Using Deep Features to Predict Where People Look

Matthias Kümmerer Matthias Bethge

Abstract

When free-viewing scenes, the first few fixations of human observers are driven in part by bottom-up attention. We seek to characterize this process by extracting all information from images that can be used to predict fixation densities (Kuemmerer et al, PNAS, 2015). If we ignore time and observer identity, the average amount of information is slightly larger than 2 bits per image for the MIT 1003 dataset. The minimum amount of information is 0.3 bits and the maximum 5.2 bits. Before the rise of deep neural networks the best models were able to capture 1/3 of this information on average. We developed new saliency algorithms based on high-performing convolutional neural networks such as AlexNet or VGG-19 that have been shown to provide generally useful representations of natural images. Using a transfer learning paradigm we first developed DeepGaze I based on AlexNet that captures 56% of the total information. Subsequently, we developed DeepGaze II based on VGG-19 that captures 88% and is state-of-the-art on the MIT 300 benchmark dataset. I will show best case and worst case examples as well as feature selection methods to visualize which structures in the image are critical for predicting fixation densities.

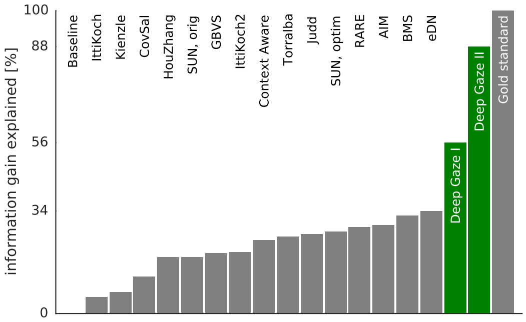

Figure 1: Information gain explained: This plot shows how much of the explainable information in the spatial fixation structure is explained by DeepGaze II compared to DeepGaze I and the models evaluated in Kümmerer et al.: Information-theoretic model comparison unifies saliency metrics, PNAS 2015. For more details on the information gain metric, see also this paper.

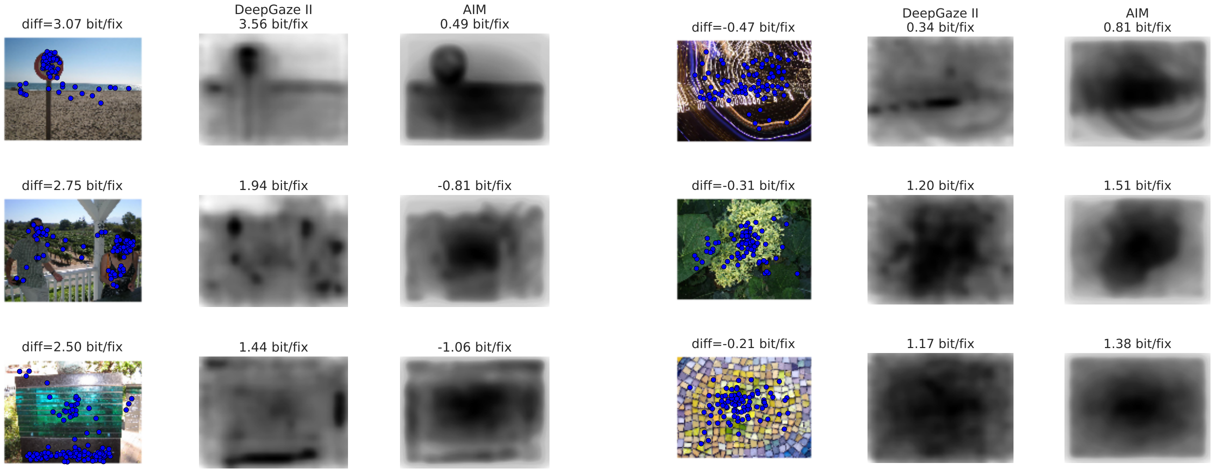

Figure 2: Examples from the MIT1003 dataset. Left block: three images with largest difference to AIM (Bruce and Tsotos, 2007) in explained information. Right block: Three images with worst difference to AIM. In each block, the left column shows the image, the middle column shows the log density from DeepGaze II, the right column shows the log density from AIM (see Kümmerer et. al 2015). The numbers above the log densities indicate how much additional information (in bit/fix) the model explains compared to the center bias baseline.