Predicting Fixations From Deep and Low-Level Features

Matthias Kümmerer Thomas S.A. Wallis Leon Gatys Matthias Bethge

Learning what properties of an image are associated with human gaze placement is important both for understanding how biological systems explore the environment and for computer vision applications. Recent advances in deep learning for the first time enable us to explain a significant portion of the information expressed in the spatial fixation structure. Our saliency model *DeepGaze II* uses the VGG network (trained on object recognition in the ImageNet challenge) to convert an image into a high-dimensional feature space which is then readout by a second very simple network to yield a density prediction. DeepGaze II is right now the best performing model for predicting fixations when freeviewing still images (MIT Saliency Benchmark, AUC and sAUC).

By retraining on other datasets, we can explore how the features driving fixations change over different tasks or over presentation time. Additionally, the modular architecture of DeepGaze II allows us to quantify how predictive certain features are for fixations. We demonstrate this by replacing the VGG network with very simple isotropic mean-luminance-contrast features and end up with a network that outperforms all previous saliency models before the models that used pretrained deep networks (including models with high-level features like Judd or eDN; Figure 2). Using DeepGaze and the Mean-Luminance-Contrast model (MLC), we can separate how much low-level and high-level features contribute to fixation selection in different situations (Figure 3).

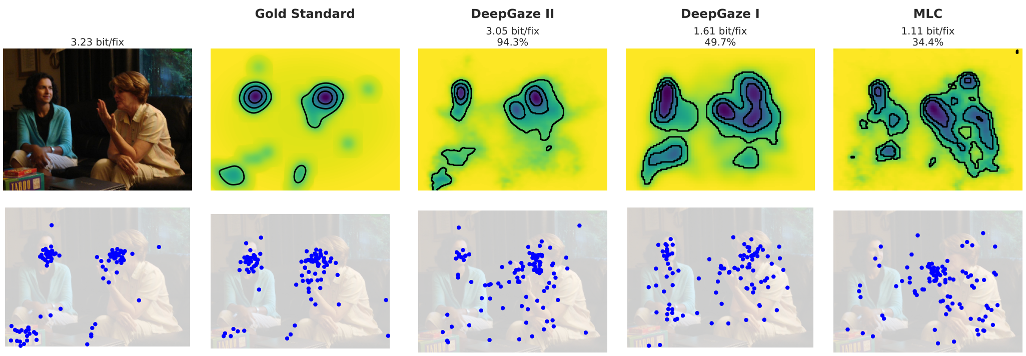

Figure 1: First row: predictions from the Gold Standard, DeepGaze II, DeepGaze I and the Mean-Luminance-Contrast model (MLC) for an example test image. Second row: empirical fixations and fixations sampled from the models.

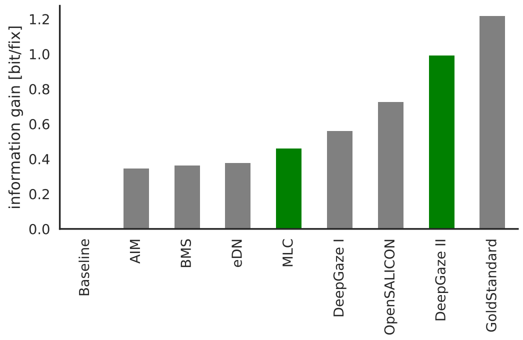

Figure 2: Performances of DeepGaze I and II and the Mean-Luminance-Contrast model (MLC) compared to a range of influential saliency models on the MIT1003 dataset.

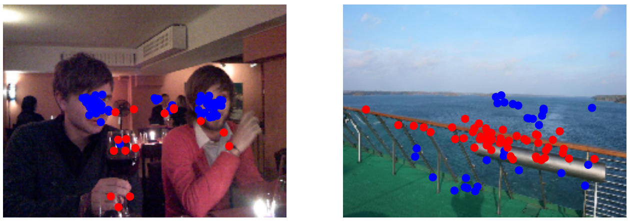

Figure 3: In the left image, the two faces are clearly the most salient spots, whereas the railing in the right image constitutes a simple edge that generates high low-level contrast. Blue dots indicate fixations that are better predicted by DeepGaze II and red dots are fixations better predicted by the MLC model. This illustrates how DeepGaze II is better in predicting fixations driven by high-level features and the MLC model is better in capturing low-level features.