Evaluating Models of Scanpath Prediction

Matthias Kümmerer Matthias Bethge

Humans visually explore the world using eye movements to point the high-resolution fovea towards whatever we consider relevant at that time. The last decades have seen a substantial efforts in formulating computational models of spatial fixation prediction and benchmarking such models. While most of the modeling efforts have been focused on spatial fixation prediction, there also has been significant work on predicting scanpaths of fixations. However, there are even more metrics for evaluating scanpath models than there are classic saliency metrics. Most of those metrics, e.g., String Edit, ScanMatch and MultiMatch, aim at comparing ground truth scanpaths to sampled scanpaths.

Here we show that such metrics result in wrong models scoring systematically better than the ground truth scanpath distribution. We take a different approach to evaluating scanpath models and bring scanpath model evaluation back to realm of spatial fixation prediction. For probabilistic models, this can be done by splitting the joint log likelihood of a scanpath $\log p((x_0,\ y_0),\ (x_1,\ y_1),\ \dots (x_N,\ y_N))$ via the chain rule into a sum of conditional log likelihoods for each fixation $\sum_i^N \log p((x_i,\ y_i)\ |\ (x_0,\ y_0),\ \dots, (x_{i-1},\ y_{i-1}))$. Compared to the evaluation of models of spatial fixation prediction, this means applying metrics to conditional, scanpath history dependent spatial fixation predictions instead of unconditional spatial fixation predictions .

To allow comparison to non-probabilistic models, we generalize this approach: internally, non-probabilistic models usually assign priority values to potential fixation locations, to, e.g., select the location of the maximum. In analogy to the conditional fixation distribution we interpret those priority values as a "conditional saliency map". This saliency map can be evaluated using classic, well-understood saliency metrics like AUC and NSS.

We apply this evaluation to a range of existing models and compute average performances per fixation on the free-viewing datasets MIT1003 (Judd et al., 2009) and CAT2000 (Borji & Itti, 2015) with the AUC and NSS metric and, for probabilistic models, log-likelihood and information gain. This allows us to quantify in a concise way how well those models capture the distribution of human scanpaths. Because we can compute prediction performance for individual fixations, we can pinpoint very precisely where and why models loose performance. Interestingly, one of the oldest models (Constrained Levy Exploration, Boccignone & Ferraro 2004) turns out one of the best performing models with respect to AUC, demonstrating the importance of principled model evaluation.

	LL [bit/fix]	IG [bit/fix]	AUC ↑	NSS
Itti&Koch (with WTA network)			0.4730	0.2706
G-Eymol			0.6620	0.5534
STAR-FC			0.6623	0.5813
MASC			0.7344	1.1252
IOR-ROI-LSTM	-46.8214	-47.7270	0.7437	0.4573
IRL			0.7482	1.1126
SaltiNet	0.7199	-0.1857	0.7905	1.1384
Center Bias	0.9057	0.0000	0.8005	1.2627
Saccadic Flow	1.1697	0.2641	0.8432	1.6029
LeMeur16	0.9314	0.0258	0.8654	2.4946
DeepGaze IIE	2.0607	1.1551	0.8959	2.7774
SceneWalk	2.1410	1.2354	0.8994	2.9145
CLE	1.8837	0.9780	0.9050	1.5216
DeepGaze III	2.4369	1.5313	0.9154	3.2450

Table 1: Performances on the MIT1003 (Judd et al. 2009) dataset. LL=log likelihood (relative to a uniform baseline model); IG: information gain (Kümmerer et al. 2015); AUC: area under the curve; NSS: normalized scanpath salience. Evaluated models are: Itti & Koch (1998), Constrained Levy Exploration (Boccignone & Ferraro 2004), SceneWalk (Engbert et al. 2015), LeMeur16 (LeMeur & Coutrot 2016), Saccadic Flow (Clarke et al. 2017), MASC (Adeli at al. 2017), IRL (Xia et al. 2017), SaltiNet (Assens et al. 2017), STAR-FC (Wloka et al. 2018), G-Eymol (Zanca et al. 2019), IOR-ROI-LSTM (Sun et al. 2019), and our own models DeepGaze IIE (Linardos et al. 2021) and DeepGaze III (Kümmerer et al. 2022). MASC, LeMeur16, SceneWalk and CLE require spatial saliency predictions, for which we use DeepGaze IIE.