DeepGaze III: Using Deep Learning to Probe Interactions Between Scene Content and Scanpath History in Fixation Selection

Matthias Kümmerer[1] (matthias.kuemmerer@bethgelab.org)
Thomas S.A.Wallis[1] (tom.wallis@bethgelab.org)
Matthias Bethge[1,2] (matthias@bethgelab.org)

[1] Institute for Theoretical Physics and Centre for Integrative Neuroscience, University of Tübingen, Germany
[2] Bernstein Center for Computational Neuroscience, Tübingen, Germany

Abstract

Many animals make eye movements to gather relevant visual information about the environment. How fixation locations are selected has been debated for decades in neuroscience and psychology. One hypothesis states that "priority" or "saliency" values are assigned locally to image locations, independent of saccade history, and are only later combined with saccade history and other constraints to select the next fixation location. A second hypothesis is that there are interactions between saccade history and image content that cannot be summarised by a single value. Here we discriminate between these possibilities in a data-driven manner.

Using transfer learning from the VGG deep neural network, we train a model of scanpath prediction "DeepGaze III" on human free-viewing eye scanpath data. DeepGaze III can either be forced to use a single saliency map or can be allowed to learn complex interactions via multiple saliency maps. We find that using multiple saliency maps gives no advantage in scanpath prediction compared to a single saliency map. This suggesest that – at least for free-viewing – no complex interactions between scene content and scanpath history exist and a single saliency map may exist that does not depend on either current or previous gaze locations.

Keywords: saliency; fixations; scanpath; deep learning; transfer learning

Introduction

How humans explore their visual environment has attracted research for many decades. A long-standing theory in the field of gaze prediction poses the existence of an image-dependent saliency map which is combined with task information and scanpath history to decide on the target of the next saccade. Different locations have been proposed for where such a map might be implemented in the brain, including V1 Zhang, Zhaoping, Zhou, and Fang (2012) and Superior Colliculus. Starting with the Feature Integration Theory implemented in the seminal model by Itti, Koch, and Niebur (1998), many models proposed different ideas how such a saliency map might be computed. The last decades have seen great growth in the number and performance of models predicting the spatial fixation distribution, with the current state-of-the-art being our model "DeepGaze II" (Kümmerer, Wallis, Gatys, & Bethge, 2017) according to the influential MIT Saliency Benchmark (saliency.mit.edu).

However, the saliency map hypthesis puts strong constraints on how fixations are selected. Interactions between saccade history and image content that cannot be summarised by a single value are not allowed. For example, if after long saccades different image features drive the next fixation than after short saccades, then it is impossible to assign a single saliency value to image locations.

In order to discriminate between these possibilities, here we move from predicting spatial fixation distributions to predicting sequences of fixations. We do so by extending our previous model DeepGaze II to predict fixation locations depending on where a subject fixated before.

Results

Model

In Figure 1b we show the architecture of DeepGaze III. DeepGaze III first encodes image content and scanpath history into spatial feature maps. The image content is encoded via deep VGG (Simonyan & Zisserman, 2014) features. For the previous scanpath history feature maps are used that encode the euclidean distance as well as the difference in x and y coordinate to the encoded fixation. These feature maps are then processed by a neural network using only 1×1 convolutions. This neural network is split into an purely image-dependent *saliency network* that computes one or multiple saliency maps, a purely scanpath dependent *scanpath network* and a final *fixation selection network* that combines the output of the previous networks. The fixation selection network outputs a single feature map that is subsequently blurred, combined with a center bias and fed through a softmax to yield the final conditional fixation density for the next fixation given the previous fixations. We train DeepGaze III on the MIT1003 dataset (scanpaths from 15 human observers, 1003 images, 3 seconds free-viewing; Judd, Ehinger, Durand, & Torralba, 2009) using maximum-likelihood training via gradient descent and tenfold crossvalidation to avoid overfitting.

Evaluation

We compare DeepGaze III to DeepGaze II and several previous scanpath models (Clarke, Stainer, Tatler, & Hunt, 2017; Adeli, Vitu, & Zelinsky, 2017; Schütt et al., 2017). In Figure 1c we evaluate the performance of DeepGaze III and the other

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0

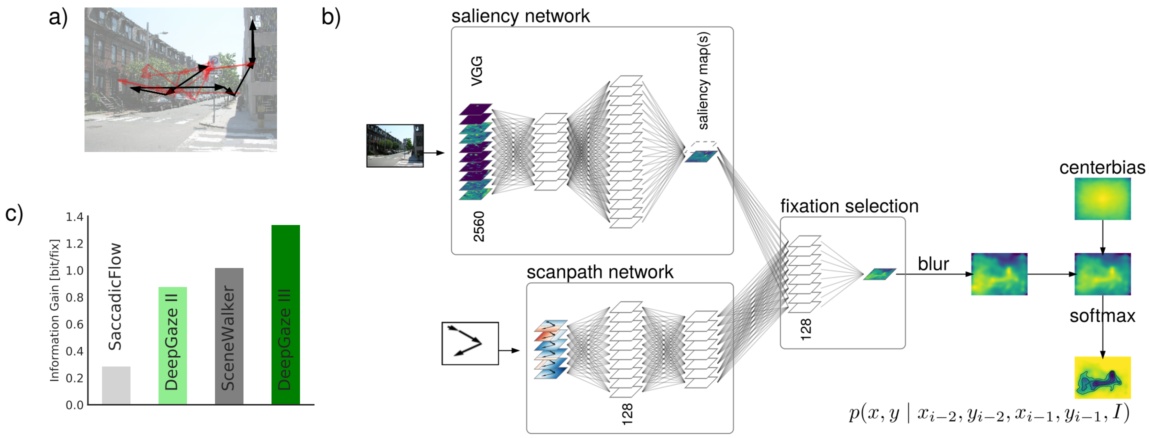

Figure 1: **a)** Humans explore images in scanpaths by making multiple fixations. **b)** Our model computes a saliency map from deep VGG features and then uses this saliency map together with information about the previous scanpath to predict possible locations of the next fixation. **c)** DeepGaze III outperforms the scanpath-independent DeepGaze II model as well as previous models of scanpath prediction.

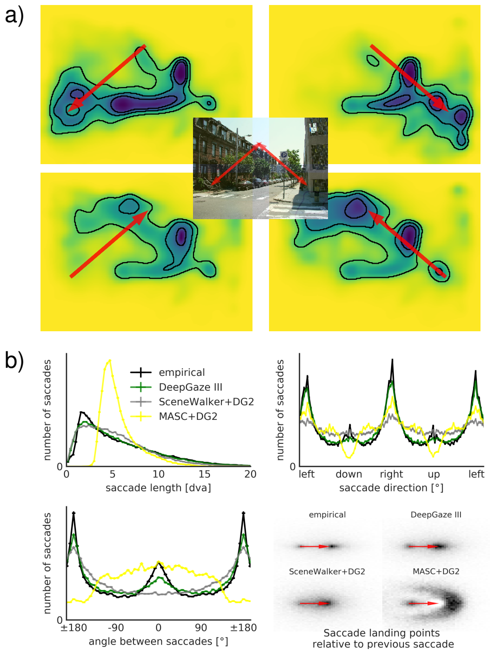

Figure 2: **a)** Model predictions are strongly influenced by the last saccade. For the same image, we show model predictions with different scanpath histories. **b)** DeepGaze III reproduces key properties of human scanpaths: distribution of saccade lengths, the tendency towards horizontal and vertical saccades and saccade angles.

models. DeepGaze III performs substantially better than the other models.

In Figure 2a we show fixation densities as predicted by the model for an example stimulus and different scanpath histories. It can be seen that the model prediction strongly depends on the scanpath history: the model favors locations close to the last fixation.

In Figure 2b we test how well DeepGaze III reproduces key properties of human scanpaths, for example a very specific distribution of saccade lengths and a tendency to favor horizontal saccades to vertical saccades and to favor vertical saccades to diagonal saccades. To this end, we sampled new scanpaths from the model and compared said statistics between the empirical data and the sampled data in Figure 2b. All properties are better reproduced by DeepGaze III than by other models.

Evidence for a Spatiotopic Free-viewing Saliency Map

All results presented above use only one single saliency map as output of the saliency network, as stated by the saliency map hypothesis in the abstract. In order to collect evidence in favor of or against that hypothesis we trained additional versions of DeepGaze III where the saliency network computes multiple saliency maps (Figure 1b, dashed feature map). Figure 3 shows that all models show very similar performance. This rules out more complicated interactions between image content and scanpath history such as the ones exemplified in the introduction and provides some evidence for the existence of a spatiotopic saliency map for free-viewing.

One might argue that our model is limited by the fact that it is not foveated. A retinotopic saliency map could show up in our model as multiple saliency maps, one that is used for the fovea and others that are used for the periphery. How-

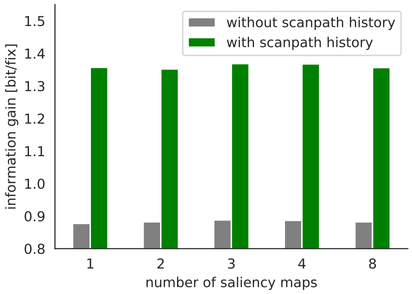

Figure 3: Whether DeepGaze III can use one or multiple saliency maps doesn't affect performance: complex interactions between scanpath history and image content don't seem to play a relevant role in fixation selection, providing some evidence for the existence of "the saliency map" of an image.

ever, since we don't find evidence against the even stronger hypothesis of a spatiotopic saliency map, this doesn't affect our results.

Since we find that the saliency map has strong high-level components (faces, text), we expect that higher brain areas that are sensitive to these objects play an important role in computing this saliency map. The saliency map could be computed downstream from these areas, or these areas could feed back to earlier areas from which the saliency map is read out.

Discussion

The present work applies deep learning to learn a probabilistic model of free-viewing human scanpaths. Using this model, interactions between scene content and recent scanpath history in fixation selection are probed with the result that no such interactions that go beyond a simple pixelwise saliency measure seem to exist.

The recent years have seen increasingly many applications of deep learning in neuroscience (Yamins et al., 2014; Hong, Yamins, Majaj, & DiCarlo, 2016; Jozwik, Kriegeskorte, Storrs, & Mur, 2017). Deep learning models as such are black boxes and it is hard to understand what the models are actually learning. For this reason, their usefulness in neuroscience is often questioned. In some cases this critique might be justified: even more for deep learning models than for classical models it is not enough to just predict the data well. Good prediction performance is merely a necessary requisite for being able to draw scientific conclusions. We want to argue that the present work showcases how deep learning can be applied in a way that tests a well-defined question and gives a clear answer: whether there are (on a functional level) interactions between scene content and scanpath history that cannot be described by a simple pixelwise saliency measure.

In order to answer this question by model comparison, there are two important factors. Firstly, the model that uses a simple pixelwise saliency measure has to be powerful enough to not be penalized simply due to the fact hat it cannot learn a sufficiently good saliency measure. Secondly, the model that uses more complicated interactions has to be able to learn quite general and arbitrary interactions. If the first model is not able to extract a good saliency measure, the second model might perform better simply because it "missuse" parts of its architecture intended for interaction modeling to learn a better saliency measure although there are no interactions. If the second model is too limited, it might just not be able to pick up existing interactions.

The deep learning based model architecture presented here is designed to circumvent exactly those problems. The architecture provides good modeling power in the form of DNNs to most parts of the models and only controls whether the models can use interactions beyond a simple saliency measure.

Acknowledgments

This research has been funded by the BMBF competence center for machine learning (FKZ 01IS18039A), the German Science Foundation (DFG; Collaborative Research Centre 1233) and the German Excellency Initiative (EXC307).

References

Adeli, H., Vitu, F., & Zelinsky, G. J. (2017). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *The Journal of Neuroscience*, *37*(6), 1453–1467.

Clarke, A. D. F., Stainer, M. J., Tatler, B. W., & Hunt, A. R. (2017). The saccadic flow baseline: Accounting for image-independent biases in fixation behavior. *Journal of Vision*, *17*(11), 12–12.

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *Computer vision, 2009 IEEE 12th international conference on* (pp. 2106–2113). IEEE.

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4789–4798).

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Reich, S., Wichmann, F. A., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, *124*(4), 505–524.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556 [cs]*.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zhang, X., Zhaoping, L., Zhou, T., & Fang, F. (2012). Neural activities in v1 create a bottom-up saliency map. *Neuron*, *73*(1), 183–192.