Predicting Eye Fixations using Convolutional Neural Networks

Nian Liu[1], Junwei Han[1*], Dingwen Zhang[1], Shifeng Wen[1] and Tianming Liu[2]
[1]School of Automation, Northwestern Polytechnical University; [2]Computer Science Department, University of Georgia

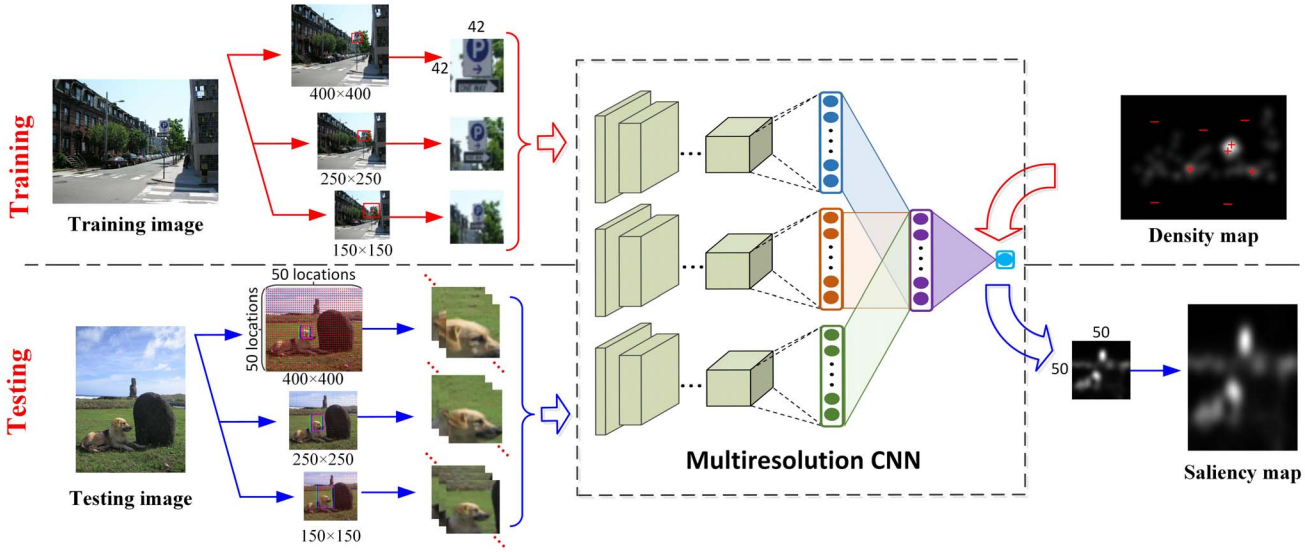

Figure 1: Diagram of our Mr-CNN based model. First, the given image is rescaled to three scales, i.e. 150×150, 250×250 and 400×400, then 42×42 sized image regions with the same center locations are extracted from the rescaled image duplicates as inputs to the Mr-CNN. We extract fixation and non-fixation image regions to train the Mr-CNN. When testing, we just evenly sample 50×50 locations per image to estimate their saliency values to reduce computation cost. The obtained down-sampled saliency map is rescaled to the original size to achieve the final saliency map.

When viewing visual scenes, human visual system has the ability to selectively locate eye fixations on some informative contents. In computer science field, researchers normally resort to computer vision techniques to quantitatively predict human eye fixations. Inspired by the biological evidence that locations distinctive from their surroundings are more likely to attract human attention, most traditional approaches typically cope with saliency modelling problem by three steps in sequence: early feature extraction, feature contrast inference, and contrast integration. However, traditional works rely on hand-crafted features and human-designed contrast inference mechanisms. Meanwhile, most works consider little top-down factors, which actually guide human eye movements after early stages of free viewing [1, 2].

To address the problems mentioned above, we combine the learning power of CNNs [3] and the contrast inference ability of multi-resolution structure, proposing to use a multiresolution convolutional neural network (Mr-CNN) to simultaneously learn all useful information for saliency detection, i.e., early feature extraction, low-level contrast inference, high-level semantics, and the integration of the two key factors. With this novel saliency-oriented architecture, our method can largely outperform other state-of-the-art methods on 4 datasets. With this model, we introduce visual saliency problem 2 insights: 1) The superior performance of our method indicates that the human visual system is more likely to process low-level contrast and high-level semantics jointly rather than separately; 2) Compared with traditional models using hand-crafted features, our deep model can uncover some new features that attract attention. The analysis of those features may offer inspirations to the understanding of visual attention.

Inspired by [4, 5], we adopt a multiresolution convolutional neural network. In details, as shown in Figure 1, we train the Mr-CNN directly from image regions centered on fixation and non-fixation locations over multiple resolutions, using raw image pixels as inputs and eye fixation attributes as labels. Benefitting from its hierarchical architecture and the purely supervised training manner, our model can learn saliency-related features with hierarchically increasing complexity in convolutional layers, instead of resorting to various hand-crafted features. These features learned with hierarchical depth can represent original image regions efficiently and discriminatively. In higher layers, the proposed Mr-CNN can learn diverse high-level top-down visual features due to its deep architecture. Meanwhile, it can also learn bottom-up saliency via combining information over multiple resolutions. Considering local image regions with the same center location but with fine-to-coarse resolutions (see the three image regions of the traffic sign in Figure 1), finer image regions are actually the central parts of coarser ones. When the deep features of both the center (the finer image region) and the context (the coarser image region) are inputted to a neural network simultaneously, the difference between them may be learned under the supervision of labels, which makes the proposed Mr-CNN have the capability to learn the bottom-up saliency, contrary to using various human-designed mechanisms in traditional models. Finally, the last logistic regression layer learns to integrate bottom-up saliency with top-down cues to predict eye fixations.

We conducted evaluation experiments on four widely used benchmark datasets. Qualitative and quantitative comparisons with other 11 state-of-the-art saliency models demonstrated the superiority of the proposed approach. In addition, we also visualized the hierarchical features learned by the proposed Mr-CNN, uncovering novel inspirations to the understanding of human visual attention mechanisms.

[1] R. Carmi, and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision research*, 46(26):4333-4345, 2006.

[2] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vision*, 7(14):4, 2007.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278-2324, 1998.

[4] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1915-1929, 2013.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.