FROM BOTTOM-UP TO TOP-DOWN: CHARACTERIZATION OF TRAINING PROCESS IN GAZE MODELING

Ron M. Hecht[1] *, Ke Liu*[2] *, Noa Garnett*[1] *, Ariel Telpaz*[1] *, Omer Tsimhoni*[2]

[1]*GM-R&D ATC-I, Israel*
[2]*GM-R&D WTC, MI, USA*

ABSTRACT

During training, artificial neural networks might not converge to a global minimum. Usually, using gradient descent, the training procedure cause the network to stroll in the high-dimensional weights' space. This stroll passes adjacently to local minima and locations in the geometry of loss landscape associated with low loss. Overall, the network moves from one low loss area to a lower loss area.

In this work, we explored those low loss areas and minima, and tried to understand them. A U-Net was trained based on a gaze prediction task. A network was presented with images of different scenes, and the purpose of the network was to predict the expected human gaze distribution over those images. The driving task was selected since it involves relatively strong goal-oriented behaviors. It was shown that the training had two stages: (1) At the beginning, the network selected area was associated with saliency distributions (bottom-up behavior); (2) Later, the network selected area had the characteristics of goal-oriented distributions (top-down behavior) and it shifted away from the saliency distributions.

***Index Terms*—**eye tracking, U-Net, saliency prediction, global minima

1. INTRODUCTION

Neural networks learn by changing their weights locations in the weight space. Usually at the beginning of training, they start at random locations, and progress iteratively toward lower loss areas in this space by using algorithm from the gradient decent family. At the end of each iteration, the networks' weights are located at a specific locations (points) in weights' space. Understanding the nature of those locations and the relations among them is a long-standing question [1] [2] [3] [4] [5].

A special derivative of the larger question relates to networks that predict eye gaze distributions over images of scenes. Generally, human gaze distribution behavior can be generated by two types of behaviors: goal-oriented behavior, and goal-free behavior [6]. Goal-free behavior is also known as bottom-up behavior or saliency behavior. In this type of behavior, the observer just looks at the environment without trying to perform any task. An example for the second type of behavior, the goal oriented (i.e.,top-down) behavior, can be driving [7] [8] [9] [10] [11]. Those types of behaviors were studied using neural networks (for review [12]).

In this work, we focus on understanding networks when training them to predict the goal-oriented behavior of driving. Specifically, the research questions to address are: (1) While training eye tracking data in a goal-oriented task (i.e., driving), can several types of locations and low loss areas be detected? (2) What are the characteristics of the gaze prediction models in a goal-oriented task reflected by the early visited locations and areas and later visited ones?

An interesting analogy emerges from developmental psychology: When compared to adults', Children's visual attentional capacities are limited [13] [14] [15] [16]. They are more likely to shift to a salient and less important areas in a scene and lose focus of their goal. It is interesting to evaluate whether similar phenomena could be observed in artificial neural network. Our assumption was that networks early in the training were like children, and networks towards the end of the training could achieve the characteristics of adults.

Our paper has the following structure. We start by laying the needed foundations. At section 2, we present our corpus and its labeling, and at section 3, we define our saliency model and saliency comparison method. Later, in section 4, the network training and architecture are discussed. Finally, the results are presented and discussed in sections 5 and 6.

2. CORPUS

In this work, we used an internally collected dataset. Table I presents the corpus by the numbers. The training data of the corpus consists of about 4,500,000 frames of gaze and about 1,500,000 frames of cameras. It was collected during 69 rides of varying duration. The rides were driven by six different drivers. The rides were conducted both during daytime and at night. Recordings were performed in urban and suburban environments. The eye gaze directions were estimated using a single camera system at 60 Hz. The data was collected over a period of several months.

TABLE I
CORPUS CHARACTERISTICS

# drivers	6
# training rides	69
# validation rides	16
#gaze training frames	4.6M
#training data length	21 hours
Gaze sample rate	60Hz
#training images	1.5M
Camera sample rate	20Hz
Lighting condition	Day, Night

Figure 1 shows several image examples that represent a spectrum of recording conditions. In addition, a validation held-out set was collected as well using the same collection mechanism. It was composed of 16 rides. The first 600 frames from each of the ride were used (10 seconds), yielding 160 seconds of data and 9600 frames. This validation set's size provides a reasonable balance between diversity of the data and feasibility of processing.

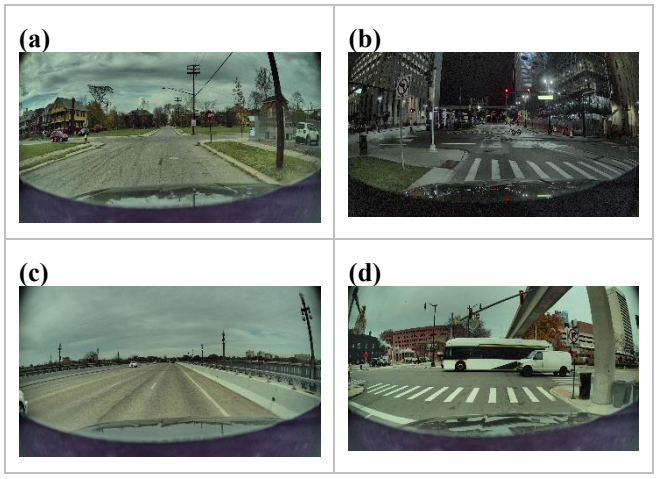

Figure 1. Recording condition (a) suburban during daylight (b) urban during night (c) bridge during daylight (d) urban during daylight.

3. SALIENCY ESTIMATION AND COMPARISON

In this section, both the saliency estimation and comparison are presented. Saliency estimation is the generation of saliency map. Saliency comparison refers to the comparisons between the network outputs and the pre-computed saliency map.

3.1. Saliency estimation

Our estimation of saliency and saliency map generation were inspired by [17] [18] [19]. Estimation of saliency distribution is known for many years. There are different methods for such estimation. Some used explicit approaches [17] [18] [19], while others were more implicitly using deep learning approach and neural network [20]. In this work, we pursued the explicit approach. Specifically, our saliency estimation had the following states:

First, four channels were generated. A single channel was generated for each color (red, green, and blue). In addition, a fourth channel was generated for the intensity. Intensity was defined as the maximal value of red, green, and blue. The channels were demoted as r, g, b, i for red, green, blue, intensity, respectively.

Each of the four channels was then down sampled twice, yielding additional channels at half and at quarter the resolution of the original image. Overall, at this stage, there were 12 channels 3×4 (original resolution, half resolution, quarter resolution) $\times$ (4 original channels).

The next step was generation of the feature maps. There were four feature maps: red, blue, intensity, orientation. The blue feature map M_b was defined as $M_b = b - 0.5g - 0.5r$. similarly, the red feature map M_r was defined as $M_r = r - 0.5g - 0.5b$. The intensity map M_i was defined as the average of the three colors: $M_i = (r + g + b)/3$. The orientation feature map M_o was generated using Gabor filters. A set of eight Gabor filters were convolved with the intensity maps and summed to generate a single output per resolution. At the end of this stage there were 12 maps: 3 M_r, 3 M_b, 3 M_i, 3 M_o at three resolutions.

The next step was Center Surround [21]. Four different filters were convolved with each of the 12 maps. The different center surround filters differ in the size of their center and surround areas. At this stage, there were 12 maps for each of the four feature map types (M_r, M_b, M_i, M_o). i.e., there are 48 maps in total.

The last stage was the combination of the 48 maps into one single map. At first, each of the 48 maps was normalized. For each map two values were extracted: ma, me – the maximal and mean value of the map, respectively. The new value of a pixel nv was based on the old value ov, and was defined as:

$$nv = \frac{ov}{ma}(ma - me)^2 \qquad (1)$$

The goal of the normalization was to boost maps where differences in the map were more dominant. After the normalization, the maps of each map were summed. Therefore, four maps, one for each feature map type (M_r, M_b, M_i, M_o) were obtained. Finally, the four images were summed with equal weights. An example of the original image and final saliency map are shown in Figure 2.

3.2. Saliency comparison

This sub-section presents our mechanism of comparing between the network outputs and the precomputed saliency maps. The input for the comparison was a single saliency map

and a single network output. The output was a saliency score that measures similarity of the network output and the saliency map. High values suggested high similarity.

(a) (b)

Figure 2. Example of an (a)original image and (b) final Saliency map.

The output of the network is composed of two matrices. One matrix holds the positive score O_p and the other the negative score O_n. A Softmax like processing was used to present the activation of each neuron: $A = e^{O_p} + e^{O_n}$. This activation is normalized by subtracting the minimum of A: $\acute{A} = A - min(A)$.

The saliency presented at the previous section was dilated by a 10x10 matrix of ones, which was denoted using S_p. Then, a negative saliency could be generated: $S_n = 1 - S_p$.

At the final step, the saliency similarity score was estimated. It measured the similarity between S_n, S_p and $\acute{A}$.

$$SC = \frac{\sum S_p \circ \acute{A}}{\sum S_p} - \frac{\sum S_n \circ \acute{A}}{\sum S_n} \quad (2)$$

where $\circ$ is the Hadamard product and SC is the final saliency similarity score.

The saliency similarity score indicated the difference between the normalized score for the saliency vs the non-saliency. This score exceeded the expected framework of cross entropy. It was created based on data. However, it was not created based on the held-out set. This set was used only during the validation. At that time the score mechanism was already finalized.

4. METHOD

The corpus used in this experiment is presented at Section 2. There, the participants, and apparatus used for data collection are described. In this section, our experiment process is presented and the dependent and independent measures.

4.1. Network training and testing procedure

The procedure presented in this work was based on two stages. The first stage was a training stage, which was followed by a score estimation stage.

At the first stage, a U-Net was trained using parameters similar to the previous work [22]. We altered the network to handle RGB images. The original images were recorded at a pixel resolution of 2688 by 1520. Later, to reduce the training time, the image's resolution was reduced to 500 by 500. The labels in this task were imbalanced, as a person can only look at few locations and cannot look at the rest of the scene. We compensated for this imbalance by using dilation to set more pixel as being viewed. This dilation was inspired by the fovea and parafovea in the human visual system. Figure 3 shows an example of an image and the dilated desired output. In addition, we compensated for the imbalance by setting class weights of $[0.01,\ 1]$ in the cross-entropy loss while training using Stochastic Gradient Descent (SGD).

A snapshot of the trained network was created every 1000 batches. Each batch was composed of 4 images and the relevant eye gaze locations on the images. We focused on the first 200 networks (i.e., 200,000 batches). Given that each batch held 4 examples, our focus was the first 800,000 training samples. We repeated this stage twice: once with a smaller dilation (150 pixels) of the real gaze distribution and once with a larger one (300 pixels). The size of the dilation was relative to the original size of the image. The network was trained using Stochastic Gradient Descent (SGD) with a cross-entropy loss function. Overall, 400 snapshots of the networks were generated: $\{1..200\} \times \{smaller, larger\}$.

The second stage was a score estimation stage. As part of this stage, the saliency score and the gaze loss of the snapshots' networks were estimated. The scores were also estimated on the validation (held-out) set. Each of the 400 snapshots was tested, and average saliency scores and eye gaze loss were estimated.

Figure 3. An example of an image and its dilated label

4.2. Dependent and independent measures

Snapshot id – independent measure –During the training, a sequence of snapshots of the trained network were recorded. The location of the snapshot in the sequence was the independent measure. It varied between 1 and 200 for each of the two dilation conditions $\{1..200\} \times \{smaller, larger\}$.

Cross entropy loss – dependent measure – This is the regular loss that was estimated on the held-out set. It measures the similarity between the network's predicted distribution over the image and the recorded gaze pattern of the driver. Given that for each pixel there are two possibilities (driver was looking at it, driver was not looking at it), we can consider it as a binary classification problem. The loss is such case is expected to be in the interval: $[0, -\ln(0.5) = 0.69]$.

Saliency comparison score – dependent measure – The saliency score is explained at section 3.2. It measures the similarity between the network's predicted distribution over the image and the saliency distribution. Positive values represent similarity of the predicted distribution to the saliency distribution.

5. RESULTS

The results were estimated based on the held-out set. The results for the smaller dilation snapshots are presented in Figure 4. The x and y axes are the gaze loss and saliency comparison score. The snapshot id, that represent the timeline, is not presented. However, it was known that loss and time tended to move together when each had its own pace. We used running average with a factor of 0.9 to smooth the data.

Three interesting areas (i.e., A,B,C) could be highlighted in Figure 4 while using small dilation (150 pixels). **Area A** – this area is related to the beginning of the training. At this area, the loss was not only maximal, it was close to the loss of random labeling (0.69). The saliency comparison scores were almost random as well and was close to zero (As can suggested by the two terms in Eq 2).

Area B – in this area, It was observed that the loss declined, suggesting that the network started to learn the goal-oriented gaze task of driving. In addition, there was an increase in the saliency score. The local minima in this area was to some extent associated with a bottom-up saliency behavior. This saliency behavior approximated the driver behavior.

Area C – in this area, a change of network behavior occurred. Intuitively, it can be viewed as a phase shift. The local minima started to deviate from the saliency like behavior, towards other local minima that are less saliency based. Here, both the loss and saliency comparison score were reduced.

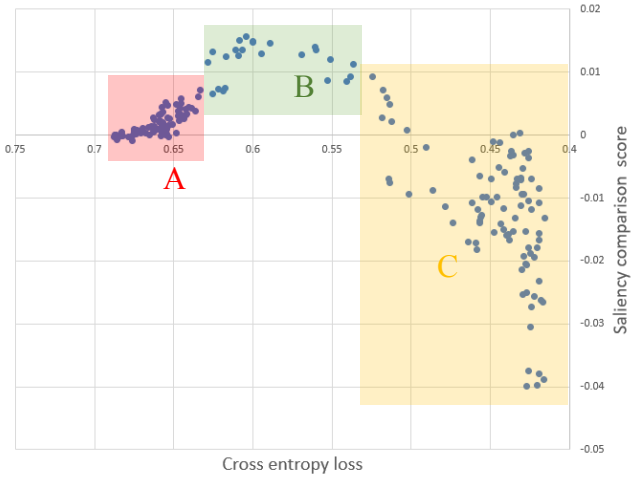

Figure 4. Saliency comparison score vs. cross-entropy loss using smaller dilation (150 pixels).

Figure 5 presents training process using larger dilation. The data was processed in the same way as the smaller dilation. There were partial similarities between this network and the one presented in the Figure 4. Area A (in Figure 4) and D (in Figure 5) shared some similarities: both areas were associated with a network at the beginning of its training. Areas B (in Figure 4) and E (in Figure 5) had relatively high similarity and high saliency comparison scores. In Figure 5, there was no area that is similar to area C (in Figure 4). This could be due to the limit of 200,000 training batches.

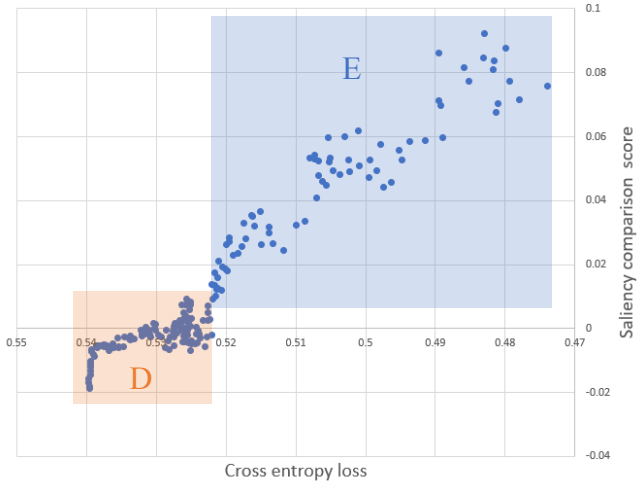

Figure 5. Saliency comparison score vs. cross-entropy loss using larger dilation (300 pixels).

Overall, in both networks, the same non-trivial behavior was observed when training for a gaze prediction of driving task. Although driving task is mainly a top-down behavior, a saliency bottom-up behavioral patterns emerged at local minima. Those minima to some extent approximated the top-down behavior by having a bottom-up pattern incorporate into them.

6. DISCUSSION

Generally, it was demonstrated that at first networks learned bottom-up models and, only later, the top-down models emerged. This finding resonates with *[5]*, which demonstrated that at first networks learned simple solutions and, only later, complex solutions were learnt. Indeed, bottom-up saliency models are considered simpler than top-down ones. It might be argued that saliency models require knowledge of a smaller area around the evaluated location.

Another interesting comparison is between our findings and children's visual attention capacity: [13] [14] [15] [16]. Children are more likely shift to a salient and less important area in a scene and lose focus of their goal relative to adults.

7. REFERENCES

[1] H. Li, Z. Xu, G. Taylor, C. Studer and T. Goldstein, "Visualizing the loss landscape of neural nets," 28 Dec 2017. [Online]. Available: https://arxiv.org/abs/1712.09913.

[2] B. Carlo, F. Pittorino and R. Zecchina, "Shaping the learning landscape in neural networks around wide flat minima," *Proceedings of the National Academy of Sciences,* vol. 117.1, pp. 161-170, 2020.

[3] R. Huang, Z. Emam, M. Goldblum, L. Fowl, J. Terry, F. Huang and T. Goldstein, "Understanding generalization through visualizations," in *arXiv preprint arXiv*, 2019.

[4] M. Van der and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* 2008.

[5] D. Kalimeris, G. Kaplun, P. Nakkiran, B. Edelman, T. Yang, B. Barak and H. Zhang, "Sgd on neural networks learns functions of increasing complexity," in *Advances in Neural Information Processing Systems*, 2019.

[6] Y. Vanunu, . J. Hotaling, M. Le Pelley and B. Newell, "How top-down and bottom-up attention modulate risky choice," *Proceedings of the National Academy of Sciences,* vol. 118, no. 39, 2021.

[7] A. Palazzi, D. Abati, F. Solera and R. Cucchiara, "Predicting the Driver's Focus of Attention: the DR (eye) VE Project," *IEEE transactions on pattern analysis and machine intelligence,* pp. 1720-1733, 2018.

[8] . T. Hwu, M. Levy, S. Skorheim and D. Huber, "Matching Representations of Explainable Artificial Intelligence and Eye Gaze for Human-Machine Interaction," arXiv preprint arXiv:2102.00179, 2021.

[9] R. M. Hecht, A. Bar Hillel, A. Telpaz, O. Tsimhoni and N. Tishby, "Information constrained control analysis of eye gazing distribution under cognitive workload," *IEEE Transactions on Human Machine Systems,* 2019.

[10] R. M. Hecht, A. Telpaz, G. Kamhi, A. Bar-Hillel and N. Tishby, "Information constrained control for visual detection of important areas," in *International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[11] T. W. Victor, J. L. Harbluk and J. A. Engstrom, "Sensitivity of eye-movement measures to in-vehicle task difficulty," *Transportation Research Part F: Traffic Psychology and Behaviour,* pp. 167-190, 2005.

[12] A. Borji, "Saliency prediction in the deep learning era: Successes, limitations, and future challenges," in *arXiv*, 2018.

[13] F. Waszak, L. Shu-Chen and B. Hommel, "The development of attentional networks: Cross-sectional findings from a life span sample.," *Developmental psychology,* vol. 46.2, p. 337, 2010.

[14] T. Sweeny, N. Wurnitsch, A. Gopnik and D. Whitney, "Ensemble perception of size in 4 5 year old children," *Developmental science,* pp. 556-568, 2015.

[15] J. Seong Taek, J. Hamid, D. Maurer and T. Lewis, "Developmental changes during childhood in single-letter acuity and its crowding by surrounding contours," *Journal of experimental child psychology,* pp. 423-437, 2010.

[16] J. Ristic and A. Kingstone, "Rethinking attentional development: reflexive and volitional orienting in children and adults," *Developmental science,* pp. 289-296, 2009.

[17] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2005.

[18] J. Harel, C. Koch and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006.

[19] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence,* pp. 1254-1259, 1998.

[20] G. Lee , Y.-W. Tai and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *IEEE conference on computer vision and pattern recognition*, 2016.

[21] . D. Jobson, Z.-u. Rahman and G. Woodell, "Properties and performance of a center/surround retinex," *IEEE transactions on image processing,* pp. 451-462, 1997.

[22] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015.