# Computational modeling of visual attention and saliency in the Smart Playroom

**Andrew Jones**
Department of Computer Science, Brown University

### Abstract

The two canonical modes of human visual attention bottom-up and top-down have been well-studied, and each has been demonstrated to be active in different contexts. Much work has investigated the computational underpinnings of visual attention, and several models have been developed to explain and predict where humans look in images. Recently, convolutional neural networks have been shown to learn useful visual features efficiently from large numbers of images, and they have demonstrated recent success in predicting visual saliency. In this work, we computationally model bottom-up and top-down attention independently in children and show how each mode contributes to childrens attention in a visual search task.

## Introduction

Historically, two complementary forms of visual attention have been identified and studied: a fast, scene-based form, and a slower, task-modulated, and deliberate form (Treisman & Gelade,1980). These are often referred to as bottom-up and top-down, respectively, and this is how the terms will be used in this report.

Several methods have been proposed to investigate the mechanisms of attention, from examining the neural underpinnings to observing the higher-level manifestation of attention. One popular method to indirectly measure visual attention is tracking eye fixation locations while subjects view static images or videos or interact with the real world. This experimental paradigm can elucidate which scene parts people find most interesting, the relative importance of color, texture, and other object properties in inducing fixations, and the temporal dynamics of attention.

In both computer vision and neuroscience, predicting the locations of humans eye fixations in an image is an important problem. This task has been termed saliency prediction, and several computational models have been proposed that predict where humans will look in an image (see (Itti & Koch,2001) for a review). Prior to the advent of deep learning and the availability of large computing power, most models made predictions based on low-level image statistics. The first such model proposed was that of Itti and Koch(Itti & Koch,2000), which integrated saliency maps for color, intensity, and orientation of an images contents.

Within the last few years, convolutional neural networks (CNNs) have proven their capacity to automatically learn meaningful features that are useful for image classification, object detection, and segmentation, among other tasks. These features have also been successfully leveraged for saliency prediction. Recently, a saliency prediction architecture called DeepGaze II was proposed (Kümmerer, Wallis, & Bethge,2016), which combines a CNNs features (VGG19) from several layers to read out saliency maps with a fully convolutional network. We refer to DeepGaze simply as DeepGaze in this report, but it should not be confused with DeepGaze IIs precursor, DeepGaze I. [4] trained and tested DeepGaze using the MIT300 Saliency Benchmark (Bylinskii et al.,2015) dataset in which the training examples are static images, and the training labels are averaged, smoothed fixation maps across several human participants. This model showed improved performance over all previous saliency prediction models in a majority of metrics, and it is currently the leader of the benchmarks leaderboard.

In this work, we present a method to computationally model two contributing forms of saliency and visual attention: a bottom-up, scene-drive contribution, and a top-down, task-driven contribution. Using deep architectures, we train a separate predictive model for each form of saliency, and we find the optimal weighting between the two to explain true fixations. We show that influence of each attentional model varies across participants and contexts. We first review the methods used, and then present results.

## Methods

### Smart Playroom

For all training and testing, we use egocentric video data collected in a developmental psychology lab, which carried out behavioral experiments in children ages 4-6. As part of a suite of testing in an experimental room termed the smart playroom, video data was collected in which the camera is head-mounted on a child participant. This playroom is designed to allow the study of childrens behavior in naturalistic environment, and it includes many toys and objects. Data was collected from 22 participants.

Importantly, in addition to the head-mounted video camera, every child participant also wears a head-mounted eye-tracker, which estimates the childs gaze location on each frame. This eye-tracker/video headset is mobile and allows the child to move about the room freely.

These videos allow for easy collection of many images within each participants session. This large number of images is helpful in this work, as training CNNs with many parameters is known to require many training samples to converge correctly.

During a given childs session, multiple experimental tasks are performed. For the purposes of this study, we focus on two of these: the free play session and the visual search session.

In the free play session, the child is allowed to explore the room freely and play with toys. No other people are present during this task, and the child is not given any specific task. As such, the model trained on data from the free play session forms our representation of bottom-up attention.

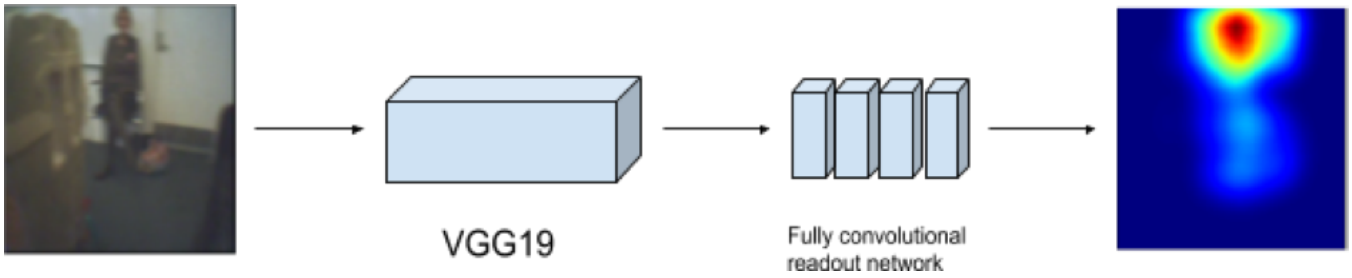


Figure 1: Overview of DeepGaze II architecture. An images features are extracted from several of VGG19s layers and are fed into a fully convolutional readout network that produces a saliency map.

The more structured visual search session is composed of several subtasks. In each subtask, the experimenter first shows a paper placard with a picture of a toy to the child, who is then instructed to retrieve the toy from the room and bring it back to the experimenter.

### Data preparation

For both the free play and visual search sessions, we trim the videos to the start and end times, which were manually annotated. We remove any frames in which the eye-trackers fixation measurement is not within the frame.

To account for the eye-trackers measurement error, we use smooth fixation maps to represent the participants gaze location on each frame. Specifically, for each frame, we center a 2-dimensional Gaussian on the fixation location, and the Gaussians variance was chosen to represent the eye-trackers listed error.

### Saliency models

We use DeepGaze to model bottom-up attention and an object detector to model top-down attention.

**DeepGaze** To model bottom-up attention, we focus on building a model that can predict saliency for images in the free play session. Specifically, we train DeepGaze (Figure 1) using frames from all subjects. In DeepGaze, features from several layers of VGG19 (conv5_1, relu5_1, relu5_2, conv5_3, and relu5_4) are concatenated and fed into a fully convolutional readout network. The readout network consists of four layers, each with 1x1 convolutions. A Gaussian blur is applied to the final layers output, after which a softmax is applied. The final output is a 112x112 normalized saliency map, where each value represents a saliency prediction value for the corresponding location in the input image. We implement the DeepGaze architecture in TensorFlow.

We train DeepGaze on frames from the free play session videos. We hypothesize that, becauase there is no task in the free play session, the child is primarily employing bottom-up attentional behavior. In each DeepGaze training session, we train the model for at least 20 epochs, and until the last five epochs show decreased performance.

To validate the DeepGaze model in the playroom setting, we first perform 10-fold cross validation across subjects with a 50/50 train/test split. We then train a model on all subjects for all dual-model (DeepGaze and object detector) tests.

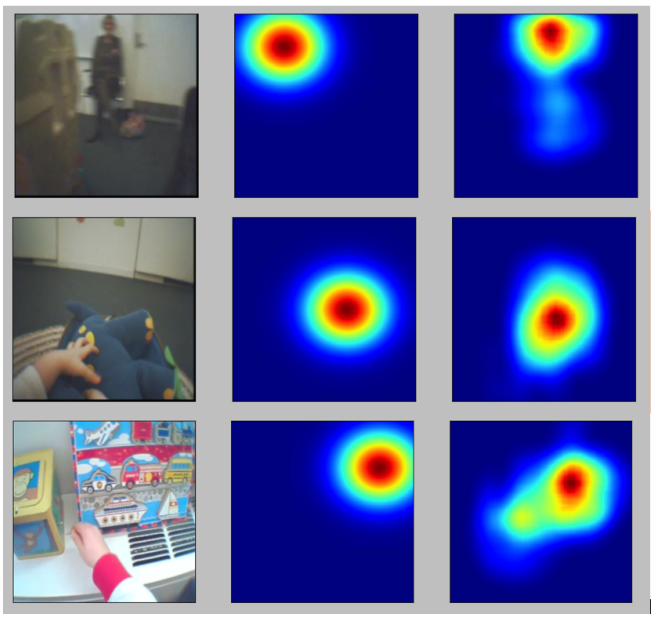

Figure 2: Example output of DeepGaze that has been trained on free play videos. Columns from left to right: initial video frame, ground truth fixation, and estimated saliency map.

**Object detector** To model top-down attention, we train an object detector to identify the toys placed in the room in the visual search session. Specifically, we train the Faster R-CNN6 (Ren, He, Girshick, & Sun,2015) architecture for multiclass detection. Object bounding box predictions were generated for each frame of the visual search videos. For each visual search subtask, bounding box predictions were only generated for the toy that was currently being searched for.

**Combined top-down and bottom-up saliency** To model the relative contribution of the bottom-up (DeepGaze) and top-down (object detector) pathways on fixations, we sought to find a linear combination of the two that could explain the true fixation maps. We fit a linear model, $\alpha * S_{BU} + \beta * S_{TD} = F$, where for each frame, $S_{BU}$ is DeepGazes estimate of a bottom-up saliency map, $S_{TD}$ is the object detectors top-down

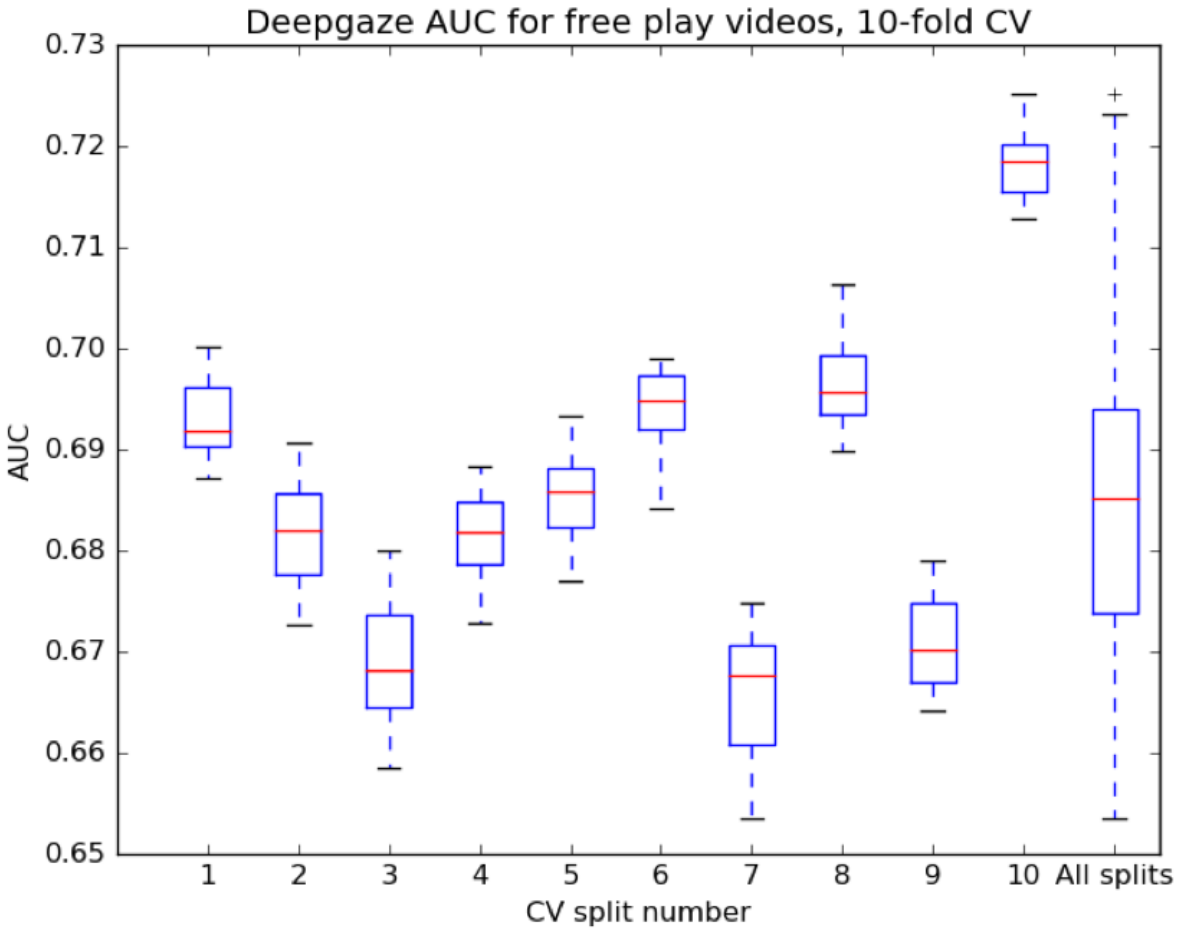


Figure 3: DeepGaze prediction AUC values for 10-fold cross-validation split across subjects in the free play session videos.

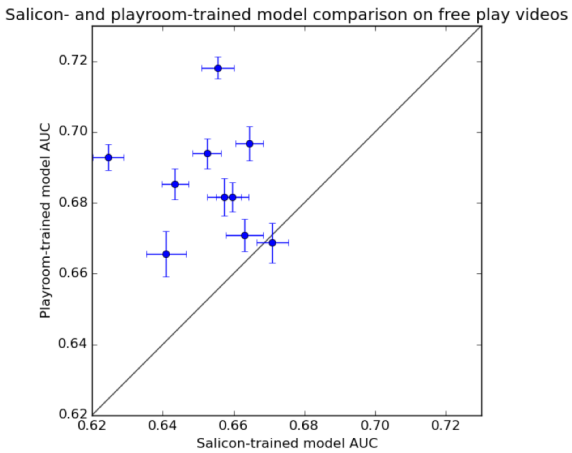


Figure 4: Comparison of prediction results from a DeepGaze model trained on smart playroom videos and a DeepGaze model trained on SALICON data.

estimate, and $F$ is the ground truth fixation map. We fit a separate model for each participant using least squares regression, obtaining weighting coefficients for all subjects.

## Results

### Baseline results for DeepGaze on free play videos

To validate DeepGaze for saliency prediction in the smart playroom video data, we performed 10-fold cross-validation across subjects, with a 50/50 train/test split in each fold using free play videos. The results, presented in Figures 3 and 4, demonstrate DeepGazes ability to predict saliency in playroom videos. As a comparison, we also trained DeepGaze on the SALICON dataset (Jiang, Huang, Duan, & Zhao,2015), which contains images and fixation maps from a free-viewing task. As demonstrated in Figure 2, predictions from a DeepGaze model trained on playroom videos outperforms DeepGaze trained on the SALICON dataset when tested on a hold-out playroom dataset.

Having validated DeepGazes ability to predict fixations in

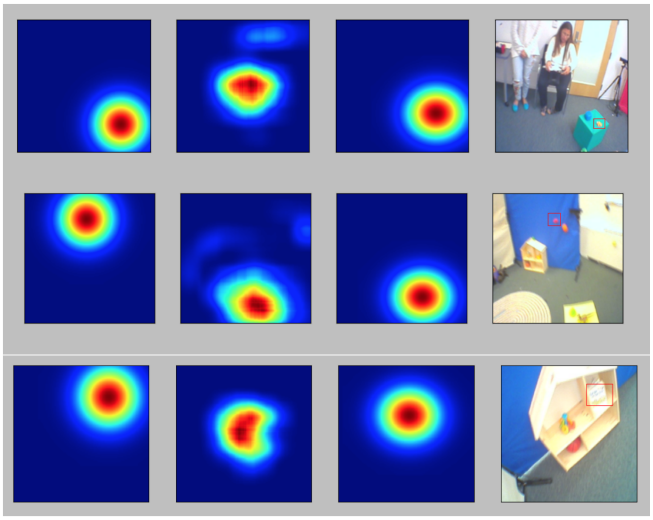

Figure 5: From left to right: Object detector annotation; DeepGaze saliency prediction; ground truth fixation; egocentric image (with bounding box on predicted object).

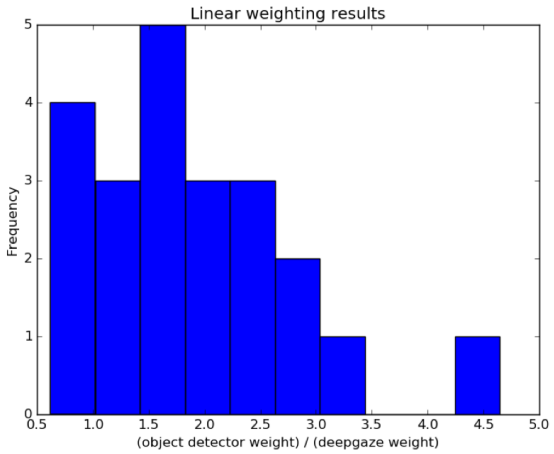


Figure 6: Histogram of results for finding optimal weight between bottom-up and top-down models. The horizontal axis is the ratio of the top-down models weight to the bottom-up models weight.

playroom videos, we next sought to combine DeepGaze and object detector predictions to try to explain ground truth fixations. Figure 5 contains example predictions for both the top-down and bottom-up models.

### Bottom-up and top-down attention weights

After computing the optimal linear weighting between the bottom-up and top-down models, we find that the subjects show a fairly wide range of weightings (Figure 6).

We next examined how these weightings vary across certain demographic conditions. We performed correlational analyses for age and socioeconomic status and performed hypothesis tests for development status (Autism Spectrum Disorder vs. typically developing children) and sex. The results are presented in Tables 1 and 2. Although none of the tests reach statistical significance, we propose that this paradigm could be useful to test more conditions and contexts in the future.

Table 1: Spearman correlations between weight ratios (top down / bottom up) and demographics. SES: socioeconomic status, as measured by income-to-needs ratio

| Variable | Spearman's rho | p-value |
|---|---|---|
| Age | -0.023 | 0.92 |
| SES | 0.174 | 0.439 |

Table 2: Hypothesis tests between weight ratios (top-down / bottom-up) and categorical demographics. ASD: Autism Spectrum Disorder; TD: Typically Developing.

| Variable | t-statistic | p-value |
|---|---|---|
| Development status (ASD vs. TD) | 1.62 | 0.155 |
| Sex | -1.11 | 0.299 |

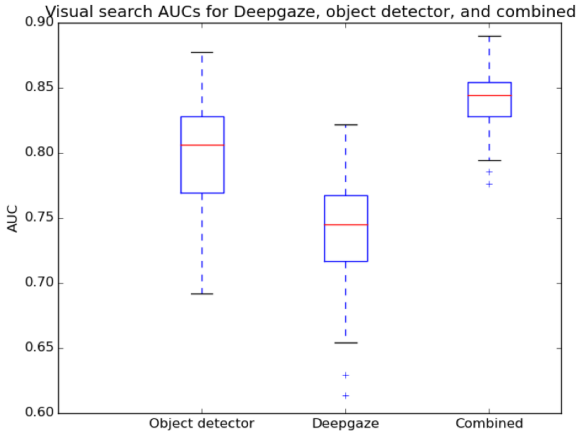


Figure 7: Saliency prediction results for the top-down (object detector), bottom-up (DeepGaze), and combined models.

## Conclusion

We present a framework for modeling top-down and bottom-up attention using saliency prediction. Our computationally-driven approach is particularly useful in the context of egocentric video, in which data can be easily collected and experiments with and without tasks can be carried out. We validate that deep features are useful for predicting eye fixations in egocentric video in children. In particular, our models can predict saliency with improved performance over models

trained on SALICON, in which fixations are measured in a passive fixation task. A limitation of our training procedure is the technical error associated with the mobile eye-tracker, which we account for in our work, but this could inhibit the precision of our measurements. In addition to saliency prediction, our framework also allows for computing the relative contribution of the bottom-up and top-down attention models to the participants' fixations in each frame. After doing so in the playroom egocentric video dataset, we observed a wide range of behavior, as measured by the weight ratios, across participants. Although our tests for relationships between saliency model weight ratios and demographics were not significant, our results using DeepGaze for saliency prediction show promise for future work using deep features to model saliency in egocentric video. In the future, this modeling paradigm could be applied in several new settings, including in adults, in participants with various developmental and psychological disorders, and in more complex tasks. In addition, we discuss one possible future improvements below.

### Temporal DeepGaze

While our saliency prediction framework operates on a frame-wise basis, our egocentric video dataset could be used to model and assess saliency across time. The dynamics of saccadic eye movements over time has been extensively studied in the neuroscience literature; for one example in the context of bottom-up and top-down attention, see (Schill, Umkehrer, Beinlich, Krieger, & Zetzsche,2001). To extend our framework, when predicting the saliency map for a given video frame, we could not only use DeepGaze features for the current frame, but also features for previous frames. One specific approach would be to train an LSTM network on DeepGaze features for sequences of $n$ frames, where $n$ could be optimized from the data. We hope to explore this approach in the future and compare its performance to our current frame-wise prediction framework.

## Acknowledgments

I am thankful for the generous support and guidance of my advisors, Professors Thomas Serre and James Tompkin throughout this project. I extend my gratitude to all members of the Serre lab for their collaboration, especially Drew Linsley, Pankaj Gupta, Vijay Veerabadran, Tarun Sharma, Lakshmi Govindarajan, and Yahya Qasem. Finally, I am also grateful for the support of Professor Dima Amso and members of the Amso lab for their efforts in data collection and processing.

## References

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., et al. (2015). *Mit saliency benchmark.*

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10-12), 1489–1506.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, *2*(3), 194.

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. In *Computer vision and pattern recognition (cvpr), 2015 ieee conference on* (pp. 1072–1080).

Kümmerer, M., Wallis, T. S., & Bethge, M. (2016). Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., & Zetzsche, C. (2001). Scene analysis with saccadic eye movements: top-down and bottom-up modeling. *Journal of electronic imaging*, *10*(1), 152–161.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.