Finding Relevant Information in Saliency Related Neural Networks

Ron M. Hecht
General Motors
Herzliya, Israel
ron.hecht@gm.com

Ariel Telpaz
General Motors
Herzliya, Israel
ariel.telpaz@gm.com

Dan Levi
General Motors
Herzliya, Israel
dan.levi@gm.com

Ronit Bustin
General Motors
Herzliya, Israel
ronit.bustin@gm.com

Gershon Celniker
General Motors
Herzliya, Israel
gershon.celniker@gm.com

Omer Tsimhoni
General Motors
Warren, MI
omer.tsimhoni@gm.com

Ke Liu
General Motors
Warren, MI
ke.liu@gm.com

Abstract

Over the last few years, many saliency models have shifted to using Deep Learning (DL). DL models can be viewed in this context as a double-edged sword. On the one hand, they boost estimation performance and on the other hand have less explanatory power than more explicit models. This drop in explanatory power is why DL models are often dubbed implicit models. Explainable AI (XAI) techniques have been formulated to address this shortfall. They work by extracting information from the network and explaining it. Here, we demonstrate the effectiveness of the Relevant Information Approach in accounting for saliency networks. We apply this approach to saliency models based on explicit algorithms when represented as neural networks. These networks are known to contain relevant information in their neurons. We estimate the relevant information of each neuron by capturing the relevant information with respect to first layer features (intensity, red, blue) and its higher-level manipulations. We measure relevant information by using Mutual Information (MI) between quantified features and the label. These experiments were conducted on a subset of the CAT2000 dataset.

1 Introduction

The cones (sensors) in the human eye are not equally distributed across the retina. Many are confined to a small central area of the retina known as the fovea. The fovea's high density of cones is unique. Given the uniqueness of the fovea, it is important to allocate it wisely to observe objects in the field of view.

In non-task-oriented gaze behavior, humans tend to implement a bottom-up approach. In this case the fovea is directed towards the most salient regions. These include bright, red or blinking regions [11]. Foveal vision can be mapped to estimate the field of view where the value of each point on the map is its level of saliency. Research on the functionalities of saliency maps goes back several decades [5][12][11][9]. Specifically, one of the goals of saliency mapping is to estimate gaze distribution as a

Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

function of a visual stimulus . Saliency prediction algorithms can be divided into several families, two of which are those that have implicit or explicit descriptions of these algorithms.
One of the first explicit saliency prediction algorithms was introduced by Koch and Ullman [11]. They explored feature-specific saliency maps (such as color, orientation, and disparity, heuristic priors towards center), and modeled their relationship with an overall saliency map. This family of algorithms also explores a second layer in that they are applied on top of the feature specific saliency maps to extract information from them. Examples include center surround, feature selection and augmentation, and probability mass concentration[7]. Higher layer algorithms have also been addressed in this family as well [9]; see [2] for a detailed review.
The implicit family of saliency prediction algorithms has gained momentum over the last few years as a result of the introduction of Deep learning (DL). Here, in what constitutes a paradigm shift, the researcher no longer constructs a saliency or higher layer algorithm and then evaluates them. Rather researchers train a multi layered Neural Network (NN) algorithm that learns the functions connecting images and their saliency distribution. The reference saliency distribution in this case was a set based on human eye gazing patterns. Participants were presented with a set of images and their eye gaze patterns and fixations were recorded. Most of these NNs had many neurons and layers. They were trained to improve the estimation of the saliency distribution.
In some cases, the NNs are initiated from parts of other networks. For example, in [12] part of the VGG-16 network was used to initiate the first stages of the net, and to some extent was used to initiate the preprocessing. One of the downsides of this approach is that it makes it more difficult to intuitively understand and explain the algorithms that are executed. Some of the explicit aspects are sacrificed to boost performance. Luckily, this downside proved to be a blessing in disguise because they prompted the invention of a set of techniques to cope with this problem. These are known as eXplainable Artificial Intelligent (XAI). They explore and explain the behavior of networks, not only qualitatively, but quantitatively as well [16]. They were so overwhelming effective [15], that they were applied in explicit models as well [1][4][13].
Here, we applied XAI to the explicit family to gain quantitative insights. In what follows, we present the behavior of the explicit model in two stages. First, we express the explicit model as an implicit model; i.e., we represent the explicit model as a neural network. Then, we evaluate the importance of each part of the network. This dual stage process generates neural networks. These networks are simple, relatively small, and have small sets of parameters. This simplicity amplifies the XAI's power and links prediction power to specific explicit features.

2 Model

As stated above, our first step was to turn the explicit saliency model into a Convolution Neural Network (CNN). We generated a CNN inspired by [11]. Specifically, our CNN has three stages (Figure 1(a)). In the first stage, the lower level saliency features are extracted, which in our case were red, blue, and intensity. The second stage was associated with derivation (Center Surround (CS), [10] Constant False Alarm Rate (CFAR) [14]). During the third stage, the signal from several saliency channels was combined into a single unified channel by summing all channels.
In most of the saliency literature, it is assumed that color related saliency maps include red, blue, and intensity related saliency maps. Let's denote these as M_r,M_b,M_i respectively. However, this description is a bit coarse. Here, we defined the three color related saliency maps to be:

$$M_i = \frac{r+g+b}{3} \tag{1}$$

$$M_b = b - \frac{r+g}{2} \tag{2}$$

$$M_r = r - \frac{g+b}{2} \tag{3}$$

where r,g, and b are the red, green, and blue values of a pixel respectively. Figure 1(a) presents a neural network implementation of the color related saliency maps as described in equations 1 - 3. The channels are defined as the linear representations of the original r,g, and b. This computation can be performed one pixel at a time, so there is no need to look at the pixel vicinity.

We use the terms CS and CFAR interchangeably. There are many possible implementations of CS, but we selected the filter presented in Figure 1(a). In the last stage, we simply summed the data from

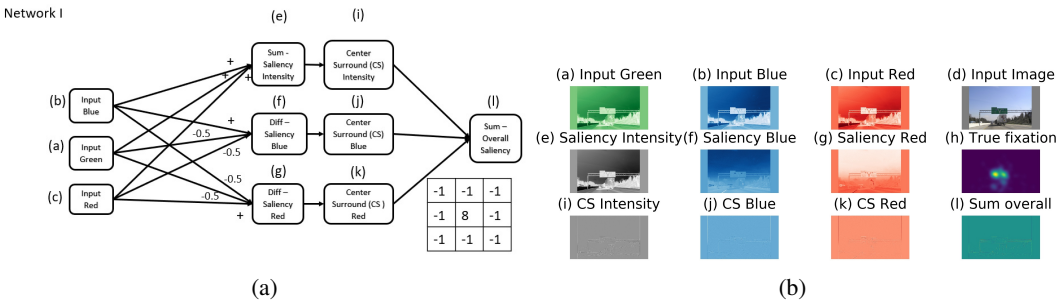

Figure 1: Diagram of Convolutional Neural Network (CNN) representing a saliency algorithm based on intensity and color and Visualization of the activation maps. (a)(b)(c) are the raw input data. (e)(f)(g) are the single feature saliency maps. (i)(j)(k) introduces center- surround to address the relative aspects of human perception. (l) is the agglomeration of all the saliency maps into a single map.

the different channels (Intensity, Red, Blue). Overall, we generated a very small CNN that only has ten neurons. This network is not only small, but simple . The perceptual field of each neuron is limited to a few values in the former layer, yielding simple filters. Although small and simple, it still conveys some of the essence of explicit saliency algorithms. Note that this network is completely linear and therefore can be collapsed. However, in its current structure it provides insights into the flow of information in it.
The network presented in Figure 1(a) has several activation maps. Figure 1(b) visualizes them for a specific example. Subplots (a), (b), (c), (d) present the input. The first three present the 3 channels of the input. The fourth subplot shows the three channels combined. Subfigures (e), (f), and (g) show the saliency maps associated with intensity, blue, and red as articulated in equations 1, 2, and 3 respectively. (i),(j), and (k) are associated with CS. (l) presents the final output of the network. (h) is the ground-truth of the fixation distribution.

3 Method

Our experiment was conducted on the CAT2000 dataset[3]. This dataset was specifically designed for saliency model training and evaluation. It is quite a large corpus (4000 images), and its training set is composed of 2000 images. We conducted our experiments on a subset of 300 images composed of the Indoor, Outdoor Natural, and Outdoor Man-made categories.
We used the eye fixations that are part of the dataset. The eye fixation data in the dataset were collected from 120 participants (80 females, median age 20). Each image was observed by 24 participants. During the data collection, each image was presented for a duration of five seconds followed by three seconds of gray screen. Eye tracking data were collected using an Eyelink-1000 device with a sample rate of 1000Hz. The image resolution was 1920×1080.

The explainability measure used in this work was the Mutual Information (MI) between the label (saliency score) and the activation values of each of the neurons (a,b,c,e,f,g,i,j,k,l) [16][8][9][17]. Both the labels and the activations are continuous values. The MI was estimated using quantization of the variables that was followed by discrete MI estimation of the discretized variables. The dynamic range of each variable was divided into 100 equally spaced bins. For each neuron activation variable, a 100×100 co-occurrence matrix was generated. It counted the co-occurrence between the discretized activation values and the discretized saliency scores. The MI values were estimated from these matrices which were then smoothed by adding one to each bin.

4 Results

Our goal in this work was to characterize the flow of relevant information in the network. This flow was measured by estimating the MI between the activation of each neuron and the desired activation of the output layer; i.e. the probability of the evaluated pixel in the fixation probability map that was estimated based on human fixations. The MI of the activations of the different neurons is presented in Figure 2 (a). Each bin in the figure represents a specific neuron. The different layers are separated

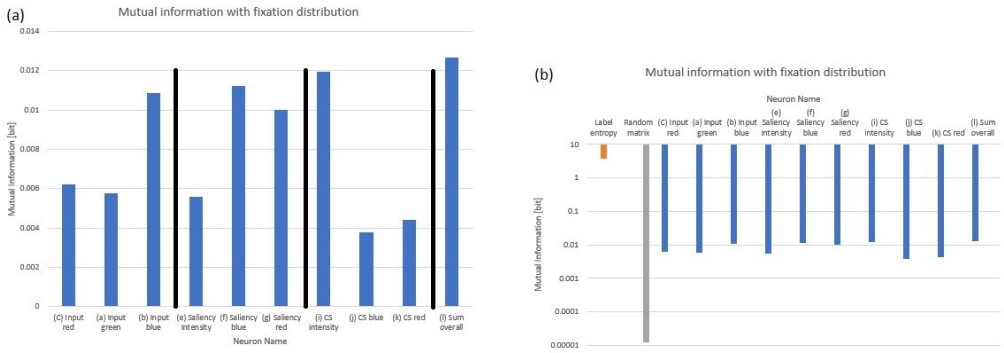

Figure 2: (a) Mutual information between neuron activations and the corresponding saliency score. (b) Activation MI alongside upper and lower bounds on a logarithmic scale. On one end the entropy of the label bounds the MI from above. At the other end, the confidence interval based on the random matrices, bounds the MI from below.

by a vertical solid black line. The layer id numbers are presented at the top, where the input layer is denoted as layer 0. Before exploring our findings, we clarify their validity in terms of significance. As stated above, about 300 images were evaluated in this experiment. The resolution of each image was about 1920×1080, yielding a set of about $1920 \times 1080 \times 300 \approx 6 \times 10^8$ samples (pixels) for the estimation of the co-occurrence matrices. A total of one hundred matrices were generated through this process. The means and standard deviations of the MI were extracted. The mean MI and standard deviations were 1.18×10^{-5} and 1.7×10^{-7} respectively. Under the i.i.d. assumption, the mean MI of the random matrices plus three standard deviations bounded the significance margin. From the MI of the random matrices, we derived the lower bound (Figure 2(b)). The entropy of the label was bound from above the MI of the label and the activation of a neuron. Here, the total amount of information (entropy) of the fixation labels according to our model and data preparation was about 3.68 bits (Figure 2(b)). Overall, we showed that the observed MI values were distant from both the upper and lower bounds. The observed MI had only about one hundredth of the information of the entropy label. This is not surprising given the extremely small receptive field of our network. In an image of 1920×1080, our receptive field is only 3×3. The lower bound is about one hundredth of the observed MI. Under the i.i.d. assumption, this difference is significant and meaningful. The entire flow of information was governed by the data processing inequality (DPI) [6]. It suggests that during data processing, data are lost in the weak sense; i.e., information cannot be gained, just not lost. The transformation from layer 0 to layer 1 was linear and reversible, suggesting that information was not lost. In the transformation to layer 2, information was gained from adjacent pixels. This introduction of a new source of information, can cause increases in data relative to the original signal. In the last layer, no new source of information was added. In this layer, the averaging across the different channels causes a concentration of information into one dimension but it also causes a loss of information as well. It is interesting to observe that the MI of the neurons in the last layer was higher than each of the neurons in the first layer. As stated earlier, we did not measure the MI directly. Here, the MI was only approximated using quantization processes which could have introduced noise to the process.

5 Conclusion

Saliency prediction neural networks have large sets of parameters, and in many cases, there is no single neuron whose removal changes the score significantly. These qualities are one of the reasons why such neural networks are referred as implicit models, and are complicated for humans to explain and understand. Earlier models had far fewer parameters, and each parameter had a humanly comprehensible explanation and hence are known as explicit models. In this work, we analyzed the flow of information within a saliency prediction neural network and explained its behavior. This type of analysis is important to overcome the implicit nature of neural networks. Specifically, we presented a method that involved building an implicit model that was not trained on samples. Its synaptic weights and perceptual fields were selected to imitate a known explicit model. In this network, we already knew the meaning and importance of each neuron. We knew that each neuron contained

relevant information. In this work, we were able to show that we could measure some of the relevant information by estimating the quantified mutual information between the neuron activation and the sample label. This suggests that such an approach can be used to explore larger saliency oriented neural networks. At the most basic level, we can measure the MI between every neuron and the saliency label. At a higher level, the MI can be measured between an entire layers in the saliency network and the saliency label. At the highest level, we can compare to internal representations in other networks.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[2] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012.

[3] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015.

[4] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[5] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[6] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[7] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[8] Ron M Hecht and Naftali Tishby. Extraction of relevant speech features using the information bottleneck method. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[9] Ron Moshe Hecht, Aharon Bar Hillel, Ariel Telpaz, Omer Tsimhoni, and Naftali Tishby. Information constrained control analysis of eye gaze distribution under workload. *IEEE Transactions on Human-Machine Systems*, 49(6):474–484, 2019.

[10] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *IEEE transactions on image processing*, 6(3):451–462, 1997.

[11] Christof Koch and Shimon Ullman. Selecting one among the many: A simple network implementing shifts in selective visual attention. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, 1984.

[12] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017.

[13] T Nathan Mundhenk, Barry Y Chen, and Gerald Friedland. Efficient saliency maps for explainable ai. *arXiv preprint arXiv:1911.11293*, 2019.

[14] Ramon Nitzberg. Constant-false-alarm-rate signal processors for several types of interference. *IEEE Transactions on Aerospace and Electronic Systems*, (1):27–34, 1972.

[15] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 5–22. Springer, 2019.

[16] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[17] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.