# Predicting the Eye Fixations with Prior Knowledge: A Bayesian Learning Architecture

Lauren Arnett and Chengzhi Mao
Columbia University in the City of New York
{lba2138,cm3797}@columia.edu

## Abstract

*Applications of tracking eye fixation location span from neuroscience and the study of human vision to advertising and human computer interaction. We look to improve upon existing models of saliency.*

## 1. Introduction

Please follow the steps outlined below when submitting your manuscript to the IEEE Computer Society Press. This style guide now has several important modifications (for example, you are no longer warned against the use of sticky tape to attach your artwork to the paper), so all authors should read this new version.

## 2. Related Work

## 3. Dataset of Eye-Tracking Data

## 4. Learning a Model

### 4.1. Baseline Model Setting

We conduct a binary classification task for each output pixel in our baseline eye fixation model. We propose a U-Net architecture, using a fixed pretrained VGG model. We take the features of 3, 8, 15, 22 layers in the VGG and feed into our fixation prediction network. We first learn a upsampling of the high level,low resolution features of the VGG, and then concatenate it with the low level, high resolution features.

### 4.2. Overcoming the Checkerboard Artifacts of Upsampling

Building the upsampling using deconvolution operation introduces checkerboard artifacts, as shown in Fig. This is partly due to the overlap of the deconvolution, according to []. We overcome this by first applying a nearest-neighbor interpolation and then normal convolution operation.

### 4.3. Learning Prior from the ImageNet

Due to the high cost of collecting eye movement data, which requires volunteers to set in front of the computer wearing an eye tracker, the number of samples in the training set is limited. We make an attempt to address this challenges by utilizing some prior knowledge of eye fixation.

We interpret eye fixation points as the places which encode the most semantic information for the image. And with this prior, we construct a model which predicts a binary mask with 0 and 1 to select the image, and then use a neural network to predict the semantic information of that after masking image.

The elements with value 1 in the predicted binary mask mimic the eye fixation point of the human, where the selection of these points should maximize the semantic information in the resulting image after training.

We denote the semantic label of a given image $X$ as $y$, the predicted mask as $M$, and the loss function to minimize $L$

$$M_i = F_1(X_i, k)$$

$$L(X, Y) = \sum_{x_i \in X} logP(y|F_2(M_i * x_i))$$

where * denotes element wise multiplication, $k$ denote the number of value 1 we need in the binary mask.

After training The $M_i$ can be interpreted as a learned prior knowledge, which is fixed during the bayesian learning scheme.

### 4.4. Incorporating Prior Improves Prediction

The main contribution of our work is a bayesian inference scheme that builds upon a prior knowledge learned externally. We denote the eye fixation ground truth as $t$,

$$L = logP(M, t|x) = log(P(M|x)) + log(t|M, x)$$

The prior M gives a good estimation of the eye fixation in advance, which enhance the optimization of this loss function.

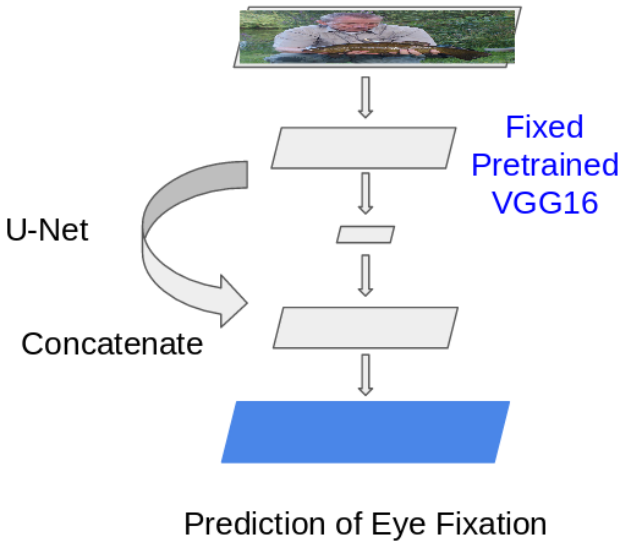


Figure 1. U-Net Architecture for our baseline model

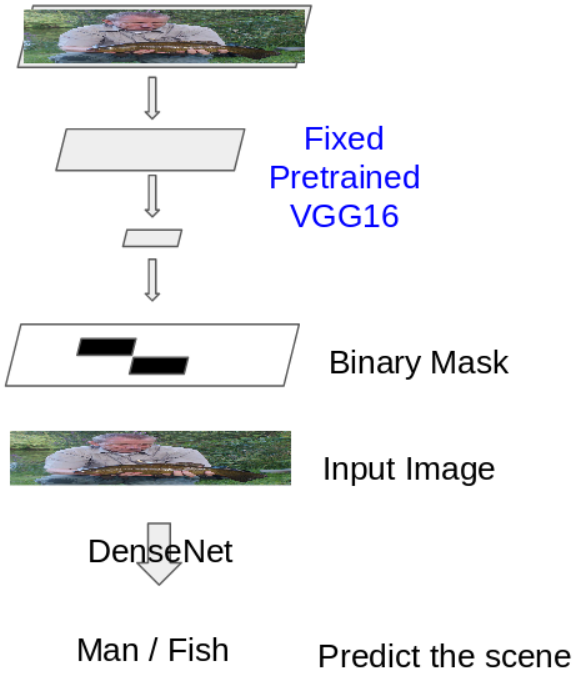


Figure 2. Architecture for learning the prior

### 4.5. Training

We train our baseline model using SGD, with learning rate of 1e-5 and weight decay of 1e-6. We train 50 epoch before we stop.

We use ImageNet data for our prior training, where the mask prediction network is based on a pretrained VGG16 network, and the network to predict semantic meaning from the masked out image is a DenseNet, which we believe have more accurate gradient information. We train this Mask prediction network using SGD with learning rate of 1e-5.

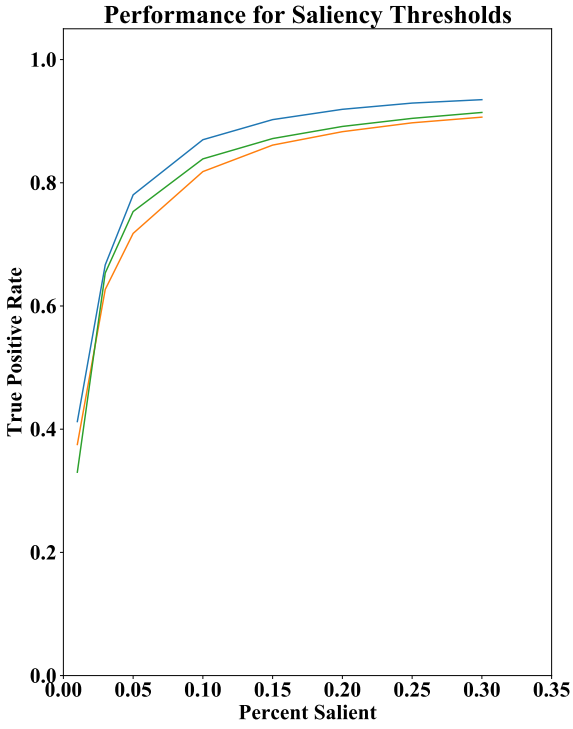


Figure 3. Example of a short caption, which should be centered.

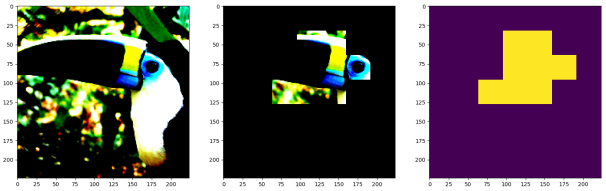

Figure 4. Example of mask learned on Imagenet

### 4.6. Performance

## 5. Conclusion

## References

[1] N. Wilming, S. Onat, J. Ossandón, A. Acik, T. C. Kietzmann, K. Kaspar, R. R. Gamiero, A. Vormberg, P. König. Data from: An extensive dataset of eye movements during viewing of complex images. https://doi.org/10.5061/dryad.9pf75. *Dryad Digital Repository*, 2017.

[2] N. Wilming, S. Onat, J. Ossandón, A. Acik, T. C. Kietzmann, K. Kaspar, R. R. Gamiero, A. Vormberg, P. König. An extensive dataset of eye movements during viewing of complex images. https://doi.org/10.1038/sdata.2016.126. *Nature Scientific Data*, 4(1):160126, 2017.

[3] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 21062113.

[4] L. Itti and C. Koch. A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention. *Vision Research*, 40(10-12):14891506, 2000.

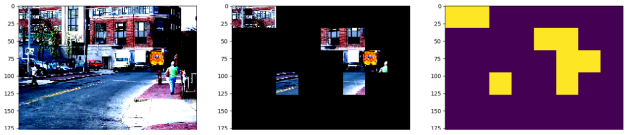

Figure 5. Example of extrapolating the prior learned on ImageNet to the eye fixation dataset