Human eye-fixation prediction based on Convolutional Neural Network in RGB images

Miao Yue, Li Wei and Li Chunle
1 Key Laboratory of ICSP, Ministry of Education, Anhui University, Hefei 230601, China
2 School of Computer Science and Technology, Anhui University, Hefei 230601, China
*Corresponding author’s e-mail: le998@sohu.com

Abstract. Using neural networks to simulate and predict human visual attention mechanism is a hot topic in the field of computer vision. In this paper, we propose an end-to-end encoder-decoder network architecture to predict the fixation mechanism of human eyes, which consists of modules composed of multiple convolutional layers with different expansion rates to capture multi-scale features in parallel. In addition, the attention module is added on the basis of the encoder network structure, and a self-attention mechanism is introduced to capture the visual feature dependency in the channel size, and the semantic interdependence in the channel dimension is modeled to predict the visual salience more accurately. In this paper, five data sets and selected examples are used to demonstrate the effectiveness of the proposed method. Our method achieves competitive and consistent results on multiple evaluation indicators on MIT1003 and CAT2000 datasets.

1. Introduction
Humans have quick search from complex visual scene and the ability to locate objects of interest[1,2], this is called visual attention this visual attention mechanism for people to process visual information in daily life is of great significance in recent years, significant detection has become a hot research topic. In computer vision, saliency detection is a basic preprocessing step for various tasks, including image redirection[3,4], target recognition[5], video compression[6], etc.[7-9] Studies of significance detection can be divided into two broad categories: significant target detection and eye-fixation. The human eye-fixation prediction purpose is to simulate the human to look at pictures of fixation point early inspired by biology, considering the significance of the visual scene may come from some of the stimulus (color size, etc.), research on human eye-fixation prediction relies on these low-level visual features, however this does not capture the high-level semantic features of the visual scene, makes the good characteristics of handmade predict significant difficult this makes deep learning to solve the problem of natural selection.

Deep learning technology, especially the convolutional neural network, due to its embedding characteristics description mechanism, for the pretreatment of the computer vision provides a practical tool for deep learning and the traditional method of the main difference is that: 1) Deep learning does not require analytical constraints, nor does it rely on prior knowledge. With the introduction of deep neural networks, deep networks have shown impressive results on a range of different perceptual tasks, such as speech recognition, natural language processing and object recognition. The ability of deep
neural networks to automatically learn complex patterns from data in a hierarchical manner makes them suitable for a wide range of problems with different data patterns where human-eye gaze prediction has also been developed. Based on the Salicon dataset[10], a large significance dataset has been published, making it possible to pre-train convolutional neural networks on gaze prediction. In 2015, Kruthiventi et al. proposed the DeepFix model[11], and they were the first to use diffuse convolutional layers to increase image resolution. Cornia et al. proposed the ML-NET model[12], which integrates the features extracted from different layers in the convolutional neural network. The model consists of three modules: a feature extraction convolutional neural network, a feature coding network and a priori learning network.

In this paper, a method of eye fixation prediction based on convolutional neural network is proposed. Modules extracted from semantic segmentation literature are used to predict fixed density maps with the same resolution as the input images. Complex representation at multiple spatial scales is necessary for accurate prediction of human gaze patterns, therefore, this article module, joined a context to the multi-scale information sampling, and using the global scene features to enhance the context module to evaluate the contribution to the overall performance, with the final result before work on two significant public benchmark are compared. In the sections that follow, we will describe our contribution to this challenging task.

2. Network Architecture
This paper proposes a CNN network structure with an ASPP module and an attention module. The ASPP(Atrous Spatial Pyramid Pooling) module, adapted from the semantic segmentation literature, is used to predict fixation density maps with the same resolution as the input images. The main contributions of this paper are described in the following sections.

![Figure 1. Architecture of our network](image)

2.1. Overview of Network Architecture
In this paper, the popular VGG16 architecture is used as the image encoder. Through the pre-trained convolution layer, more and more complex features are extracted along the hierarchical structure. The VGG16 trunk was modified to meet the requirements of intensive prediction tasks by removing feature undersampling in the last two maximum pooling layers, which still allowed us to initialize the model with pre-trained weights because the number of trainable parameters remained unchanged. Previous work has demonstrated the effectiveness of this approach in significance prediction problems. After
that, the feature layer output by the last three layers of the convolutional neural network is processed by the attention module, and then the acquired attention features are combined with the convolutional coding layer and processed by the ASPP module. The ASPP module uses several convolutional layers with different expansion factors to obtain multi-scale image information in parallel, thereby increasing the receptive field to compensate for higher resolution. Fig.1 shows the overall architectural design described in this section.

2.2. Channel Attention Module

We propose to add a channel attention module[13] to the encoder network structure to model the semantic interdependence in the channel dimension. The channel mapping for each high-level feature can be thought of as a class-specific response, with different semantic responses associated with each other. The channel attention module attempts to obtain the final feature of each channel, that is, the feature-weighted data of all channels and original features. It models remote semantic dependencies between feature mappings. It emphasizes class-dependent feature mappings and helps to improve feature differentiation. At the same time, the channel attention module selectively emphasizes the interdependent channel mapping by integrating all the related features of the channel mapping.

\[
x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^{C} \exp(A_i \cdot A_j)}
\]

(1)

Where \(x_{ji}\) represents the mapping of the \(i\)-th channel to the \(j\)-th channel, and matrix multiplication of \(X\) and \(A\) transpose. Sum the result with \(A\) to get the final output \(E\) (CxHxW).

\[
E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j
\]

(2)

By mining the interdependence between channel mappings, we can emphasize the interdependent feature mappings and improve the feature representation of specific semantics. Therefore, we build a channel attention module to explicitly model the dependencies between channels.

It does not add too many parameters, but it effectively enhances the feature representation. We will convert the output of the attention module into features of the same size through the convolution layer, and perform feature sum to complete feature fusion. Finally, the prediction graph will be generated through the ASPP module. We don't use cascading because it requires more GPU memory. Our attention module is simple and can be plugged directly into an existing convolutional neural network flow. It does not add too many parameters, but it effectively enhances the feature representation.
2.3. ASPP Module
The ATRous Spatial Pyramid Pooling (ASPP) module samples the given input in parallel with the void convolution at different sampling rates[14], which is equivalent to capturing the context of the image at multiple ratios. This allows us to obtain the result of expanding the receiver field without taking the next sampling.

In this work, we deployed three convolution layers in parallel, with kernel sizes of 3x3, dilatation rates of 4, 8, and 12, as well as a 1x1 convolution layer that could not learn new spatial dependencies but nonlinearly combined existing feature mappings. The image-level context is represented as the output after global average pooling (that is, the tensor terms across two spatial dimensions are averaged to one value), and then bilinear upsampling is used to obtain the same resolution as all other representations, followed by another point-by-point convolution operation. Each of the five branches of the module contains 256 filters, resulting in a characteristic map of 1280 tensors. Finally, the combined output is forwarded to a 1x1 convolution layer with 256 channels, resulting in the output of an aggregate tensor containing 1280 feature mappings.

2.4. Deconvolution
Decoder means that in order to restore the resolution of the original image, the extracted features are processed through a series of convolution layers and upper sampling layers. This article uses three upsampling blocks, including a bilinear scaling operation that doubles the number of rows and columns and is followed by a convolutional layer with a kernel size of 3X3. Compared with deconvolution, this setup has been shown to prevent checkerboard artifacts in the upsampled image space. In addition to improving the resolution of the entire decoder, the number of channels per block is reduced by half, resulting in 32 feature mappings. The final network layer converts the activation into a continuous significance distribution by applying triple convolution.

3. Experiment and Analysis

3.1. Experimental details and data sets
In this paper, the proposed model is implemented and trained on Tensorflow1.14.0, a deep learning library based on Python, and the performance of the model is evaluated and tested using Matlab R2018a. The proposed encoder decoder model was evaluated on publicly available eye movement data sets, qualitative and quantitative results were obtained, and the common indicators for the prediction performance of the significance model were summarized. Finally, we report on the contribution of our architectural design choices and benchmark the overall results against the baseline of computer vision and related work.

In this paper, three most popular fixation point data sets are used: SALICON, MIT1003 and CAT2000.

Salicon dataset[10]: This dataset consists of 10,000 training instances and 5,000 validation instances, all images taken from the Microsoft Coco dataset. The fixation points in this data set are not recorded by an eye tracker, but simulated by a mouse click. The authors demonstrate a high degree of similarity between the results of mouse markers simulating human eye fixations and those recorded by eye trackers. Due to its large size, this paper first trains the model on Salicon and then fine-tunes the fixed predictions on other datasets with the same optimization parameters. This widely used method has been shown to improve the accuracy of eye movement estimation.

MIT1003 data set[15]: This data set consists of 1003 randomly selected images from Flickr and Labelme. The truth values of fixation points corresponding to the images were recorded by 15 volunteers through an eye tracker.

CAT2000 data set[16]: The data set consists of 2000 training instances and 2000 test instances, with a total of 4000 images of 20 categories. Categories include: Cartoon, Art, Satellite Images, Low Resolution Images, Indoor, Outdoor, Stick Drawing, etc. There are 200 images for each category.
3.2. Evaluation Criteria
In this study, linear linear correlation coefficient (CC), AUC, normalized scanning path significance (NSS) and similarity (SIM) were used as indicators for eye fixation prediction.

CC: CC is a statistical method that usually measures the linear correlation between two random variables $P$ and $G$.

$$CC = \frac{\sigma(P, G)}{\sigma(P) \times \sigma(G)}$$

(3)

NSS: NSS is a metric designed specifically to measure the average normalized significance between two fixed graphs. We know that $P$ and $G$ and $N$ are the number of gazes that the human eye makes. NSS can be calculated as follows:

$$NSS = \frac{1}{N} \sum_{i=1}^{N} G_N(i)$$

(4)

Sim: Measure the similarity between two distributions, represented by a histogram. The SIM is calculated as the sum of the minimum values at each pixel after normalization of the input mapping.

AUC: Is defined as the area under the rate of change indicator (ROC) curve and is widely used in significance model evaluation maps. Given an image and the truth graph of its fixation points, the fixation points and other fixation points are regarded as positive and negative sets respectively. Then, the threshold value was used to divide the significance map into significant area and non-significant area, and the ROC curve was drawn. The area under the curve was calculated as the AUC score.

3.3. Experimental results and Comparison
In the experimental model comparison, we selected six models ITTI [17], GBVS [18], Judd Model [19], BMS [20], CAS [21], eDN [22], for comparison. We randomly selected 7 pictures and 4 models for comparison experiment. The comparison of experimental prediction maps is shown in Fig.3.

![Figure 3. Qualitative results and comparison to the state of the art. validation images from MIT1003 dataset](image)

In this paper, the proposed encoder decoder model is evaluated on the publicly available eye movement data set MIT1003, and qualitative and quantitative results are obtained. Table 1 shows the evaluation results of this model and BMS, eDN, CAS, Judd Model, GBVS and ITTI models in
MIT1003. The experimental results show that this method achieves the best evaluation results on most parameters of AUC, SAC, SIM, CC and NSS, and NSS and CC have the greatest improvement.

We quantitatively compared our method with existing models on MIT1003 and CAT2000 test sets. The results of MIT300 and CAT2000 datasets are shown in Table 1 and Table 2, respectively. In the MIT1003 data set, our method achieves the best results in all indicators compared with other models. However, in the CAT2000 dataset, our method achieves the best results in other indicators except AUC, but does not achieve significant performance improvement in AUC. This can be explained by the fact that the AUC indicator is mainly based on true positive results, with no obvious penalty for false positive results. Thus, fuzzy significance maps can achieve higher AUC values, although visually very different from truth maps.

As we noted, our network was able to predict high salient values for human-faced objects and other major cues. It can also generate good saliency maps when the image does not contain areas of strong saliency, such as when the saliency is concentrated in the center of the scene or when the image depicts the landscape. We note that from a qualitative point of view, the model can sometimes infer the relative importance of different people in the same scenario, and as discussed in, significance models still have difficulty replicating human behavior.

### 4. Conclusion

We propose a new salience attention model, based on encoder-decoder architecture, which can predict the fixation of human eyes on natural images. Through extensive evaluation, we verified that our model achieved the most advanced results on the two most important saliency prediction datasets. The ASPP module in the model integrates multi-scale information and global context based on semantic feature representation, and the channel attention module models the semantic interdependence in the channel dimension, which significantly improves the results of the model on the eye-tracking data set qualitatively and quantitatively. Finally, we contribute to further research work by releasing the source code and pre-trained models of our architecture. This model is computationally lightweight compared to the previous significance approach.
References

[1] Irwin, D. E. (1991). Information integration across saccadic eye movements. Cognitive Psychology, 23(3), 420–456.

[2] Jonides, J., Irwin, D. E., & Yantis, S. (1982). Integrating visual information from successive fixations. Science, 215(4529), 192–194.

[3] CHEN L Q, XIE X, FAN X, et al. A visual attention model for adapting images on small displays[J]. Multimedia Systems, 2003, 9(4):353-364.

[4] SETLUR V, TAKAGI S, RASKAR R, et al. Automatic image retargeting[C]. International Conference on Mobile and Ubiquitous Multimedia, Mum 2005, Christchurch, New Zealand, December. DBLP, 2004:4.

[5] WALTHER D, ITTI L, RIESENHUBER M, et al. Attentional selection for object recognition — a gentle way[J]. Allergologia Et Immunopathologia, 2002, 42(1):85–87.

[6] HADIZADEH H, BAJIC I V. Saliency-Aware Video Compression[J]. IEEE Transactions on Image Processing, 2014, 23(1):19.

[7] MAHADEVAN V, VASCONCELOS N. Biologically Inspired Object Tracking Using Center-surround Saliency Mechanisms[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(3):541-554.

[8] HAN J, NGAN K N, LI M, et al. Unsupervised extraction of visual attention objects in color images[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2006, 16(1):141-145.

[9] HU X, WU H, ZHANG Y, et al. Flower image retrieval based on saliency map[C]. International Symposium on Computer, Consumer and Control. IEEE, 2014:304-307.

[10] JIANG M, HUANG S, DUAN J, et al. SALICON: saliency in context[C]. Journal of Visual Communication and Image Representation. IEEE, 2015:1072-1080.

[11] KRUTHIVENTI S S S, AYUSH K, BABU R V. DeepFix: a fully convolutional neural network for predicting human eye fixations[J]. arXiv preprint arXiv:1510.02927, 2015.

[12] CORNIA M, BARALDI L, SERRA G, et al. A deep multi-Level network for saliency prediction[J]. arXiv preprint arXiv:1609.01064, 2016.

[13] Fu J, Liu J, Tian H, et al. Dual Attention Network for Scene Segmentation[J]. 2018.

[14] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the international conference on artificial intelligence and statistics (pp. 249–256).

[15] Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In Proceedings of the international conference on computer vision (pp. 2106–2113).

[16] Borji, A., & Itti, L. (2015). CAT2000: A large scale fixation dataset for boosting saliency research. arXiv preprint arXiv:1505.03581.

[17] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, “Superpixel-based spatiotemporal saliency detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[18] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, “Discriminative multi-instance multitask learning for 3D action recognition,” IEEE Trans. Multimed, vol. 19, no. 3, pp. 519–529, Mar. 2017.

[19] C. Gong et al., “Saliency propagation from simple to difficult,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 2531–2539.

[20] J. Zhang and S. Sclaroff, “Saliency detection: A Boolean map approach,” in Proc. IEEE Int. Conf. Comput. Vis., Jun. 2013, pp. 153–160.

[21] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 10, pp. 1915–1926, Oct. 2012.
[22] E. Vig, M. Dorr, and D. Cox, “Large-scale optimization of hierarchical features for saliency prediction in natural images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 2798–2805.