DLSU Research Congress 2022
De La Salle University, Manila, Philippines
July 6 to 8, 2022

An Empirical Study on the Misprediction of Text Regions in Modelling Human Attention

Karl D. Galapon[1], Francine S. Carlos[1*], Johanna L. Vendiola[1], Keona A. Bocol[1], Nathan M. Lim[1], Macario O. Cordel II[2]

[1] *Senior High School, De La Salle University*
[2] *College of Computer Studies, De La Salle University*
** Corresponding Author/s: francine_carlos@dlsu.edu.ph*

Abstract: Advances in the computer vision field have yielded saliency models that predict human attention. Current saliency models using deep neural networks have displayed excellent results in the prediction of visual aesthetics, sentiments, and memorability. However, these recent saliency models are non-inclusive of human cognition, such as recognizing semantically meaningful, or simply, informative text. As a result, predicted attention maps often underestimate the saliency of information-dense or informative text. To address this, we compare the predicted attention in text regions with fixation maps generated from actual human eye fixations. Our descriptive and inferential studies reveal that (1) informative text attention level underprediction can only be observed for informative text regions that do not coexist with non-informative text regions. Nevertheless, (2) the saliency model can identify the location of information text regions. These insights on how different objects compete for human attention can be used towards designing a better human attention model.

Key Words: visual saliency, saliency models, attention misprediction, informative text

1. INTRODUCTION

Humans perform image viewing to selectively sort the most relevant features in a scene while limiting the visual attention on other details (Harel et al., 2007). Significantly distinct elements called salient stimuli attract human visual attention and allow humans to concentrate on important information in a scene and accomplish specific tasks. Visual attention is classified into two approaches: bottom-up and top-down. On the one hand, the bottom-up attention is an externally induced process where information is selected based on highly noticeable features of the stimuli. Thus, bottom-up saliency detection models use low-level visual attributes, such as brightness, color, and texture, in generating saliency. On the other hand, top-down attention is internally induced, where information is actively sought out in the scene based on the current task. The top-down attention uses high-level and context-dependent visual attributes, such as the object of human action and gaze, familiar faces, and text[1] information for image saliency detection (Banitalebi-Dehkordi et al., 2016).

Recently, in image saliency detection, many studies have been leaning towards the prediction of high-level image attributes (Fan et al., 2020). Saliency models have turned to using deep neural networks (DNN), which have displayed excellent results in predicting visual realism, visual memorability, visual aesthetics, and visual sentiment. For example, Chen and his colleagues (Chen et al., 2014) introduced DeepSentiBank, a DNN model that classifies visual sentiment concepts by detecting emotions portrayed in the images and utilizing training data to describe the images through an adjective-noun pair (e.g., curious deer, playful dog, and

[1]Ming Jiang et al., SALICON: Saliency in Context, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 1072-1080.

tired eyes). However, current saliency models still appear to disregard one potentially important feature in images: the informative and non-informativeness of text regions. This work quantifies the misprediction of saliency models in predicting attention levels of informative text regions in an image through identifying image text regions and determining the attention score (AS) for each, comparing the ground truth attention map and predicted saliency maps to determine commonly mispredicted image regions with informative text, and characterizing the mispredicted text regions and comparing it with other correctly predicted text regions. This work highlights the misprediction of the presence of informative text in an image, which may greatly affect how human attention is modeled in future studies. Accurately predicting human text detection and attention in natural scenes is necessary to provide the observer with details that are critically useful in their current situation. For this work, fixation maps refer to the given human attention maps from the datasets where the photos were obtained, saliency models refer to the human attention prediction model used to generate "prediction maps", saliency map refers to the maps generated using the saliency models, and fixation location refers to a region in an image where a human focuses their attention upon viewing an image.

2. RELATED WORKS

Since the dawn of deep learning architectures, saliency models have experienced an abrupt increase in attention prediction performance[2] (MIT/Tuebingen Saliency Benchmark, n.d.). These saliency models such as the SALICON (Jiang, et al., 2015), DeepFix (Kruthiventi et al., 2017), and Salnet (Chen et al, 2020), are trained as a whole, thus, disregarding scene semantics that could affect human attention.

First, Bylinskii et al., (2016) showed that saliency models underestimate the prediction of informative text regions, as illustrated in Fig. 1. This is because current saliency models are mostly data-driven without consideration for human cognitive characteristics (Wu et al., 2020), particularly the detection of informative texts. Since not all texts are equal in attracting visual attention, the saliency of a certain text region relies on its informativeness. More importantly, the large gaps in the predictive performances between traditional saliency models and current DNN-based saliency models indicate that higher-level attributes in images do not match with the ground truth fixations made by human observers (Bylinskii et al., 2016). Particularly, the misprediction of attention in an image with text is common in current saliency models, presumably because previous studies on human attention prediction treated images as a whole instead of using image regions. In the process, high-level image attributes were ignored, particularly the presence of informative text in an image, which may have affected the prediction of human visual attention.

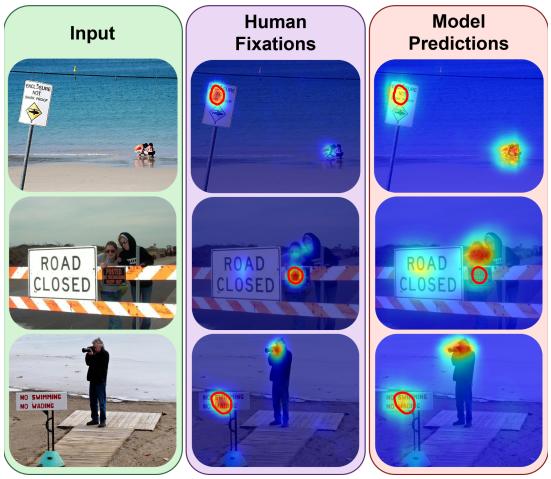

Figure 1. The Saliency in Context (SALICON) model prediction underestimated the attention levels of the text regions (see first and third column). These text regions contain important context-dependent information that attract human attention as reflected in the fixation map (see second column).

Additionally, they studied mispredictions of these saliency models on high-level features including human faces, animals, and text. Motivated by this study, Fan and her colleagues (2020) designed a computational model that predicts high-level image attributes in digital images affecting human perceptions. Their work showed that human perception is influenced by sentiment, memorability, and aesthetics. Furthermore, it was found that natural objects induce just as much excitement as human faces. It covered holistic cues, color information, and semantics of digital images. Meanwhile, Cordel et al. (2019) conducted a study that focused on other high-level features that possibly affect human attention, i.e. the object sentiment. Different objects with varying sentiments receive attention from an individual. Object sentiments, having an influence on human perception, should be accounted for in saliency models.

Though these previous works provide an initial study on high-level attributes that can improve human attention prediction, none of these are focused on informative text. Our work studies human attention prediction on images with text regions.

[2] https://saliency.tuebingen.ai/

3. METHODOLOGY

3.1 Dataset

An assortment of 238 images containing text with corresponding fixations, and ground truth attention maps or heatmaps, are gathered from the following datasets: CAT2000 (Borji & Itti, 2015), Object and Semantic Images and Eye-tracking (OSIE) (Xu et al., 2014), and EMOtional attention dataset (EMOd) (Fan et al., 2018). Image samples should have at least one text region that is legible and written in English language. The groundtruth fixation map and the predicted saliency map using current saliency models are then compared to quantify the mispredictions. We use SALICON to generate the predicted saliency map through Ubuntu 20.04 Machine. SALICON is a neural network model that generates saliency maps by applying convolutional neural networks at fine and coarse image scales. A more detailed discussion on SALICON's architecture can be found at Huang et al. (2015). The implementation uses Caffe framework, trained using the SALICON dataset.

Figure 2. Informative texts provide information about the image's context, such as the cost of squash or the safety precautions of a hazard (see first column, solid line box). Non-informative texts may contain information, but they are irrelevant to the context or do not introduce new information. Examples are shown in the second column, i.e. a number in a baseball game and a label for a dog food bowl.

3.2 Text region annotation and mask

We use labels to define criteria in annotating text regions, *i.e.* informative text and non-informative text. Informative text regions provide information that are context-dependent, which are subjectively determined by the annotator. Otherwise, they are considered as non-informative text regions. Shown in Fig. 2 are samples of informative and non-informative text regions.

For each text region, a bounding box (bbox) is created using LabelImg[3]. These bboxes' information and annotations of each text region are saved as *xml* files and are the basis for the generation of text region masks. The masks are needed to extract the ground truth and predicted attention level of a text region. To generate the mask, *xml* ElementTree python library is used to extract the coordinates of all bounding boxes from the saved *xml* files. Then, an array of zeros, whose size is of the same size as the corresponding image, is generated. Finally, values inside the bounding box coordinates, i.e. text region, are set to 1.0.

3.3 Attention level of text regions

The purpose of the masks is to isolate parts of the attention maps (ground truth or prediction) that correspond to the text regions, as illustrated in Fig. 3. First, we scale the fixation maps to the masks' sizes using *bilinear interpolation*. Note that some fixation maps are stored as arrays with size different from the size of the stimuli. Since their dimensions are equal, attention maps and masks can be multiplied, zeroing out all attention levels outside the bounding box.

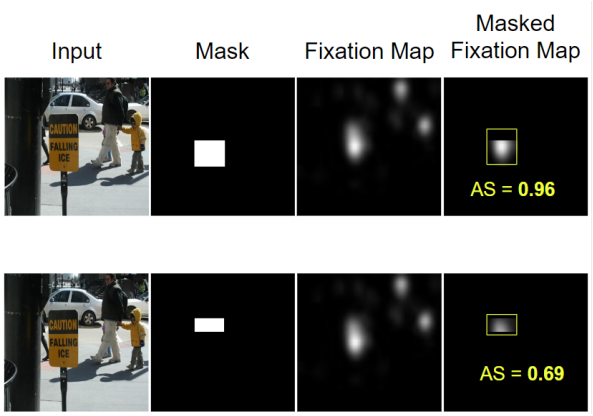

Figure 3. Masking of attention maps through pixel multiplication and determination of AS as the normalized maximum pixel value.

The attention levels of the remaining region are then determined using the attention score (AS) metric. AS is the maximum normalized pixel value of a fixation map or saliency map, T, as described in Eq. (1).

$$AS = \frac{max(T)}{255} \quad (1)$$

[3] https://github.com/tzutalin/labelImg

(a) (b)

(c)

Figure 4. The attention scores of images were grouped into images with informative text regions only, images with non-informative text regions only, and images with co-occurring informative and non-informative text regions. (a) Sample image wherein PINF contains the AS of an image with purely informative text regions. (b) Sample image wherein PNNF includes the AS of an image with purely non-informative text regions. (c) Sample image wherein CINF and CNNF were both based on the co-occurring informative and non-informative text regions of an image. The red solid and broken line boxes indicate informative and non-informative text regions, respectively.

Using the NumPy library, the maximum pixel value was obtained. However, since the pixel values of the mask range from 0-255, and the fixation maps range from 0-1, the maximum pixel value was within 0-255. Thus, we obtained the AS by dividing the maximum pixel value by 255, achieving a normalized 0-1 range. The general equation is shown in equation (1). The AS of text regions from ground truth fixation maps are called the *ground truth AS*. Similarly, the predicted attention levels of text regions are collected and called the *prediction AS*. An AS value closer to 1 indicates that a text region is more salient, while a value closer to 0 indicates that a text region is less salient.

3.4 Statistical analysis

To confirm if the differences between the ground truth and predicted attention levels are statistically significant at some confidence level, hypothesis testing is conducted. The confidence level used for this study is 95% (α=0.05). Since the values within each group are not known to be distributed normally and each group is independent, the non-parametric test called *Mann-Whitney U-test* is used to compare the means of any chosen two groups.

For the statistical analysis, the gathered AS were grouped into four. The first group, called PINF, contains AS of the informative text regions from images having only informative text. The second group, called PNNF, consists of AS of non-informative text regions from images with only non-informative text. Lastly, the co-occurring groups consist of AS of text regions from images having both informative and non-informative text. The AS groups of informative and non-informative text regions in the co-occurring group are called CINF and CNNF, respectively. Refer to Figure 4.

4. RESULTS AND DISCUSSION

4.1 Descriptive statistics

The groupings and their respective descriptive AS data analysis are summarized in Table 1. Using the AS of the ground truth fixation and predicted saliency maps of the 238 images containing a total of 468 text regions, descriptive statistics of all groups (PINF, PNNF, CINF, and CNNF) of both ground truth and prediction AS were calculated.

Table 1. Descriptive statistics of the AS extracted from the ground truth and predicted attention maps for all types of text regions. μ means the mean, σ means standard deviation and N is the number of samples. Δp is the absolute difference between PINF and PNNF. Δc is the absolute difference between CINF and CNNF.

		PINF	PNNF	Δp	CINF	CNNF	Δc
Contains informative text?/ extracted AS?		Yes/ Yes			Yes/ Yes	Yes/ No	
Contains non-informative text?/ extracted AS?			Yes/ Yes		Yes/ No	Yes/ Yes	
Ground truth attention map	μ	0.75	0.57	0.18	0.69	0.45	0.24
	σ	0.29	0.34		0.30	0.33	
	N	111	214		61	82	
Predicted attention map	μ	0.66	0.57	0.09	0.70	0.54	0.16
	σ	0.19	0.20		0.18	0.18	
	N	111	214		61	82	

First, the ground truth groups are analyzed. The AS of informative and non-informative text regions both decrease when they co-occur. From the mean AS of the different groups, the average AS of regions from PINF

images is reduced from 0.75 to 0.69 when they co-occur (CINF column).

The same occurred in regions from non-informative text regions. The average AS in PNNF is reduced from 0.57 to 0.45 in CNNF. However, the difference in attention levels between the two types of text regions increases when they co-occur in an image (see Δp and Δc). This implies that the competition for human attention is more evident when the two types coexist in the same image. It is also observed that higher attention score is received by informative text regions vs. non-informative text regions.

The same observation is also seen in predicted attention maps, when the gap between the AS of the informative and non-informative text regions is compared (see Table 1 Δp columns for the predicted attention maps). Lower AS occurred when text was non-informative. Both purely informative, PINF, and co-occurring informative, CINF, text regions attract more attention than their non-informative counterparts, PNNF and CNNF. Furthermore, AS of the informative text regions are generally higher than the non-informative text regions.

However, when the level of AS in ground truth is compared with the predicted attention map for images with informative text only (see Table 1 PINF column), the predicted AS is lower. This is not seen on the AS levels of the text informative text regions that co-occur with non-informative text regions. This suggests that the observation of Bylinskii and her colleagues (2016), that *informative text regions are underpredicted in saliency prediction models, can only be observed for informative text regions that do not coexist with non-informative text regions*. These non-informative text regions presumably benchmark the attention level induced by text regions, in both the ground truth and predicted attention maps.

The interestingness of informative text regions, as reflected by its AS and are mispredicted by attention models, seemingly comes from how these saliency models learn weights as a whole by extractacting low level features contour, color, and contrast and not the context.

In addition, fine grain analysis on the number of text regions with high value of AS shows that informative text regions with high AS (≥ 0.5) are more common than low AS (< 0.5). And this is observed in both ground truth and predicted AS. Refer to Figure 5. Thus, *in terms of the number of high AS values, although the saliency models mispredicted the attention levels of informative text regions, it does identify the location of informative text regions.*

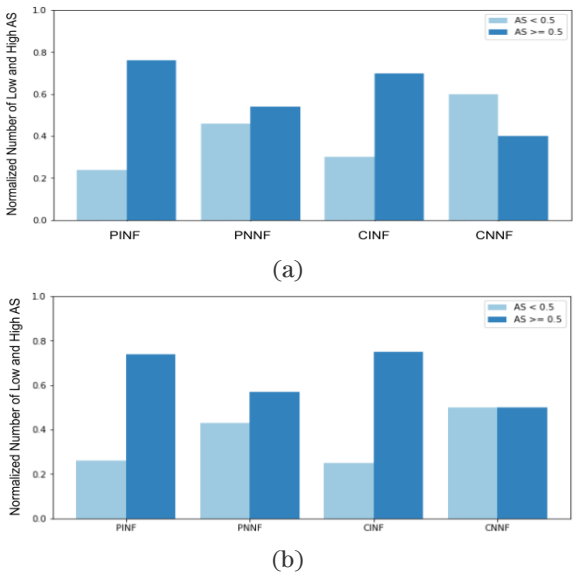

Figure 5. Histogram of the ground truth AS (a) and predicted AS (b) of text regions. Note that PINF and CINF – both the informative text regions AS, have a much higher number of images with high AS.

4.2 Hypothesis testing

Hypothesis testing is crucial to determine if the earlier differences discussed were statistically significant. Our data has no assumption on the normality of its distribution, thus, the Mann Whitney U-test, a non-parametric test, was used. For clarity, our null hypothesis (H_0) is that there is no significant difference between the two groups.

We conducted two statistical tests to check if there is significant difference between the (1) attention levels of informative text regions and non-informative text regions (see Table 2) and the (2) text regions of informative and non-informative texts (see Table 3 and Figure 6).

Table 2. Mann-Whitney U-test on AS of Informative and Non-informative Text in all categories

	AS Group	Statistical Significance
Ground Truth AS	PINF vs. PNNF	$p < 0.01$
	CINF vs. CNNF	$p < 0.01$
Predicted AS	PINF vs. PNNF	$p < 0.01$
	CINF vs. CNNF	$p < 0.01$

Table 3. Average normalized area in a bounding box for all types of text regions .Normalized area is the number of pixels of the text region divided by the amount of pixels in an image. A normalized area of 1 means the text region occupies the entire image.

	PINF	PNNF	CINF	CNNF
μ	0.062	0.025	0.053	0.018
σ	0.090	0.053	0.073	0.023
N	111	214	61	82

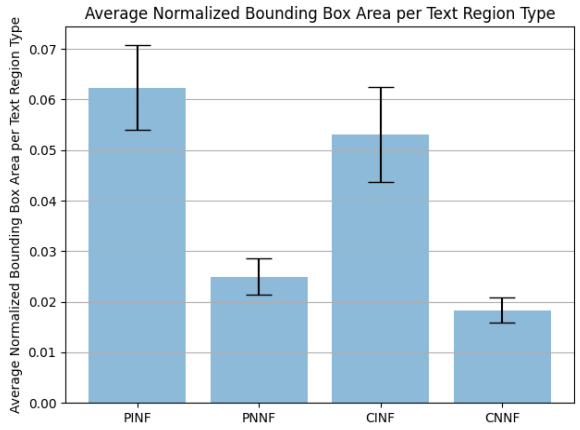

Figure 6. Average bounding box area (length×width) normalized by the size of the image (length×width). Informative text regions, PINF-Area and CINF-Area, are significantly larger than other text regions without informative text, with statistical significance ($p < 0.01$).

From Table 2, the *p*-value derived from the tests were all less than 0.01 ($p < 0.01$). Thus, the alternative hypothesis was accepted for all comparisons. There exists significant differences between: Ground Truth PINF vs. PNNF, Ground Truth CINF vs. CNNF, Prediction PINF vs. PNNF, and Prediction CINF vs. CNNF. Since for all comparisons, the mean AS values of the informative text is greater than the non-informative text, informative text was statistically generally more salient in terms of AS score, be it sole, co-occurring, ground truth or prediction.

The ground truth attention levels of informative text regions differ from non-informative text regions due to their higher information density and saliency. Currently, no computer vision procedure distinguishes informative text from the others.

The main differences observed were the number of keywords per region and the area per bounding box. The pixels in a text region were calculated by multiplying the length and widths derived from the bounding box coordinates. Hypothesis testing using the Mann-Whitney U-test was done for each variable.

From Table 3, the number of pixels occupied by informative co-occurring and non-co-occurring regions was numerically larger than the non-informative text regions. These results allude to the possibility that informative texts occupied more than non-informative texts. However, hypothesis testing was required to confirm these differences statistically.

Furthermore, the null hypothesis was rejected after evaluating the *p*-values of both tests ($p < 0.01$), see Table 3. This result means that there exists a significant difference between the bounding box areas of informative text and non-informative text. That is, the areas of informative text regions were statistically higher than that of non-informative ones.

5. CONCLUSIONS

Current attention models are trained end-to-end and thus, do not provide insights on how high-level information, such as informative text regions, affects human attention. Images with text regions were manually selected. After AS of each text region was calculated, the predicted value was compared to the ground truth value and statistically analyzed.

The following findings were revealed: Human attention on text regions is greater when only informative text is present. When both informative and non-informative text regions are present in an image, attention level is lower, implying attention competition. Images that contain only *informative* text regions possess a greater attention level than those images with only *non-informative* text regions. When these informative and non-informative text regions co-occur, their ground truth attention levels are lower, with informative text regions having higher than non-informative text regions' attention level. Lastly, gaps in saliency model's ability to distinguish informative and non-informative text were found. Therefore, this competition dynamic between texts presents considerable potential for a design of a saliency model that encodes informative text regions and integrates this information into saliency prediction.

7. REFERENCES

Banitalebi-Dehkordi, A., Pourazad, M. T., & Nasiopoulos, P. (2016). A learning-based visual saliency prediction model for stereoscopic 3D video (LBVS-3D). *Multimedia Tools and Applications, 76*(22), 23859–23890.

Borji, A., & Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581.*

Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., & Durand, F. (2016). Where Should Saliency Models Look Next?. *In the European Conference on Computer Vision (ECCV 2016).*

Chen, T., Borth, D., Darrell, T., & Chang, S. F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint.*

Chen, F., Jiang, Y., Zeng, X., Zhang, J., Gao, X., & Xu, M. (2020). PUB-SalNet: A Pre-Trained Unsupervised Self-Aware Backpropagation Network for Biomedical Salient Segmentation. *Algorithms*, *13*(5), 126.

Cordel, M., Fan, S., Shen, Z., & Kankanhalli, M. (2019). Emotion-Aware Human Attention Prediction. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Fan, S., Shen, Z., Jiang, M., Koenig, B. L., Xu, J., Kankanhalli, M. S., & Zhao, Q. (2018). Emotional Attention: A Study of Image Sentiment and Visual Attention. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Fan, S., Koenig, B.L., Zhao, Q., & Kankanhalli, M. S. (2020). A Deeper Look at Human Visual Perception of Images. *SN Computer Science, 1*, 58.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in Neural Information Processing Systems, 19*, 545-552.

Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting Deep Neural Networks. 2015 IEEE International Conference on Computer Vision (ICCV).

Jiang, M., Huang, S., Duan, J., & Zhao Q. (2015) SALICON: Saliency in Context, In proc. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1072-1080.

Kruthiventi, S. S., Ayush, K., & Babu, R. V. (2017). DeepFix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing, 26*(9), 4446–4456.

Lin, T. (2015). LabelImg. Git code. https://github.com/tzutalin/labelImg.

Wu, S., Fan, S., Shen, Z., Kankanhalli, M., & Tung, A. K. (2020). Who You Are Decides How You Tell. *In Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 4013–4022.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision, 14*(1), 1–20.

Yang, J., & Yang, M.-H. (2017). Top-Down Visual Saliency via Joint CRF and Dictionary Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(3), 576–588.