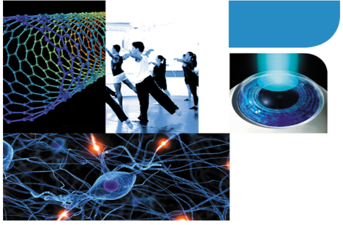

XXV Congresso de Iniciação Científica da Unicamp

October 18 to 20 Campinas | Brazil

Efficient Visual Attention with Deep Learning

Erik de Godoy Perillo*, Esther Luna Colombini.

Abstract

The high volume of visual data usually available for autonomous applications contains information that is mostly irrelevant. Humans perform sensorial filtering through a mechanism called attention. In this work, we present a new fully convolutional neural network 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.

Key words:

Machine Learning, Computer Vision, Attentional Processes.

Introduction

Vision is a key component for intelligent agents that interact with the world. Even for humans, it is impossible to process all visual information received every second: we have attention, a filtering mechanism that directs our cognition to what is important at a given time.[1] This is motivation for a work towards artificial visual attention.

Visual salience systems aim at, given an image, produce a salience map where pixels close to white represent areas that are more prone to be looked first by humans on original image. Recent systems using convolutional neural networks showed to be very effective.[2] However, such systems are computationally expensive, partly because they use networks designed and pre-trained originally for other tasks such as classification.

In this work, we aim at developing a new, efficient fully convolutional neural network designed specifically for salience detection, exploring concepts from psychology not found to be used in recent models.

Results and Discussion

Image 1 illustrates the proposed network architecture. It extracts increasingly higher-level features from the input image. Layer blocks 1, 2, 3 apply 3x3 convolutions followed by ReLU activation and max-pooling, downsampling image by a factor of 2. Block 4 uses 8 inception layers that apply convolutions of filter sizes 1x1, 3x3, 5x5, and pooling in parallel, allowing for a combination of information from different scales of the image, an important process for salience detection.

We apply pre-processing techniques not found to be used in similar models but considered to be important in the context of salience: first, we use the LAB colorspace, which is more closely related to human color perception than RGB.[3] We also apply normalization per image rather than per dataset: in the context of salience, we consider more appropriate to use the global context of the image. Prior experiments showed better performance using such techniques.

Image 2 shows one example of salience map generated by our model. We evaluated our model on MIT300 benchmark (table 1).[2] It obtained state-of-the-art scores while having around one fourth of the number of parameters of other similar models.

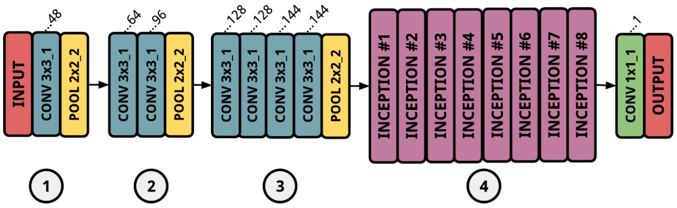

Image 1. Overview of the proposed network architecture.

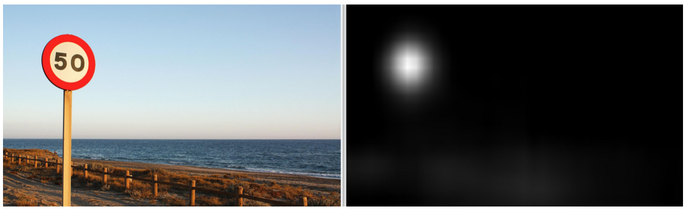

Image 2. Salience map generated by model from image.

Table 1. Proposed model results on MIT300 benchmark compared to other state of the art models.

Model	N. Params	AUC Judd	NSS	Sim
DeepFix	~16.7 M	0.87	0.78	0.67
Salicon	~14.7 M	0.87	0.74	0.60
Proposed	**3.7 M**	**0.85**	**0.71**	**0.62**
ML-Net	~15.4 M	0.85	0.69	0.60

Conclusions

The proposed network architecture and data pre-processing designed for visual salience showed to be effective, yielding state-of-the-art results on MIT300 benchmark with around 3/4 less parameters than similar models. We now aim at extending the model for videos.

[1] Treisman, A. M.; Gelade, G. A feature-integration theory of attention. **1980.**

[2] Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; Torralba, A. MIT salience Benchmark. **2016.**

[3] Frintrop, S. VOCUS: a visual attention system for object detection and goal-directed search. **2005.**