SALIENCY PREDICTION BASED ON NEW DEEP MULTI-LAYER CONVOLUTION NEURAL NETWORK

Dandan Zhu[1]*,Ye Luo*[1,*]*, Xuan Shao*[1]*, Laurent Itti*[2]*, Jianwei Lu*[1,3,*]

[1]School of Software Engineering, Tongji University
[2]Dept. of Computer Science and Neuroscience Program, University of Southern California
[3]Institute of Translational Medicine, Tongji University

ABSTRACT

Recent advances in saliency detection have utilized deep learning to obtain high-level features to detect salient regions. These advances have demonstrated superior results over previous works that utilize hand-crafted low-level features for saliency detection. In this paper, we propose a new multi-layer Convolutional Neural Network (CNN) model to learn high-level features for saliency detection. Compared to other methods, our method presents two merits. First, when performing features extraction, apart from the convolution and pooling step in our method, we add Restricted Boltzmann Machine (RBM) into the CNN framework to obtain more accurate features in intermediate step. Second, in order to deal with case of non-linear classification, we add the Deep Belief Network (DBN) classifier at the end of this model to classify the salient and non-salient regions. Quantitative and qualitative experiments on three benchmark datasets demonstrate that our method performs favorably against the state-of-the-art methods.

Index Terms— Convolutional Neural Network, Restricted Boltzmann Machine, Deep Belief Network, Saliency Detection

1. INTRODUCTION

Visual attention is indispensable in our everyday life and it can allocate limited cognitive resources to the most important scene and suppress unimportant information. To simulate human visual attention process, computational saliency estimation models are built. In recent years, the estimated visual saliency got broad applications in image retargeting, object detection, human behavior detection, etc.

For a long time, visual saliency estimation models [1] are generated based on extracting low-level features and all kinds of methods are named bottom-up methods. However, bottom-up saliency estimation methods based on low-level features can't fully explain the mechanism of visual attention. In order to solve the problems in bottom-up methods using only low-level features, the effects by high-level features are explored in [2–6] methods. However, saliency estimation methods relying on high-level features by object detectors makes the models more complex and infeasible in implementation.

In order to overcome the aforementioned limitations in high-level features extraction, deep convolution neural network (CNN) [7–10] models have achieved significant success in computer vision since CNN can automatically extract high-level features from the original images, and is more effective than traditional hand-crafted feature extraction methods [11, 12]. Although CNN based models show its superiority on saliency detection especially on extracting high-level salient features, methods of this kind also have drawbacks. At first, in the current CNN architecture, the input data is convolved and pooled as the input of the next layer. However, not all the convoluted data is useful to the next layer. In other words, information is redundant and features on the previous layer can be further processed such that more useful features obtained can initiate more efficient next layer's training. Moreover, in the traditional CNN based methods for recognition tasks or classification problems, after a fully connected layer, it is usually directly followed by a simple SVM classifier [13–15]. Although SVM is simple and efficient, it is still a linear classifier and cannot handle the nonlinear cases.

In order to solve these problems existing in CNN framework based saliency detection methods, in this paper, we propose a new multi-layer CNN model to extract/learn high-level features for saliency detection. The whole architecture of our model can be found in Figure 1. The input image is first preprocessed by sparse coding, then the sparse representation of the image is as the input and sent to the four-layer CNN network which has been trained for high-level feature extraction from natural images. At every layer, except for the traditional units: convolution and pooling, we add another component Restricted Boltzmann Machine (RBM) to further extract the features after the convolution and before the pooling. Moreover, at the end of our model, after two fully-connected layers, rather than SVM, Deep Belief Network (DBN), a multi-layer RBM network, is used. We replace SVM with DBN, mainly because DBN is unsupervised and can be applied to unlabeled data; besides it is a fast, greedy learning algorithm that can find a fairly good set of parameters, even in deep networks with millions of parameters and many hidden layers.

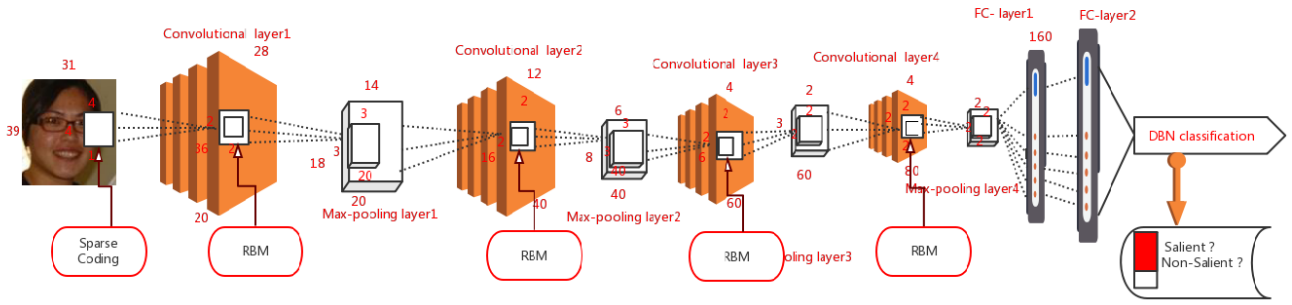


Fig. 1. Architecture overview of our proposed multi-layer CNN based saliency prediction model. A complete convolution layer consists of convolution and RBM operations. The feature map is generated in the pooling layer by max-pooling operation. A DBN network for training is fully connected to the output of the network to predict saliency.

2. OUR APPROACH

In this section, we present the proposed method, which includes three parts: a new multi-layer CNN network to learn high-level features, DBN classifier training and the visual saliency prediction by the learned high-level features and the trained DBN classifier for the testing image. We will introduce them one by one as following.

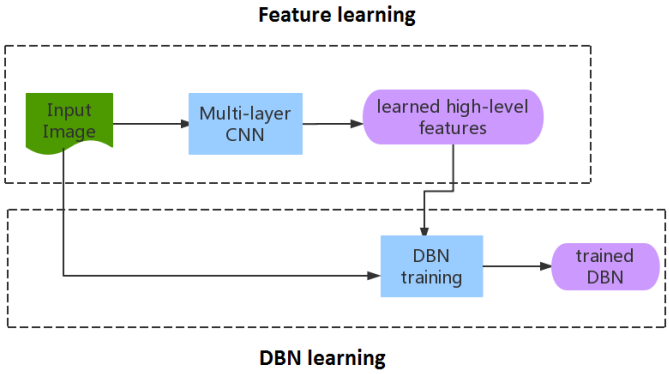


Fig. 2. Illustration of the pipeline of the proposed high-level feature learning and the DBN classifier training.

2.1. High-level Feature Learning

High-level feature learning for natural images is the key part of our method. Before introducing the specific structure of the new multi-layer CNN network, we first introduce the flow chart of the feature learning process.

2.1.1. Pipeline of the Feature Learning

As Figure 2 shows, high-level features are learned or extracted together with training of the CNN network. In order to learn good features, pre-training is performed in advance.

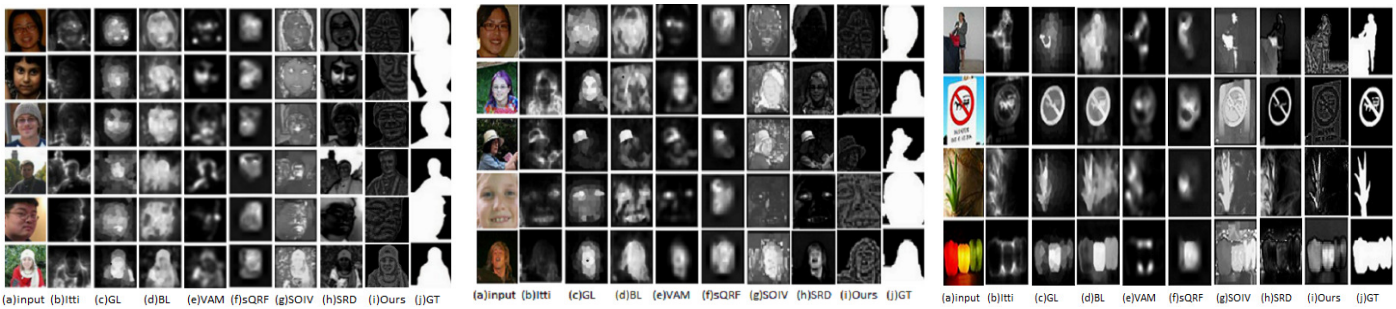
Pre-training In order to better initialize the training of our proposed feature learning framework, as [16] did, pre-training is first performed with the assumption that useful features are always from salient regions. We collect the salient regions from the eye fixation points for each dataset. Specifically, a cumulative eye fixation map from all the subjects for an image is convolved with a Gaussian mask. Then a 100×100 square bounding box centered on the maximum local response point is cropped, as the salient region of this image which we expect to collect. With the collected salient regions, sparse coding is only performed on the salient regions. Then, the sparse representation is sent to the first layer's convolution unit, followed by RBM and max pooling operations.

Training Similar to the pre-training step, the training of the proposed CNN network is performed with features from the pre-training step as the input on the same collected salient regions.

2.1.2. The Architecture of the New CNN Network

The following describes the composition of our proposed CNN network at one layer. It consists of sparse coding, convolution, Restricted Boltzmann Machine (RBM) and pooling operations. Sparse coding produces a "sparse representation" [24] of visual perception information. This sparse representation can effectively reduce the data redundancy and can be a better input data to CNN network for feature learning. Convolution indicates: We align the center of the convolution window with each pixel, and weight all the pixels covered by the window as the response of the convolution process to this pixel on this image. Pooling is a process of aggregating specific features of a region into one single feature in an image. In our method, max-pooling operation is used to select the maximum value of specific features in the pooling window.

RBM A Restricted Boltzmann Machine is a generative stochastic artificial neural network that can learn a probability distribution over its set of inputs. As its name implies, a RBM is a bipartite graph which restricts a pair of nodes from each of the two groups of units (commonly referred to as the "visible" and "hidden" units respectively) may have a symmetric connection between them; and there are no connections between nodes within a group. In our pa-



(a) MIT dataset (b) FIFA dataset (c) MIT300+MSRA-B dataset

Fig. 3. Visual comparisons of our results with the state-of-the-art methods on MIT, FIFA and MIT300+MSRA-B dataset. The ground truth (GT) is shown in the last column. Our method produces saliency maps closest to the ground truth. We compare our method against saliency model (Itti [17]) , global and local cues (GL [18]) , Bootstrap Learning (BL [19]) , visual attention mechanism (VAM [20]) , saliency qualification random field (sQRF [21]) , segmenting salient objects from images and video (SOIV [22]) , salient region detectors (SRD [23]).

per, we adopt two layers' (i.e. one visible layer and one hidden layer) RBM. The connection between the visible units $v \in R^m$ and the hidden units $h \in R^n$ is associated with a weight matrix $W \in R^{m \times n}$. Thus, the parameters of RBM are $\theta = \{W, a, b\}$, where $a \in R^m$ is the bias (or offsets) for the visible unit v and $b \in R^m$ is the bias (or offsets) for the hidden unit h. After several iterations, more accurate features after the convolution operation in CNN network can be extracted via RBM.

2.2. DBN Classifier Training

DBN [8] is a generative graphical model, composed of multiple layers of hidden units and one layer of visible units. It also can be viewed as a composition of several RBM networks. Specifically, the first RBM is sufficiently trained in the first place. Then the hidden units at this layer are passed to the second RBM as visible units, and the second RBM is trained and stacked together with the first trained RBM. The process is repeated until the top two layers. The top two layers have undirected connections and are usually enabled with the discriminative capacity. Therefore, after learning the parameters of lower layers, a DBN can be further trained in a supervised way to perform classification. The expected classification results are compared with the output of the network, and the difference between the output and the expected results is used to fine-tune the previously learned RBM parameters (i.e. weights).

2.3. Saliency Prediction

Given an image, the saliency is predicted with the trained DBN network. To obtain the high-level features of this image, the whole image is as input and sent to the trained multi-layer CNN network. In other words, sparse coding is conducted to every patch of the image, followed by convolution, RBM and pooling. The feature matrix at the last layer (i.e. the two-fully connected layer) is as the input to the DBN network. After many iteration of running, a weight matrix corresponding to the high-level features of this image is obtained, and the saliency value of this patch can be obtained via multipling the weight with the features.

$$S = G \circ max(w^T x, 0), \tag{1}$$

where w is learned parameters of the DBN classifier, and x are the high-level features matrix extracted by the well trained CNN network. Denotes G the Gaussian masking template, which is used to tackle the center bias issue in human visual mechanism [25].

3. EXPERIMENTS AND DISCUSSION

In this section, experimental results are reported to validate the proposed saliency estimation model. Firstly, we compare our method with seven state-of-the-art methods on three public datasets qualitatively and quantitatively. Besides, in order to analyze and evaluate the performance of our method, we present the effectiveness analysis by adding RBM component in CNN for high-level feature extraction. Moreover, to show the superiority of using DBN for saliency prediction, the performance comparisons between the proposed CNN network with RBM and the proposed CNN network with SVM are also presented.

3.1. Performance Comparisons with State-of-the-art Methods

The results of our experiments are compared with Itti [17], GL [18], BL [19], VAM [20], sQRF [21], SOIV [22], SRD [23]. The results of Itti, VAM, SOIV and SRD are based on low-level features, and sQRF, BL and SOD are using high-level features. We compare with them quantitatively and qualitatively.

Quantitative Comparison Figure 3 shows the saliency map generated by our method and seven other methods on MIT, FIFA and MIT300+MSRA-B dataset. Experimental results have shown that our methodscan generate high quality saliency map, clearly outperforming the existing methods. Through subfigure b in Figure 4, we visualize the results from

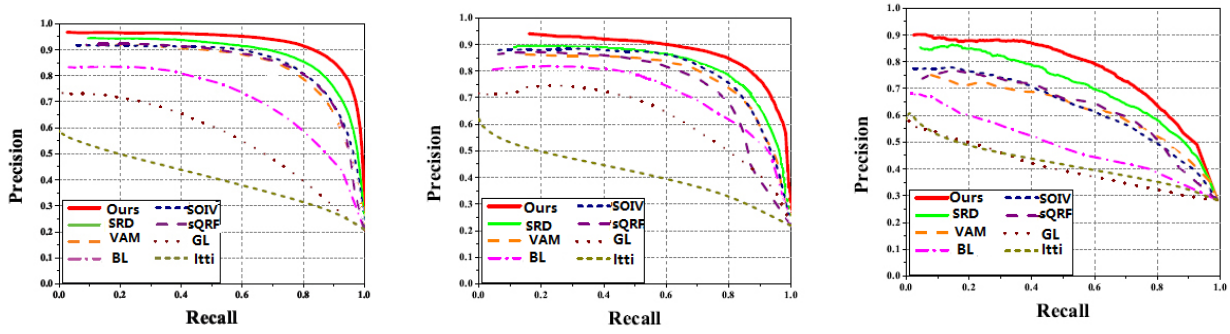


Fig. 4. Quantitative comparison the precision-recall curves of different methods on 3 datasets: (a) MIT dataset (b) FIFA dataset (c) MIT300+MSRA-B.

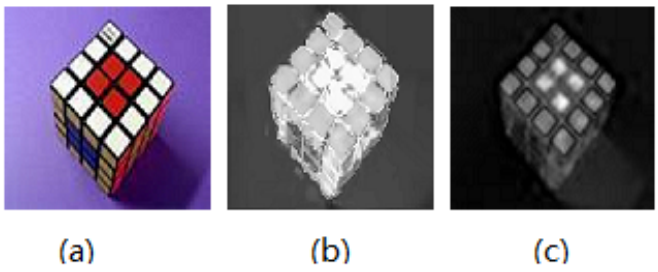


Fig. 5. Illustration the benifit by using RBM in CNN for high-level feature learning. Final saliency prediction results are provided. (a) original image (b) CNN model without RBM (c) CNN model with RBM. CNN model with RBM gets better saliency results in terms of clear salient object boundary and detailed object information (better viewed in color).

various diffcults cases including multiple salient objects (row 4), toching boundary examples (row 5) and face images of different postures and locations (row 1-6). We can see that the saliency map obtained by our method have complete object contours and accurate salient features.

Qualitative Comparison As shown in the Figure 4, our method achieves the highest precision in the entire recall range in three datasets. The results of P-R curves demonstrate that the proposed method outperforms the state-of-the-art methods. On the MIT dataset, the proposed method obtains the highest precision value of 98.5%, which is 3.5% higher than the best one (95% in the SRD method).

3.2. CNN with RBM vs CNN without RBM

In order to show the effectiveness by adding RBM component in CNN network for high-level feature extraction, we perform the performance comparisons on the MIT300+MSRA-B dataset for the CNN with RBM and the CNN without RBM. Except for RBM, all other settings of the CNN network and the DBN classifier are kept untouched. Sampled results of saliency prediction by CNN with RBM and CNN without RBM are shown in Figure 5, we can see that, four red stickers in the middle of Rubik's Cube can be clearly detected by CNN network with RBM and the edge is more distinct. However, edges detected by the CNN network without RBM are comparatively blurred. Therefore, CNN network with RBM achieved better performance in detecting saliency.

3.3. Performance Comparisons with DBN and SVM on Saliency Prediction

In order to demonstrate the superiority of using DBN for saliency prediction with our proposed CNN network, the performance of F-measure is compared on MIT dataset. From the experimental results, our method achieve great performance improvement from 83.57% to 92.85% by replacing SVM with DBN for saliency prediction.

4. CONCLUSION

In this paper, we have introduced a new saliency model based on a multi-layer CNN framework and demonstrated its superiority of the learned high-level features to predict visual saliency. Different from the other methods, when doing features extraction, apart from the convolution and pooling in CNN, extra operation RBM is added into the CNN framework to obtain more accurate features in intermediate step. Meanwhile, in order to deal with the case of non-linear classification, we add the DBN classifier at the end of this model to classify the salient and non-salient regions. As a future work, we are planning to explore various CNN architectures to further improve the performance of saliency estimation.

5. ACKNOWLEDGMENT

This work was supported by the General Program of National Natural Science Foundation of China (NSFC) under Grant No.61572362. This research was also partially supported by the General Program of National Natural Science Foundation of China (NSFC) under Grant No.81571347.

References

[1] C Koch and S Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry.," *Human Neurobiology*, vol. 4, no. 4, pp. 219–27, 1985.

[2] M. Feng, A. Borji, and H. Lu, "Fixation prediction with a combined model of bottom-up saliency and vanishing point," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–7.

[3] Laurent Itti and Christof Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[4] P. K. Podder, M. Paul, T. Debnath, and M. Murshed, "An analysis of human engagement behaviour using descriptors from human feedback, eye tracking, and saliency modelling," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, Nov 2015, pp. 1–8.

[5] Moran Cerf, Jonathan Harel, Wolfgang Einhuser, and Christof Koch, "Predicting human gaze using low-level saliency combined with face detection," *Advances in Neural Information Processing Systems*, vol. 20, pp. 241–248, 2007.

[6] Tilke Judd, Krista Ehinger, Frdo Durand, and Antonio Torralba, "Learning to predict where humans look," *Proceedings*, vol. 30, no. 2, pp. 2106–2113, 2009.

[7] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June*, 2009, pp. 609–616.

[8] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, no. Jan, pp. 1–40, 2009.

[9] Gayoung Lee, Yu Wing Tai, and Junmo Kim, "Deep saliency with encoded low level distance map and high level features," 2016.

[10] Lijun Wang, Huchuan Lu, Ruan Xiang, and Ming Hsuan Yang, "Deep networks for saliency detection via local estimation and global search," in *Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[11] M Riesenhuber and T Poggio, "Hierarchical models of object recognition in cortex.," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–25, 1999.

[12] Y-Lan Boureau, Jean Ponce, and Yann LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 111–118.

[13] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Saliency detection by multi-context deep learning," in *Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[14] Xiao-Xiao Niu and Ching Y. Suen, "A novel hybrid cnnsvm classifier for recognizing handwritten digits," *Pattern Recognition*, vol. 45, no. 4, pp. 1318 – 1325, 2012.

[15] Hongyang Li, Huchuan Lu, Zhe Lin, Xiaohui Shen, and Brian L. Price, "LCNN: low-level feature embedded CNN for salient object detection," *CoRR*, vol. abs/1508.03928, 2015.

[16] Chengyao Shen and Qi Zhao, "Learning to predict eye fixations for semantic contents using multi-layer sparse network," *Neurocomputing*, vol. 138, pp. 61–68, 2014.

[17] Laurent Itti, Christof Koch, Ernst Niebur, et al., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] Na Tong, Huchuan Lu, Ying Zhang, and Xiang Ruan, "Salient object detection via global and local cues," *Pattern Recognition*, vol. 48, no. 10, pp. 3258–3267, 2015.

[19] Na Tong, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, "Salient object detection via bootstrap learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1884–1892.

[20] X.D. Gu, Z.B. Chen, and Q.Q. Chen, "Salience estimation for object-based visual attention model," Feb. 26 2013, US Patent 8,385,654.

[21] Richard M Jiang and Danny Crookes, "Deep salience: Visual salience modeling via deep belief propagation.," in *AAAI*, 2014, pp. 2773–2779.

[22] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*. IEEE, 2009, pp. 1597–1604.

[23] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä, "Segmenting salient objects from images and videos," in *European Conference on Computer Vision*. Springer, 2010, pp. 366–379.

[24] David H Hubel and Torsten N Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[25] Diane M. Beck and Sabine Kastner, "Top-down and bottom-up mechanisms in biasing competition in the human brain," *Vision Research*, vol. 49, no. 10, pp. 1154–1165, 2009.