Visual Attention: From Classical to Modern Approaches

Steffen Schneider
Technical University of Munich
steffen.schneider@tum.de

Abstract

Predicting the ways in which image locations draw the attention of humans gives important insights into the visual system and the way in which humans access image contents. The notion of saliency is a popular research item in both neuroscience and lately, also in deep learning research. As attention based gating is incorporated in signal processes, predicting saliency to identify important parts of the image is also interesting from a techical perspective. In this paper, the implementation of a variant of the Itty Koch model [4] is discussed as an example for a historical approach to saliency. The approach will be compared to a simple data-driven adaptation approach as well as the state-of-the art model for visual saliency, DeepGazeII [6] according to the MIT300 benchmark. The models will be evaluated on a custom dataset of four photographs, three artificial visual stimuli as well as three video files.[1][2]

1 Introduction

Saliency models usually process an image or a sequency of images and produce a saliency map.

2 Methods

In this section, we discuss different methods to compute saliency maps.

To provide a unique mathematical framework, I will consider the architectural similarities of classic approaches such as the Itty Koch model [**?**] and modern appraoches such as Deep Gaze [].

Overall, when given an image $x \in \mathbb{R}^{W \times H \times C}$, many saliency models first use a feature extractor $\mathcal{F} : \mathbb{R}^{W \times H \times C} \mapsto \mathbb{R}^{W \times H \times K}$ to convert an image with C channels into K feature maps. Afterwards, we use an affine transformation $W : \mathbb{R}^K \times \mathbb{R}^K \mathbb{R}^K \mapsto \mathbb{R}$ applied pointwise to combine the features, followed by a non-linearity $\phi : \mathbb{R} \mapsto \mathbb{R}$ that computes the final saliency map.

In the following, we discuss choices for the feature extractor $\mathcal{F}$, the transformation W and the non-linearity ϕ.

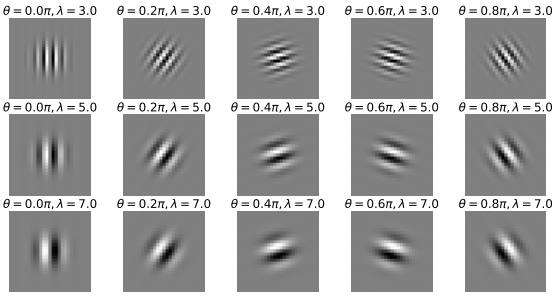

Figure 1: Gabor wavelets constructed with $\psi = \pi/2, \sigma = 3, \gamma = 1$, variable angle θ and wavelength λ.

2.1 Feature Extraction

Goal of the feature extraction mechanism is the computation of K features maps from the C input channels of the image.

2.1.1 Baseline Method

As a baseline, we use a simplified version of the model propsed by [**?**]. The model consists of a feature extraction pipeline as well as a mechanisms for feature weighting and computation of the final saliency map.

Intensity Bright spots within the image more likely trigger a saliency response. Therefore, as one score, we compute the channel mean $\sum_c x^{(c)}$ as one feature map.

Edge filters Gabor wavelets [**?**] are widely used in image processing and feature extraction. They are also present in the first layers of deep neural networks [], underlining their importance in early visual processing.

Gabor wavelets are parametrized as a composition of a Gaussian and Sinusoidal kernel. A real valued gabor kernel is given as

[1] The implementation in Python is available at https://github.com/stes/saliency

[2] This report is part of a lab of the Neuroengineering program at TU Munich.

$$g(x, y|\theta, \psi, \lambda, \sigma, \gamma) = \exp\left(\frac{x'^2 + (\gamma y')^2}{-2\sigma^2}\right) \cos\left(\frac{2\pi}{\lambda} + \psi\right),$$

with

$$x' = x\cos\theta + y\sin\theta$$
$$y' = -x\sin\theta + y\cos\theta.$$

The parametrization includes the angle (filter rotation) θ, phase offset ψ, wavelength λ, standard deviation σ and skew γ.

Gabor filters with different settings for these parameters are depicted in figure 1.

Color opponent xxx

Color bias Humans seem to be attracted to a certain set of colors, which is why a color bias is implemented.

Color distance can be defined in different ways.

2.1.2 Deep Neural Networks

As deep learning becomes increasingly popular in image recognition [7], it seems natural to adapt it to the task of saliency estimation. In this report, I will focus on the work and models proposed by the Bethge lab [6, **?**, 5], that are currently highest ranking on the MIT300 benchmark [2]. Deep Gaze I [5] and Deep Gaze II [6] make use of the feature maps computed by a deep neural network, in this case the VGG network [8].

In general, the overall approach of using a deep neural network is not fundamentally different from the previously discussed approaches. Considering the framework presented above, a VGG model takes the role of $\mathcal{F}$, extracting feature maps based on the image, with a fixed set of parameters estimated by training on ImageNet data [8].

A readout network is then used to estimate a saliency map from the network outputs [6], taking over the role of W and ϕ.

2.1.3 Temporal Approaches

Background Extraction For videos with a static background, a simple background extraction scheme can be used by computing the foreground components as the difference between each image and the temporal average. Smoothing is applied to yield the final saliency map S_B:

$$S_B^t = G_\sigma\left(x^t - \frac{1}{N}\sum_{\tau=1}^{N} x^\tau\right) \quad (1)$$

Optical Flow Optical flow algorithms estimate a vector field $(v_{i,j})_{i=1\ldots W, j=1\ldots H} \in \mathbb{R}^{W\times H\times 2}$ from a sequence of input images, where each $v_{i,j}$ denotes the direction in which the pixel (i, j) is expected to move in the next timestep. Optical flow can be computed using the approaches by Lucas and Kanade or more recently, using convolutional neural networks. The computed optical flow map could be used as an input to the computation of the saliency maps. In this work however, a simpler approach will be taken: Given two images x^t and x^{t+1}, an approximation to the temporal derivative will be used as a proxy for $\|v\|$:

$$\|v\| \approx G_\sigma(x^{t+1}) - G_\sigma(x^t), \quad (2)$$

where a gaussian smoothing function G_σ was applied to alleviate high frequency noise between the images.

2.2 Saliency Responses

So far, we discussed different ways to extract feature maps of a given image. In this section, an approach to combine feature extraction mechanisms into a saliency maps will be dicussed.

2.3 Metrics

As evaluation metrics, the commonly used mean-squared error (MSE) as well as the NSS score will be used, common for visual saliency [3, 2, 1].

Median squared error Given the human gaze fixation data median, $\mu' = (i, j)$ in pixels coordinates, and the maximum salient response $\hat{\mu}$ from the model we compute the total error of all frames N as:

$$\mathrm{MSE}(\hat{\mu}, \mu') = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_i - \mu'_i)^2$$

NSS Given a binary map of fixation locations F (human data) and the saliency map S (response from the model), the NSS measure is computed as:

$$\mathrm{NSS}(S, F) = \frac{1}{N}\sum_{i=1}^{N}\overline{S}_i \odot F_i$$

Where $N = \sum_i F_i$ is the total number of fixated pixels and $\overline{S} = \frac{S-\mu(S)}{\sigma(S)}$ is the normalized saliency map. A value of 0 means means that it is chance (i.e., random), positive values shows correspondence and negative values anti-correspondence.

3 Experiments

In this section, results for applying the previously introduced models are provided for both static images and video data. For the exact parameters, please consider the provided reference implementation along with the supplementary material.

3.1 Itty Koch Model

Results for the Itty Koch Model applied to a collection of photographs as well as artificial stimuli is provided in figure 2. The exact reaction to the artificial stimuli (rightmost images in the figure) is largely dependent on how the top-down modulation weights, i.e., the function W, is implemented. For instance, for the fifth image depecting both highlighted and rotated instances of a "5", this directly influences whether the saliency response at the red five or the rotated 5 is more prominent.

3.2 Sequential Fixation with static images

In this section, the implementation of a sequential attention mechanism is demonstrated.

A simple strategy which will be implemented here is using the maximally attended spot within the current saliency map $S^{(t)}$ to direct the gaze, i.e.,

$$f^{(t+1)} = \arg\max_{i,j} S^{(t)}, \tag{3}$$

followed by

$$S^{(t+1)} = (1 - G_\sigma(\delta_{i,j})) \odot S^{(t)}, \tag{4}$$

where $\delta_{i,j}$ is an image with pixel (i, j) being the only non-zero pixel set to 1. Results for sequential fixation are depicted in Figure 3 and for more images in the appendix, Figure 4.

3.3 Sequential Trajectory of gaze fixation on a video

The DeepGazeII and ICF models from [6] were used on all videos for an upper baseline [3]. For a lower baseline, two models using temporal differences or background extraction were used. For the human baseline, an 18-fold cross validation was run by using the trajectory of one subject as the "prediction" of a model and tested against the remaining 17 subject's gaze trajectories. Mean and standard deviation of this human baseline is also reported.

A graphical comparison is given in Figure **??**, the exact values are given in Table **??**.

[3] For this work, the tensorflow implementation provided at `deepgaze.bethgelab.org` was integrated into the experimental framework

4 Discussion

The classic, but slightly simplified Itty Koch model only relying on hand-crafted features could already explain a range of test stimuli, such as color bias and selection of high frequency information in the image using gabor wavelets.

Replacing the hand-crafted feature extraction mechanisms with features learned for object classiciation, further improvement is possible. On temporal data, using the difference between saliency responses of different models to detect changes in saliency was used in this report.

Improvements are likely if the used models are explicetly designed to take into account temporal information (e.g., by the use of temporal filters or optical flow estimation).

5 Supplementary Material

The implementation of the models along with supplementary figures is available at `github.com/stes/saliency`. Implementation was performed in Python, using Tensorflow for running the Deep Gaze II and ICF [6] models. The Itty Koch model was not fully implemented according to the original work [4], however the overall model structure is consistent.

References

[1] Ali Borji and Laurent Itti. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. *CVPR 2015 workshop on "Future of Datasets"*, 2015.

[2] Zoya Bylinskii, Tilke Judd, Ali Borji, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT Saliency Benchmark. http://saliency.mit.edu/, 2015.

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016.

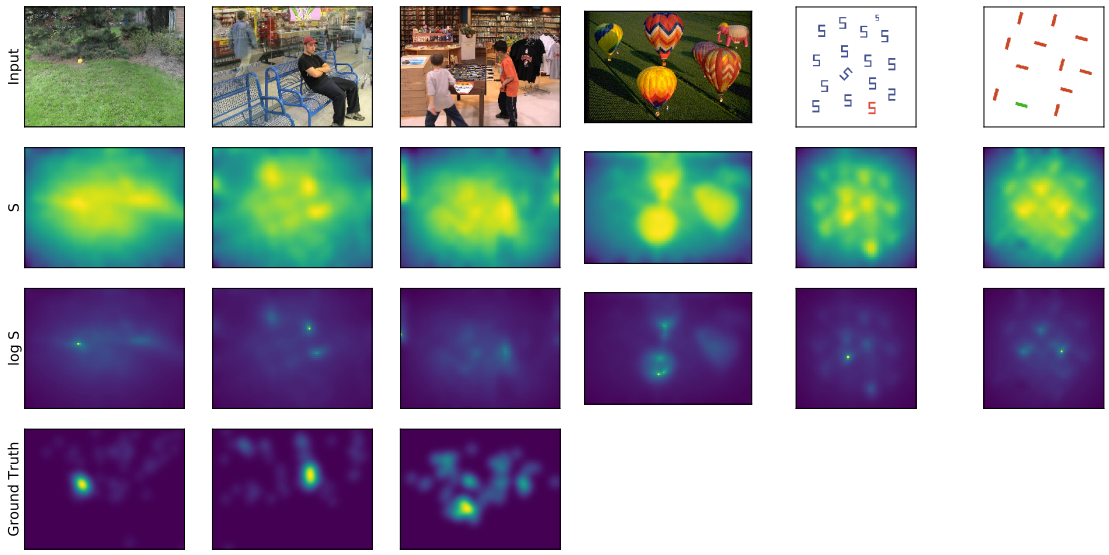

Figure 2: Results for the Itty Koch [**?**] model. As a simple strategy to get more focussed saliency map, we propose to compute the negative logarithm of the negative saliency output S, resulting in sharp peaks in the saliency map. With this representation, corrospondance to the ground truth becomes clear even when the saliency output is blurred substantially by the model.

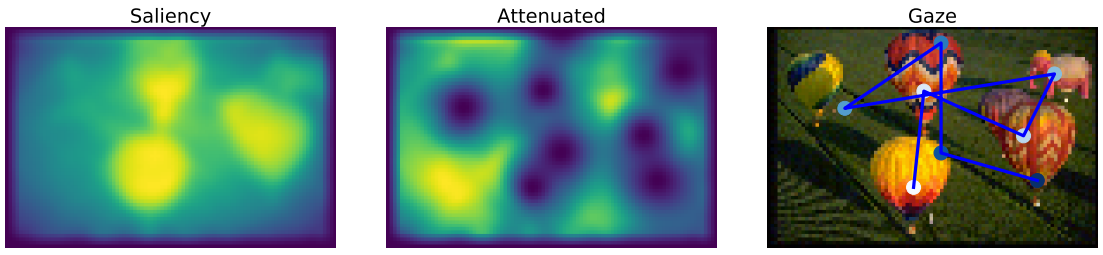

Figure 3: Sequential fixation using sequential attention. Beginning from the saliency map, the point of maximum saliency is sampled. Afterwards, a region around this point is attenuated before the next point is sampled. A full list of prediction for different images can be found in the appendix figure 4.

[4] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert hifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[5] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. 2014.

[6] Matthias Kümmerer, Thomas S A Wallis, Leon A Gatys, and Matthias Bethge. Understanding Low-and High-Level Contributions to Fixation Prediction. pages 4799–4808, 2017.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. jan 2015.

[8] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014.

A Sequential Fixation

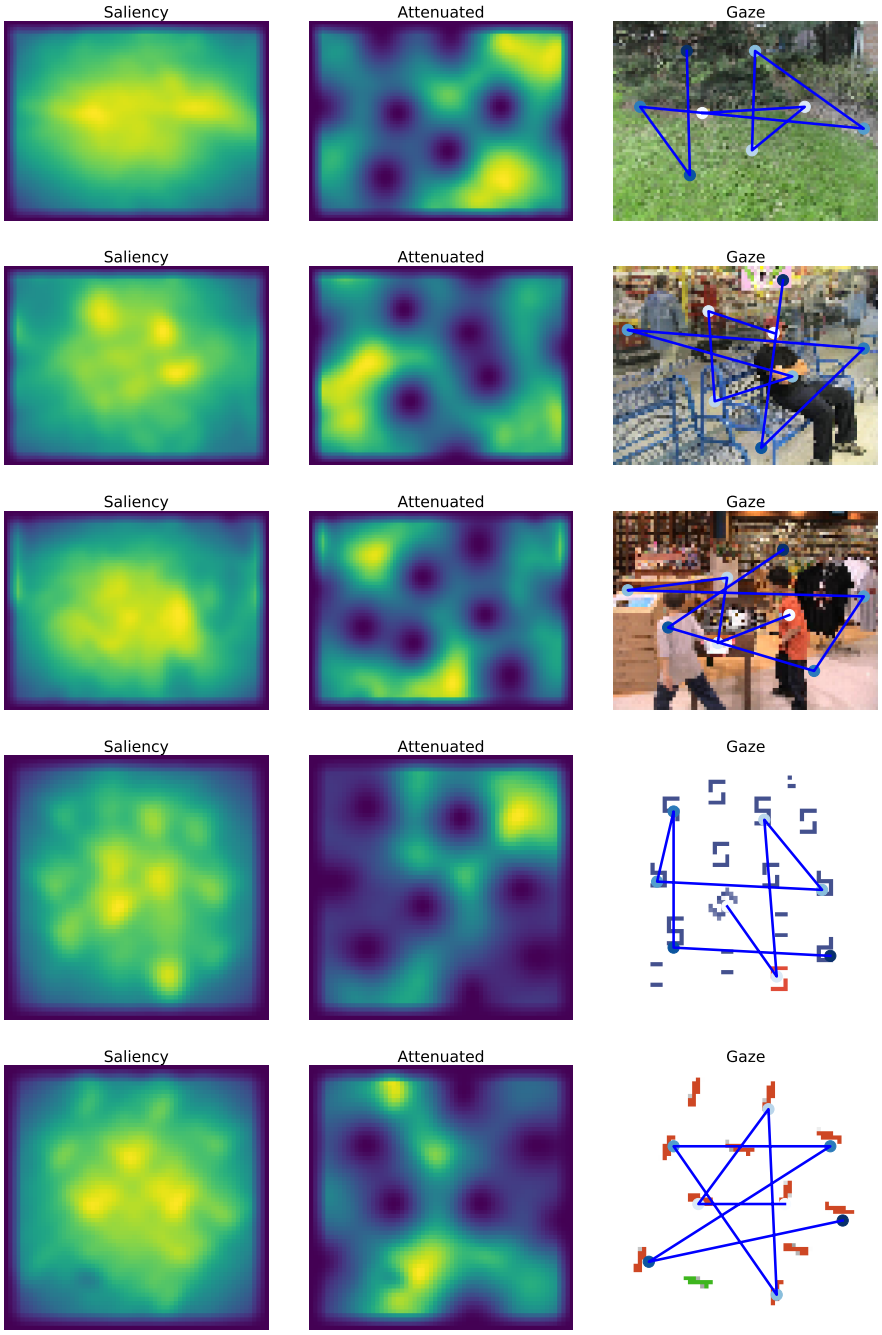

Figure 4: Sequential fixation for all images, including test stimuli. The fixation sequence is color-coded from first fixation (White) to last fixation (Blue).