Interpretable Image Recognition by Constructing Transparent Embedding Space (Supplementary Material)

Jiaqi Wang, Huafeng Liu*, Xinyue Wang, Liping Jing*
School of Computer and Information Technology, Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University, Beijing, China
{jiaqi.wang, huafeng, xinyuewang, lpjing}@bjtu.edu.cn

1. Additional experiments on CUB-200-2011

Orthogonality analysis. The correlation of concept representations is evaluated with cosine distance in ProtoPNet [1] and TesNet and the correlation matrix is visualized in Figure 1. In subfigure (a) for ProtoPNet, the concept representations are correlated with each other. In contrast, in subfigure (b, c) for TesNet, the concept representations are highly uncorrelated (orthogonal) for each other and orthonormality loss enhances uncorrelation.

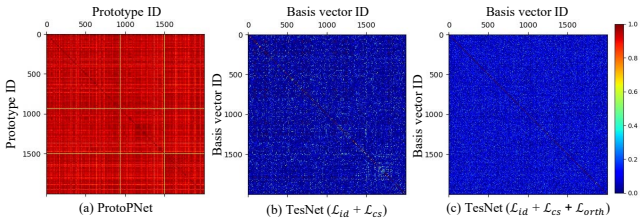


Figure 1. The correlation matrix of concepts (cosine distance between concept representations) in ProtoPNet and TesNet.

Visualization comparison. The reasoning process of TesNet and ProtoPNet shown in Figure 2, ProtoPNet learns all the same concepts "wing" but TesNet learns the different concepts "wing", "head" and "fur" among the three concepts with the highest contribution. TesNet outperforms ProtoPNet to learn disentangled concepts.

2. More examples of how TesNet classifies birds

In this section, we provide more examples of how our TesNet classifies images on CUB-200-2011. Figure 3 provides examples of how our TesNet correctly classifies a previously unseen image of the Evening Grosbeak with different CNN base architectures. Figure 4 provides more examples of how our TesNet correctly other previously unseen images. In each subfigure, the left side presents the 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 a closely related class.

Figure 3 demonstrates how our TesNet correctly classifies an Evening Grosbeak with different CNN base architec-

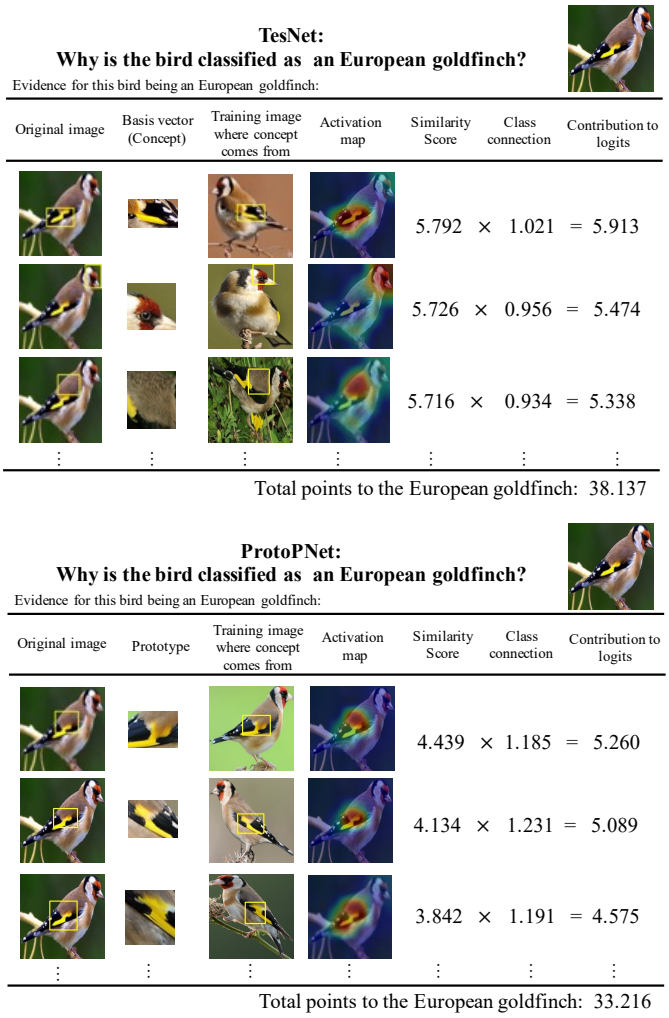


Figure 2. The visualization comparsion in ProtoPNet and TesNet.

*Corresponding authors.

Table 1. Ablation study on cropped car images of Stanford Cars dataset

Method	VGG16	VGG19	ResNet34	ResNet152	Dense121	Dense161
TesNet ($\mathcal{L}_{id}$+$\mathcal{L}_{cs}$)	89.7 $\pm$ 0.2	90.0 $\pm$ 0.2	90.2 $\pm$ 0.2	90.8 $\pm$ 0.2	90.2 $\pm$ 0.1	91.0 $\pm$ 0.1
TesNet($\mathcal{L}_{id}$+$\mathcal{L}_{cs}$+$\mathcal{L}_{orth}$)	88.8 $\pm$ 0.4	89.2 $\pm$ 0.2	90.8 $\pm$ 0.3	89.8 $\pm$ 0.3	90.5 $\pm$ 0.2	91.2 $\pm$ 0.2
TesNet($\mathcal{L}_{id}$+$\mathcal{L}_{cs}$+$\mathcal{L}_{orth}$+$\mathcal{L}_{ss}$)	**90.3 $\pm$ 0.2**	**90.6 $\pm$ 0.2**	**90.9 $\pm$ 0.2**	**92.0 $\pm$ 0.2**	**91.9 $\pm$ 0.3**	**92.6 $\pm$ 0.3**

tures. In particular, each TesNet has learned the prototypical black and white wings of the Evening Grosbeak and is able to associate the black and white wings of the previously unseen given image to the basis concept of the Evening Grosbeak. Each TesNet also learned basis concepts such as the bright yellow forehead and golden abdomen of the Evening Grosbeak. Our TesNet accumulates the evidence presented by the comparison with all basis concepts and concludes that the bird is an Evening Grosbeak.

Figure 4 provides more reasoning processes on CUB-200-2011. Specifically, TesNet is able to learn the prototypical red legs of the Red-Legged Kittiwake, white plumes above the eyes of the Rhinoceros Auklet, and the bright yellow throat of the Cape May Warbler. Our TesNet is able to pick out a similar patch that contains the basis concept on the previously unseen image, and also highlight this basis concept in the activation map of the given image.

3. Additional experiment on Stanford Cars

In this section, we conduct an ablation study on the Stanford Cars to evaluate the components in embedding space learning. As shown in Table 1, the orthonormality loss slightly reduces the classification performance at most 1%, while the subspace separation loss slightly improves the performance at most 2.2%.

4. More examples of how TesNet classifies cars

In this section, we provide examples of how our TesNet classifies a previously unseen image of the Tesla Model S Sedan 2012, the Rolls-Royce Phantom Sedan 2012, and BMW 1 Series Couple 2012 in Figure 5. In these examples, our TesNet learns the basis concepts of the front include the car logo, headlights, doors, and so on. Moreover, we can look the evidence for this car being a BMW 1 Series Convertible 2012 in Figure 5 (c) to show that even if there is a different perspective between the front of the given image and the basis concept, our model can pick out the prototypical concept in the given image.

5. Implementation details

In this section, we describe our choice of hyperparameters and other training details.

5.1. Training parameters

In the experiments , we set the coefficient of the compactness loss, the separation loss, the orthonormality loss, the subspace separation loss to 0.8, -0.08, 10^{-4}, 10^{-7} during the stage of embedding space learning and we set the coefficient of the L_1-regularization term to 10^{-4} during the concept based classification.

In the stage of embedding space learning, we started with a "warm-up" stage, where we loaded and froze the pre-trained weights and biases, and focused on warming the two additional convolutional layers and basis vectors in the subspace layer for the first 5 epochs. We used the Adam optimizer and the learning rate we used in the "warm-up" is 3×10^{-3}. Subsequently, we trained all the convolutional layers and the subspace layer jointly with 10^{-4} learning rate by Adam optimizer for the weights which were pretrained on ImageNet, and 3×10^{-3} learning rate for the two additional convolutional layers and the subspace layer. We decayed the learning rate by a factor of 0.1 every 5 epochs. In the stage of concept based classification, we optimized the last layer by Adam optimizer with a learning rate of 10^{-4} for 20 iterations.

5.2. Training software and platform

We implemented our TesNet using Pytorch and all experiments were run on 4 NVIDIA GeForce RTX 2080ti GPUs.

References

[1] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32:8930–8941, 2019. 1

Why is the bird classified as an Evening Grosbeak?

Evidence for this bird being an Evening Grosbeak :

Original image	Basis vector (Concept)	Training image where concept comes from	Activation map	Similarity Score	Class connection	Contribution to logits
				6.531	× 0.940	= 6.139
				6.283	× 0.834	= 5.240
				4.566	× 0.892	= 4.072
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Total points to the Evening Grosbeak: 29.913

Evidence for this bird being a White Throated Sparrow:

Original image	Basis vector (Concept)	Training image where concept comes from	Activation map	Similarity Score	Class connection	Contribution to logits
				4.890	× 1.019	= 4.982
				3.967	× 1.042	= 4.133
				2.356	× 1.067	= 2.513
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Total points to the White Throated Sparrow: 11.931

(a) VGG16-based TesNet

Why is the bird classified as an Evening Grosbeak?

Evidence for this bird being an Evening Grosbeak :

Original image	Basis vector (Concept)	Training image where concept comes from	Activation map	Similarity Score	Class connection	Contribution to logits
				5.786	× 1.006	= 5.820
				5.521	× 0.967	= 5.338
				4.923	× 0.871	= 4.287
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Total points to the Evening Grosbeak: 28.796

Evidence for this bird being a White Throated Sparrow:

Original image	Basis vector (Concept)	Training image where concept comes from	Activation map	Similarity Score	Class connection	Contribution to logits
				5.095	× 1.104	= 5.625
				4.309	× 1.090	= 4.697
				2.163	× 1.121	= 2.425
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Total points to the White Throated Sparrow: 15.110

(b) VGG19-based TesNet

Why is the bird classified as an Evening Grosbeak?

Evidence for this bird being an Evening Grosbeak :

Original image	Basis vector (Concept)	Training image where concept comes from	Activation map	Similarity Score	Class connection	Contribution to logits
				5.228	× 1.038	= 5.427
				4.713	× 0.983	= 4.633
				4.101	× 0.990	= 4.060
⋮	⋮	⋮	⋮	⋮	⋮	⋮

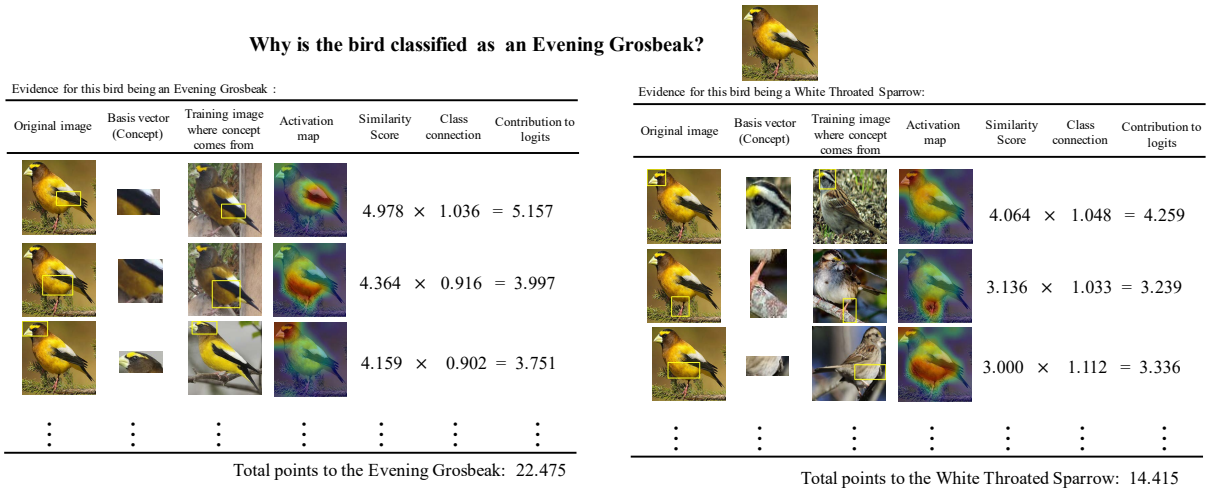
Total points to the Evening Grosbeak: 25.414

Evidence for this bird being a White Throated Sparrow:

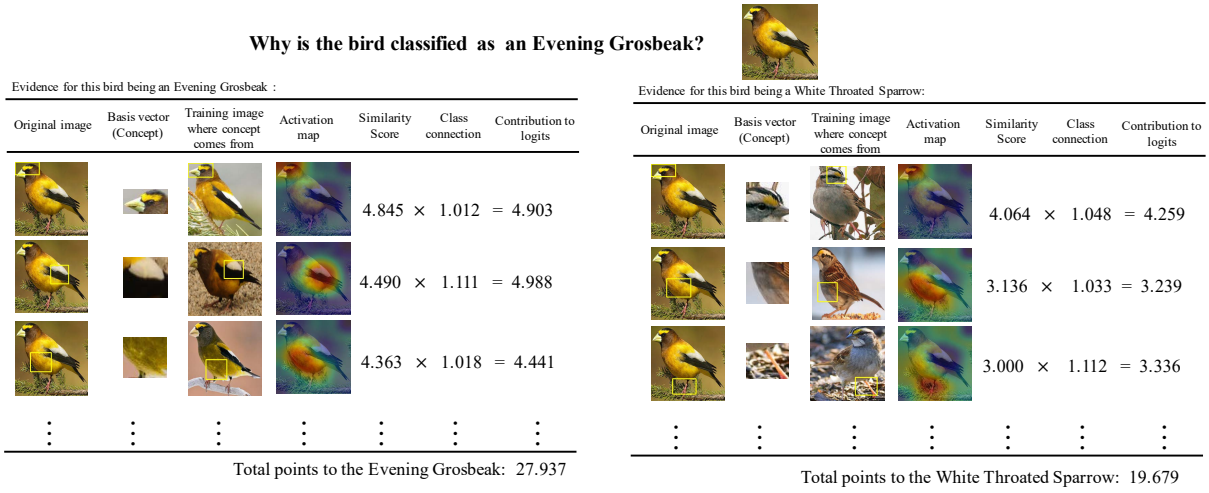
Original image	Basis vector (Concept)	Training image where concept comes from	Activation map	Similarity Score	Class connection	Contribution to logits
				4.643	× 1.054	= 4.894
				3.762	× 1.058	= 3.980
				3.708	× 1.028	= 3.812
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Total points to the White Throated Sparrow: 16.106

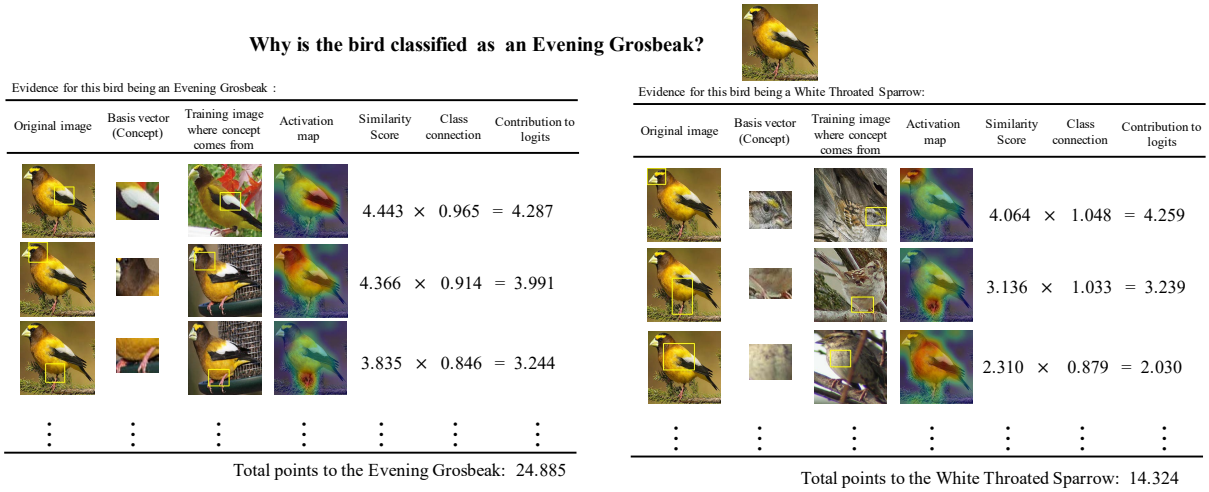
(c) ResNet34-based TesNet



(d) ResNet152-based TesNet

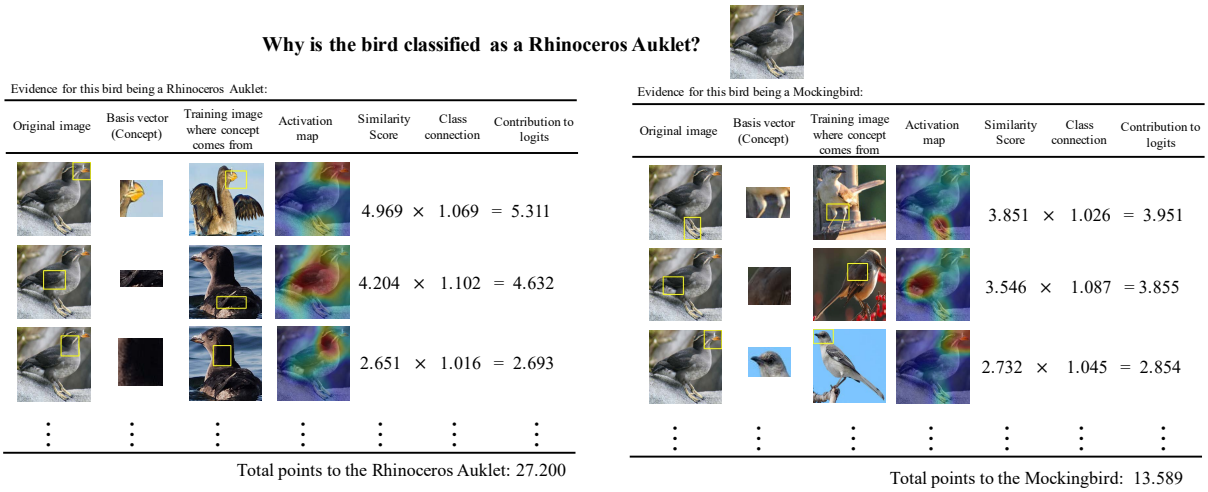


(e) DenseNet121-based TesNet

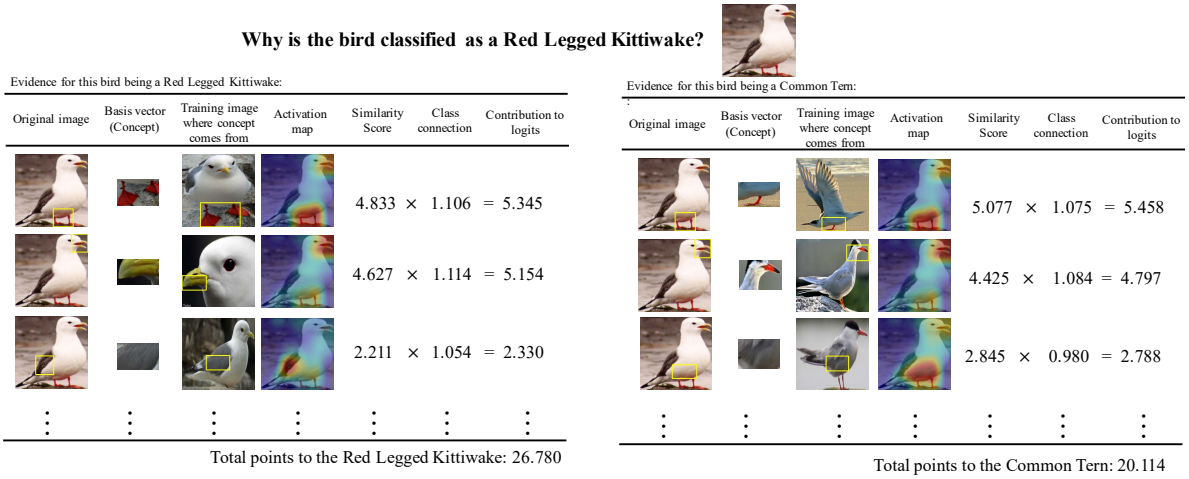


(f) DenseNet161-based TesNet

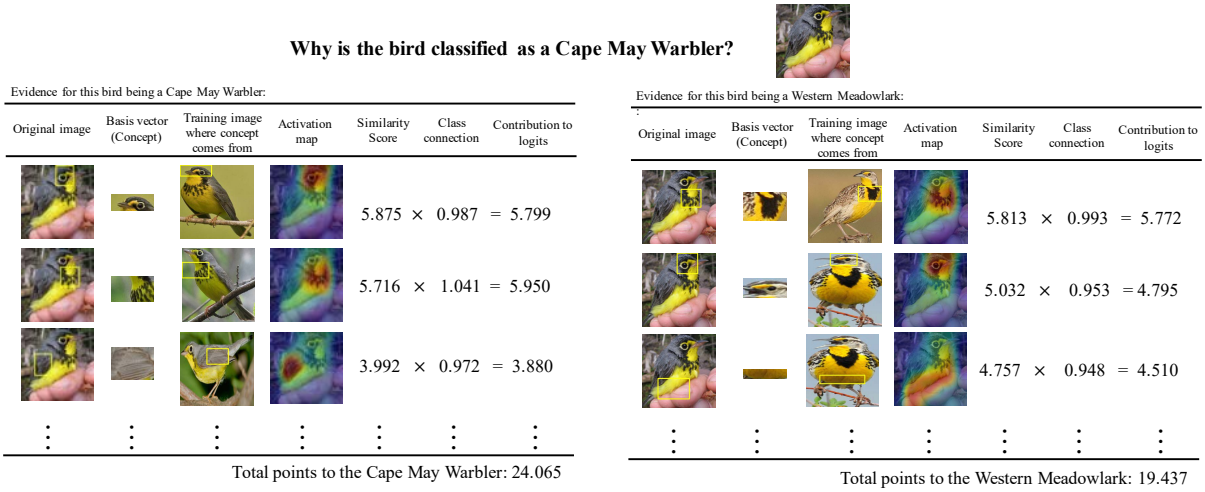
Figure 3. How our TesNet classifies an Evening Grosbeak on different CNN base architectures.



(a) How our TesNet classifies a Rhinoceros Auklet.

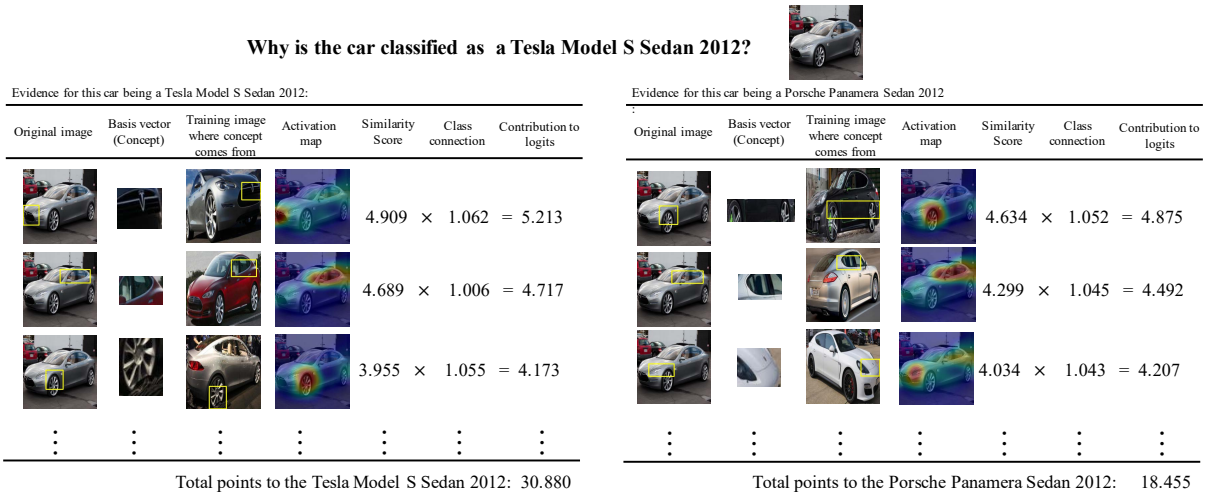


(b) How our TesNet classifies a Red Legged Kittiwake.

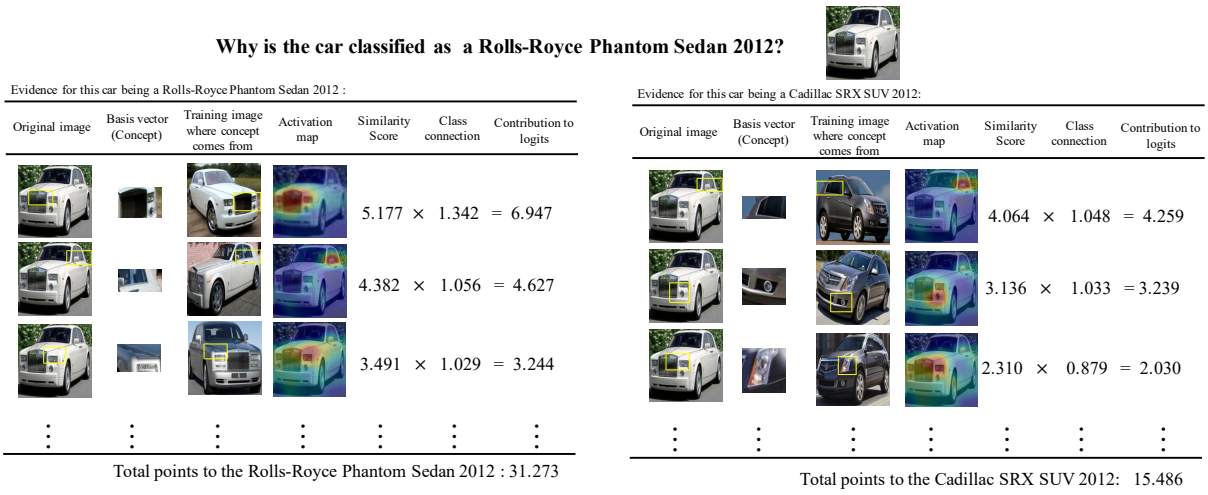


(c) How our TesNet classifies a Cape May Warbler.

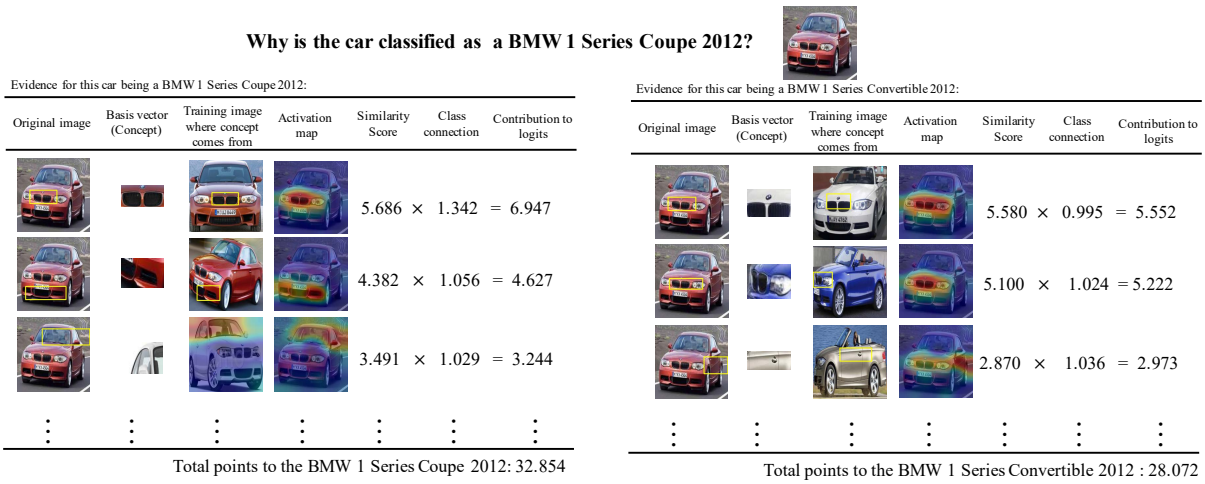
Figure 4. More reasoning processes on CUB-200-2011 produced by TesNet.



(a) How our TesNet classifies a Tesla Model S Sedan 2012.



(b) How our TesNet classifies a Rolls-Royce Phantom Sedan 2012.



(c) How our TesNet classifies a BMW 1 Series Coupe 2012.

Figure 5. More reasoning processes on Stanford Cars produced by TesNet.