim to pinpoint parts of the image that are important and similar to prototypical parts of images from a class.

Our network architecture includes a prototype layer that replaces the conventional inner product with the squared L^2 distance computation. This is not new [9], but in our work we require the filters to be *identical* to the latent representation of some training image patch. This added constraint allows us to interpret the filters as prototypical parts of images from different classes, and also necessitates a more specialized training procedure for our network. We implement the squared L^2 distance computation using the conventional inner product convolution, as described in [22].

Our work also relates closely to other prototype classification techniques in machine learning [25, 3, 14, 5]. It relates closely to Branson et al. (2014) [4], who used a greedy facility location algorithm to identify a set of prototypes for use in bird species identification. However, their prototypes are whole images in the pixel space and do not involve *parts* of images. Our work relates most closely to Li et al. [19], who proposed a network architecture that builds case-based reasoning into a neural network. However, their model requires a decoder for visualizing prototypes, and when trained on datasets of natural images such as CUB-200-2011, the decoder fails to produce realistically looking prototype images. In contrast, our model does not require a decoder for prototype visualization. It projects (“pushes”) the latent representations of prototypes onto the closest latent representations of training image patches, and

uses those training image patches for prototype visualization. The removal of the decoder also facilitates the training of our network, leading to better explanations and better accuracy. Unlike Li et al., whose model requires the prototypes to have exactly the same shape as the latent representations of images, the prototypes in our model can have much smaller spatial dimensions than the latent representations of images in general, which means that *our prototypes are prototypical parts of images*. This allows for more fine-grained comparisons because different parts of an image can now be compared to different prototypes. Moreover, in our work, we also associate each prototype with a class, and use a different training objective (the clustering cost and the separation cost are both new to our work) and a more elaborate training scheme than Li et al. to cluster image patches of a particular class around the prototypes of the same class, while separating image patches of different classes. The result is a more meaningful latent space for comparison with prototypical parts, which also leads to improved explanations and improved accuracy over Li et al.

2. Case Study 1: Bird Species Identification

In this case study, we will introduce the architecture and the training procedure of our interpretable network in the context of bird species identification, and provide a detailed walk-through of how our network classifies a new bird image and explains its prediction. We trained and evaluated our interpretable network using the CUB-200-2011 dataset [36] of color images of 200 bird species. Since the dataset has only about 30 images per class, we performed offline data augmentation using random rotation, skew, shear, distortion, and left-right flip to enlarge the training set, so that each class has approximately 1200 training images. We trained our network on both full images and images cropped using the bounding box annotations provided with the dataset.

2.1. Network Architecture

Figure 2 gives an overview of the architecture of our interpretable network. Our network consists of a regular convolutional neural network f , whose filters and biases are collectively denoted by α_{conv} , followed by a prototype layer g_p and a fully connected layer h with weight matrix α_h and no bias. For the regular convolutional network f , we use the first 13 convolutional layers (with max-pooling) of the VGG-16 network [31] pretrained on ImageNet [7], followed by two additional 1×1 convolutional layers in our experiments. Given an input image x (such as the clay colored sparrow in Figure 2), the convolutional layers of our model extract useful features $f(x)$ for prediction. Let $H \times W \times D$ be the shape of the convolutional output $f(x)$. The network learns m prototypes $P = \{\mathbf{p}_j\}_{j=1}^m$, whose shape is $H_1 \times W_1 \times D$ with $H_1 \leq H$ and $W_1 \leq W$. Since the

depth of each prototype is the same as that of the convolutional output but the height and the width of each prototype is less than or equal to those of the convolutional output, each prototype will be used to represent some prototypical activation pattern in a *patch* of the convolutional output, which in turn corresponds to some prototypical image patch in the original pixel space. Hence, each prototype \mathbf{p}_j can be understood as the latent representation of some prototypical *part* of some bird image in this case study. As a schematic illustration, the first prototype \mathbf{p}_1 in Figure 2 corresponds to the head of a clay colored sparrow, and the second prototype \mathbf{p}_2 the head of a field sparrow. Given a convolutional output $\mathbf{z} = f(\mathbf{x})$, the j -th prototype unit $g_{\mathbf{p}_j}$ in the prototype layer g_p computes the squared L^2 distances between the j -th prototype \mathbf{p}_j and all patches of \mathbf{z} that have the same shape as \mathbf{p}_j , and inverts the distances into similarity scores. The result is an activation map of similarity scores whose value indicates how strong a prototypical part is present in the image. This activation map preserves the spatial relation of the convolutional output (e.g. the upper-left value in the activation map is the similarity score between the upper-left patch of \mathbf{z} and the prototype), and can be upsampled to the size of the input image to produce a heat map that identifies which part of the input image is most similar to the learned prototype. The activation map of similarity scores produced by each prototype unit $g_{\mathbf{p}_j}$ is then reduced using global max pooling to a single similarity score, which can be understood as how strongly a prototypical part is present in some patch of the input image. In Figure 2, the similarity score between the first prototype \mathbf{p}_1 , which corresponds to the head of a clay colored sparrow, and the most activated (upper-right) patch of the input image of a clay colored sparrow is 1.993, and the similarity score between the second prototype \mathbf{p}_2 , which corresponds to the head of a field sparrow, and the most activated patch of the input image is 0.697. This shows that our model finds that the head of a clay colored sparrow has a stronger presence than that of a field sparrow in the input image. Mathematically, the prototype unit $g_{\mathbf{p}_j}$ computes the following:

$$g_{\mathbf{p}_j}(\mathbf{z}) = \max_{\tilde{\mathbf{z}} \in \text{patches}(\mathbf{z})} \log \left(1 + 1 / (\|\tilde{\mathbf{z}} - \mathbf{p}_j\|_2^2 + \epsilon) \right).$$

Hence, if the output of the j -th prototype unit $g_{\mathbf{p}_j}$ is large, then there is a patch in the convolutional output that is very close to the j -th prototype in the latent space, and this in turn means that there is a patch in the input image that has a similar concept to what the j -th prototype represents.

Finally, the m similarity scores produced by the prototype layer g_p are multiplied by the weight matrix α_h in the fully connected layer h to produce the output logits for classification.

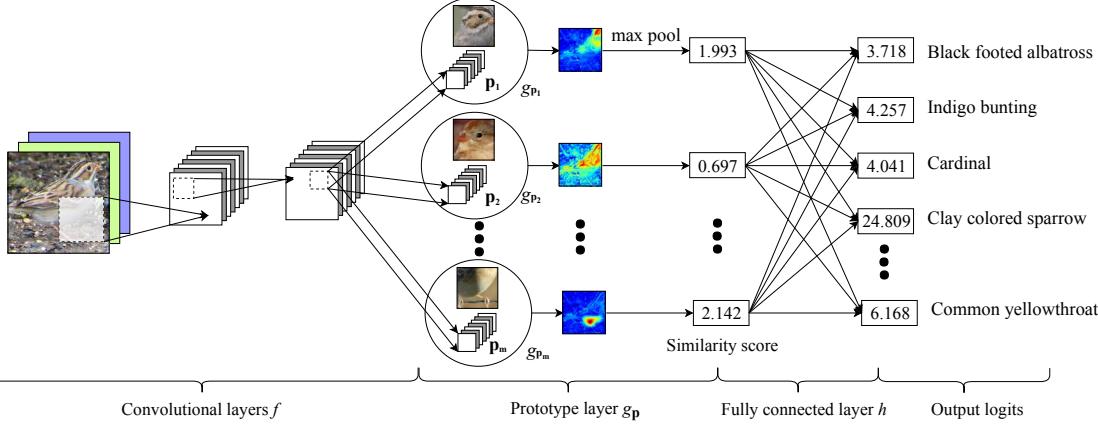


Figure 2. The network architecture.

2.2. Training Algorithm

The training procedure of our interpretable network is divided into three stages: (1) stochastic gradient descent of the layers before the fully connected last layer h ; (2) projection of the prototypes \mathbf{p}_j onto the closest latent representations of training image patches from the same class as that of \mathbf{p}_j ; (3) convex optimization of the last layer h .

2.2.1 Stochastic Gradient Descent of Layers before the Last Layer

In the first training stage, we aim to learn a meaningful latent space where the most important patches for classifying images are clustered around prototypes associated with their own classes, and those important patches from different classes will be separated into distinct clusters. To achieve this goal, we jointly optimize the convolutional layers f , and the prototype layer g_p , which stores a set of prototypes $\mathbf{P} = \{\mathbf{p}_j\}_{j=1}^m$, using stochastic gradient descent, while keeping the fully connected last layer h fixed. Let $D = [\mathbf{X}, \mathbf{Y}] = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training set of images, with labels $y_i \in \{1, \dots, K\}$ for $i \in \{1, \dots, n\}$. We allocate a pre-determined number of prototypes for each class, so that every class will be represented by some prototypes in the final model, and no class will be left out. Let $\mathbf{P}_k \subseteq \mathbf{P}$ be the subset of prototypes that are allocated to class k : these prototypes should capture the most relevant parts or semantic concepts for identifying images of class k .

The optimization problem we aim to solve in this training stage is:

$$\begin{aligned} \min_{\mathbf{P}, \alpha_{\text{conv}}} & \text{CrossEntropy}(h \circ g_p \circ f(\mathbf{X}), \mathbf{Y}) \\ & + \lambda_1 \text{Clst}(\mathbf{P}, \mathbf{X}, \mathbf{Y}) + \lambda_2 \text{Sep}(\mathbf{P}, \mathbf{X}, \mathbf{Y}). \end{aligned}$$

The cross entropy loss penalizes misclassification on the training data and encourages accuracy. The cluster cost

(Clst) and the separation cost (Sep) are defined by

$$\begin{aligned} \text{Clst}(\mathbf{P}, \mathbf{X}, \mathbf{Y}) &= \frac{1}{n} \sum_{i=1}^n \min_{j: \mathbf{p}_j \in \mathbf{P}_{y_i}} \min_{\tilde{\mathbf{z}} \in \text{patches}(f(\mathbf{x}_i))} \|\tilde{\mathbf{z}} - \mathbf{p}_j\|_2^2 \\ \text{Sep}(\mathbf{P}, \mathbf{X}, \mathbf{Y}) &= -\frac{1}{n} \sum_{i=1}^n \min_{j: \mathbf{p}_j \notin \mathbf{P}_{y_i}} \min_{\tilde{\mathbf{z}} \in \text{patches}(f(\mathbf{x}_i))} \|\tilde{\mathbf{z}} - \mathbf{p}_j\|_2^2. \end{aligned}$$

For a given training image of class k , the cluster cost penalizes *large* distance between the closest pair of patches and class k prototypes, and the separation cost penalizes *small* distance between the closest pair of patches and non-class k prototypes. Hence, the minimization of the cluster cost encourages each training image to have some patch whose latent representation is close to at least one prototype associated with its own class, while the minimization of the separation cost encourages every patch of a training image to stay away from the prototypes *not* associated with its own class in the latent space. These terms shape the latent space into a clustering structure, which facilitates the L^2 distance based classification of our network.

In this training stage, we also fix the fully connected last layer h with weight matrix α_h as follows. Let $\alpha_h^{(k,j)}$ be the (k, j) -th entry in α_h that corresponds to the weight connection between the output of the j -th prototype unit $g_{\mathbf{p}_j}$ and the logit of class k . Given a class k , we set $\alpha_h^{(k,j)} = 1$ for all j with $\mathbf{p}_j \in \mathbf{P}_k$ and $\alpha_h^{(k,j)} = -0.5$ for all j with $\mathbf{p}_j \notin \mathbf{P}_k$. Intuitively, the positive connection between a class k prototype and the class k logit means that similarity to a class k prototype increases the predicted probability that the image belongs to class k , and the negative connection between a non-class k prototype and the class k logit means that similarity to a non-class k prototype decreases the predicted probability that the image belongs to class k . By fixing the fully connected last layer h in this way, we can force the network to learn a meaningful latent space because if a patch of a class k image is too close to a non-class

k prototype in the latent space, it will decrease the predicted probability that the image belongs to class k and, as a result, increase the cross entropy loss in the training objective.

Note that both the separation cost and the negative connection between a non-class k prototype and the class k logit encourage prototypes of class k to represent semantic concepts that are characteristic of class k but not of other classes: if a class k prototype represents a semantic concept that is also present in a non-class k image, this non-class k image will highly activate that class k prototype , and this will be penalized by increased (i.e. less negative) separation cost and increased cross entropy (as a result of the negative connection). The separation cost is new to this paper, and it has not been explored by previous works that involve prototype learning (cf. [4, 19]).

2.2.2 Projection of Prototypes

In order to visualize the prototypes as training image patches, we project (“push”) each prototype \mathbf{p}_j onto the latent representation of some training image patch after the first training stage is complete. In this way we can conceptually equate each prototype with a training image patch, and visualize it by finding that training image patch in the original pixel space. Note that we cannot simply push each prototype \mathbf{p}_j onto the closest latent representation of training image patches from any class, because we are associating each prototype with some prototypical part of a particular class. Therefore, we push each prototype \mathbf{p}_j onto the closest latent representation of training image patches from the *same* class as that of \mathbf{p}_j . Mathematically, this projection stage amounts to setting each prototype \mathbf{p}_j to $\arg \min_{\tilde{\mathbf{z}} \in \mathcal{Z}_j} \|\tilde{\mathbf{z}} - \mathbf{p}_j\|_2^2$, where \mathcal{Z}_j is the set of training image patches from class j with $\mathbf{p}_j \in \mathbf{P}_k$ in the latent space:

$$\mathcal{Z}_j = \{\tilde{\mathbf{z}} : \tilde{\mathbf{z}} \in \text{patches}(f(\mathbf{x}_i)) \text{ for all } i \text{ with } y_i = k, \text{ where } k \text{ satisfies } \mathbf{p}_j \in \mathbf{P}_k\}.$$

2.2.3 Convex Optimization of Last Layer

In the last training stage, we perform a convex optimization on the weight matrix α_h of the fully connected last layer h . The goal of this stage is to adjust the last layer connection $\alpha_h^{(k,j)}$ between the output of the j -th prototype unit $g_{\mathbf{p}_j}$ and the logit of class k , so that for all k and for all j with $\mathbf{p}_j \notin \mathbf{P}_k$, our final model has the sparsity property $\alpha_h^{(k,j)} \approx 0$ (remember that this was fixed at -0.5 in the previous stages). This sparsity property is desirable because having this property means that our model relies less on a *negative* reasoning process of the form “this bird is of class k' because it is *not* of class k (it contains a patch that is *not* prototypical of class k),” to reach its final prediction. Mathematically, the optimization problem we aim to solve

in this training stage is as follows:

$$\min_{\alpha_h} \text{CrossEntropy}(h \circ g_{\mathbf{p}} \circ f(\mathbf{X}), \mathbf{Y}) + \lambda \sum_{k=1}^K \sum_{j: \mathbf{p}_j \notin \mathbf{P}_k} |\alpha_h^{(k,j)}|.$$

This optimization is convex because we fix all the parameters from the convolutional and prototype layers. This also further improves accuracy after prototype projection without changing the learned latent space and prototypes.

2.3 Accuracy

The accuracy of our interpretable network is compared to that of the baseline VGG-16 model and previous works on interpretable image classification models in Table 1.

Table 1. Accuracy comparison

Model	Acc.	Acc. (on bb)	Part super- vision
Our network	71.4	74.5	Not needed
VGG-16 *	70.4	76.4	NA
Part R-CNN [41]	73.9	76.4	Required
Part-stack.CNN [13]	-	76.6	Required
Pose-norm. CNN [4]	75.7	-	Required
GoogLeNet-GAP [43]	63.0	70.5	Not needed

* The accuracy figures for VGG-16 are provided by [20].

The “Acc.” column in Table 1 gives the accuracy of the respective model trained without using bounding box annotations of the dataset, and the “Acc. (on bb)” column gives the accuracy of the respective model trained on images cropped using the bounding box annotations. As we can see, the accuracy of our interpretable network is comparable with that of the baseline (non-interpretable) VGG-16 model: there is no loss in accuracy when trained without using bounding box annotations, and the loss in accuracy is less than 2% when trained on cropped images. More strikingly, when compared against other interpretable image classification models that require supervised learning of parts [41, 13, 4], our network can still achieve comparable accuracy (and find meaningful parts in most cases) *without heavy supervision to locate the most important parts for classification during model training*. This is desirable because a lot of real-world datasets do not come with annotated parts, and it is expensive to construct such fine-grained annotations. For some applications, fine-grained annotations could also be incomplete or error-prone. When compared against a non-posthoc attention-based model that also requires no supervised learning of parts, such as the GoogLeNet-GAP model with built-in class activation maps [43]), our network can achieve better accuracy on this dataset. We do not compare the accuracy of our interpretable network with other non-interpretable methods of more complex architecture (e.g. [20]), which can achieve better accuracy on this dataset.

The accuracy figures of our interpretable network are obtained from training our network on images of size 224×224 , using 2000 prototypes (10 per class) of shape $1 \times 1 \times 512$. We also trained our network on cropped bird images of size 336×336 , using 2400 prototypes (12 per class) of the same shape: the accuracy of this model is 74.0%. Training our network using a larger image size has some added advantages: (1) the image patch (receptive field) corresponding to each prototype is relatively smaller compared to the original image, so direct visualization of the image patch corresponding to each prototype is often enough to see what structural part each prototype is looking for; (2) the map of similarity scores produced by each prototype unit has a larger spatial dimension, and is more likely to generate a smoother activation map for the prototype after upsampling. In the following sections, we use the model we trained on 336×336 cropped bird images for analysis.

2.4. Reasoning Process of Our Network

Figure 3 shows the reasoning process of our interpretable network in reaching a classification decision on a test image of a Lincoln sparrow at the top of the figure. Given this test image \mathbf{x} , our model first extracts useful features using the convolutional layers f . The extracted features $f(\mathbf{x})$ are then compared against the learned prototypes \mathbf{p}_j . In particular, for each class k , our network asks the question – Is the given image a bird of class k ? The network tries to find evidence for the given image being a bird of class k by comparing patches of the given image with each learned prototype \mathbf{p}_j of class k (i.e. $\mathbf{p}_j \in \mathbf{P}_k$) in the latent space. For example, in Figure 3 (left), the network tries to find evidence for the given image being a Lincoln sparrow, by comparing its patches with each learned prototype of the Lincoln sparrow class in the latent space: the top three activated prototypes of the Lincoln sparrow class are visualized in the “Prototype” column. This comparison produces a map of similarity scores between patches of the given image and each learned prototype, which can be upsampled and superimposed on the original image to see which part of the given image is activated by each prototype. As shown in the “Activation Map” column in Figure 3 (left), the first prototype of the Lincoln sparrow class activates on the wing/body of the given bird, and the second prototype on the head of the given bird. This means that the network finds a high similarity between the wing/body of the given bird and the *prototypical* wing/body of a Lincoln sparrow, and a high similarity between the head of the given bird and the *prototypical* head of a Lincoln sparrow. The network then uses the maximum similarity score between patches of the given image and each prototype as the similarity score between the given image and each prototype. Thus, we can interpret the similarity score between the given image and a prototype as a measure of how much the concept represented by

the prototype is present in *some* part of the given image. In Figure 3 (left), the prototypical wing/body of a Lincoln sparrow is present in some part of the given image with a similarity score of 3.758, and the prototypical head of a Lincoln sparrow is present in some part of the given image with a similarity score of 2.282. These similarity scores are then weighted by the last layer connection between the prototypes and the class we are considering (Lincoln sparrow), to obtain the points contributed by the prototypes to the given image being classified as a Lincoln sparrow. In Figure 3 (left), the first prototype contributes 3.938 points, and the second 3.101 points, and so on, to the given image being classified as a Lincoln sparrow. The reasoning process is similar when the network tries to find evidence for the given image being a bird of some other class: an example is shown in Figure 3 (right). In Figure 3, the total points contributed by the prototypes to the given bird being classified as a Lincoln sparrow is 22.749, and the total points contributed by the prototypes to the given bird being classified as a Henslow sparrow is 14.891. These total points are precisely the output logits of the Lincoln sparrow class and the Henslow sparrow class. Finally, the network decides that the given image must be the class of the highest total points, i.e. Lincoln sparrow. More examples of how our network classifies can be found in the supplement.

We want to emphasize that the prototype activation maps produced by our model are fundamentally different from the class activation maps in [43]. Our prototype activation maps are associated with prototypes: they are visualization of which parts in the original image are similar to particular prototypes. The class activation maps are associated with classes: they are visualization of which parts in the original image are important for it being classified as particular classes. Also, our prototype activation maps are produced by the squared L^2 distance computation, whereas the class activation maps are produced by a weighted combination of the conventional inner product.

2.5. Analysis of Latent Space

In this section, we analyze the structure of the latent space learned by our interpretable network. We demonstrate that the nearest prototypes of a given image are mostly prototypes associated with the class of the image, and the nearest patches of a given prototype mostly come from those images in the same class as that of the prototype. Moreover, we show that similar parts are consistently highlighted when the original images containing the nearest patches of a given prototype are passed through the network to generate the activation maps.

Figure 4 shows the three nearest prototypes to a test image of a ringed kingfisher and also to a test image of a Cerulean warbler. For a given image, we define its nearest prototype as the one that forms the closest patch-prototype

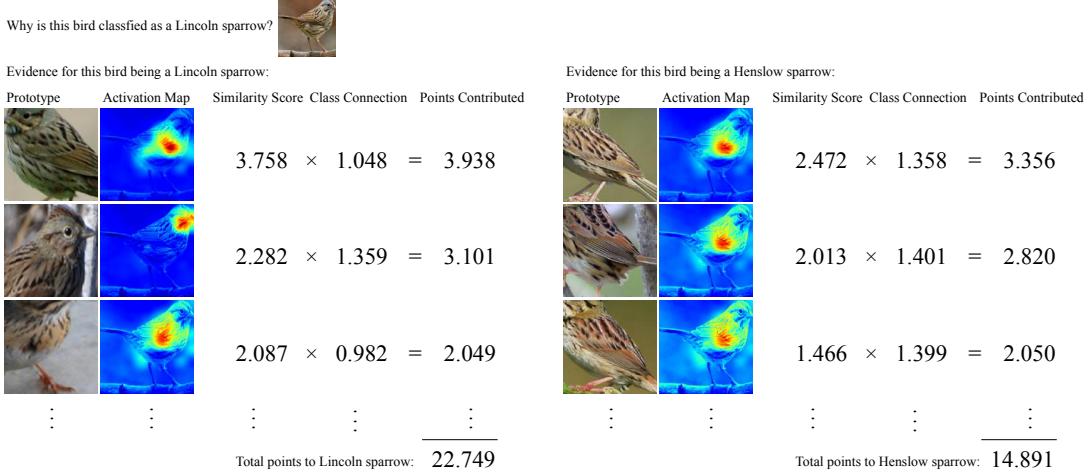


Figure 3. The reasoning process of our network in deciding the species of a bird (top).

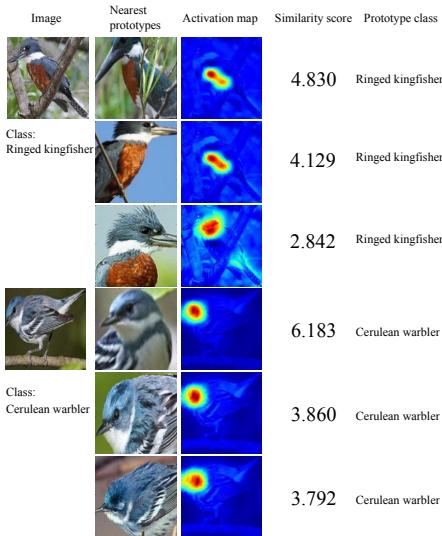


Figure 4. Nearest prototypes of two test images.

pair in the latent space, over all patches of the given image. As we can see from Figure 4, the nearest prototypes for each of the two test images come from the same class as that of the image. The activation map beside each prototype provides clue for which patch of the original image is closest to the prototype in the latent space: in the top example (ringed kingfisher) of Figure 4, the patch around the chest of the bird in the test image must be close to the top two prototypes in the latent space, and the patch around the neck of the bird must be close to the third prototype in the latent space, because those regions are the most activated in the respective activation maps. This shows that the latent space learned by our interpretable network does have a clustering structure, where the most relevant patches for classification are close to some prototypes of the same class.

Figure 5 shows the nearest image patches (in the latent

space) of two prototypes from both the training and the test set. As we can see, the nearest image patches to the first prototype in the figure all contain the white chest of a parakeet auklet, and the nearest image patches to the second prototype all contain the body of a lazuli bunting. To further evaluate which part or semantic concept each prototype is detecting, we generated an activation map for each image patch on each prototype as follows – For each image patch, we passed the original image containing the patch through the network to generate a map of similarity scores between the patches and the prototype of interest. We then upsampled the map of similarity scores to the size of the original image to generate the activation map, and cropped the activation map at the same location as the image patch. The result is the activation map for the prototype of interest on the image patch. This activation map gives us an idea of what each prototype is looking for: the activation maps for the first prototype in Figure 5 on its nearest image patches all highlight the white chest of a parakeet auklet, and the activation maps for the second prototype on its nearest image patches all highlight the body of a lazuli bunting. This demonstrates that our interpretable network is able to learn a meaningful latent space where similar parts or semantic concepts are clustered together, with only weak supervision.

The supplementary material contains more examples of the nearest prototypes of given images and the nearest image patches of prototypes. It also includes a t-SNE visualization [21] of the latent space learned by our network.

3. Case Study 2: Breast Cancer Detection

In this section, we will explore the possibility of applying our interpretable network to high-stakes decision making, where interpretability is the key to whether we can trust the predictions made by a machine learning model. We will use breast cancer detection as our high-stakes application

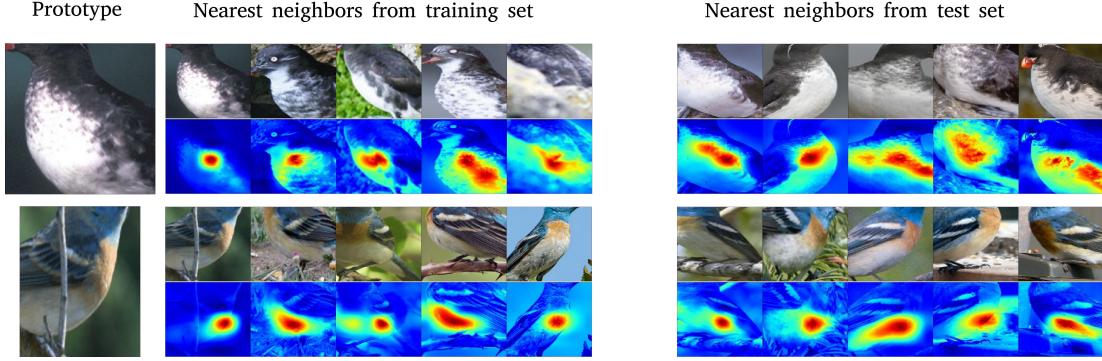


Figure 5. Nearest image patches (in the latent space) of two prototypes from both the training and the test set. The activation map below each image patch shed lights on which part or semantic concept the prototype is most likely detecting: it is obtained by passing the original image containing the patch through the network to generate a map of similarity scores, which is then upsampled to the size of the original image and cropped at the same location as the image patch.

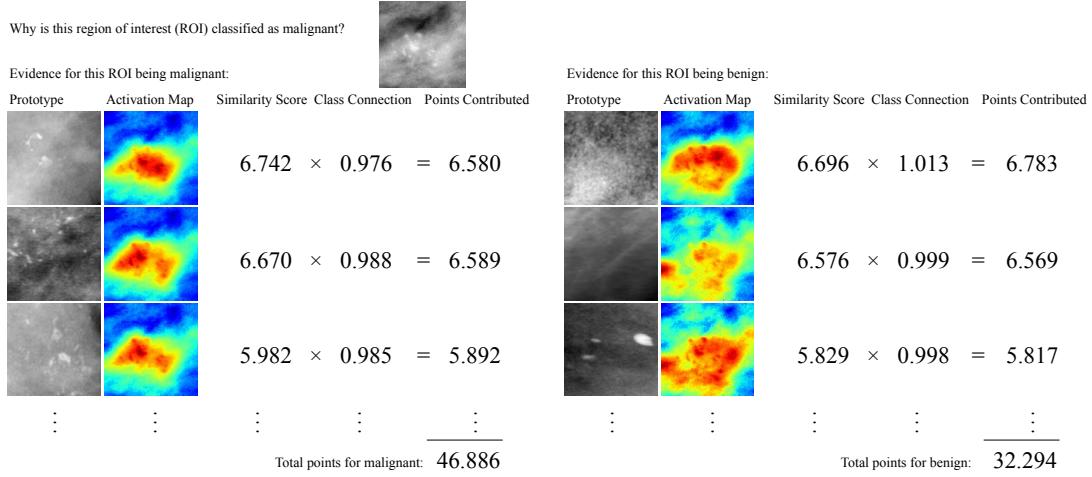


Figure 6. The reasoning process of our network in deciding if the given region of interest contains benign or malignant tumors.

example. We trained and evaluated our interpretable network using the CBIS-DDSM dataset [6, 17, 16]. This is a dataset of mammograms of benign and malignant breast tumors. While it is relatively easy to identify a region of interest (ROI) from a given mammogram, the task of classifying whether an ROI contains a benign or malignant tumor is more difficult. Hence, instead of looking at the entire mammogram, we cropped out the ROI in each mammogram by placing a 224×224 square box at the center of each provided ROI mask. The original dataset contains 1520 training images (1001 benign, 519 malignant) and 328 test images (193 benign, 135 malignant). Since the dataset has a small number of training images, we performed offline data augmentation using random rotation, zoom, and left-right flip to enlarge and balance the training set, so that each class has 10000 training images.

We trained our interpretable network on the CBIS-DDSM dataset using a similar architecture and training algorithm as we did on the CUB-200-2011 dataset. The

test accuracy of our interpretable network is 82.6%, which is comparable with that of the baseline (non-interpretable) VGG-16 model (84.0% as reported in [28]). The area under the receiver operating characteristic curve (AUROC) of our network is 0.86, which is also comparable with that of VGG-16 (0.89 as reported in [37]).

Figure 6 shows the reasoning process of our interpretable network in reaching a classification decision on a test mammogram ROI at the top of the figure. As we can see, our network considers both the evidence for and against the given ROI containing malignant tumors, by comparing the given ROI with prototypical cases of malignant and benign tumors. Our network also highlights the parts in the given ROI that it thinks is similar to the prototypical cases, as shown by the activation maps. Observe that both the given ROI and the top three activated malignant prototypes (Figure 6, left) contain a similar distribution of irregular calcifications, and the region of irregular calcifications in the given ROI is highlighted in the activation maps for the three

prototypes. On the other hand, both the given ROI and the most activated benign prototype (the first prototype in Figure 6, right) contain a similarly large volume of calcifications, even though the prototype comes from a training image of benign tumors. The similarity scores from the comparison with prototypes are weighted by the contributions of those prototypes to the class being considered, and then summed to give a final score for the class. In Figure 6, the final score for the given ROI containing malignant tumors is higher than the final score for it containing benign tumors, so the ROI is (correctly) classified as malignant. This case study will be presented in a separate work. We hope that our work will give both physicians and patients a perspective on evidence from similar mammograms.

4. Conclusion

In this work, we have defined a form of interpretability in image processing (*this looks like that*) that agrees with the way humans describe their own reasoning in classification. We have presented a network architecture that accommodates this form of interpretability, where the comparison of image parts to learned prototypes is integral to the way our network classifies new examples. We have described our specialized training algorithm, and applied our technique to bird species identification and breast cancer detection.

References

- [1] J. Ba and D. Kingma. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 19
- [2] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3319–3327. IEEE, 2017. 2
- [3] J. Bien and R. Tibshirani. Prototype Selection for Interpretable Classification. *Annals of Applied Statistics*, 5(4):2403–2424, 2011. 2
- [4] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 2, 5
- [5] Chenyue Wu and Esteban G. Tabak. Prototypal analysis and prototypal regression. *CoRR*, abs/1701.08916, 2017. 2
- [6] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppenel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013. 8
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 3
- [8] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing Higher-Layer Features of a Deep Network. Technical Report 1341, University of Montreal, June 2009. Also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montreal, Canada. 2
- [9] K. Ghiasi-Shirazi. Generalizing the convolution operator in convolutional neural networks. *arXiv preprint arXiv:1707.09864*, 2017. 2
- [10] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2
- [11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2
- [12] G. E. Hinton. A Practical Guide to Training Restricted Boltzmann Machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012. 2
- [13] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1182, 2016. 2, 5
- [14] B. Kim, C. Rudin, and J. Shah. The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1952–1960, 2014. 2
- [15] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 609–616, 2009. 2
- [16] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4:170177, 2017. 8
- [17] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin. Curated breast imaging subset of ddsm. *The Cancer Imaging Archive*, 2016. 8
- [18] T. Lei, R. Barzilay, and T. S. Jaakkola. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2
- [19] O. Li, H. Liu, C. Chen, and C. Rudin. Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2, 5
- [20] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015. 5
- [21] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7, 16
- [22] K. Nalaie, K. Ghiasi-Shirazi, and M.-R. Akbarzadeh-T. Efficient implementation of a generalized convolutional neural networks based on weighted euclidean distance. In *Computer and Knowledge Engineering (ICCKE), 2017 7th International Conference on*, pages 211–216. IEEE, 2017. 2

- [23] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 3387–3395, 2016. [2](#)
- [24] P. O. Pinheiro and R. Collobert. From Image-Level to Pixel-Level Labeling With Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015. [2](#)
- [25] C. E. Priebe, D. J. Marchette, J. G. DeVinney, and D. A. Socolinsky. Classification Using Class Cover Catch Digraphs. *Journal of classification*, 20(1):003–023, 2003. [2](#)
- [26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. [2](#)
- [27] 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 *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. [2](#)
- [28] L. Shen. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *arXiv preprint arXiv:1708.09427*, 2017. [8](#)
- [29] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1143–1151, 2015. [2](#)
- [30] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICLR) Workshop*, 2014. [2](#)
- [31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [3](#)
- [32] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. [2](#)
- [33] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. [2](#)
- [34] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. [2](#)
- [35] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel Recurrent Neural Networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1747–1756, 2016. [2](#)
- [36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. [2, 3](#)
- [37] J. Wu, D. Peck, S. Hsieh, V. Dialani, C. D. Lehman, B. Zhou, V. Syrgkanis, L. Mackey, and G. Patterson. Expert identification of visual primitives used by cnns during mammogram classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105752T. International Society for Optics and Photonics, 2018. [8](#)
- [38] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The Application of Two-Level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 842–850. IEEE, 2015. [2](#)
- [39] J. Yosinski, J. Clune, T. Fuchs, and H. Lipson. Understanding Neural Networks through Deep Visualization. In *In ICML Workshop on Deep Learning*, 2015. [2](#)
- [40] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2014. [2](#)
- [41] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014. [2, 5](#)
- [42] Q. Zhang, Y. N. Wu, and S.-C. Zhu. Interpretable Convolutional Neural Networks. *arXiv preprint arXiv:1710.00935*, 2017. [2](#)
- [43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016. [2, 5, 6](#)
- [44] B. Zhou, Y. Sun, D. Bau, and A. Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. [2](#)

Supplementary Material

5. More Examples of How Our Interpretable Network Classifies Birds

In this section, we provide more examples of how our interpretable network classifies previously unseen images of birds.

Figures 7 through 12 give six examples of how our interpretable network correctly classifies a previously unseen image of a bird and how our network explains its prediction. In each of these figures, the left side presents evidence for the given bird belonging to the class with the highest logit, and the right side presents evidence for the given bird belonging to the class with the second highest logit. We shall give some general observations regarding the ways in which our network thinks that the given image is similar to the prototypical cases. The detailed reasoning process of our network has been explained in our main paper, and will not be repeated here.

Figure 7 demonstrates how our interpretable network correctly classifies an image of a Baltimore oriole. In particular, our network thinks that the golden chest of the given bird is similar to the prototypical chest of a Baltimore oriole, as evidenced by the activation maps of the two most activated Baltimore oriole prototypes (the first two prototypes in Figure 7, left): both of these prototypes correspond to the characteristic golden chest of a Baltimore oriole, and their activation maps on the given image also highlights the golden chest of the given bird. Our network also thinks that the black and white striped wing of the given bird is similar to the prototypical wing of a Baltimore oriole, as shown by the activation map of the third most activated Baltimore oriole prototype (the third prototype in Figure 7, left). On the other hand, our network thinks that there is some resemblance between the golden chest of the given bird and the more yellowish chest of a hooded oriole, as shown by the activation maps of the hooded oriole prototypes (Figure 7, right), but the resemblance is not as strong as that between the golden chest of the given bird and the prototypical chest of a Baltimore oriole, as shown by the generally smaller similarity scores between the given bird and the hooded oriole prototypes. Not surprisingly, when our network accumulates the evidence presented by the comparison with all the prototypes, it sees that the evidence for the given bird being a Baltimore oriole is the strongest, and concludes that the bird is a Baltimore oriole.

Figure 8 demonstrates how our interpretable network correctly classifies an image of a house wren. In particular, our network thinks that the wing of the given bird is similar to the prototypical wing of a house wren (Figure 8, left). On the other hand, our network thinks that there is some small resemblance between the wing/head of the given bird and the prototypical wing/head of a rock wren (Figure 8, right).

Figure 9 demonstrates how our interpretable network correctly classifies an image of a Kentucky warbler. In particular, our network thinks that the yellow and black striped head of the given bird is very similar to the prototypical head of a Kentucky warbler (Figure 9, left). Our network also thinks that the yellow chest of the given bird bears some resemblance to the prototypical yellow chest of a tropical kingbird (Figure 9, right).

Figure 10 demonstrates how our interpretable network correctly classifies an image of a pied-billed grebe. In particular, our network thinks that the back and the head of the given bird are similar to the prototypical back and the prototypical head of a pied-billed grebe (Figure 10, left). Our network also thinks that there is some resemblance between the back of the given bird and the prototypical back of a eared grebe (the first two prototypes in Figure 10, right), and the resemblance between the neck of the given bird and the prototypical neck of a eared grebe is very small (the third prototype in Figure 10, right).

Figure 11 demonstrates how our interpretable network correctly classifies an image of a western meadowlark. In particular, our network thinks that the wing and the chest of the given bird is similar to the prototypical wing and the prototypical chest of a western meadowlark (Figure 11, left). Our network also thinks that there is some resemblance between the wing of the given bird and the prototypical wing of a grasshopper sparrow (Figure 11, right), and not surprisingly, both species have wings that are gray and black striped.

Figure 12 demonstrates how our interpretable network correctly classifies an image of a common tern. Different species of terns are strikingly similar, and it is difficult for humans to distinguish them. In this case, our network thinks that there is almost as much evidence for the given bird being a common tern as that for the given bird being a black tern – this is shown by the very close logits (total points) for the common tern class and for the black tern class. Our network thinks that the wings of the given bird are more similar to the prototypical wings of a common tern than to the prototypical wings of a black tern, as shown by the slightly higher similarity scores between the given bird and the common tern prototypes (whose activation maps highlight the wings of the given bird). Note that the second and the third most activated common tern prototypes on the given image (the second and the third prototypes in Figure 12, left) are the same: this results from the projection of each prototype onto the closest latent representation of training image patches from the prototype’s designated class (described in Section 2.2.2 in the main paper) – in this case, the closest training patches to both prototypes are the same before the projection stage, and consequently both prototypes are projected onto the same patch in the latent space. This means that some of the learned prototypes in our network

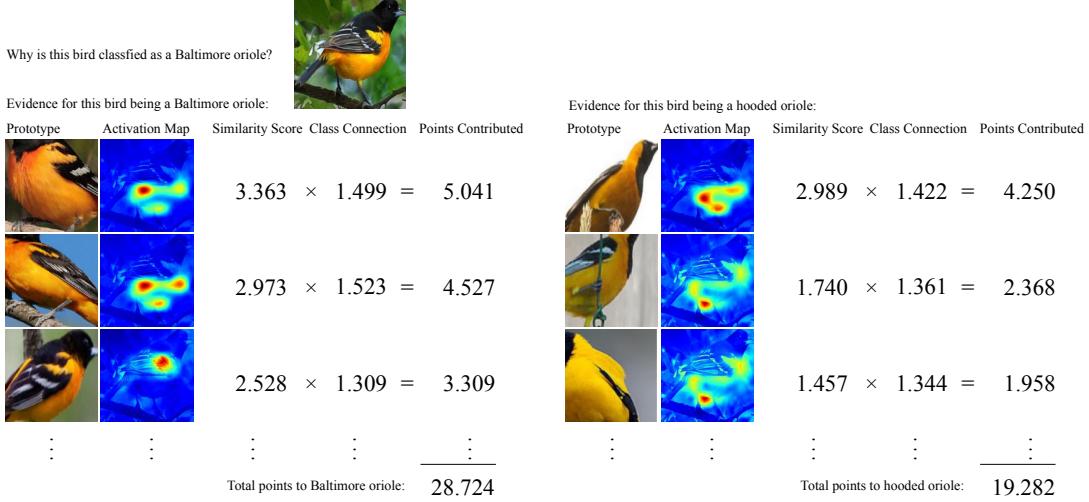


Figure 7. How our interpretable network correctly classifies an image of Baltimore oriole.

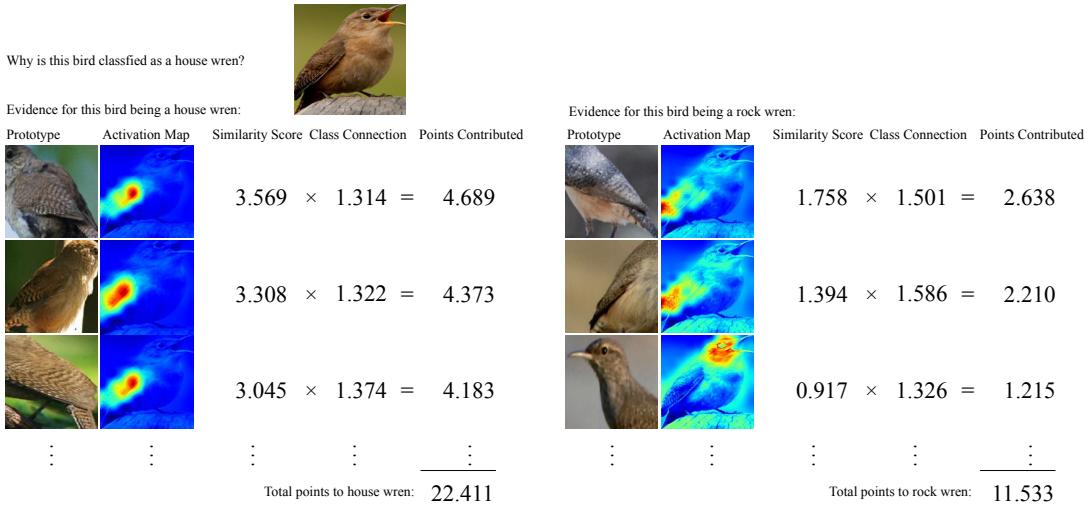


Figure 8. How our interpretable network correctly classifies an image of house wren.

are repeated. However, this is not a problem because we can conceptually understand the repeated prototypes as one prototype, with its weight connection to each class in the fully connected last layer being the sum of the weight connections of those repeated prototypes to that class. Thus, we can understand the second and the third common tern prototypes in Figure 12 (left) as one common tern prototype with class connection $1.323 + 1.323 = 2.646$. This also means that the actual number of prototypes used by our interpretable network is in general *less* than the pre-determined number of prototypes when the network architecture is specified.

Figures 13 and 14 give two examples when our network mistakes the identity of the given bird – the transparency of our network means that when our network makes a mistake, we are able to see the reasoning behind its misclassification. In each of these figures, the left side presents evidence for

the given bird belonging to the predicted class, and the right side presents evidence for the given bird belonging to the actual class.

Figure 13 shows why our network identifies the given bird as a prothonotary warbler instead of its true identity – a Wilson warbler. As we can see, our network thinks that the wing of the given bird is similar to the prototypical wing of a prothonotary warbler (the first two prototypes in Figure 13, left). More interestingly, our network also thinks that the head of the given bird is more similar to the prototypical head of a prothonotary warbler than to a Wilson warbler: this is shown by the third prothonotary warbler prototype (Figure 13, left), whose activation map on the given image highlights the head of the given bird, having a higher similarity score to the given image than the first and the third Wilson warbler prototypes (Figure 13, right), whose activation maps on the given image also highlight the head of

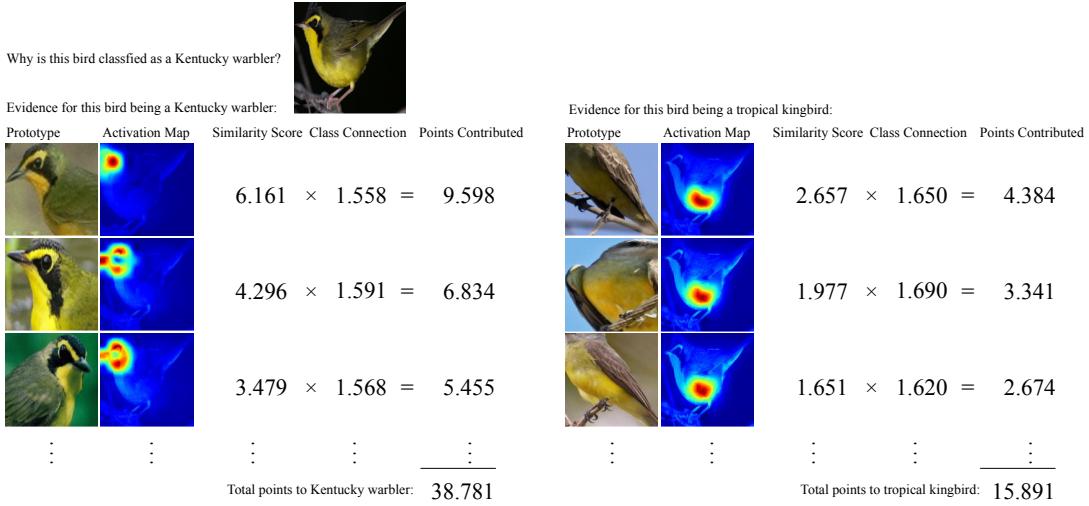


Figure 9. How our interpretable network correctly classifies an image of Kentucky warbler.

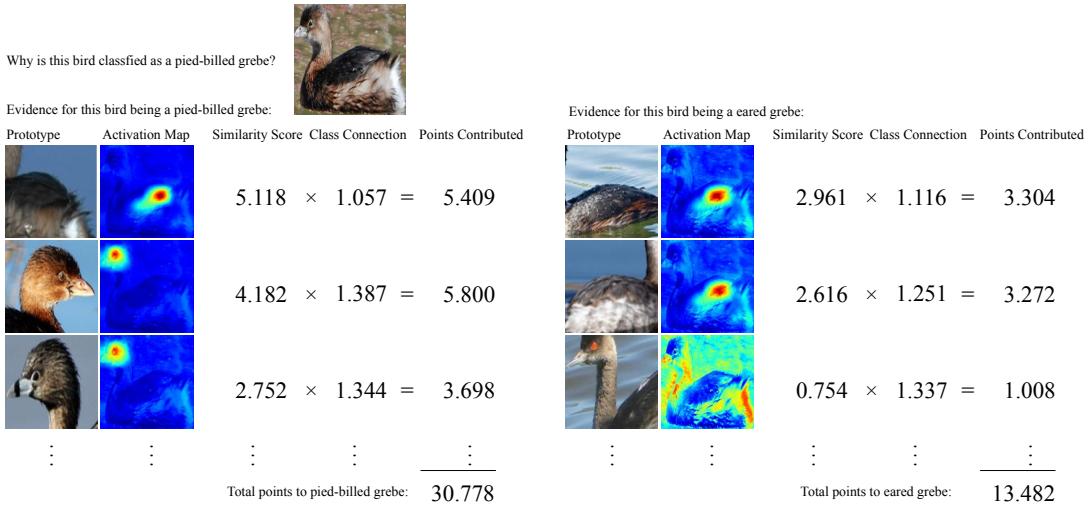


Figure 10. How our interpretable network correctly classifies an image of pied-billed grebe.

the given bird. In the end, the network finds more evidence for the given bird being a prothonotary warbler than being a Wilson warbler.

Figure 14 shows why our network identifies the given bird as a red cockaded woodpecker instead of its true identity – a downy woodpecker. By looking at the activation maps and the similarity scores, we see our network thinks that the wing of the given bird is more similar to the prototypical wing of a red cockaded woodpecker than to the prototypical wing of a downy woodpecker, and it finds more evidence in the end for the given bird being a red cockaded woodpecker than being a downy woodpecker.

6. More Examples of Nearest Prototypes of Given Images

In this section, we provide more examples of the nearest prototypes of given test images.

Figure 15 shows the three nearest prototypes to each of the six test bird images (the activation map beside each prototype provides clue for which patch of the original image is closest to the prototype in the latent space). For a given image, we define its nearest prototype as the one that forms the closest patch-prototype pair in the latent space, over all patches of the given image. As we can see from Figure 15, the nearest prototypes for each of these test images generally come from the same class as that of the image. There are some exceptions: for example, the third nearest prototype for the cardinal in Figure 15 corresponds to the body of a summer tanager, and the third nearest prototype for the

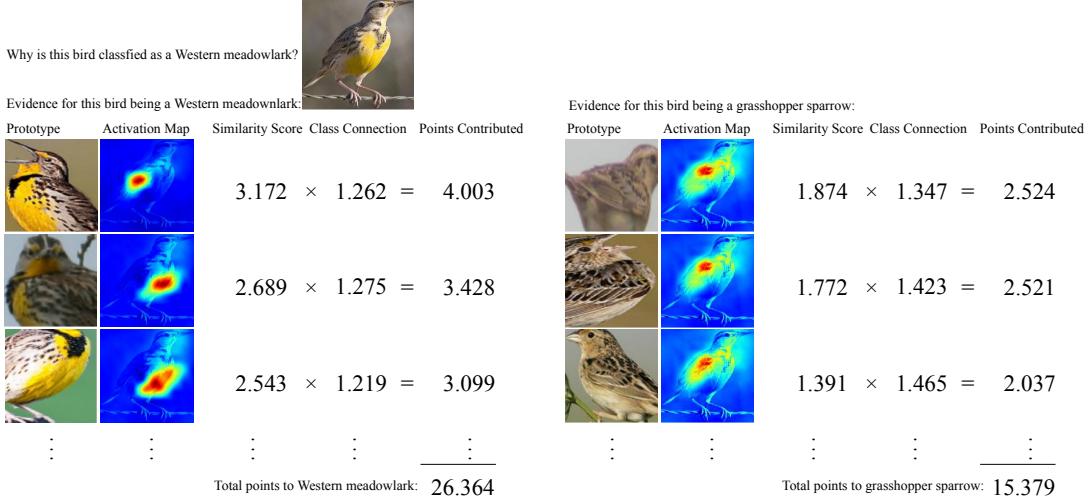


Figure 11. How our interpretable network correctly classifies an image of western meadowlark.

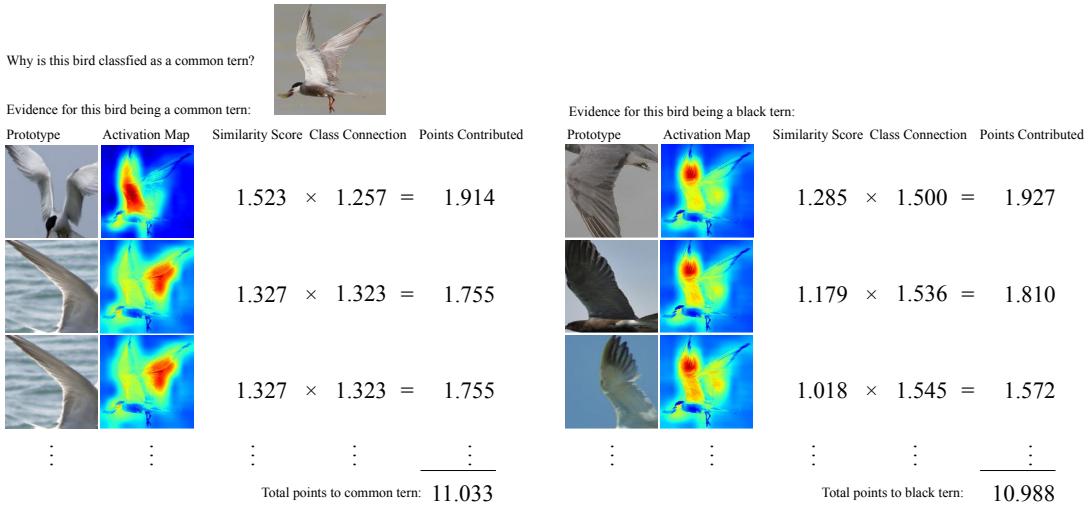


Figure 12. How our interpretable network correctly classifies an image of common tern.

blue winged warbler corresponds to the wing of a Cerulean warbler. This is understandable, because a cardinal has a red body much like that of a summer tanager, and a blue winged warbler has blue and white-striped wings much like those of a Cerulean warbler. Hence, it is not too surprising that the red body of the cardinal in Figure 15 is close to the prototypical red body of a summer tanager (as shown by the strong activation around the body of the cardinal in the activation map), and the blue and white-striped wing of the blue winged warbler is close to the prototypical wing of a Cerulean warbler in the latent space (as shown by the strong activation around the wing of the blue winged warbler in the activation map). This shows that the latent space learned by our interpretable network does have a clustering structure, where semantically similar patches that are relevant for classification are clustered together.

Figure 16 gives two examples that show the limitations

of our approach. The example on the left shows that our method sometimes chooses latent representations corresponding to image background as prototypes and clusters image patches around these background prototypes in the latent space: as shown in Figure 16 (left), the green background in the downy woodpecker image looks like that in a bay breasted warbler image or a ruby throated hummingbird image. While the green background does in some ways help our network classify (e.g. the presence of a green background means that the bird most likely lives in the woods, and is less likely a water or sea bird), it may sometimes cause our network to misclassify (because there are many bird species that live in the woods). The example on the right of Figure 16 shows that sometimes the nearest prototypes of an image may not come from the image's own class. This usually happens because of strong resemblance among semantic parts of different classes: as shown in Fig-

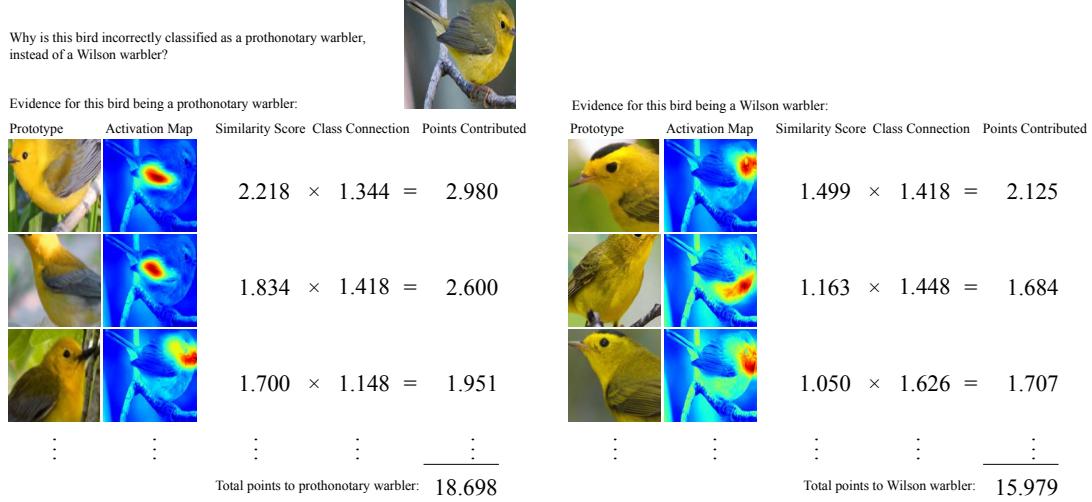


Figure 13. How our interpretable network mistakes a Wilson warbler as a prothonotary warbler.

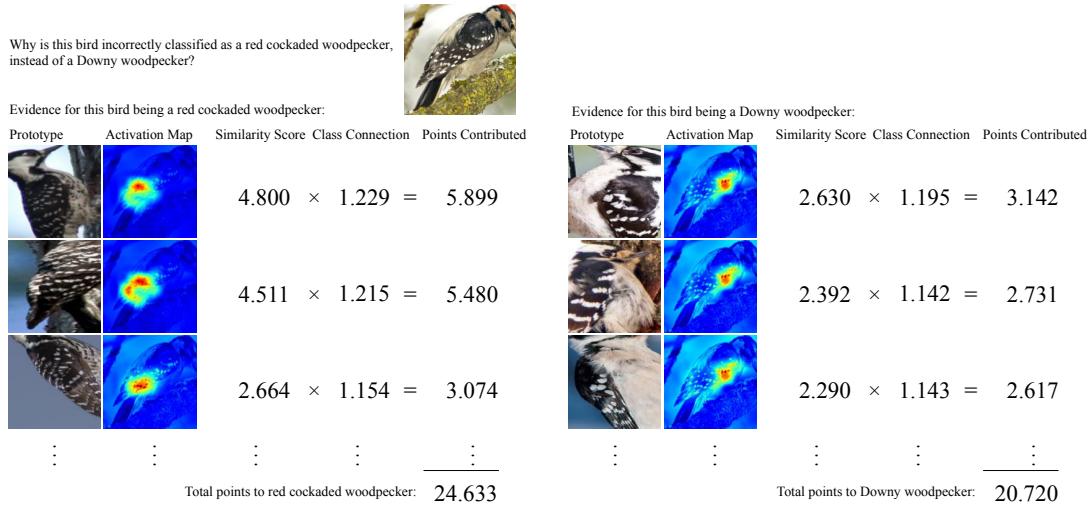


Figure 14. How our interpretable network mistakes a downy woodpecker as a red cockaded woodpecker.

ure 16 (right), the wing of the yellow bellied flycatcher bears strong resemblance to the prototypical wing of a black capped vireo, a white eyed vireo, or an Acadian flycatcher. However, these limitations do not detract from our finding that semantically similar patches are close in the latent space: in the former example, the image patches corresponding to green background are close in the latent space, and in the latter example, the patches corresponding to similar wings are close in the latent space, albeit from different classes.

7. More Examples of Nearest Image Patches of Given Prototypes

Figure 17 shows the nearest image patches (in the latent space) of ten prototypes from both the training and the test set, along with an activation map for each prototype

on each image patch (how these activation maps were generated has been described in our main paper). As we can see, the nearest image patches to each prototype in the figure all contain similar bird parts, and the activation maps for each prototype on its nearest image patches all highlight similar structures of birds (e.g. wings). This demonstrates that our interpretable network is able to learn a meaningful latent space where similar parts or semantic concepts are clustered together, with only weak supervision.

8. t-SNE Visualization of Latent Space

The examples of the nearest prototypes of given images and those of the nearest image patches of given prototypes presented in the previous sections already demonstrate that the learned latent space of our network has a clustering structure where semantically similar patches are clustered

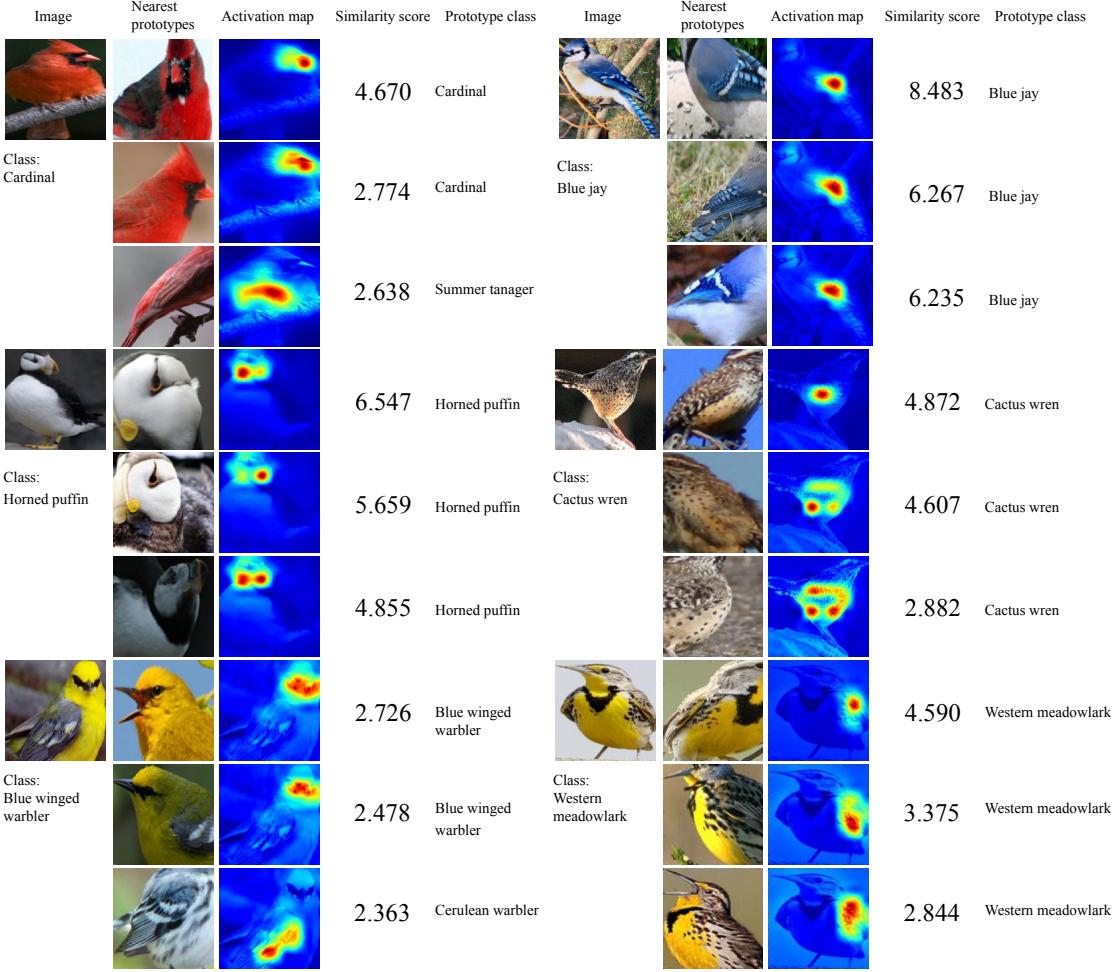


Figure 15. Nearest prototypes of six test images.

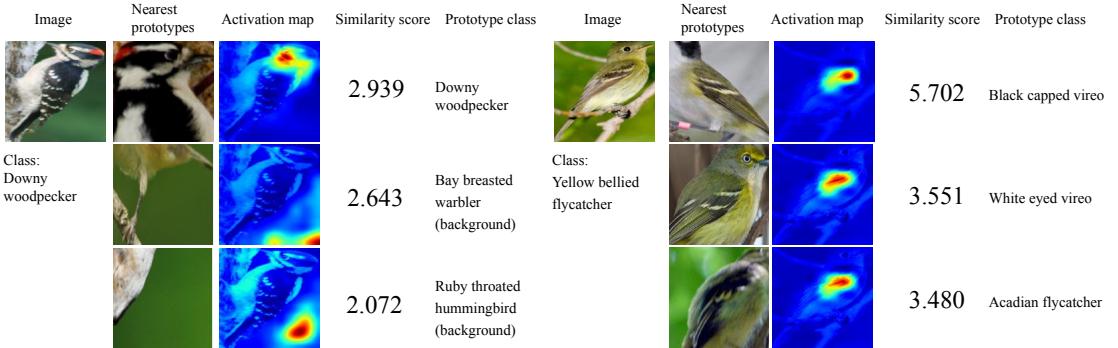


Figure 16. Nearest prototypes of two test images: the example on the left shows that some prototypes correspond to image background and these background prototypes could be some of the nearest prototypes of a given image; the example on the right shows that sometimes the nearest prototypes of an image may not come from the image’s own class.

around prototypes. In this section, we provide another visualization of the latent space using t-distributed stochastic

neighbor embedding, or t-SNE [21]. For the sake of visual clarity, we use only the first five species of birds in the test

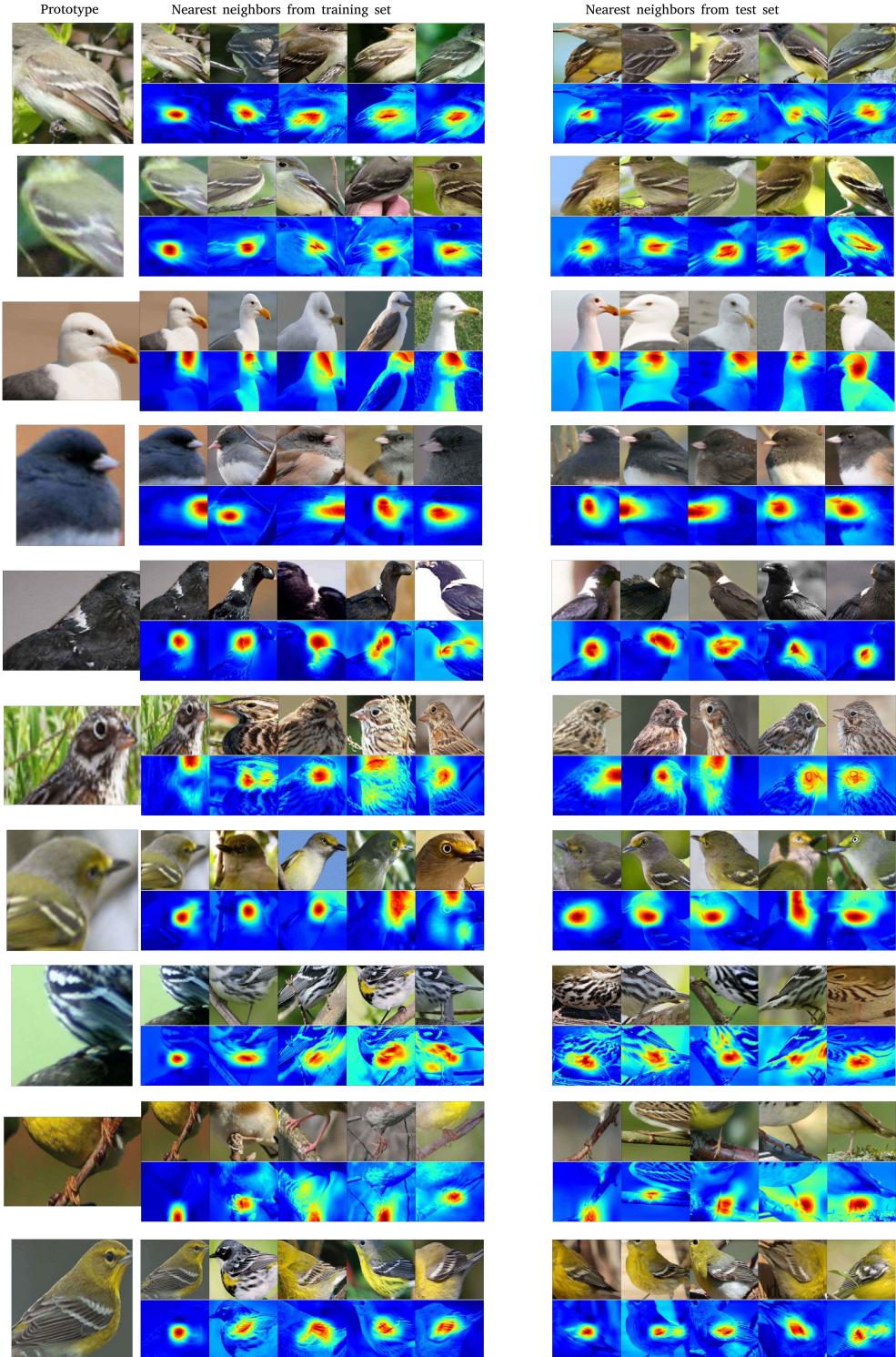


Figure 17. Nearest image patches (in the latent space) of ten prototypes from both the training and the test set. The activation map below each image patch shed lights on which part or semantic concept the prototype is most likely detecting: it is obtained by passing the original image containing the patch through the network to generate a map of similarity scores, which is then upsampled to the size of the original image and cropped at the same location as the image patch.

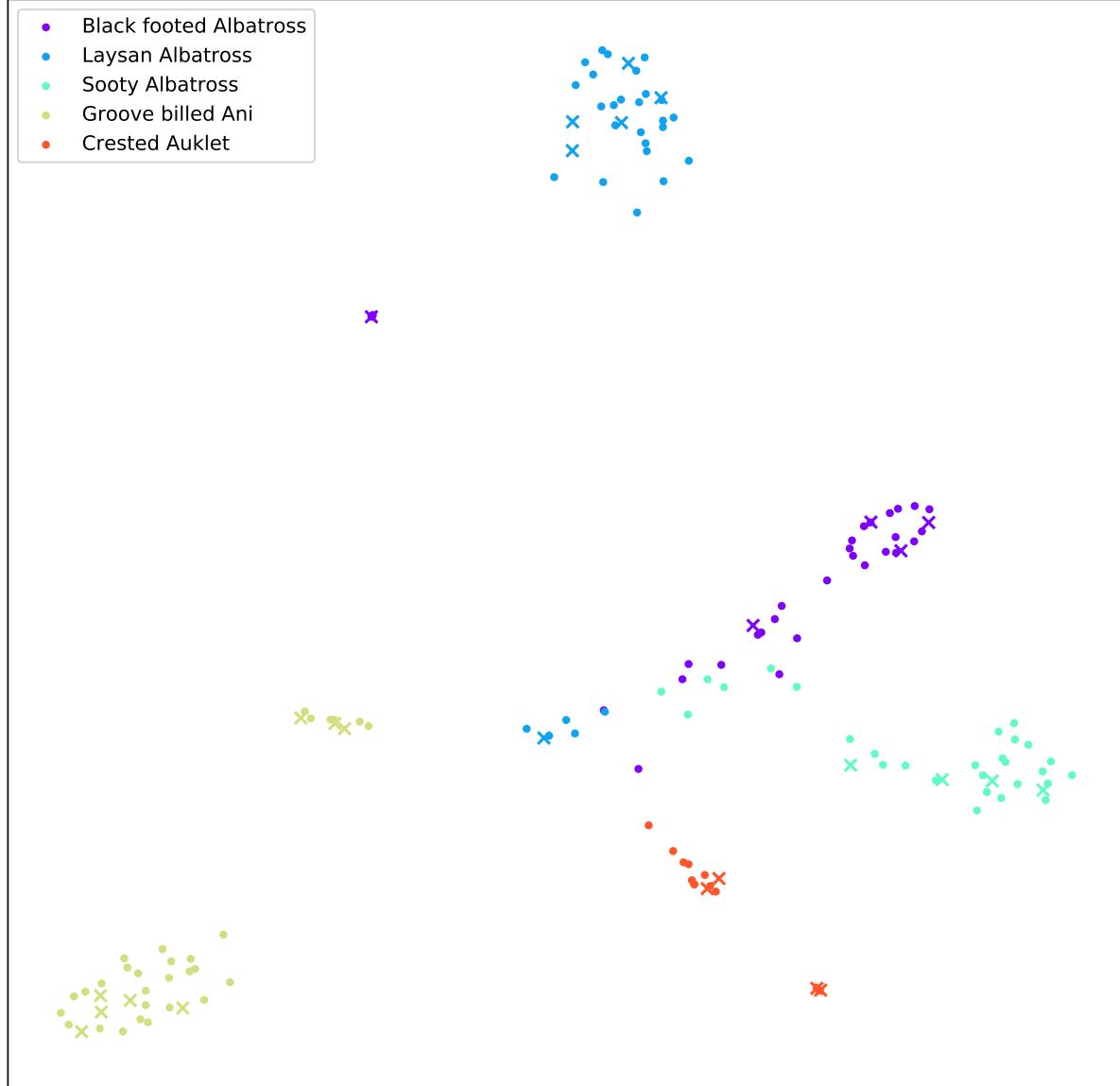


Figure 18. t-SNE visualization of latent representations of test images from five classes. Each dot represents the closest patch of a test image to some prototype, and each cross represents a prototype. Each color represents a different class identity.

set in our visualization. Since we require only one patch from every image to be close to one of the prototypes of its own class in our training objective (see the definition of the cluster cost in Section 2.2.1 in the main paper), for each test image, we embed only the latent representation of the patch that is closest to one of the prototypes of its own class for visualization. Figure 18 shows the t-SNE visualization of the latent representations of the test image patches: each dot in the figure represents the closest patch of a test image to some prototype of its own class, and each cross represents a prototype. Figure 19 shows the same t-SNE embedding but with dots and crosses replaced by the image patches: the images with rectangular bounding boxes are the test im-

age patches, and the images with sawtooth bounding boxes are the visualizations of the prototypes. As shown in Figures 18 and 19, the test image patches are clustered around the prototypes of their respective classes, and the clusters from different classes are in general well separated from each other.

9. Training Details

In this section, we describe the hyperparameters we used to train our interpretable network on the CUB-200-2011 dataset.

In the first training stage (stochastic gradient descent of the layers before the fully connected last layer, described in

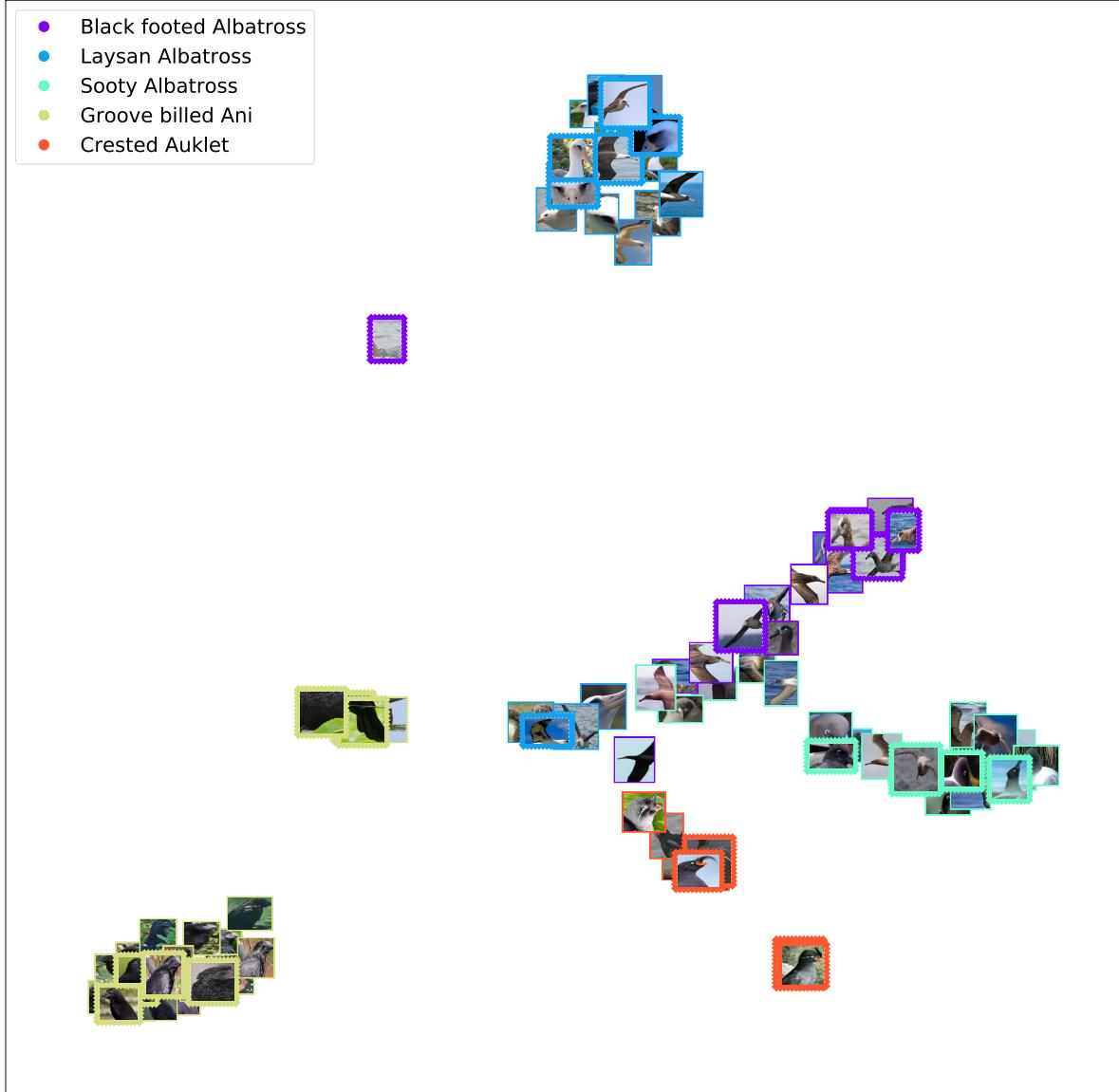


Figure 19. t-SNE visualization of latent representations of test images from five classes. The images with rectangular bounding boxes are the test image patches, and the images with sawtooth bounding boxes are the visualizations of the prototypes. Each color represents a different class identity. This figure shows the same t-SNE embedding as Figure 18, but with dots and crosses replaced by the image patches.

Section 2.2.1 in the main paper), we set the coefficient λ_1 of the cluster cost in the training objective to 1. We set the coefficient λ_2 of the separation cost to 0.1 when we trained our network on 224×224 full/cropped bird images, and to 0.25 when we trained our network on 336×336 cropped bird images. We also used weight decay on the convolutional layers of our network, with coefficient 10^{-3} . In our experiments, we divided this training stage into two sub-stages. In the first sub-stage, we loaded the pre-trained weights and biases of the VGG-16 network into the first 13 convolutional layers of our network and fixed these layers, and trained only

the two additional 1×1 convolutional layers as well as the prototype layer with learning rate 3×10^{-3} , for 5 epochs. In the second sub-stage, we trained all the convolutional layers and the prototype layer jointly, using 10^{-4} learning rate for the first 13 convolutional layers and 3×10^{-3} learning rate for the two additional convolutional layers and the prototype layer, for another 5 epochs. We then trained these layers for one more epoch using 1/10 of the original learning rates for the respective layers. We used Adam optimization [1] in this training stage.

In the third training stage (convex optimization of the last

layer, described in Section 2.2.3 in the main paper), we set the coefficient λ of the sparsity term in the training objective to 10^{-4} . We used Adam optimization with a learning rate of 10^{-4} to optimize the last layer.

The hyperparameters we used to train our network on the CBIS-DDSM dataset are similar, except that in the first training stage, we set both the coefficient λ_1 of the cluster cost and the coefficient λ_2 of the separation cost in the training objective to 10^{-3} , and we trained all the convolutional layers and the prototype layer jointly from the beginning, using Adam optimization with 10^{-5} learning rate for the first 13 convolutional layers and 3×10^{-4} learning rate for the two additional convolutional layers and the prototype layer. In the third training stage, we set the coefficient λ of the sparsity term in the training objective to 10^{-3} , and used Adam optimization with a learning rate of 10^{-4} to optimize the last layer.