Investigating Temporal Capabilities of the Standard Saliency Model

Liya Merzon[1,2] (presenting author), Georgiy Zhulikov[1], Tatiana Malevich[1,3], Sofia Krasovskaya[1], Joseph MacInnes[1]

1 National Research University Higher School of Economics
2 Aalto University
3 Werner Reichardt Centre for Integrative Neuroscience

The study investigates the Saliency model of Itti & Koch (2000) and its ability to represent both temporal and spatial aspects of visual attention. This model has maintained its popularity for its biological plausibility and its focus on attentional theory (Itti & Borji, 2013). The model decompose images into bottom-up attentional features (e.g. color, intensity and orientation) and combines them in a spatial saliency map. A simplified version of neuronal population, based on leaky integrate-and-fire (LIF) neurons builds up on the saliency map and predicts when and where attention will be directed next (Itti & Koch, 2000).

The model's performance in predicting the spatial locus of attention has been well studied (Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002; Itti & Borji, 2013), however, the temporal aspect of the attentional shift has been neglected, despite the temporal aspects of the LIF layer. The aim of the study was to investigate the final layer of the model, winner-take-all (WTA) layer in order to test its ability to make an accurate temporal prediction of events (latency of saccades / duration of fixations). This layer of neurons uses LIF model, which is a simplified neuronal activation model able to predict neuronal spikes (Itti & Koch, 2001).

The current study included 3 experiments. For all of the experiments, only the very first fixation on the image was used to avoid confusion related to influence by Inhibition of Return. As the ground truth for all experiments we used data from the previous study (Gordienko, 2016). The first, smaller dataset, included data, collected on 44 pictures and consisted of 782 first fixations (29528 fixations in total). This dataset was used in Experiment 1. The second, larger, dataset included data, collected on 91 pictures: 1593 first fixations (60186 fixations in total); it was used in the Experiments 2-3.

In Experiment 1 the default parameter space from Walther and Koch (2006) was tested against the human data. The prediction of the model was significantly different from the ground truth, the result of Kolmogorov-Smirnov test allows to dismiss the null hypothesis that the two were sampled from the same distribution *(KS-test: $D = 0.58067$, $p < 2.2e\text{-}16$)*. The default parameter space, therefore, cannot be considered as acceptable for modeling temporal aspects of saccadic movements, so the optimization process was launched.

In Experiment 2 the LIF parameters were optimized by genetic algorithm (Davis, 1991) and Nelder–Mead method (Nelder & Mead, 1965; MathWorks, 2018), using a combination of z-test and ks-test statistics as the fitness function. The best parameters were found by Nelder-Mead algorithm, however it didn't pass even z-test *(z-test statistics: $z = 3.1117$, $p = 0.00186$, ks-statistics: $D = 0.15857$, $p = 1.305e\text{-}09$)*. Thus, no tested combination of parameters yielded a match to human temporal data.

For the third experiment we used sum of ks-test on each image as optimization function to force the model use noise component and learn within image variability of fixation durations. Further investigation showed that LIF relies on variability of input images instead of fixation selection itself as is the case with humans. Human data on one separate image produce a reaction time distribution, however, Experiment 3 revealed, that the model produced as maximum two different values per image, and used only differences in saliency between images to match the ground truth. Apparently, the standard Saliency Model has important limitation and doesn't include enough variability to produce a distribution of reaction times.