

Supplementary Material — Human Attention in Fine-grained Classification

Yao Rong¹

yao.rong@uni-tuebingen.de

Wenjia Xu²

xuwenjia16@mails.ucas.ac.cn

Zeynep Akata^{1,3}

zeynep.akata@uni-tuebingen.de

Enkelejda Kasneci¹

enkelejda.kasneci@uni-tuebingen.de

¹ University of Tübingen, Germany

² University of Chinese Academy of Sciences, Beijing, China

³ Max Planck Institute for Intelligent Systems, Tübingen, Germany

In this document, we provide technical details about our data collection and experiments. First, we explain how we set the standard deviation of the Gaussian distribution in the Human Attention (HA) saliency map generation and show more analyses on gaze data including the relationship between human fixation points and the discriminative attributes of birds. In addition, quantitative and qualitative comparisons between the model explanations (MEs) and HA are demonstrated. In the second section, we introduce implementation details (e.g. sliding window sizes) in the Gaze Augmentation Training (GAT).

1 CUB-GHA

1.1 HA Saliency Map Generation

Figure 1 illustrates a human observing an image on the eye-tracker display. As mentioned in the paper, we post process every fixation location as a Gaussian distribution $N(\mu, \sigma^2)$ on the HA saliency map, where σ is 75 pixels (in the display's resolution). We calculate the standard deviation σ as follows. In our experiment setup, the distance d between the human eye and the eye-tracker display is 60 cm, and the visual angle θ is set to 2° following [4]. In this case, $l = \tan 2^\circ \cdot d = 21$ mm. According to the settings of display, in the horizontal direction the length of the display is 530 mm and the resolution is 1920 pixels. Therefore, we can get that $l = 21$ mm covers approximately 75 pixels on the display. We set 75 pixels as the standard deviation with the image rescaled to the display resolution (1920×1080). The saliency map is rescaled to its original size afterwards.

1.2 Gaze Data Analysis

In this section, we validate that the attributes discovered by our collected human gaze data are discriminative for the fine-grained classification. CUB includes ground-truth attributes for

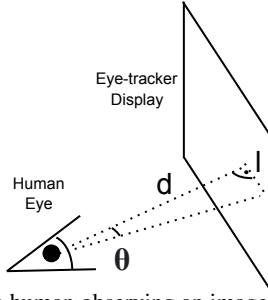


Figure 1: Illustration of a human observing an image on the eye-tracker display.

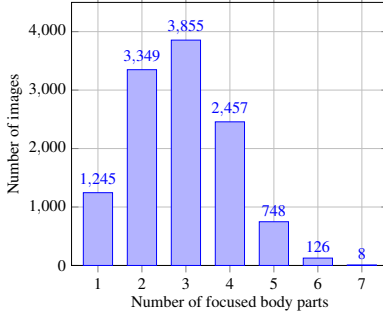


Figure 2: Histogram of the number of focused bird body parts in CUB-GHA. **Y-axis** refers to the amount of images with the certain number of parts (**X-axis**).

each image and they are 312-dimension binary vectors. We use them to conduct the ground-truth discriminative attributes of each bird class in the dataset. There are 100 comparison pairs in the data collection experiments, and we compare each image from the first class with every image in the second class. For instance, if there are M images in the first class and N in the second one, there are in total $M \cdot N$ combinations between the two classes. For each combination, we conduct a comparison attribute vector where 1 is set if that attribute entry is the same for both images, or 0 if not the same, i.e. the comparison attribute vector is also a 312-dimension binary vector. We sum $M \cdot N$ comparison attribute vectors together to have one 312-dimension vector representing the ground-truth discriminative attribute for these two classes. For instance if the attribute `has-wing-color::brown` in the comparison vector is 354, it means that the attribute `has-wing-color::brown` differs in the 354 image pairs. In the end, we group the attributes into seven body parts (head, beak, breast, belly, back, wing, leg). For example, we sum up all the attribute values in the comparison vector that are related to the wing, and the sum represents the difference of the wing between the given two classes. The body part with the highest sum is the most discriminative body part between the two classes.

When our participants look at the image, they always focus on the discriminative body parts of the bird. The body part which human gaze falls in should contain the largest number of different attributes between the compared two classes. With the help of body part center coordinates in each image, we can assign every fixation (collected for this image from five participants) to its nearest body part according to the distance between the center coordinate and the fixation coordinate. In Figure 2, we show the histogram of the number of focused bird body parts on the whole CUB-GHA dataset. We see that there are three body parts

focused by humans in 3855 images. Most of the images (92.52%) include less than five parts focused in the dataset. In very few images, our participants view all seven parts of the bird. In each image, we sum up the duration of fixations belonging to one body part and use it to represent the amount of human attention on that part. A longer duration sum indicates more attention participants have paid. We rank the seven body parts for each image according to the duration sums and calculate the rate that the top- k focused body parts hit the most discriminative one (which is conducted from the ground-truth attributes). The hit rate is shown in Table 1. From the results, we see that our participants discover the most discriminative body part in 84.4% of the images correctly. Within four parts that participants consider to be important for the classification, the ground-truth distinct body part is found in 98.3% of the images. This result shows that human gaze data in CUB-GHA hints on discriminative body parts/attributes in the classification.

Top-k	1	2	3	4
Hit rate (%)	84.40	93.60	97.18	98.31

Table 1: Hit rate of the most discriminative body part. Top- k refers to the k longest focused body parts by humans in CUB-GHA.

1.3 Comparison between ME and HA

In this section, we provide more details and results of comparing MEs and HA. We use the KAR (keep and retrain) procedure [14] and the concrete procedure works as follows: given an input image $I \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 3}$ and the importance estimation map $A \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 1}$, where H and W represent the width and height of input image, respectively; A can be the HA or ME saliency map. We construct a mask $M \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 1}$ to filter the pixels in I . First, we sort A in a descending order to A^R according to the attention values. Then we binarize A by taking the top p percent of pixels in A^R as one and others as zero:

$$M(x, y) = \begin{cases} 1.0, & \text{if } (x, y) \in P \\ 0.0, & \text{otherwise} \end{cases},$$

where P are the indices of top ranked p percent pixels. We apply the mask M to filter the corresponding image I in the training and testing set: $I' = M \odot I$, so that only the top p percent of the most important features are observed by the network. After such a modification of the dataset, we train a new model and compare the test accuracy. This procedure aims at evaluating whether the important feature estimated by A (i.e. model or human attention) is critical to the classification or not. A good estimation A encodes important features in a small amount of pixels. In other words, a higher accuracy with such small amount of pixels indicates that the given features are more important. We generate the new dataset using an insertion percentage $p = [5, 10, 15, 20, 15, 30, 50, 70, 90]$ and train the vanilla ResNet-50 [15] using the same hyper-parameters as in the baseline training. We run this procedure three times independently from random initialization for each estimation map and report the average accuracy on the test set.

Figure 3 illustrates the qualitative results of the modified images using HA and MEs. These differences can be observed when using 5% and 10% as insertion percentages. If we compare HA and MEs, they focus on a similar area after 20% pixels are inserted: the wing

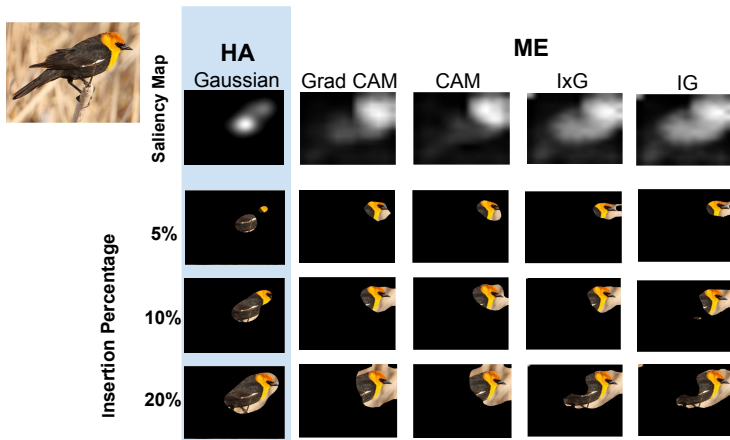


Figure 3: Modified images in the Keep and Retrain procedure. The pixels are inserted according to the importance in the estimation maps. **Top to bottom:** importance estimation maps (saliency maps), modified images using top 5%, 10% and 20% important pixels in saliency maps. **Left to right:** HA, ME-Gradient-based CAM (Grad-CAM) [10], Class Activations Maps (CAM) [11], InputXGradient (IxG) [12] and Integrated Gradients (IG) [13].

and head parts. When comparing among MEs, CAM and IxG are similar to Grad-CAM and IG, respectively. In this example, Grad-CAM/CAM pays more attention to the head, while IG/IxG focuses more on the body. From the qualitative comparison results, we see that the HA and MEs estimate different parts of the bird as being the most important ones for the classification task, especially regarding the first 10% important regions.

We also conduct a quantitative similarity comparison between HA and MEs. We evaluate on different metrics: Kullback-Leibler divergence (KL-D), correlation coefficient (CC) and similarity (SIM), which are often used in comparisons of how similar two images are [14]; rank-correlation (Rank-Co) as introduced in [15]; shuffled AUC metric (sAUC) evaluating every pixel in saliency maps as a classification task; information gain (IG) measuring the performance over a baseline [16, 17]. CAM is very similar to Grad-CAM, e.g. Grad-CAM achieving 0.565 on CC and 1.242 on KL-D, while CAM achieving 0.563 and 1.248, respectively. Additionally, we observe IG and IxG achieving similar performances on these metrics, i.e. 0.699 for IG v.s. 0.694 for IxG on CC, and 1.318 for IG v.s. 1.310 for IxG on KL-D. These similarities can be seen from the qualitative results as well. From all different metrics, we see that the Grad-CAM tends to be the most similar to HA, as Grad-CAM achieves the highest scores in all six metrics. This is consistent with the results from the KAR that Grad-CAM achieves the best performance among all MEs.

	KL-D ↓	CC ↑	SIM ↑	Rank-Co ↑	sAUC ↑	IG ↑
CAM	1.248	0.563	0.399	0.761	0.460	0.938
Grad-CAM	1.242	0.565	0.415	0.761	0.508	1.376
IG	1.318	0.546	0.361	0.699	0.436	0.921
IxG	1.310	0.543	0.375	0.694	0.461	1.001

Table 2: Similarity comparison between MEs and HA saliency map. (↓: the lower the better; ↑: the higher the better.)

GAT Experiments

	Small	Medium	Large
CUB-GHA	(123,134) (134,123) (123,123) (134,134)	(174,190) (190,174) (174,174) (190,190)	(246,264) (269,246)
CXR-Eye	(87,95) (95,87) (95,95) (87,87)	(123,135) (135,123) (123,123) (135,135)	(180,190) (190,180)

Table 3: Sliding window size used in GAT.

Concrete sliding windows sizes (w, h) used for each dataset in GAT experiments are listed in Table 3. For the CUB-GHA dataset, we choose the sliding window sizes based on the averaged size of bird bounding boxes: the width is 246 and the height is 269 if images are resized to 448×448 . Therefore, we use 246 and 269 as sizes for the large scale. The medium window size is conducted using the factor of $\frac{\sqrt{2}}{2}$ to have the half of the bounding box area, i.e. we use 174 and 190 as window size options. The factor used in the small scale is 0.5. For the CXR-Eye dataset, we choose 0.8 and 0.85 as factors with respect to the resized image size 224×224 for the large window size, i.e. two options are 180 and 190. Similarly, factors for the medium window size are 0.55 and 0.6. The small window sizes are scaled based on the medium window sizes by the factor of $\frac{\sqrt{2}}{2}$. The motivation of using different sliding window sizes is to get different parts which are discriminative for the classification. To avoid very similar cropped areas, we choose 0.25 as the iou threshold in the non-maximum suppression. Table 4 lists the ablation study of using different numbers of cropped areas (k) in the augmentation training on two datasets. (2,2,2) denotes that two cropped areas are picked up from each window scale to form the augmentation training set. We choose (2,3,4) as the final setting since it gives relatively better results on both datasets. Figure 4 illustrates the augmentation images using the setting (2,3,4) in three sets of window scales on both datasets.

(L,M,S)	CUB (%)	CXR (%)
(2,2,2)	87.50	71.03
(2,3,2)	88.06	71.58
(2,3,3)	88.00	71.86
(2,3,4)	88.00	72.21

Table 4: Results of using different window size settings on CUB-GHA and CXR-Eye. The number of windows used in large, medium and small size is shown on the left. The accuracy is in %.

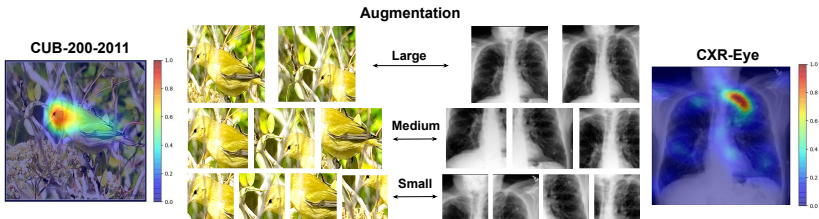


Figure 4: Illustration of cropped images used in the Gaze Augmentation Training. **Left and Right:** HA saliency maps used for augmentation on CUB-GHA and CXR-Eye. **Middle:** cropped images in three scales (large, medium and small).

References

- [1] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *TPAMI*, 2018.
- [2] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 2017.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [4] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *NeuIPs*, 2019.
- [5] LAI Qiuxia, Salman Khan, Yongwei Nie, Sun Hanqiu, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 2020.
- [6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.
- [7] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*. PMLR, 2017.
- [8] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*. PMLR, 2017.
- [9] Joan N Vickers. *Perception, cognition, and decision training: The quiet eye in action*. Human Kinetics, 2007.
- [10] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.