Where Should Saliency Models Look Next?

Zoya Bylinskii[1(✉)], Adrià Recasens[1], Ali Borji[2], Aude Oliva[1], Antonio Torralba[1], and Frédo Durand[1]

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA
{zoya,recasens,oliva,torralba,fredo}@mit.edu
[2] Center for Research in Computer Vision, University of Central Florida, Orlando, USA
aborji@crcv.ucf.edu

Abstract. Recently, large breakthroughs have been observed in saliency modeling. The top scores on saliency benchmarks have become dominated by neural network models of saliency, and some evaluation scores have begun to saturate. Large jumps in performance relative to previous models can be found across datasets, image types, and evaluation metrics. Have saliency models begun to converge on human performance? In this paper, we re-examine the current state-of-the-art using a fine-grained analysis on image types, individual images, and image regions. Using experiments to gather annotations for high-density regions of human eye fixations on images in two established saliency datasets, MIT300 and CAT2000, we quantify up to 60% of the remaining errors of saliency models. We argue that to continue to approach human-level performance, saliency models will need to discover higher-level concepts in images: text, objects of gaze and action, locations of motion, and expected locations of people in images. Moreover, they will need to reason about the relative importance of image regions, such as focusing on the most important person in the room or the most informative sign on the road. More accurately tracking performance will require finer-grained evaluations and metrics. Pushing performance further will require higher-level image understanding.

Keywords: Saliency maps · Saliency estimation · Eye movements · Deep learning · Image understanding

1 Introduction

Where human observers look in images can provide important clues to human image understanding: where the main focus of the image is, where an action or event is happening in an image, and who the main participants are. The collection of human eye movements can help highlight image regions of interest to

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46454-1_49) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part V, LNCS 9909, pp. 809–824, 2016.
DOI: 10.1007/978-3-319-46454-1_49

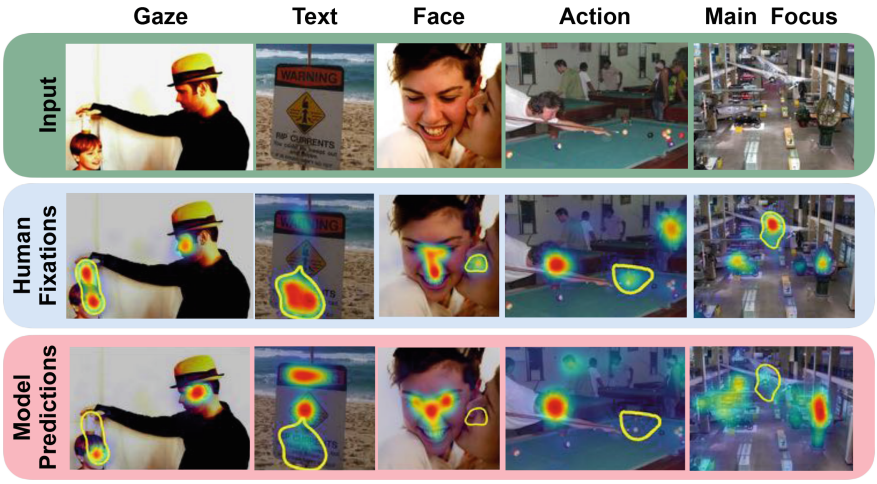

Fig. 1. Recent progress in saliency modeling has significantly driven up performance scores on saliency benchmarks. On first glance, model detections of regions of interest in an image appear to approach ground truth human eye fixations (Fig. 3). A finer-grained analysis can reveal where models can still make significant improvements. High-density regions of human fixations are marked in yellow, and show that models continue to miss these semantically-meaningful elements. (Color figure online)

human observers, and models can be designed to make computational predictions. The field of saliency estimation has moved beyond the modeling of low-level visual attention to the prediction of human eye fixations on images. This transition has been driven in part by large datasets and benchmarks of human eye movements (Fig. 1).

For a long while, the prediction scores of saliency models have increased at a stable rate. The recent couple of years have seen tremendous improvements on well-established saliency benchmark datasets [1]. These improvements can be attributed to the resurgence of neural networks in the computer vision community, and the application of deep architectures to saliency estimation. As a result, a large number of neural network based saliency models have emerged in a short period of time, creating a large gap in performance relative to traditional saliency models that are based on hand-crafted features, and learning-based models that integrate low-level features with object detectors and scene context [2–6]. Neural network-based models are trained to predict saliency in a single end-to-end manner, combining feature extraction, feature integration, and saliency value prediction.

These recent advances in the state-of-the-art and the corresponding saturation of some evaluation scores motivate the questions: Have saliency models begun to converge on human performance and is saliency a solved problem? In this paper we provide explanations of what saliency models are still missing, in order to match the key image regions attended to by human observers. We argue

that to continue to approach human-level performance, saliency models will need to discover increasingly higher-level concepts in images: text, objects of gaze and action, locations of motion, and expected locations of people in images. Moreover, they will need to reason about the relative importance of image regions, such as focusing on the most important person in the room or the most informative sign on the road. In other words, more accurately predicting where people look in images will require higher-level image understanding. In this paper, we examine the kinds of problems that remain and what will be required to push performance forward.

2 Related Work

Computational modeling of bottom-up attention dates back to the seminal works by Treisman and Gelade [7] (Feature Integration Theory), the computational architecture by Koch and Ullman [8] and the bottom-up model of Itti et al. [9]. Parkhurst and Neibur were the first to measure saliency models against human eye fixations in free-viewing tasks [10]. Followed by this work and the Attention for Information Maximization model of Bruce and Tsotsos [11], a cascade of saliency models emerged, establishing saliency as a subarea in computer vision. Large datasets of human eye movements were constructed to provide training data, object detectors and scene context were added to models, and learning approaches gained traction for discovering the best feature combinations [2–6]. Please refer to [12,13] for recent reviews of saliency models.

One of the first attempts to leverage deep learning for saliency prediction was Vig et al. [14], using convnet layers as feature maps to classify fixated local regions. Kümmerer et al. [15] introduced the model DeepGaze, built on top of the AlexNet image classification network [16]. Similarly, Liu et al. [17] proposed the Multiresolution-CNN model in which three convnets, each on a different image scale, are combined to obtain the saliency map. In the SALICON model [18], CNNs are applied at two different image scales: fine and coarse. The SALICON dataset, a large-scale crowd-sourced mouse movement dataset, made available to the saliency community for training new deep models [18], has led to the emergence of a number of other neural network models. For instance, DeepFix [19] is a fully convolutional neural network built on top of the VGG network [20] and trained on the SALICON dataset to predict pixel-wise saliency values in an end-to-end manner. DeepFix has additionally been fine-tuned on MIT1003 [3] and CAT2000 [21]. Pan et al. [22] also trained two architectures on SALICON in an end-to-end manner: a shallow convnet trained from scratch, and a deeper one whose first three layers were adapted from the VGG network (SalNet). Other saliency models based on deep learning have been proposed for salient region detection [23–26]. In this paper, we focus on predicting eye fixations rather than detecting and segmenting salient objects in scenes.

While deep learning models have shown impressive performance for saliency prediction, a finer-grained analysis shows that they continue to miss key elements in images. Here we investigate where the next big improvements can come from.

3 Evaluating Progress

We perform our evaluation on two datasets from the well-established MIT Saliency Benchmark [1]. We use the data from this benchmark because it has the most comprehensive set of traditional and deep saliency models evaluated. The **MIT300** dataset [27] is composed of 300 images from Flickr Creative Commons and personal collections. It is a difficult dataset for saliency models, as images are highly varied and natural. Fixations of 39 observers have been collected on this dataset, leading to fairly robust ground-truth to test models against. The **CAT2000** dataset [21] is composed of 2000 images from 20 different categories, varying from natural indoor and outdoor scenes to artificial stimuli like patterns and sketches. Images in this dataset come from search engines and computer vision datasets [28,29]. The test portion of this dataset, used for evaluation, contains the fixations of 24 observers.

As of March 2016, of the top 10 (out of 57) models evaluated on MIT300, neural network models filled 6 spots (and the top 3 ranks) according to many metrics[1]. DeepFix [19] and SALICON [18], both neural network models, hold the top 2 spots. The CAT2000 dataset, a recent addition to the MIT benchmark, has 19 models evaluated to date. DeepFix is the best model on the CAT2000 dataset overall and on all 20 image categories. BMS (Boolean map based saliency) [30] is the best-performing non neural network model across both datasets.

A finer-grained analysis on MIT300 (see Supplemental Material) shows that on a per-image level, DeepFix and SALICON alternate in providing the best prediction for ground-truth fixations. In the rest of the paper, our analyses are carried out on these models. Performances of these models on the MIT benchmark according to the benchmark metrics are provided in Table 1. We supplement

Table 1. Scores of top-performing neural network models (DeepFix, SALICON) and best non-neural network model (BMS) on MIT300 Benchmark. Top scores are bolded. Lower scores for KL and EMD are better. There has been significant progress since the traditional bottom-up IttiKoch model, but a gap remains to reach human-level performance. Chance and human limit values have been taken from [1,32].

Saliency model	AUC ↑	sAUC ↑	NSS ↑	CC ↑	KL ↓	EMD ↓	SIM ↑	IG ↑
Human limit	0.92	0.81	3.29	1	0	0	1	1.80
DeepFix [19]	**0.87**	0.71	**2.26**	**0.78**	0.63	**2.04**	**0.67**	0.67
SALICON [18]	**0.87**	**0.74**	2.12	0.74	**0.54**	2.62	0.60	**0.71**
BMS [30]	0.83	0.65	1.41	0.55	0.81	3.35	0.51	0.22
IttiKoch[a]	0.75	0.63	0.97	0.37	1.03	4.26	0.44	-0.15
Chance	0.50	0.50	0	0	2.09	6.35	0.33	-1.67

[a]Implementation from http://www.vision.caltech.edu/~harel/share/gbvs.php.

[1] As of July 2016, 8 of the top 10 (out of 62) models on MIT300 are neural networks.

Fig. 2. Saliency model ranking is preserved when evaluating models on this subset of 10 images as when evaluating them on the whole 300-image benchmark. These images help to accentuate differences in model performance. These images contain people at varying scales, as well as text (small here) amidst distracting textures.

these scores with a measure of Information Gain (IG) as suggested in [31,32]. Definitions and interpretations of these metrics are provided in [32].

To begin to explore where these recent large gains in performance are coming from, we visualize the most representative dataset images in Fig. 2. We define representative images as those that best preserve model rankings when tested on, compared to the whole dataset. Our correlation-based greedy image selection is described in the Supplemental Material. We find that a subset of $k = 10$ images can already rank the saliency models on the MIT benchmark with a Spearman correlation of 0.97 relative to their ranking on all dataset images. These images help to accentuate differences in model performance. By visualizing the predic-

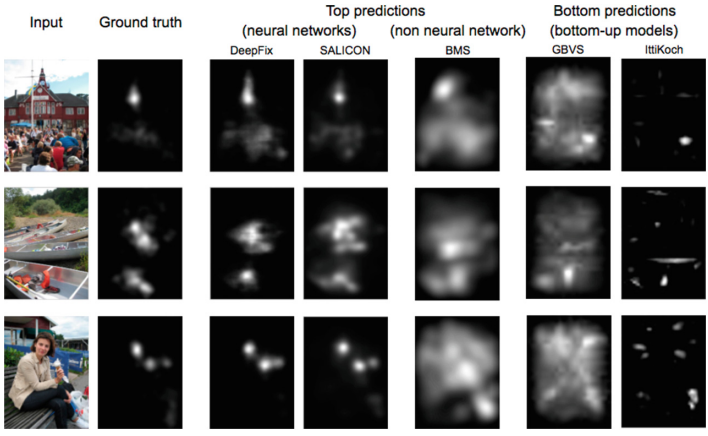

Fig. 3. Some of the best and worst model predictions on a few of the representative images from Fig. 2. Unlike traditional bottom-up models, recent neural network models can discover faces, text, and object-like features in images, prioritizing them over textures and low-level features appropriately, to better approximate human fixations.

tions of some of the top and bottom models on these images (Fig. 3), we can see that driving performance is a model's ability to detect people and text in images in the presence of clutter, texture, and potentially misleading low-level pop-out.

4 Quantifying Where People and Models Look in Images

To understand where models might fail, we must first understand where people look. Our goal is to name all the image regions lying beneath the high-density locations in fixation heatmaps. We computed fixation heatmaps aggregated over all observers on an image (39 observers in the MIT300 dataset, for a robust ground truth). Then we thresholded these ground truth heatmaps at the 95th percentile and collected all the connected components. This produced an average of 1–3 regions per image for a total of 651 regions.

The resulting region outlines were plotted on top of the original images and shown to Amazon Mechanical Turk (MTurk) participants with the task of selecting the labels that most clearly describe the image content that people look at (Fig. 4a). The labels provided for this task were not meant to serve as an exhaustive list of all objects, but to have good coverage of label types, with sufficient instances per label. If an image contained multiple image regions, only one would be displayed to participants at a time. Participants could select out of 15 different label categories as many labels as were appropriate to describe a region. For each image region, we collected labels from a total of 20 participants. Majority vote was used to assign labels to regions. A region could have multiple labels in case of ties. For further analyses, related labels (e.g. "animal face", "part of an animal", etc.) were aggregated to have sufficient instances per label type (see Supplemental Material). Not all regions are easily nameable, and in these cases participants could select the "background" or "other" labels. To account for these image regions to which simple labels could not be assigned, a second question-based MTurk task was deployed, described in the next section.

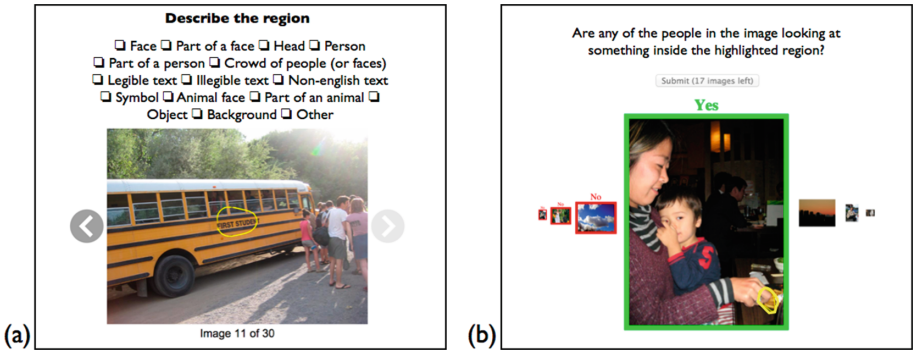

Fig. 4. Two types of Mechanical Turk tasks were used for gathering annotations for the highly-fixated regions in an image. These annotations were then used to quantify where people look in images.

4.1 What Do Models Miss?

Given labels for all the highly-fixated image regions in the MIT300 dataset, we intersected these labeled regions with the saliency maps of different computational models. To determine if saliency models made correct predictions in these regions, we calculated whether the mean saliency in these regions was within the 95-th percentile of the saliency map for the whole image. We then tallied up the types of regions that were most commonly under-predicted by models. In Table 2 we provide the error percentages, by region type, where saliency models assigned a value less than the percentile threshold to the corresponding regions. Our analyses are performed over DeepFix and SALICON models on the MIT300 dataset, and on DeepFix on the CAT2000 dataset (additional analyses in the Supplemental Material). The four categories chosen from the CAT2000 dataset are ones that contain natural images with a variety of objects and settings.

About half the failure modes are due to misdetections of parts of people, faces, animals, and text. Such failure cases can be ameliorated by training models on more instances of faces (partial, blurry, small, non-frontal views, occluded), more instances of text (different sizes and types), and animals. However, the labels "background", "object", and "other" assigned to image regions by MTurk participants originally accounted for about half of model errors on MIT300.

A second MTurk task was designed to better understand the content found in these harder-to-name image regions. Participants were asked to answer binary questions, such as whether or not a highlighted region in an image is an object of gaze or action in the image (see Fig. 4b and Supplemental Material). The results of this task allowed us to further break down model failure modes, and account for 60 % of total mispredictions on MIT300 and 39 %-98 % of mispredictions on four categories of CAT2000. The remaining failure modes (labeled "other") vary from image to image, caused by low-level features, background elements, and other objects or parts of objects that are not the main subjects of the photograph, nor are objects of gaze or action. Later in this paper, the most common failure modes are explored in greater detail. Examples are provided in Fig. 6.

4.2 What Can Models Gain?

With the region annotations obtained from our MTurk tasks, we performed an analysis complementary to the one in Sect. 4.1. Instead of computing model misses across image regions of different types, here we estimate the potential gains models could have if specific image regions were correctly predicted. A region is treated as a binary mask for the image, and a modified saliency map is computed as a combination of the original saliency map and ground truth fixation map. For each region type (e.g. "part of a person", "object of gaze"), we compute modified saliency maps. We replace model predictions in those regions with ground truth values obtained from the human fixation map (e.g., Fig. 5, top row). Figure 5 provides the score improvements of the modified models on the MIT300 benchmark. This analysis is meant to provide a general sense of the possible performance boost if different prediction errors are ameliorated.

Table 2. Labels for under-predicted regions on MIT300 and CAT2000 datasets. Percentages are computed over 681 labels assigned to 651 regions (some regions have multiple labels so percentages do not add up to 100 %). See Fig. 6 for visual examples.

Dataset	MIT300		CAT2000			
Model	DeepFix	SALICON	DeepFix			
Image category	All		Social	Action	Indoor	Outdoor
Part of main subject	31 %	36 %	49 %	68 %	12 %	24 %
Unusual element	18 %	16 %	33 %	63 %	8 %	8 %
Location of action/motion	16 %	16 %	67 %	78 %	8 %	11 %
Text	16 %	13 %	6 %	5 %	8 %	29 %
Part of a person	15 %	14 %	23 %	37 %	8 %	5 %
Possible location for a person	15 %	7 %	6 %	24 %	10 %	11 %
Object of action	14 %	15 %	27 %	51 %	0 %	3 %
Object of gaze	11 %	11 %	50 %	44 %	0 %	0 %
Part of a face	6 %	8 %	46 %	7 %	0 %	0 %
Part of an animal	5 %	5 %	3 %	10 %	0 %	0 %
Other	40 %	40 %	3 %	2 %	61 %	37 %

We include performance boosts of Normalized Scanpath Saliency (NSS) and Information Gain (IG) scores, which follow the distribution of region types in Table 2. The complete set of scores is provided in the Supplemental Material. It is important to note that the Area under ROC Curve (AUC) metrics have either saturated or are close to saturation. The focus of saliency evaluation should turn instead towards metrics that can continue to differentiate between models, and that can measure model performances at a finer-grained level (Sect. 5).

4.3 The Importance of People

A significant part of the regions missed by saliency models involve people (Table 2): people within the salient region, or people acting on or looking at a salient object. In this section we provide a deeper analysis of the images containing people. To expand our analysis, we annotated all the people's faces in the MIT300 images with bounding boxes. This provided a more complete set of annotations than the regions extracted for the MTurk labeling tasks, where only the top 1–3 most highly-fixated regions per image were labeled. In this section we compute the importance of faces in an image following the approach of Jiang et al. [18]: given a bounding box for an object in an image, the maximum saliency value falling within the object's outline is taken as the object's importance score (the maximum is a good choice for such analyses as it does not scale with object size). This will be used to analyze if saliency models are able to capture the relative importance of people in scenes.

Across the images in MIT300 containing only one face (53 images), the face is the most highly fixated region in 66 % of the images, and the DeepFix model

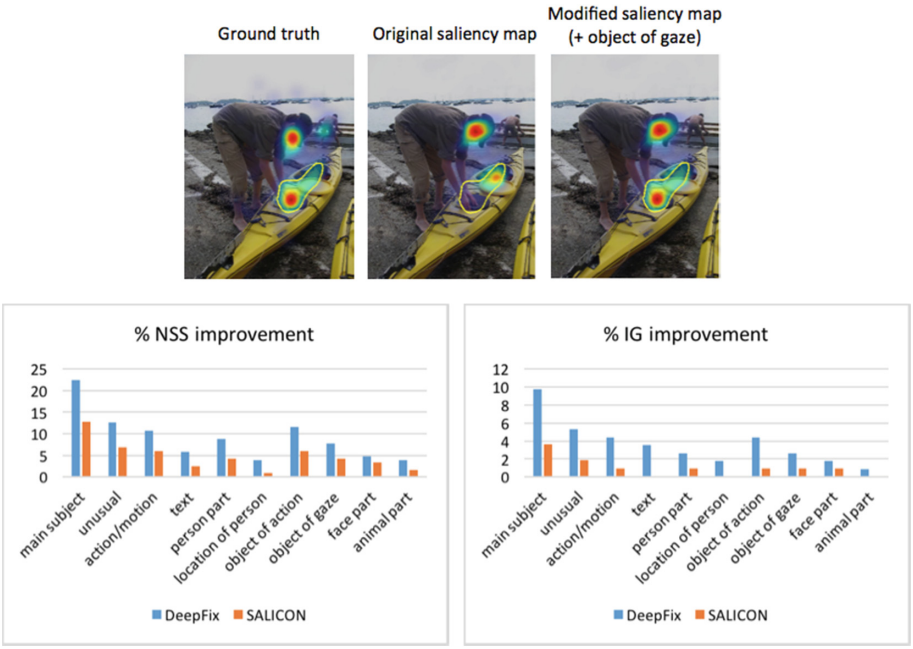

Fig. 5. Improvements of DeepFix and SALICON models on MIT300 if specific regions were accurately predicted. Performance numbers are over all 300 benchmark images, where regions from the ground truth fixation map are substituted into each model's saliency maps to examine the change in performance (top row). The percentage score improvement is computed as a fraction of the score difference between the original model score and the human limit (from Table 1).

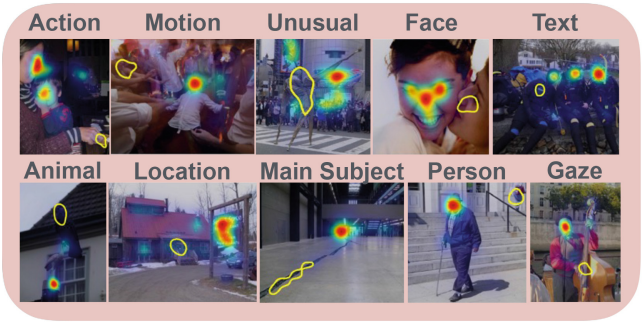

Fig. 6. Regions often fixated by humans but missed by computational models.

correctly predicts this in 77 % of these cases. Out of the 53 images with faces, the saliency of the face is underestimated (by more than 10 % of the range of saliency values) by the DeepFix model in 15 cases, and overestimated in 3 cases. In other words, across these images, the DeepFix model does not assign the

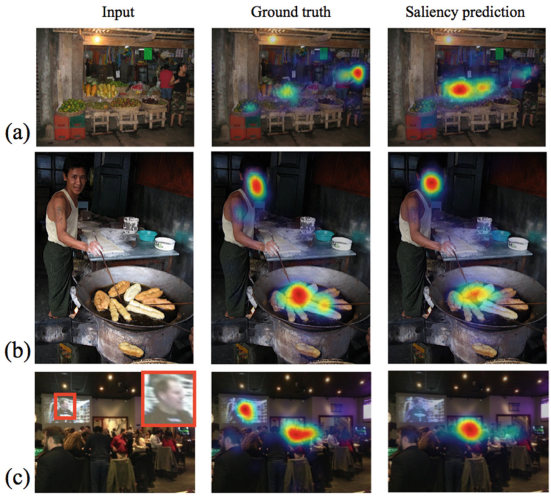

Fig. 7. Saliency prediction failure cases for faces: (a) Face saliency is underestimated when faces are small, non-frontal, or not centered in an image; (b) Sometimes the actions in a scene are more salient to human observers than the participants, but saliency models can overestimate the relative saliency of the faces; (c) Face detection can fail on depictions (such as in posters and photographs within the input images) which often lack the context of a body, or appear at an unusual location in the image.

correct relative importance to the face relative to the rest of the image elements in a third of the total cases. Some of these examples are provided in Fig. 7 and in the Supplemental Material. Note that the importance of faces extends to depictions of faces as well: portraits or posters containing human faces in images. Human attention is drawn to these regions, but models tend to miss these faces, perhaps because they are lacking the necessary context to discover them.

Similarly to the analysis in Sect. 4.2, here we quantify the performance boost of saliency models if the saliency of faces were always correctly predicted. We used the same procedure: to create the modified saliency map for an image we assign the ground truth saliency value to the bounding box region and the predicted output of the model to the remaining part of the image. The DeepFix model's Normalized Scanpath Saliency (NSS) score on the MIT300 benchmark improves by 7.8 % of the total remaining gap between the original model scores and human limit, when adding ground truth in the face bounding boxes. Information Gain (IG) also goes up 1.8 %. A full breakdown of all the scores is provided in the Supplemental Material. Improving the ability of models to detect and assign correct relative importance to faces in images can provide better predictions of human eye fixations.

4.4 Not All People in an Image Are Equally Important

Considering images containing multiple faces, we measure the extent to which the computational prediction of the relative importance of the different faces matches human ground-truth fixations. For all the faces labeled in an image, we use the human fixation maps to compute the importance score for each face, and analogously we use the saliency map to assign a predicted importance score to the same faces. Since both fixation and saliency maps are normalized, each face in an image will receive an importance score ranging from 0 to 1. A score of 1 occurs when the face bounding box overlapped a region of maximum density in the corresponding fixation/saliency map. Interpreted in terms of ground truth, this is the face that received the most fixations.

Across the images with more than one visible face, the average Spearman correlation between the ground truth and predicted face importance values is 0.53. This means that for many images, the relative ordering assigned by the saliency model to people does not match the importance given by human fixations. As depicted in Fig. 8, discovering the most important person in the image is a task that requires higher-level image understanding. Human participants tend to fixate people in an image that are central to a depicted action, a conversation, or an event; people who stand out from the crowd (based on some high-level features like facial expression, age, accessories, etc.).

Fig. 8. Although recent saliency models have begun to detect faces in images with high precision, they do not assign the correct relative importance to different faces in an image. This requires an understanding of the interactions in an image: who is participating in an action and who has authority. Facial expressions, accessories/embellishments, facial orientation, and position in a photo also contribute to the importance of individual faces. We assign an importance score to each face in an image using the maximum ground truth (fixation) or predicted (saliency) density in the face bounding box. These importance scores, ranging from 0 to 1, are included above each bounding box.

4.5 The Informativeness of Text

In the MIT300 and CAT2000 datasets, most text, large or small, has attracted many human fixations, with regions containing text accounting for 7 % of all highly-fixated image regions. While text has been previously noted as attracting human visual attention [33], not all text is equal. The informativeness of text in the context of the rest of the image, or the interestingness of the text on its own can affect how long individual observers fixate it, and what proportion of observers look at it. There are thus a number of reasons why the human ground truth might have a high saliency on a particular piece of text, and some of those reasons depend on understanding the text itself - something that computational models currently lack (Fig. 9).

To expand our analysis on text regions, we annotated all instances of text present in the MIT300 dataset with bounding boxes. The DeepFix model's NSS scores improves by 7.8 % of the total remaining gap between the original model scores and human upper bound, when adding ground truth in the text bounding boxes. Its IG score improves by 4.4 %. A full breakdown of all the scores is provided in the Supplemental Material. Overall, an accurate understanding of text is another step towards better predictions of human fixations.

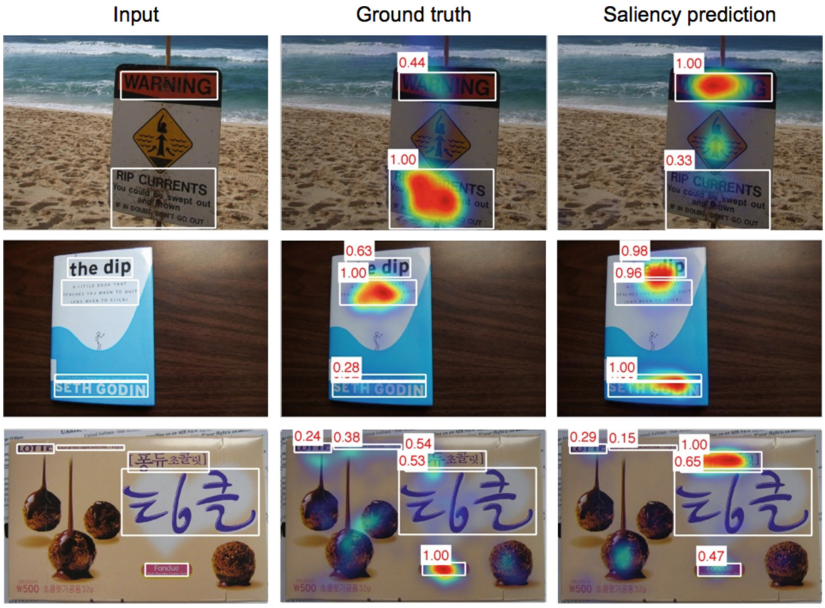

Fig. 9. Example images containing text that receive many fixations by human observers, but whose saliency is under-estimated by computational models. Text labels can be used to give the observer more information. For instance, the description of a warning or a book are more informative to observers than the warning or book title itself. These regions receive more eye fixations. The informativeness of text also depends on the context of the observer: most observers fixated the only piece of English text on the box of chocolates.

4.6 Objects of Gaze and Action

Another common source of missed predictions are objects of gaze and/or action. These are objects or, more generally, regions in an image that are looked at or interacted with by one or more persons in an image. In Fig. 10, we include 4 images from the MIT300 dataset that include objects of gaze missed by both DeepFix and SALICON. In the last column of Fig. 10 we also show the predictions that can be made possible by a computational model specifically trained to predict gaze [34]. For each person in an image, this model predicts the scene saliency from the vantage point of the individual selected (details in Supplemental Material). Training saliency models to explicitly follow gaze can improve their predictive power of modeling the saliency of the entire scene [35].

The gaze-following model only works when gaze information can be extracted from the orientation of the head and, if visible, the location and orientation of the eyes. However, the orientation of the body and location of body parts (specifically the hands) can provide additional clues as to which objects in an image are relevant from the vantage point of different people in the image, even if not fully visible. Detecting such objects of action remains a problem area for saliency models (some failure cases are provided in the Supplemental Material).

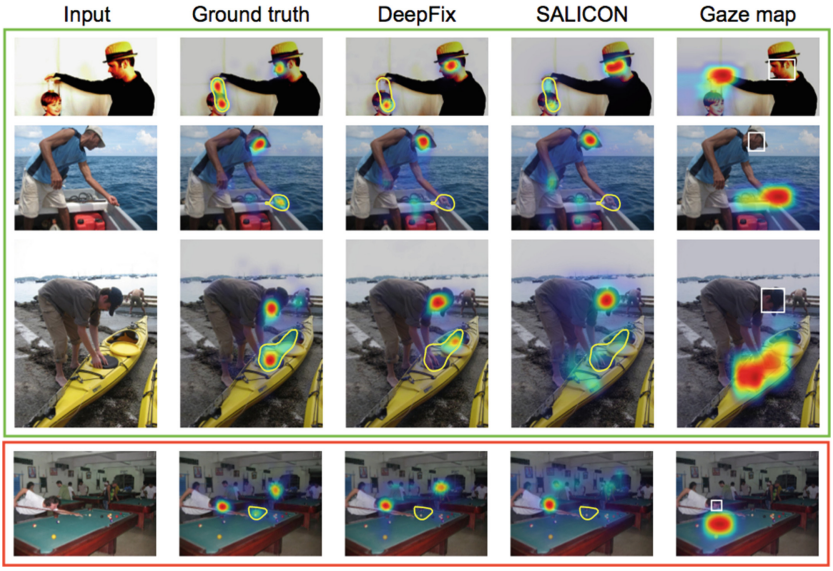

Fig. 10. Both top neural network saliency models perform worse on these images than on any other images in the MIT300 dataset labeled with objects of gaze. The yellow outlines highlight high-density regions in the ground truth fixation map that were labeled by MTurk participants as regions on which the gaze of someone in the image falls. A model that explicitly predicts the gaze of individuals in an image can locate these objects of gaze [34]. The last row is a failure of the gaze-following model, requiring an understanding of actions that is beyond just gaze. (Color figure online)

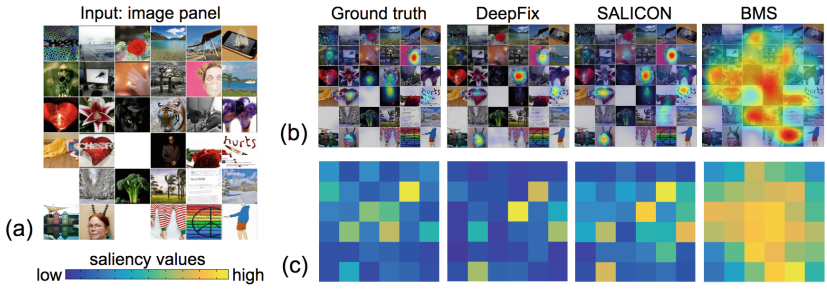

Fig. 11. A finer-grained test for saliency models: determining the relative importance of different sub-images in a panel. (a) A panel image from the MIT300 dataset. (b) The saliency map predictions given the panel as an input image. (c) The maximum response of each saliency model on each subimage is visualized (as an importance matrix).

5 Conclusion

As the number of saliency models grows and score differences between models shrink, evaluation procedures should be adjusted to elucidate differences between models and human eye movements. This calls for finer-grained evaluation metrics, datasets, and prediction tasks. Models continue to under-predict crucial image regions containing people, actions, and text. These are precisely the regions with greatest semantic importance in an image, and become essential for saliency applications like image compression and image captioning. Aggregating model scores over all image regions and large image collections conceals these errors. Moreover, traditionally favored saliency evaluation metrics like the AUC can not distinguish between cases where models predict different relative importance values for different regions of an image. As models continue to improve in detection performance, measuring the relative values they assign to the detected objects is the next step. This can be accomplished with metrics like the Normalized Scanpath Saliency (NSS) and Information Gain (IG), which take into account the range of saliency map values during evaluation [32]. Finer-grained tasks like comparing the relative importance of image regions in a collection or in a panel such as the one in Fig. 11 can further differentiate model performances. Finer-grained datasets like CAT2000 [21] can help measure model performance per image type.

Recent saliency models with deep architectures have shown immense progress on saliency benchmarks, with a wide performance gap relative to previous state-of-the-art. In this paper we demonstrated that a finer-grained analysis of the top-performing models on the MIT Saliency Benchmark can uncover areas for further improvement to narrow the remaining gap to human ground truth.

Acknowledgments. This work has been partly funded by an NSERC PGS-D Fellowship to Z.B., La Caixa Fellowship to A.R., NSF grant #1524817 to A.T., and a Toyota Grant to F.D.

References

1. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: MIT saliency benchmark. http://saliency.mit.edu/
2. Kienzle, W., Wichmann, F.A., Franz, M.O., Schölkopf, B.: A nonparametric approach to bottom-up visual saliency. In: Advances in Neural Information Processing Systems, pp. 689–696 (2006)
3. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE 12th International Conference on Computer Vision, pp. 2106–2113 (2009)
4. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 438–445 (2012)
5. Xu, J., Jiang, M., Wang, S., Kankanhalli, M.S., Zhao, Q.: Predicting human gaze beyond pixels. J. Vis. **14**(1), 1–20 (2014)
6. Zhao, Q., Koch, C.: Learning a saliency map using fixated locations in natural scenes. J. Vis. **11**(3), 9 (2011)
7. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**(1), 97–136 (1980)
8. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurbiology **4**, 219–227 (1985)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 1254–1259 (1998)
10. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vis. Res. **42**(1), 107–123 (2002)
11. Bruce, N., Tsotsos, J.: Attention based on information maximization. J. Vis. **7**(9), 950 (2007)
12. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 185–207 (2013)
13. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. IEEE Trans. Image Process. **22**(1), 55–69 (2013)
14. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2798–2805 (2014)
15. Kümmerer, M., Theis, L., Bethge, M.: Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. arXiv preprint (2014). arXiv:1411.1045
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
17. Liu, N., Han, J., Zhang, D., Wen, S., Liu, T.: Predicting eye fixations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 362–370 (2015)
18. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: saliency in context. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
19. Kruthiventi, S.S., Ayush, K., Babu, R.V.: Deepfix: A fully convolutional neural network for predicting human eye fixations. arXiv preprint (2015). arXiv:1510.02927
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint (2014). arXiv:1409.1556

21. Borji, A., Itti, L.: Cat2000: A large scale fixation dataset for boosting saliency research. arXiv preprint (2015). arXiv:1505.03581
22. Pan, J., Sayrol, E., Giro-i-Nieto, X., McGuinness, K., O'Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
23. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1265–1274 (2015)
24. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5455–5463 (2015)
25. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3183–3192 (2015)
26. Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J.: Deepsaliency: multi-task deep neural network model for salient object detection. IEEE Trans. Image Process. **25**(8), 3919–3930 (2016)
27. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. In: MIT Technical report (2012)
28. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: IEEE International Conference on Computer Vision (ICCV), pp. 1331–1338 (2011)
29. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492 (2010)
30. Zhang, J., Sclaroff, S.: Saliency detection: a boolean map approach. In: IEEE International Conference on Computer Vision (2013)
31. Kümmerer, M., Wallis, T.S., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. Proc. Nat. Acad. Sci. **112**(52), 16054–16059 (2015)
32. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint (2016). arXiv:1604.03605
33. Cerf, M., Frady, E.P., Koch, C.: Faces and text attract gaze independent of the task: experimental data and computer model. J. Vis. **12**(10), 1–15 (2009)
34. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems, pp. 199–207 (2015)
35. Soo Park, H., Shi, J.: Social saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4777–4785 (2015)