Visual Attention with Deep Learning

Erik Perillo[1], **Esther Luna Colombini**[1]

[1]Institute of Computing (IC) – University of Campinas (Unicamp)
Caixa Postal 6176 – 13.084-971 – Campinas – SP – Brazil

erik.perillo@gmail.com, esther@ic.unicamp.br

Abstract. *Vision is a key element in one's process of understanding the world. The high volume of sensorial data is however problematic because most of the information is often irrelevant. Humans realize sensorial filtering by what we call attention. We propose the application of Deep Learning for obtaining a visual salience system which behaves similarly to humans. We built a new convolutional neural network with relatively simple architecture, yielding a performance level consistently among the best ten state of the art models in MIT300 benchmark.*

Resumo. *A visão é elemento-chave no processo de entender o mundo para um ser. Entretanto, a alta quantidade de dados sensoriais é problemática, havendo muitas vezes irrelevância de informação. Nos seres humanos, há um filtro sensorial realizado pela atenção. Propomos a aplicação de Deep Learning para a obtenção de um sistema de saliência visual que se comporte como o dos seres humanos. Construímos uma nova rede neural convolucional de arquitetura relativamente simples e com um desempenho que a coloca consistentemente entre os dez melhores modelos estado da arte no MIT300 benchmark.*

1. Introduction

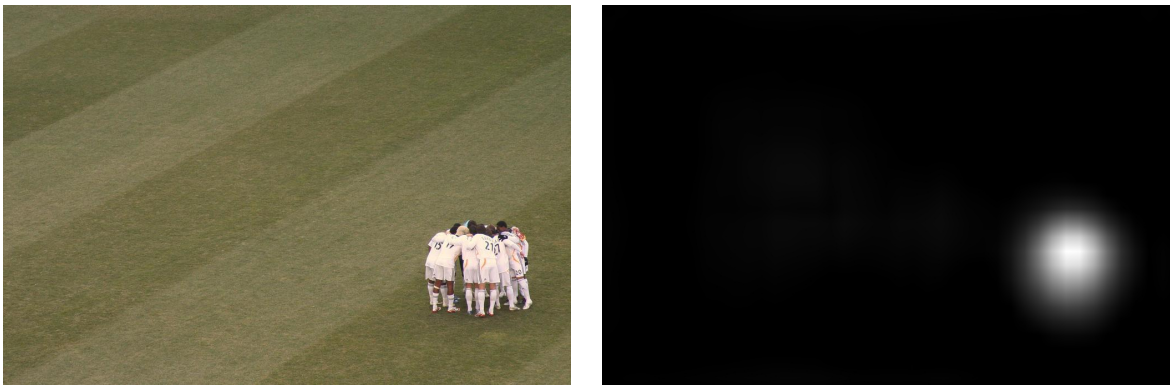
One of the most challenging unsolved problems in Artificial Intelligence is vision. It is fundamental for the conception of systems that interact with the real, physical world. Such systems would be useful for applications that involve robotics and tasks in domestic houses, industry and agriculture, so there is great potential for the benefit of society.

Vision is remarkably data and computationally intensive: In humans, approximately half the brain is involved in vision-related tasks [Fixott 1957]. Even our brains can't handle all the sheer amount of sensorial information that we receive every second: We have attention, a fundamental mechanism that, among other functionalities, filters out irrelevant information – either visual or from other senses– and helps us focus our cognitive processes on what is important at a given moment. These facts are a strong evidence that, in order to solve vision, we need to have attention.

1.1. What is visual saliency?

Visual saliency can be defined as the delimitation of a certain spatial region on an image for further cognitive processing [Treisman and Gelade 1980]. Psychologists have been studying for at least half a century what makes we direct our visual focus to a certain region and how we do it.

The phenomenon of visual saliency may emerge from two fundamentally different processes: There is *top-down* attention, an internal stimuli of the agent (e.g. find a red apple in a tree because of hunger, which will automatically make red be reconizable more easily on the scene), and *bottom-up* attention, an external process which captures the agent's attention from abrupt changes in visual stimuli or high contrast. It has been shown that some patterns are naturally visually salient for humans. Some of them are: color contrast (specially green-red and yellow-blue), luminance contrast, orientation contrast, high frequency movement [Colombini 2014]. In this work, we focus on *bottom-up* attention.



(a) (b)

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

Visual salient regions on images are usually represented by *saliency maps*. Such maps are grayscale images generated such that areas with pixels close to white express high saliency of the same region on the original image, whereas regions with pixels close to black represent a region of low saliency. Datasets with pairs image-map are usually obtained by colleting eye-fixation data from humans that looked at the images.

1.2. 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. Many of them are based on theoretical models of attention, such as the *Feature Integration Theory* [Treisman and Gelade 1980] and *Guided Search* [Wolfe et al. 1989]. *Vocus* [Frintrop 2005] is a computational model that extracts features such as color contrast, orientation and luminance contrast from different scales of the image to produce saliency maps in the style of figure 1b.

A rapid change of paradigm ocurred around 2015 when *Deep Learning* techniques showed to be extremely effective in the generation of saliency maps. Models such as *Salicon* [Jiang et al. 2015] showed that applying convolutional neural networks with weights initialized from networks used for image classification, e.g. *VGG-16* [Simonyan and Zisserman 2014] could yield maps very similar to those generated from humans. Later works further explored this idea. *ML-net* [Cornia et al. 2016] uses the output from different layers of *VGG-16* to use information from various dimensions

Block number	Filter name	Number of filters
1	conv 1	48
2	conv 1	64
2	conv 2	96
3	conv 1	128
3	conv 2	128
3	conv 3	144
3	conv 4	144

Block number	conv 3x3 reduce	conv 3x3	conv 5x5 reduce	conv 5x5	conv 1x1	pool
1	32	64	24	48	64	64