Learning High-Level Concepts by Training A Deep Network on Eye Fixations

NUS National University of Singapore

NUS Graduate School for Integrative Science & Engineering (NGS)

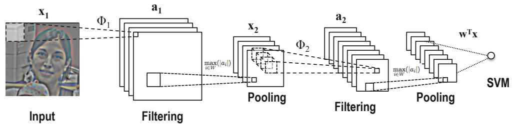
Chengyao Shen, Mingli Song, Qi Zhao

Abstract

Visual attention is the ability to select visual stimuli that are most behaviorally relevant among the many others. It allows us to allocate our limited processing resources to the most informative part of the visual scene. In this work, we learn general high-level concepts with the aid of selective attention in a multi-layer deep network. Greedy layer-wise training is applied to learn mid- and high- level features from salient regions of images. The network is demonstrated to be able to successfully learn meaningful high-level concepts such as faces and texts in the third-layer and mid-level features like junctions, textures, and parallelism in the second-layer. Unlike object detectors that are recently included in saliency models to predict semantic saliency, the higher-level features we learned are general base features that are not restricted to one or few object categories. A saliency model built upon the learned features demonstrates its competitive power in object/social saliency prediction compared with existing methods.

The Model

General Structure



1. Input Layer:
   Whitened to have zero-mean and unit variance in each channel
   $$\mathbf{x}_1 = \frac{\mathbf{x} - \bar{\mathbf{x}}}{var(\mathbf{x})}$$
2. **Filtering Layer:**
   Sparse Coding for Feature Learning and Inference
   $$E = \|\mathbf{x} - \Phi\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1$$
   $$\Phi = \arg\min_{\Phi,\mathbf{a}} E \qquad \mathbf{a} = \arg\min_{\mathbf{a}} E$$
3. **Pooling Layer:**
   Max-pooling to provide invariance to translation and scale change
   $$\mathbf{z} = \max_{i \in W}(|a_i|)$$
4. **SVM Layer:**
   Linear SVM for feature integration and saliency prediction
   $$s = g \circ \max(\mathbf{w}^T\mathbf{x}, 0)$$

Methods

1. **Feature Learning**: Greedy layer-wise training on 100×100 salient patches extracted from MIT fixation [1] and FIFA[2] datasets.
2. **Training and Saliency Prediction**:
   - Train a two-class linear SVM using the responses of salient patches and non-salient patches extracted from the datasets.
   - Predict Saliency using full images in the datasets as inputs.
3. **Feature Visualization:**
   - First Layer Feature: Visualize its weight in direct.
   - Higher Layer Feature:
     - Compute the effect receptive sizes of a higher layer neuron in input space.
     - Crop the regions of 36 top responsive input stimuli throughout the full images in database.
     - Average these input stimuli (*optional*).

Experiments & Results

Datasets

MIT fixation dataset [1]:

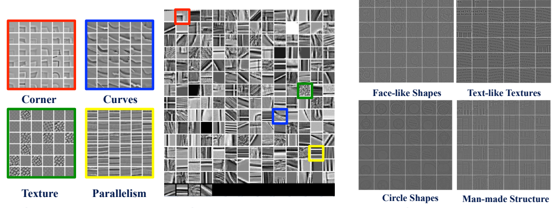
- 1003 images with a variety of objects (mostly 36° × 27°)
- Fixation data collected from 15 subjects
- Largest ever dataset with eye fixations

FIFA (Fixations on Faces) dataset [2]:

- 181 colored natural images (28° × 21°)
- Fixation data collected from 8 subjects
- Most of the images contain faces different sizes and postures

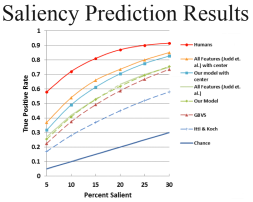
Results on MIT fixation dataset:

Feature Visualization

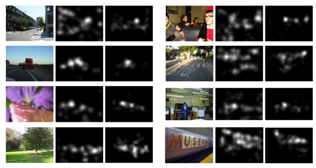


2nd Layer Features (Top 36) 2nd Layer Features (Average) 3rd Layer Features (Top 36)

Saliency Prediction Results



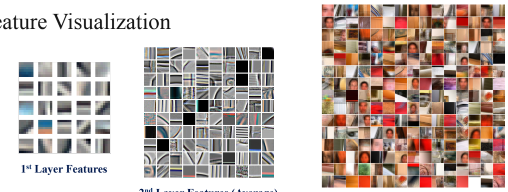
ROC curve (MIT on 2nd layer responses)



Qualitative examples with good performance (Left: original, Mid: Prediction, Right: Ground Truth)

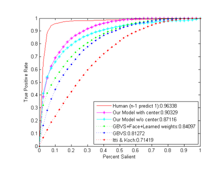
Results on FIFA dataset:

Feature Visualization

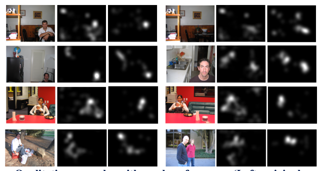


1st Layer Features 2nd Layer Features (Average) 3rdLayer Features (Average)

Saliency Prediction Results



ROC curve (FIFA on 3rd layer responses)



Qualitative examples with good performance (Left: original, Mid: Prediction, Right: Ground Truth)

Conclusion

Contributions

- Learning out meaningful high-level visual features on human fixations.
- The first saliency model that attempt to utilize hierarchies of features learned from natural images to tackle the problem of object/social saliency.

Future Works

- Improve current model in feature learning and parameter tuning.
- Extend the work to dynamic scene, learning invariant feature with temporal coherence and mimic human daily visual experience in feature learning.

Reference

[1]. T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2106–2113, IEEE, 2009.

[2]. M. Cerf, E. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experi- mental data and computer model," *Journal of Vision*, vol. 9, no. 12, 2009.