Unsupervised Neural Architecture for Saliency Detection

Natalia Efremova* and Sergey Tarasenko
Plekhanov’s University,
Stremyanny per., 36, 117997, Moscow, Russia
{natalia.efremova,infra.core}@gmail.com

Abstract. We propose a novel neural network architecture for visual saliency detections, which utilizes neurophysiologically plausible mechanisms for extraction of salient regions. The model has been significantly inspired by recent findings from neurophysiology and aimed to simulate the bottom-up processes of human selective attention. Two types of features were analyzed: color and direction of maximum variance. The mechanism we employ for processing those features is PCA, implemented by means of normalized Hebbian learning and the waves of spikes. To evaluate performance of our model we have conducted psychological experiment. Comparison of simulation results with those of experiment indicates good performance of our model.

Keywords: neural network models, visual saliency detection, Hebbian learning, Oja’s rule, psychological experiment

1 Introduction

Precise processing of only important regions of visual scenes is one of essential properties of human visual system. Extraction of important regions is referred to as selective attention: “The mechanism in the brain that determines which part of the multitude of sensory data is currently of most interest is called selective attention” [1] (p. 6).

To date, two type of attention have been distinguished: bottom-up and top-down. Bottom-up attention is believed to be triggered only by characteristics of the visual scene. Top-down attention is hypothesized to be driven by higher cognitive processes (individual’s previous experience, expectation and current goals).

In this study, we only concentrate on bottom-up selective attention, which is also referred to as saliency. Frintrop et al. [1] (p. 6) defines saliency as follows: “Regions of interest that attract our attention in a bottom-up way are called salient and the responsible feature for this reaction must be sufficiently discriminative with respect to surrounding features”.

* Authors contributed to the paper equally. S.Tarasenko is independent researcher.
In this paper, we introduce a unsupervised neural architecture, which detects the salient regions in visual scenes in neurophysiologically plausible way on the one hand, and in the computationally inexpensive manner on the other hand. To achieve this goal, we use three main neurophysiological findings which comprise a neurophysiologically plausibility of our model: 1) findings by Knierim and Van Essen [2]; 2) waves of spikes [9,10,11]; and 3) regularised Hebbian learning [4,8,5].

It was illustrated in the seminal work by Knierim and Van Essen [2] that activation of neurons in primary visual area V1 of macaque monkey depends not only on the actual stimulus, presented in the receptive field (RF) of a given neuron (stimulus 1), but also on how different are the stimuli presented to neighbour neurons from the stimulus 1. The work by Knierim and Van Essen [2] is considered to provide insight into saliency mechanism on the neural level [16].

Mechanism of attention based on waves of spikes has been proposed by Van Rullen and Thorpe [11]. They described the rapid and reliable mechanism of processing natural input scenes by ventral stream of the visual system, based on the relation between stimulus saliency and spike relative timing. According to their theory, waves of spikes are generated in the retina in response to a visual stimulation and the obtained information is transferred further explicitly in its spatio-temporal structure: the most salient information is represented by the first spikes over the population. While propagating through a hierarchy of visual areas, this spike wave is regenerated at each processing stage, where its temporal structure can be modified either by the selectivity of the cortical neurons, or lateral interactions and or top-down attentional influences from higher order cortical areas. Since we are interested only in the bottom-up processes, we do not take into account the last two influences.

The next mechanism we utilize is principal component analysis (PCA) that was proven to be effective instrument for saliency detection [?]. However, we suggest more neurophysiologically plausible mechanism for implementation of PCA: we employ Oja’s rule to extract the first principle component in a neurophysiologically plausible way [8].

The rest of the study is organised as following: first, we describe the architecture of the proposed model and the implementation of the above mechanisms in our model. Next, we introduce the psychological experiment with human subjects and comparison of experimental results with results produced by our model. Finally, we evaluate the performance of our model and discuss the future work.

2 Model Description

In this section, we describe in details each computational principle we employ, and the way how all three principles are integrated together to deliver simple neurophysiologically plausible mechanism of saliency detection.
2.1 Unfold Static Image in Time by using Waves of Spikes approach

The neurons, whose preferred stimulus is closer to the actual visual input, fire faster [11]. Therefore, the higher is the activation of neuron the faster it fires. In [17](p. 1518), the following approach for modelling such process was suggested: “the neurons with unit activation (response is 1 on the normalized scale) fire first. Then the neurons with activation greater or equal to 1-ε (ε = 0.1), excluding previously fired neurons, will fire and so on. Thus, the neurons with the same activation level form ‘waves of spikes’. Consequently, the original retinal image is unfolded in a time domain.”

In the case of color intensity, we consider that the signals corresponding to the higher color intensities are produced faster than signal corresponding to the lower color intensities. First the highest color intensity (unit amplitude on the normalized scale) signal will be generated.

To employ the principle of waves of spikes in this study, we consider color RGB image as three independent R-, G-, and B- channels. Then we decompose signal of each channel into 10 intensity levels starting from 0 with step ε. Thus, the first level contains color intensities from the interval [0, 0.1), second level contains intensities from the interval [0.1, 0.2), etc.

Therefore it is possible to decompose entire color image into several layers (level of intensities) for each of three color channels.

![Fig. 1. Black and white image decomposition into intensity layers: a) original black and white; b) 3d decomposition using intensity layers.](image)

After such decomposition, we split each intensity layer for each color channel into 16 x 16 pixels patches. Each 16 x 16 patch is considered to be a RF of a visual neuron, which extracts the first (main) principle component by means of neurophysiologically plausible analogue of PCA. In the next section, we consider such neurophysiologically plausible analogue in greater details.
2.2 PCA for saliency and neural implementation of PCA: Hebbian learning and Oja’s rule

PCA was used by Zhou et al. [6] in conjunction with distance computation and histogramming methods to detect salient regions. The goal of our study is to provide neurophysiologically plausible algorithm (PCA itself is not plausible enough for our purposes). On the other hand, Hebbian learning [4] and normalized Hebbian learning (Oja’s rule) [8] are biologically plausible analogues of PCA. Hebbian learning and Oja’s rule deliver the first principle component into weights of a given neuron tuned by such learning rule.

Hebbian learning. Hebb has suggested that certain cortical cell populations behave as functional units with coordinated activity patterns in accordance with changing synaptic strength [4]. It was assumed that the strength of a synapse (connection) between neuron 1 and neuron 2 increases, when firing in neuron 1 is followed by firing in neuron 2 with a very small time delay.

The discrete Hebbian learning rule for a single neuron can be presented in the form of eq. (1):

$$\Delta w(k+1) = \mu [y(k)x(k) - \alpha w(k)]$$

where $y$ is output of the neuron, $w = (w_1, ..., w_n)$ is vector of synaptic weights, $\mu$ and $\alpha$ are constants regulating learning and forgetting rates, respectively, and $k$ indicates the iteration number. The output $y$ can be calculated as $y = w^T x$, where $w = (w_1, ..., w_n)$ is a vector of synaptic weights, $x = (x_1, ..., x_n)$ is vector of inputs, and $n$ is a total number of inputs of the given neuron.

The equation for calculation of the synaptic weights update can be formulated as following

$$w(k+1) = (1 - \alpha \mu) w(k) + \mu y(k)x(k)$$

Oja’s learning rule. The main shortcoming of this learning rule is unlimit growth of synaptic weights. To overcome this shortcoming, Oja [8] has proposed normalized Hebbian or Oja’s rule:

$$w(k+1) = w(k) + \mu(y(k)x(k) - y(k)^2 w(k))$$

The final weight vector after learning will represent the first (main) principle component of the presented data. In this study, we employ Oja’s rule to evaluate main principle component of the color intensity presented in each patch.

2.3 Determine salient regions

In the work by Knierim and Van Essen [2], it was shown that neural responses depend on the activation of surrounding neurons. To incorporate this property into our model, we consider that the weight vector of a given neuron serves as a input into the surrounding neurons, after Oja’s learning has been completed.

In other words, the same neurons first takes its input from a RF (16x16 patch) for the purpose of Oja’s learning. After learning is completed, it takes...
inputs from all surrounding neurons with stimulus being synaptic weights of surrounding neurons. In this case, there are eight possible connections between the given neuron (at the center) and surrounding neurons.

As all synaptic weight vectors are the first principle components (orthonormal vectors), the output of a neuron is cosine of an angle between two principle components. We consider the cosine of such angle to characterize the degree of similarity between two principle components. In other words, by measuring the cosine of angle between two synaptic weights (the first principle components) we evaluate how similar are original stimuli at the RFs of the two given neurons.

We set the decision rule to distinguish between similar and dissimilar stimuli: if \( w_i^T w_j < 0.1 \), then we consider two stimuli to be dissimilar, where \( w_i \) and \( w_j \) are synaptic weights after Oja's learning of neurons \( i \) and \( j \), respectively.

To process a stimulus in the RF of a given neuron, we compute first principle component for each color channel for the given patch. Then we calculate the total number of dissimilar weight vectors in a surrounding neurons for each color channel separately. If the number of dissimilar weight vectors across all intensity level for a certain color channel is greater than a threshold, we consider that given stimulus is a salient region for the given color channel. In this study, we use threshold value equal to 10.

Finally, throughout the color channels we calculate frequency of a certain patch to contain salient regions. At this point we cut off the patches with frequencies less than an expected value by chance. The expected value is calculated as a ratio of total number of patches containing salient region across color channels and intensity layers to the total number of patches to cover the figure.

3 Psychological Experiment

3.1 Method

Objective. In this study, we aim to show that the type of model that we propose can predict the location of salient regions in natural scenes and home scenes in the similar way human subjects do. To evaluate performance of our model, we have conducted a psychological experiment to collect the data about performance of human subjects.

Participants. All participants were undergraduate students (ages 18–23) at the Plekhanov's University, who received course credit for participation. All participants had normal or corrected-to-normal vision. In total 18 subjects were tested, including 11 males and 7 females.

Procedure. During the experiments, we asked participants to select all the regions, which attracts their attention on the given set of images. The images are taken from the database, provided by Bruce and Tsotos [13,14]. In particular, we used 21 images numbered 20 to 40 in Bruce original enumeration. In this study we enumerate images starting from 1 to 21 beginning with image 20.

All the subjects were given the instructions to select all the regions of the image that attract their attention. They were ask to do it by using mouse to
encompass the Region Of Interest (ROI) with a color curve on computer screen. Each ROI was defined as an area within the mouse selection.

3.2 Results of Psychological Experiment

For the purpose of further analysis, we have discarded results of two subjects. One subject has not selected any region in either of images. The regions, selected by another subject, were too large for our purposes (over 30 percent of a visual scene). We have also discarded some results of other participants for the same reason.

Finally, we have obtained from 12 to 16 individual ROI maps for each image. To obtain final integrated ROI map we have superimposed individual ROI map for a single image produced by all subjects. Therefore integrated ROI map illustrates frequencies of selection of specified ROIs in a visual scene.

The sample integrated ROI map for one of the images is presented in Fig. 2. It demonstrates the selected regions for all of the tested subjects (heat map), where red color represents the most frequently selected region of the scene and the blue color corresponds to the least frequently selected region of the scene. This integrated ROI map is for image 32 in Bruce and Tsotos [13-14] original enumeration.

![Fig. 2. The results of the psychological experiment for one of the images](image)

4 Evaluation of the proposed model

To evaluate performance of our proposed model, simulation results produced by our model were compared with those of the human subjects. On Fig. 3 the simulation results are compared with the results of the psychological experiment for image 32. Complete results of model performance for each image vs integrated ROI map is presented in Table 2.
To evaluate the performance of our model, we compared the salient regions selected by our model with ROI maps. To measure performance, we calculate precision and recall rates. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of retrieved relevant instances [15]. We consider total number of patches, which contain at least one non-zero value of ROI map value, to be positive samples.

The total number of patches, which are selected by our model, and contain at least one non-zero value of integrated ROI map, we consider as true positives ($tp$). While the patches selected by our model, but containing only zero values, we consider as false positive ($fp$). The number of patches, which contain at least one non-zero value, and were not selected by our model, we consider to be false negative ($fn$).

Then precision is calculated as $\text{precision} = \frac{tp}{tp + fp}$. Recall is calculated as $\text{recall} = \frac{tp}{tp + fn}$.

The integrated ROI maps are frequency maps that illustrate how many participants have selected particular regions. Therefore the regions, selected by the majority of subjects (with higher frequencies) can be considered more important than others. Therefore, detection of more frequently selected regions demonstrate better model performance than detection of less frequent ones.

To take this idea of region importance due to selection frequency into account, we also calculate weighted precision and weighted recall along with standard ones. By normalizing the integrated ROI maps, we obtain probability density of region selections. Therefore, the sum of all non-zero values should equal to 1. This probability distribution represents all positive samples. In this context, precision is a part of probability density contained in patches, which were selected as salient ones by our model. We refer to precision defined in this way as weighted precision.

However, calculation of recall is not that straightforward, because all false positive patches contain zero values. To calculate weighted recall, we first add
Table 1. Evaluation of model performance

| Image Number | Recall  | Precision | Weighted Recall | Weighted Precision |
|--------------|---------|-----------|-----------------|--------------------|
| Image 1      | 0.925   | 0.315     | 0.944           | 0.375              |
| Image 2      | 0.914   | 0.275     | 0.938           | 0.486              |
| Image 3      | 0.797   | 0.431     | 0.827           | 0.398              |
| Image 4      | 0.956   | 0.302     | 0.967           | 0.511              |
| Image 5      | 0.728   | 0.399     | 0.774           | 0.567              |
| Image 6      | 0.631   | 0.162     | 0.701           | 0.284              |
| Image 7      | 0.416   | 0.448     | 0.486           | 0.425              |
| Image 8      | 0.330   | 0.434     | 0.360           | 0.562              |
| Image 9      | 0.369   | 0.489     | 0.419           | 0.404              |
| Image 10     | 0.505   | 0.426     | 0.537           | 0.460              |
| Image 11     | 0.658   | 0.218     | 0.734           | 0.411              |
| Image 12     | 0.973   | 0.185     | 0.982           | 0.472              |
| Image 13     | 0.533   | 0.143     | 0.617           | 0.339              |
| Image 14     | 0.865   | 0.450     | 0.887           | 0.599              |
| Image 15     | 0.868   | 0.423     | 0.887           | 0.478              |
| Image 16     | 0.872   | 0.147     | 0.898           | 0.273              |
| Image 17     | 0.953   | 0.169     | 0.970           | 0.451              |
| Image 18     | 0.926   | 0.173     | 0.949           | 0.273              |
| Image 19     | 0.743   | 0.520     | 0.764           | 0.556              |
| Image 20     | 0.849   | 0.371     | 0.864           | 0.379              |
| Image 21     | 0.686   | 0.383     | 0.715           | 0.508              |

1 to all the non-normalized integrated ROI map values. Then we calculate the sum of all integrated ROI elements in $tp$ and $fp$ patches.

The results of calculation of recall, precision, weighted recall and weighted precision are presented in Table 1.

For all images, the value of weighted recall is higher that of standard recall. The weighted precision values is higher than standard precision values for all images except for images 3, 7 and 9. The mean recall and precision values are 0.734 and 0.327, respectively. The mean weighted recall and weighted precision values are 0.772 and 0.439. This results indicate that on average our model selects more frequent salient regions.

Overall results suggest that our model was quite accurate at predicting human judgements of Regions Of Interest in the visual scenes.

5 Discussion and Conclusion

In this study, we have proposed a neurophisiologically plausible neural architecture for saliency detection. The main specifics of our model is that it is contracted in a completely unsupervised manner. The computational principles, which were utilized for building this model, are considered to be employed by the human brain. Finally, our model demonstrates good performance of saliency detection based on only local information and is less computationally greedy then other proposed algorithms.
The observation of model performance illustrates that the proposed model is especially good in detecting salient regions surrounded by homogeneous areas. However, model produces a lot of false positive responses when dealing with high-contrast textures surfaces if they are extensively presented on the image (e.g. grass, carpets, structured bricks etc.). This restriction of a model is a subject for future work.

Table 2. ROI maps and model performance

References

1. Frintrop, S., Rome, E., Christinsen, H.: Computational Visual Attention Systems and their Cognitive Foundations: A Survey. ACM, Vol. 7, No. 1, pp. 1-46, 2010
2. Knierim, J. J. and Van Essen, D. Neural Responses to Static Texture Patterns in Area V1 of the Alert Macaque Monkey. *Journal of Neurophysiology*, vol. 67, no. 4, pp. 961 - 980, 1992.
3. Itti, L., Koch, C., and Niebur, E. 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.
4. Hebb, D. O. *The organization of Behaviour*, Wiley: New York, 1949.
5. Rumelhart, D. E., and McClelland, J. L. *Parallel Distributed Processing*, MIT Press: Cambridge, MA, 1986.
6. Zhou J., Jin Z. and Yung J., Multiscale saliency detection using principle component analysis. *World Congress on Computational Intelligence*, pp. 1457 - 1462, 2012.
7. Frintrop, S., Garcia, M.G., Cremers, A.: A Cognitive Approach for Object Discovery. *The proceedings of the ICPR 2014 Conference*, pp.2329-2334, 2014.
8. Oja, E. A simplified neuron model as a principle component analyzer, *Journal of Mathematical Biology*, vol. 15, pp. 267-273, 1982.
9. Thorpe S. J., Spike arrival times: a highly efficient coding scheme for neural networks, In R. Eckmüller, G. Hartmann & G. Hauske (Eds.), *Parallel processing in neural systems and computers*, vol. 1, pp. 91-94, 1990.
10. Thorpe S., Fize D. and Marlot C., Speed of processing in the human visual system, *Nature*, vol. 381, no. 1, pp. 520-522, 1996.
11. Van Rullen R. and Thorpe S. J., Surfing a spike wave down the ventral stream, *Vision Research*, vol. 42, no.23, pp. 2593-2615, 2002.
12. Oja, E. Neural networks, principle components and subspaces, *International Journal of Neural Systems*, vol.1., pp. 61 - 68, 1989.
13. Bruce, Niel D.B., Tsotsos, John K., A Statistical Basis for Visual Field Anisotropies. *Neurocomputing*, vol. 69, no. 10-12, pp. 1301-1304, 2006.
14. Bruce, Niel D.B., Tsotsos, John K., Saliency, Attention, and Visual Search: An Information Theoretic Approach. *Journal of Vision* 9:3, p1-24, 2009, [http://journalofvision.org/9/3/5/](http://journalofvision.org/9/3/5/). doi:10.1167/9.3.5.
15. van Rijsbergen C.J. Information retrieval. Butterworths, 1979, 208 p.
16. Li, Zh. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, vol. 16, no. 1, 2002, pp. 9 - 16.
17. Tarasenko, S. S. A General Framework for Development of the Cortex-like Visual Object Recognition System: Waves of Spikes, Predictive Coding and Universal Dictionary of Features. In *Proceedings of The 2011 International Joint Conference on Neural Networks*, pp. 1515 - 1522, 2011.