Neurocomputing 138 (2014) 61–68

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Learning to predict eye fixations for semantic contents using multi-layer sparse network

Chengyao Shen [a], Qi Zhao [b,*]

[a] NUS Graduate School for Integrative Sciences and Engineering (NGS), National University of Singapore, 117456, Singapore
[b] Department of Electrical and Computer Engineering, National University of Singapore, 117576, Singapore

ARTICLE INFO

Article history:
Received 2 August 2012
Received in revised form
29 July 2013
Accepted 16 September 2013
Available online 1 April 2014

Keywords:
Semantic saliency
Gaze prediction
Sparse coding
Deep learning

ABSTRACT

In this paper, we present a novel model for saliency prediction under a unified framework of feature integration. The model distinguishes itself by directly learning from natural images and automatically incorporating higher-level semantic information in a scalable manner for gaze prediction. Unlike most existing saliency models that rely on specific features or object detectors, our model learns multiple stages of features that mimic the hierarchical organization of the ventral stream in the visual cortex and integrate them by adapting their weights based on the ground-truth fixation data. To accomplish this, we utilize a multi-layer sparse network to learn low-, mid- and high-level features from natural images and train a linear support vector machine (SVM) for weight adaption and feature integration. Experimental results show that our model could learn high-level semantic features like faces and texts and can perform competitively among existing approaches in predicting eye fixations.

1. Introduction

Visual attention is a fundamental process of our visual system that happens in our everyday life. It enables us to allocate our limited processing resources to the most informative part of the visual scene. Visual attention has been studied in different areas such as psychology, neurosciences and computer vision. Various computational models, which are called saliency models, have been proposed according to the psychological and neurobiological findings in this area.

Most saliency models follow the "Feature Integration Theory" (FIT) [1–3] framework which suggests low-level visual feature maps such as luminance, color, orientation and motion to compute saliency map and predict human eye fixations [4,5]. These models work well to a certain extent, but are usually insufficient in predicting accurate eye fixations, especially when the scene contains strong semantic objects such as faces, texts, or other socially meaningful contents [6,7].

To overcome this so-called "semantic gap", many improved computational models [7–10] have been proposed to better predict human fixations by integrating higher-level features (e.g., a common practice is to add specific object detectors) into the original low-level feature based models. However, regarding the fact that there are thousands of object categories existing in our daily life, simply adding detectors would make the saliency models more complex and even infeasible in implementation. Hence, a unified framework that could naturally integrate features at various levels is desirable.

Recent advances on deep learning and unsupervised feature learning [11–13] provide us a useful tool for this unified feature integration. Deep learning models are usually multilayer generative networks trained to maximize the likelihood of input data. With sparse priors on the responses of each layer, hierarchies of target-relevant features or bases with increasing complexity could be learnt out in an unsupervised way from a large amount of input data through greedy layer-wise training. After feature learning, multiple levels of sparse representations can then be generated as the efficient coding of the inputs. Such properties of deep learning models are attractivethat they resemble early processing of the primate visual system [14,15].

In this paper, we build our new saliency model upon the deep learning framework in the hope to learn saliency-relevant features from natural images and predict eye fixations that is related to object and semantic contents [16]. The model is built with three layers of filtering units and pooling units stacking together followed by a linear SVM to integrate the top-level feature map into the saliency map. To mimic the images projected to the fovea during eye fixations, the model is first pre-trained on salient regions from the MIT eye tracking dataset [7] and Fixations on Faces (FIFA) [17] dataset for feature learning. Then a SVM training

* Corresponding author. Tel.: +65 6516 6658.
E-mail address: eleqiz@nus.edu.sg (Q. Zhao).

http://dx.doi.org/10.1016/j.neucom.2013.09.053
0925-2312/

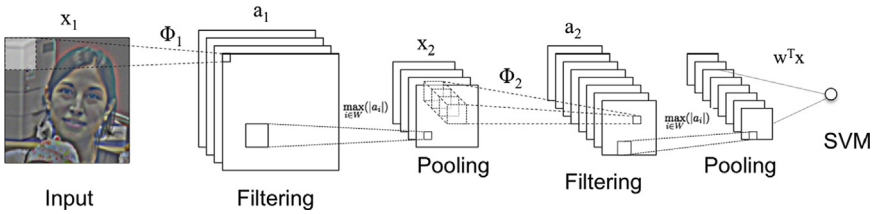

Fig. 1. Architecture of the multi-layer sparse network model, 'Filtering' layers correspond to the feature maps generated by convolutional sparse coding operations and 'Pooling' layers correspond to the feature maps generated by max-pooling operations. A linear SVM is fully connected to the output of the network to train the saliency model.

is performed on the responses of salient and non-salient regions from the datasets to learn the weight of each feature map. After training, the model is applied to test images from the same datasets and saliency maps are generated by organizing each response in a small region to a map. Experimental results show that the model is competitive among existing models to predict gaze.

The main contributions of our work are as follows:

1. We learn meaningful high-level visual features using the principled framework of deep networks by modeling the way humans sample the visual scene and we show that this way of sampling plays an important role in the learning of these features.
2. We propose a unified feature integration framework for saliency detection that could integrate low-, mid- and high-level features learned from natural images.

The rest of the paper is organized as follows. In Section 2, we review related works on saliency detection and deep network. We then present the model of multi-layer sparse network and the way of training and testing the model in Section 3. In Section 4, experiments are conducted on MIT eye tracking and FIFA datasets and both quantitative and qualitative results are given. Section 5 concludes the paper.

2. Related works

In recent years, due to the limitation of classical saliency model based on low-level features [6,7], there have been growing interests in modeling eye fixations by integrating mid-/high-level features [17,9,7,18]. Cerf et al. [17] refine Itti and Koch's model [4] by adding a face detector. Zhao and Koch [9] further improve Itti and Koch's model [4] by using a least square technique to learn the weights of face and low-level feature maps from different eye tracking datasets. In Judd et al.'s work [7], low-level features including statistics of local orientations, luminance and colors, mid-level features such as a horizon line detector, and high-level features such as a face detector and a person detector are integrated by a linear SVM to predict where humans look. Based on Judd et al.'s work, Lu et al. [18] further improve the saliency computation by including Gestalt cues such as convexity, symmetry and surroundedness into their model. All these works indicate that mid-/high-level features play an important role in predicting human fixations, but there still lacks a unified framework that could integrate various low-, mid- and high-level features that have been mentioned or not mentioned above.

Also closely related are deep learning models that could learn higher-level features from natural images. In one seminal work [11], Lee et al. show that, by training on well-aligned images from the Caltech 101 dataset [19], hierarchies of representations which correspond to object parts and objects could be learned with a convolutional Restricted Boltzmann Machine (RBM). In [12], Zeiler et al. propose a hierarchical sparse network in which each layer reconstructs the input and shows that edges, junctions, and even object parts can be learned out from the images that contain objects. In one recent work [13], Le et al. build a three-layer deep auto-encoder and prove that neurons representing faces, human bodies, and cats can be learned out in a fully unsupervised way on images sampled from 10 million YouTube videos. These models all validate that, by training on natural images, meaningful high-level features can be learned out using a deep network. However, none of them has considered the influence of visual attention on the feature learning in deeper levels. Furthermore, compared with existing works, our model is able to learn out meaningful high-level neurons in relatively few samples with the aid of eye fixations.

3. The multi-layer sparse network framework

In this section, we describe the hierarchical model that is used to learn features from natural images and predict visual saliency. The general structure of the model is shown in Fig. 1, which is composed of multiple layers of filtering and pooling sublayers stacking together (here we only show two layers for the brevity of illustration) and a linear SVM at the end to generalize the responses of the network to visual saliency.

This hierarchical model can be seen as a natural extension of previous hierarchical models such as Neocognitron [20], HMAX [21,22] and Convolutional Neural Network [23] that aim to model the hierarchical structure of ventral stream[1] in the visual cortex. This structure is also a common structure employed by many recent deep learning models [11–13].

3.1. Sparse coding and unsupervised feature learning

Sparse coding is an unsupervised scheme that learns to represent input data using a small set of bases (or features). It is the core computational algorithm in our model and constitutes the basic unit for the filtering layer.

The idea of sparse coding originates from Barlow's principle of redundancy reduction [25], which states that a useful goal of sensory coding is to transform the input in such a manner that reduces the redundancy of the input stream. In its original form of modeling image patches [26], it can be described as a generative image model as

$$E = \|\mathbf{x} - \Phi\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1 \quad (1)$$

[1] The division of "Ventral Stream" and "Dorsal Stream" is a widely accepted concept to the function of primate visual cortex. The ventral stream (also called "What Pathway") is related to object recognition and form representation and is found to have a hierarchical structure with larger receptive field size, and more complexity along the stream from V1 to AIT [24]. The"Dorsal Stream" (also called "Where Pathway") deals with the guidance of actions and localization of object.