Deep Learning Based Visual Attention Models for Salient Object Recognition

Cristina Rodrigues Melício

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. Alexandre José Malheiro Bernardino
               Prof. José Alberto Rosado dos Santos-Victor

Examination Committee
Chairperson: Prof. João Fernando Cardoso Silva Sequeira
Supervisor: Prof. Alexandre José Malheiro Bernardino
Member of the Committee: Prof. Maria Margarida Campos da Silveira

June 2019
Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.
Acknowledgments

I would first like to thank my supervisors, Professors Alexandro Bernardino and José Santos-Vitor, for their great guidance and insightful comments throughout the course of this thesis.

Secondly, I must thank Rui Figueiredo, for always being available to discuss whichever issues I had.

A sincere thank you goes also to my college friends, both for their companionship and for keeping me on track when tough times were leading me astray.

Last but not least, I would like to thank my family, in particular my mother and father, for molding me into who I am today.
Resumo

Os métodos de processamento de imagem sofreram grandes avanços nas últimas décadas. No entanto, ainda são computacionalmente muito dispendiosos. Por esta motivo, atualmente têm sido pesquisadas abordagens biologicamente inspiradas de forma a tornar o processamento de uma cena mais eficiente. Em particular, neste trabalho, propomos uma estrutura geral computacional inspirada pela visão humana. Esta framework é capaz de realizar tarefas de reconhecimento de objetos combinando Redes Neurais Convolucionais (CNNs) com técnicas de visão foveal. Ela integra duas metodologias dissociadas de atenção visual que podem ser realizadas sequencialmente: (1) um modelo de Saliência Foveal capaz de orientar o foco de atenção para as regiões de interesse, e (2) um modelo de Focagem Piramidal para realizar o reconhecimento de objetos quando o sistema já fixou o alvo. Usando uma partição do conjunto de dados de validação do ILSVRC 2012 e uma rede GoogLeNet pré-treinada, realizamos vários testes para avaliar o trade-off entre o desempenho do reconhecimento e o custo computacional e que parâmetros dos modelos o influenciam. Demonstramos que, usando o método de Saliência Foveal, há um aumento significativo no desempenho de classificação para objetos não centrados ao realizar pelo menos duas iterações sacádicas. Quanto ao modelo de Focagem Piramidal, embora o custo computacional aumente linearmente com o número de níveis de pirâmide, ele alcança consistentemente melhores resultados do que a abordagem clássica de redimensionar a imagem para o tamanho da rede. Esses resultados promissores destacam a importância do desenvolvimento de soluções visuais biomiméticas para tarefas visuais.

Palavras-chave: Atenção Visual, Redes Neurais Convolucionais, Reconhecimento de Objetos, Foveação, Saliência, Pirâmide Gaussiana
Abstract

Image processing methods have made great strides in the last decades. Nonetheless, they are still very computationally expensive. For this reason, in order to more efficiently process a scene, biologically inspired approaches are a currently being researched. In particular in this work, we propose a general computational framework inspired by human vision. This framework is capable of performing object recognition tasks by combining Convolutional Neural Networks (CNNs) with foveal vision techniques. It integrates two dissociated visual attentional methodologies that can be performed sequentially: (1) a Foveal Saliency model capable of orienting the focus of attention to the regions of interest, and (2) a Pyramidal Focus model to perform object recognition when the system has already fixated the target. Using a partition of the ILSVRC 2012 validation data set, and a pre-trained GoogLeNet network, we conducted several tests to evaluate the trade-off between the recognition performance and the computational cost, and which model parameters influence it. We demonstrated that by using the Foveal Saliency method there is a significant increase on the classification performance for non-centered objects when using at least two saccade iterations. As for the Pyramid Focus model, although the computation time increases linearly with the number of pyramid levels, it consistently outperforms the classical approach of resizing the image to the size of the network. These promising results highlight the importance of developing biomimetic visual solutions for visual tasks.

Keywords: Visual Attention, Convolutional Neural Networks (CNNs), Object Recognition, Foveation, Saliency, Gaussian Pyramid
## Contents

Declaration ............................................................... iii
Acknowledgments ......................................................... v
Resumo ................................................................. vii
Abstract .............................................................. ix
List of Tables .......................................................... xiii
List of Figures .......................................................... xv
Nomenclature ........................................................... xvii

1 Introduction ......................................................... 1
  1.1 Motivation ....................................................... 1
  1.2 Objectives ....................................................... 2
  1.3 Thesis Outline .................................................. 2

2 Visual Attention Background ...................................... 5
  2.1 Visual Attention ................................................ 7
    2.1.1 Unit of Attention ........................................ 7
    2.1.2 Overt and Covert Attention ................................ 7
    2.1.3 Bottom-up and Top-down Attention ......................... 8
  2.2 Psychological Theories of Visual Attention .................. 8
    2.2.1 Feature Integration Theory ................................ 8
    2.2.2 Guided Search Theory ..................................... 9
    2.2.3 Saliency Map Model ......................................... 9

3 Deep Learning Supporting Concepts ............................ 11
  3.1 Deep Neural Networks (DNNs) ................................. 11
    3.1.1 Backpropagation algorithm .............................. 13
    3.1.2 Overfitting ............................................... 14
  3.2 DNNs in Image Classification ................................ 14
    3.2.1 Convolutional Neural Networks (CNNs) .................. 15

4 Related Work ..................................................... 17
  4.1 Convolutional Neural Networks for Salient Object Detection .. 17
List of Tables

6.1 Correspondence of the smallest images dimensions that can generate $N_L$ pyramidal levels. ................................................................. 34

7.1 Classification Error (in %) as a function of the image dimensions for the top-5 and top-1 errors for our Pyramidal Focus (PF) method and for the baseline solution. ............................................ 42
List of Figures

2.1 Visual areas and paths of the human brain ........................................... 5
2.2 Distribution of cones and rods along the retina of the human eye ............ 6
2.3 Most influential psychological theories for visual attention .................. 9
2.4 Koch and Ullman Saliency Map Model .................................................. 10
3.1 Artificial Neuron structure ................................................................. 12
3.2 Schematic of an artificial neural network ............................................ 12
3.3 Example architecture of a CNN ......................................................... 15
3.4 Layout of Convolutional and Pooling layers ....................................... 16
5.1 Foveal Saliency Method Scheme ........................................................ 23
5.2 Gaussian Kernel with $\sigma = 1.$ ......................................................... 24
5.3 Artificial Foveal Visual system scheme .............................................. 25
5.4 Pyramidal Focus method scheme for object recognition ....................... 27
6.1 Baseline architecture ........................................................................... 30
6.2 Example of images with different foveal visual parameters as: foveation point $(u_0, v_0)$ and fovea size $f_0.$ .................................................. 32
6.3 Example of the weakly supervised object detection for the same image with the ground truth red fox ....................................................... 33
6.4 Examples of our Pyramid Focus method ............................................. 34
6.5 ImageNet partition dataset characteristics ......................................... 35
7.1 ImageNet partition dataset distribution of the center of the bounding boxes in the resized image domain ................................................. 38
7.2 Classification Error (in %) as a function of the initial foveation point $(u_0, v_0)$ .......................... 38
7.3 Classification Error (in %) as a function of the fovea size $f_0$ over 6 model iterations. ... 40
7.4 JPEG Compression Reduction Gain (in %) between foveated and cartesian images. ... 41
7.5 Classification Error (in %) as a function of the ratio between the ground truth bounding box area and the images resolution ........................... 42
Nomenclature

Acronyms

ANN  Artificial Neural Network
CNN  Convolutional Neural Network
DNN  Deep Neural Network
EMA  Exponential Moving Average
FIT  Feature Integration Theory
FS   Foveal Saliency Method
GPU  Graphics Processing Unit
GS   Guided Search Model
ILSVRC ImageNet Large Scale Visual Recognition Challenge
IOR  Inhibition Of Return
PF   Pyramidal Focus Method
ReLU Rectified Linear Unit
WTA  Winner-Take-All

Parameters

\((u_0, v_0)\)  First foveation point
\(\alpha\)  Degree of weighting decrease on the exponential moving average
\(\theta\)  Segmentation mask threshold
\(f_0\)  Size of the high acuity region
\(N_L\)  Number of Gaussian pyramid levels on the Foveal Saliency Method
\(Q\)  JPEG Quality factor
\(S\)  Score vector of prediction
Chapter 1

Introduction

1.1 Motivation

The visual acuity attained by the human eye is not uniform. In fact, the resolution of the captured images is higher in the fovea, the small central region of the eye, decaying drastically as we approach the periphery [1]. This anatomical structure acts as a space-variant sensor that samples images with detailed information only at the point of gaze, coding progressively less information further away from it.

This non-uniform visual sampling leads to the need of moving the eyes in order to process an entire scene. However, the amount of visual stimuli in the real world that reaches the eyes (0.1 to 1 Gbits/s for each) is too high for the available brain resources, making it impossible to process an entire scene with the high resolution part of the eye [2]. Hence, humans have mechanisms of visual attention used to reduce the amount of information that is processed. These mechanisms select only the relevant parts of an image and prioritize them in time. These parts are known as the salient parts.

In cognitive psychology [3], visual attention is considered to operate as a two-stage process. In the first stage, attention is distributed uniformly over the visual field and the extraction of saliency information is performed in parallel. In the second stage, the attention is focused to a specific area of the visual scene, and processing is performed in a serial fashion.

Likewise, in computer vision, robots share similar resource limitations when processing image information in real-time [4]. Conventional computational vision solutions are based on the increase of the number of pixels in an image which results in an increase of the amount of raw data to be processed. Common camera sensors sample the real world uniformly, capturing unnecessary parts of the visual field that can be distracting. The main differences between these systems and the human vision system are the sensor topology and the system feedback.

A new set of techniques known as deep learning has led to exciting new developments in the field
of artificial intelligence [5]. Among these techniques Convolutional Neural Networks (CNNs) are of particular interest for computer vision. These networks are inspired biologically on the connectivity pattern between neurons existing on the visual cortex of mammals. Thus, they have made the development of sophisticated systems for specialized vision tasks possible. Despite all this progress, the problem of creating more general visual models that replicate human vision remains challenging. With this goal in mind, much effort has been made towards understanding the efficient human vision system and designing algorithms based on foveal inspired sensors and selective attentional mechanisms of feedback.

### 1.2 Objectives

The main goal of this work is to study the need for human inspired vision systems capable of performing object detection and recognition when computational resources are limited. To this end, we developed two methodologies, anatomically and physiologically inspired by the human visual attention mechanism. These two systems integrate a foveal inspired way of sampling the image, an active vision strategy to control the gaze, and a selective visual mechanism to analyse only the relevant regions of the image. Moreover, the brain processing is mimicked by the use of pre-trained CNNs for feature extraction in order to choose where to look and to replicate the high level cognitive object recognition processing.

In our work we developed two dissociated visual attention methodologies that represent a proof-of-concept of an integrated biomimetic solution for detecting object classes present in images: (1) a Foveal Saliency model capable of orienting the focus of attention to the regions of interest, and (2) a Pyramidal Focus model to perform object recognition when the system has already fixated the target. We intend to evaluate the trade-off between the recognition performance and the computational cost, and which model parameters influence it.

This thesis was based on published works: the paper “Object Detection and Localization with Artificial Foveal Visual Attention” in Proceeding of the 8th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics 2018 [6]; and the poster “CNN based Object Classification with Multiresolution Foveated Pyramid” on IROS 2018 Workshop “Unconventional Sensing and Processing for Robotic Visual Perception” [7].

### 1.3 Thesis Outline

The remainder of this thesis is organized as follows: in chapter 2 we overview the visual attention concepts and some physiological theories that inspired the proposed attentional frameworks, then in chapter 3 we detailed the support concepts of deep learning useful for the methodologies implementation. In chapter 5 we describe in detail the proposed methodologies, the foveal saliency method in section 5.1 and the pyramidal focus method in section 5.2. In chapter 6 we explain the experiments, implementation aspects and and the metrics used to analyse the results. Then, in chapter 7 we quantitatively evaluate
our models according to the object recognition performance and to the computational cost trade off. Finally, in chapter 8 we wrap up with the conclusions, our contribution and ideas for future work.
Chapter 2

Visual Attention Background

Vision is one of the five senses that allows organisms to perceive and interact with the environment, enabling a greater knowledge of the world. Visual perception is the ability to interpret our surroundings through the visible light reflected on objects present in the environment.

In humans, and a number of other mammals, visual perception arises when the light is captured by the eye and projected onto the retina, which is light-sensitive membrane in the back of the eye. The retina, due to its specialized photoreceptive cells, acts as a transducer for the conversion of light into neuronal signals. These signals are transmitted by the optic nerve, from the retina to the visual cortex. The structure of the visual cortex occurs in a columnar organization, meaning that neuronal cells within each layer perpendicular to the surface of the cortex have similar response properties. This structure contributes to an increase in the complexity of the neuronal representations of the captured images.

![Figure 2.1: Visual areas and paths of the human brain](http://philosophy.hku.hk/courses/cogsci/ncc.php)
The visual cortex is divided into the primary visual cortex and secondary visual areas, as shown in Figure 2.1. The former is the region where the visual signals that are captured by the eyes end up. After leaving the primary visual cortex, the visual information is divided into two pathways which are characterized by specialized processing [8]. The dorsal stream passes through the parietal cortex, which is responsible for spatial localization and communication with the regions associated with eye movement towards areas of interest; and the ventral stream passes through the inferior temporal cortex, which is mainly responsible for the recognition and identification of visual stimuli.

The amount of visual information that reaches the human eyes is quite high, so the resources that the brain has at its disposal are not enough to process it all simultaneously. Therefore, in the retina the photoreceptive cells are not uniformly distributed, i.e. are not at their maximum density over the entire visual field, as shown in Figure 2.2. There are two types of photoreceptive cells also known as cones and rods. The former are concentrated at a very large density in a small central region of the retina, the fovea, and are responsible for daytime vision being sensitive to color light. The latter are located with a decreasing density towards the periphery and are responsible for night vision [1]. This anatomic distribution is the reason behind images being processed with higher resolution in the fovea, as well as the fact that the resolution decays drastically when approaching the periphery of the retina. The non-uniform distribution of the number of photoreceptors leads to the need to move the gaze towards regions of interest, making the image of the visual stimulus fall on the fovea in order to be processed with higher resolution. This process is called foveation. Thus, in addition to the human eye anatomy, there are also functional mechanisms, such as visual attention, that contribute to reduce the information to be processed by the brain.

Figure 2.2: Distribution of cones and rods along the retina of the human eye. (a) Density of the receptors in degrees relative to the fovea. (b) Cones are concentrated in the fovea while the rods are absent and are at their highest density in the periphery. There are no receptors in the blindspot.

Source: https://foundationsofvision.stanford.edu/chapter-3-the-photoreceptor-mosaic/
2.1 Visual Attention

In everyday life, visual scenes usually contain a huge amount of visual information, about 0.1 to 1 Gbits per second [2], that cannot be analyzed simultaneously due to limited resources of the visual system. The concept of visual attention refers to the cognitive operations that allow humans to efficiently perceive that information.

Over the years, several definitions of visual attention have been proposed. However, the one that is most widely accepted today is selective attention. This is due to the attempt of solving the problem of the lack of cognitive resources. This concept consists of processing with higher detail only the relevant sub-regions of the visual field, called focus of attention, which are focused through selection mechanisms. Visual attention is sometimes compared with a spotlight in a dark room [9], where the fovea is associated to that spotlight. Thus, by moving it around, one can obtain an impression of the room contents and this corresponds to scanning the scene with quick eye movements.

2.1.1 Unit of Attention

A visual stimulus is something present in the visual field that can trigger our attention depending on the degree of importance. The more relevant ones are denominated salient. According to the mechanism of selective attention visual stimuli are organized and processed in descending order of saliency making attention a sequential process. The relevance of visual stimuli can be influenced in three different ways: by the spatial location of objects and the knowledge one has about the world (space-based attention) [10]; by certain features of the object, such as color, size, orientation, field of motion, regardless the object location (feature-based attention) [11]; or by the specific object structure (object-based attention) [12]. These types of attention are not mutually exclusive and usually we focus on more than one region of interest simultaneously. For example, when we look for a friend at a party that we know loves to dance, we place our spatial-based attention on the dance floor, the feature-based attention in the blonde color of his hair, and the object-based attention in his familiar body structure.

2.1.2 Overt and Covert Attention

The orientation of the focus of attention to process regions of interest is called overt attention which is associated with the movement of the eyes, head and body onto those directions. However it is also possible to attend to peripheral areas without moving the eyes. This phenomenon, termed covert attention, was first described by Helmholtz in 1894. These two types of attention are not independent and can occur simultaneously. For example, when driving it is important to have eyes directed towards the road while paying peripheral attention to abnormal movements out of focus.

In visual search tasks, covert attention and saccadic eye movements (rapid, simultaneous movements of both eyes in the same direction produced by overt attention) usually have a joint role: a salient
feature triggers the attention and through a saccade the eye fixates the region of interest enabling perception with higher resolution. However, the eccentricity effect created by the physical structure of the retina, i.e. a higher resolution in the center relative to the periphery, makes it difficult to detect elements in peripheral areas. Another effect is the inattentual blindness that can occur when an unexpected object appears in the fully visible visual field and yet is not perceived because the attention is focused on another task or object.

2.1.3 Bottom-up and Top-down Attention

According to James [13], there are two major categories of factors that drive attention to certain objects or locations: bottom-up and top-down factors. Bottom-up factors are driven by stimuli generated by discriminative features of the visual scene. If a feature stands out from its surroundings, it is automatically and involuntarily perceived. This suggests that the saliency of visual field features is formed in the brain before the focused attention itself by a pre-attentive process. Some features are intrinsically more salient in a given context, for example a black ball in the middle of white balls (the salient feature is the color). On the other hand, top-down factors are generated by a goal and are influenced by knowledge, expectations and objectives [9]. Even if objects appear in known settings, the attention driven by these factors is slower because it requires focal attention.

Yarbus presented in 1967 [1] an example of top-down attention, showing that eye movements depend on the task performed. The experiment consisted in asking individuals to observe the same scene (a room with a family in which an unexpected visit appears), but under different conditions, i.e. different questions such as "estimate the financial circumstances of the family", "how old are people?" or simply "freely examine the scene".

2.2 Psychological Theories of Visual Attention

Over the years many psychological theories have been suggested in order to explain and understand the selective visual attention mechanisms. In this section, we address the theories which have been most influential for computational systems and then we present a computational architecture mainly followed in computer vision and robotics for visual attention.

2.2.1 Feature Integration Theory

Feature Integration Theory (FIT), proposed by Treisman and Gelade [14], suggests that when perceiving a stimulus, features are "registered early, automatically, and in parallel, while objects are identified separately" and at a later stage of processing.

During the pre-attentive stage, the basic features, such as color, orientation, intensity and contrast are extracted simultaneously and registered in feature maps, through a bottom-up process. These
Figure 2.3: Most influential psychological theories for visual attention. (a) Feature Integration Theory (FIT). Basic features are recorded in parallel pre-attentively on specific topographic maps that highlight the corresponding image locations. Those features maps are combined in a single master map of locations that is scanned through focal attention allowing to obtain the perception of the object. (b) Guided Search Model. Visual scene is filtered into feature map categories according to bottom-up and top-down information. The activation map results from the weighted sum of those features maps. Then attention is directed sequentially to the regions in descending order of activation.

2.2.2 Guided Search Theory

The Guided Search Model (GS) developed by Wolfe [15], despite being complementary to FIT differs from it since it assumes the pre-attentive stage includes not only the bottom-up parallel saliency extraction registered in feature maps, but also top-down information from prior knowledge and expectations. While in the FIT feature maps correspond to specific characteristics, in GS there are feature map categories, for example, the color map combines the activation of green, red, blue, and yellow. The top-down factors trigger certain locations in the feature maps, for example in the color feature map they activated the specific color of the object due to our prior knowledge. This feature maps are weighted according to top-down factors and summed up in an activation map (Figure 2.3 (b)) that represents the likelihood that the corresponding location contains the target. Then, in descending order of activation, each peak is inspected to search for it.

2.2.3 Saliency Map Model

Most biologically inspired bottom-up computational models [16] follow the architecture proposed by Koch and Ullman [17]. This architecture (Figure 2.4), motivated by the FIT and GS theories described above, is based on the idea that the various feature maps (registered in a preattentive stage) are combined...
into a single map called saliency map. This map, analogous to the FIT master map of locations, is a two-dimensional scalar map that represents topographically the visual conspicuity in which the enhance regions of the map correspond to salient regions on the image. This theory was designed to control the mechanism of covert attention. Thus, it states that attention is focused on the location that corresponds to the maximum in this map through the Winner-Take-All (WTA) principle. Then one applies a strategy of Inhibition Of Return that allows the focus of attention to shift to the next salient region.

Figure 2.4: Koch and Ullman Saliency Map Model. The basic features are recorded in parallel in a preattentive phase in feature maps that are aggregated in a saliency map. This map corresponds to a topographic map that contains information from the various feature maps highlighting the areas of the visual field according to their relevance. The attention is directed to the location that corresponds to the maximum of this signal through the principle Winner-Take-All (WTA). The WTA is equivalent to the operator finding the maximum of the overhang map, and allows you to select the location to focus on. Then the Inhibition Of Return strategy is applied so that the focus of attention shifts to the next most salient zone 3.

---

3Source [3]
Chapter 3

Deep Learning Supporting Concepts

In recent years there have been great advances in the field of artificial intelligence provided by a new set of learning techniques known as Deep Learning. These advances, according to LeCun, Bengio and Hinton [5], were only achieved due to the development of more powerful, dedicated hardware such as Graphics Processing Unit (GPUs) and the creation of large sets of labeled data [18, 19]. Deep Learning is a useful framework for many different machine learning algorithms to model data with a high level of abstraction through the implementation of deep artificial neural networks inspired by animal brains.

Machine learning is a category of algorithms that allows computers systems to effectively perform a specific task without explicit instructions. Those algorithms are mathematical models that receive input data and based on it use statistical analysis to make predictions or decisions.

3.1 Deep Neural Networks (DNNs)

Artificial Neural Networks (ANNs) are biologically inspired models that, through the association of simple computational units called neurons, attempt to estimate the value of functions. These units were first introduced by McCulloch and Pitts [20] and their structure is shown in Figure 3.1.

The output, $y$, of these unit corresponds to the result of the Heaviside activation function given by

$$H(n) = \begin{cases} 
0, & \text{if } n < \theta \\
1, & \text{if } n \geq \theta 
\end{cases}$$

(3.1)

which sets the output to 1 or 0 depending on the value of the threshold. This function is applied to the dot product between the inputs, $x_i$, and its weights, $w_i$, summed with a bias, $b$. Nowadays, other options exist for the activation function such as linear, sigmoid, tanh, ReLu, depending on the purpose of the network.
Figure 3.1: Artificial Neuron structure. The output of the unit, $y$, corresponds to the linear combination of inputs, $x_i$, respectively weighted by $w_i$, and summed with a bias $b$. The result is activated by an activation function, $f$.

Figure 3.2: Scheme of an artificial neural network with two fully connected intermediate layers.

An ANN is a collection of neuron units (Figure 3.2) that transmit a signal by transforming it through a series of hidden layers. In each hidden layer all neurons are fully connected to the ones in the previous layer and are completely independent, being able to perform different activation functions. The last fully-connected layer is called the “output layer” and in classification settings it represents the class scores. A DNN is an ANN with multiple layers between the input and the output layers. From layer to layer there is an increase in the abstraction level of the input data representation.

Each connection between units has a weight that is adjusted by the learning process and can increase or decrease the signal strength. During the training process, weights gradually converge to certain values, so inputs produce the desired outputs. This network training can be supervised or unsupervised:

- In supervised learning we have a set of correspondences between input and output vectors that form the training set. Each input vector of this set is fed into the network and, according to a cost function, the error depends on the difference between the output obtained and the desired one is calculated. The internal weights of the network are adjusted to minimize the value of the cost function.
In unsupervised learning we have the sequence of inputs but not the desired outputs. The network can only learn intrinsically to characterize the distribution or structure of the data.

The choice of the error function depends on factors such as the learning type (supervised, unsupervised, reinforcement) and the activation function. For example, when performing supervised learning on a multiclass classification problem, the activation function of the last layer is typically a $\text{softmax}$ and the natural error function that derives from it is the cross entropy function $[21]$. The $\text{softmax}$ function is defined as

$$
p_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)} \quad (3.2)
$$

where $p_j$ is the class probability from unit $j$ computed with its input $x_j$ over all $k$ index classes. Cross entropy is defined as

$$
C = - \sum_j d_j \log(p_j) \quad (3.3)
$$

where $d_j$ represents the desired probability for unit $j$ and $p_j$ the probability given by the $\text{softmax}$ output.

### 3.1.1 Backpropagation algorithm

Backpropagation $[22, 23]$ is an algorithm for supervised learning of feedforward neural networks using the gradient descent method. The goal of this algorithm is: given an ANN and a training set, find the network parameters, $w$, that minimize the error function. That being said the algorithm requires three aspects:

- a training dataset consisting of a set with $N$ input-output pairs $X = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $y_i$ is the desired output associated with the input sample $x_i$.

- a feedforward neural network with a set of parameters $w$ composed by the weights between nodes and biases. It includes the parameters $\omega_{ij}^k$, that corresponds to the weight between node $j$ in layer $k$ and node $i$ in layer $k - 1$, and the bias $b_i^k$ for node $i$ in layer $k$ will be denoted also as $\omega_{0i}^k$.

- an error function $E(X, w)$ that defines the difference between the desired output $y_i$ and the output obtained $\hat{y}_i$ from a feedforward pass with the input $x_i$ and a particular set of parameters $w$.

Training an ANN with the gradient descent algorithm requires the computation of the gradient of the error function $E(X, w)$ with respect to parameters $\omega_{ij}^k$ at each iteration. Thus, it is necessary to guarantee the continuity and differentiability of the error function. This can be achieved by choosing proper activation functions since the outputs of the network is just a composition of these functions. Then, through the gradient descent method the weights can be updated according to

$$
w^{n+1} = w^n - \eta \frac{\partial E(X, w^n)}{\partial w^n} \quad (3.4)
$$
by adding an increment, weighted by the learning rate, $\eta$, to the parameters $w$ of the previous iteration $n$, in the opposed direction of the gradient.

The backpropagation algorithm proceeds in the following steps, assuming a random initialization of the parameters $\omega_{k}^{ij}$. First, there is a feedforward phase to calculate the outputs $o_{k}^{j}$ of each unit $j$ in layer $k$ and therefore the network output $\hat{y}_{d}$, for each input-output pair $(x_{d}, y_{d})$. Then, in the backward phase, one calculates the partial derivatives $\frac{\partial E_{d}}{\partial \omega_{k}^{ij}}$. The cumulative result of the backpropagation node $j$ of layer $k$ is called backpropagation error and is given by $\delta_{k}^{j}$. Through the backpropagation rule

$$\frac{\partial E_{d}}{\partial \omega_{k}^{ij}} = o_{k}^{j} - 1 \times \delta_{k}^{j}$$

one obtains the partial derivatives of the error function for each weight $\omega_{k}^{ij}$ connecting node $i$ in layer $k - 1$ to node $j$ in layer $k$. Then, one has to combine the individual gradients for each pair $\frac{\partial E_{d}}{\partial \omega_{k}^{ij}}$ to get the total gradient $\frac{\partial E(X, w)}{\partial \omega_{k}^{ij}}$. Finally, through the gradient descent equation (3.4) the weights are updated.

### 3.1.2 Overfitting

One of the biggest challenges when training a neural network is overfitting [24]. This phenomenon consists in having a very small error in the training set, however when evaluating the network with new data it performs very poorly. This happens because the network has memorized the training examples, but it has not learned to generalize to new situations. This is an even more significant issue in deep learning, where neural networks have large numbers of layers containing many neurons. To detect if there is overfitting, one can split the initial dataset into separate training and test subsets. Thus one can perform the training process with the training set and then evaluate the network performance with the test set.

Additionally, in order to prevent overfitting, one can use techniques such as cross validation [25], early stopping, regularization (L1/L2 regularization [26] or dropout [27]). Dropout [27] proposed by Srivastava consists in ignoring, at each training iteration, some connections of the network according to a predefined probability. This forces the weights of other connections to interact with different random sample units. After each training iteration, ignored units are reset to their original weight. This is the most used regularization technique because it not only allows the network to learn more robust parameters that generalize better new data, but it also allows for a significant improvement of the training speed.

### 3.2 DNNs in Image Classification

Image classification is a classical computer vision task tackled by this kind of algorithms. It consists of assigning a class or probability of belonging to a class to an input image from a set of fixed classes. This task can be very challenging due to the amount of intra-class variability. This problem can be mitigated by using DNNs, instead of manually selected abstraction features (such as edges, textures, colors), since they can learn to extract high-level features with increasing abstraction as the number of layers
Figure 3.3: Example architecture of a CNN composed by: four convolutional layers that extract specific features in features activation maps; two Pooling layers where the size of the input volume is reduced; and a last fully connected layer that returns the output vector of the CNN.

However, DNNs are characterized for being over-parameterized, requiring a huge amount of labelled data. This data acquisition process often requires manually annotations that can be exhausting and expensive. Thus, in some computer vision problems such as object detection, segmentation, key points detection, performing a weakly supervised learning process from weaker annotations can be useful, for instance localize object bounding boxes only with image class labels (without any bounding box annotations).

3.2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a class of deep artificial neural networks that are biologically inspired on the visual cortex of mammals. These architectures were first introduced with Neocognitron [28] and have been widely applied in computer vision tasks such as object classification [29] and object detection [30].

The cells of the visual cortex [31] are sensitive to small regions of the visual field acting as local filters, allowing the exploration of local spatial correlations. Since in natural images adjacent pixels are more correlated than those that are further apart it would be impractical to have fully connected layers. This is justified due to every neuron is connected to all neurons in the previous, since such architecture does not take the spatial structure of images into account.

CNN architectures exploit spatially local correlations by enforcing a sparse local connectivity pattern between neurons of adjacent layers: each neuron is connected to only a small region, the receptive field, of the previous volume. Another advantage of these architectures is parameter sharing that allows a reduction of the number of training parameters and contributes to translational invariance. The intuition behind this is the fact that if a feature detector is helpful in one part of the image, it is also probably useful in another part of the image. Thus parameters are shared along the spatial dimensions (width and height) of the representation.
Normally CNNs input are image data in the form of multiple 2D vectors containing pixel intensities, therefore neurons of each layer are arranged to process volumes of 3 dimensions: width, height and depth (number of color channels). The architecture of a CNN (Figure 3.3) is typically a stack of different types of layers that apply different transformations to the representation with the objective of extracting task-specific features. Usually the layers are as follows:

- **Convolutional Layer**: It is the principal component of a CNN architecture (Figure 3.4 (a)). It is composed of filters with a small receptive field, to extract specific features. Depending on the hyperparameters such as number of filters, kernel size, stride and padding, convolutions are computed from the dot product between the entries of the filter and the input volume thus producing a 2-dimensional activation map of that filter.

- **ReLu Layer**: This layer have a linear rectifier activation function, given by \( \max(0, x) \), after the convolution in order to introduce a nonlinearity to the results obtained.

- **Pooling Layer**: The purpose of it is to progressively downsample the representation along spatial dimensions (width and height) to reduce the amount of parameters and computation in the network, and hence controlling overfitting. This layer is periodically inserted in-between successive convolutional layers. The most used types of Pooling are: the Average Pooling, which calculates the average of each region; and the Max Pooling which determines the maximum present in a neighborhood. (Figure 3.4 (b)).

- **Fully Connected Layer**: In this layer each neuron is connected to all activations in the previous layer. This is typically placed at the end of the network to compute the class scores of the image. By assigning a **softmax** activation function the outputs can be interpreted as probabilities of the initial image belonging to a certain class.

![Figure 3.4: Layout of Convolutional and Pooling layers. (a) Layout of a Convolutional layer with 12 filters (b) Numeric example of the representation spatial reduction performed by the Average Pooling and the Max Pooling.](image)

Some of the most used CNN architectures [32] are inspired by the LeNet [23] network that follows the simple stacking structure of 7 convolutional layers interspersed with a pooling layer and then a last fully connected layer to perform the final classification.
Chapter 4

Related Work

Salient object detection in an image is a task performed by human visual attention mechanisms. These mechanisms orient the gaze to regions that stand out from the environment in the visual field.

Humans in a pre-attentive stage are able to detect visually distinctive regions in an effortless and parallel way across the entire visual field. These filtered regions, so called salient, are then, in a focused attention stage, inspected in a finer scale in order to extract high-level information and perceive objects present in the scene. These visual attention mechanisms have been studied not only by cognitive scientists, but also, more recently by the computer vision community. This interest is due to the fact they allow to find salient objects that can represent a scene efficiently, responding to complex vision problems.

4.1 Convolutional Neural Networks for Salient Object Detection

Convolutional Neural Networks (CNNs) are able to provide high-level and multiscale feature representations of the input. Thus they have been applied to many computer vision problems such as object recognition [29, 33], semantic segmentation [34], edge detection [30], among others. Recently, it has been shown that CNNs are also efficient in the task of salient object detection [35]. This is because they are capable of capturing the most salient regions of an image without having any prior knowledge about it, such as segmentation-level information, thanks to their feature extraction quality.

4.1.1 Weakly Supervised Object Detection

Weakly supervised learning is a way of learning supervised models when there is incomplete ground-truth training data. Because it is very time consuming to manually do bounding boxes annotations this weakly supervised training has become an emerging research topic in the task of object detection. In this case, a model is trained to predicted the location of the bounding-boxes only with image-level labels
that identify the presence or absence of objects [36].

Simonyan et al. [37] proposed a weakly supervised object segmentation way of obtaining the object bounding boxes in a image. Having a pre-trained CNN the method is based on obtaining a relation between category score and the input image. That is made through the computation and then back-propagation of the error gradient of the class score until the input layer. Therefore, they obtain a class saliency map that represents the importance of each pixel to predict the class backpropagated.

Cao et al. proposed a method called Look and Think Twice [38] which had the objective to detect and localize different objects in cluttered images. The method was inspired by the way of generating saliency maps proposed by the previous work. Thus, by making use of a pre-trained CNN, after the first glimpse in a coarse scale the regions of interest are cropped in rectangular patches and than inspect again in a more fine scale. In order to obtained the patch crops they perform two passages in the network, a firsts feedforward pass to make a first prediction of the class labels that might be present in the scene. Then, based on the set of the most probable object class labels predicted, the method computes imagespecific saliency maps for each of the labels. Next, a segmentation mask is computed by thresholding the saliency map. Therefore, in a weakly supervised object detection process the bounding boxes of the possible object locations are the up-right rectangles that contain all the activated points in the segmentation mask. Finally, a second feedforward is performed with the bounding boxes that were cropped in order to make a re-classification.

4.2 Foveal Attention Mechanisms

Foveal sampling can be a way of reducing the information encoded in the image when compared to the typical uniform sampling at maximum resolution. This reduction increases as the field of view and the maximum resolution increases.

The term foveal, as explained in chapter 2, comes from human-vision anatomical inspiration, where the sampling resolution is higher in the center small region of the eye - the fovea - and decays towards the periphery.

4.2.1 Multiscale Foveation

Objects in the real world can be perceived in a continuous domain of scales that depend on their own size or the distance to which the observer is. This feature makes it useful to study representations that can encode several scales. One of those is called multiscale foveation that tries to simulate the foveation mechanism. It corresponds to a class of methods that can decomposed an image in a composition of different scale representations.
The foveated pyramid is based on the standard Laplacian pyramid way of encoding an image proposed in [39]. This technique is used for image compression where the differences between successive higher blurred image-levels, generated in the Gaussian pyramid, are saved. In the foveated pyramid each level is weighting by a window of size of the topmost level, so that pixels outside that window can be neglected. Thus, as the level of the pyramid increases the resolution of the image decreases. In this way, the spatial non-uniform way of sampling the field of view is mimicked, simulating a foveated image.

Geisler et al. [40] applied this foveated architecture to video compression in order to transmit and visualize images when the bandwidth is reduced. They allow a moving fovea, whose foveation point is obtained from the user-controlled by a pointing device (e.g., a mouse or an eyetracker). Due to rectangular patch cropping, i.e. rectangular weighting windows with binary values, some spatial edged artifacts can arise. Thus, to eliminate this problem they proposed not only a smooth decay in each pyramid level obtained by a “foveation point interpolation” but also, a raised-cosine blending across levels of the pyramid.
Chapter 5

Methodologies

In this chapter we will present two methodologies for object recognition, anatomically and physiologically inspired by the human visual attention mechanism. They are influenced by the FIT, already described in section 2.2. Briefly, it suggests that object recognition is made in two stages: (1) a pre-attentive stage where the features are perceived automatically and in parallel across the visual field followed by (2) a focused attention stage to sequentially process the stimuli locations to identify the objects present in the scene.

We proposed a Foveal Saliency (FS) method, that we will described in section 5.1, capable of integrating the two described psychological stages. The FS method extracts salient features and then orients the focus to those regions of interest. Then, we will detailed, in section 5.2, our Pyramidal Focus (PF) method that increases the recognition accuracy when the visual system finally fixates a target.

To mimic the human visual attention these two methods are based on pre-trained CNN architectures to help recognize the object. However, these CNN architectures are not trained end-to-end. The models take advantage of pre-trained CNN architectures that are already able to classify thousands of objects [32, 41, 42]. Distinct to the traditional way of using these networks, we employ different artificial foveal systems to replicate the human anatomy of the retina. Although these two methods are presented here in a dissociated way, they represent a proof of concept of an integrated biomimetic solution for detecting object classes present in images. We assume that the images have only one object of considerable dimensions and for this reason we denoted our work detection of salient objects.

5.1 Foveal Saliency Method

The FS method used in this work corresponds to the one previously developed by us in [6]. The correspondent scheme, presented in Figure. 5.1, can be decomposed into two phases. First, in a bottom-up fashion, the model performs a first object classification, through a feedfoward CNN pass with a foveated image created by an artificial foveal visual system. Then, in a weakly supervised process, through
feature saliency extraction, the model computes the next foveation point by performing a feedback propagation in the CNN according to top-down information. These two phases are sequentially repeated to refine the object recognition task.

One of the bottlenecks of classical saliency models is feature selection. The amount of features that exist to be extracted is overwhelming and choosing just a few to have a generalized overhang prediction is quite complicated. For example, Itti's model [43] extracts pre-defined features through linear filtering such as colors, intensity and orientations that are combined in order to identify the salient location of the object. To overcome this bottleneck, feature selection is employed through a pre-trained CNN model that has learned to extract high-level features in order to produce a more abstract representation of the input.

This methodology is an active vision system capable of improving object detection and recognition that tries to replicate the human saccadic mechanism and the foveal eye anatomy. It was mainly based on the work developed previously by Almeida [44] that was inspired in the object recognition process proposed in Look and Think Twice by Cao [38].

Cao's method attempts to mimic the human visual strategy described in FIT, where after the first glimpse of a scene some relevant parts are inspected with focused attention to identify the object. The process consists in first analysing the full image at a coarse scale through feedforward CNN classification to obtain a set of the most probable object classes. Then in a weakly supervised fashion by feedback propagation of each top predicted classes, the salient regions of the possible object locations are patch cropped and re-classified in a zoomed-in scale. However, when recognizing objects with attention it is important to have some context of a scene, so in Almeida's work [45] instead of cropping a bounding box around the object, a human like foveated image centered on the object is presented. In this latter work the re-classification scheme proposed in the former is also used.

5.1.1 Artificial Foveal Visual System

The purpose of this artificial foveal visual system is to generate foveated images that are able to replicate the non-uniform distribution of the receptive fields in human eyes. This is represented by the first block in Figure 5.1. It was inspired by the Laplacian Pyramid technique proposed by Burt and Adelson [39] for image compression, which is extremely fast and easy to implement and has been applied to real-time image processing and pattern recognition.
First, the artificial foveal visual system begins by generating a Gaussian Pyramid. A Gaussian Pyramid is a collection of images, all arising from a single original image, that are successively lowpass filtered using a Gaussian average (gaussian blur). Each subsequent image level, \( g_{k+1} \), is filtered through the convolution of the previous level, \( g_k \), with a \( 5 \times 5 \) Gaussian mask:

\[
g_{k+1}(u,v) = \omega * g_k(u,v) = \sum_{s=-2}^{2} \sum_{t=-2}^{2} \omega(s,t) g_k(u-s,v-t)
\]  

(5.1)

where \( u \) and \( v \) are the image level coordinates. The Gaussian mask \( \omega \) is computed from the isotropic 2D Gaussian kernel \( G(x,y,\sigma) \) with zero mean:

\[
G(x,y,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}}
\]  

(5.2)

where \( \sigma \) is the Gaussian standard deviation and also corresponds to the cut-off frequency of the filtering. The size of the mask influences the amount of blur: a larger size corresponds to a larger convolution mask and thus to a greater degree of filtering. The Gaussian mask is obtained by discretizing the Gaussian kernel. This theoretically requires an infinitely large convolution kernel, as the Gaussian distribution is non-zero everywhere. Fortunately, the distribution decays abruptly, approaching zero about three standard deviations from the mean. Thus, one can limit the kernel size to contain only values within that interval. Therefore, the discrete approximation of a Gaussian function with \( \sigma = 1 \), presented in Figure 5.2, is a \( 5 \times 5 \) mask.
Since the 2D Gaussian kernel is isotropic, i.e. circularly symmetric, it is separable into $x$ and $y$ components. This means that the 2D convolution can be decomposed in a series of two 1D Gaussian convolutions, one in the horizontal direction and one in the vertical direction. This represents a significant reduction in the computational cost. Also, the iterative pyramid generation is equivalent to convolving the original image with gaussian kernels with $\sigma_k = 2^{k-1}\sigma_1$ where $k$ is the level of the pyramid and $\sigma_1$ is the standard deviation from the first level. In this way, the frequency and therefore the sample interval is reduced by an octave from level to level. This results in a reduction of each image resolution by a factor of 2 along each coordinate direction.

Secondly, the Laplacian pyramid levels are computed as the differences between adjacent Gaussian levels. They represent the error images and result from the up sample of the Gaussian images in order to have compatible resolution to enable the computation of pixel-wise differences.

In the next step, to mimic human vision with a high resolution in a fovea of size $f_0$ and a lower resolution in the rest of the retina, each Laplacian level is multiplied by an exponential weighting kernel of the form

$$k(u, v, f_k) = e^{-\frac{(u-u_0)^2 + (v-v_0)^2}{2f_k^2}}$$

(5.3)

where $f_k = 2^k f_0$ is the exponential kernel standard deviation at the $k$-th level and $(u_0, v_0)$ is the foveation point which defines the focus of attention. Finally by expanding and summing all the resulting images the foveated image is created (Figure 5.3).
5.1.2 Iterative Weakly Supervised Object Detection

In our model we use a weakly supervised detection process described in [38] since it is a very efficient way of obtaining a possible object location. The process uses class-specific saliency maps to encode the possible location of the object of a given class in an image, and thus can be used for object detection. These saliency maps are extracted through a single backward pass in a CNN trained only on the image labels, so no additional annotations are required (such as object bounding boxes or segmentation masks).

Image-Specific Class Saliency Extraction

According to Simonyan’s findings [37], it is possible to obtain an image-specific class saliency map via a feedback propagation in the network. This part of the method is present in the second block of the scheme in Figure 5.1. Given an image $I$ and a class $c$, the CNN classification output score $S_c(I)$ is highly non-linear, therefore it is useful to linearize it through a first-order Taylor approximation in the neighborhood of $I$ as

$$S_c(I) \approx G_c^T I + b$$  \hspace{1cm} (5.4)

where $b$ is the bias of the model. The term $G_c = \frac{\partial S_c(I)}{\partial I}$ can be viewed as a measure of how likely it is that pixels of image $I$ are important for the classification of a class $c$ and therefore can help detect that class in the image. The pixel derivatives are found by back-propagation until the first input layer, that correspond to the input image. The back-propagated error values are the difference between the output of the CNN softmax layer and the desired output vector, that corresponds to assigning 1 to the element associated with the specific class we want to detect and 0 to all the other inputs. $G_c$ defines the class specific saliency map of the image $I$.

Since the images used are RGB, a single class saliency value for each pixel $M_c(i,j)$ is obtained by
taking the maximum magnitude of $G_c$ across all colour channels $l$,

$$M_c(i,j) = \max_{l \in \{\text{rgb}\}} |G_c(i,j,l)|. \quad (5.5)$$

**Next Foveation Point Control**

The next foveation point consists in the center of the possible object location and the foveation is made on the original image. This is shown in the last block (bottom-right) of Figure 5.1. The object bounding box is obtained by computing the segmentation mask by selecting the pixels of the saliency map $M_c$ with a value higher than a certain threshold, and setting the rest of the pixels to zero. Thus, one is able to define the tightest bounding box covering the stain of non-zero saliency values, obtaining a guess of the localization of the object. This bounding box corresponds to the minimal up-right rectangle for the set of non-zero saliency values. The class score prediction of each iteration is the score vector obtained after running the CNN model.

### 5.2 Pyramidal Focus Method

The PF method used in this work follows the one previously developed by us in [7]. Assuming the target was already fixated and centered in the visual field, this method corresponds to the focused attention stage described in FIT that allows for an improvement of the performance on the object recognition task. This architecture takes advantage of the human-inspired foveal vision based on low-pass multiresolution pyramids [40].

#### 5.2.1 Artificial Foveal Visual System

The PF model is achieved by Gaussian pyramid coding [39], already explained in section 5.1. In order to have more levels for each image, each successive pyramid level contains half of the pixel number of the previous level. This means that the downsampling factor is $\sqrt{2}$ in each coordinate direction, and not the typical downsample factor of 2 used in Gaussian Pyramids. Considering an $n \times m$ input image, the number of Gaussian pyramid levels, $N_L$, is given by the maximum number of times one can reduce the image by a factor of $\sqrt{2}$ in the smaller image dimension, until it reaches a value $N$, where $N \times N$ is the size of the input layer of the CNN pre-trained architecture. Formally, it is represented by

$$N_L = \min \left[ \max \{N_L \in \mathbb{Z} : n2^{-N_L/2} \leq N\}, \max \{N_L \in \mathbb{Z} : m2^{-N_L/2} \leq N\} \right]. \quad (5.6)$$

To mimic the human foveal vision each pyramid level is patch cropped in the center with the input size of the network. In this way, the first level of the pyramid represents the fovea with the same resolution as the input image, that decreases as the pyramid level increases (Figure 5.4).
Figure 5.4: Pyramidal Focus method scheme for object recognition. The first level ($l_1$) is the original image, and each level downscales the image by a factor of 2. Then each image level is cropped in the center with the same input size of the CNN architecture and loaded into the pre-trained network. The output probabilities of the softmax layer from each pyramid crop are averaged through EMA to obtain a final joint prediction.

### 5.2.2 Joint Object Recognition

In order to identify the object present in the image, each pyramid crop is loaded into a pre-trained network and the output probabilities of the softmax layer from each pyramid crop are averaged with an exponential moving average (EMA) to obtain a final joint prediction. EMA is applied recursively to each increasing level of the pyramid $n$ until it reaches the level $N_L$ according to

$$S(n) = \begin{cases} 
Y_1, & n = 1 \\
\alpha Y_n + (1 - \alpha) S_{n-1}, & n \geq 1 
\end{cases}$$

(5.7)

where $Y_n$ is the score vector of the prediction with the crop of the pyramid level $n$ and $S(N_L)$ is the score vector of the model prediction of each image. The coefficient $\alpha$ represents the degree of weighting decrease, a constant smoothing factor between 0 and 1. The greater the constant $\alpha$ the greater the penalty of the scores of the prediction with the crop of the lower pyramid level. Therefore, higher levels of the pyramid that have more context information have more emphasis on the final prediction than lower levels. This implementation is better than a simple average filtering mainly for two reasons: (1) the context, since higher levels contain more context information useful for recognition then lower levels; (2) the object size, since for larger objects the crops from lower levels of the pyramid in some images do not catch the entire object.
Chapter 6

Experiments

Both Foveal Saliency (FS) and Pyramidal Focus (PF) methods are primarily intended to detect and recognize salient objects in an image. In this chapter, we will address some implementation issues and the tests applied to our models in order to evaluate their classification performance and computational cost. Therefore, in section 6.4 we will detail the aspects concerning FS method and in section 6.5 the ones concerning the PF method. We will also outline the evaluation metrics in section 6.1 and the dataset used to perform the tests in section 6.2. Also, the baseline solution with which we will compare our results is explained in section 6.3.

6.1 Evaluation Metrics

To evaluate the performance of our models we compute the top-1 classification error according to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) of 2012 metrics [18]. For each image we produce a list of at most \( n \) object class labels in descending order of confidence. If we consider \( n > 1 \) the model is not penalized for detecting multiple objects in the image that are in fact present but not included in the ground truth class labels. Thus, for each image we consider an error if the ground truth does not match with any of the top-\( n \) class labels predicted. Formally, for an image with a single ground truth class label \( C_i \), there are \( n \) associated predictions \( c_{i1}, \ldots, c_{in} \) that are considered correct if \( c_{ij} = C_i \) for some \( j \) in \( j = 1 \ldots n \). Then let the error of a prediction \( d_{ij} = d(c_{ij}, C_i) \) be 1 if \( c_{ij} \neq C_i \) and 0 otherwise. The classification error of a model is given by the average error over all \( N \) images

\[
\text{Error} = \frac{1}{N} \sum_{i=1}^{N} \min_j d_{ij} .
\] (6.1)

We will also estimate the computational cost associated with the object recognition. We define it as the amount of information, expressed simply by the number of bytes encoded in the image, that is going to be processed. This estimation is necessary to evaluate the trade-off between the performance and the computational cost associated with the algorithms. What we mean by this is that we want to first understand the amount of information processed by each model is greater or not than the information
Figure 6.1: Baseline architecture. Typical scheme of the image classification task with a uniform image resolution in a generic CNN architecture.

processed by the baseline solution. Secondly, in case it is greater, if we can achieve better classification performance.

6.2 Dataset

The dataset used to perform our tests was a partition of the ILSVRC 2012 validation dataset [18]. ImageNet is a large visual dataset of over 15 million labeled images that are part of about 22 thousand categories. The annual challenge started in 2010 and uses a subset of ImageNet formed by around 1000 images in each of the 1000 categories. The ILSVRC 2012 dataset was previously divided into training, validation and test images. The validation and test data consist of 50K and 100K images hand labeled but only validation labels data were released. These 150K images (validation and test) were not part of the training data that is formed by 1.2 million images containing the 1000 categories.

The size of the images that can be part of this partition is constrained by the PF model properties, since the method needs to generate at least two levels of the Gaussian Pyramid with a downscale factor of $\sqrt{2}$ in each dimension. So, in order to have consistent results across models and to be able to compare them, our dataset partition is composed of the 7100 images selected from the ILSVRC 2012 validation dataset that have a resolution greater than $454 \times 454$.

6.3 Baseline Solution

The baseline solution is the CNN architecture with which we compare our methodologies’ performance. As explained in chapter 5, our models integrate pre-trained CNN architectures. For matters of consistency, we chose to use always a GoogleNet [41, 46]. Thus, our baseline is a single traditional feedfoward pass in that network for an image with uniform resolution. The original image has to be first resized to the resolution of the input layer which is $227 \times 227$.

GoogLeNet is the winner of the ILSVRC of 2014 achieving a top-5 error of 6.67%. This network is 22 layers deep organized in modules denoted by Inception modules that allow parallel operations such as pooling and convolutions. These modules help reduce the number of network parameters to
4 million. Moreover, the last layer performs an average pooling instead of being fully connected, which also eliminates a large number of parameters.

6.4 Foveal Saliency Method

In this section we will present the tests performed to the FS model in order to analyse which parameters affect the object recognition performance. We begin by describing the steps of which the model consists:

1. Resize the original image to the CNN architecture input size.

2. Foveate the resized image centered in a foveation point \((u_0, v_0)\) with a certain fovea size \(f_0\) via the artificial foveal visual system.

3. Run the CNN model with the foveated image and predict the top \(k\) class labels with a feedforward pass.

4. For each of the top \(k\) class labels, compute each object's bounding boxes with top-down backpropagation.

5. For each of the \(k\) object location proposals foveate the original image with the foveation point in the center of each bounding box and predict again the top \(k\) class labels with a feedforward pass.

6. Given the total \(k \times k\) labels and the corresponding confidences, sort them in descending order and choose the top \(k\) as final prediction.

7. Repeat from 4.

In the implementation phase we had to choose the values of some hyperparameters. In the artificial foveal visual system we set the Gaussian standard deviation of the first level as \(\sigma_1 = 1\) and the number of levels to be five since these values generate a good decay of the blur employed around the foveation point. Then, differing from the previous work [44], where the first foveation point \((u_0, v_0)\) was fixed to be on the image center, we made it a free parameter. In this way we are able to identify objects laying not only on the center but also in the periphery of the image, thus mimicking human-like vision. The fovea size, \(f_0\), was also set to be a free parameter. In Figure 6.2 we represent different resulting images from our artificial foveal visual system with different fovea sizes \(f_0\) and foveation point \((u_0, v_0)\).

For the weakly supervised detection procedure, we had to apply a threshold to the image-specific class saliency map in order to obtain a segmentation mask. Since this parameter does not influence the classification error we set it to \(\theta = 0.65\). This value was chosen because it is the one that generates a better compromise between the number of salient pixels in the saliency map and the correct bounding box (discussed in [44]). Also, we chose to do the backpropagation phase with the top-5 class labels, setting \(k = 5\). Therefore, we obtain the five possible bounding boxes associated with each class label.
In Figure 6.2 we present the weakly supervised detection for an image whose ground truth is *red fox*. We show the image-specific saliency map, followed by the segmentation mask, and the bounding box of the possible object locations. This was obtained via backpropagation of the top five class labels from the first prediction (*kit fox, red fox, grey fox, arctic fox* and *coyote*) without applying the artificial foveal system to the input image. As explained in section 5.1.2 the saliency map highlights the pixels in the image that are more important for classifying the class that was backpropagate. Therefore, each saliency map and respective segmentation mask is different. Since the ground truth label is *red fox* the second image is the one that has more pixels highlighted in the snout, showing that those features are very relevant for intra-class distinction.

The experiments performed had the goals of analysing the effect of the foveation point \((u_0, v_0)\), the fovea size \(f_0\) and the iteration number on the classification performance of the model. In order to understand how the foveation point of the first feedforward pass influences the classification error, we made it vary along a 11 by 11 grid on the resized image domain. Therefore, for each foveation point we foveate with different fovea sizes, compute the classification error, and then average it. Afterwards, we do this for all images and average again, obtaining the classification error for one foveation point. Finally, we repeat this process for all the points considered on the grid.

To analyse how the fovea size affects the classification error we varied \(f_0\) between 0 and 130 and we ran the model for 6 iterations for \((u_0, v_0) = (113, 113)\) and other 16 foveation points spread in a grid over the resized image. In the previous work [44], the foveation point was fixed to be in the image center, but in natural vision tasks objects can be located in the periphery. As our model iteratively converges to the possible object locations we want to compare the performance for the first foveation point in \((u_0, v_0) = (113, 113)\) against the other chosen positions. For this latter case the classification error is an average over the 16 different foveation positions.
6.5 Pyramidal Focus Method

In this section we will present the tests performed to the PF model in order to analyse which parameters affect the object recognition performance. We begin by describing the steps in which the model consists:

1. Compute $N_L$ Gaussian pyramid levels according to the image resolution.

2. Patch crop each pyramid level image in the center with the size of the CNN input layer. This pyramidal representation consists of the foveated image.

3. Run the CNN model for each $N_L$ cropped images.

4. Average the output probabilities obtained from the feedforward classification with each image crop to obtain a joint prediction.

We begin by presenting some aspects of the implementation phase. In the joint object recognition part we set the weighting factor of the exponential moving average $\alpha = 0.6$. In Fig. 6.4 we show three examples of the PF method whose ground truth labels are *trimaran*, *rock beauty* and *tusker*. On the left we present the resized images followed by the patch crops generated by the model. The difference in the number of columns, as explained in section 5.2, is due to the original image dimensions that provide a different number of crops, according to equation (5.6). Therefore, in table 6.1 we present from which image dimensions are the $N_L$ pyramid levels generated.
Figure 6.4: Three examples of our Pyramid Focus method whose ground truth labels are *trimaran*, *rock beauty* and *tusker*. On the left are the resized images followed by the patch crops generated by the model. The difference in the number of columns is due to the original image dimensions that provide different number of crops.

<table>
<thead>
<tr>
<th>$N_L$</th>
<th>$n \times m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>454 $\times$ 454</td>
</tr>
<tr>
<td>3</td>
<td>643 $\times$ 643</td>
</tr>
<tr>
<td>4</td>
<td>908 $\times$ 908</td>
</tr>
<tr>
<td>5</td>
<td>1285 $\times$ 1258</td>
</tr>
</tbody>
</table>

Table 6.1: Correspondence of the smallest images dimensions that can generate $N_L$ pyramidal levels.

One of the experiments performed on the model had the objective of analysing the relation between the classification performance and the number of pyramid levels generated. We also study the behavior of our model relative to the ratio between the object size and the image resolution, where the object size is considered to be the mean areas of the ground-truth bounding boxes. In Figure 6.5, we show the distribution of the used dataset characteristics. In Figure 6.5 (a) there is the distribution of the images according to their dimensions and the number of pyramid levels that can be generated. In Figure 6.5 (b) we show the number of images as a function of the ratio between the object size and the image resolution.

The PF attention works under certain conditions. We assume that images have already been submitted to an active vision system in order to focus the fovea on the region of interest. Thus the object can be processed in detail with a focused attention that is replicated with this method. This method represents an improvement on the classification performance when compared to the previous one. However, there is a trade off between the improved performance and the increase in the computational cost that must be studied.
Figure 6.5: ImageNet partition dataset characteristics. (a) Distribution of the image dimensions. (b) Distribution of the ratio of the object ground truth bounding box/image resolution.
Chapter 7

Results

In this chapter it will be presented the results of the experiments detailed in chapter 6. Therefore, in section 7.1 we will discuss the results concerning the Foveal Saliency (FS) method. We will begin by analysing the classification performance according to the initial foveation point, the fovea size and the iteration number. We will also discuss the trade off between the classification performance and the computational cost. Then, in section 7.2 we will address the results for Pyramidal Focus (PF) method with respect to different aspects such as the number of pyramid levels and the ratio between the object size and the image resolution. We will also study the computational cost trade-off.

7.1 Foveal Saliency Method

7.1.1 Foveation Point

In Figure 7.2 we show the classification performance both for top-1 and top-5 metrics as a function of the foveation point for the first and second iterations. The ImageNet dataset is mostly composed of images with centered objects, as is demonstrated by the Figure 7.1, where the distribution of the center of the bounding boxes in the resized image domain is. Therefore, the classification error is minimal in the center, and it increases with the distance to it. Also, from the first feedforward classification to the second we verify that there is an improvement in the performance and even the first foveation points on the corners manage to achieve better recognition error in the second iteration.

7.1.2 Fovea Size

In Figure 7.3 we show both top-1 and top-5 classification errors as a function of the fovea size, \( f_0 \), for 6 iterations of the model. We also compare the results considering just a single first foveation point on the image center \((u_0, v_0) = (113, 113)\) (Figure 7.3 (a, b))) and considering a set of several points spread over the resized image domain (Figure 7.3 (c, d)). The shadow on those bottom graphs represents the uncertainty associated to those points. Moreover, there is the baseline solution error for the dataset, which we include in order to understand if our model outperforms it.
Figure 7.1: ImageNet partition dataset distribution of the center of the bounding boxes in the resized image domain.

Figure 7.2: Classification Error (in %) as a function of the initial foveation point \((u_0, v_0)\). For each point the error was computed over all \(f_0\). On the top are the top-1 and top-5 errors for the first foveation, whereas on the bottom are the top-1 and top-5 errors for the second foveation.
On one hand, one of the first aspects we can infer from all the graphs shown is that the error decreases as the fovea size increases, reaching the baseline performance. This was an expected result since smaller fovea sizes represent more blur in the foveated image and therefore less context in the recognition. On the other hand, as was already discussed in [6, 44], analyzing the graphs of the center foveation (Fig. 7.5 (a, b)), it appears that there is practically no reduction of the classification error over iterations, whereas when considering a free initial foveation (Fig. 7.5 (c, d)) we verify that our model has a significant gain from the first to the second iteration. If the first foveation is not focused on the location of the object (which in real life is the case, because when we want to find an object we do not know where it is) our model is useful, as it shows the importance of having an active vision system. Also, another way of demonstrating the need for an active vision in this scenario is the fact that with just a single iteration we could not reach the baseline.

From the first iteration, independent of the foveation point, a saliency map is extracted. This map contain important features of images regardless of the applied blur. Although these features are not fine enough to obtain a correct recognition, they are sufficient to obtain a possible location of the object. Therefore, in a second iteration, we already have the foveation point focused on the possible object location, and thus we achieve our best prediction for the fovea size. From this iteration on there is no performance gain. This shows two things: (1) that our model can converge to its best recognition in two iterations, and (2) that the value it converges to is the limit of the performance for each fovea size, i.e, no matter how many iterations we perform the model saturates at around the second.

The value where the error of our model and the baseline match depends on if we are considering the top-1, around $f_0 = 90$, or the top-5, around $f_0 = 80$. On one hand, this means that there is also a saturation of the information needed to recognize the object. We can thus achieve the baseline performance without processing all the information contained in the original image. This happens since the artificial foveal visual system reduces the information compressed in the image, since it applies a blur filter and an exponential weighting. We can thus achieve the baseline performance without processing all the information contained in the original image.

In FS method we are not considering any weighting from previous iterations in the current prediction, thus we conclude that it is never going to reach a better performance than the baseline solution. This is because at most, in each iteration the model processes the resized resolution image (that is processed by the baseline solution), the model is only able to approach the baseline solution. However, the objective of this experiment was to show the relevance of having an active vision system capable of directing the focus of attention to the relevant regions of the image, in analogy with human vision.

Having the focus of attention directed to the object, we are able to obtain better classification performance with the the PF model, whose results we will discuss later.
Figure 7.3: Classification Error (in %) as a function of the fovea size $f_0$ over 6 model iterations. On the top are the top-1 and top-5 errors for a single first foveation point on the image center $(u_0, v_0) = (113, 113)$, whereas on the bottom are the top-1 and top-5 errors considering several foveation points spread over a grid. The shadow around the curves represent the uncertainty of the foveation points.

7.1.3 Computational Cost

As the FS method processes images that are first low-passed by a Gaussian filter and then exponentially weighted in order to create foveated images, one can expect that these images have less information and thus can be more compressed than the respective resized image (with original resolution). The information compression that arises from employing our artificial foveal visual system depends on the fovea size ($f_0$) and on the image resolution ($N \times N$). The latter we considered fixed, since all images in our experiments had to be first resized to $227 \times 227$ so that we could load them into the network. Therefore, through the jpeg algorithm [47] we computed the compression gain as a function of the fovea size. This compression gain was obtained from the division between the information encoded on the foveated images and the information encoded on the respective original images, for different quality factors. The quality factor is a weight between 1 and 100 that is associated to the image quality generated by the jpeg algorithm. Thus, the higher the quality factor, the better is the image quality.
In Figure 7.4 where these results are presented, we can verify that the compression also reaches a saturation fovea, that is consistent with the saturation point of the classification errors. This means that the reason why the model achieves the same performance as the baseline is that for those foveas the amount of information contained in the images is practically the same.

A compression gain of 50% means that the computational cost of processing two foveated images is equivalent to processing the respective original resized image. The best curve to analyse, i.e. with the higher compression, is $Q = 90$, where for a fovea size of $f_0 = 40$ we can achieve that compression gain of 50%. However, for that fovea value the model never reaches the baseline solution performance, as it is shown in Figure 7.5. Therefore, we conclude that in terms of efficiency the FS model is worse than a single feedforward classification, meaning that for the same computational cost we only achieve a classification error of 25.6%, for top-5, and 49.3%, for top-1, whereas the baseline error is 10.8%, for top-5, and 29.0%, for top-1.

### 7.2 Pyramidal Focus Method

#### 7.2.1 Number of Pyramid Levels

In Table 7.1 we present the classification error as a function of the intervals of image dimensions. We verify that in almost all cases (except one of the intervals) our PF model outperforms the baseline error. This is because, although CNNs are robust against scale changes (they are trained for each object class with different scales), the need of resizing the image before loading it in the network can produce distortions in the object. Therefore, as the PF model does not employ any resizing to the level images we have always the original object proportions, improving the recognition of the CNN.
Table 7.1: Classification Error (in %) as a function of the image dimensions for the top-5 and top-1 errors for our Pyramidal Focus (PF) method and for the baseline solution. The shaded cells point out when our model achieves better results than the baseline.

![Table 7.1](image)

Figure 7.5: Classification Error (in %) as a function of the ratio between the ground truth bounding box area and the image resolution. On the left is the top-5 error and on the right is the top-1 error.

### 7.2.2 Ratio Object Size/Image Resolution

In Figure 7.5 we verify that for smaller objects, relative to the image resolution, our PF model works better than the usual traditional single feedforward classification. It is also shown that the bigger the object is relative to the image resolution, the smaller the classification error. This is due to the fact that when the object occupies most of the image, there are usually no distracting objects, contributing to a more accurate recognition. Moreover, objects with a higher ratio can only be fully processed in the latest crops, whereas objects with lower ratios start being processed with a more fine scale in lower levels.

### 7.2.3 Computational Cost

The computational cost of this PF model is linearly proportional to the number of pyramid levels \( N_L \) that a image can generate. Nevertheless, the classification performance is better for almost all numbers of pyramid levels. This means that if we are willing to accept an increase in the computational cost, a simple way of improving the classification results for high resolution images is to use a pyramid scheme such as our model.
Chapter 8

Conclusions

In this thesis we developed a general computational framework, inspired by human vision, that is capable of performing the task of object recognition by combining Convolutional Neural Networks (CNNs) with foveal vision techniques. This framework integrates two dissociated visual attentional methodologies that can be performed sequentially.

The first one consists of a Foveal Saliency method capable of replicating the human- nonuniform way of sampling the image and the active strategy of orienting the gaze to regions of interest using attentional mechanisms. The second one is a Pyramidal Focus method that improves the object recognition after the target is centered in the visual field.

For the Foveal Saliency method we were able to conclude that it is necessary to have successive saccades, because in real scenarios, where objects can be anywhere in the image, the results show that the classification performance improves significantly from the first to the second feedforward pass. Additionally, as no performance gain was obtained in consequent iterations, it seems that the method is capable of orienting the gaze to the target in two saccades.

Furthermore, we concluded that the classification performance depends on the fovea size, which is imposed by the artificial foveal visual system. However, the performance falls short of the baseline solution (a single feedforward pass on a CNN whose input is the uniform resolution image) for smaller foveas, where the information reduction is higher. It was also observed that as the fovea size increases performance reaches a saturation point. This suggests that it is not necessary to store and transmit all the information encoded in the image in order to achieve the model's best performance. However, the information reduction obtained was not enough to demonstrate computational gains.

For the Pyramidal Focus method we were able to obtain favorable results in the classification performance where the model outperforms the baseline solution. We conclude that if we are willing to accept an increase in the computational cost, this methodology is a simple and effective way of improving the
classification accuracy after centering the objects in the field of view.

Finally, we conclude that our results are promising. Although our computational efficiency was not favorable, since the CNN used was trained to receive images with uniform resolution, we presented a biomimetic visual system that is capable of orienting the gaze to regions of interest and then performing a target classification with foveal focused attention that outperforms the baseline performance. Therefore, in the future, in order to achieve a computational gain, we intend to leverage log-polar like transformations with more compact neural network architectures trained to classify images more efficiently. As opposed to the Cartesian representation of images, Log-polar geometry resembles the structure of the retina, where we are able to have higher sampling rates on the central part of the retina - the fovea - and thus, mimicking more efficiently a foveal sensor.
Bibliography


