Learning Object Representations for Modeling Attention in Real World Scenes

Alex Schwarz, Frederik Beuth, Fred H Hamker, Chemnitz University of Technology, Germany

Many models of visual attention exist, but only a few have been shown with real-world scenes. Beuth, F. and F. H. Hamker, 2015, *NCNC* proposed such a model, which is a succeeding model of Hamker, F. H., 2005, *Cerebral Cortex*. We show how the object representations in this model (Fig. 1) have been learned in a biological plausible way. To learn the representations, we use the principle of temporal continuity as it has been hypothesized that the primate brain uses temporal continuity for the development of invariant objects representations (Földiák, P., 1991, *Neural Computation*). The idea is that on the short time scale of stimuli presentation, the visual input is more likely to originate from the same object, rather than a different one. However, temporal continuity models have been typically used in simple scenes composed of bars, but not in a real-world task.

$$\tau_w \frac{\partial w}{\partial t} = u_t \cdot v_{t-1} - \alpha \cdot w \cdot v_{t-1}^2, \text{ with: } u = \left(r^{\text{V1}} - \theta^{\text{V1}}\right),\ v = \left(r^{\text{HVA}} - \theta^{\text{HVA}}\right)^+ \tag{1}$$

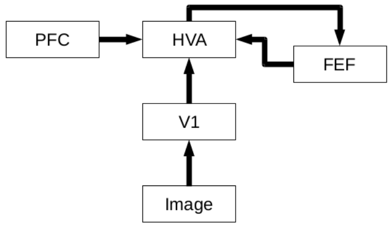

Figure 1: Schematic layout of the model. The image is filtered in V1, and the connection between V1 and HVA will be learned. A feature-based attention signal is send from prefrontal cortex (PFC) to the object-view specific cells in a higher visual area (comparable to IT), altering the strength of their response. The loop to the frontal eye field (FEF) will then find the correct location of the given object and propagate it back to the corresponding location in HVA.

Testing a learning rule (Eq. 1), based on Földiák, P., 1991, *Neural Computation*, with real-world objects, we observed that the HVA cells learn very unspecific object representations. A cell learns whenever its activity exceeds the postsynaptic threshold (θ^{HVA}), which is traditionally the population mean. As this value is relatively low, the cells learn not only for their preferred stimuli, but also for partially preferred ones. We solve this problem by introducing a high postsynaptic threshold: $\theta^{\text{HVA}} = \Psi \cdot \max r^{\text{HVA}}$. The parameter Ψ controls how similar a stimulus has to be, compared to the preferred one, to get learned. With the new threshold method, the cells react more specific for their preferred object than without the threshold (Fig. 2 upper vs. lower row).

However, the novel object representations still learn the background along with the object, whereas an object representation independent of the background would be desirable. It is currently unclear how the human brain learns background invariant object representations. Suggestions cover the usage of disparities or motion, whereas we propose that temporal continuity alone is powerful enough: The background changes much more often than the object, thus the learning rule should not learn connections from the background region. Yet, we observed that inhibitory weights were learned mistakenly. We found that the reason is an difference in the learning speed for weights with positive or negative weight changes. Due to this, we introduce a normalization of τ to ensure that the speed for both cases is balanced. With this novelty, the rule learns zero weights for V1 neurons representing background regions as shown in a miniature model (Fig. 3 a vs. b).

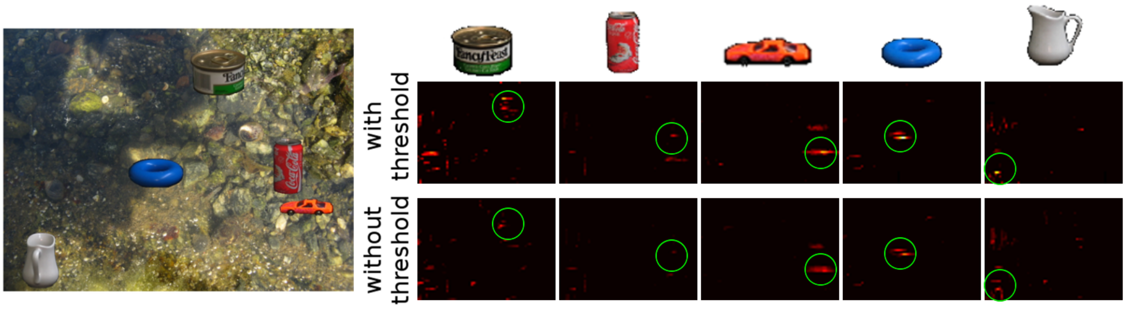

Figure 2: Excitation of 5 chosen HVA cells, screening a given image. It is visible that in the threshold condition ($\Psi = 0.9$), each cell responds highest to its preferred object in the scene (green circle), and respond much less to other objects and the background. In the condition without threshold, the cells react weaker to its preferred object and much broader in the scene, indicating a less meaningful object representation.

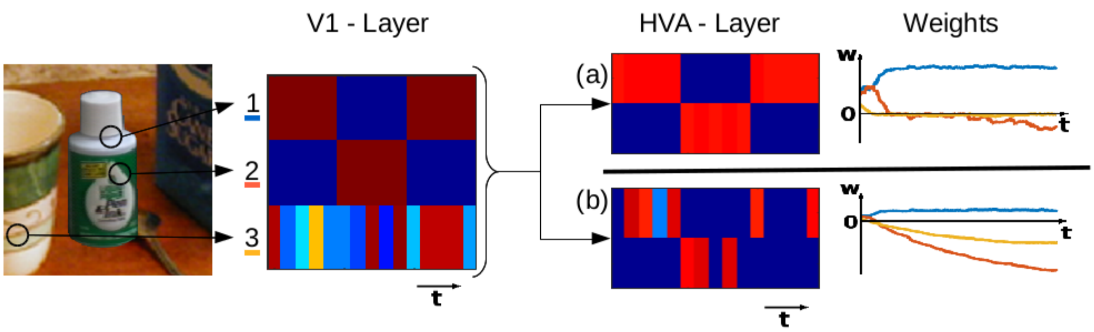

Figure 3: Responses of a miniature version of the model. It visualizes three V1 cells, two HVA cells and the synaptic weights towards the first (upper) HVA cell for the complete learning time. The first two V1 cells encode the features of two alternating-presented objects, while the third cell represents a changing background. **a)** With normalization of τ, the HVA cell responds correctly every time when its preferred stimulus is active because the synaptic weight from the background cell is zero (yellow line). **b)** Whilst without normalization, the HVA cell is inhibited whenever the background is active as this weight is negative.