Selecting Maximally-Predictive Deep Features to Explain What Drives Fixations in Free-Viewing

Matthias Kümmerer Thomas S.A. Wallis Matthias Bethge

Recent advances in deep learning have allowed to predict a substantial amount of the explainable information in the spatial fixation distribution in natural images. For example, our model DeepGaze II uses deep features from the VGG deep neural network trained on object recognition as image representation and combines them in a simple pixelwise nonlinear way to predict a fixation density.

However, while these models are very successful at predicting fixations, they are mainly black boxes and therefore not very good at explaining what drives fixations. Here we address this problem by selecting features that are maximally predictive for fixations in a stepwise fashion (Baddeley & Tatler 2006). Starting from a version of DeepGaze II without any VGG features (a pure centerbias model), we first search for the VGG feature that maximally improves model performance when added to this model. Subsequently, we continue the greedy search strategy by looking for the next feature to add that yields the biggest improvement. To reduce computational effort of the feature selection, we use a search strategy that iteratively trains models on subsets of candidate features, selects most promising subsets and splits them into smaller partitions. We call this strategy partition search.

The search results in an ordered list of features where it can be clearly quantified what additional contribution each feature gives to fixation prediction performance while still allowing nonlinear interactions between features via the readout architecture of DeepGaze II. A single VGG feature that is sensitive to very general popout (including objects) already explains more than half of the performance of DeepGaze II. Adding a text sensitive feature, a face sensitive feature and a “geometry” feature yields a model that uses only 4 VGG features and performs at 85% of the full model (see Figures for details).

The deep VGG features we found in this work turned out to have surprisingly intuitive interpretations. We are currently extending the model to use features that are interpretable by design, such as BagNets (Brendel et al. 2019) that allow to determine how localized the features are that drive fixational eye movements.

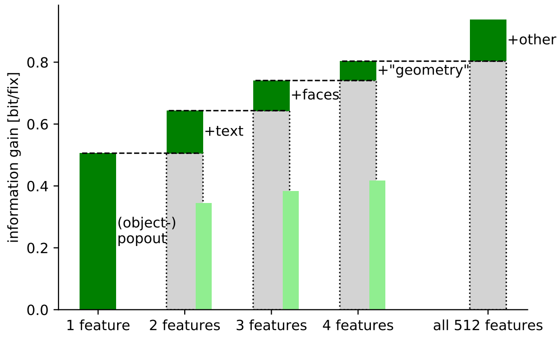

Figure 1: Fixations can be predicted with high precision using a few interpretable VGG features. Each feature, when added to the previous model, improves performance by a certain amount (dark green bar) that can be attributed to the patterns the feature is sensitive to. The light green bars show how well the added feature performs when used without any other features: there is a substantial amount of shared information between the features.

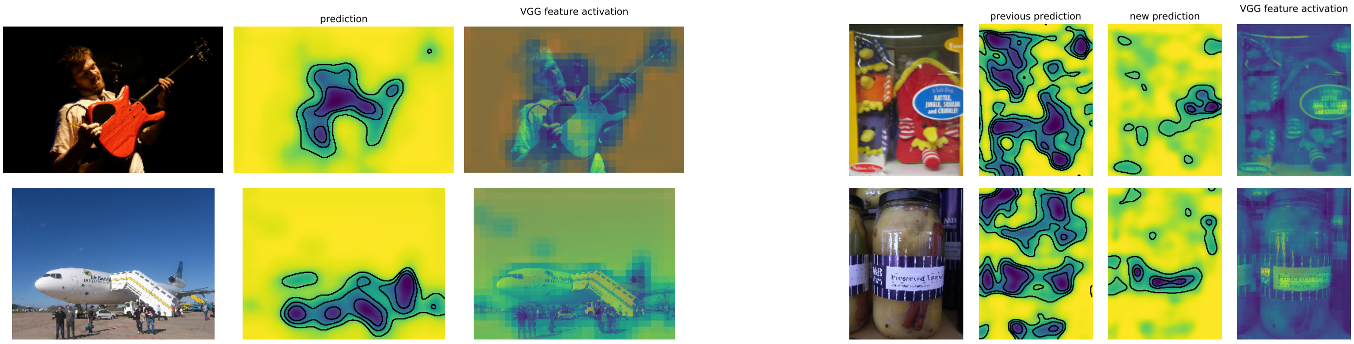

Figure 2: The VGG feature best at predicting fixations is sensitive to a very general form of popout (left) and explains more than half of the full DeepGaze II performance. The next best feature improves performance to nearly 70% of the full model and is sensitive to text (right).