Prediction of Human Fixations Movement Using Eye Gaze Data and LSTM model

Enming Zhang

Research School of Computer Science, Australian National University
U6542688@anu.edu.au

Abstract. In order to understand human's ability to distinguish manipulated pictures and locating the manipulated area, many different pieces of research have been conducted based on eye gaze data, and different statistical analysis methods. While tracking human's fixations on different pictures, the fixation's coordinates will be recorded in line with the time series. Instead of making predictions on the manipulations, this paper builds a long short-term memory (LSTM) network model and trains the model using a series of human fixations' coordinates to see whether the trajectory of fixation can be predicted. A pruning method is applied to the model as well to reduce the number of neurons in the hidden layer. The result indicates that a well-trained neural network model can generate a prediction with low accuracy. For a neural network model with a relatively small size, the pruning method based on distinctiveness produces a very little effect.

Keywords: Long short-term memory, human fixation, network reduction, distinctiveness pruning

1 Introduction

Eye gaze tracking data is wildly used in many different types of researches including human perception, embodied cognition, and human behaviour analysis[1, 2]. To find the correlation between human's eye gaze with their perception of whether a photo has been manipulated or not, Caldwell et al. designed an experiment which combined with eye gaze tracking and verbal questioning, to make comparisons between the image information the participants received and the judgments they made[3]. The result of this experiment shows that the participants' overall ability to distinguish a manipulated image is poor, however, the increased attention, which could be reflected by eye gaze tracking data in different aspects, tends to indicate a higher accuracy when picking up those manipulated photos. This interesting finding may indicate that while observing a picture, the movement of human fixations can follow some patterns. Judd, Durand and Torralba's research made comparisons on different models which predict human fixations, mentioned that the visual system is affected by bottom-up and top-down mechanisms [4]. These internal mechanisms make up the ground theory of the prediction. Since the network model has a large scale, the training process is very time-consuming. After building and training the standard

neural network, The network pruning technique raised by Gedeon and Harris[5] when implementing image compression using a single layer network has been applied, to see whether this network reduction technique will reduce the cost of training the original neural network model.

2 Preparation and Methods

2.1 Dataset Description and Pre-processing

The first dataset I use in this paper is a subset of the original dataset retrieved from the experiment conducted by Caldwell et al. [3]. It contains the X and Y coordinates of 80 participants' fixations, plus the corresponding time durations. All these fixations data were retrieved during the period while participants were observing 5 pictures which have been manipulated. Since different image may have different attractions, the first step of data pre-processing is to group the fixation data using the image id, then draw the distribution of fixations on each image.

Fig. 1. The image viewed by participants, from left to right: 10, 11,12,13,14 [3].

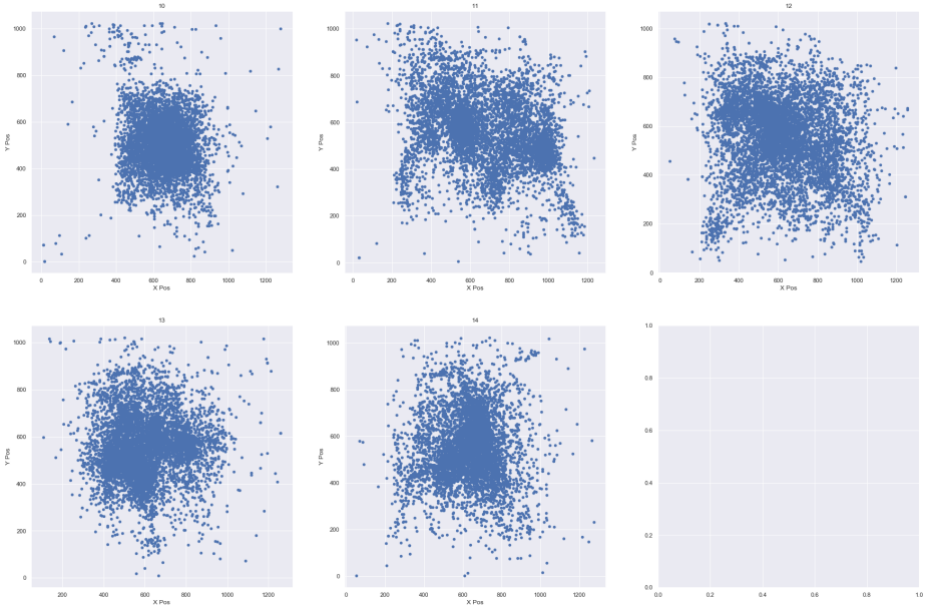

Fig. 2. The fixations' distribution on 5 manipulated images.

According to Fig 2, while participants were looking at images, one of the most outstanding commonalities is that most of the fixations will land upon the central area of the image. When it comes to image 11, there are more fixations focusing on the right part of this image, which is also the place where the researchers added the Queen's figure. This phenomenon might indicate that this manipulation is more conspicuous, Caldwell et al. [3] also mentioned this in the paper.

The second dataset used in this paper is the coordinates of the manipulated areas in these images, to observe whether there are any associations between the fixations and the manipulations, all of the scatter plots in Fig 2 are overlapped with the manipulations in one single plot. The result shows that there's no evident correlation between the location of fixations and manipulations except image 11.

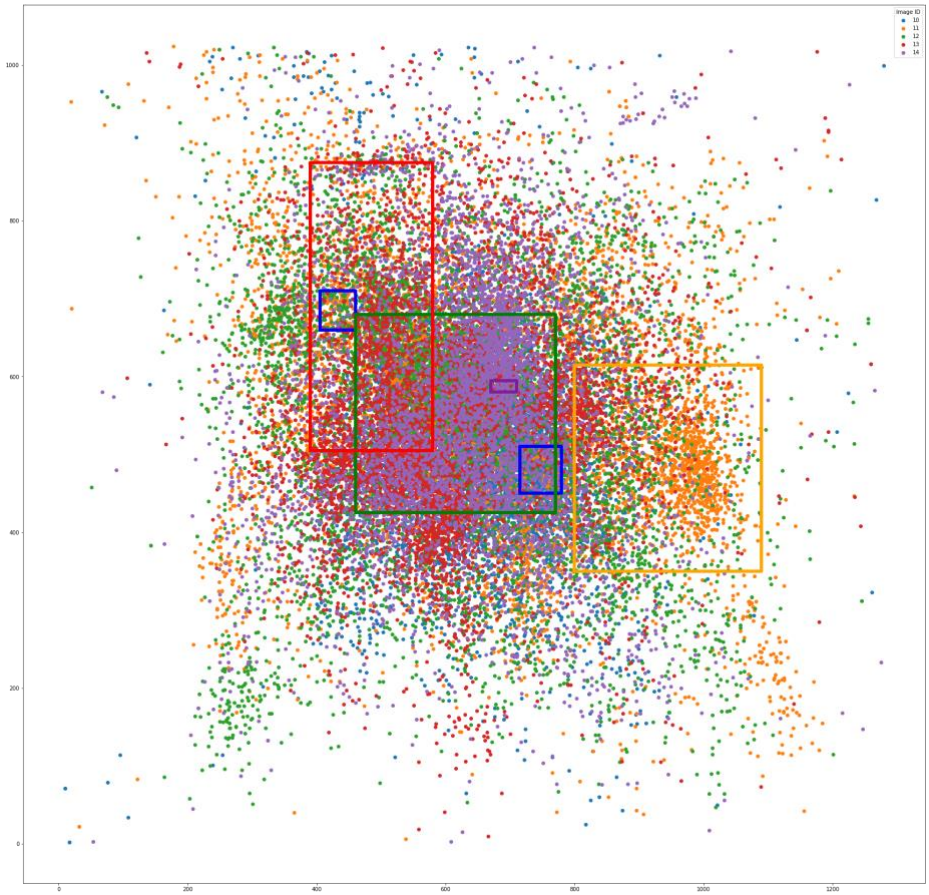

Fig. 3. The positional relation between fixations and manipulations. Each colour represents an image, the rectangles cover the areas that have been manipulated.

While training the LSTM network using the first dataset, only the X and Y positions are kept since these are the only two features that the network model needs to take as inputs. The order of fixations is rearranged by the time sequence provided in the first dataset. Then all of the fixations are grouped by the image ID. The LSTM network model will take 5 groups of fixations and make predictions for 5 images separately.

2.2 LSTM Neural Network

LSTM neural network is a special Recurrent Neural Network (RNN) which contains memory blocks in its hidden layer. The memory cells in these memory blocks will store relevant information about the next event within a small time window. LSTM neural network model has been widely used in dealing with time series and related data. In this paper, the LSTM model takes only one input: the x or y coordinate, and

will generate one output. This model will run 5 iterations and make predictions on each image to control other variables which can be brought by the image itself. After making predictions on both two coordinates, the datasets will be combined together and generate one single dataset for the location distribution.

This implementation has two main disadvantages. First, since the hyperparameters are set up manually, it's very likely that some redundant hidden neurons exist in the neural network model. As Gedeon and Harris mentioned, the toughest part while building the model during practice is to decide the number of hidden units[5]. A sufficient number of hidden neurons could make the model get well trained. Whereas the duplication will lead to a higher cost of time, storage and memory. In this case, to optimize the neural network model, finding the redundant neurons and cut them off becomes an essential task. Second, instead of taking a pair of coordinates together as a double input, this model only takes one coordinate as the input and it only generates one output. This approach is to reduce the time and memory cost but it can also bring negative impact on the accuracy. Increasing the dimension of input could be an important task that will be implemented in the future.

Activation function. In line with Gedeon and Harris[6], I choose to use Sigmoid function (formula 1) as the activation function for the neural network model. According to Jain et al., Sigmoid function is effective in providing a smooth non-linear decision boundary[7].

$$Sigmoid(x) = \frac{1}{1 - e^{-x}}. \tag{1}$$

2.3 Distinctiveness Analysis

The method I use to find a similar neuron in this experiment is based on the distinctiveness analysis theory raised by Gedeon and Harris[6]. This algorithm created a vector for each neuron which has the same dimensions as the input patterns. To determine the similarity of two neurons, it provided a way of calculating the angle, see formulae listed below.

$$angle(i,j) = tan^{-1}\left(\sqrt{\frac{\sum_p^{pats} sact(p,i)^2 * \sum_p^{pats} sact(p,j)^2}{\sum_p^{pats}\left(sact(p,i) * sact(p,j)\right)^2}} - 1\right) \tag{2}$$

Where

$$sact(p,h) = activation\ (p,h) - 0.5 \quad \textbf{(3)}$$

Since the Sigmoid function will always get an output in range 0 to 1, the second formula normalizes the result to the range -0.5 to 0.5, which enlarges the angular range to 0° to 180°. This method prescribed that if the angular separation is smaller than 15°, these two neurons will be treated as duplicate and one of them has to be removed. Moreover, two vectors whose angular separation is larger than 165° will be both removed. One of the benefits of this method is that it doesn't require further training, which makes the testing process easier.

3 Results and Discussion

After the first training process, the average training loss is lower than 0.0001 and all the predicted fixations are concentrated in the middle of each image. The main reason for this phenomenon is the normalization operation. Since the value of coordinate is large, normalizing the coordinates into [-1,1] before sending them into the LSTM model makes a huge damage to the performance, since the inputs are very focused. After removing the normalization, the average training loss increased a lot. The overall accuracy of the prediction is lower than 40%, the comparisons between predictions and actual fixations can be seen in Fig 4. The green dots represent the actual fixations while the red dots represent the prediction results.

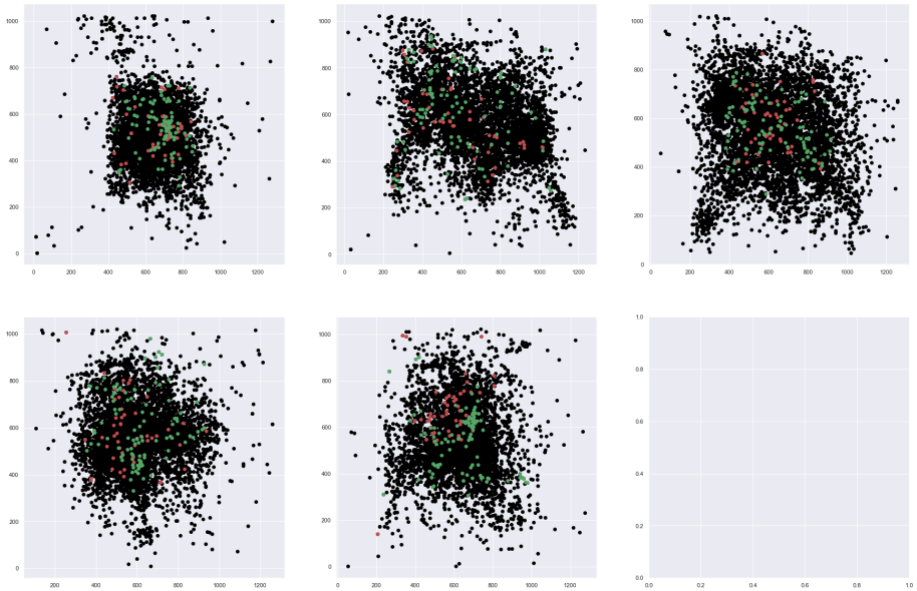

Fig. 4. The comparisons between predictions and actual fixations for each image.

According to the result, the model presents the best performance on predicting fixations for image 10, while it has the lowest accuracy on image 13 and 14. An interesting thing is that the prediction fixations in these two figures are more fare away from the actual result, however, these predictions are closer to the manipulated areas. One possible reason behind this is that the timing window is too small to let the model can't capture the pattern correctly. Another possible reason is that this model is overfitting. The manipulated area in these two images might be found by the participants, so the model tends to make the predicted location close to the area which has been noticed. The overall prediction result matches the expectation.

After finishing the design and training process, I start the pruning process by designing an angle calculator based on formulae 2 and 3. This is the Instead of the output of the hidden layer, I retrieve the output of each hidden neuron, and normalized the value to range -0.5 to 0.5. Unlike the previous standard artificial neural network model, the pruning process becomes much more difficult while working on this model, since the hidden layer size increases to 100. The pruning process fails to bring a contribution to the LSTM model's process.

4 Conclusion and Further Work

This paper implements an LSTM neural network model and applies a pruning method based on the theory of distinctiveness analysis. According to the result and the performance analysis results, the original prediction model's accuracy could be evaluated as low, besides those possible reasons mentioned above, another reason which could explain this result is that to exactly match the time order is too hard. For example, if the actual location for the fixation 3921 is (650,200) , even if the prediction drop off at (650,200) one step later, the overall accuracy will be damaged. Overall, the LSTM model didn't show a well performance in making prediction of fixation moving trajectory.

The pruning process, on the other hand, didn't bring benefit to the model's prediction. The most important approach I would like to implement in the future is to increase the dimension of the input and make it be able to accept both two coordinates as input. My assumption is that this could improve the model's overall performance. Also, since this pruning method's target is the neuron, I plan to apply weight pruning methods on the same model and make performance analysis based on the result comparison.

5 References

1. Cegala, D.J., Sokuvitz, S., Alexander, A.F.: An investigation of eye gaze and its relation to selected verbal behavior. Human Communication Research 5, 99-108 (1979)
2. Macrae, C.N., Hood, B.M., Milne, A.B., Rowe, A.C., Mason, M.F.: Are you looking at me? Eye gaze and person perception. Psychological science 13, 460-464 (2002)
3. Caldwell, S., Gedeon, T., Jones, R., Copeland, L.: Imperfect understandings: a grounded theory and eye gaze investigation of human perceptions of manipulated and unmanipulated digital images. In: Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science. (Year)
4. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. (2012)
5. Gedeon, T., Harris, D.: Progressive image compression. In: [Proceedings 1992] IJCNN International Joint Conference on Neural Networks, pp. 403-407. IEEE, (Year)
6. Gedeon, T., Harris, D.: Network reduction techniques. In: Proceedings International Conference on Neural Networks Methodologies and Applications, pp. 119-126. (Year)
7. Jain, A.K., Mao, J., Mohiuddin, K.M.: Artificial neural networks: A tutorial. Computer 29, 31-44 (1996)