Model-based decoding of bottom-up visual attention in the human brain

Samuel Nastase & Hsin-Hung Li
Machine Learning and Statistical Data Analysis
Dartmouth College
2013

Introduction

At any given time, the brain must cope with nearly infinite sensory information. Attention can be understood as a computational strategy for effectively filtering this torrent of incoming data. More specifically, a critical function of bottom-up attentional mechanisms in the human visual system is to direct the fovea toward important, or salient, objects or events in our environment. Itti and Baldi (2009) developed a neurally-inspired computational model of bottom-up visual attention that computes saliency maps for naturalistic movie stimuli. In brief, the model employs stacked banks of Gabor filters that capture the functional properties of early visual neurons by decomposing movie frames according to several feature maps, including color opponency, orientation contrasts, flicker, and motion energy. Temporal and spatial Bayesian surprise metrics (essentially the Kullback-Liebler divergence) are then computed for each of these feature maps based on the previously observed frames of the movie. This model predicts human observers' gaze shifts at 72% accuracy in individual subjects, and 84% accuracy for gaze shifts common across all subjects. In recent years, several computational neuroscience groups have been able to use neurally-inspired models capturing certain stimulus features to accurately predict brain activity for novel stimuli (Huth, Nishimoto, Vu, & Gallant, 2012; Mitchell et al., 2008; Nishimoto et al., 2011). Nonetheless, these model-based decoding methods have not yet been applied to the neural basis of bottom-up visual attention.

The current study has three principle objectives:

1. Neural validation of the saliency model
2. Localization of saliency processing to specific cortical structures
3. Evaluation of inter-subject commonalities in saliency processing

Method

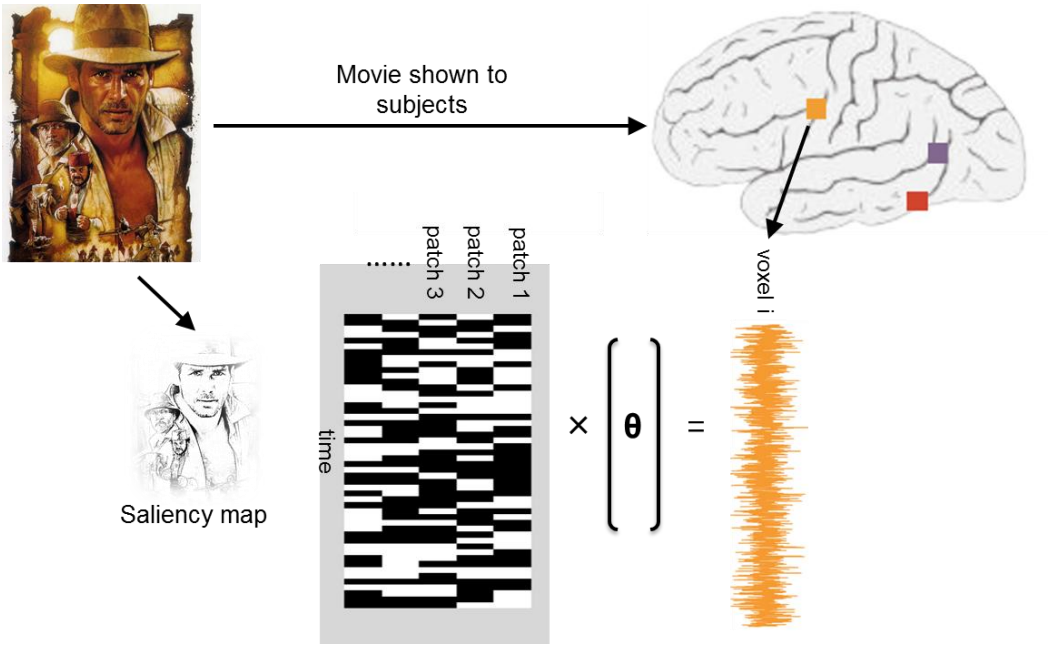
Functional MRI was used to index BOLD (blood oxygenation level-dependent) responses in the brain while 11 subjects viewed a full-length action film (*Indiana Jones: Raiders of the Lost Ark*). For each subject, data consisted of three-dimensional brain images segmented into 3 mm^3 "voxels" (volume pixels), each containing a time series of signal values sampled at 3 s intervals (i.e., TR = 3 s) over the course of the movie (totaling 2,205 time points). This data set is not yet publicly available, although related material can be found at http://www.pymvpa.org.

Prior to implementing the attentional model, we will use the hyperalignment procedure (see Haxby et al., 2011, for details) based on half of the movie data to functionally align each subject's anatomically idiosyncratic responses into a common, high-dimensional space. The purpose of this is to later evaluate whether model parameters derived from one subject's data remain valid across subjects. All subsequent analyses—i.e., model training, cross-validation—will be based on the other half of the movie data to avoid hyperalignment overfitting.

We plan to implement Itti and Baldi's model in order to compute surprise metrics for each patch of each movie frame, which will then be averaged across 3 s segments of the movie in order to match the TR of the fMRI data. Then, a regularized finite impulse response regression model will be

estimated for each cortical voxel recorded in each subject's brain (see Figure for a schematic representation). That is, linear regression will be performed for each voxel to learn the optimal model parameters for predicting that voxel's activity over the course of the movie. During this training procedure, model output will be convolved with a hemodynamic response function to simulate the lag and temporal smearing typical of the BOLD response. The resulting model weights describe how effectively the saliency maps computed by the model drives neural responses recorded for a given voxel. For model training, only data from the half of the movie not used for hyperalignment will be used. Furthermore, this half of the movie will then be broken into 18 s segments for the purpose of cross validation. A leave-two-out cross-validation scheme will be used such that model parameters will be derived from all but two movie segments (Mitchell et al., 2008).

Figure 1: Schematic depicting voxelwise model estimation.



To test model efficacy, the model will then be given the two held out segments of movie data for which it will produce predicted voxel time series for the each voxel in the brain. The predicted brain images will then be compared to the actual brain images for the held out movie segments via their cosine similarity, following Mitchell et al. (2008). The expected classification accuracy across all held out movie segments if the model were performing at chance level would be 0.50. If overall classification accuracy within each subject exceeds statistical significance, this would suggest that the saliency model effectively captures how the brain processes salient visual input.

In order to localize saliency processing in the brain, two methods will be used. First, the most accurately predicted voxels can be localized by projecting the level of correlation between predicted brain images and actual brain images for novel movie segments into brain space. Using standard cluster-based thresholding on the correlation values, this statistical parametric map (SPM) would depict locations in the brain where visual saliency is represented, and, in addition, the level of correlation between predicted and actual activity patterns. We expect that this approach will capture the frontal eye fields (FEF) and several structures in the dorsal visual stream, including the intraparietal sulcus (IPS). The second approach (Huth et al., 2012) goes a step further by projecting the estimated model weights into brain space, thresholded for accurately predicted

voxels. If color scales are applied to the patches of the movie stimulus, perhaps corresponding to direction and eccentricity, this SPM would reflect the allocentric spatial tuning (with regard to the patches of the movie stimulus) of voxels responsible for computing visual saliency. That is, this SPM would convey to what spatial regions of the film stimulus voxels are tuned. Unlike retinotopy, which is retinocentric and requires fixation throughout the experiment, this would reflect an allocentric, or "screen-centered", visual coordinate space.

Finally, we plan to evaluate whether the neural signature of saliency processing is conserved across individuals. In the analyses discussed so far, a separate model will be estimated for each subject, based on the correspondence between their individual brain activity and the movie stimulus. Cortical topographies from individual to individual are highly variable and inter-subject classification based on anatomical alignment has been notoriously difficult. Hyperalignment (Haxby et al., 2011) extracts subjects' functional data from anatomical space into a high-dimensional, abstract space, then aligns these functional data using the Procrustes transformation. By transforming each subject's functional data into a common, high-dimensional space, we will derive hyperalignment parameters by which an individual subject's functional data can be transformed through the common space and projected into any other hyperaligned subject's native space. This procedure typically allows for inter-subject classification at accuracies comparable to within-subject classification. There are two possible strategies for accomplishing inter-subject classification. First, saliency models can be estimated as previously described, then validated in a left out subject. Alternatively, it may be possible estimate the model parameters in the common space, then project these common parameters back into individual subject space for validation. As the hyperalignment procedure was very recently developed, to our knowledge, no group has successfully classified novel stimuli (via a featural model) in novel subjects (via hyperalignment).

Schedule

By the milestone (February 19, 2013), we plan to finish preprocessing the raw fMRI data and complete the hyperalignment procedure. Furthermore, by this date we intend to have a working replication of Itti & Baldi's attentional model and the code needed to optimize model parameters via linear regression. Given time constraints and the complexity of Itti & Baldi's model, we may elect to implement a stripped down version of the model that collates fewer feature maps (e.g., only orientation contrast). Some additional adjustments to the model will be required for our particular movie stimulus, as well as the addition of a hemodynamic impulse response term so that the output of the model reflects the BOLD response. At completion (March 7, 2013), we will have implemented tests on the estimated parameters, performed inter-subject classification, and prepared a poster and final essay conveying our results.

References

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., . . . Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron, 72*(2), 404-416.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron, 76*(6), 1210-1224.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision research, 49*(10), 1295-1306.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *science, 320*(5880), 1191-1195.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology, 21*(19), 1641-1646.