Learned Region Sparsity and Diversity Also Predict Visual Attention

Zijun Wei*
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794
zijwei@cs.stonybrook.edu

Hossein Adeli*
Department of Psychology
Stony Brook University
Stony Brook, NY 11794
hossein.adelijelodar@stonybrook.edu

Gregory Zelinsky
Department of Psychology
Stony Brook University
Stony Brook, NY 11794
gregory.zelinsky@stonybrook.edu

Minh Hoai
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794
minhhoai@cs.stonybrook.edu

Dimitris Samsaras
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794
samaras@cs.stonybrook.edu

Abstract

Learned region sparsity has achieved state-of-the-art performance in classification tasks by exploiting and integrating a sparse set of local information into global decisions. The underlying mechanisms resemble how people sample information from the image with their eye-movements when making similar decisions. In this paper we enhance the learned region sparsity model with the biologically plausible mechanism of Inhibition of Return, to impose diversity on the selected regions. We investigated whether these mechanisms of sparsity and diversity correspond to visual attention by testing our model on three different types of visual search tasks. We report state-of-the-art results in predicting the location of human visual attention, even though we only trained on image-level labels without object location annotation. Notably the enhanced model's classification performance remains the same as the original. This work sheds some light on the possible visual attention mechanisms in the brain and argues for inclusion of attention-based mechanisms for improving computer vision techniques.

1 Introduction

Visual spatial attention refers to the narrowing of processing in the brain to particular objects in particular locations so as to mediate everyday tasks. A widely used paradigm for studying visual spatial attention is visual search, where a desired object must be located and recognized in a typically cluttered environment. Visual search is accompanied by observable estimates—in the form of gaze fixations—of how attention samples information from a scene while searching for a target. Efficient visual search requires prioritizing the locations of features of the target object class over features at locations offering less evidence for the target [33]. Computational models of visual search typically

*Both authors contributed equally to this work

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

estimate and plot goal directed prioritization of visual space as *priority maps* for directing attention [34]. Note that this form of target directed prioritization is different from the *Saliency* modeling literature where the image feature contrast is used to predict fixation behavior during free-viewing of scenes [17].

The field of fixation prediction is highly active and growing [2], although it was not until fairly recently that attention researchers have begun to use the sophisticated object detection techniques developed in the computer vision literature [9, 19, 33]. The dominant method used in visual search literature to generate priority maps for detection has been the exhaustive detection mechanism [9, 19]. Using this method, an object detector is applied on the image to provide bounding boxes which are then combined, weighted by their detection score, to generate the priority map [9]. While these models have had success in predicting behavior, training these detectors requires human labeled bounding boxes, which are only approximations of the object fixated. These boxes are expensive and very laborious to collect but also prone to individual annotator differences.

An alternative method to model visual attention is to test whether the core mechanisms that drive the performance of computational models in their task corresponds to attention mechanisms in the brain that are central to performing the same task [25]. To this end, a new class of models have been developed as part of a larger class of attention-inspired models applied to tasks from image captioning [32] to hand writing generation [16] where selective spatial attention mechanisms have been shown to emerge [1, 26]. By constraining the visual input to be gated similar to human gating of visual input by fixations, these models are able to localize or "attend" selectively to the most informative regions of an input image while ignoring the irrelevant visual input [26, 1]. The built in attention mechanism enables the model of [32] trained only on generating captions to bias the visual input so as to gate only relevant information when generating each word to describe an image. The priority maps were then generated to show the mapping of attended image areas to generated words. While these new models show attention-like behavior, to our knowledge none have been used to predict actual human attention behavior.

The current work bridges the behavioral and computer vision literatures by using a classification model that has biologically plausible constraints to create a priority map for the purpose of predicting the allocation of spatial attention measured by changes in fixation. The specific image-category classification model that we use is called Region Ranking SVM (RRSVM). This model achieves state-of-the-art performance on a number of classification tasks by learning categorization with locally-pooled information from input images. This model works by imposing sparsity on selected image areas that contribute to the classification decision much like how humans would prioritize the visual space to sample only a sparse set of areas in the image to detect and recognize an object category [4, 27]. We believe that this analogy between sparse sampling and attention makes this model a natural candidate for predicting attention behavior in visual search tasks. It is worth noting that this model was originally created for object classification and not localization, hence no object localization data is used to train it, unlike standard fixation prediction algorithms.

There are two contributions of our work. First, we show that the RSSVM model approaches state-of-the-art in predicting the fixations made by humans searching for the same targets in the same images. This means that a model trained solely for the purpose of image classification, without any localization data, is also able to predict the locations of fixations that people make when searching for the to-be-classified objects. Second, we incorporate the biologically plausible constraint of Inhibition of Return [13] which we model by requiring a set of diverse (not-too-much overlapping) sparse regions in RRSVM. Incorporating this constraint, we are able to significantly improve the performance of an already powerful image classification model for fixation prediction (up to 21% in error reduction). Interestingly we do not incur classification performance cost. By building this bridge, we hope to show how automated object detection might be improved by the inclusion of an attention mechanism, and how a recent attention-inspired approach from computer vision might illuminate how the brain prioritizes visual information for the efficient direction of spatial attention.

2 Region Ranking SVM

We review here Region Ranking SVM (RRSVM) [31]. The main problem addressed by RRSVM is image classification, which aims to recognize the semantic category of an image, such as whether the image contains a certain object (e.g., car, cat) or portrays a certain action (e.g., jumping, typing). RRSVM evaluates multiple local regions of an image, and subsequently outputs the classification

decision based on a sparse set of regions. This mechanism is noteworthy. It is very different from approaches that aggregate information from multiple regions indistinguishably (e.g., [24, 30, 23, 7]).

RRSVM assumes the training data consists of images $\{\mathbf{B}_i\}_{i=1}^n$ and associated binary labels $\{y_i\}_{i=1}^n$ indicating the presence or absence of the visual concept of interest. To account for the uncertainty of each semantic region in an image, RRSVM considers multiple local regions. The number of regions can differ between images, but for brevity, assume each image has the same number of regions. Let m be the number of regions for each image, and d the dimension of each region descriptor. RRSVM represents each image as a matrix $\mathbf{B}_i \in \Re^{d \times m}$, but the order of the columns can be arbitrary. RRSVM jointly learns a region evaluation function and a region selection function by solving the following optimization problem:

$$\underset{\mathbf{w},\mathbf{s},b}{\text{minimize}}\ \lambda||\mathbf{w}||^2 + \sum_{i=1}^{n}(\mathbf{w}^T\Gamma(\mathbf{B}_i;\mathbf{w})\mathbf{s} + b - y_i)^2 \tag{1}$$

$$\text{s.t. } s_1 \geq s_2 \geq \cdots \geq s_m \geq 0, \tag{2}$$

$$h(\Gamma(\mathbf{B}_i;\mathbf{w})\mathbf{s}) \leq 1. \tag{3}$$

In the above formulation, $h(\cdot)$ is the function that measures the spread of the column vectors of a matrix: $h([\mathbf{x}_1, \cdots, \mathbf{x}_n]) = \sum_{i=1}^n \left|\left|\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\right|\right|^2$. $\mathbf{w}$ and b are the weight vector and the bias term of an SVM classifier, which are the parameters of the region evaluation function. $\Gamma(\mathbf{B};\mathbf{w})$ denotes a matrix that can be obtained by rearranging the columns of the matrix $\mathbf{B}$ so that $\mathbf{w}^T\Gamma(\mathbf{B};\mathbf{w})$ is a sequence of non-increasing values. Vector $\mathbf{s}$ is the weight vector for combining the SVM region scores for each image; this vector is common to all images of a class.

The objective of the above formulation consists of the regularization term $\lambda||\mathbf{w}||^2$ and the sum of squared losses. This objective is purely based on classification performance. However, note that the classification decision is based on both the region evaluation function (i.e., $\mathbf{w}, b$) and the region selection function (i.e., $\mathbf{s}$), which are simultaneously learned using above formulation. What is interesting is that the obtained $\mathbf{s}$ vector is always sparse. An experiment [31] on the ImageNet dataset [29] with 1000 classes shows that RRSVM generally uses 20 regions or less (from hundreds of local regions considered). This intriguing fact prompts us to consider the connection between sparse region selection and visual attention. Would machine-based discriminative localization somehow relate to human attention in visual search? It turns out that there is a strong connection, as will be shown in the experiment section. This connection can be further reinforced if RRSVM is extended to incorporate *Inhibition of Return* in the region selection process, which is the subject of the next section.

3 Incorporating Inhibition of Return into Region Ranking SVM

An important characteristic of visual search behavior is Inhibition of Return: there is a lower chance of re-fixating on or nearby already attended areas, possibly mediated by lateral inhibitory mechanism [17, 21]. This mechanism, however, is not currently enforced in the formulation of RRSVM. In particular, the spatial relationship between selected regions is not considered in RRSVM. RRSVM usually selects a sparse set of regions, but the selected regions can overlap and cluster on a single image area.

Inspired by Inhibition of Return, we consider an extension of RRSVM where non-maxima suppression is incorporated into the process of selecting regions. This mechanism will select the local maximum for nearby activation areas (potential fixation location) and discard the rest (non-maxima nearby locations). The biological plausibility of non-maxima suppression has been discussed in previous work and it was shown to be a plausible method for allowing the stronger activations to stand out (see [22, 8] for details).

To incorporate non-maxima suppression in the framework of RRSVM, we propose to replace the region ranking procedure $\Gamma(\mathbf{B};\mathbf{w})$ of RRSVM (Eq. 1) by $\Psi(\mathbf{B}_i;\mathbf{w},\alpha)$, a procedure that ranks and subsequently returns the list of regions that do not significantly overlap with one another. In particular, we use intersection over union to measure overlapping, and α is the threshold for tolerable overlap

(we set $\alpha = 0.5$ in our experiments). This leads to the following optimization problem:

$$\underset{\mathbf{w},\mathbf{s},b}{\text{minimize}}\ \lambda||\mathbf{w}||^2 + \sum_{i=1}^{n}(\mathbf{w}^T\Psi(\mathbf{B}_i;\mathbf{w},\alpha)\mathbf{s} + b - y_i)^2 \tag{4}$$

$$\text{s.t. } s_1 \geq s_2 \geq \cdots \geq s_m \geq 0, \tag{5}$$

$$h(\Psi(\mathbf{B}_i;\mathbf{w},\alpha)\mathbf{s}) \leq 1. \tag{6}$$

The above formulation can be optimized in the same way as RRSVM in [31]. It will yield a classifier that makes decision based on a sparse and diverse set of regions. Sparsity is inherited from RRSVM, and location diversity is attained using non-maxima suppression. Hereafter, we refer to this method as Sparse Diverse Regions classifier (SDR).

4 Experiments and Analysis

We present here empirical evidence showing that learned region sparsity and diversity can also predict visual attention. We first describe the implementation details of RRSVM and SDR. We then consider attention prediction under three conditions: (1) single-target present, that is to find one category instance presented in the stimuli image; (2) target absent, i.e., searching for a target category that is absent in the stimuli; and (3) multiple-targets present, i.e., searching for multiple object categories where at least one category is present in the stimuli. Experiments are performed on three datasets POET [28], PET [14] and MIT people search (MIT900) [9], which are the only available datasets for object search tasks.

4.1 Implementation details of RRSVM and SDR

Our implementation of RRSVM and SDR is similar to [31], but we consider more local regions. This yields finer localization map without changing the classification performance. As in [31], the feature extraction pipeline is based on VGG16 [30]. The last fully connected layer of VGG16 is removed and the remaining fully connected layer is converted to a fully convolutional layer. To compute feature vectors for multiple regions of an image, the image is resized and subsequently fed into VGG16 to yield a feature map with 4096 channels. The size of the feature map depends on the size of the resized image, and each feature map corresponds to a subwindow of the original image. By resizing the original image to multiple sizes, one can compute feature vectors for multiple regions of the original image. In this work, we consider 7 different image sizes instead of the three sizes used by [30, 31]. The first three resized images are obtained by scaling the image isotropically so that the smallest dimension equals to either 256, 384, or 512. For brevity, assuming the width is smaller than the height, this yields three images with dimensions $256 \times a, 384 \times b$, and $512 \times c$. We consider four other resized images with dimensions $256 \times b, 384 \times c, 384 \times a, 512 \times b$. These image sizes correspond to the consideration of local regions that have aspect ratio of either 2:3 or 3:2, while the isotropically resized images yield square local regions. Additionally, we also consider horizontal flips of the resized images. Overall, this process yields from 700 to 1000 feature vectors, each corresponding to a local image region.

The RRSVM and SDR classifiers used in the following experiments are trained on the trainval set of PASCAL VOC 2007 dataset [12] unless otherwise stated. This dataset is distinct from the datasets used for evaluation. For SDR, the non-maxima suppression threshold is 0.5, and we only keep the top ranked regions that have non-zero region scores ($s_i \geq 0.01$). To generate a priority map, we first associate each pixel with an integer indicating the total number of selected regions covering that pixel. To create the priority map [5], we apply a Gaussian blur kernel to the integer valued map, with the kernel width tuned on the validation set.

To understand how learned region sparsity and diversity correspond to human attention, we compare the generated priority map with the fixation density map. We use the Area Under ROC Curve (AUC), the most commonly used metric for visual search task evaluation [6]. We use the publicly available implementation of the AUC evaluation from the MIT saliency benchmark [5]. More specifically, we use the AUC-Judd implementation for its better approximation.

4.2 Single-target present condition

We consider visual attention in single-target present condition using the POET dataset [28]. This dataset is a subset of PASCAL VOC 2012 dataset [11], and it has has 6270 images from 10 object

Table 1: AUC scores on POET and PET test sets

Model	POET aero	bike	boat	cat	cow	table	dog	horse	mbike	sofa	**mean**	PET multi-target
SDR	**0.87**	**0.85**	**0.83**	**0.89**	**0.88**	**0.79**	**0.88**	**0.86**	0.86	**0.77**	**0.85**	**0.83**
RCNN	0.84	0.83	0.79	0.84	0.81	0.76	0.83	0.80	**0.87**	0.76	0.82	0.77
CAM [36]	0.86	0.78	0.78	0.88	0.84	0.74	0.87	0.84	0.83	0.67	0.82	0.65
AnnoBoxes	0.85	0.86	0.81	0.84	0.84	0.79	0.80	0.80	0.88	0.80	0.83	0.82

categories (aeroplane, boat, bike, motorbike, cat, dog, horse, cow, sofa and dining table). The task of two-alternative forced choice on object categories was used in eye tracking data collection to approximate visual search tasks. Eye movement data were collected from 5 subjects for each image. A mean of 5.7 fixations per image were recorded from subjects. We randomly selected one third of the images for each category to compile a validation set for tuning the width of the Gaussian blur kernel for all categories. The rest are used as test images.

For each test image, we compare the priority map generated for the selected regions by RRSVM and the human fixation density map. The overall correlation is high, yielding the mean AUC score of 0.81 (on all images of 10 object classes). This is intriguing because RRSVM is set up to optimize the classification performance only; somehow, joint classification and discriminative localization relates to human attention in a visual search task. By incorporating Inhibition of Return to RRSVM, we observe stronger correlation with human behavior. The mean AUC score obtained by SDR is 0.85.

The left part of Table 1 shows AUC scores for individual categories of the POET dataset. We compare the performance of other attention prediction baselines. All recent fixation prediction models [9, 20, 33] apply object category detectors on the input image and combine the detection results to create the priority maps. Direct comparison is infeasible at the moment as we do not have access to all of the code and datasets required. However, our RCNN [15] baseline, which is the state-of-the-art object detector on this dataset should improve the pipelines of these models. To account for possible localization errors and multiple object instances, we keep all the detections with detection score greater than a threshold. This threshold is chosen to maximize the detector's F1 score, which is the harmonic mean between precision and recall. We also consider a variant method where only the top detection is kept, but the result is not as good (see supplementary material). We also consider the recently proposed weakly-supervised object localization approach of [36], which is denoted as CAM in Table 1. We use the released model to extract features and train a linear SVM on top of the features. For each test image, we weigh a linear sum of local activations to create an activation map. We normalize the activation map to get the priority map. We even compare SDR with a method that directly uses the annotated object bounding boxes to predict human attention, which is denoted as AnnoBoxes in the table. For this method, the priority map is created by applying a Gaussian filter to a binary map where the center of the bounding box over the target(s) is set to 1 and everywhere else 0. Notably, the methods for comparison here are strong models for predicting human attention. RCNN has an unfair advantage over SDR because it has access to localized annotation for train data, and AnnoBoxes even assumes the availability of object bounding boxes for test data. As can be seen from Table 1, SDR significantly outperforms the other methods. This strong empirical evidence suggests that the learned region sparsity and diversity can predict human attention. Fig. 1 shows some randomly selected results of SDR on test images.

Note that the incorporation of Inhibition of Return into RRSVM and the consideration of more local regions do not alter the classification performance. Evaluated on the test set of PASCAL VOC 2007 dataset, the RRSVM that uses local regions corresponding to 3 image scales (as in [31]), using more regions with different aspect ratios (as explained in Sec. 4.1), or incorporating the NMS mechanism (i.e., SDR) all achieve a mean AP of 92.9%. SDR, however, is significantly better than RRSVM in predicting fixations during search tasks, raising mean AUC scores from 0.81 to 0.85. Also note that the predictive power of SDR is quite stable over changes of α: for aeroplane on the POET dataset, the AUC scores are the same (0.87) when α rangs from 0.5 to 0.7.

Figure 2 shows some examples highlighting the difference between the regions selected by RRSVM and SDR. As can be seen, incorporating non-maxima suppression encourages greater dispersion of the sparse areas as opposed to a more clustered distribution in RRSVM. This in turn better predicts attention when there are multiple instances of the target concept in the display.

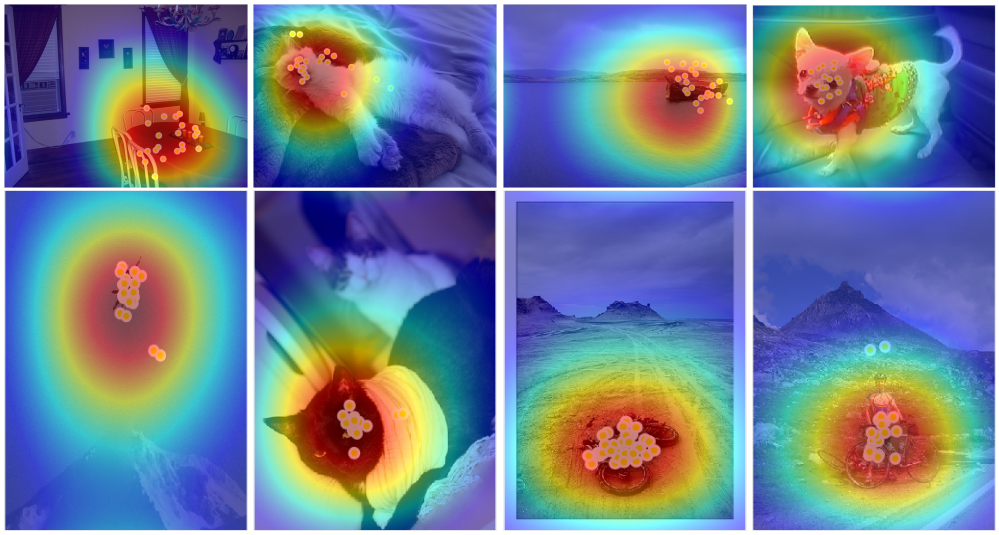


Figure 1: **Priority maps generated for SDR on the POET dataset.** Warm colors represent high values. Dots represents human fixations. Best viewed on a digital device.

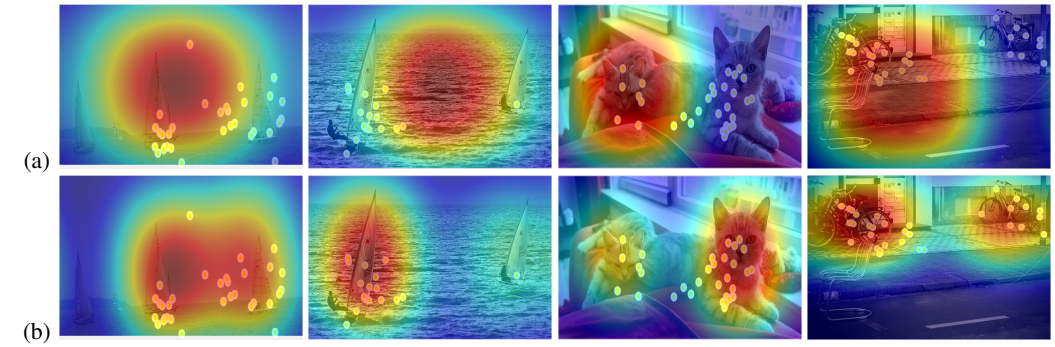
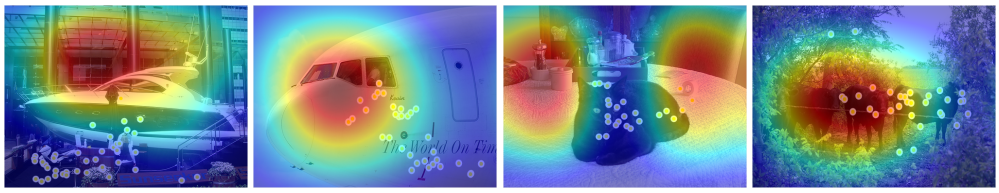


Figure 2: **Comparison between RRSVM and SDR on the POET dataset.** (a): priority maps created by RRSVM, (b): priority maps generated by SDR. SDR better capture fixations for multiple instances of the target categories. The KL Divergence scores between RRSVM and SDR are also reported in the bottom row.

Figure 3 show representative cases where the priority maps produced by SDR are significantly different from human fixations. The common failure modes are: (1) failure to locate the correct region for correct classification (see Fig 3a); (2) there are distracting non-target concepts in the scene, such as text (3b) or faces (3c); (3) failure to attend to multiple instances of the target categories. Tuning SDR using human fixation behavior data [18] and combining SDR with multiple sources of guidance information [9] including saliency and scene context could mitigate some of the model limitations.

4.3 Target absent condition

To test whether SDR is able to predict people's fixation when the search target is absent, we run experiments on 456 target absent stimulus from the MIT900 dataset [9]. The eye tracking data was collected by asking human observers to search for people in real world scenes. Eye movement data was collected from 14 viewers who made roughly 6 fixations per image on average. We pick a random subset of 150 images to tune the Gaussian blur parameter and report the results on the remaining 306 images. We noticed that the size and pose of the people in these images are very different from the training examples in VOC2007, which may lead to poor SDR classification performance. In order to address this issue, we augment the training set of SDR with 456 images from MIT900 that contain people. The added training examples are distinct from the target-absent images for evaluation.



(a) motorbike (b) aeroplane (c) diningtable (d) cow

Figure 3: **Failure cases**. Representative images where the priority maps produced by SDR are significantly different from human fixations. The caption under each image is the object category to search for. The modes of failure are: (a) failure in classification; (b) and (c) existence of a more attractive object (text or face); (d) co-occurrence of multiple objects. Best viewed on digital devices.

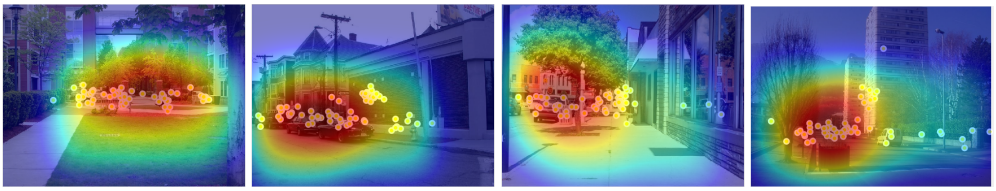


Figure 4: **Priority map predictions using SDR on the MIT target absent stimulus.** Warm colors represent high probabilities. Dots represents human fixations. Best viewed on a digital device.

On these target absent cases, SDR achieves an AUC score of 0.78. As a reference, the method of [9] also achieves AUC of 0.78. But the two methods are not directly comparable because [9] uses a HOG-based person detector that was trained on a much larger dataset with location annotation.

Some randomly selected results from the test set are shown in Figure 4. Interestingly, SDR looks at regions that either contain person-like objects or are likely to contain persons (e.g., sidewalks). This is predictive of where people attend to.

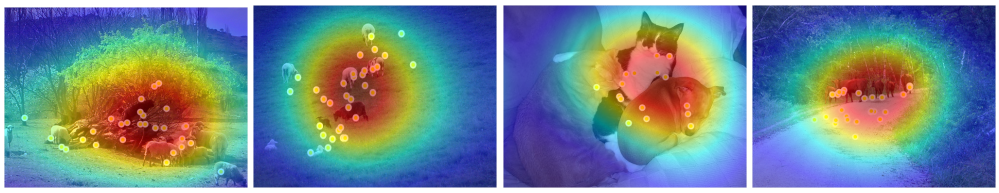
4.4 Multiple-target attention

We considered human attention for a visual search task with multiple targets. The experiments were performed on the PET dataset [14]. This dataset is a subset of PASCAL VOC2012 dataset [10], and it contains 4135 images from 6 animal categories (cat, dog, bird, horse cow and sheep). Image viewers were instructed to find **all** the animals in the image. Eye movement data were collected from 4 human viewers with roughly 6 fixations from each viewer per image. We excluded the images that contained people to avoid ambiguity with the animal category. We also removed the images that were shared with the PASCAL VOC 2007 dataset to ensure no overlap between training and testing data. This yielded a total of 3309 images from which a random set of 1300 images were selected for tuning the Gaussian kernel width parameter and the remaining 2309 images were used for testing.

To model the search for multiple categories in an image, for all methods except AnnoBoxes, we applied six animal classifiers/detectors simultaneously to the test image. For each classifier/detector of each category, a threshold was selected to achieve the highest F_1 score on the validation data. The prediction results are shown in Tab. 1. SDR significantly outperforms other methods. Notably, CAM performs poorly on this dataset, this might be the due to low classification accuracy of that model (83% mAP on VOC 2007 test set as opposed to 93% in SDR). Some randomly selected results are shown in Fig. 5.

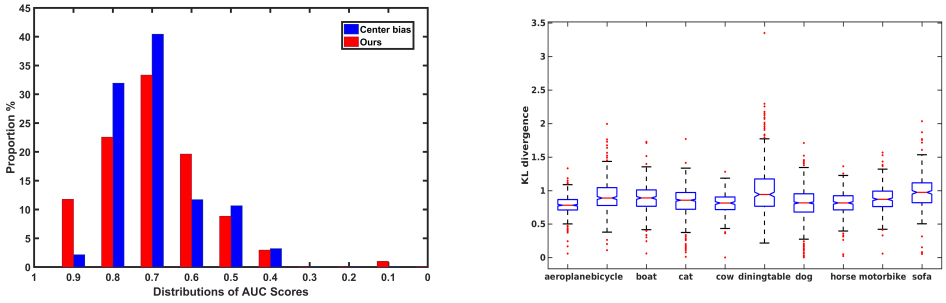
4.5 Center Bias

For the POET dataset, some of the target objects are quite iconic and in the center of the image. Under this case, a simple center bias map might be a good predictor of the fixations. To test this, we generated priority maps by setting the center of the image to 1 while everywhere else 0 and then applied Gaussian filter with sigma tuned on the validation set. This simple Center Bias map achieved AUC score of 0.84, which is even higher than some of the methods presented in Tab. 1. This prompts us to analyze whether the good performance of SDR is due to center bias.



(a) dog and sheep (b) cows and sheep (c) dog and cat (d) cows

Figure 5: **Visualization of SDR prediction on the PET dataset**. Note that the high classification accuracy ensures that more reliable regions are detected.



(a) (b)

Figure 6: (a): Red bars: the distribution of AUC scores returned by SDR for which the AUC scores returned by center bias are under 0.6. Blue bars: the distribution of AUC scores returned by center bias where their scores returned by SDR are under 0.6. (b): The box plot for the distributions of KL divergence from center bias to SDR scores on each class in POET dataset. The KL divergence distribution revealed that the priority maps created by center bias are largely different from the ones created by SDR.

An intuitive way to address the CB problem would be to use Shuffled AUC (sAUC) [35]. However, sAUC favors true positives over false negatives and gives more credit to off-center information [3], which may lead to biased results. This is especially true when the datasets are center-biased. We computed sAUC scores for the methods used in this paper on the POET dataset and found: RCNN(0.61), AnnoBox(0.61), CAM(0.65), SDR(0.64), and Inter-Observer(0.70). SDR outperforms AnnoBox and RCNN by 3% and is on par with CAM. Also note that sAUC for IO is 0.70, which suggests center bias in POET (the sAUC score of IO on MIT300 [18] is 0.81) and raises a concern that sAUC might be misleading for model comparison using this dataset. This prompts us to analyze whether the good performance of SDR is due to center bias.

To address the concern of center bias, we show in Fig. 6 that the priority maps produced by SDR and Center Bias are quite different. Fig. 6a plots the distribution of the AUC scores for one method for images where the AUC scores of the other method is low (< 0.6). These spread-out distributions indicate the low correlation between the errors of the two methods. Fig. 6b shows the box plot for the distribution of KL divergence [6] between the priority maps generated by SDR and Center Bias. For each category, the mean KL divergence value is high, indicating the large difference between RRSD and Center Bias. For qualitative intuition of KL divergence in these distributions, see Figure 2.

The center bias effect in PET and MIT900 is not as pronounced as in POET because there are multiple target objects in PET images and the target objects in MIT900 dataset are relatively small. On these datasets, Center Bias achieves AUC scores of 0.78 and 0.72 respectively. These numbers are significantly lower than the results obtained by SDR, which are 0.82 and 0.78 respectively.

5 Conclusions and Future Work

We have introduced a classification model based on sparse and diverse region ranking and selection, which is trained only on image level annotations. We then provided experimental evidence from visual search tasks under three different conditions to support our hypothesis that these mechanisms might be analogous to visual attention processes in the brain.

While this work is not the first using computer vision models to predict where human look in visual search tasks, it is the first to show that core mechanisms that drive the high performance of computational models in their task also predict the human visual attention behavior. We hope to shed some light on the possible implementation of attention mechanisms in the brain and on improving current computer vision systems by analogy of human visual processing systems.

There are several directions of future work. The first is creating a dataset for visual search tasks to mitigate the center bias effect and prevent trivial search tasks such as the ones in iconic images. The second will be to incorporate other sources such as the center bias effect, bottom-up saliency map and contextual information into the current model to better predict human attentions in visual search tasks.

References

[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv*, 2014.

[2] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *TPAMI*, 2013.

[3] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *2013 IEEE International Conference on Computer Vision*, pages 921–928. IEEE, 2013.

[4] N. D. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3):5–5, 2009.

[5] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/.

[6] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *arXiv*, 2016.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv*, 2014.

[8] P. Dario, G. Sandini, and P. Aebischer. Robots and biological systems: Towards a new bionics? In *NATO Advanced Workshop on Robots and Biological Systems*, 2012.

[9] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 2009.

[10] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.

[11] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. www.pascal-network.org/challenges/VOC/voc2012/workshop/, 2012.

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[13] J. H. Fecteau and D. P. Munoz. Salience, relevance, and firing: a priority map for target selection. *Trends in cognitive sciences*, 10(8):382–390, 2006.

[14] S. O. Gilani, R. Subramanian, Y. Yan, D. Melcher, N. Sebe, and S. Winkler. Pet: An eye-tracking dataset for animal-centric pascal object classes. In *Multimedia and Expo (ICME)*, 2015.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[16] A. Graves. Generating sequences with recurrent neural networks. *arXiv*, 2013.

[17] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10):1489–1506, 2000.

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[19] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17(6-7):979–1003, 2009.

[20] A. Kannan, J. Winn, and C. Rother. Clustering appearance and shape by learning jigsaws. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

[21] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[22] I. Kokkinos, R. Deriche, T. Papadopoulo, O. Faugeras, and P. Maragos. Towards bridging the Gap between Biological and Computational Image Segmentation. Research Report RR-6317, INRIA, 2007.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[24] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[25] T. S. Lee and X. Y. Stella. An information-theoretic framework for understanding saccadic eye movements. In *NIPS*, pages 834–840, 1999.

[26] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014.

[27] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, 1993.

[28] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *ECCV*. 2014.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

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

[31] Z. Wei and M. Hoai. Region ranking svms for image classification. In *CVPR*, 2016.

[32] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv*, 2015.

[33] G. J. Zelinsky, H. Adeli, Y. Peng, and D. Samaras. Modelling eye movements in a categorical search task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 2013.

[34] G. J. Zelinsky and J. W. Bisley. The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*, 2015.

[35] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008.

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *arXiv*, 2015.