Task-Dependent Attention Allocation through Uncertainty Minimization in Deep Recurrent Generative Models

Kai Standvoss (kstandvoss@gmail.com)
Donders Institute for Brain, Cognition, and Behavior, Radboud University, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Silvan Quax (s.quax@donders.ru.nl)
Donders Institute for Brain, Cognition, and Behavior, Radboud University, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Marcel A. J. van Gerven (m.vangerven@donders.ru.nl)
Donders Institute for Brain, Cognition, and Behavior, Radboud University, Montessorilaan 3, 6525 HR, Nijmegen, The Netherlands

Abstract

Allocating visual attention through saccadic eye movements is a key ability of intelligent agents. Attention is both influenced through bottom-up stimulus properties as well as top-down task demands. The interaction of these two attention mechanisms is not yet fully understood. A parsimonious reconciliation posits that both processes serve the minimization of predictive uncertainty. We propose a recurrent generative neural network model that predicts a visual scene based on foveated glimpses. The model shifts its attention in order to minimize the uncertainty in its predictions. We show that the proposed model produces naturalistic eye-movements focusing on salient stimulus regions. Introducing the additional task of classifying the stimulus, modulates the saccade patterns and enables effective image classification. Given otherwise equal conditions, we show that different task requirements cause the model to focus on distinct, task-relevant regions. The results provide evidence that uncertainty minimization could be a fundamental mechanisms for the allocation of visual attention.

Keywords: neural networks; visual attention; eye movements; uncertainty; prediction

Visual Attention

Vision is the most dominant sense in human perception and has evolved to rapidly extract relevant information within an observed scene, and filter irrelevant input in a dynamically changing environment. The eyes' anatomy favors the selective processing of relevant stimuli through a central foveal region that processes information with high resolution and a periphery that has a lower resolution. Overt attention — moving an attended region of the visual field into the fovea for heightened processing — is a central selection mechanism to make sense of a visual scene. Through saccadic eye movements, different parts of a scenery are brought into focus, allowing for efficient processing of the surrounding using only a limited bandwidth information channel (Itti & Koch, 2001). Given the relevance of the visual sense and the fundamental role of attention as a mechanism for neural information processing, eye movements and visual attention remain an active field of research (Moore & Zirnsak, 2017).

A central research question is how the brain allocates its attention and decides which locations to sample through saccades when observing a visual scene. Low-level saliency features are strong predictors for human saccade patterns (Soltani & Koch, 2010). Additionally, attention allocation is strongly modulated by task requirements in a top-down fashion (Gilbert & Li, 2013). While both mechanisms have been studied extensively, fewer accounts have investigated their interaction within a comprehensive framework (Moore & Zirnsak, 2017).

One attempt to reconcile top-down and bottom-up attention mechanisms argues that the brain performs saccades in order to actively maximize information gain or conversely minimize uncertainty (Renninger, Verghese, & Coughlan, 2007). In free viewing, these regions of maximal uncertainty correspond to salient stimulus regions which are harder to predict given their surrounding, while in a task setting, uncertainty is shaped by task relevant prior expectations. From a predictive coding perspective, attention is allocated in order to minimize the uncertainty in the brain's predictions (Mirza, Adams, Mathys, & Friston, 2016).

While biologically accurate models like Mirza et al. (2016) can make precise predictions about human behavior and help to understand the neural mechanisms involved in visual processing, their level of abstraction often makes it difficult to interpret the fundamental components at play. A different approach that operates on much higher abstractions are artificial neural network (ANN) models that only loosely connect to computations in the brain. Nevertheless, these models have successfully revealed relevant computational principles of neural information processing (Güçlü & van Gerven, 2015).

So far, ANN models used to explain visual attention have been largely driven by a machine learning focus in order to improve computer vision applications (Mnih, Heess, Graves, et al., 2014). Other approaches that more closely address the pathways in the brain require strong supervision and do not account for different task conditions (Adeli & Zelinsky, 2018).

Here, we propose a deep ANN architecture to investigate

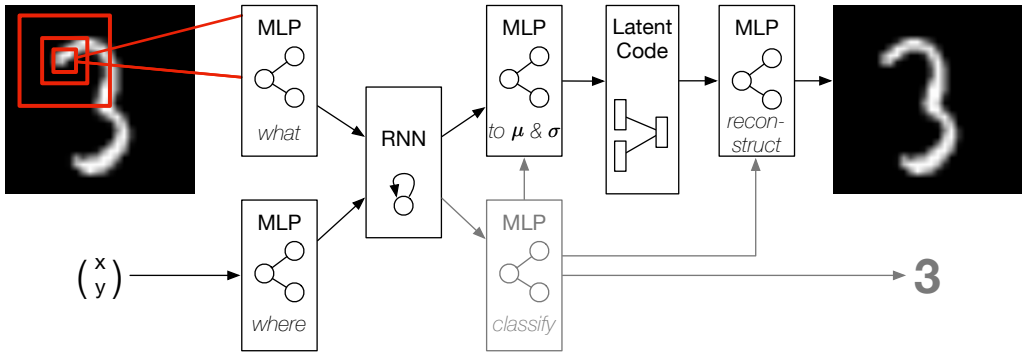

Figure 1: The recurrent generative neural network architecture. Input to the network are the currently attended coordinates and the foveated glimpse. Feature representations of both are integrated in a recurrent layer. The recurrent representation is used to parameterize a Gaussian latent representation. Samples from the latent code are used to generate predictions by the decoder. In the classification experiments, the recurrent representation is additionally used to train a classifier (gray). The classification output is used as input to the latent network and the decoder, disentangling the latent representation into class and style variables.

task-dependent visual attention allocation on an interpretable level of abstraction. Using a generative recurrent latent variable model trained on different tasks, we show that minimizing the uncertainty in the model's predictions of a visual stimulus, based on a series of foveated saccades, constitutes an effective mechanism to learn relevant saccade paths. The proposed model encompasses both bottom-up and top-down processing and can be trained in unsupervised and supervised settings.

Methods

In order to validate whether uncertainty minimization constitutes a useful strategy to allocate visual attention, different experiments were performed. In all experiments, a recurrent ANN architecture was trained to perform a series of saccadic "eye-movements", revealing a limited field of view of a visual stimulus. *Hard attention* was implemented by cropping a foveal part of the scene around the attended location and peripheral patches of increasingly lower resolution. Each patch had twice the side-length of the previous one and all patches were resized to the resolution of the fovea (cf. Figure 2, first row). To process the whole stimulus, the model thus had to integrate the information of the saccade sequence. In all experiments, the model consisted of a latent variable encoder-decoder architecture that was trained as a Variational Autoencoder (Kingma & Welling, 2013) with the objective to reconstruct the full visual scene based on the foveated *glimpses*. The model is depicted in Figure 1. The encoder part consisted of a *what*- and *where*-pathway that were integrated within a recurrent processing layer as proposed by Mnih et al. (2014). The what-pathway extracts relevant features of the currently observed glimpse and the where-pathway represents the currently attended location. Both modules were instantiated as multi-layer perceptrons (MLPs) that feed into a layer of recurrent units. The recurrent activation was used to parameterize the mean and variance of a Gaussian latent distribution. Using the reparameterization trick, the model can effectively sample from the approximate posterior $p(z|x)$ with latent variables z and inputs x. The sampled latent code was then used by an MLP decoder to reconstruct the full input image.

The model predicted the image after each saccade. The next saccade location was the pixel with the highest model uncertainty. Uncertainty was determined by drawing T samples from the latent representation and generating a prediction of the image per sample. The variance in these predictions was taken as the pixel-wise model uncertainty. The model performed a fixed number N of saccades after which the model weights were updated. In all experiments, the MNIST dataset (LeCun, 1998) was used.

Results

Prediction as Objective

The first experiment was used to test whether the model learns to perform useful saccades with prediction as the only optimization criterion without further supervision. As the model only ever sees a limited part of the image, it has to generate a saccade path that samples the stimulus sufficiently to accurately reconstruct the target. The target constitutes a 28×28 MNIST digit whereas the model receives only 4×4 crops centered around the attended location and low resolution peripheral patches of sizes 8×8 and 16×16. For each training stimulus, the first saccade was focused on the center of the image. The model then performed $N = 5$ saccades with $T = 5$ forward passes through the decoder to determine the predictive uncertainty.

Figure 2 shows an example saccade trajectory for a test stimulus. The rapidly decreasing input resolution towards the periphery requires the model to effectively sample the full image. It can be seen that the model predictions get succes-

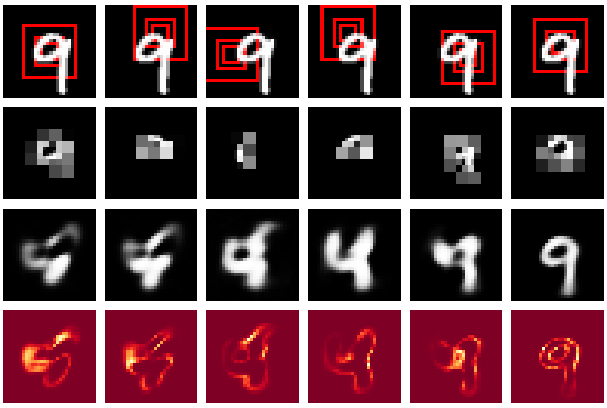

Figure 2: Example episode of the model reconstructing a test stimulus. The first row shows the target image and the currently attended location. Red squares represent different levels of resolution, with the smallest patch corresponding to the full resolution fovea and the larger patches to the lower resolution periphery. The second row displays the current input to the network. The third row shows the model's prediction after receiving the current input and the fourth row shows the pixel-wise model uncertainty obtained through sampling with lighter regions corresponding to higher uncertainty.

sively better with the number of saccades. It can be further seen that class properties are not determined for the first saccades and remain ambiguous until the last saccade. Especially after the final saccade, model uncertainty is highest for edges of the predicted stimulus. The mean model variance decreases per saccade.

In Figure 3 the saccade paths are visualized as heatmaps. Figure 3a) shows on which parts of the stimulus the model focuses within the first three saccades, for two example digits. It can be seen that the model learns to follow the digit contours. Figure 3b) shows the focus points of the model during the last two saccades, closer to the point at which the model has to make a final prediction and the loss is calculated. Here, the focus of the model shifts more towards the center of the digit.

Supervised Learning

The second experiment tested whether the learned saccades are useful for classification. For that, the model was trained to predict the digit class next to reconstructing the image. The reconstruction criterion was still required to obtain pixel-wise uncertainties to guide saccades. A separate MLP was trained for classification, using the recurrent feature representations as input. Its output was compared to the target class and was additionally used as input to the decoder. Uncertainty was calculated by sampling both the latent style representation as well as the class prediction. Using only the uncertainty to guide saccades, the model obtained a test error of 2.25% without decrease in reconstruction quality and no further hyperparameter optimization. It was observed that, with addition

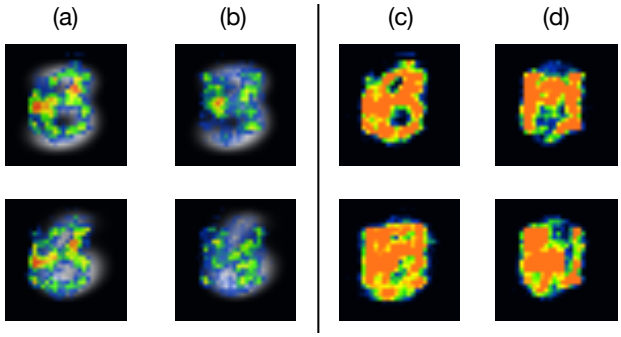

Figure 3: Heatmaps of saccade targets over the test set. Heatmaps were generated separately for the first three and last two saccades in order to account for temporal patterns in the saccade paths. a) Heatmap of the first three saccades over all threes (above) and sixes (below) in the test data. b) Heatmap of the last two saccades. c) Heatmap of the first three saccades over all digits combined for experiment I without classification (above) and experiment II with classification (below). d) Heatmaps of first three saccades over all digits for experiment II, with target class 3 (above) and target class 6 below.

of the classification task, the model produced earlier reconstructions that more clearly resemble a digit. To test whether the introduction of the classification task changed the saccade pattern in order to commit to a scene reconstruction more quickly, the fixation targets for the first three saccades over all test digits are visualized as heatmaps in Figure 3c). The upper heatmap shows the saccade pattern for the model without classification while the lower image displays the saccades of the model that is trained with classification as an additional task. It can be seen that the models' sampling strategies are different. The classification model more broadly covers the whole extent of the average digit location.

Task-Dependent Modulation

The previous experiment revealed changes in the saccade patterns through introduction of an additional task. However, the classification model introduced new network components. In order to further investigate how task requirements modulate gazing behavior for models employing the same architecture, the final experiment used two binary classification tasks that require the model to decide whether the observed stimulus corresponds to a target class. One network was trained to differentiate threes from all other digits and the second one to decide between sixes and non-sixes. This simple task was efficiently learned with an accuracy of 99.5% and 99.25% for sixes and threes, respectively. Figure 3d) displays the saccade targets for the two task conditions. Both models focus on different regions of the input during the first three saccades. The two final saccade patterns are not displayed as they more strongly resemble each other and cover the stimulus regions more uniformly.

Discussion

Attention is a fundamental mechanism of neural information processing and focusing on relevant stimulus regions through saccadic eye movements is a key ability of humans to interact within a dynamically changing environment. While both bottom-up influences on attention allocation as well as top-down task modulation are well studied, their interaction is not yet fully understood (Moore & Zirnsak, 2017). A parsimonious reconciliation of these two processes posits that both serve the common objective to minimize uncertainty about the sensed world (Renninger et al., 2007). Exploring this hypothesis, we proposed an artificial neural network model that makes predictions of an observed scene and performs saccades in order to minimize uncertainty in its predictions. Uncertainty in this model is represented implicitly by sampling from a generative model of the observed scenes. As different task requirements affect the model's latent representations, the model uncertainty is affected both by bottom-up saliency features of the observations as well as top-down task demands.

The first experiment showed that the simple heuristic to select the pixel with the highest uncertainty under the objective of scene reconstruction leads to sensible saccade paths. In that way, a model can be trained without supervision to sample a target stimulus. The model learned to follow digit contours along the edges, which also correspond to the most salient stimulus parts. Investigating the temporal order of the saccades revealed that the initial saccades were mostly targeted at regions just outside of the digits while later saccades focused more on the digit centers. One explanation for this observation could be that earlier glimpses served to determine the stimulus class while later saccades improve the prediction just before the stimulus target is provided. This could also explain why task differences in experiments II and III were most prominent in the earlier saccades.

Introducing a supervised task to the model showed that the proposed attention mechanism can also be used for classification with only a limited number of fixations. The learned saccade paths yield better classification scores than random fixations. Further research is required to see whether the approach can compare to other machine learning methods like RAM (Mnih et al., 2014) for more complex tasks. However, without the need for reinforcement learning to train the attention mechanisms it could offer a valuable alternative.

The classification task modulated the saccade pattern of the model, especially for early saccades. With prediction as the only optimization criterion the saccades mostly resemble digit contours. The classification task caused the model to more quickly reconstruct stimuli of the predicted class. In order to do so, the first saccades of the classification model might be better suited to discern between classes. In all experiments, the model particularly often focused on a location at the left center, which might be the most discriminating stimulus region.

The final experiment revealed differences in saccade behavior with the only difference to the model being the target class. For the task of classifying threes, the network primarily focused on the left and right edges as well as the top of the digits. As the three has two open sides to the left and two closed curves to the right, these could be the most discriminative regions. Similarly, for the sixes the model mostly focuses on the bottom left which is the location of the distinctive loop of the digit six.

Altogether, the results showed that minimizing predictive uncertainty is a useful mechanism to effectively sample a stimulus, that can account both for bottom-up as well as top-down influences. Further research is required to show that this model can be used for more naturalistic stimuli and in the presence of distracting inputs. In order to learn more about the suitability of the proposed model to predict human fixations, eye-tracking experiments could be performed to compare the saccade patterns. The results can be seen as initial evidence that uncertainty minimization could underly neural models of attention and as guidance for future research to better understand the intricate interaction between different attentional processes.

Acknowledgments

We would like to thank L. Goerke and L. Grossberger for the helpful discussions, creative ideas and general support.

References

Adeli, H., & Zelinsky, G. (2018). Learning to attend in a brain-inspired deep neural network. *arXiv preprint arXiv:1811.09699*.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

LeCun, Y. (1998). The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Frontiers in Computational Neuroscience*, *10*, 56.

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in Neural Information Processing Systems* (pp. 2204–2212).

Moore, T., & Zirnsak, M. (2017). Neural mechanisms of selective visual attention. *Annual Review of Psychology*, *68*, 47–72.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, *7*(3), 6–6.

Soltani, A., & Koch, C. (2010). Visual saliency computations: mechanisms, constraints, and the effect of feedback. *Journal of Neuroscience*, *30*(38), 12831–12843.