Proceedings of the 21st Biennial International Congress on Infant Studies
June 30-July 3, 2018

Where and Why Infants Look: A recurrent neural network for the development of visual attention

Yukie Nagai[1] and Niyati Rawal[2]
[1]National Institute of Information and Communications Technology, [2]Osaka University

Visual attention of infants develops as they grow. The scan paths of young infants are scattered over a visual stimulus, whereas the scan paths of older infants and adults converge to a certain pattern. For example, 13.5-month-olds, but not 6-month-olds, tend to attend the eyes and mouth and repeatedly shift their attention between the features when observing an image of a human face (Kato and Konishi, 2013). Individuals with autism spectrum disorder (ASD), in contrast, do not show such a common pattern. Some people with ASD look only at the mouth and locations without any feature, whereas other people look at the eyes and surrounding areas but not the mouth. Their scan paths exhibit a large divergence among individuals (Pelphrey et al., 2002). An open question here is how humans do (not) acquire the common scan path over development.

We introduce a recurrent neural network based on reinforcement learning (Mnih et al., 2014) to account for the underlying neural mechanism for visual attention (see Figure 1). The network receives an image $\boldsymbol{x}_t$ and the attention location $\boldsymbol{l}_{t-1}$ at the previous time step t-1 as input and learns to estimate the category c_t of the image and the attention location $\boldsymbol{l}_t$ at t. Our key idea is twofold: (a) Infants would learn to determine where to attend in order to gain more information from both local and global features of visual input, and (b) integrate the obtained information spatially and temporally in order to optimize the subsequent attention. The scan paths of older infants and adults show repetitive shift between prominent features. They seem to collect detailed information from the features and integrate them to form a global picture of the input. We suggest that two architectures are required to replicate the above behavior: (a) reinforcement learning of visual tasks that require both local and global features of input and (b) a recurrent connection to integrate and internally maintain the obtained information. The network shown in Figure 1 comprises these two architectures. It is trained through reinforcement learning with two types of reward functions r_t: local and global tasks, and the internal state $\boldsymbol{h}_t$ of the network is conveyed to the next time step t+1 through the recurrent connection.

We conducted experiments to examine whether the proposed network reproduced the developmental change observed in infants. We designed two tasks: one to estimate the emotion portrayed in a face image, which would require local features of the image (e.g., shape of the eyes and mouth), and the other to estimate the head orientation represented in a face image, which would need global features of the image (e.g., location of the eyes and mouth). Our experiments demonstrated infant-like development of visual attention produced by the network. The network trained for the emotion estimation acquired local but close attention to the eyes and/or mouth (see Figure 2 (a)). The scan paths were rather similar to those of individuals with ASD. The network trained for the estimation of head orientation acquired wider attention shift while gazing not only at the eyes and mouth but also the ears (see Figure 2 (b)). This task seemed to drive the network to obtain global information regardless of what information to be included. There results suggest that the combination of the local and global tasks is crucial to draw typical development of visual attention.

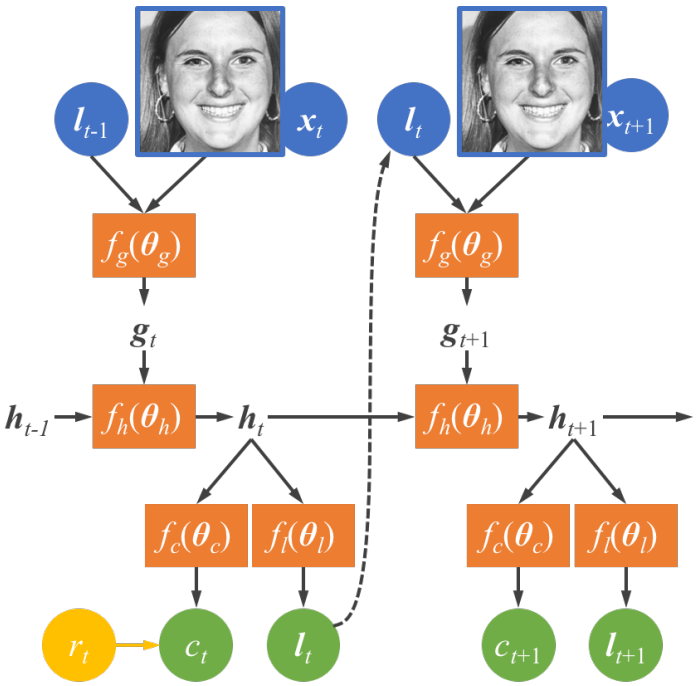
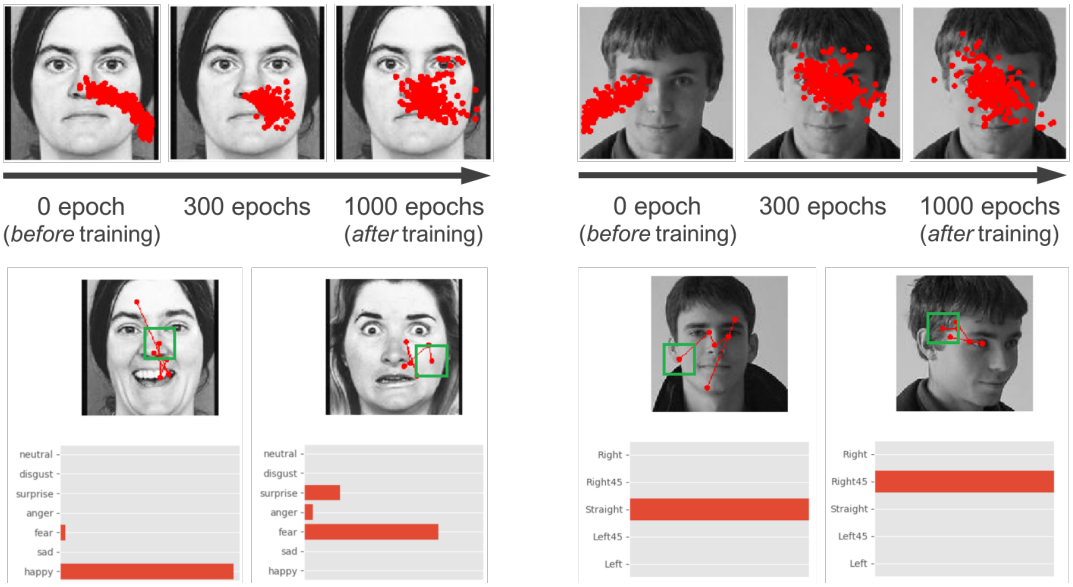


Figure 1: A recurrent neural network for the development of visual attention (modified from (Mnih et al., 2014)). It receives an image $\boldsymbol{x}_t$ and the attention location $\boldsymbol{l}_{t-1}$ at the previous time step t-1 and learns to estimate the category of the image c_t and the attention location $\boldsymbol{l}_t$ at t. The network modules (f_g, f_h, f_c, and f_l) are trained in order to maximize the reward r_t given for the category estimation. The recurrent architecture maintaining the internal state h_t enables the network to leverage the history of attention information.



(a) Result for learning to estimate the emotion. The attention covers local areas including only the mouth or an eye, although the accuracy of emotion estimation was high enough.

(b) Result for learning to estimate the head orientation. The attention covers wider areas including both the eyes and mouth, while the head orientation was accurately estimated.

Figure 2: Experimental results showing the location of visual attention by the neural network. The upper demonstrates the developmental change over learning (0, 300, and 1000 epochs), and the lower depicts the scan path after learning with the probability of estimation.