HSI: Human Saliency Imitator for Benchmarking Saliency-Based Model Explanations

Yi Yang[1], **Yueyuan Zheng**[1, 2], **Didan Deng**[1, 3], **Jindi Zhang**[1], **Yongxiang Huang**[1], **Yumeng Yang**[2], **Janet H. Hsiao**[2, 4, 5], **Caleb Chen Cao**[1]

[1] Huawei Research Hong Kong
[2] Department of Psychology, University of Hong Kong
[3] Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology
[4] State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong
[5] Institute of Data Science, University of Hong Kong
yang.yi4@huawei.com, mercuryzheng@connect.hku.hk, ddeng@connect.ust.hk, {zhangjindi2, huang.yongxiang2}@huawei.com, {alicey01, jhsiao}@hku.hk, caleb.cao@huawei.com

Abstract

Model explanations are generated by XAI (explainable AI) methods to help people understand and interpret machine learning models. To study XAI methods from the human perspective, we propose a human-based benchmark dataset, i.e., human saliency benchmark (HSB), for evaluating saliency-based XAI methods. Different from existing human saliency annotations where class-related features are manually and subjectively labeled, this benchmark collects more objective human attention on vision information with a precise eye-tracking device and a novel crowdsourcing experiment. Taking the labor cost of human experiment into consideration, we further explore the potential of utilizing a prediction model trained on HSB to mimic saliency annotating by humans. Hence, a dense prediction problem is formulated, and we propose an encoder-decoder architecture which combines multi-modal and multi-scale features to produce the human saliency maps. Accordingly, a pretraining-finetuning method is designed to address the model training problem. Finally, we arrive at a model trained on HSB named human saliency imitator (HSI). We show, through an extensive evaluation, that HSI can successfully predict human saliency on our HSB dataset, and the HSI-generated human saliency dataset on ImageNet showcases the ability of benchmarking XAI methods both qualitatively and quantitatively.

Introduction

In the past decades, there have been continuing breakthroughs in the capability of machine learning models. The rapidly developed deep neural networks (DNNs) have shown great predictive ability on various learning tasks. However, DNNs are built upon deep structures and non-linear functions, in which information representation and decision rationale are not explicitly observable. As a result, DNN models are criticized for not being understandable, raising concerns about the deployment in critical systems such as medical diagnosis (Bakator and Radosav 2018) and autonomous driving (Grigorescu et al. 2020).

Due to the cognitive gap, explainable artificial intelligence (XAI), which is a discipline focused on bringing interpretability to machine learning, is emerging in recent years (Gunning 2017). Towards the goal of XAI, methods have been designed to generate explanations for reasoning about model behavior. Since the rise of convolutional neural networks (CNNs), numerous saliency-based explanation methods have been proposed for explaining an image classifier's output. As shown in Figure 1, important image pixels/regions for the classification result are identified by popular XAI methods, i.e., Guided Backpropagation (GBP) (Springenberg et al. 2015), Grad-CAM (Selvaraju et al. 2017) and RISE (Petsiuk, Das, and Saenko 2018).

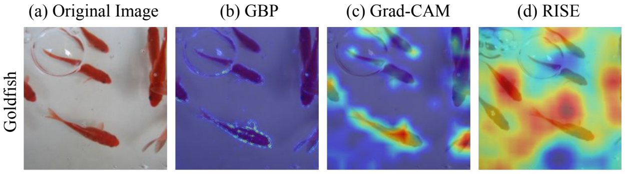


Figure 1: (a) A sample image that is classified as 'goldfish'. (b-d) Saliency-based explanations of the listed XAI methods. Best viewed in color.

As different XAI methods are emerging, a question arises: how do we assess the quality of model explanations? Different dimensions to design XAI metrics, e.g., human trust and task performance, have been proposed in recent research of cognitive science (Hoffman et al. 2018; Hsiao et al. 2021d). However, practical experiments for these metrics are highly labor-intensive, as the judgements or states of human subjects need to be measured for each explanation. A recent work (Mohseni, Block, and Ragan 2021) proposed to collect human-grounded important features and quantitatively benchmark XAI methods. The intuition of this benchmarking idea is that good explainers should identify features of importance with similar strategies to humans.

In this paper, we propose to use eye-tracking techniques to measure human attention on image data features as the grounded saliency map. Research in cognitive science has shown that eye movement can reflect humans' underlying cognition objectively. Our preliminary study verifies that our

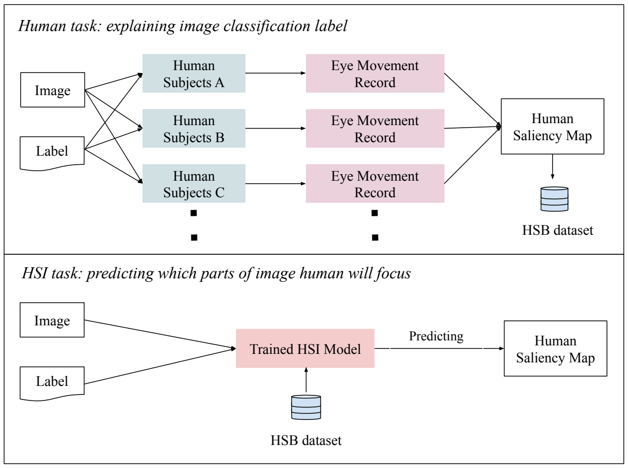


Figure 2: Overview of human task for benchmark dataset annotation and HSI task for predicting human attention.

data collection method and user task design can obtain annotations relevant to the important features for the task of image classification. We construct a human saliency benchmark, HSB, the first eye-tracking based human attention dataset for the evaluation of model explanations and make it publicly available for the related research[1]. Moreover, we propose to model the generation of human saliency annotation in the eye-tracking experiments. As a possible replacement of real human subjects to annotate new data, this model is aimed to provide a trade-off between the cost of crowdsourcing experiments and the accuracy of annotations.

As illustrated in Figure 2, our modeling task aligns with the human task in the eye-tracking experiment. We formulate a new dense prediction task, where the input is multimodal (image and text label) information and the output is an input-sized fixation map showing the spatial distribution of human attention. In spite of that the encoder-decoder architecture has been proven to be effective in dense prediction tasks, including saliency prediction, they mainly focus on learning the relationship between visual features and output. The problem of exploiting multimodal information and generating label-specific human attention is barely studied.

To accomplish the new task, we adapt a vision-language pre-training model CLIP (Radford et al. 2021) to encode the multimodal input. Inspired by (Rao et al. 2022), the image and text are embedded separately with the pre-trained encoders, and a pixel-text alignment is calculated at several stages of embedding. The alignment fuses the visual and textual features. We then fuse multi-level images feature maps with the pixel-text alignment and feed them to a simple decoder to predict human saliency maps.

Conventionally, pre-training and fine-tuning strategies are widely used for the attention prediction task due to the scarcity of eye-tracking data. We consider pre-training our model on SALICON (Jiang et al. 2015) dataset, which is the largest eye movement dataset containing human attention on the images from a large-scale object detection dataset, COCO dataset (Lin et al. 2014b). However, the collection of the SALICON dataset involves passive viewing, i.e., asking subjects freely viewing the pictures, instead of goal-directed viewing, such as image classification and explanation in our dataset collection. To better utilize the SALICON dataset and the multimodal correlations, we propose to construct pseudo-labelled fixations/saliency maps on the dataset for model to learn the knowledge of label-specific attention in the pre-training stage. We also tailor our loss functions based on a mathematical model to learn both of them.

To prove the effectiveness of our model, we design extensive evaluations. First, we evaluate the fine-tuned model on our benchmark dataset and obtain high prediction accuracy. Second, we test the model on the ImageNet validation set, and the results show that the generated human attention maps have good visual quality. We also conduct an XAI evaluation study by comparing similarity between the saliency maps generated by HSI and XAI methods, and find that RISE (Petsiuk, Das, and Saenko 2018) ranks the first place among the three XAI methods.

The main contributions of our paper are as follows:

- We collect, to the best of our knowledge, the first attention-based human saliency benchmark for evaluating XAI from the cognitive science perspective: whether the saliency maps generated by XAI explainers are close to human saliency maps for explanation.
- We propose a model architecture which utilizes the image-vision multi-modal encoder, CLIP, to better align the human attentions with the location of the objects intended for classification.
- We design a novel pretraining-finetuning method to address the scarcity of eye-tracking data in training the prediction model for human saliency annotation.

Related Work

In this section, we first introduce representative saliency-based XAI methods. Next, we briefly review current human attention datasets and their applications. Lastly, we review and compare human attention prediction models with ours.

Saliency-Based Model Explanations

Different strategies are employed to generate model explanations. They can be grouped into two categories. **Backpropagation**: they calculate gradients and assume that important regions are at locations with high gradient magnitude. **Perturbation**: they perturb the input samples and assume that the occlusion of important regions will result in drop of output probability. In this paper, we select three representative methods to evaluate.

Guided Backpropagation (GBP): The vanilla backpropagation method generates noisy saliency maps (Simonyan, Vedaldi, and Zisserman 2014). Therefore, GBP sets negative gradient entries to zero while backpropagating the output through a ReLU unit (Springenberg et al. 2015). The

[1] The HSB datasets including the original images, human attention saliency mask, and human attention saliency overlay can be found here: https://OSF.IO/F3BAW/.

saliency map mostly captures important edges in images (Figure 1b).

Grad-CAM: It uses the class-specific gradient in the last convolutional layer of a CNN model. The method is a generalization of Class Activation Mapping (CAM) (Zhou et al. 2016) and can be applied to CNN models without a global average pooling layer (Selvaraju et al. 2017). The saliency map focuses more on important regions rather than edges (Figure 1c).

RISE: The input image is masked randomly and fed into the model to get a prediction. The final saliency map is the weighted sum of all random masks, with the corresponding prediction scores as weights (Petsiuk, Das, and Saenko 2018). The saliency map is smooth but irregular noises may exist due to the distribution bias of masks (Figure 1d).

Human Attention Datasets and Applications

Previous studies have commonly used passive viewing paradigms to collect human attention data for human attention prediction (e.g., Wang and Shen 2017; Borji and Itti 2015; Jiang et al. 2015). More recent studies have also developed task-driven human attention datasets for training AI models to predict human attention in similar tasks, such as DR(eye)VE dataset with eye gaze data in naturalistic driving settings (Alletto et al. 2016) and CUB-VWSW dataset for image classification (Karessli et al. 2017). Zheng and his colleagues (Zheng et al. 2018) further showed that training an AI model with both task-specific and task-free human attention could achieve state-of-art prediction performance.

In addition to prediction purposes, human attention data have been used to enhance AI models' performance (Lai et al. 2020). For example, in more fine-grained classification tasks, using human attention in the same task during model training is shown to facilitate the focus on discriminative regions for the task and boost the models' performance (Rong et al. 2021).

Human attention data also play an important role in XAI research, in particular to be compared with saliency-based XAI for evaluation purposes. For example, Hwu et al. (2021) compared heatmap-based XAI methods, LRP, with human attention data and reported a higher similarity to task-driven attentive human attention than inattentive attention. Note however that cogntive science research has consistently shown that human attention during image viewing are both task-specific (e.g., Borji and Itti 2015; Kanan et al. 2015) and person-specific (e.g., Hsiao et al. 2021a; An and Hsiao 2021; Hsiao et al. 2021b; Hsiao, Liao, and Tso 2022). As the purpose of XAI is to provide explanations, comparing saliency-based XAI with human attention that are associated with better performance during an explanation task will provide more insights into the quality of XAI.

Human Attention Prediction

In the literature, the computational model for human attention prediction is tackled by saliency models. These models predict the distribution of human fixations in the form of a saliency map, where a brighter pixel value indicates it has higher probability of gaining human attention.

Feature representation of the input image is an essential problem for saliency models. Early attempts adopted handcrafted features (e.g., intensity, color, and edge orientation) to present images. With the development of deep learning, different model structures have been proposed to improve the capabilities of feature representation. For example, (Wang and Shen 2018; Kümmerer et al. 2017; Reddy et al. 2020; Kroner et al. 2020) explored the combination of multi-resolution features, (Cornia et al. 2018) applied recurrent architecture and (Lou et al. 2022) integrated transformer to refine the learnt features.

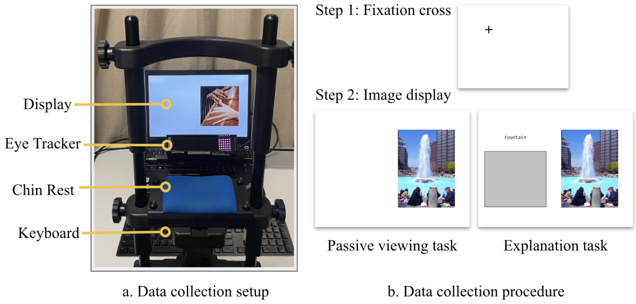
Despite the models are increasingly complex, none of the previous work explored the label-specific prediction task and the integration of text information into image features. Our model differentiates itself from existing methods by leveraging a recent vision-language pre-training model to encode and combine multi-modal input features.

Eye Tracking Experiment and Dataset

Pilot Study: Passive Viewing vs. Explanation Tasks

Here we examined whether human eye movement patterns differ between passive viewing and explanation tasks in viewing images with a single foreground object category.

Methods 20 participants (17 females), aged 18-31 years (M = 23.18, SD = 4.04), were recruited from a local university to complete an explanation task and a passive viewing task in two separate sessions (Figure 3). All participants had normal or corrected-to-normal vision. During the first session, they freely viewed 160 images one at a time, each for 5 seconds. During the second session, they were presented with the same 160 images one at a time together with a category label and asked to type explanations on why the foreground object could be labeled with the given category in a textbox. Their dominant eye was tracked during both sessions. The 160 images were from 20 object classes, including ant, zebra, horse, lion, jellyfish, snail, lemon, mushroom,



a. Data collection setup

b. Data collection procedure

Figure 3: (a) Eye tracker set-up: We used an EyeLink Portable Duo eye tracker (SR Research) to record eye movements. (b) Data collection procedure: Step 1 shows the presentation of a fixation cross, where a class label would appear in the explanation task. Step 2 shows the display for passive viewing and explanation tasks respectively. In the explanation task, participants typed their explanations in the gray textbox.

corn, pizza, laptop, cellphone, microphone, sofa, broom, shovel, umbrella, harp, tennis ball, and fountain. Each class had 8 images. Images of horse and sofa were obtained from PASCAL VOC, while the rest 144 images were obtained from ImageNet. The selected classes were from human basic level categories (Markman and Wisniewski 1997) and also commonly used as output categories of image classification AI models. The selected images constituted a representative set with different levels of foreground object complexity and background saliency. All images were resized to fit into a 400 × 520 pixel frame on a blank canvas for image presentation and data analysis. The experiment was conducted using E-Prime Extensions for EyeLink on a 255 mm × 195 mm laptop with 1024 × 768 pixel resolution. The images were presented in 9.51 degrees of visual angle (dva) horizontally and 12.36 dva vertically on the screen. A chinrest was placed at a 60 cm viewing distance to minimize head movement.

Results Eye fixations on the image area during the tasks were analyzed using Eye Movement analysis with Hidden Markov Models (EMHMM) (Chuk, Chan, and Hsiao 2014) with co-clustering (Hsiao et al. 2021c), since it allowed us to quantify eye movement pattern similarities among individuals or conditions across images with varying layouts. Specifically, a participant's eye movements for each image when doing a task were summarized using an HMM, with person-specific regions of interest (ROIs) and transition probabilities among the ROIs. In each HMM, the optimal number of ROIs (within a preset range 1 to 10) was determined using the variational Bayesian method. Thus, for each image, there were 40 different eye movement patterns/HMMs, corresponding to 40 different participant-task combinations. Co-clustering clustered these participant-task combinations into two groups such that the combinations in each group had similar eye movement patterns to one another across the 160 images. In each group, a representative HMM was generated for each image (with the number of ROIs set to be the median number of ROIs among the individual HMMs), resulting in two representative eye movement pattern groups derived from these 40 participant-task combinations. As shown in Figure 4, the pattern group on the top, referred to as the explorative pattern group, had large ROIs with eye fixations widely distributed on different features. In contrast, the pattern group on the bottom, referred to as the focused pattern group, had small ROIs focusing on a few key features. The two pattern groups differed significantly: Data from the explorative group were significantly more likely to be generated by the representative explorative HMMs than the representative focused HMMs, t(27) = 8.478, p $<$ .001, d = 1.602, and vice versa for data from the focused group, t(11) = 8.499, p $<$ .001, d = 2.453 (Chuk, Chan, and Hsiao 2014). Following previous studies (e.g., Chan et al. 2018; Zheng, Ye, and Hsiao 2022), we quantified each participant's eye movements in each task along the dimension contrasting the two representative pattern groups using EF scale, defined as $(E - F)/(|E| + |F|)$, where E and F represent the log-likelihood of the participant's eye movement data being generated by the explorative and focused pattern group respectively. A more positive EF scale indicates a higher similarity to the explorative pattern group.

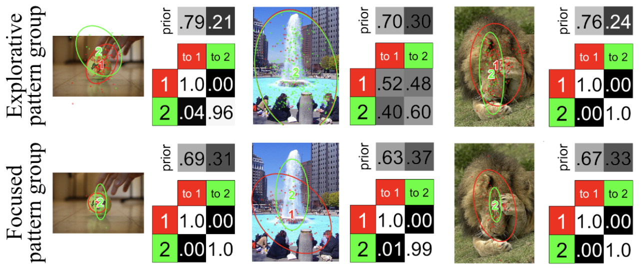


Figure 4: Pilot study: Explorative and focused pattern groups resulting from co-clustering with example images and HMMs. Ellipses show ROIs as 2-D Gaussian emissions. The table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. The image shows the ROI assignments of the raw fixations. Best viewed in color.

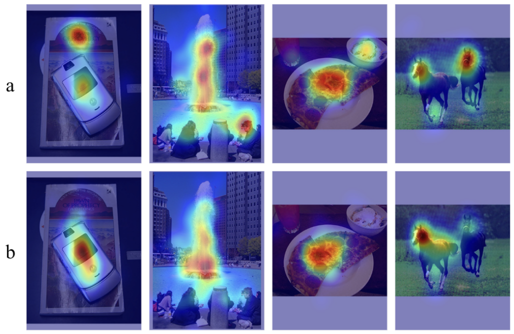


Figure 5: Saliency map (with a Gaussian distribution with SD equivalent to 0.5 dva or 21 pixels applied on each fixation) comparisons between (a) passive viewing and (b) explanation tasks. Best viewed in color.

When comparing participants' EF scale between the two tasks, we found that their eye movements were more similar to the focused pattern group, t(19) = -5.01, p $<$ .001, d = -1.12, in the explanation task than the passive viewing task. In other words, human attention differed significantly between the two tasks (Figure 5). Accordingly, transfer learning may be necessary for generating explanation-oriented XAI benchmark based on a large-scale passive viewing dataset.

HSB Dataset from Human Explanation Behavior

Here we aimed to collect human attention data from an explanation task to generate an XAI benchmark with eye movement patterns that are associated with better human explanation performance. We recruited additional 42 participants from the University of Hong Kong to complete the explanation task in Human study 1, resulting in a total of 62 participants (52 females), aged 18 to 37 (M = 22.5, SD = 3.8). Their English proficiency was assessed using a standardized test LaxTale (Lemhöfer and Broersma 2012).

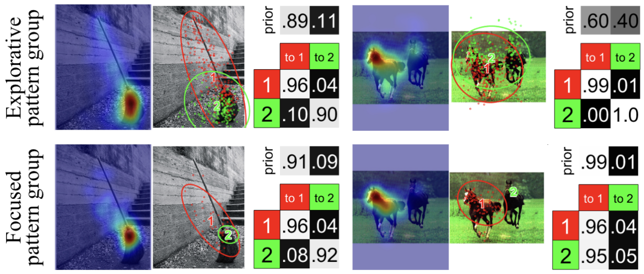


Figure 6: HSB dataset: Explorative and focused pattern groups resulting from co-clustering with example images, saliency map and HMMs. For each image, the figure on the left shows the saliency map with a Gaussian distribution (SD = 0.5 dva) applied on each fixation. The figure on the right shows the HMM model with the ROI assignments of the raw fixations. Ellipses show ROIs and the table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. Best viewed in color.

Results EMHMM with co-clustering on the explanation eye movement data showed similar explorative and focused pattern groups (Figure 6). The two groups differed significantly: data from the explorative pattern group were more likely to be generated by the representative explorative HMMs, t(46) = 15.47, p $<$ .001, d = 2.26, and vice versa for those from the focused group, t(14) = 6.94, p $<$ .001, d = 1.79. Each participant's eye movement pattern in the task was then quantified in EF scale. The quality of their explanations was evaluated by two data scientists with expertise in computer vision using a 7-point Likert scale. The two raters had a good inter-rater reliability (Cronbach's alpha = .858). The average rating was used as participants' explanation performance. Partial correlation analysis controlling for English proficiency showed a significant correlation between EF scale and explanation performance, r(59) = .48, p $<$ .001, indicating that participants with a more explorative eye movement pattern had better explanation performance. Accordingly, here we used fixations from the explorative pattern group to produce human attention benchmark for XAI.

HSI: Computational Model

Our HSI model is aimed to solve a new dense prediction task, where the input is multi-modal (image and text label) information. In this section, we elaborate the design of our model architecture and the training method for utilizing the knowledge of a large-scale human attention dataset.

Architecture

Our model takes the encoder-decoder structure which is common for dense prediction. The overall architecture is shown in Figure 7.

Multi-Modal Encoder We refer to a recent contrastive language-image pre-training model CLIP (Radford et al. 2021) to encode text and image information. Through pre-training on 400 million image-text pairs with constrastive objectives, the embedding spaces of visual and language are aligned for its encoders.

CLIP consists of two alternative image encoders, i.e., ResNet-50 (He et al. 2016) or ViT (Dosovitskiy et al. 2021). Here we leverage the ResNet-based encoder (denoted as G_{image}) to extract multi-resolution image features. Given an input image, we take feature maps at the last two stages of the encoder, which are denoted as $\mathbf{u_4}$ and $\mathbf{u_5}$. They are the outputs of the fourth and the fifth convolutional blocks. Later, we down-sample $\mathbf{u_4}$ to $\mathbf{u_4}'$ by the factor of two, so that $\mathbf{u_4}'$ can be concatenated with other feature maps.

Originally, CLIP outputs an embedding $\bar{z} \in \mathbb{R}^{1024}$ which aligns with the language embedding space. Instead, we extract the feature map from another output of the final attention pooling layer $\mathbf{z} \in \mathbb{R}^{1024\times7\times7}$, as it retains spatial information and can be regarded as a language-compatible feature map (Rao et al. 2022).

For text encoder G_{text} , the input label l is first formulated to $T(l)$ as is suggest by (Radford et al. 2021). T is a default prompt template as "`a photo of a [`l`]`". The encoder consists of a Transformer (Vaswani et al. 2017) and outputs text embedding $s \in \mathbb{R}^{1024}$. Note that G_{text} is fixed during our whole experiments.

Furthermore, we use a score map calculated from the low-resolution image embedding $\mathbf{z}$ and the text embedding s to represent the relations between image pixel information and the text information. Note that $\mathbf{z}$ consists of 49 pixels, and each pixel has an embedding $z_{i,j} \in \mathbb{R}^{1024}$. The inner product between the normalized $z_{i,j}$ and the normalized s represents the cosine similarity between the image pixel at (i, j) location and the text. We denote it as $a_{i,j}$.

$$a_{i,j} = \langle \frac{z_{i,j}}{\|z_{i,j}\|}, \frac{s}{\|s\|} \rangle. \tag{1}$$

We use $\mathbf{a}$ to denote the score map where the element at (i, j) location is $a_{i,j}$. $\mathbf{a} = \mathcal{I}(\mathbf{z}, s)$, where $\mathcal{I}$ denotes the function in Equation 1.

Finally, we merge all the feature maps (i.e., $\mathbf{u_4}'$, $\mathbf{u_5}$, $\mathbf{z}$) and the score map $\mathbf{a}$, by passing them through a concatenation layer at the end of the multi-modal encoder.

Decoder We employ a minimal decoder architecture with only convolutional layers. $G_{decoder}$ consists of a convolutional layer with 3×3 kernels , a convolutional layer with 1×1 kernels and a sigmoid function. We use bilinear upsampling to produce the final output $\mathbf{p} \in \mathbb{R}^{1\times224\times224}$.

Training Method

Given the same image, we think a person will generate distinct saliency maps when performing two different tasks: one is passive viewing, another is explanatory classification task specified to an object in this image. The gaze map in the latter case is conditioned on the object, and may vary when the conditioned object changes. Therefore, we propose the marginal saliency map for the first task (passive viewing) and the conditional saliency map for the second task (object

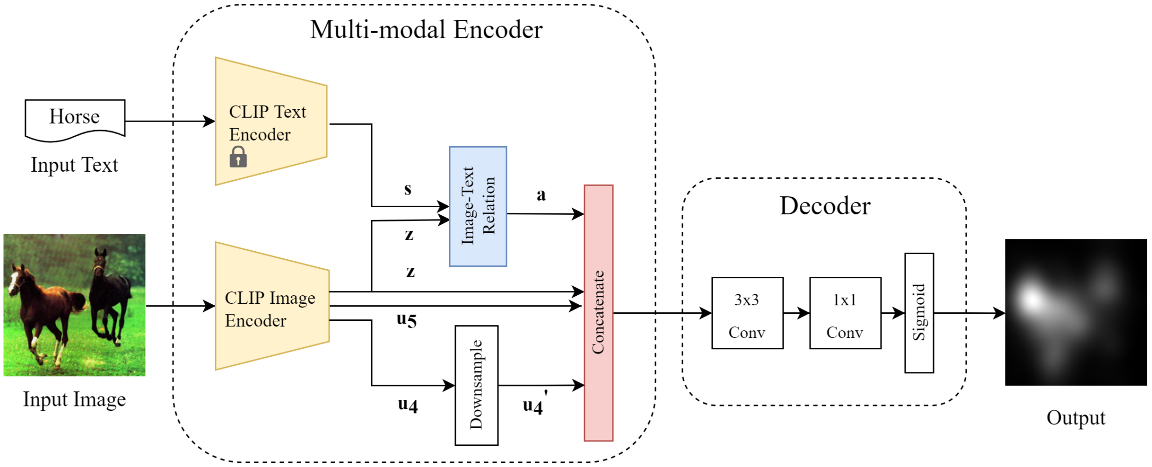


Figure 7: Overview of our model architecture. A multi-modal encoder is designed to extract multi-resolution visual information and text information. A decoder network takes the concatenated feature maps to produce predicted human saliency map.

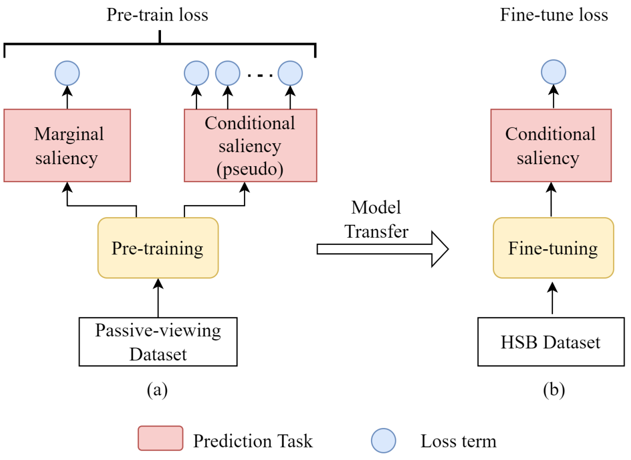


Figure 8: Overview of the training pipeline.

classification). We also analyse their relations and tailor our loss functions to learn both of them.

The pipeline of our approach is shown in Figure 8. First, we pre-train our model on the passive-viewing saliency dataset. Although the passive-viewing dataset does not have conditional saliency map ground truth, we can create pseudo labels, so that our model can predict both the marginal and the conditional saliency maps. After pretraining, we finetune our model on the collected dataset in order to make the predictions of conditional saliency map closer to real human attentions.

Conditional vs. Marginal Saliency Map We suppose an input image contains N objects that may attract human attentions. We denote each object as obj_n and the set of objects in this image is denoted as $\mathcal{T}_{obj} = \{obj_n\}_{n=1}^{N}$. The marginal saliency map is the saliency when human are freely viewing all the objects in the image. We denote it as $P(S|x_{i,j})$, which means the probability of human attention assigned to this pixel located at (i, j). In the contrast, the conditional saliency map is the saliency conditioned on a certain object: $P(S|obj_n, x_{i,j})$. It can be interpreted as the probability of human attention assigned to the pixel at (i, j) location when the target object is obj_n. We show that the marginal saliency map be can expressed as the summation of conditional saliency maps multiplied with $P(obj_n|x_{i,j})$:

$$P(S|x_{i,j}) = \sum_{obj_n \in \mathcal{T}_{obj}} P(S|obj_n, x_{i,j})P(obj_n|x_{i,j}). \tag{2}$$

$P(obj_n|x_{i,j})$ is the probability of finding obj_n at the (i, j) pixel location. Although we do not have the ground truth of $P(obj_n|x_{i,j})$, we think it is reasonable to use $\mathbf{a}$ to derive a proxy for $P(obj_n|x_{i,j})$. This is because $\mathbf{a}$ represents the similarity of the text embedding (the name of the object) and the image embedding at (i, j) location. The proxy for $P(obj_n|x_{i,j})$ is represented by $\tilde{P}(obj_n|x_{i,j})$.

$$\tilde{P}(obj_n|x_{i,j}) = Up[\sigma(\mathcal{I}(\mathbf{z}, \mathbf{s}))_n \times \sigma(\mathcal{I}(\mathbf{z}, s_n))_{i,j}] \tag{3}$$

$\mathbf{z} \in \mathbb{R}^{1024\times7\times7}$ is the image feature map, s_n is the text embedding of obj_n. $\mathcal{I}(\mathbf{z}, s_n))_{i,j}$ is defined as the score map between the image feature map and the text embedding at (i, j) location.

$\sigma(\cdot)$ is the softmax function. $\sigma(\mathcal{I}(\mathbf{z}, \mathbf{s}))_n = \frac{exp(\mathcal{I}(\mathbf{z},s_n))}{\sum_{t=1}^{N} exp(\mathcal{I}(\mathbf{z},s_t))}$ is the softmax function applied on the dimension of the number of objects. $\sigma(\mathcal{I}(\mathbf{z}, s_n))_{i,j} = \frac{exp(\mathcal{I}(\mathbf{z},s_n))_{i,j})}{\sum_i \sum_j exp(\mathcal{I}(\mathbf{z},s_n))_{i,j})}$ is the softmax function applied on the dimension of the number of pixels.

$Up[\cdot]$ is the bilinear upsampling function in order to match the tensor size for element-wise multiplication.

For the conditional saliency map $P(obj_n|x_{i,j})$, we do not have the ground truth in the pretraining stage. However, we propose a pseudo-label construction method to generate the saliency maps conditioned on every objects in an image. The pseudo-label construction method will be elaborated in the next section.

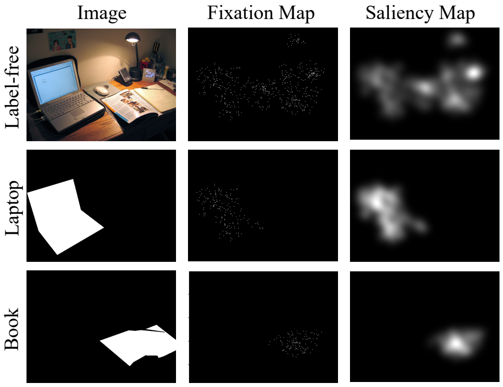


Figure 9: Examples from the constructed dataset.

Pre-train Dataset Construction We aim to construct a pseudo-labeled dataset from a passive-viewing dataset named SALICON (Jiang et al. 2015). SALICON is currently the largest public dataset for passive-viewing eye movement data, containing 10,000 training images, 5,000 validation images and 5,000 testing images. For each image, SALICON provides a pair of fixation map and a saliency map annotation. Due to the large volumn of annotated data, SALICON dataset has been widely employed to pre-train DNN models for saliency predictions.

The images in SALICON are taken from the Microsoft COCO (MS COCO) dataset (Lin et al. 2014a), which contains rich contextual information. It consists of object annotations (bounding box and super-pixel segmentation) information for 80 object categories.

A simple method (Algorithm 1) is developed to generate the pseudo fixation and saliency maps on the objects using the segmentation information. We denote the ground truth of passive viewing saliency maps and corresponding fixations as Q_{pv} and F_{pv}. Suppose the ground truth of object classification saliency map and corresponding fixations are Q_{obj_n} and F_{obj_n}, we denoted the pseudo saliency/fixations as $\tilde{Q}_{obj_n}$ and $\tilde{F}_{obj_n}$. For the threshold parameters, we set $T_a = 0.1$, $T_r = 0.05$ based on empirical study. The sigma for gaussian filter is set to 19, which is same to the original SALICON dataset. Figure 9 shows an original image in SALICON and our generated pseudo fixation and saliency maps for each object in the image.

Loss Function For dense prediction task, the loss function compares the output with the ground truth. Recent studies (Cornia et al. 2018) show that a combination of saliency evaluation metrics in loss function can improve model performance on visual saliency prediction.

Inspired by Reddy et al. (2020); Jia and Bruce (2020), we adopt three most popular metrics, i.e., Kullback-Leibler Divergence (KLdiv), Linear Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS) to construct our loss function. Denote a predicted saliency map as P, the ground truth of human saliency map as Q, and the ground truth fixation map only contains binary values as F. The overall loss function is combined as follows:

$$\begin{aligned} L(P,Q,F) =& \lambda_1 L_{KLdiv}(P,Q) + \lambda_2 L_{CC}(P,Q) \\ &+ \lambda_3 L_{NSS}(P,F) \end{aligned} \tag{4}$$

where λ_1, λ_2, λ_3 are the weights of each metric.

KLdiv measures the dissimlarity between two distributions, i.e., the predicted and ground-truth saliency map in this case:

$$L_{KLdiv}(P,Q) = \sum_i Q_i log(\epsilon + \frac{Q_i}{P_i + \epsilon}) \tag{5}$$

where i denotes the location of pixels in a saliency map and ϵ is a regularization term. The KLdiv is a dissimilarity metric and a lower value indicates a better result.

CC is the Pearson's correlation coefficient and treat the saliency maps as random variables. Formally,

$$L_{CC}(P,Q) = \frac{cov(P,Q)}{cov(P) \times cov(Q))} \tag{6}$$

where $cov(P,Q)$ denotes the covariance of P and Q.

NSS computes the average of the normalized saliency values at each fixation location. The metric is defined by:

$$L_{NSS}(P,F) = \frac{1}{N} \sum_i P_i \times F_i \tag{7}$$

where $N = \sum_i F_i$ and $P = \frac{P-\mu(P)}{\sigma(P)}$. $\mu(P)$ and $\sigma(P)$ respectively represent the mean and standard deviation of the predicted saliency map.

Pretraining and Finetuning In our pretraining stage, given a single image sample, we have the ground truth $\mathbf{q}_{pv}, \mathbf{f}_{pv}$ and pseudo labels $\tilde{\mathbf{q}}_{obj_n}, \tilde{\mathbf{f}}_{obj_n}, \forall n \in [1, N]$.

We take the output of decoder $\mathbf{p}$ and treat it as the prediction of the conditional saliency map. The first term of loss function in the pretraining stage is for condition saliency map prediction:

$$L_1 = \frac{1}{N} \sum_n L(\mathbf{p}_n, \tilde{\mathbf{q}}_{obj_n}, \tilde{\mathbf{f}}_{obj_n}) \tag{8}$$

N is the total number of objects we consider in the input image. $\mathbf{p}_n$ is the conditional saliency map prediction when the object is obj_n.

The second term of loss function in the pre-training stage is for marginal saliency map prediction. We first obtain the predicted marginal saliency map as $\mathcal{H}(\mathbf{p})$.

$$\mathcal{H}(\mathbf{p}) = \sum_n \mathbf{p}_n \times \tilde{P}(obj_n). \tag{9}$$

This expression is derived from Equation 2 and Equation 3. We use $\mathcal{H}(\mathbf{p})$ and the ground truth labels to calculate the second term of the loss function:

$$L_2 = L(\mathcal{H}(\mathbf{p}), \mathbf{q}_{pv}, \mathbf{f}_{pv}) \tag{10}$$

The overall loss function we use for pretraining is:

$$L_{pretrain} = L_1 + L_2 \tag{11}$$

Algorithm 1: Pseudo Saliency/Fixations Construction

Input: image $\mathbf{i}$, binary fixation map $\mathbf{f}_{pv}$, segmentation binary masks $\{\mathbf{m}_{obj_n}\}_{n=1}^{N}$ for N objects in image $\mathbf{i}$.
Parameter: threshold for ratio of area size with attention T_a, threshold for ratio of fixations T_r.
Output: pseudo labeled saliency maps $\mathcal{Q} = \{\tilde{\mathbf{q}}_{obj_n}\}$, pseudo fixation maps $\mathcal{F} = \{\tilde{\mathbf{f}}_{obj_n}\}$.

```
create two empty sets: $\mathcal{Q}$, $\mathcal{F}$.
for $n = 1$ to $N$ do
    Find $\tilde{\mathbf{f}}_{obj_n} = \mathbf{m}_{obj_n} \cdot \mathbf{f}_{pv}$
    $\gamma$ = the area of $\mathbf{m}_{obj_n}$.
    $\eta = sum(\mathbf{f}_{pv})$.
    if $\frac{1}{\gamma} \cdot sum(\mathbf{m}_{obj_n}) \geq T_a$ and $\frac{1}{\eta} \cdot \tilde{\mathbf{f}}_{obj_n} \geq T_r$ then
        APPEND $\tilde{\mathbf{f}}_{obj_n}$ to $\mathcal{F}$
        $\tilde{\mathbf{q}}_{obj_n} = Guassian\_filter(\tilde{\mathbf{f}}_{obj_n}, sigma)$
        APPEND $\tilde{\mathbf{q}}_{obj_n}$ to $\mathcal{Q}$
    end if
end for
return $\mathcal{Q}$, $\mathcal{F}$
```

In the finetuning stage, we are no longer interested in predicting the marginal saliency map. We have the ground truth labels of object classification in the finetuning stage, and we denote them as $\mathbf{q}_{obj_1}, \mathbf{f}_{obj_1}$ since we only have one object label for one image in finetuning stage. The loss function we use in the finetuning stage is:

$$L_{finetune} = L(\mathbf{p}_1, \mathbf{q}_{obj_1}, \mathbf{f}_{obj_1}) \tag{12}$$

Experiments and Results

Dataset

HSB We perform model evaluation on our HSB dataset. Due to the scarcity of data, we split the image samples into 5 folds and perform cross-validation. To balance the data distribution, we require each fold to contain at least one sample for each class.

Experimental Setup

The optimizer we used is Adam (Kingma and Ba 2015). The learning rate was initially $1e^{-4}$. For pretraining, we trained the model for 20 epochs, and decreased the learning rate by a factor of 10 after every 10 epochs. After pretraining, we selected the model weights performing the best on the validation set of the pretraining dataset, and used it as the initial weights for our finetuning in every fold.

In the finetuning stage, we performed five-fold cross validation, For each fold, we loaded the best model weights from pretraining stages as the initial weights. We then trained the model for 20 epochs, and decreased the learning rate by a factor of 10 after every 10 epochs. We took the model from the final epoch and evaluated its performance on the validation set of each fold.

In both pretraining and finetuing, we chose the weight of each loss term in Equation 4 as $\lambda_1 = 1, \lambda_2 = -1, \lambda_3 = -1$.

Metric

The evaluation metrics we used were KL Divergence (**KLdiv**), Pearson's correlation coefficient **CC**, Normalized Scanpath Saliency (**NSS**), and Similarity (**SIM**). The similarity metric is computed from the predicted saliency map and the ground truth saliency map. **SIM** is defined by:

$$SIM(P, Q) = \frac{1}{W \times H} \times \sum_{i}^{W} \sum_{j}^{H} min\left(\frac{p'_{i,j}}{\sum_i \sum_j p'_{i,j}}, \frac{q'_{i,j}}{\sum_i \sum_j q'_{i,j}}\right),$$

where $\mathbf{p}' = \frac{\mathbf{p} - min(\mathbf{p})}{max(\mathbf{p}) - min(\mathbf{p})}$, and $\mathbf{q}' = \frac{\mathbf{q} - min(\mathbf{q})}{max(\mathbf{q}) - min(\mathbf{q})}$.

Results

Qualitative Evaluation Figure 10 shows some qualitative examples of saliency maps predicted by our HSI model with respect to the ground-truth human saliency annotations. Overall, HSI presents similar distribution of saliency patterns to real humans in our explanation task.

With and Without Finetuning We first pretrained our model on the pseudo labels and ground truth labels of the SALICON dataset. Then, we conducted two experiments: with and without finetuning, before the five-fold cross validation. The evaluation metrics are shown in Table 1.

Finetuning	CC	KLdiv	NSS	SIM
w.	**0.8819**	**0.2162**	**2.3505**	**0.7563**
w.o	0.7317	0.5168	1.9357	0.6012

Table 1: The five-fold cross validation metrics (average values are reported). w. represents the experiment with finetuning. w.o. represents the experiment without finetuning.

From Table 1, we find that with finetuning, our model performed better on every metric compared with model without finetuning. This indicates that there is a gap between the SALICON dataset passive-viewing saliency maps and the saliency maps in HSB dataset. In addition, our constructed pseudo labels may not mimic the the saliency maps in HSB dataset perfectly, thus finetuning is necessary.

Ablation Study

In the ablation study, we considered the usage of image-text relations and the necessity of pretraining.

Image-Text Relations In our first experiment, we compared a baseline model which only consists of the image encoder G_{image} and the decoder $G_{decoder}$. The inputs to the decoder of this baseline model are visual feature maps, without the score map $\mathbf{a}$. This baseline model is a simplified version of our proposed model where the image-text relations are not considered. We pretrained this baseline on the SALICON dataset with ground truth labels, and then finetuned it on the HSB dataset. Neither pretraining or finetunig used the objects information.

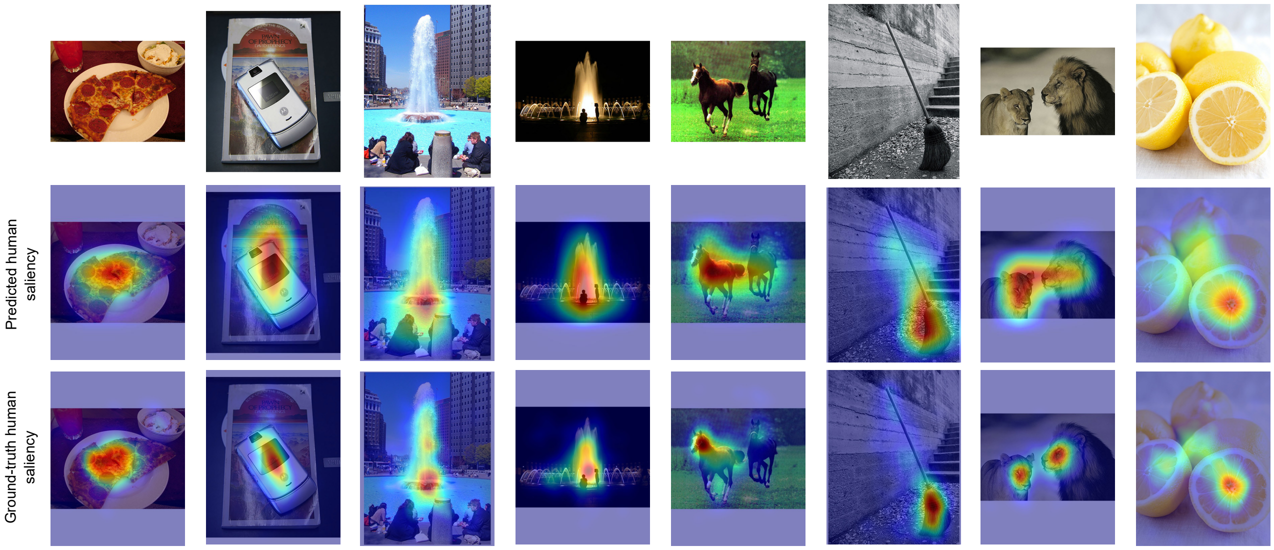


Figure 10: Visual comparison of the predicted saliency map and ground truth. Best viewed in color.

Approach	CC	KLdiv	NSS	SIM
baseline	0.8754	0.2239	2.3366	0.7533
ours	**0.8819**	**0.2162**	**2.3505**	**0.7563**

Table 2: The five-fold cross validation metrics (average values are reported).

In Table 2, we compared the baseline model and our proposed model. We find that using the image-text relations in the training losses functions and the model architecture as our proposed, we achieved better performance than the baseline model.

Necessity of Pretraining In our second experiment, we tested the necessity of pretraining for the HSB dataset with the aforementioned baseline model. Since the HSB dataset is a very small dataset, our intuition is that training from scratch will easily lead to overfitting to the training data. We compared the evaluation metrics of two baseline models: one was intialized randomly, another was initialized from the best model of pretraining. From Table 3, we notice that when initialized with the best model weights in pretraining, the baseline model outperformed the counterpart with random initial weights by a large marginal for every metric. This indicates the necessity of pretraining for a small saliency dataset like the HSB dataset.

Init. weights	CC	KLdiv	NSS	SIM
random	0.7729	0.4539	1.9538	0.6626
pretrain	**0.8754**	**0.2239**	**2.3366**	**0.7533**

Table 3: The five-fold cross validation metrics (average values are reported). “random” indicates the model weights were initialized randomly. “pretrain” indicates the model weights were initialized from the best model in pretraining.

HSI-Generated Human Saliency Map

In this section, we generate human saliency maps on images out of HSB dataset and show the capability of using HSI to replace human-based experiments in benchmarking saliency-based model explanations.

Dataset Preparation

We select the validation set of ImageNet 2012 classification dataset for human saliency map generation. In total, 1000 images are randomly sampled from each of the 1000 ImageNet classes. Then, we apply our trained model HSI on the images to obtain the predicted human saliency maps. The result is denoted as HSI map (Figure 11b).

The XAI saliency maps (Figure 11c-d) are generated for the images on classification labels. As is introduced in Section 2, three XAI methods, i.e. GBP (Springenberg et al. 2015), Grad-CAM (Selvaraju et al. 2017) and RISE (Petsiuk, Das, and Saenko 2018), are selected for demonstration. The model explanations are generated using a pretrained ResNet-50 from PyTorch (Paszke et al. 2019), which achieved 95.434% Top-5 accuracy on ImageNet classification task.

Qualitative Comparison

Here, we qualitatively compare the HSI map and XAI saliency maps. Overall, the HSI maps (Figure 11a) are smooth, continuous and tend to focus on center parts in the image. Moreover, the highlighted features are selective and representative for the target label, showing a similar attention strategy that human employed in our explanation task.

In contrast, highlight features on XAI maps are more distributed. In detail, each method has their own characteristic in terms of saliency patterns, e.g., GBP highlights pixels along edges (Figure 11c), Grad-CAM (Figure 11d) and RISE (Figure 11e) highlights some abnormal regions which are off-center.

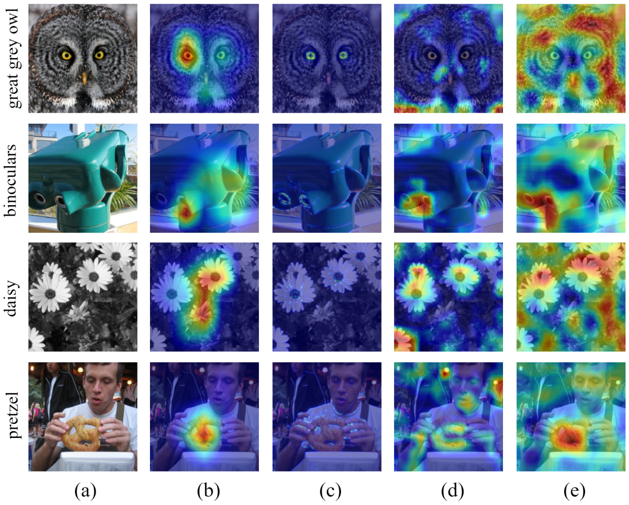


Figure 11: Visual comparison of HSI-generated human saliency map (b) and XAI saliency maps. Three XAI methods are (c) GBP, (d) Grad-CAM, (e) RISE. Best viewed in color.

Quantitative Comparison

Here we quantify the difference of saliency maps between XAI and human (HSI). Following the evaluation procedure in (Mohseni, Block, and Ragan 2021), we obtain a pair of saliency maps for one test image and calculate the pixel-wise Mean Absolute Error (MAE) between them. Then, we calculate MAE over the entire test set and report the average score. Table 4 reported the score for each XAI methods and their ranking in ascending order of the score. Based on this ranking, RISE is regarded as the best XAI method in terms of the pixel-wise comparison with human (HSI) saliency map.

Ranking	Method	MAE
1	RISE	0.44
2	Grad-CAM	0.70
3	GBP	0.89

Table 4: MAE scores for XAI methods benchmarked on HSI-generated dataset.

Discussion

Previous research mainly used attention datasets from passive viewing for human attention prediction. However, our pilot human study showed that human attention differs between passive viewing and explanation tasks greatly. People showed more focused eye movement patterns (and lower entropy) in the explanation task than in the passive viewing task as assessed using EMHMM. More specifically, people attended to information irrelevant to the foreground object category more often in passive viewing than explanation. To provide a more accurate human attention benchmark for XAI, we used an explanation task for our HSB. We further examined the relationship between individual differences in eye movement pattern and explanation performance using EMHMM, aiming to use those associated with better explanation performance as the benchmark. Our analysis showed that people who adopted a more explorative eye movement pattern during explanation had better explanation performance (as evaluated by data scientists). We thus used eye movement data from the explorative pattern group in the explanation task to generate our HSB. In other words, we have considered the latest findings from cognitive science research that human attention is both task- and person-specific, and accordingly generated our HSB using eye movements associated with better performance in the explanation task.

In the past, various computational models based on encoder-decoder architecture have been studied for human attention prediction. However, the model can only learn from image features to predict human attention, which fits to the passive-viewing human attention only. In HSB dataset, the studied human attention is driven by both image and label features, and thus we believe it is important for the encoder to have the ability of receiving and combining text features and image features. To our best knowledge, our encoder is the first to utilize the pre-training knowledge of V&L model in human saliency prediction. Although the model architecture is minimal, we show that the proposed model can successfully reproduce human saliency map in HSB dataset.

In the XAI evaluation experiment, we selected MAE as the metric to evaluate three XAI methods, as the metric was used in previous benchmarking research. We leave it for further work to study different similarity (and dissimilarity) metrics and their consistency in ranking XAI methods.

Conclusion

In this paper, we proposed a human attention benchmark for saliency-based explanations and a computational model to generate human saliency map for relieving labor cost in crowdsouring experiments. For HSB dataset, we considered both task- and person-specific nature of human attention, and used eye movements associated with better performance in the explanation task for human saliency imitation. Our model, HSI, extracts multi-modal features with V&L pretrain knowledge and successfully reproduces human attention saliency maps on HSB dataset. The study demonstrates the ability of using well-trained model to generate huamn saliency-based annotation, breaking through the constraint of the high cost of human data collection. Based on the advantage of HSI, further study of XAI evaluation can be conducted on a large scale of human benchmark dataset.

Acknowledgements

We thank Jennifer Ruoxi QI for data collection and eye movement data analysis and Yannie Tian Xin LIM for data collection. The eye tracker used was supported by RGC of Hong Kong (No. C7129-20G to Hsiao).

References

Alletto, S.; Palazzi, A.; Solera, F.; Calderara, S.; and Cucchiara, R. 2016. DR(Eye)Ve: A Dataset for Attention-Based Tasks With Applications to Autonomous and Assisted Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

An, J.; and Hsiao, J. H. 2021. Modulation of mood on eye movement and face recognition performance. *Emotion*, 21(3).

Bakator, M.; and Radosav, D. 2018. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3): 47.

Borji, A.; and Itti, L. 2015. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. *CoRR*, abs/1505.03581.

Chan, C. Y.; Chan, A. B.; Lee, T.; and Hsiao, J. H. 2018. Eye-movement patterns in face recognition are associated with cognitive decline in older adults. *Psychonomic bulletin & review*, 25(6): 2200–2207.

Chuk, T.; Chan, A. B.; and Hsiao, J. H. 2014. Understanding eye movements in face recognition using hidden Markov models. *Journal of vision*, 14(11): 8–8.

Cornia, M.; Baraldi, L.; Serra, G.; and Cucchiara, R. 2018. Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model. *IEEE Trans. Image Process.*, 27(10): 5142–5154.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Grigorescu, S.; Trasnea, B.; Cocias, T.; and Macesanu, G. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386.

Gunning, D. 2017. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, 2(2): 1.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608.

Hsiao, J. H.; An, J.; Zheng, Y.; and Chan, A. B. 2021a. Do portrait artists have enhanced face processing abilities? Evidence from hidden Markov modeling of eye movements. *Cognition*, 211: 104616.

Hsiao, J. H.; Chan, A. B.; An, J.; Yeh, S.-L.; and Jingling, L. 2021b. Understanding the collinear masking effect in visual search through eye tracking. *Psychonomic Bulletin & Review*, 28(6): 1933–1943.

Hsiao, J. H.; Lan, H.; Zheng, Y.; and Chan, A. B. 2021c. Eye Movement analysis with Hidden Markov Models (EMHMM) with co-clustering. *Behavior Research Methods*, 53(6): 2473–2486.

Hsiao, J. H.; Ngai, H. H. T.; Qiu, L.; Yang, Y.; and Cao, C. C. 2021d. Roadmap of Designing Cognitive Metrics for Explainable Artificial Intelligence (XAI). arXiv:2108.01737.

Hsiao, J. H.-w.; Liao, W.; and Tso, R. V. Y. 2022. Impact of mask use on face recognition: an eye-tracking study. *Cognitive Research: Principles and Implications*, 7(1): 1–15.

Hwu, T.; Levy, M.; Skorheim, S.; and Huber, D. 2021. Matching Representations of Explainable Artificial Intelligence and Eye Gaze for Human-Machine Interaction. *CoRR*, abs/2102.00179.

Jia, S.; and Bruce, N. D. B. 2020. EML-NET: An Expandable Multi-Layer NETwork for saliency prediction. *Image Vis. Comput.*, 95: 103887.

Jiang, M.; Huang, S.; Duan, J.; and Zhao, Q. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1072–1080.

Kanan, C.; Bseiso, D. N.; Ray, N. A.; Hsiao, J. H.; and Cottrell, G. W. 2015. Humans have idiosyncratic and task-specific scanpaths for judging faces. *Vision research*, 108: 67–76.

Karessli, N.; Akata, Z.; Schiele, B.; and Bulling, A. 2017. Gaze embeddings for zero-shot image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4525–4534.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kroner, A.; Senden, M.; Driessens, K.; and Goebel, R. 2020. Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*, 129: 261–270.

Kümmerer, M.; Wallis, T. S. A.; Gatys, L. A.; and Bethge, M. 2017. Understanding Low- and High-Level Contributions to Fixation Prediction. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 4799–4808. IEEE Computer Society.

Lai, Q.; Khan, S.; Nie, Y.; Sun, H.; Shen, J.; and Shao, L. 2020. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23: 2086–2099.

Lemhöfer, K.; and Broersma, M. 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, 44(2): 325–343.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014a. Microsoft COCO: Common Objects in Context. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, L. 2014b. Microsoft COCO: Common Objects in Context. In *ECCV*. European Conference on Computer Vision.

Lou, J.; Lin, H.; Marshall, D.; Saupe, D.; and Liu, H. 2022. TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494: 455–467.

Markman, A. B.; and Wisniewski, E. J. 1997. Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1): 54.

Mohseni, S.; Block, J. E.; and Ragan, E. D. 2021. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. In Hammond, T.; Verbert, K.; Parra, D.; Knijnenburg, B. P.; O'Donovan, J.; and Teale, P., eds., *IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021*, 22–31. ACM.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, 151. BMVA Press.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Reddy, N.; Jain, S.; Yarlagadda, P.; and Gandhi, V. 2020. Tidying Deep Saliency Prediction Architectures. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021*, 10241–10247. IEEE.

Rong, Y.; Xu, W.; Akata, Z.; and Kasneci, E. 2021. Human Attention in Fine-grained Classification. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, 150. BMVA Press.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. A. 2015. Striving for Simplicity: The All Convolutional Net. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, W.; and Shen, J. 2017. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5): 2368–2378.

Wang, W.; and Shen, J. 2018. Deep Visual Attention Prediction. *IEEE Trans. Image Process.*, 27(5): 2368–2378.

Zheng, Q.; Jiao, J.; Cao, Y.; and Lau, R. W. 2018. Task-driven webpage saliency. In *Proceedings of the European conference on computer vision (ECCV)*, 287–302.

Zheng, Y.; Ye, X.; and Hsiao, J. H. 2022. Does adding video and subtitles to an audio lesson facilitate its comprehension? *Learning and Instruction*, 77: 101542.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2921–2929. IEEE Computer Society.