Efficient Visual Attention with Deep Learning

Anonymous CVPR submission

Paper ID ****

Abstract

The high volume of visual data contains information that is mostly irrelevant for intelligent agents. Humans perform sensorial filtering through a mechanism called attention. In this work, we present a new fully convolutional neural network architecture designed for detecting visual salience. Experiments carried out with the MIT300 benchmark presented state-of-the-art performance and a parameter reduction of 3/4 compared to similar models.

1. Introduction

One of the most challenging unsolved problems in Artificial Intelligence is vision. However, it is fundamental for the conception of systems that interact with the real physical world. Such systems would be useful for applications in areas like domestic services, industry and agriculture, with great potential for the benefit of society.

Vision is remarkably data and computationally intensive. In humans, approximately half of the brain is involved in vision-related tasks [4]. Even our minds can not handle all the sheer amount of sensorial information received every second. In order to deal with this amount of data, humans have attentional systems, a fundamental mechanism that, among other functions, filters out irrelevant information – either visual or from other senses– and helps us focusing our cognitive processes on what is important at a given moment. These facts are a strong evidence that, in order to help solving the vision problems, attention should be applied.

Visual attention can be defined as the delimitation of a certain spatial region on an image for further cognitive processing [12]. The phenomenon emerges from two fundamentally different processes: the *top-down* mechanism that implements our longer-term cognitive strategies by biasing attention according to one's interests (e.g. find a red apple in a tree because of hunger, which will make red be more recognizable on the scene), and the *bottom-up* mechanism [2], a process generated through external stimuli that captures one's attention from its conspicuousness level. In this work, we focus on the latter, also named visual saliency.

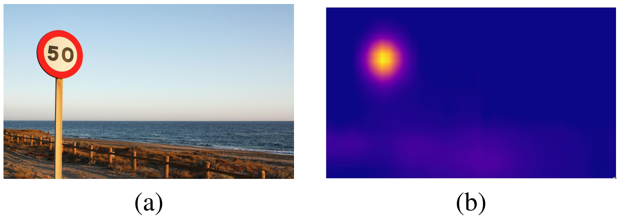
(a) (b)

Figure 1. Example of visual saliency. b) is the salience map where brighter pixels (warmer colors) represent regions more salient to humans on the original image a).

Visually salient regions on images are usually represented by *saliency maps* (Figure 1). In these maps, images are generated such that areas with high-valued pixels express high saliency on the original image, whereas regions with low-valued pixels represent low saliency. Datasets with such maps are obtained by collecting eye-fixation data from humans while observing the scenes.

1.1. Related work

Early computational models of visual saliency were generally built based on filtering of images for extraction of a pre-selected set of features considered important for *bottom-up* attention. *Vocus* [5] is a computational model that extracts features shown to be naturally salient to humans such as color/luminance contrast and orientation from different scales of the image.

A rapid change of paradigm occurred around 2015 when *Deep Learning* techniques showed to be very effective in the generation of saliency maps. Models such *Salicon* [6] demonstrated that the use of convolutional neural networks with weights initialized from image classification networks, e.g. *VGG-16* [10] could considerably increase the similarity of computed maps to those generated from humans. *ML-Net* [3] uses the output of different layers of *VGG-16*, combining them in many dimensions and various levels of abstraction. *DeepFix* [8] extends a pre-trained model with new layers that account for global features and center bias, whereas *Salnet* [9] explores two models that are simple yet provide good results. These models are usually evaluated

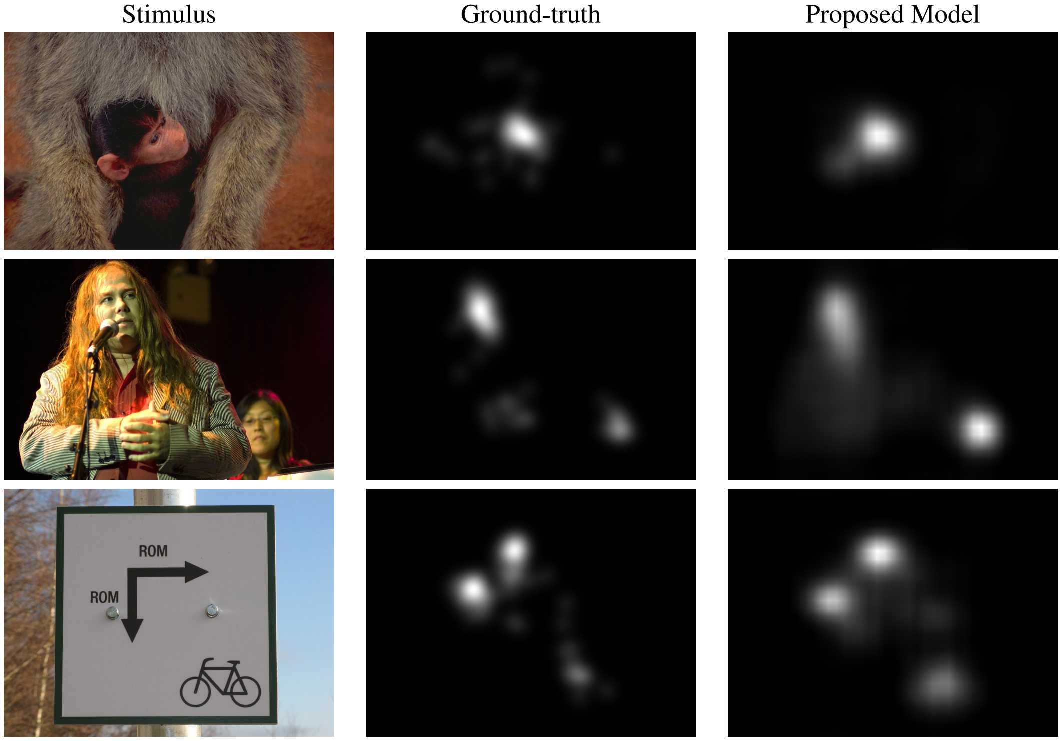

Figure 4. Examples of predictions made by our model.

Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 2

[12] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognit Psychol*, 1980. 1

Table 1. Number of filters used in each inception block.

Block	pool	conv 1×1	3×3 reduce	conv 3×3	5×5 reduce	conv 5×
1	96	128	96	192	58	96
2	64	128	80	160	24	48
3	64	128	80	160	24	48
4	64	128	96	192	28	56
5	64	128	96	192	28	56
6	64	128	112	224	32	64
7	64	128	112	224	32	64
8	112	160	128	256	40	80

Table 2. State of the art models and metric scores on *MIT300 benchmark*.

Model	Num. parameters	AUC-Judd ↑	CC ↑	NSS ↑	Sim ↑
Infinite humans	-	0.92	1.0	3.29	1.0
DeepFix	≈16.7 million	0.87	0.78	2.26	0.67
Salicon	≈14.7 million	0.87	0.74	2.12	0.60
Proposed Model	**3.72 million**	**0.85**	**0.71**	**1.98**	**0.62**
ML-Net	≈15.4 million	0.85	0.69	2.07	0.60
SalNet	25.8 million	0.83	0.57	1.51	0.52